Towards Context-aware Image Semantic Representation via Modality Relational Reasoning and Embedding

Xuri Ge

Submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

November 2024

Abstract

Representation learning is a machine learning technique aimed at automatically discovering the most informative features of raw data, transforming it into a representation that captures the essential characteristics relevant to a specific task. Instead of relying on manual feature engineering, representation learning enables models to learn these features directly from the data, often leading to more accurate and robust performance across various artificial intelligence (AI) applications. In contexts like computer vision (CV) or natural language processing (NLP), etc., representation learning helps models understand complex, high-dimensional data by focusing on meaningful patterns and structures within the input. This approach is fundamental for enabling deep learning models to generalize effectively and adapt to diverse challenges in real-world scenarios.

Unlike other modalities such as text or speech with explicit semantic expressions, image data is inherently complex and ambiguous, requiring the extraction of more complex spatial and contextual information. In particular, factors such as the diversity and complexity of entities and corresponding relations and the ambiguity of semantic expressions make it more challenging to accurately capture and represent the features of images. In unimodal visual representation learning or multimodal joint representation learning that includes vision, visual representation learning presents unique challenges. Consequently, effective visual representation learning demands more sophisticated techniques to overcome these challenges and achieve robust performance.

This thesis is geared towards context-aware image semantic representation learning via modality relational reasoning and embedding methods. Our research aims to advance understanding and methodologies of combining contextual relationship information from a uni-visual modality or multiple joint modalities to enhance visual semantic representations. Two different tasks are studied in depth, namely unimodal facial action unit (FAU) recognition and multimodal image-sentence retrieval (ISR). We explore the effectiveness of various visual relational reasoning and embedding approaches in these two tasks. On the one hand, we explore the effectiveness of relational reasoning and information transfer between different muscle regions to improve the final visual facial representations in the FAU recognition task. We first propose a biLSTM-based implicit relational reasoning and embedding method with skipping connections (Skip-BiLSTM) and verify the effectiveness of relational reasoning for face representation. Then, we explore the encoding of explicit muscle relations into muscle features and propose a Graph Neural Network

(GNN) model with local-global interactions to further enhance the face representation capability. In our latest work, we introduce language-guided supervision for FAU recognition, which introduces language-level local and global relational reasoning for face representation learning, and we achieve better AU recognition performance in the final.

On the other hand, we explore the effectiveness of different multimodal relationship reasoning and encoding approaches to improve representation learning ability, especially for complex images, in multimodal interaction tasks. We first explore the contribution of a novel multimodal tree-structured relational reasoning and embedding to the multimodal feature representation learning in the image-sentence retrieval task. Moreover, we introduce scene recognition for semantic relational preprocessing of complex image scenes and utilize graph convolutional neural networks (GCNs) for further relational reasoning and embedding (termed relationship-aware GCNs), which further improves the multimodal feature representation capability, especially for complex visual representations. Finally, we explore the effectiveness of a semantic and spatial relation-based salient object enhancement approach within the visual modality for image-sentence retrieval during multimodal alignment optimization.

Experimental results demonstrate that visual representation learning based on relational reasoning and embedding can effectively promote the visual feature representation ability and further enhance the semantic and relational expression of fundamental visual features, whether for unimodal FAU recognition or multimodal image-sentence retrieval tasks.

Contents

List of Tables

List of Figures

Acknowledgements

During my four years of pursuing a Ph.D. at the University of Glasgow, I have been blessed with the support of countless individuals, without whom I could not have maintained my composure throughout this journey. The path of a Ph.D. is one filled with both opportunities and challenges, and my motto, "Seize every opportunity and embrace every challenge," has guided me through it all. The entire experience has been a beautiful, valuable, and unforgettable memory that will benefit me for a lifetime.

First of all, I would like to express my deepest gratitude to my supervisor, Professor Joemon M. Jose. He is a very important person in my doctoral career. He taught me academic knowledge and professional experience. He often promoted my academic achievements in the academic community and allowed me to explore in many research fields. He also gave me many opportunities, project applications, student guidance, etc. Because of him, I can complete the transition to an academic and become a qualified thesis advisor.

Next, I would like to express my special thanks to my girlfriend Songpei Xu, who is an excellent doctoral student at the University of Glasgow. It is no exaggeration to say that without her, I might not even have the opportunity to come to Glasgow. She made me seize the opportunity to study for a doctorate at the University of Glasgow. Songpei has always been a good partner of mine, supporting and encouraging me in all aspects. Our mutual help and support in life is the driving force that keeps me going. In academics, we are also partners, exploring interdisciplinary knowledge together. When my paper was rejected, she often comforted me and affirmed my work, which is extremely important to me.

In addition, I am grateful to my collaborators, including Fuxiang Tao and Junchen Fu, Xin Xin, Kaiwen, etc., as well as the related professors such as Prof. Anna Esposito, who have greatly assisted me in getting my papers published. I also want to thank Dr. Gerardo Aragon-Camarasa and Dr. Nicolas Pugeault for serving as my annual progress reviewers, providing constructive feedback that helped me complete my thesis more effectively. I am also thankful to Mrs Helen Border for the opportunity to work as a teaching assistant and for the support that helped me learn how to disseminate knowledge.

I must also extend my heartfelt thanks to my family. Without their unwavering support and encouragement, I would not have been able to complete my studies. They have always stood by me during difficult times, providing the strength I needed to persevere.

Finally, I would like to acknowledge my friends in Glasgow, both past and present. I am grateful to Qiyuan Wang, Zejian Feng, Yingying Huang, and Weiyun Wang for the joyful moments we shared. These happy times will forever remain in my memory.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work, under the supervision of Professor Joemon M. Jose.

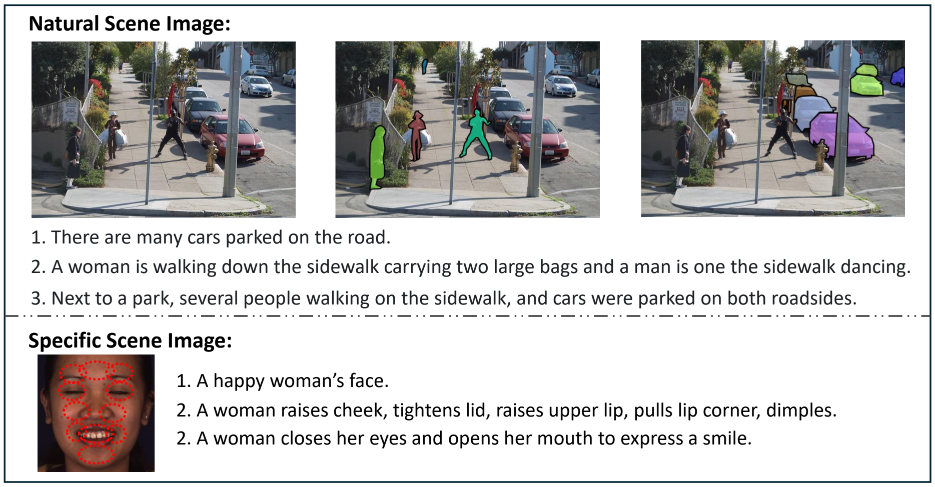
Chapter 1

Introduction

1.1 Background

With the advent of the digital information age and the rapid development of computer multimedia technology, the volume of various types of media data, including images, texts, videos, and speeches, has increased dramatically (J. Tang et al., 2015). Representation learning (RL) (Bengio, Courville, & Vincent, 2013) plays a crucial role in automatically extracting useful information from these raw data, transforming them into features for subsequent intelligent models and applications. These applications include intelligent search models, automatic recommendation systems, and intelligent diagnosis, facilitating efficient filtering and in-depth exploration of vast data collections. Therefore, effective representation of these data has become a critical area of foundational research.

Multimodal representation learning has made significant strides in the research community, with numerous advanced methods being proposed. In particular, the transformer-based models (Tao, Ge, Ma, Esposito, & Vinciarelli, 2023; Vaswani et al., 2017) are proposed to improve the text and speech representation learning abilities significantly. For example, in text representation learning, the latest models such as BERT (Devlin, Chang, Lee, & Toutanova, 2018) and GPTs (Radford, Narasimhan, Salimans, Sutskever, et al., 2018) perform well in capturing semantic nuances and contextual information. In speech representation learning, techniques utilizing large-scale models, such as Wav2Vec (Schneider, Baevski, Collobert, & Auli, 2019) and Wav2Vec2.0 (Baevski, Zhou, Mohamed, & Auli, 2020), have successfully modeled temporal dependencies and phonetic variations, achieving state-of-the-art performance in many related tasks. Similarly, the proposal of Convolutional Neural Networks (CNNs) (Krizhevsky, Sutskever, & Hinton, 2012) and Vision Transformers (Dosovitskiy et al., 2020) has greatly contributed to the advancement of visual feature representation. However, despite the progress made by CNNs and other deep learning models of image representation learning, it is more challenging to achieve similar contextual relationship modeling compared to modeling textual modal data with explicit contextual semantic expressions. The high dimensionality of image representations and the variability of contexts require visual models that can capture more complex patterns and contextual information compared to text and explicitly structured modalities. Unlike text modality, which has certain regular sequence structures and explicit semantics, images cover a wide range of changes in scale, orientation, lighting conditions, relationships between objects, etc.



1. There are many cars parked on the road.
2. A woman is walking down the sidewalk carrying two large bags and a man is one the sidewalk dancing.
3. Next to a park, several people walking on the sidewalk, and cars were parked on both roadsides.

1. A happy woman’s face.
2. A woman raises cheek, tightens lid, raises upper lip, pulls lip corner, dimples.
2. A woman closes her eyes and opens her mouth to express a smile.

Figure 1.1: Comparisons of image and corresponding text expressions.

This study focuses on image representation learning, a fundamental problem in various image and image-related applications. Image representation learning is more challenging than other modalities due to the inherent complex contextual semantic expression of images. For example, as shown in Figure 1.1, explicit sentences have unambiguous semantic representations to describe the content of images, due to the explicit syntax and clear content expression. However, images of either natural scenes or specific scenes have more diverse semantic representations which are difficult to be fully covered by partial textual descriptions, due to the complexity of inter-object relationships, foreground-background relationships and contexts. Therefore, representing the contextual semantics of images is a more challenging fundamental problem to provide better feature representation for computer vision tasks.

In recent years, basic image representation learning has witnessed substantial advancements, driven by the development of various neural network architectures. The foundations of these advances are convolutional neural networks (CNNs) and transformer-based models, which have significantly enhanced the ability to extract basic appearance features from images. Among the pioneering architectures are VGGNet (Simonyan & Zisserman, 2014), GoogleNet (Szegedy et al., 2015), ResNets (K. He, Zhang, Ren, & Sun, 2016), Vision-Transformer (Dosovitskiy et al., 2020), Swin-Transformer (Z. Liu et al., 2021), etc., each contributing uniquely to the progress in computer vision and related multi-modal tasks. These methods represent fundamental appearance features of images by training classification models on corresponding benchmarks. For example, ResNets (K. He et al., 2016) inserts shortcut connections as novel residual blocks into a depth convolutional neural network, which can be used to train fundamental classification models for multi-scene images, including ImageNet (Deng et al., 2009; Russakovsky et al., 2015) and CIFAR-10 (Krizhevsky, Hinton, et al., 2009). These residual connections allow the input original information to be directly propagated to the output layer, thereby alleviating the vanishing gradient problem caused by network depth (The vanishing gradient problem happens because, as the network depth increases, the gradients shrink as they are propagated backwards through many layers, leading to slow learning or the network failing to learn altogether.). Therefore, ResNets (K. He et al., 2016) become the mainstream image stem extraction network, facilitating the various vision-related tasks. Besides, the recent advent of transformer-based architectures, specifically the Vision Transformer (ViT) (Dosovitskiy et al., 2020) and Swin-Transformer (Z. Liu et al., 2021), etc., represent a paradigm shift in image representation learning. For example, ViT model (Dosovitskiy et al., 2020) is proposed for larger-scale image representation learning, which directly splits the image into multiple patches for subsequent self-attention-based[1] transformer encoder (Vaswani et al., 2017). These fundamental image representation backbones greatly improve image feature representation capabilities and facilitate the development of downstream vision tasks.

While fundamental image representation learning methods have achieved impressive performance in extracting basic appearance features from images, a significant gap remains in effectively capturing the semantic contextual relationships between objects or regions within images. These relationships are crucial for understanding the deeper semantic structure of visual scenes, which often goes beyond mere object recognition. For example, as shown in Figure 1.1, the important subject-verb-object structural relationship information exists in images of natural scenes, such as “Cars park on the road" and “people walk down the sidewalk", etc. Similarly, in another image domain, such as the human faces, the natural co-activation of muscles, such as the activated "Cheek" and "Upper lip" muscles, reflects inherent biomechanical connections that contribute to the expression and meaning conveyed by the face. These semantic relationships establish context-aware linkages between fine-grained regions of an image, providing richer and more precise information that can significantly enhance image representation. Such contextual semantics are essential for accurately depicting the relationships between various elements within an image, leading to a more holistic understanding of the visual content. However, conventional image representation learning models often overlook these intricate relationships, focusing primarily on global features rather than their interconnections.

To address the above issues, there is a growing need to integrate relational reasoning and embedding techniques into the existing frameworks of image representation learning. By building on the foundational representations provided by models like ResNets (K. He et al., 2016) and Swin-Transformer (Z. Liu et al., 2021), relational reasoning and embedding can enhance

[1] Self-attention is a mechanism that allows the model to learn different attention weights for the importance of different elements, such as patches of image or words of text, within a sequence by relating each element to every other element, enabling the model to capture dependencies and context over long distances, and it is widely used in applications like natural language processing and computer vision, particularly in Transformer models for tasks such as translation and image recognition.

the fine-grained representation capabilities of images. This approach involves explicitly or implicitly modeling the interactions between different regions or objects within an image, allowing the model to understand and represent the context in which these elements exist. The resulting representations would not only capture the visual appearance of individual objects but also the meaningful relationships between them, leading to more distinguishable and contextually aware image representations. Incorporating relational reasoning and embedding into image representation learning has the potential to significantly improve performance across various computer vision tasks, particularly those that rely on understanding the interactions between multiple objects or regions. For example, in unimodal vision tasks such as image recognition, and multimodal tasks such as image captioning, image-sentence retrieval, etc., the ability to model and embed semantic relationships can lead to more accurate and contextually relevant outputs. As research continues to explore this direction, the development of models that effectively combine basic feature extraction with advanced relational reasoning and embedding will be key to advancing the state-of-the-art in image representation and related applications.

In this thesis, we contend that novel modality-based relational reasoning and embedding methods can significantly enhance context-aware semantic image representation, thereby improving the performance of a range of tasks. We substantiate this by exploring two distinct yet complementary tasks across different modalities: unimodal facial action unit (FAU) recognition and multimodal image-sentence retrieval. This comprehensive investigation verifies the importance of modality-based relational reasoning and embedding for enriched image context representation. On the one hand, we focus on unimodal FAU recognition, where modality-based relational reasoning and embedding are employed to learn context-aware image representations that capture subtle facial action details, boosting recognition accuracy and robustness. On the other hand, we examine multimodal image-sentence retrieval, where capturing context-aware semantic representations becomes more complex, as effective alignment between visual and textual modalities is crucial. Here, we investigate how relational reasoning and embedding techniques enhance the model's ability to interpret and align visual semantics with corresponding text, addressing challenges specific to cross-modal understanding. These two tasks both rely on accurately understanding and embedding context-aware image semantics, although they diverge in modality interaction. In FAU recognition, a unimodal task, relational reasoning is applied within the visual modality to learn fine-grained facial representations. In contrast, image-sentence retrieval, as a multimodal task, requires aligning complex contextual representations between visual and textual modalities, making the challenge of capturing nuanced semantics even more pronounced. This cross-modal complexity highlights the need for sophisticated relational reasoning approaches, which we argue are instrumental in both tasks for capturing and leveraging contextual cues effectively. Therefore, exploring different tasks with different modalities can fully verify the importance of modality-based relational reasoning and encoding for learning and enhancing image context representations.

1.2 Thesis Statements

The overall statement of this thesis is that relational reasoning and embedding based on fundamental image representations can leverage the semantic relationship structure of image contexts to provide finer-grained and more discriminative image representations for downstream tasks. In particular, based on the original fundamental image representation, inner-modal implicit or explicit relational reasoning can promote the feature saliency of important objects by embedding richer contextual information based on constructed relationships, thereby improving the recognition ability of image representations in unimodal vision tasks (e.g., facial action unit recognition). Moreover, cross-modal guided relational reasoning and embedding, such as matched image-sentence pairs, is also an effective facilitator of contextual semantic information propagation between relevant objects, which can improve image structural representation and reduce semantic ambiguity in multimodal vision-related tasks (e.g., image-sentence retrieval).

1.3 Thesis Contributions

The summarises of our contributions are described below:

I. Visual Relational Reasoning and Embedding for Facial Action Unit Recognition

We first explore the effectiveness of multiple kinds of novel visual relational reasoning and embedding methods for structural context-aware visual representation learning to improve unimodal facial action unit (FAU) recognition. Specifically, the contributions are as follows:

1. We design *novel LSTM-based implicit relational reasoning and embedding models* (called **LGRNet** (Ge, Wan, et al., 2021) and **ALGRNet** (Ge, Jose, et al., 2023)) for face image representation learning, which can better improve face feature representation abilities by linking the accurate localized AU muscles with potential associations for unimodal FAU recognition and its application.

2. We introduce *the explicit natural prior relationships into a novel GNN-based relational reasoning and embedding network* (named **MGRR-Net** (Ge, Jose, Xu, Liu, & Han, 2024)) for FAU recognition, where the explicit prior relationships are statistics of AU co-occurrences in the dataset, improving naturally occurring linkages between muscles.

3. We propose *the external explicit relationship guidance from the joint textual AU description generation* for FAU recognition with explainable capabilities (named **VL-FAU (Ge, Fu, et al., 2024)**), where each local textual AU description contains explicit relationships among relevant facial muscle states. The joint learning of vision and language provides unique textual semantic supervision for each AU state thus enhancing feature distinguishability between AUs due to the accurate semantic relational description of AU states.

II. Visual Relational Reasoning and Embedding for Image-Sentence Retrieval

We further explore the effectiveness of novel relational reasoning and embedding models for complex visual semantic representation learning in multimodal image-sentence retrieval (ISR). Different from unimodal FAU recognition, ISR contains two modalities, i.e. image and text, where text contains clearer semantic expression than image. Based on this characteristic, we propose a series of context-aware structural visual feature representation models with relational reasoning and embedding from another modality. Outlined below are the detailed contributions made by these parts:

1. We construct *the intrinsic structure along with relations for images and sentences, i.e. visual and textual structural-semantic trees*, to improve the structured representation capabilities based on the corresponding fundamental visual/textual representations, especially complex image representation. To ensure the accuracy of semantic content in constructed structure trees, we leverage the explicit context-aware semantic and structural supervision extracted from the corresponding text sentences, which provide clear semantic expression, for entity and relation predictions. Finally, the proposed novel Structured Multi-modal Feature Embedding and Alignment model (named **SMFEA** (Ge, Chen, et al., 2021)) facilitates higher performances of image-sentence retrieval by maximizing the semantic and structural similarity of construct semantic trees between the image and corresponding sentences.

2. We propose two novel *intra- and inter-modal relational semantic enhanced interaction methods* between objects of images and words of sentences (called **CMSEI** (Ge, Chen, Xu, Tao, & Jose, 2023) and ***Hire*** **(Ge, Chen, et al. (2024) Under Review)**) for image-sentence retrieval. Our main contributions lie in exploring and integrating explicit relationships of salient objects into visual representations, and further improving cross-modal relationship reasoning in image representations guided by corresponding accurate text sentences. In particular, we leverage the pre-trained object detection model with scene graph generation to detect the inter-object relationships in images, containing the explicit spatially relative positions and semantic relationships among the object regions. We then propose a relationship-aware GCN model to enhance the object region representations with their relationships. In particular, cross-level interactive attention is proposed to model the correlations between the fragments and the instance.

3. We further propose a novel visual Semantic-Spatial Self-Highlighting Network (termed 3SHNet in (Ge, Xu, et al., 2024)) to enhance image representation during image and sentence alignment by object and spatial saliency guided by segmentation information. specifically, 3SHNet highlights the salient identification of prominent objects and their spatial locations within the visual modality, thus allowing the integration of visual semantics–spatial interactions and maintaining independence between two modalities. 3SHNet

utilizes the structured contextual visual scene information from segmentation to conduct the local (region-based) or global (grid-based) guidance and achieve accurate hybrid-level retrieval. It also guarantees efficiency and generalization of retrieval due to modality independence.

1.4 Thesis Structures

The remainder of this thesis is organised as follows:

- Chapter 1 provides the introduction of our thesis. It includes the background of our research content and the thesis statements with thesis contributions. Chapter 2 provides the background of image representation learning for corresponding vision-related tasks, including unimodal image representation learning and multimodal image representation learning. In addition, we detail the related works of image feature representation learning in unimodal facial action unit recognition and multimodal image-sentence retrieval, which are the focus of this thesis.

- Part I describes a set of visual relational reasoning and embedding models for the facial action unit (FAU) recognition task. We explore the effectiveness of different relational reasoning and feature enhancement methods for local-level and global-level facial feature representation and also study the interpretability of the predictions. Specifically, it includes the following sections:
  ALGRNet: an adaptive local global relational network in Chapter 3. It can adaptively mine the context of well-defined facial muscles by a novel skip-BiLSTM module and further enhance the visual details of facial appearance and texture with a feature fusion&refining module.
  MGRR-Net: a multi-level graph relational reasoning network in Chapter 4. MGRR-Net constructs the dynamic interactions among local-global features from multiple perspectives in a Graph Neural Network (GNN). Each layer of MGRR-Net performs a multi-level (i.e., region-level, pixel-wise, and channel-wise level) feature learning. After multiple iterations, we finally obtain richer contextual facial representations than fundamental features.
  VL-FAU: a vision-language joint learning network in Chapter 5. VL-FAU aims to reinforce AU representation capability and language interpretability through the integration of joint multimodal tasks, i.e. FAU recognition (vision task) and language generation (language task). Through this, the facial representation achieves a higher distinguishability among different AUs and different subjects. And compared with mainstream FAU recognition methods, VL-FAU can provide local- and global-level interpretability of language descriptions with the AUs' predictions.

- Part II describes a set of visual relational reasoning and embedding models for image-sentence retrieval (ISR). We explore the effectiveness of different visual relational reasoning and embedding models for the multimodal task, i.e. cross-modal retrieval (named image-sentence retrieval). It includes the following sections:
  SMFEA: structured multi-modal feature embedding and alignment for ISR in Chapter 6. SMFEA creates a novel multi-modal structured module with a shared context-aware referral tree to obtain consistent multi-modal representation in both semantics and structural spaces.
  ***Hire*:** hybrid-modal interaction with multiple relational enhancements for ISR in Chapter 7. *Hire* correlates the intra- and inter-modal semantics between objects and words with implicit and explicit relationship modeling. It can better leverage the contextual information of the object representation in images based on the inter-object relationships that match the corresponding sentence with rich contextual semantics.
  3SHNet: boosting ISR via visual semantic–spatial self-highlighting in Chapter 8. 3SHNet is proposed for high-precision, high-efficiency and high-generalization image–sentence retrieval via highlighting the salient objects and their spatial locations within the visual modality. This integration effectively combines object regions with the corresponding semantic and position layouts derived from segmentation to enhance the visual representation. And the modality-independence guarantees efficiency and generalization.

- Chapter 9 includes final conclusions and future work.

1.5 Supporting Publications

Most of the thesis generalizes and builds on the following publications accepted by various international conferences and journals, as follows:

- **Xuri Ge**, Pengcheng Wang, Hu Han, Joemon M. Jose, Zhilong Ji, Zhongqin Wu and Xiao Liu. "Local global relational network for facial action units recognition." 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2021. (Chapter 3)

- **Xuri Ge**, Joemon M. Jose, Pengcheng Wang, Arunachalam Iyer, Xiao Liu and Hu Han. "ALGRNet: Multi-Relational Adaptive Facial Action Unit Modelling for Face Representation and Relevant Recognitions." IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM), 2023. (Chapter 3)

- **Xuri Ge**, Joemon M. Jose, Songpei Xu, Xiao Liu and Hu Han. "MGRR-Net: Multi-level Graph Relational Reasoning Network for Facial Action Unit Detection." ACM Transactions on Intelligent Systems and Technology (TIST), 2024. (Chapter 4)

- **Xuri Ge**, Junchen FU, Fuhai Chen, Shan An, Nicu Sebe and Joemon M. Jose. "Towards End-to-End Explainable Facial Action Unit Recognition via Vision-Language Joint Learning." Proceedings of the 29th ACM International Conference on Multimedia (ACM MM), 2024. (Accepted) (Chapter 5)

- **Xuri Ge**, Fuhai Chen, Joemon M. Jose, Zhilong Ji Zhongqin Wu and Xiao Liu. "Structured multi-modal feature embedding and alignment for image-sentence retrieval." Proceedings of the 29th ACM International Conference on Multimedia (ACM MM), 2021. (Chapter 6)

- **Xuri Ge**, Fuhai Chen, Songpei Xu, Fuxiang Tao and Joemon M. Jose. "Cross-modal semantic enhanced interaction for image-sentence retrieval." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (WACV), 2023. (Chapter 7)

- **Xuri Ge**, Fuhai Chen, Songpei Xu, Fuxiang Tao, Jie Wang and Joemon M. Jose. "Hire: Hybrid-modal Interaction with Multiple Relational Enhancements for Image-Text Matching." ACM Transactions on Intelligent Systems and Technology (TIST). 2024. (Under Review) (Chapter 7)

- **Xuri Ge**, Songpei Xu, Fuhai Chen, Jie Wang, Guoxin Wang, Shan An and Joemon M. Jose. "3SHNet: Boosting image–sentence retrieval via visual semantic–spatial self-highlighting." Information Processing & Management (IP&M). 2024. (Chapter 8)

Chapter 2

Related Work

In this chapter, we provide an overview of image representation learning, which can be divided into two key aspects: unimodal task based image representation learning and multimodal task based image representation learning. We first introduce the common methods of image representation learning in unimodal visual tasks, such as object detection, object recognition, etc. In this thesis, we focus on developing image representation learning in unimodal facial action unit (FAU) recognition, so we fully studied the basic techniques and existing works in FAU recognition (Section 2.2). In addition, we further analyse and summarise the research on image representation learning in multimodal tasks. Moreover, we further introduce the existing works and analyse the problems for multimodal image-sentence retrieval (Section 2.3).

2.1 Image Representation Learning

2.1.1 Image Representation Learning in Unimodal Tasks

In mainstream unimodal vision tasks, e.g. object detection (Y. Chen, Li, Sakaridis, Dai, & Van Gool, 2018; Ren, He, Girshick, & Sun, 2015), object segmentation (Y. Li, Hou, Koch, Rehg, & Yuille, 2014; Milletari, Navab, & Ahmadi, 2016) and object recognition (Dosovitskiy et al., 2020), etc., there are two main streams of image representation learning methods: (i) the traditional convolutional neural networks (CNNs), such as LeNet (LeCun, Bottou, Bengio, & Haffner, 1998), GoogleNet (Szegedy et al., 2015), AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan & Zisserman, 2014), and ResNet (K. He et al., 2016), etc. (ii) the latest transformer-based networks, such as ViT (Dosovitskiy et al., 2020), DeiT (Touvron et al., 2021), and Swin-Transformer (Z. Liu et al., 2021), etc.

For the former, the CNN-based networks have a great advantage in extracting low-level features and visual structures for images. These low-level features constitute the image texture and apparent structure at the patch level, such as key points, lines, and some basic contour information. For instance, as an early CNN model, LeNet (LeCun et al., 1998) achieved efficient feature extraction and classification by introducing convolutional layers and pooling layers, and pioneered deep learning in the field of image processing. LeNet was very successful in the application of the MNIST dataset (LeCun & Cortes, n.d.) and is widely used in handwritten digit recognition systems. However, the LeNet structure (LeCun et al., 1998) is relatively simple and the model depth is shallow. So the ability to learn and extract image feature representations is generally poor, especially for complex images. More recently, ResNet (K. He et al., 2016) introduced a residual structure that protects the integrity of information by directly passing input information to the output. This simplifies learning objectives and difficulty, and to a certain extent solves the problem of information loss during information transmission, thereby enabling the design of a deeper network structure and improving the representation capabilities of feature learning. However, CNN-based networks perform local perception through convolution kernels. Although the receptive field can be expanded by increasing the number of layers and using larger convolution kernels, they still tend to extract local-level features, making it harder to directly capture global contextual information at a distance.

To address the above issues, the transformer structure (Vaswani et al., 2017) is introduced from the natural language processing (NLP) community to the computer vision (CV) field. In NLP, a transformer (Vaswani et al., 2017) captures the dependencies between features at all sequence positions through the self-attention mechanism. This means that no matter how far apart two elements are in the sequence, the transformer can directly calculate the relationships between them, thereby better understanding the global context. To employ the transformer (Vaswani et al., 2017) in computer vision tasks, the images are split into sub-region patches, where each patch is the equivalent of a word token in sentences, and then fed into the self-attention modules (Vaswani et al., 2017) for adaptive attention-aware feature extraction. For instance, the widely used vision transformer (ViT) (Dosovitskiy et al., 2020) is the earliest model that directly applies transformers to image classification and is constantly being expanded by researchers (Carion et al., 2020; Z. Liu et al., 2021; Touvron et al., 2021). The ViT model (Dosovitskiy et al., 2020) only used the encoder in the transformer structure (Vaswani et al., 2017) to extract image representations. Specifically, it directly divides the image into fixed-size patches and then obtains the patch embedding through linear projecting. After that, it performs feature relationship reasoning and aggregation based on the self-attention mechanism (Vaswani et al., 2017) in the encoder and finally uses the image classification objective function for whole network optimization. Although ViT (Dosovitskiy et al., 2020) pioneered the application of transformers (Vaswani et al., 2017) in the field of computer vision, there are still two serious problems as follows. (i) The complexity of images leads to large variations in objects of different images, which makes it challenging for vision transformers to achieve high performance in different scenarios. (ii) The image resolution is high and there are many pixels, resulting in a large amount of computation for ViT models based on global self-attention learning. Recently, Swin Transformer (Z. Liu et al., 2021), a general hierarchical vision backbone based on multi-level sliding window attention-ware calculations, is proposed to address the above issues. It introduced window attention mechanisms with patch merging to save a certain number of computing parameters in a hierarchical framework. On the one hand, Swin Transformer (Z. Liu et al., 2021) limits the calculation of attention within each window, thereby reducing the amount of calculation. On the other hand, it introduces an operation, called shifted window, to improve the information interaction between different windows, thereby achieving the same globality as ViT (Dosovitskiy et al., 2020) and maintaining the advantage of localized properties similar to CNNs.

With the rapid development of CV technology, various fundamental backbones for image representation learning have been rapidly developed and iterated, which provides a powerful fundamental feature representation capability for related visual tasks. However, the latest research (Messina, Amato, Carrara, Falchi, & Gennaro, 2018; J. Yu et al., 2020) show that these deep frameworks (Dosovitskiy et al., 2020; K. He et al., 2016; Z. Liu et al., 2021; Simonyan & Zisserman, 2014) have difficulty understanding the images with complex scenes or fine-grained scenes, especially understanding and capturing the spatial and temporal relationship between objects. *To this end, in this thesis, we focus on relational reasoning to improve the image representation ability based on these fundamental features.* It is beneficial to improve the high-level image representation capability for downstream computer vision tasks, such as image understanding (F. Chen et al., 2019; Z.-M. Chen, Wei, Wang, & Guo, 2019; Ge, Chen, Shen, & Ji, 2019; Hwang et al., 2018), facial expression recognition and analysis (Y. Chen, Chen, Wang, Wang, & Liang, 2021; Ge, Wan, et al., 2021; Z. Liu, Dong, Zhang, Wang, & Dang, 2020; T. Song, Chen, Zheng, & Ji, 2021), etc., and has received increasing attention in recent years. For example, Relational-CBIR (Messina et al., 2018) introduced the spatial relations among the visual objects for image representation learning, leading an excellent visual question-answering performances. Z.-M. Chen et al. (2019) introduced a multi-label Graph Convolutional Network (GCN) to effectively capture the correlations between visual object labels and to improve the visual representation ability for the final image object classification. These methods improve the high-level visual representation capability for complex images. Moreover, G. Li, Zhu, Zeng, Wang, and Lin (2019) incorporated an extra relational knowledge-graph for facial muscle relationship constructing and a GCN-based relationship refinement module for the enhancement of fine-grained facial region representation based on the fundamental facial representation.

Through the above extensive research, we can find that relational reasoning can further improve the fundamental visual representation capabilities in unimodal computer vision tasks. In this thesis, we explore multiple novel relational reasoning architectures for facial action unit (FAU) recognition, which is a fundamental research problem with extensive research attention due to its wide use in facial expression analysis.

2.1.2 Image Representation Learning in Multimodal Tasks

Multimodal tasks cover computer vision (CV), natural language processing (NLP) and speech processing, such as visual question-answering, image-sentence retrieval, image captioning, etc. They have recently received widespread attention in recent years due to their wide application scenarios and usage value in the real world, where image representation learning also plays an important role. Different from image representation learning in unimodal tasks, multimodal tasks usually combine multimodal representation learning via learning and embedding the relevant different modality information to improve the representation ability of each modality. It can jointly optimize multi-modality representations through multimodal tasks, rather than just improving unimodal representation. In this thesis, we focus on the exploration of image representation learning in multimodal tasks. Specifically, it can be divided into three main kinds: intra-modal interactive enhancement, cross-modal interactive enhancement and hybrid-modal interactive enhancement.

Most earlier works used independent intra-modal interactive processing of images and other modalities, such as sentences and audio, to improve the fundamental representation capabilities of each modality in multimodal tasks. For instance, VSE∞ (J. Chen et al., 2021) introduced a sample learnable GRU-based embedding pooling strategies in visual and textual branches, respectively, to further improve the contextual semantic representations of fundamental image and text features via pairwise cross-modal alignment optimization. Although simple feature recombination may not have a significant effect in unimodal vision tasks, joint learning in multimodal tasks and optimization of multimodal objective functions make it effective. Latest, CLIP (Radford et al., 2021) leveraged natural language supervision to train a better visual model on large-scale image-text pairs via contrastive learning strategy[1], where the matched image-text pairs are positive samples and others are negative samples. This powerful multimodal base model for zero-shot transfer learning greatly improves multimodal tasks and most unimodal tasks by linking images and text together during the training process, which can obtain formally independent but feature-associated multimodal representations. In addition, DVSA in (Karpathy, Joulin, & Fei-Fei, 2014) first adopted R-CNN to detect salient objects and aligned the image and sentence by evaluating the similarities between word-level textual features in sentences and region-level visual features in images. The above methods use a modality-independent dual-tower structure for feature encoding and enhancement. They have a significant efficiency guarantee i.e., they maintain independence between modalities to reduce the complexity of interactive calculations. However, cross-modal interactions are not considered, which is an effective way to further improve the modal alignment capability, especially to enhance the complex image representation capability guided by rich contextual texts.

[1] Contrastive learning in image-sentence retrieval trains a model by pulling together paired image-text representations (positive samples) and pushing apart mismatched ones (negative samples) in the embedding space, improving the alignment between visual and textual modalities.

On the other hand, fine-grained cross-modal interactive enhancement for multimodal tasks has attracted extensive research attention in multimedia and computer vision due to its promising applications, e.g., multimodal retrieval (H. Chen et al., 2020; Diao, Zhang, Ma, & Lu, 2021a; Ge, Chen, et al., 2023), cross-modal translation (R. Zhao et al., 2024; Zhou et al., 2021), multimodal emotion recognition (Goncalves & Busso, 2022; Mittal, Bhattacharya, Chandra, Bera, & Manocha, 2020; T. Shi, Ge, Jose, Pugeault, & Henderson, 2024), etc. Different from independent intra-modal interaction enhancement methods, cross-modal interaction enhancement provides more complementary or useful feature information between multiple modalities for multimodal alignment. For instance, SGRAF (Diao et al., 2021a) proposed a cross-modal attention-aware alignment method to improve the fine-grained object-word correspondences for image-text matching, which contributes to cross-modal similarity representations and meanwhile reduces the disturbance of less meaningful alignments. Cross-modal interactions are important in improving the high-level facial representation based on multi-modality features in visual-audio emotion recognition. For example, DE-III (T. Shi et al., 2024) proposed a transformer-based cross-modal enhancement module based on the optical-flow enhanced visual features and audio features to improve the cross-modal attention-aware multi-modal representation for visual-audio emotion recognition. These fine-grained cross-modal interaction enhancement methods can utilize the complementarity of multimodal features and the characteristics of multimodal feature alignment to further refine useful feature information and play an important role in enhancing the representation of complex images for different multimodal tasks.

Recently, many works (Fu et al., 2024; Gong, Liu, Rouditchenko, & Glass, 2022; Qu, Liu, Wu, Gao, & Nie, 2021; Wei, Zhang, Li, Zhang, & Wu, 2020; Q. Zhang, Lei, Zhang, & Li, 2020) have tried to combine the intra- and cross-modal interactive enhancements to highlight the high-level intra-modal semantic representation and improve the fine-grained inter-modal correspondences. For example, UAVM (T. Shi et al., 2024) proposed a unified audio-video emotion recognition framework, containing two independent-modality transformers (Vaswani et al., 2017) (audio-transformer and video-transformer) for intra-modal feature enhancements and a multi-modal shared transformer (Vaswani et al., 2017) for cross-modal information complementary. For visual-language tasks, such as image-sentence retrieval, DIME (Qu et al., 2021) proposed a dynamical learning interaction pattern through soft-path decisions in a 4-layer network, where each layer contains two intra-modal and two inter-modal interaction strategies, respectively. These hybrid-modal interactive enhancement methods improve the representation capabilities of each modality from multiple perspectives and can improve the representation of one modality through the guidance of another modality. Moreover, in the above vision-related multimodal tasks, we observe that hybrid-modal interactions are very useful in improving the representation ability of complex images. Both the feature interaction within the modality and the feature guidance between the modalities improve the high-level context representations of images.

In this thesis, we focus on the new methods of image representation learning in a multimodal task, i.e. image-sentence retrieval (ISR). We explore the novel intra-modal, inter-modal and hybrid-modal interactive enhancement methods for ISR, in particular, to improve the high-level contextual semantic representation of complex natural scene images through visual relational reasoning and embedding.

2.2 Facial Action Unit Recognition

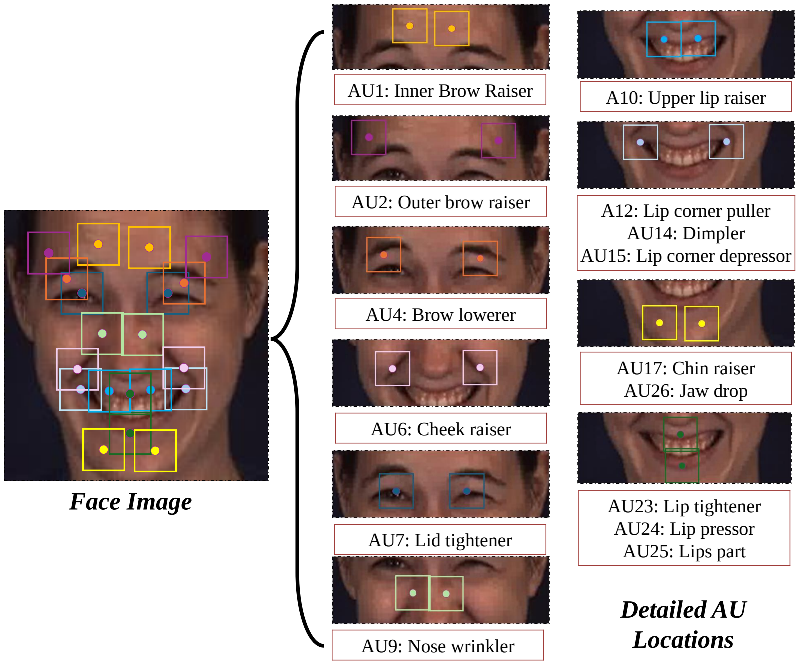


Figure 2.1: Examples of 15 popular AU region definitions with corresponding descriptions, which are widely used in the literature (Shao et al., 2021).

We will explore the effectiveness of multiple novel visual relational reasoning and embedding structures in an unimodal vision task–***Facial Action Unit (FAU) Recognition***. In this section, we introduce the related works of FAU recognition, specifically for the facial image representation learning methods. Automatic facial AU recognition is a task that detects the movement of a set of facial muscles, as shown in Figure 2.1. It has been studied for decades due to its wide potential applications in diagnosing mental health issues (Rubinow & Post, 1992; J. Shi et al., 2019), improving e-learning experiences (X. Niu et al., 2018), detecting deception (X. Li, Komulainen, Zhao, Yuen, & Pietikäinen, 2016), etc. To predict the activation states of multiple AUs, FAU recognition is treated as a multi-label classification problem. Earlier works directly predicted the AU states by a shared multi-label classifier, which however hard to focus exactly on the muscle area corresponding to each AU category. In fact, each AU has an accurate muscle area definition, and independent representation between AUs can improve the discriminative ability of AU representation. To address this issue, most existing methods adopt multiple independent AU branches for corresponding AU categories in a unified network, leading to good performances on FAU recognition. From this perspective, most existing methods can be roughly categorized into two groups: global-level facial representation learning for FAU recognition (Y. Li, Huang, & Zhao, 2021; P. Liu, Zhou, et al., 2014; X. Niu, Han, Yang, Huang, & Shan, 2019; Sankaran, Mohan, Lakshminarayana, Setlur, & Govindaraju, 2020; Shao et al., 2019) and local-level facial representation learning for FAU recognition (Chang & Wang, 2022; W. Li, Abtahi, & Zhu, 2017; C. Ma, Chen, & Yong, 2019; X. Niu, Han, Yang, et al., 2019; Shao et al., 2018, 2021; K. Zhao, Chu, De la Torre, Cohn, & Zhang, 2016).

2.2.1 Global-level Image Representation Learning for FAU Recognition

Global-level image representation learning for FAU recognition implies directly extracting global facial features through CNN-based or Transformer-based extraction networks, and performing feature refinement and enhancement for identifying AU categories based on these global-level features. Some works (Y. Li et al., 2021; P. Liu, Zhou, et al., 2014; X. Niu, Han, Yang, et al., 2019; Sankaran et al., 2020; Shao et al., 2019) predicted the activation state of each AU by directly extracting global face features via CNNs. For instance, LP-Net (X. Niu, Han, Yang, et al., 2019) using an LSTM model (Hochreiter & Schmidhuber, 1997) combines the patch features from grids of equal partition made by a global Convolutional Neural Network (CNN). Y. Li et al. (2021); Shao et al. (2019) proposed sequential or parallel channel and spatial attention learning mechanisms to explore the attention-aware global representation of each face, based on the pre-trained feature extractors (e.g. InceptionV3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016)). To explore the relationship among different AUs, Jacob and Stenger (2021) was the first method to introduce the transformer-based relational reasoning and embedding module into the FAU recognition task, achieving an excellent AU recognition performance. SRERL (G. Li et al., 2019) incorporated the AU knowledge graph as extra guidance for enhancing facial representations of all AU branches based on the fundamental global-level face representation extracted from VGG-16 (Simonyan & Zisserman, 2014).

Recent studies (X. Li, Zhang, Zhang, et al., 2023; Luo, Song, Xie, Shen, & Gunes, 2022; Z. Wang et al., 2024) employed the pre-trained transformer-based feature extraction backbones to improve the fundamental global-level image representation capabilities. For example, ME-GraphAU (Luo et al., 2022) adopted a more powerful pre-trained backbone extractor, i.e. Swin-Transformer (Z. Liu et al., 2021), to extract the whole facial representation as the global-level feature, and then constructed an AU relationship graph in a GatedGCN model for multi-branch AU representation refinement of final AU recognition. Latest MCM (X. Zhang, Yang, Wang, Li, & Yin, 2024) fused multiple modality representations, including the types of RGB, depth and thermal, extracted from the pre-trained ViT-based extractors (Dosovitskiy et al., 2020) by a novel channel and spatial masked AutoEncoder, which are optimized by a reconstruction object function to improve the face representation ability.

While progress has indeed been achieved through the utilization of global representations, the advancement remains constrained by the rudimentary nature of the coarse-grained features. Some of the above methods improve the image representation learning and relational reasoning for FAU recognition from a global perspective, which however suffers from the challenges of accurate localization of muscle areas corresponding to AUs, leading to potential interference from some irrelevant regions.

2.2.2 Local-level Image Representation Learning for FAU Recognition

To further model the accurate AU representation, local-level face representation learning is widely used to focus on the accurately corresponding patches with AU muscle regions. In the past, such issues were addressed by extracting AU-related features from regions of interest (ROIs) centered around the associated facial landmarks (Shao et al., 2018, 2021; K. Zhao, Chu, De la Torre, et al., 2016), which provide more precise muscle locations for AUs and lead to a better AU recognition performance. For instance, Jaiswal and Valstar (2016) proposed to use domain knowledge and facial geometry to pre-select a relevant image region (as a patch) for a particular AU and feed it to a convolutional and bi-directional Long Short-Term Memory (LSTM) (Graves & Schmidhuber, 2005) neural network. Recently, JAA (Shao et al., 2018) and JÂANet (Shao et al., 2021) proposed attention-based deep models to adaptively select the highly-contributing neighbouring pixels of initially predefined muscle region for joint facial AU recognition and face alignment, where the face alignment model is used to detect landmarks for specific AU region localization. However, all the above methods focused only on independent regions without considering the correlations among different AU areas to reinforce and diversify each other.

Recent works (Ge, Jose, et al., 2024; T. Song, Cui, Wang, Zheng, & Ji, 2021; T. Song, Cui, Zheng, & Ji, 2021) focus on capturing the relations among AUs for local AU representation enhancement, which can improve robustness compared to single-patch features or global face features. For instance, Z. Liu et al. (2020) applied the spectral perspective of graph convolutional network (GCN) for AU relation modelling, which also needed an additional AU correlation reference extracted from EAC-Net (W. Li, Abtahi, Zhu, & Yin, 2018). However, these methods need prior knowledge of co-occurrence probability in different datasets to construct the fixed relation matrix instead of dynamically updating for different expressions and individuals. T. Song, Cui, Zheng, and Ji (2021) proposed a performance-driven Monte Carlo Markov Chain to generate graphs from the global face, which, however, also captures some irrelevant regions affecting the performance. T. Song, Cui, Wang, et al. (2021) emphasised the learning of important local facial regions based on probabilistic graphs and obtained better facial appearance features by emphasizing important local facial regions via LSTM (Graves & Schmidhuber, 2005).

Based on the comparison of existing methods, we can see that local-level image representation learning with relational reasoning and embedding demonstrates their significant effectiveness for the unimodal FAU recognition task. Therefore, in this thesis, we further explore the effectiveness of image relational reasoning and embedding to enhance local- and global-level image representation learning for FAU recognition and propose a series of new models.

2.2.3 The Datasets and Evaluation Metrics of FAU Recognition

In Chapters 3, 4, 5 of this thesis, we investigate the proposed methods for facial action unit (AU) recognition and evaluate their effectiveness on two widely-used benchmark datasets: BP4D (X. Zhang et al., 2014) and DISFA (Mavadati, Mahoor, Bartlett, Trinh, & Cohn, 2013). These datasets are commonly employed in the field of AU recognition and provide a standard basis for comparing model performance.

BP4D (X. Zhang et al., 2014): The BP4D dataset consists of 328 facial videos collected from 41 participants (23 females and 18 males), each of whom was recorded in 8 different emotional sessions. This dataset provides a rich source of labeled data, containing approximately 140K frames with AU labels across 12 AUs, following previous research protocols (W. Li et al., 2017; Shao et al., 2021, 2019). BP4D offers a diverse range of facial expressions, making it suitable for developing robust AU recognition models.

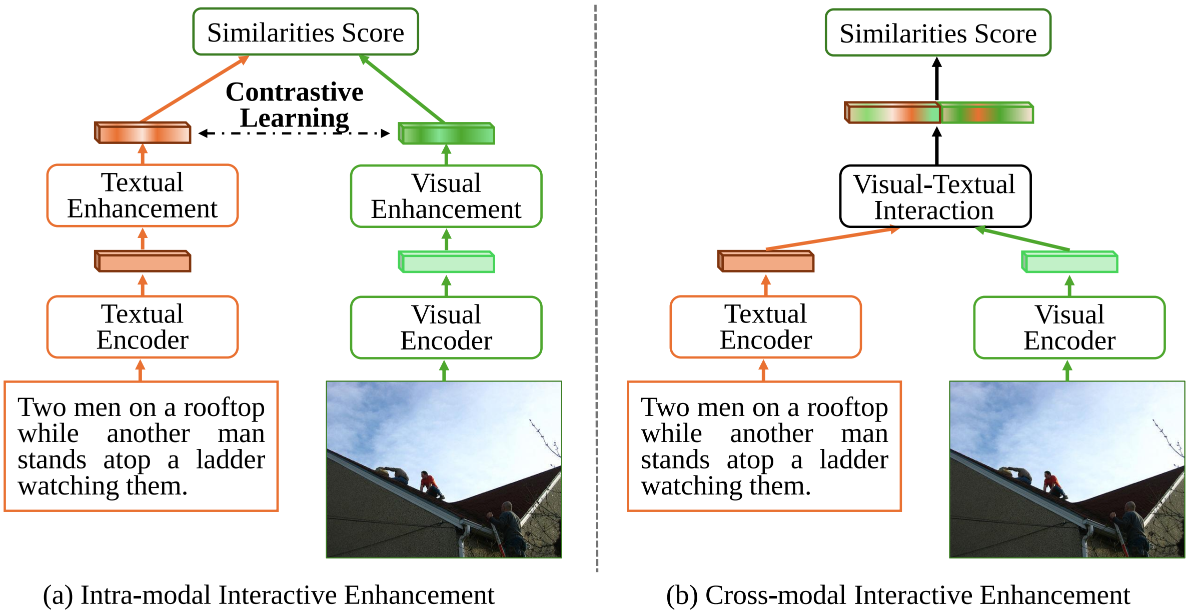
DISFA (Mavadati et al., 2013): The DISFA dataset includes videos from 27 participants (12 females and 15 males), with each participant having one video of 4,845 frames. Similar to BP4D, DISFA contains AU labels; however, the number of AUs is limited to 8, following the experimental setup used in (W. Li et al., 2017; Shao et al., 2021). Compared to BP4D, DISFA presents additional challenges due to its more variable lighting conditions and slightly different experimental protocols, making it a valuable dataset for assessing model robustness under diverse conditions.

For both datasets, we adopt a 3-fold subject-exclusive cross-validation protocol, in line with prior research (Shao et al., 2021). This protocol ensures that data from the same subject do not appear in both training and test sets, providing a rigorous evaluation of the generalizability of the model.

Evaluation Metric. To assess the performance of our proposed methods, we utilize the F1 score (%), which is a widely accepted metric for classification tasks and frequently used in AU recognition studies. The F1 score, calculated as the harmonic mean of Precision (P) and Recall (R) using the formula $\mathrm{F1} = 2\mathrm{PR}/(\mathrm{P}+\mathrm{R})$, provides a balanced measure that accounts for both false positives and false negatives, making it particularly suitable for AU recognition.

For comparison, we compute the F1 score across all AUs in the BP4D and DISFA datasets and report the average (denoted as Avg.), with percentages omitted for simplicity. Additionally, we calculate the F1 score for each facial paralysis grade in a separate evaluation on the FPara dataset, which is detailed introduced in 3.3.1. The use of the F1 score enables a consistent comparison of our approach against state-of-the-art methods, demonstrating its effectiveness across multiple levels of AU recognition and facial analysis tasks.

2.3 Image-Sentence Retrieval


Similarities Score
Contrastive Learning
Textual Enhancement
Visual Enhancement
Textual Encoder
Visual Encoder
Two men on a rooftop while another man stands atop a ladder watching them.
Similarities Score
Visual-Textual Interaction
Textual Encoder
Visual Encoder
Two men on a rooftop while another man stands atop a ladder watching them.

(a) Intra-modal Interactive Enhancement (b) Cross-modal Interactive Enhancement

Figure 2.2: Comparisons of different multimodal representation learning.

Cross-modal retrieval, *a.k.a* image-sentence retrieval, plays an important role in real-world multimedia applications, *e.g.*, queries by images in recommendation systems, or image-sentence retrieval in search engines. Image-sentence retrieval aims at retrieving the most relevant images (or sentences) given a query sentence (or image). For example, in Google search, we can directly use text descriptions to retrieve the corresponding images. It has attracted increasing research attention recently (H. Chen et al., 2020; J. Chen et al., 2021; Faghri, Fleet, Kiros, & Fidler, 2017; Frome et al., 2013; Huang, Wu, Song, & Wang, 2018; K.-H. Lee, Chen, Hua, Hu, & He, 2018; C. Liu et al., 2020; H. Liu et al., 2018; Qu et al., 2021; H. Wang et al., 2020; L. Wang, Li, & Lazebnik, 2016). The key issue of image-sentence retrieval lies in jointly learning the visual and textual representations and capturing the effective alignment to guarantee their similarity between the matched image and sentence. To this end, existing image-sentence retrieval works mainly adopt two schemas to learn the visual and textual representations, i.e. modality-independent representation retrieval (J. Chen et al., 2021; Cheng, Zhu, Qian, Wen, & Liu, 2022; Faghri et al., 2017; Frome et al., 2013; Ge, Chen, et al., 2021; Karpathy et al., 2014; L. Wang et al., 2016; Wen, Gu, & Cheng, 2020) and the cross-modal interaction retrieval (H. Chen et al., 2020; Huang et al., 2018; K.-H. Lee et al., 2018; C. Liu et al., 2019; Qu et al., 2021; K. Zhang, Mao, Wang, & Zhang, 2022). The former as shown in Figure 2.2 (a) can extract and refine image representation offline first during the inference process due to the independence between multi-modality branches and has a faster response, but it is a coarse-grained multimodal feature alignment way; the latter as shown in Figure 2.2 (b) can better align image and text modalities due to the complex interaction structure between modalities, but in actual application, it requires paired image-text input for inference which caused larger online calculation.

2.3.1 Modality-independent Representation Retrieval

Most earlier works (Fang et al., 2015; Frome et al., 2013; Kiros, Salakhutdinov, & Zemel, 2015; Mao et al., 2014; Vendrov, Kiros, Fidler, & Urtasun, 2015; L. Wang, Li, Huang, & Lazebnik, 2018; L. Wang et al., 2016) used independent processing of images and sentences within two branches to obtain a holistic representation of images and sentences. Typically, traditional approaches (Faghri et al., 2017; Frome et al., 2013; L. Wang et al., 2016; Zheng et al., 2020) model the cross-modal alignment on an instance level by directly extracting the global instance-level features of the visual and the textual modalities via Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) respectively, and estimate the visual-textual similarities based on the global features, as shown in Figure 2.2 (a). However, these approaches utilized coarse-grained global image representations to match fine-grained contextual textual representations, which makes the representational power of visual modalities insufficient to match the semantic richness of textual representations. Inspired by the detection of object regions, many studies (Karpathy & Fei-Fei, 2015; Karpathy et al., 2014) started to use the pre-extracted salient object region features to represent images. And fine-grained region-level image features and word-level text features are constructed and aligned within the modalities, respectively. For instance, DVSA in (Karpathy & Fei-Fei, 2015) first adopted R-CNN (Ren et al., 2015) to detect salient objects and inferred latent alignments between word-level textual features in sentences and region-level visual features in images. Furthermore, Huang et al. (2018) proposed an image semantic concept extraction module to predict the explicit semantic concepts for each image, improving the semantic representation ability of images when aligning with corresponding sentences. The above methods improve the representation ability within the modality, especially the visual feature enhancement, but they ignore the modeling of the relationship between objects within the image, which is more conducive to the accurate expression of image semantics.

To take full advantage of high-level objects and words semantic information, many recent methods (Diao et al., 2021a; Ge, Chen, et al., 2021; K. Li, Zhang, Li, Li, & Fu, 2022; Nam, Ha, & Kim, 2017; Y. Wu, Wang, Song, & Huang, 2019) exploited the relationships between the objects in an image and words in a sentence to help the global embedding of images and sentences, respectively. For instance, CAMERA (Qu, Liu, Cao, Nie, & Tian, 2020) introduced a new context-aware multi-view summarization network based on an adaptive gating self-attention module to exploit the intra-modal context for images and sentences to integrate the multiple visual object-level features for images and textual word-level features for sentences. VSRN (K. Li, Zhang, Li, Li, & Fu, 2019) proposed to incorporate the semantic relationship information into visual and textual features by performing object or word relationship reasoning by graph convolutional networks (GCNs), capturing key concepts of a scene. ReSG (X. Liu, He, Cheung, Xu, & Wang, 2022) introduced multiple relationship-enhanced semantic graphs for both images and sentences to further improve the high-level contextual semantics by learning their implicit locally semantic concepts of corresponding instances and their semantic relationships in a visual relationship-enhanced graph and a textual relationship-enhanced graph. In addition, they combined three types of loss functions to optimize the ranking objective functions, including individual image hard-negative triplet ranking loss, sentence hard-negative triplet ranking loss and image-sentence hard-negative triplet ranking loss. The above methods demonstrate the potential of relational reasoning and embedding to improve the representation ability of images and texts during the multimodal alignment process, but these implicit instance-based relational reasoning and embedding have certain reasoning errors and logical confusion, leading to contextual semantic ambiguity. For sentences with explicit contexts, it is more challenging to accurately express image contents with complex contextual semantics. For example, a clear semantic structure of an image – "dog –> play –> ball" – can not be reasoned and embedded as "ball –> play –> dog". To ameliorate this issue, structured relation guiding and embedding are introduced for multimodal representation learning, greatly addressing the semantic ambiguity problem of multimodal encoding, especially for complex image representation learning. For instance, (B. Shi, Ji, Lu, Niu, & Duan, 2019) integrated a scene concept graph as a rich common-sense prior knowledge information into the visual encoder, which provides strong contextual signals for semantic image understanding. However, intra-modal interactive enhancement improves the cross-modal retrieval performance via relationship interactions between the objects of image and words of texts, which, however, fails to capture the fine-grained correspondence between objects and words. This implies that fine-grained cross-modal semantic alignment has not been fully explored, especially the correspondence between entities and their attributes in image contexts and in text contexts, thus making it difficult to further improve cross-modal retrieval performance.

In this thesis, we will explore relational reasoning and embedding for modality-independent representation image-sentence retrieval, especially to address various challenges in complex images, such as structured contextual relational reasoning and salient object feature enhancement.

2.3.2 Cross-modal Interaction Retrieval

Other popular retrieval schemes exploit the fine-grained cross-modal interactions (H. Chen et al., 2020; Ji, Wang, Han, & Pang, 2019; K.-H. Lee et al., 2018; W.-H. Li, Yang, Wang, Song, & Li, 2021; Nam et al., 2017; Qu et al., 2020, 2021; Q. Zhang et al., 2020) to improve the visual-textual semantic alignments. Different from the modality-independent representation retrieval, cross-modal interactions explore the fine-grained correspondence of the cross-modality local-level representations, i.e. employing cross-modal attention mechanisms to establish connections between image regions and sentence words. For instance, SCAN model (K.-H. Lee et al., 2018) proposed a novel Stacked Cross Attention Network to construct both image-to-text attention and text-to-image attention interactions. DIME (Qu et al., 2021) adopted a multi-layer multiple cross-modality interaction framework to learn fine-grained cross-modal correspondences for representation alignments by cross-modal attention-aware region/word aggregating and region-sentence/word-image correspondence learning. The above interactive methods based on implicit cross-modal attention mechanisms optimize the correspondence between the two modalities and adaptively learn the fine-grained attention within the instance-level representations of two modalities, thereby improving multimodal alignment capabilities. However, these implicit cross-modal attention mechanisms ignore the structured contextual relationship in the representation of the two modalities, especially the visual feature representation.

Recently, structured contextual semantic feature representation learning, i.e., instance-level contextual embedding considering their relationship structures, can further improve multimodal feature representation. Especially for images with complex contextual backgrounds, it can further improve the semantic representation capability and reduce the ambiguity of content expression based on fine-grained feature representation. This challenge is more significant in image representation because textual sentences already contain rich explicit contextual structures, while images have complex and diverse contexts. Similar to the context modelling within an independent modality as above mentioned in Section 2.3.1, the implicit or explicit structured relational reasoning and embedding in cross-modal interactions is also a focus of current research (K. Li et al., 2019; C. Liu et al., 2020; S. Long, Han, Wan, & Poon, 2022; S. Wang, Wang, Yao, Shan, & Chen, 2020). This is of great help in representing multiple modality representations in a unified embedding space, especially in reducing the semantic gap between multi-modal representations, e.g. images and texts. For example, some studies (K. Li et al., 2019; C. Liu et al., 2020) employed GCNs to improve the implicit relationship interactions and integrate different item representations by a learnable graph structure. These methods only focus on the context-aware connections between the relevant objects of images and words of sentences, ignoring the explicit structure of these entities. To address this issue, recent studies (X. Dong, Zhang, Zhu, Nie, & Liu, 2022; S. Long et al., 2022; Y. Wu et al., 2023) introduced the explicit semantic relational structure to further improve the structural contextual representation ability based on the relational reasoning and embedding architectures, such as GCNs, etc. For instance, GraDual (S. Long et al., 2022) proposed two semantic graph based modules for visual and textual modalities, where the visual scene graphs and textual semantic graphs are extracted from the off-the-shelf scene graph extractors, to help the interactions between modalities. Moreover, RCRN (Y. Wu et al., 2023) leveraged the rich textual contexts to construct a language graph for structured relational reasoning and embedding of visual representations, which can encode the representations of two modalities into a joint embedding space and narrow their semantic gaps. These methods further provide contextual semantic information for multimodal interaction in image-sentence retrieval, thereby improving the accuracy of attention-aware interactions and the distinguishability of object/word features. Furthermore, some latest studies (X. Dong et al., 2022; Pei, Zhong, Yu, Wang, & Lakshmanna, 2023; S. Wang et al., 2020) tried to combine the relationship-aware structure information with a further feature enhancement, which can boost the image-sentence retrieval performances with rich context semantic representation. For instance, SGSIN (Pei et al., 2023) proposed a scene graph semantic inference network to explore the intra-modal relational semantic information between visual and textual scene graphs in image-sentence alignment and adaptively aggregates salient semantic similarities using multimodal self-attention mechanisms.

In this thesis, we explore novel frameworks, combining the implicit and explicit intra-modal semantic relational reasoning and embedding strategy with cross-modal interactions, to improve the multiple modality representations, especially for complex image scenes, in image-sentence retrieval.

2.3.3 The Datasets and Evaluation Metrics of Image-Sentence Retrieval

Dataset. To validate the effectiveness of our approach to the image-sentence retrieval task, we undertake comprehensive experimentation on two widely recognized datasets, *i.e.,* MS-COCO (T.-Y. Lin et al., 2014) and Flickr30k (Young, Lai, Hodosh, & Hockenmaier, 2014).

MS-COCO: There are over 123,000 images in MS-COCO. Following the splits of most existing methods (H. Chen et al., 2020; C. Liu et al., 2020; Qu et al., 2020), there are 113,287 images for training, 5,000 images for validation, and 5000 for validation testing. On MS-COCO, we report results on both 5-folder 1K and full 5K test sets, which are the average results of 5 folds of 1K test images and the results of full 5K test set, respectively. The full 5K test set is more challenging due to its large size.

Flickr30K: There are over 31,000 images in Flickr30K with 29, 000 images for the training, 1,000 images for the testing, and 1,014 images for the validation. Since Flickr30K is smaller in diversity than MS-COCO, we initialize the network with the well-trained model from MS-COCO for further fine-tuning instead of directly training the model on Flickr30K.

In both benchmarks, five sentences for each image are supplied, each originating from a different AMT worker.

Evaluation Metrics. Following the mainstream (J. Chen et al., 2021; Faghri et al., 2017; Ge, Chen, et al., 2023; W. Li, Su, et al., 2023), we evaluate the numerical efficacy of all approaches by the widely employed recall metrics, Recall@Q (Q=1, 5, 10), indicating the percentage of ground-truth instances successfully matched among the top Q rankings. Furthermore, we follow the standard rSum metric to calculate the summation of all six recall rates, thereby substantiating the comprehensive performance assessment.

$$rSum = \underbrace{(Recall@1 + Recall@5 + Recall@10)}_{(Image-to-Sentence)} + \underbrace{(Recall@1 + Recall@5 + Recall@10)}_{(Sentence-to-Image)}, \tag{2.1}$$

I. Visual Relational Reasoning and Embedding for Facial Action Unit Recognition

As a fundamental research problem, facial action units (AU) recognition is beneficial to facial expression recognition and analysis, and has received increasing attention in recent years. In this part, we introduce a series of visual relational reasoning and embedding approaches for unimodal facial action unit (FAU) recognition, which can significantly improve the visual representation of faces and thus model recognition performance. We propose a variety of novel relational reasoning and encoding structures, such as Adaptive Local Global Relational Network (named ALGRNet (Ge, Jose, et al., 2023; Ge, Wan, et al., 2021)) in Chapter 3, Multi-level Graph Relational Reasoning Network (named MGRR-Net (Ge, Jose, et al., 2024)) in Chapter 4 and the joint learning strategy by vision recognition and language generation (named VL-FAU (Ge, Fu, et al., 2024)) in Chapter 5, and we explored their contributions to learning facial image representation. In addition, we also verified the capabilities of the proposed models in some real-world applications, such as facial paralysis estimation, further demonstrating the generalization of these models. Extensive experiments on the widely used standard AU datasets show that the proposed approaches achieve superior performance than the state-of-the-art methods.

Chapter 3

Adaptive Local Global Relational Network

Facial action units (AUs) represent the fundamental activities of a group of muscles, exhibiting subtle changes that are useful for various face analysis tasks. One practical application in real-life situations is the automatic estimation of facial paralysis. Many existing facial action unit (FAU) recognition approaches often enhance the AU representation by combining local features from multiple independent branches, each corresponding to a different AU. However, such multi-branch combination-based methods usually neglect potential mutual assistance and exclusion relationships between AU branches or simply employ a pre-defined and fixed knowledge-graph as a prior. In addition, extracting features from pre-defined AU regions of regular shapes, i.e. fixed region locations and sizes, limits the representation ability of AUs. This is because different individuals tend to have different facial differences. To this end, we have developed a new model to recognise the activation state of AUs that deals with rich, detailed facial appearance information, such as texture, muscle status, *etc.* Specifically, a novel **A**daptive **L**ocal-**G**lobal **R**elational **N**etwork (ALGRNet) is designed to adaptively mine the context of well-defined facial muscles and enhance the visual details of facial appearance and texture, which can be flexibly adapted to facial-based tasks, *e.g.,* FAU recognition and facial paralysis estimation. ALGRNet consists of three key structures: (i) an adaptive region learning module that identifies high-potential muscle response regions, (ii) a skip-BiLSTM that models the latent relationships among local regions, enabling better correlation between multiple regional lesion muscles and texture changes, and (iii) a feature fusion&refining module that explores the complementarity between the local and global aspects of the face. We have extensively evaluated ALGRNet to demonstrate its effectiveness using two widely recognized AU benchmarks, BP4D and DISFA. Furthermore, to assess the efficacy of FAUs in subsequent applications, we have investigated their application in the identification of facial paralysis. Experimental findings obtained from a facial paralysis benchmark, meticulously gathered and annotated by medical experts, underscore the potential of utilizing identified AU attributes to estimate the severity of facial paralysis.

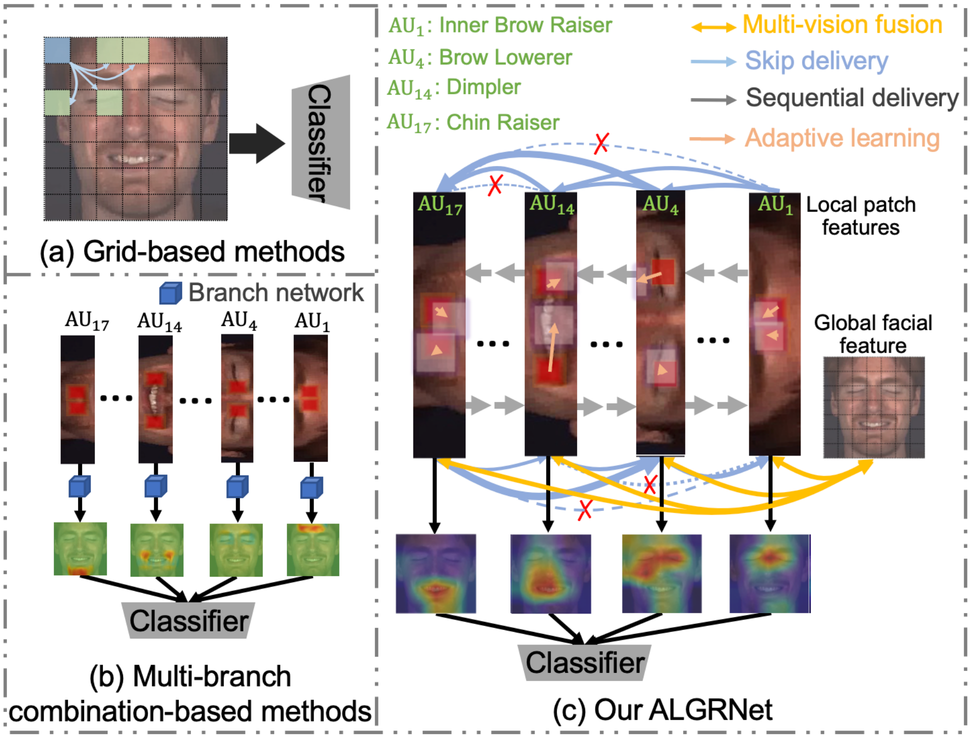


Figure 3.1: Illustration of the different schemes for AU recognition: (a) the traditional grid-based feature extraction and classification, (b) the popular multi-branch combination-based detection methods, and (c) ALGRNet method: ALGRNet, in comparison with (a) and (b), adaptively adjusts the AU areas in terms of different individuals based on detected landmarks, exploits mutual facilitation and inhibition of region-based multiple branches through a novel bidirectional structure with skipping gates and refines their irregular representations guided by the global facial feature.

3.1 Introduction

Deep learning based facial analysis tasks, such as facial recognition and facial expression recognition, aim to extract facial visual features that capture the intricate facial appearance and texture information using well-crafted Convolutional Neural Networks (CNNs). Many existing methods (Hossain, Jamal, Noshin, & Khan, 2022; Hsu, Kang, & Huang, 2018; X. Liu et al., 2020; Storey, Jiang, Keogh, Bouridane, & Li, 2019) directly extract a global facial representation from an entire face image through CNNs to perform subsequent recognition tasks. However, accurately localizing the relevant muscle regions that contribute significantly becomes challenging, thus hindering the utilization of potentially responsive muscle regions in specific facial analysis tasks, such as facial expression recognition etc. Facial expression recognition has wide potential applications in facial paralysis estimation (J. Dong et al., 2008), diagnosing mental disease (Rubinow & Post, 1992), improving e-learning experiences (X. Niu et al., 2018), detecting deception (Feldman, Jenkins, & Popoola, 1979), assisting teaching in education (Butt & Iqbal, 2011; Sathik & Jonathan, 2013), *etc.*

Recently, facial action units (AUs) have been defined to represent the precise muscle activities that capture detailed facial information. Initially, AUs are used in the Facial Action Coding System (FACS) (Ekman & Rosenberg, 1997), which can manually code nearly any anatomically possible facial expression via different groups of specific AUs. To this end, as a fundamental research problem, FAU recognition is beneficial to facial state recognition and analysis and has received increasing attention in recent years. However, AU recognition is challenging because of the difficulty in identifying the subtle facial changes caused by AUs. Looking from a biological perspective, the activation of AU corresponds to the movement of facial muscles, which inspired earlier works (Y. Li, Wang, Zhao, & Ji, 2013; Tong & Ji, 2008) to design hand-crafted features to represent the appearance of different local facial regions. However, hand-crafted features are not discriminative enough to represent the facial morphology due to their shallow natures. Hence, in recent years deep learning-based AU recognition methods have been studied to enhance the AU's feature representation.

Many existing automatic FAU recognition methods aim to enhance facial feature representation by combining local features from multiple independent branches that are related to regions of different AUs. Some grid-based deep learning frameworks (P. Liu, Han, Meng, & Tong, 2014; P. Liu, Zhou, et al., 2014) incorporate regional (patch-based) Convolutional Neural Network (CNN) features from a face with equal grids, as shown in Figure 3.1 (a). For instance, the scheme in X. Niu, Han, Yang, et al. (2019) combines local CNN features from equal partition grids by an LSTM (Hochreiter & Schmidhuber, 1997). However, dividing images into fixed grids leads to numerous issues: (i) it is difficult to focus exactly on the muscle area corresponding to each AU; (ii) ROIs for AUs with irregular shapes may not be well represented by grid-based features. To address the above issues, recent popular multi-branch combination-based methods (Shao et al., 2021, 2019; K. Zhao, Chu, De la Torre, et al., 2016) fuse global or local features from independent AU branches based on the corresponding muscle region detection, to refine the features for AUs with irregular regions, as shown in Figure 3.1 (b). For instance, a multi-branch end-to-end framework is proposed in Shao et al. (2018) to combine the features from independent branches for individual AU-related muscle regions according to some predefined attention maps based on detected landmarks.

While the multi-branch combination based AU recognition methods show their effectiveness in local AU feature fusion, there are still limitations in modeling their mutual relationship as well as the local-global context. On one hand, the multiple patches related to individual AUs, may have a strong positive or negative latent correlation in most expressions Here, if multiple AUs jointly affect the target AU category, it is defined as a positive correlation (mutual assistance), otherwise negative correlation (mutual exclusion). For example, adjacent AU2 (“Outer Brow Raiser") and AU7 (“Lid Tightener") will be activated simultaneously when scaring. And non-adjacent AU6 (“Cheek Raiser") and AU12 (“Lip Corner Puller") will be activated simultaneously when smiling. In addition, some AUs may not to be activated simultaneously, *e.g.*, we cannot simultaneously stretch our mouth (“AU20") and raise our cheek (“AU6"). Several recent works (T. Song, Cui, Zheng, & Ji, 2021; T. Song, Zheng, Song, & Cui, 2018; S. Wang, Peng, & Ji, 2018; Y. Wu & Ji, 2016) have focused on capturing the interactions among different AUs for local feature enhancement, considering the relationship of multiple facial patches to achieve better robustness than using a single patch. For instance, the studies (G. Li et al., 2019; X. Niu, Han, Shan, & Chen, 2019; T. Song, Cui, Wang, et al., 2021) incorporated AU knowledge-graph derived from statistical benchmarks to provide additional relational guidance for enhancing facial region representation. Another study (Z. Liu et al., 2020) utilized the spectral perspective of graph convolutional network (GCN) to model the AU relationship, requiring an additional AU correlation reference extracted from EAC-Net (W. Li et al., 2018). Despite the improvement of the introduced AU relationship modelling, these methods rely on the prior knowledge of AU correlation to define a fixed graph to exploit useful information from correlated AUs. Other studies (T. Song, Chen, et al., 2021; T. Song, Cui, Zheng, & Ji, 2021) employed an adaptive graph to model AU relationships based on global features, but they overlooked the local-global feature interactions that enhance the distinguishability of AUs by exploiting the complementary global details. Furthermore, these methods ignored the physiological phenomenons that adjacent related muscles often exhibit high potential correlation due to muscle linkage, and the relationship between non-adjacent related muscles may vary across different expressions and individuals. To address the above issues and inspired by the biological phenomenons, we argue that capturing the interactive information delivery between patch-based branches, such as sequential/skipping delivery of adjacent/non-adjacent related regions, is important for enhancing the representation of AU features. In addition, the global face feature provides important cues to refine the limited regular patch features, which is important for dealing with irregular muscle shapes. This is because the local AU patches may not cover the entire face, and other non-AU regions may also be activated due to muscle linkage. To the best of our knowledge, the above two key issues are left unexploited in the literature.

3.1.1 Motivation

Motivated by the aforementioned considerations, we present a novel approach, i.e. an adaptive local-global relational network (ALGRNet), utilizing a flexible and innovative end-to-end framework for automatically recognising FAU states. To accommodate facial variations across individuals, we first introduce an adaptive region learning module that detects landmarks and corresponding offsets for accurate AU muscle localization. This aspect becomes crucial due to individual facial differences. Drawing inspiration from physiological phenomena, which suggest that adjacent related muscles tend to exhibit high potential correlation. In contrast, non-adjacent corresponding muscles may display variations in different expressions and individuals. To this end, we design a skip-BiLSTM to capture implicit interactive information exchange among patch-based branches (each AU corresponds to one branch) via multiple connections, *i.e.*

Partial Facial Action Units	
Description	**Facial Muscle**
Inner brow raiser	frontalis (pars medialis)
Outer brow raiser	frontalis (pars lateralis)
Brow lowerer	corrugator supercilii
Cheek raiser	orbicularis oculi
Nose wrinkler	levator labii superioris alaeque nasi
Upper lip raiser	levator labii superioris
Lip corner puller	zygomaticus major
Lip corner depressor	triangularis
Chin raiser	mentalis
Lips part	orbicularis oris

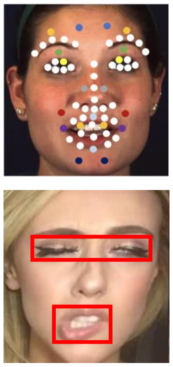


Figure 3.2: The descriptions and corresponding facial muscles of the partial facial AUs. The first row of images is the definitions of AU centers based on the detected landmarks on facial AU recognition methods (Shao et al., 2018, 2021) and the second is a facial paralysis patient with the detected bounding boxes of potential muscle lesions from (Hsu, Huang, & Kang, 2018). It is clear to observe that the AU regions can cover most areas of potential muscle lesions.

sequential and skipping connections. These connections effectively capture the potential relationships of assistance and exclusion among the sequential branches, with the ability to adjust transfer within the BiLSTM (Graves & Schmidhuber, 2005) for adjacent patches, while distant patches are connected via skipping-type gates. As each AU branch is treated independently and equally, this skip connection method minimizes information loss compared to traditional BiLSTM. Subsequently, we introduce a novel feature fusion and refining module to enhance the local features obtained from the skip-BiLSTM, guided by global grid-based features. In contrast to previous basic feature fusion methods (Ge et al., 2019; Shao et al., 2021), our gated fusion architecture in the feature fusion and refining module effectively supplements global information, including non-AU region information, for each local AU region. This is crucial because different AUs may prioritize different global information. Finally, after the relational reasoning and embedding for AU representations and feature enhancement from ALGRNet, we can obtain a better AU representation ability.

In addition, AUs offer independent interpretation and accurate localization, making them valuable for various higher-order decision-making processes beyond facial expression recognition, such as mental disease diagnosis (Rubinow & Post, 1992), depression analysis (Reed, Sayette, & Cohn, 2007), and deception detection (Feldman et al., 1979). As depicted in Figure 3.2, AUs capture fine-grained facial behaviours and possess inherent properties of symmetry and flexibility, inspiring exploration in higher-order decision-making tasks. Therefore, AU-based automatic facial paralysis recognition can leverage rich facial representation and inherent properties of AUs in symmetry and flexibility by combining well-defined muscle regions (similar to AUs), one of the facial biometrics' challenging and meaningful applications. To this end, to further explore the application ability of our proposed ALGNet, we apply it to the facial paralysis estimation. The features learned by ALGRNet can be utilized either for AU recognition through a multi-branch classification network or seamlessly integrated into a facial paralysis estimation classifier with minimal adjustments to the AU positions. We are the pioneers in investigating the effectiveness of an end-to-end deep learning-based AU recognition model for predicting the severity of facial paralysis. The intrinsic characteristics of AUs, such as their ability to represent facial features vigorously, exhibit a degree of symmetry and flexibility, making them suitable for aiding in the automated diagnosis of patients with facial palsy. Existing methods (Barrios Dell'Olio & Sra, 2021; Gaber, Taher, Abdel Wahed, Shalaby, & Gaber, 2022) have demonstrated the feasibility of this approach, but they lack robust AU recognition capabilities. We propose the novel ALGRNet, achieving exceptional performance in FAU recognition and highlighting its impressive representation abilities and its potential to be seamlessly applied to facial paralysis estimation.

3.1.2 Contribution

Our contributions can be summarized as follows:

- We propose a novel end-to-end AU recognition model (named ALGRNet) that combines adaptive local facial muscle features, their relationships, and local-global contexts to improve facial representation, which leads to improved robustness in AU recognition. In addition, we thoroughly explore utilizing a novel AU recognition model to assess the severity of facial paralysis, which provides sufficient evidence of the better applicability of both tasks on the proposed novel and general facial status analysis method. ALGRNet offers flexibility and applicability in face paralysis diagnosis, which is a pioneering effort in developing a well-designed model for this purpose.

- A new adaptive region learning module is proposed to improve the accuracy of muscles corresponding to action units and accommodate symmetric muscle region biases due to individual or lesion differences, thereby further improving the robustness and flexibility of the model.

- We propose a novel skip-BiLSTM module based on the natural physiological phenomenons to improve the representation of local AUs by modelling the mutual assistance and exclusion relationships of individual AUs via multiple inter-muscular connections, i.e. sequential and skipping. And a new gated feature fusion&refining module, filtering information that contributes to the target AU, even non-defined AU areas, is further designed to facilitate more discriminative local AU feature generation.

- The proposed ALGRNet achieves new state-of-the-art on two AU recognition benchmarks, i.e., BP4D and DISFA. Notably, we achieve superior performance to baselines on a collected facial paralysis dataset (named ***FPara***), which validates the potential of our ALGRNet for facial paralysis estimation.

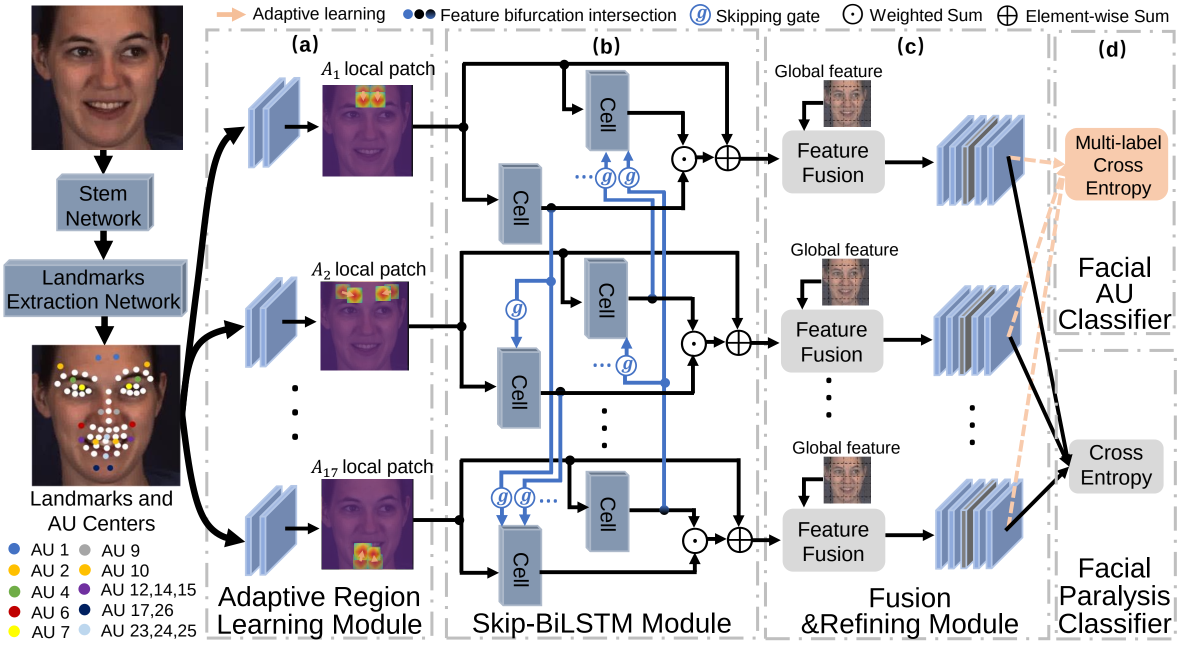


Figure 3.3: The overall architecture of the proposed ALGRNet for facial paralysis estimation or AU recognition. The location definition of salient muscle regions (as new AUs) for facial paralysis estimation is detailed in Figure 3.4 and the definition used for AU recognition is from Shao et al. (2021). We utilize a simple landmark localization network to detect the landmarks and two linear-based network to learn the offsets and scaling factor of AU centers, which are used to compute local AU patches. We then feed the features into the novel multi-branch network with a skip-BiLSTM module and a feature fusion&refining module, with each branch corresponding to an AU (each AU contains two relatively symmetrical muscle areas). The skip-BiLSTM module explores positive and negative relations among different AU branches by different information delivery options. The feature fusion&refining module in each branch helps the local AU region to fit irregular shape guided by the global grid-based feature. Finally, a multi-label binary-classifier for AU recognition is employed to predict individual AU activation probabilities and a multi-class classifier for facial paralysis estimation is used to predict the grade of facial paralysis.

3.2 The Proposed Method – ALGRNet

The framework of the proposed ALGRNet for AU recognition is presented in Figure 3.3. It is composed of four main modules, i.e., adaptive region learning module (Subsection 3.2.2) for adaptive muscle region localisation, a skip-BiLSTM module (Subsection 3.2.3) for mutual facilitation and inhibition modelling, a feature fusion&refining module (Subsection 3.2.4) for refining features of irregular muscle regions, and a multi-classifier module (Subsection 3.2.1) for predicting the grade of facial paralysis.

3.2.1 Overview of ALGRNet

Similar to Corneanu, Madadi, and Escalera (2018); Shao et al. (2018, 2021), we also employ a multi-branch network for the multi-label facial AU recognition task and we are the first to investigate the application of a multi-branch facial AU recognition network on the facial palsy estimation task. However, in contrast to previous methods, we believe that exploiting the relationship between multiple patches related to symmetrical muscle areas plays a crucial role in building a robust model. In addition, due to the diversity of expression, lesion extent, and individual characteristics, we also attempted to learn adaptive muscle region offsets and scaling factors for each AU region. To this end, we design three modules (adaptive region learning module, skip-BiLSTM module, and feature fusion&refining module) based on the established multi-branch network that can fully exploit inter-regional and local-global interactions.

We first adapt a hierarchical and multi-scale region learning network from Shao et al. (2018) as our stem network, which is used to extract the grid-based global feature and the local muscle region features. However, unlike the predefined muscle regions based on the detected landmark in Shao et al. (2018), we add two simple networks combined with the previous face alignment network, named adaptive region learning module (detailed in Section 3.2.2), to learn the offsets and scaling factors for each region adaptively. After that, local patches $A = \{A_1, A_2, ..., A_n\}$ are computed from the learned locations and their features $V = \{v_1, v_2, ..., v_n\}$ can be extracted through the stem network, where n is the numbers of selected patches. For simplicity, we do not repeat the detailed structure of the stem network here.

In our ALGRNet, we design a novel implicit relasional reasoning and embedding module, –skip-BiLSTM– (detailed in Section 3.2.3), to address the lack of sufficient delivery of local patch information between individual branches. Skip-BiLSTM can transmit information in two ways (sequential delivery and skipping delivery) in both two directions (forward and backward), in contrast to the traditional sequence spreading of LSTM. The sequential delivery of information enables full exploration of the contextual relationships between adjacent patches. The skipping delivery highlights the interaction of information from non-adjacent related patches. After skip-BiLSTM, we get a set of the local patch features $S = \{s_1, s_2, ..., s_n\}$, which are expected to have all the valuable information from adjacent and non-adjacent AU patches.

Furthermore, a novel feature fusion&refining module (detailed in Section 3.2.4) is developed to deal with irregular muscle areas, which can refine the local patches to obtain salient micro-level features for the global facial feature G. Finally, the new patch-based representations $R = \{r_1, r_2, ..., r_n\}$ for AUs are obtained by integrating local muscle features and global facial features.

This work integrates face alignment and AU recognition (or facial paralysis estimation) into an end-to-end learning model. We aim to learn all the parameters jointly by minimizing face alignment loss and facial paralysis estimation loss (or facial AU recognition loss) over the train-

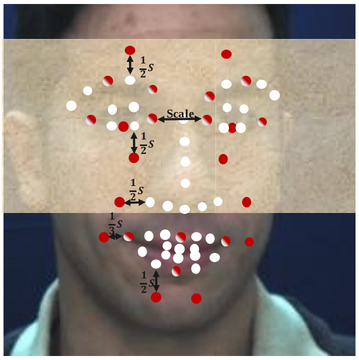


Figure 3.4: New definitions for the 12 locations of muscle centers of facial paralysis estimation, which are marked in red or mixed red. The detected landmarks are marked in white or mixed red. “Scale” denotes the distance between two inner eye corners.

ing set. The face alignment loss is defined as:

$$\mathscr{L}_{align} = \frac{1}{2d_o^2}\sum_{i=1}^{m}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2], \tag{3.1}$$

where (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ denote the ground-truth (GT) coordinate and corresponding predicted coordinate of the i-th facial landmark, and d_o is the ground-truth inter-ocular distance for normalization.

In this chapter, following Shao et al. (2021), we also regard facial AU recognition as a multi-label binary classification task. It can be formulated as a supervised classification training objective as follows,

$$\mathscr{L}_{au} = -\frac{1}{n}\sum_{i=1}^{n} w_i[p_i \log\hat{p}_i + (1 - p_i)\log(1 - \hat{p}_i)], \tag{3.2}$$

where p_i denotes the GT probability of occurrence for the i-th AU, which is 1 if occurrence and 0 otherwise, and $\hat{p}_i$ denotes the predicted probability of occurrence. w_i is the data balance weights, which is employed in Shao et al. (2018). Moreover, the loss of facial paralysis estimation is formulated as:

$$\mathscr{L}_{par} = -w_i q \mathrm{Log}(\hat{q}), \tag{3.3}$$

where q and $\hat{q}$ are the label and predicted probability for the facial paralysis grades, respectively. w_i is the data balance weights obtained by counting the different classes in the training set. Finally, we optimize the whole end-to-end network by minimizing the joint loss function $\mathscr{L} = \mathscr{L}_{au}(\text{or } \mathscr{L}_{par}) + \lambda\mathscr{L}_{align}$ over the training set.

3.2.2 Adaptive Region Learning Module

Instead of the predefined muscle regions based on landmarks, we use two simple fully connected networks to adaptively learn the offsets and scaling factors for all muscle regions respectively. Specially, we utilize an efficient landmark extraction network after the stem network to extract the landmarks $L = \{l_1, l_2, ..., l_m\}$ (m is the numbers of landmarks) similar to Shao et al. (2021), including three convolutional layers connected to a max-pooling layer. Simultaneously, two networks containing two fully-connected layers are used to get the adaptive offsets $O = \{o_1, o_2, ..., o_{2n}\}$ and scaling factors $E = \{e_1, e_2, ..., e_n\}$ respectively. According to the learned landmarks, offsets and scaling factors, local patches A are calculated. In particular, we first use the same rules in Shao et al. (2021) to get the locations of target muscle area centers based on the detected landmarks and then update the locations by adding the learned offsets. Please note that, we change the predefined muscle region centers, as shown in Figure 3.4[1], based on the detected landmarks when we apply ALGRNet on facial paralysis estimation. When defining the new salient muscle regions as new AUs (note that each AU contains two muscle regions), we maintain its roughly symmetrical distribution on faces. Different from Shao et al. (2021), we make the scaling factor E learnable rather than a fixed value, where e_i is the width ratio between the region of AU_i and whole feature map. After that, we generate an approximate Gaussian attention distribution for each AU region following Shao et al. (2018). Finally, based on the learned regions, local patch features V are extracted via the stem network.

3.2.3 Skip-BiLSTM

Figure 3.3 (b) shows the detailed structure of our implicit relational reasoning and embedding module (**skip-BiLSTM module**) for contextual and skipping relationship learning. Specifically, we extract a set of local patch features $V = \{v_1, v_2, ..., v_n\}$ from the stem network, and feed them to skip-BiLSTM. Distinct from the prior works (X. Niu, Han, Yang, et al., 2019), we regard the multiple patches as a sequence structure from top to bottom, which can transfer information by a Bi-directional LSTM based model (Graves & Schmidhuber, 2005) with our skipping-type gate. Different from traditional BiLSTM (Graves & Schmidhuber, 2005), our skip-BiLSTM can directly calculate the correlation between a target AU and all other AUs. For the t-th patch ($t > 1$), the extracted feature v_t is used to learn the weights with forward hidden states $H = \{h_1, ..., h_{t-1}\}$ by the skipping-type gates, which can determine the correlation coefficient between past AUs and current AU. And then the new states $\hat{H} = \{\hat{h_1}, ..., \hat{h}_{t-1}\}$ and v_t are fed into the t-th forward cell in the skip-BiLSTM to learn the association weights, which can promote the transfer of relevant AUs information. The above process can be formulated as:

$$\overrightarrow{h_t} = \mathrm{Cell}(\sum_{j=1}^{t-1} \overrightarrow{\hat{h}_j}, v_t), \tag{3.4}$$

$$\overrightarrow{\hat{h}_j} = \overrightarrow{h_j} f_j, \tag{3.5}$$

$$f_j = \sigma(\mathrm{GAP}(W_j(\overrightarrow{h_j} v_t))), \tag{3.6}$$

[1] Due to patient confidentiality agreements, we cannot show real patients with facial palsy. This example image is from BP4D.

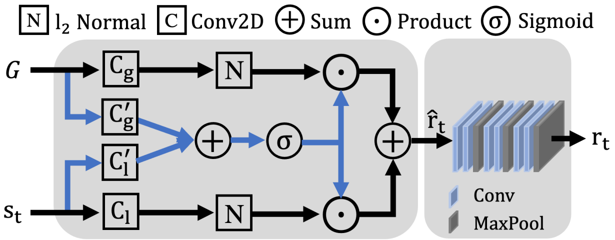


Figure 3.5: The architecture of our feature fusion&refining module is guided by the global face feature.

where Cell$(\cdot)$ indicates the basic ConvLstm cell (X. Shi et al., 2015), and GAP denotes the global average pooling operation. W_j is the parameters of mapping function, in which we used Conv2D. σ denotes sigmoid function. We obtain the t-th patch feature for backward delivery, which follows the identical forward method as:

$$\overleftarrow{h_t} = \text{Cell}(\sum_{j=t+1}^{n} \overleftarrow{h}_j, v_t), \tag{3.7}$$

In order to fully promote the information interactive among individual AUs, the final representation for each patch is computed as the average of the hidden vectors in both directions, as well as the original patch feature:

$$s_t = v_t + (\overrightarrow{h_t} + \overleftarrow{h_t})/2, \tag{3.8}$$

3.2.4 Feature Fusion&Refining Module

To exploit the useful global face feature, we design a gated fusion architecture and a refining architecture (F&R) that can selectively balance the relative importance of local patches and global face grids. We add these two architectures on each local AU branch because different local muscles may focus on different global details. The grid-based global face feature G is extracted using a simple CNN with the same structure as the face alignment network (Shao et al., 2021). As shown in Figure 3.5, after obtaining the learned t-th local patch feature, it is fused with the grid-based global feature G by the fusion architecture, which can be formulated as:

$$\alpha = \sigma(\mathrm{C}_\mathrm{g}^{'}G + \mathrm{C}_\mathrm{l}^{'}s_t), \tag{3.9}$$

$$\hat{r}_t = \alpha \odot ||C_g G||_2 \oplus (1-\alpha) \odot ||C_l s_t||_2, \tag{3.10}$$

where σ is the sigmoid function, and $||\cdot||$ denotes the l_2-normalization. $\mathrm{C}_*^{'}$ and C_* denote the Conv2D operation. $\oplus$ denotes the element-wise weighted sum of $||C_g G||_2$ and $||C_l s_t||_2$ according to the learned gate vector α.

The final local fusion feature s_t for t-th patch refined by our F&R module is shown in Figure

3.5. F&R module contains three blocks. Each block consists of two convolutional layers and a max-pooling layer. Then multi-patch features R are sent to the multi-label binary classifier to calculate the occurrence probabilities of individual AUs.

3.3 Experiments

3.3.1 Facial Paralysis Dataset

Table 3.1: Overview information of our collected facial paralysis dataset.

Grade	Normal	Low	Medium	High
Num. of Video	20	29	20	20
Num. of Frame	9049	16970	11019	10547

To evaluate the effectiveness of our ALGRNet for facial paralysis severity estimation, we exploited a facial paralysis dataset from NHS, named **FPara** (the details in Table 3.1), which consists of 89 videos of facial paralysis patients performing various types of facial paralysis exercises inline with the House-Brackmann (H-B) scale (House, 1985). Each of the videos consisted of facial paralysis patients performing a set of exercises, such as raising eyebrows, closing eyes gently, closing eyes tightly, scrunching up face and smiling, etc. Please note that all videos do not include patient rest time and remove some pauses, thus ensuring that our frame-based classification method can be fully applied. They were part of a previous study on facial paralysis with patient consent for research (O'Reilly, Soraghan, McGrenary, & He, 2010). These videos are assigned an H-B scale from 1 to 6, and 1 being normal and 6 being severest with no body movements. We then further split into four grades, such as normal (H-B score=1), low (H-B score=2), medium ($3\leq$H-B score≤ 4) and high ($5\leq$H-B score≤ 6) grades. FPara data is summarised in Table 3.1. All facial paralysis grades are evaluated using subject exclusive 3-fold cross-validation, where two folds (about 80%) are used for training and the remaining one is used for testing (about 20%).

3.3.2 Implementation Detail

Our model is trained on a single NVIDIA Tesla V100 GPU with 32 GB memory. The whole network is trained using PyTorch (Paszke et al., 2019) with the stochastic gradient descent (SGD) solver, a Nesterov momentum (Sutskever, Martens, Dahl, & Hinton, 2013) of 0.9 and a weight decay of 0.0005. The learning rate is set to 0.01 initially with a decay rate of 0.5 every 2 epochs. Maximum epoch number is set to 20. To enhance the diversity of training data, aligned faces are further randomly cropped into 176×176 and horizontally flipped. Regarding the face alignment network and stem network, we set the value of the general parameters to be the same as Shao et al. (2021). The filters for the convolutional layers in refining architecture are used

Table 3.2: Performance comparisons on F1-frame score of diverse AU recognition for 12 AUs on BP4D. All values are in %. * means the method employed pretrained model on additional dataset, such as ImageNet and VGGFace2, etc., so we do not compare. The first and second places are marked with the bold font and underline, respectively.

Method	AU Index												**Avg.**
	1	2	4	6	7	10	12	14	15	17	23	24	
DSIN	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.8	38.8	41.6	58.9
MLCR	42.4	36.9	48.1	77.5	**77.6**	83.6	85.8	61.0	43.7	**63.2**	42.1	55.6	59.8
CMS	49.1	44.1	50.3	**79.2**	74.7	80.9	88.2	63.9	44.4	60.3	41.4	51.2	60.6
LP-Net	46.9	45.3	55.6	77.1	76.7	83.8	87.2	63.3	45.3	60.5	**48.1**	54.2	61.0
JAA-Net	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
ARL	45.8	39.8	55.1	75.7	**77.2**	82.3	86.6	58.8	47.6	62.1	47.4	**55.4**	61.1
JÂA-Net	**53.8**	47.8	**58.2**	**78.5**	75.8	82.7	88.2	63.7	43.3	61.8	45.6	49.9	62.4
HMP-PS*	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	49.9	54.5	63.4
DML*	52.6	44.9	56.2	79.8	80.4	85.2	88.3	65.6	51.7	59.4	47.3	49.2	63.4
ALGRNet (Ours)	51.2	**48.2**	57.3	77.9	76.4	**84.9**	**88.2**	**64.8**	**50.8**	**62.8**	47.6	51.9	**63.5**

3×3 convolutional filters with a stride 1 and a padding 1. In this study, all of the mapping Conv2D operations are used 1×1 convolutional filters with a stride 1 and a padding 1. The dimensionality of hidden state in ConvLstm cell is set to 64. The filters for the convolutional layers in ConvLstm cell are the same as refining architecture. λ is set to 0.5 for the jointly optimizing of AU recognition (or facial paralysis estimation) and face alignment. The ground-truth annotations of 49 landmarks of training data is detected by SDM (X. Xiong & De la Torre, 2013). Different from JÂA-Net (Shao et al., 2021), we averaged the predicted probability of the local information and the integrated information as the final predicted activation probability for each AU, rather than simply using the integrated information of all the AUs. The main difference between our ALGRNet applying to facial palsy and AU recognition lies in the final classifier, where facial paralysis is four categories and AU recognition is two categories per AU.

3.3.3 Overall Performance of Facial AU recognition

We compare the proposed ALGRNet for FAU recognition with several single-image based baselines in Table 3.2 and Table 3.3, including DSIN (Corneanu et al., 2018), CMS (Sankaran, Mohan, Setlur, Govindaraju, & Fedorishin, 2019), MLCR (X. Niu, Han, Shan, & Chen, 2019), LP-Net (X. Niu, Han, Yang, et al., 2019), JAANet (Shao et al., 2018), ARL (Shao et al., 2019), and JÂA-Net (Shao et al., 2021), etc. The performances of the baselines in Table 3.3 and 3.2 are their reported results. For a more comprehensive display, we also show methods (marked with $*$) (T. Song, Chen, et al., 2021; T. Song, Cui, Zheng, & Ji, 2021; S. Wang, Chang, & Wang, 2021) that use additional data, such as ImageNet (Deng et al., 2009) and VGGFace2 (Cao et al., 2018), etc., for pre-training. Due to the fact that our stem network only consists of a few simple convolutional layers, even if we pre-trained on additional datasets, it is unfair compared to pre-training on deeper feature extraction networks, such as ResNet50 (K. He et al., 2016). In

Table 3.3: Performance comparisons on F1-frame score of diverse AU recognition for 8 AUs on DISFA. All values are in %. The first and second places are marked with the bold font and underline, respectively.

Method	AU Index								**Avg.**
	1	2	4	6	9	12	25	26	
DSIN	42.4	39.0	68.4	28.6	46.8	70.8	90.4	42.2	53.6
CMS	40.2	44.3	53.2	**57.1**	50.3	73.5	81.1	59.7	57.4
LP-Net	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
JAA-Net	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
ARL	43.9	42.1	63.6	41.8	40.0	**76.2**	**95.2**	66.8	58.7
JÂA-Net	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
HMP-PS*	21.8	48.5	53.6	56.0	58.7	57.4	55.9	56.9	61.0
DML*	62.9	65.8	71.3	51.4	45.9	76.0	92.1	50.2	64.4
ALGRNet (Ours)	**63.8**	**65.4**	**73.6**	44.5	**54.1**	74.0	94.7	**69.9**	**67.5**

Table 3.4: Performance comparisons on F1-frame score (in %) of diverse facial paralysis estimation for 4 grades on FPara.

Method	Facial Paralysis Grades				**Avg.**
	Normal	Low	Medium	High	
ResNet18	99.8	50.7	47.7	67.9	66.5
ResNet50	99.9	53.9	54.7	71.4	70.0
Transformer-based	100	**63.0**	58.6	68.7	72.6
JÂA-Net	100	55.9	62.8	72.5	72.8
ALGRNet (Ours)	**100**	55.9	**72.1**	**73.2**	**75.4**

fact, our results are still excellent compared with them, which demonstrates the superiority and effectiveness of our proposed learning scheme. We omit the need for additional modal inputs and non-frame-based models (P. Liu, Zhang, Yang, & Yin, 2019; H. Yang, Wang, & Yin, 2020).

Quantitative comparison on BP4D: We report the performance comparisons between our ALGRNet and baselines on BP4D in Table 3.2. As it can be observed, our ALGRNet significantly outperforms all the other methods in terms of F1-frame score and achieves the first and second places for most of the 12 AUs annotated in BP4D. Our ALGRNet achieves 1.1% higher average F1-frame score compared with the latest state-of-the-art method JÂA-Net.

Quantitative comparison on DISFA: We also report the performance of our proposed ALGRNet on DISFA. Table 3.3 shows the performance of our ALGRNet is the best in terms of average F1 score compared with all baselines. And our approach significantly outperforms all other methods for most of the 8 AUs annotated in DISFA. Compared with the existing end-to-end feature learning and multi-label classification methods DSIN (Corneanu et al., 2018) and ARL (Shao et al., 2019), the average F1-frame score of our proposed ALGRNet get 13.9% and 8.8% higher, respectively. Moreover, compared with the multi-branch combination-based state-of-the-art method JÂANet (Shao et al., 2021), our ALGRNet achieves 4.0% improvements in terms of average F1-frame score.

3.3.4 Overall Performance of Facial Paralysis Estimation

Different from facial AU recognition, the existing deep-learning based facial paralysis estimation methods are rare, so we apply currently popular deep learning classification methods, such as the ResNet (K. He et al., 2016) and Transformer(Vaswani et al., 2017), on our collected facial paralysis dataset (FPara). Besides, we also compare it with the state-of-the-art AU recognition approach, JÂA-Net (Shao et al., 2021). Specially, we evaluate the following methods:

- ResNet18 and ResNet50 (K. He et al., 2016): These methods use different depth layers based on ResNet to model the input face images, which are similar to A. Song, Wu, Ding, Hu, and Di (2018).

- Transformer-based method (Valanarasu, Oza, Hacihaliloglu, & Patel, 2021): This baseline is motivated from self-attention and uses the Transformer (Vaswani et al., 2017) architecture. The output of the Transformer-based encoder (Valanarasu et al., 2021) is treated as the latent representation for the input of the multi-label AU classifier.

- JÂA-Net (Shao et al., 2021): This is a recently proposed multi-branch combination-based AU recognition method, which can extract precise local muscle features thanks to a joint facial alignment network.

The first and second places are marked with the bold font and "_", respectively.

Quantitative comparison on the collected FPara: Facial paralysis estimation results by different methods on our FPara are shown in Table 3.4. It has been shown that our ALGRNet outperforms all its competitors with impressive margins. Specifically, JÂANet is the latest state-of-the-art method which also joint AU recognition and face alignment into an end-to-end multi-label multi-branch network. Compared to the facial paralysis estimation model based on the state-of-the-art AU recognition method JÂA-Net (Shao et al., 2021), our ALGRNet achieves 2.6% improvements in terms of average F1 score. The main reason lies in our ALGRNet overcomes the problem of non-transferable information between branches in the JÂA-Net and adaptively adjusts the muscle regions corresponding to the AUs. Moreover, the average F1 score of our ALGRNet get 2.8% higher compared to the currently popular Transformer-based approach (Valanarasu et al., 2021).

The eventual experimental results of our ALGRNet demonstrate that it is successful in boosting AU detection accuracy on BP4D and DISFA and having high generalization ability on our new facial paralysis dataset.

3.3.5 Ablation Studies

To fully examine the impact of our proposed adaptive region learning module, skip-BiLSTM module and feature fusion&refining module, we conduct detailed ablative studies to compare different variants of ALGRNet for AU detection on DISFA and facial paralysis estimation on FPara.

Table 3.5: Ablation study of ALGRNet for 8 AUs on DISFA and for 4 grades on FPara. All values are in %.

Methods	Setting			AU Index								**Avg.**	Paralysis Grade				**Avg.**
	S-B	F&R	Ada	1	2	4	6	9	12	25	26		Nor.	Low	Med.	Hig.	
w/o full				47.1	61.1	66.3	44.7	52.2	74.9	92.2	66.2	63.1	99.8	54.6	64.1	70.9	72.3
w/o F&R	√			62.6	64.2	72.4	42.3	49.9	**76.1**	93.5	72.6	66.7	100	54.7	66.2	72.6	73.3
w/o S-B		√		58.7	65.2	73.5	43.9	53.5	72.2	94.1	64.7	65.7	99.9	55.1	65.3	71.3	72.9
w/ Bi		√		61.1	58.4	70.9	45.5	47.9	74.9	92.5	70.8	65.2	99.8	57.1	67.3	72.8	74.3
w/o Ada	√	√		62.6	64.4	72.5	**46.6**	48.8	75.7	94.4	**73.0**	67.3	100	**57.8**	68.7	72.0	74.6
ALGRNet	√	√	√	**63.8**	**65.4**	**73.6**	44.5	**54.1**	74.0	**94.7**	69.9	**67.5**	**100**	55.9	**72.1**	**73.2**	**75.4**

Effects of adaptive region learning module

To cancel out the adaptive region learning (indicated w/o Ada), we follow the same experiment setting as Shao et al. (2021) (It means each scaling factor *e* is set to 0.14.) to predefined muscle region based on the detected landmarks for each AU/PAU. In Table 3.5, ALGRNet decreases its F1 score to 74.6% and 67.3% on the collected FPara and DISFA respectively. Our whole ALGRNet may show slightly lower accuracy than the method without using adaptive region learning. This is because of the severe data imbalance issues of individual classes. After using adaptive region learning, our method may sacrifice the accuracies of a few AUs (or grades) while improving the overall accuracy.

Effects of skip-BiLSTM

In Table 3.5, when the skip-BiLSTM module is removed (indicated by w/o S-B), ALGRNet (without adaptive region learning module) shows an absolute decrease of 1.7% and 1.6% in the average F1 score for facial paralysis estimation on FPara and AU detection on DISFA, respectively. In addition, to explicitly validate the effectiveness of our skipping operation, we use the basic BiLSTM (Graves & Schmidhuber, 2005) (indicated by w/ Bi) instead of skip-BiLSTM for information sequential transfer across different branches in the ALGRNet (also with Fusion&Refining module), ALGRNet obtains lower average F1 scores of 74.3% and 65.2% on FPara and DISFA, respectively. The performance reduction verifies that roughly defining the relationships between branches related to AU symmetry regions from top to bottom may not be the best way to model the real relationships between AUs. Notably, skipping operation can significantly improve performance, suggesting that our skip-type gates play an important role in our model.

Table 3.6: Mean error (lower is better) results of different face alignment models on BP4D, DISFA and FPara. All values are in %.

Methods	BP4D	DISFA	FPara
JÂA-Net	3.80	3.87	**5.15**
ALGRNet	**3.78**	**3.29**	5.18

Effects of feature fusion&refining module

Without the fusion&refining module (indicated by w/o F&R in Table 3.5 for facial paralysis estimation and AU detection, respectively), we directly conduct classification over the output of skip-BiLSTM. The average F1 score drops from 74.6% to 73.3% on FPara and from 67.3% to 66.7% on DISFA, due to the lack of supplementary information from the global face for each patch. In addition, we simply fuse the global features to the local AU features following Shao et al. (2021), due to the lack of effective information filtering, the average F1 score drops from 74.6% to 73.9% on FPara and from 67.3% to 66.9% on DISFA. This suggests that the refined local region features from the proposed fusion&refining module, guided by the grid-based global features, significantly contribute to our model.

Finally, after simultaneously removing all the proposed adaptive region learning module, skip-BiLSTM and fusion&refining module (marked by w/o full in Table 3.5), a significant performance degradation in facial paralysis estimation and AU detection can be observed, *i.e.*, a 3.1% drop on FPara and a 4.4% drop on DISFA in terms of average F1 score. This sufficiently demonstrates that the potential mutual assistance and exclusion relationships between the adaptive AU patches, complemented by the global facial features, can significantly improve the performance of facial AU detection. Furthermore, for facial paralysis estimation, the adaptive local-global interaction based on symmetrical muscles (PAUs) greatly enhances the semantic representation of facial context, obtaining accurate semantic information from potential lesion regions and contextual relational help from the global face.

3.3.6 Results for Face Alignment

We integrate face alignment and facial paralysis estimation into our end-to-end ALGRNet, which can benefit each other as they are coherently related. For example, detected landmarks can help the model focus on the exact location of regions with high probability of muscle lesions as PAU patches. As shown in Table 3.6, compared with baseline method JÂA-Net (Shao et al., 2021), our ALGRNet performs comparably to baseline on FPara and better on BP4D and DISFA. The robustness of the adaptive region learning module allows our ALGRNet to outperform JÂA-Net in facial paralysis estimation and AU detection, even if sometimes with slightly lower landmark detection accuracy.

3.3.7 Visualization of Results

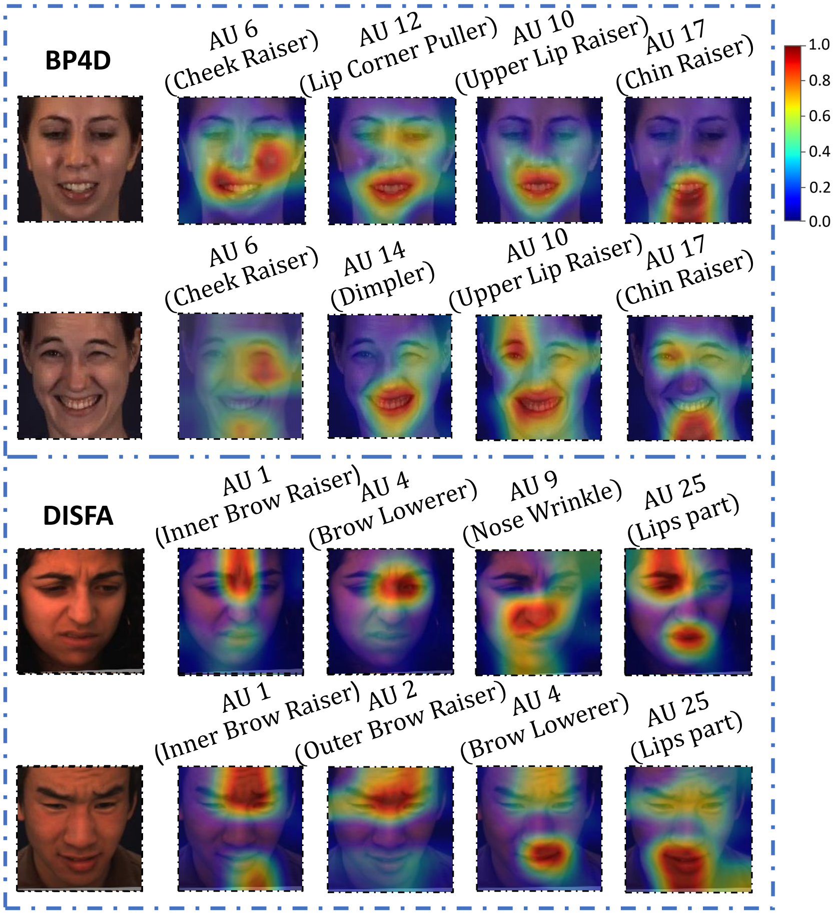


Figure 3.6: Class activation maps that show the discriminative regions for different AUs in terms of different expressions and individuals on DISFA and BP4D datasets.

For a clearer and adequate display, four examples of the learned class activation maps of ALGRNet (the outputs of F&R module) from two different datasets are given in Figure 3.6, two of which are from BP4D and two are from DISFA, containing visualization results of different genders with different AU categories. Through the learning of ALGRNet, not only the concerned AU regions can be accurately located, but also the positive (in red) or negative (in blue) correlation with other AU areas can be established and other details of the global face can be supplemented. This obviously improves the flaws of the excessive localisation of JÂA-Net (Shao et al., 2021) and the negative influence of unrelated regions of ARL (Shao et al., 2018). In addition, it also adapts well to irregular muscle areas for different AUs. The heatmaps for the same AU category in the different examples are broadly consistent but also vary slightly by the individual, demonstrating that our ALGRNet can learn certain rules across different datasets and adaptively adjust to different samples. In addition, Figure 3.7 shows four examples of the learned class activation maps of ALGRNet (the input of classifier) corresponding to different patients. It suggests that our method can mine the relationship between related muscle regions while accurately locating the muscle regions where the underlying disease occurs, thus enhancing the contextual detail of the face representation.

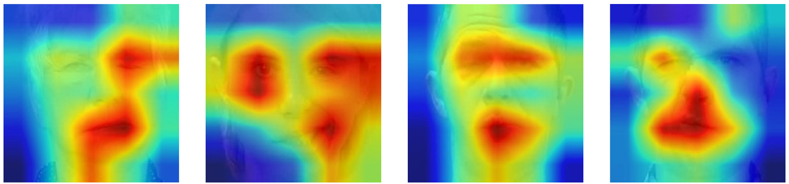


Figure 3.7: Class activation maps that show the discriminative regions for different patients with different expressions on FPara datasets. Due to patient confidentiality agreements, we process patient images with strong transparency.

3.4 Discussion and Limitation

ALGRNet, an advanced facial representation stem network based on adaptive facial action units with multiple relational reasoning and embedding modules, offers several notable advantages. Firstly, ALGRNet demonstrates outstanding performance in AU detection, showcasing its remarkable facial representation capabilities. This enables its application in a wide range of higher-order decision-making processes. Secondly, we have demonstrated that the features learned by ALGRNet can be effectively utilized either for AU recognition through a multi-branch classification network or seamlessly integrated into a facial paralysis estimation classifier with minimal adjustments to the AU positions. Through identifying symmetrical AUs, we have developed an effective facial palsy detector. This pioneering work explores the effectiveness of an end-to-end deep learning-based AU detection model in predicting the severity of facial paralysis.

Limitations. ALGRNet is based on automatic mining of implicit inter-muscle relationships, ignoring the inherent relationship modeling that exists due to the natural linkage of facial muscles. In addition, modeling the relationship between AUs through a variant of traditional Bi-LSTM requires high complexity because each AU uses a complex LSTM cell and performs bidirectional and skip propagation.

3.5 Conclusion

This chapter introduces ALGRNet, an innovative adaptive local-global relational network designed for detecting facial action unit states and also in estimating the severity of facial paralysis through AU detection. ALGRNet capitalizes on the precision and adaptability of muscle region localization and leverages the comprehensive facial semantic feature representation offered by AU detection models. By harnessing the interactive relationships and interplay between adaptive and symmetrical muscle regions, ALGRNet effectively captures the dynamic nature of these regions across various expressions and individual characteristics. ALGRNet employs a skip-BiLSTM mechanism to facilitate efficient information embedding and exchange based on implicit relational reasoning, allowing for seamless transfer of local muscle features while modelling the potential assistance and exclusion relationships among AU branches. Furthermore, a novel feature fusion and refining module is incorporated into each branch, promoting the synergy between local features and grid-based global features while accommodating irregular muscle regions. We substantiate the effectiveness of our approach by conducting comprehensive experiments on two widely utilized benchmarks for AU detection. Furthermore, we have successfully applied AU detection to the detection of facial paralysis by identifying symmetrical Action Units (PAUs). Our experiments on a benchmark specifically designed for facial paralysis estimation highlighted the remarkable superiority of our method in accurately estimating the severity of facial paralysis.

Chapter 4

Multi-level Graph Relational Reasoning Network

Many methods that perform well on automatic FAU recognition primarily focus on modelling various AU relations between corresponding local muscle areas or mining global attention-aware facial features; however, they neglect the dynamic interactions among local-global features. Although the ALGRNet presented in Chapter 3 also somewhat constructs local and global relational interactions, it still has obvious limitations. One is that the relationship modeling of AUs is performed using complex LSTM-based cells without considering the natural muscle linkage, and the other is that the global facial features are only crudely integrated into each AU representation from a single perspective. We argue that encoding AU features just from one perspective may not capture the rich contextual information between regional and global face features, as well as the detailed variability across AUs, because of the diversity in expression and individual characteristics. In this chapter, we propose a novel Multi-level Graph Relational Reasoning Network (termed *MGRR-Net*) for facial AU recognition. Each layer of MGRR-Net performs a multi-level (*i.e.*, region-level, pixel-wise and channel-wise level) feature learning. On the one hand, the region-level feature learning from the local face patch features via graph neural network can encode the correlation across different AUs with prior knowledge initiation. On the other hand, pixel-wise and channel-wise feature learning via graph attention networks (GAT) enhance the discrimination ability of AU features by adaptively recalibrating feature responses of pixels and channels from global face features. The hierarchical fusion strategy combines features from the three levels with gated fusion cells to improve AU discriminative ability. Extensive experiments on DISFA and BP4D AU datasets show that the proposed approach achieves superior performance than the state-of-the-art methods.

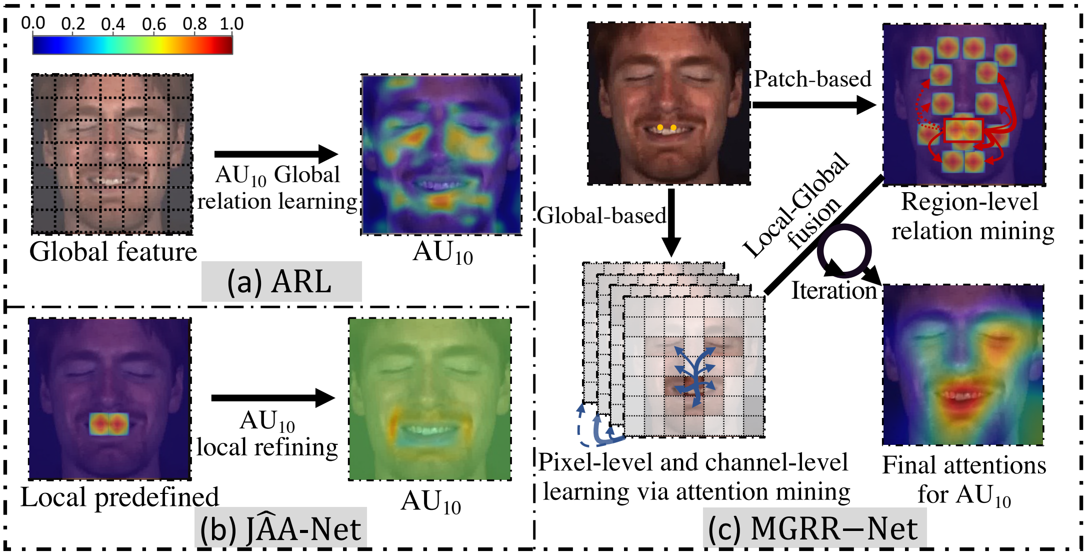


Figure 4.1: Comparisons between the proposed method and two state-of-the-art methods in AU feature learning, and the corresponding visualized activation maps for AU10 (Upper Lip Raiser / Levator labii superioris). (a) ARL (Shao et al., 2019) performs global feature learning, (b) JÂANet (Shao et al., 2021) learns from predefined local regions based on the landmarks, and (c) multi-level feature learning from both local regions and global face regions (best viewed in color).

4.1 Introduction

The key issue of facial AU recognition lies in obtaining a better facial appearance representation by improving the feature discriminative ability of local AUs and global features from the whole face. On the one hand, region-level dynamic AU relevance mining based on facial landmarks accurately detects the corresponding muscles and flexibly models the relevance among muscle regions. It is different from the existing methods focusing on extracting features for a single AU region (Shao et al., 2018, 2021) or a predefined fixed graph representing prior knowledge (G. Li et al., 2019). Although there have been many methods (Z. Liu et al., 2020; T. Song, Chen, et al., 2021; T. Song, Cui, Wang, et al., 2021; T. Song, Cui, Zheng, & Ji, 2021) on modelling relationships between AU regions, this issue still needs to be addressed effectively. On the other hand, due to the differences in expressions, postures and individuals, fully learning the responses of the target AU in the global face can better capture the contextual differences between different AUs and complement more semantic details from the global face. For instance, Shao et al. (2018, 2021) simply concatenated the global features extracted from the whole face via CNNs with all local AU features for input into the final classifier. However, it is difficult for all these methods to learn the sensitivity of the target AU within the global face and supplement enough semantic details from the global face representation in terms of different expressions, postures and individuals. To the best of our knowledge, how to better respond globally to each AU remains unexploited in existing works (G. Li et al., 2019; Z. Liu et al., 2020; Luo et al., 2022; Shao et al., 2021).

4.1.1 Motivation

Motivated by the above insights, we propose a novel technique for facial AU recognition called MGRR-Net. Our main innovations lie in three aspects, as shown in Figure 4.1 (c). Firstly, we introduce a dynamic graph to model and reason the relationship between a target AU and other AUs. The region-level AU features (as nodes) can accurately locate the corresponding muscles. Secondly, we supplement each AU with different levels (channel- and pixel-level) of attention-aware details from global features, which greatly improves the distinction between AUs. Finally, we iteratively refine the AU features of the proposed multi-level local-global relational reasoning layer, which makes them more robust and more interpretable. Different from the existing GNN-based approaches (G. Li et al., 2019; Z. Liu et al., 2020; Luo et al., 2022; X. Niu, Han, Shan, & Chen, 2019; T. Song, Chen, et al., 2021; T. Song, Cui, Zheng, & Ji, 2021) that utilize complex GCNs (Kipf & Welling, 2016) to enhance the distinguishability of AUs by constructing AU relationships, however, we supplement each AU with different perspectives (channel- and pixel-level) of attention-aware details from global features, making it possible to achieve the same purpose in a basic GNN and solve a certain over-smoothing issue. In particular, we extract the global features by multi-layer CNNs and precise AU region features based on the detected facial landmarks, which serve as the inputs of each multi-level relational reasoning layer. A simple region-level AU graph is constructed to represent the relationships by the adjacency matrix (as edges) among AU regions (as nodes), initialized by prior knowledge and iteratively updated. We propose a method to learn channel- and pixel-wise semantic relations for different AUs at the same time by processing them in two separate efficient and effective multi-head graph attention networks (MH-GATs) (Veličković et al., 2018). Through this, we model the complementary channel- and pixel-level global details. After these local and global relation-oriented modules, a hierarchical gated fusion strategy helps to select more useful information for the final AU representation in terms of different individuals.

4.1.2 Contribution

The contributions of this work are as follows:

- We propose a novel end-to-end iterative reasoning and training scheme for facial AU recognition, which leverages the complementary multi-level local-global feature relationships to improve the robustness and discrimination for AU recognition;
- We construct a region-level AU graph with the prior knowledge initialization and dynamically reason the correlated relationship of individual AUs, thereby improving the robustness of AU recognition;

- We propose a GAT-based model to improve the discrimination of each local AU patch by supplementing multiple levels of global features;

- The proposed MGRR-Net outperforms the state-of-the-art approaches for AU recognition on two widely used benchmarks, *i.e.,* BP4D and DISFA, without any external data or pre-trained models.

4.2 Related Work on GNNs for FAU Recognition

Integrating graphs with deep neural networks have recently been an emerging topic in deep learning research. GCNs have been widely used in many applications such as human action recognition (Yan, Xiong, & Lin, 2018), emotion recognition (T. Song et al., 2018), social relationship understanding (Z. Wang et al., 2018) and object parsing (Liang, Shen, Feng, Lin, & Yan, 2016), which can improve robustness compared to single-patch features or global face features. SRERL (G. Li et al., 2019) proposed to apply a gated graph neural network (GGNN) with the guidance of AU knowledge-graph on facial AU recognition. MLCR (X. Niu, Han, Shan, & Chen, 2019) embedded the relations among AUs through a predefined GCN to enhance the local semantic representation. AU-GCN (Z. Liu et al., 2020) applied the spectral perspective of graph convolutional network (GCN) for AU relation modelling, which also needed an additional AU correlation reference extracted from EAC-Net (W. Li et al., 2018). However, these AU recognition methods require a fixed predefined graph from different datasets when applying GGNN or GCN. T. Song, Chen, et al. (2021); T. Song, Cui, Zheng, and Ji (2021) applied an adaptive graph to model the relationships between AUs based on global features, ignoring local-global interactions. Moreover, these approaches usually ignored or simply fused the local and global information for each AU without considering the importance (important and non-important) of features. ME-GraphAU (Luo et al., 2022) learned a unique AU graph to explicitly describe the relationship between AUs, where each AU is simply represented from the same full face representation via a fully connected layer and a global average pooling. Although this method explores the global face features to some extent, it relies on strong global feature extraction benchmarks and lacks accurate localization of local muscle areas and discriminable feature representation via local-global interaction. Recently, a novel graph attention network with multi-head (MH-GAT) leverages masked self-attentional layers to operate on graph-structured data, which shows high computational efficiency.

As far as we know, there has been no work attempting to obtain better feature representation by multiple interactions between local AU regions and the global face, which we believe is an important cue to boost facial AU recognition performance with more fine-grained information and higher diversity of expressions. To this end, our proposed MGRR-Net automatically models the relevance among the facial AU regions by a dynamic matrix as a graph and supplements each AU patch with multiple levels of global features to improve the variability. Multiple layers of iterative refinement significantly improve the AU discrimination ability.

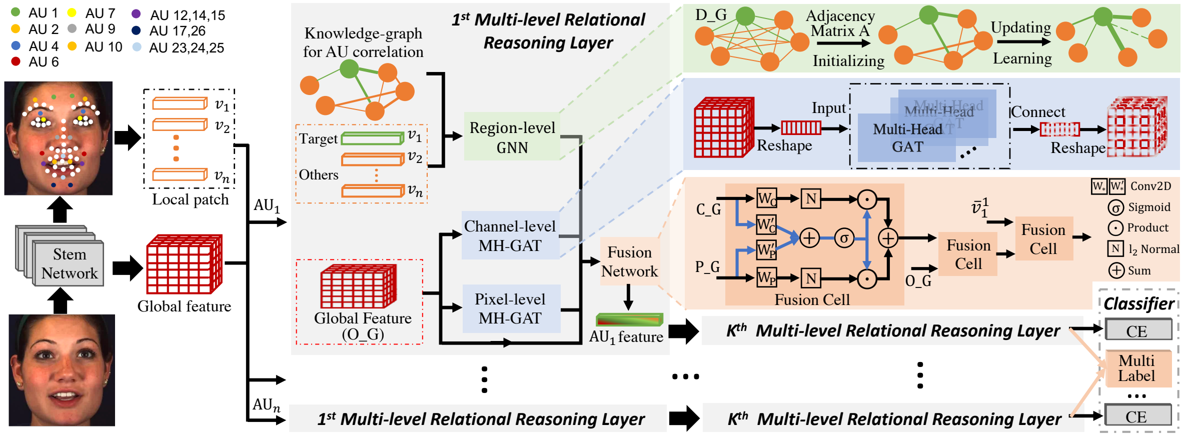


Figure 4.2: The overall architecture of the proposed MGRR-Net for facial AU recognition. Given one face image, the region-level features of local AU patches are extracted based on the detected landmarks from an efficient landmark localization network. The original global feature is extracted from the same shared stem network. Then the region-level GNN initialized with prior knowledge is applied to encode the correlation between different AU patches. Two separate MH-GATs are adopted to get two levels of global attention-aware features to supplement each AU. Finally, multiple levels of local-global features are fused by a hierarchical gated fusion strategy and refined by multiple iterations (best viewed in color).

4.3 The Proposed Method – MGRR-Net

As shown in Figure 4.2, the proposed approach consists of two core modules in each relational reasoning layer, *i.e.*, region-level local feature learning with relational modelling, and global feature learning with channel- and pixel-level attention. A hierarchical gated fusion network is designed to combine multi-level local and global features as the new target AU feature. Finally, after multiple layers of iterative refinement and updating, the AU features are fed into a multi-branch classification network for AU recognition. For clarity, the main notations and their definitions throughout the chapter are shown in Table 4.1.

4.3.1 Global and Local Features Extraction

Given a face image I, we adapt a stem network from the widely used multi-branch network (Ge, Wan, et al., 2021; Shao et al., 2018) to extract the original global feature O_G and further obtain the AU regions based on the detected landmarks. Different from the W. Li et al. (2018), our stem network contains a face alignment module for automatic face landmark detection, facilitating end-to-end training of our method. All branches share the stem network to reduce training costs and the complexity of network training. In particular, a hierarchical and multi-scale region learning module in the stem network extracts features from each local patch with different scales,

Table 4.1: Main notations and their definitions.

Notation	Definition
I	a facial image
S	a set of detected landmarks
O_G	the original global feature
m	the number of detected landmarks
V	a set of calculated patch features
v_i	the feature of i-th patch
n	the number of calculated patches corresponding to AUs
D_G	a fully-connected graph for AU relationship construction
A	a learnable adjacency matrix
a_i	the activation status of the i-th AU
P_{ij}	the coefficient between i-th and j-th AU
P,C	a set of pixel- and channel-level features
P_G	the pixel-level attention-aware global feature
C_G	the channel-level attention-aware global feature
L	the number of parallel attention layers
K	the number of relational reasoning layers
$\bar{v}_i^k$	the feature of i-th AU patch after k-th reasoning layer
GFC	a gated fusion cell
(x_i, y_i)	the ground-truth coordinate of the i-th facial landmark
$(\hat{x}_i, \hat{y}_i)$	the predicted coordinate of the i-th facial landmark
d_o	the ground-truth inter-ocular distance
p_i	the ground-truth occurrence probability of i-th AU
$\hat{p}_i$	the predicted occurrence probability of i-th AU

thus obtaining multi-scale representations. A series of landmarks $S = \{s_1, s_2, ..., s_m\}$ with length m are detected by an efficient face alignment module similar to Ge, Wan, et al. (2021); Shao et al. (2021), including three convolutional blocks connected to a max-pooling layer. According to the detected landmarks, local patches are calculated, and their features $V = \{v_1, v_2, ..., v_n\}$ are learned via the stem network, where n is the number of selected AU patches. For simplicity, we do not repeat the detailed structure of the stem network here.

4.3.2 Multi-level Relational Reasoning Layer

After we get the original global feature O_G for a face and the local region features $V = \{v_1, v_2, ..., v_n\}$ for AUs, a multi-layer multi-level relational reasoning model is introduced to automatically explore the relationship of individual local facial regions and supply two levels of global information. Figure 4.2 shows the detailed structure of the 1^{st} multi-level relational reasoning layer.

Region-level Local Feature Relational Modeling

Different from the predefined fixed AU relationship graph in G. Li et al. (2019), we construct a fully-connected graph D_G for all AUs, where the region-level features $V = \{v_1, v_2, ..., v_n\}$

constitute the nodes, and a learnable adjacency matrix A constitutes the edges at each layer to represent the possibility of AU co-occurrence (co-activated or non-activated). In this scheme, the AUs with no co-occurrence or low co-occurrence relationship in the training set will not be completely ignored, like G. Li et al. (2019); Z. Liu et al. (2020); X. Niu, Han, Shan, and Chen (2019). During the training process, we utilize prior knowledge to initialize A to assist and constrain model learning. Specifically, the dynamic graph D_G comprises nodes (the local region features $V = \{v_1, v_2, ..., v_n\}$) and edges (the relationship matrix A among AUs). Following X. Niu, Han, Shan, and Chen (2019), we calculate the relationship coefficients between AUs from datasets to initialize the adjacency matrix A. The statistical prior knowledge serves as the initial relationship, allowing suppression of the edges with low correlation and speeding up the relationship learning. The relationship coefficient A_{ij} between the i-th and j-th AU can be formulated as:

$$P_{ij} = \frac{1}{2}(P(a_i = 1|a_j = 1) + P(a_i = 0|a_j = 0)), \tag{4.1}$$

$$A_{ij} = |(P_{ij} - 0.5) * 2| \tag{4.2}$$

where a_i=1 denotes i-th AU is activated and 0 otherwise, $|\cdot|$ means absolute value function. From Eq.4.1 and Eq.4.2, $P(a_i$=1$|a_j$=1)=0.5 means that when j-th AU is activated, the probability of occurrence is equal to the no occurrence for i-th AU. It indicates that the activation of j-th AU could not provide useful information for the i-th AU, and therefore no edge is connected.

Attention-aware Global Features Learning

We argue that complementary global feature can improve the discrimination between AUs, which also alleviates the over-smoothing issue (Z. Chen et al., 2023)[1] in graph neural networks for local relationship modelling. To this end, we employ two separate high-efficiency GAT models (Veličković et al., 2018) to perform channel- and pixel-level attention-aware global features from original deep visual features in order to handle expression and subject diversities. Specifically, we reshape the original global feature O_G $\in \mathbb{R}^{(\mathrm{c,w,h})}$ into a set of channel-level features $\{C_1, ..., C_c\}, C_i \in \mathbb{R}^{w*h}$. Similarly, by reshaping pixel dimensions and keeping channel dimension of O_G from a convolution layer to reduce the parameters, we get a set of pixel-level features $\{P_1, ..., P_{w'*h'}\}, P_i \in \mathbb{R}^c$. The attention coefficient α_{ij} between channel- or pixel-level features is calculated in GAT, which can be formulated as (Here we take the process of channel-level

[1] In Graph Neural Network (GNN), the over-smoothing issue refers to the fact that as the number of network layers increases, the node embeddings tend to be similar, or even identical, causing the model to lose the ability to distinguish between different nodes. This problem weakens the expressive power of the model, especially in scenarios where subtle structures or complex relationships in the graph need to be recognised. The oversmoothing problem mainly stems from the propagation mechanism of GNNs, i.e., each layer of GNN aggregates node features with their neighbouring node features. As the number of layers increases, the transmission distance of the features expands, and the information gradually spreads to the whole graph, which leads to the features of different nodes becoming more and more similar.

attention-aware features as an example.):

$$\alpha_{ij} = \frac{\exp(U_q C_i (U_k C_j)^T / \sqrt{D})}{\sum_{\mathrm{o}\in\Omega_\mathrm{i}} \exp(U_q C_i (U_o C_o)^T / \sqrt{D})}, \tag{4.3}$$

where U_q, U_k, U_o are the parameters of mapping from $w * h$ to D and Ω_i denotes neighborhoods of C_i. $\sqrt{D}$ acts as a normalization factor. Following Vaswani et al. (2017); Veličković et al. (2018), we also employ multi-head dot product by L parallel attention layers to speed up the calculation efficiency. The overall working flow is formulated as:

$$\begin{aligned} \bar{C_i} &= \mathrm{ReLU}(\textstyle\sum_{\mathrm{o}\in\Omega_\mathrm{i}} U_c ||_l^L (\alpha_{io}^l * C_i)), \\ \alpha_{ij}^l &= \frac{\exp(U_q' C_i (U_k' C_j)^T / \sqrt{d})}{\sum_{\mathrm{o}\in\Omega_\mathrm{i}} \exp(U_q' C_i (U_o' C_o)^T / \sqrt{d})}, \end{aligned} \tag{4.4}$$

where U_c is the mapping parameter, U_q', U_k', U_o' map the feature dimension to $1/L$ of the original, $||$ means concatenation, and d equals D/L. Finally, the new channel-level attention-aware global feature C_G=$\{\bar{C_i}\}$ is reshaped to the same domination with O_G. With the same process on pixel-level features $\{P_1, ..., P_{w'*h'}\}$, we can get the final pixel-level attention-aware global features P_G after a deconvolution layer behind of a GAT with multi-head (MH-GAT).

Hierarchical Fusion and Iteration

We iteratively refine the i-th target AU feature of the proposed multi-level relational reasoning layer K times, which obtains other correlated local, and regional information and provides rich global details in each layer. The process can be formulated as:

$$\bar{v}_i^k = W_i^k v_i^k + \textstyle\sum_i^n (A_{ij}^k W_j^k v_j), \tag{4.5}$$

where W^k is the mapping parameter and A_{ij}^k means the learnable correlation coefficient between AU$_i$ and AU$_j$ at k-th layer. We then use a hierarchical fusion strategy by a gated fusion cell (GFC) to complement the global multi-level information for each updated AU feature at k-th layer as follows:

$$\bar{v}_i^{k+1} = \mathrm{GFC}(\bar{v}_i^k, \mathrm{GFC}(\mathrm{O\_G}^k, \mathrm{GFC}(\mathrm{C\_G}^k, \mathrm{P\_G}^k))), \tag{4.6}$$

We define the operation of GFC as follows:

$$\mathrm{GFC}(\mathrm{C\_G}^k, \mathrm{P\_G}^k) = \beta \odot ||W_C^k \mathrm{C\_G}^k||_2 + (1-\beta) \odot ||W_P^k \mathrm{P\_G}^k||_2, \tag{4.7}$$

$$\beta = \sigma(W_C^{k'} \mathrm{C\_G}^k + W_P^{k'} \mathrm{P\_G}^k), \tag{4.8}$$

where σ is the sigmoid function, and $||\cdot||$ denotes the l_2-normalization. $W_*^{k'}$ and W_*^k denote the Conv2D operation.

4.3.3 Joint Learning

A multi-label binary classifier is used to classify the AU activation state, which adopts a weighted multi-label cross-entropy loss function (denoted as CE in Figure 4.2) as follows,

$$\mathscr{L}_{au} = -\frac{1}{n}\sum_{i=1}^{n} w_i[p_i \log \hat{p}_i + (1-p_i)\log(1-\hat{p}_i)], \tag{4.9}$$

where p_i and $\hat{p}_i$ denote the ground-truth and predicted occurrence probability of the i-th AU, respectively; w_i is the data balance weights used in JAA (Shao et al., 2018). Furthermore, we also minimize the loss of AU category classification $\mathscr{L}_{int}$ by integrating all AUs information, including the refined AU features and the face alignment features, which is similar to the processing of $\mathscr{L}_{au}$.

We jointly integrate face alignment and facial AU recognition into an end-to-end learning model. The face alignment loss is defined as:

$$\mathscr{L}_{align} = \frac{1}{2d_o^2}\sum_{i=1}^{m}[(x_i-\hat{x}_i)^2 + (y_i-\hat{y}_i)^2], \tag{4.10}$$

where (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ denote the ground-truth coordinate and corresponding predicted coordinate of the i-th facial landmark, and d_o is the ground-truth inter-ocular distance for normalization in JÂANet (Shao et al., 2021). Finally, the joint loss of our MGRR-Net is defined as:

$$\mathscr{L} = (\mathscr{L}_{au} + \mathscr{L}_{int}) + \lambda \mathscr{L}_{align}. \tag{4.11}$$

where λ is a tuning parameter for balancing.

4.4 Experiments

In this section, we conduct extensive experiments to evaluate the proposed MGRR-Net. Especially the dataset and training strategy are first introduced. Then, MGRR-Net is compared with state-of-the-art FAU detection approaches quantitatively. Finally, we qualitatively analyze the results in detail.

4.4.1 Training Strategy

Our model is trained on a single NVIDIA RTX 2080Ti with 11 GB memory. The whole network is trained with the default initializer of PyTorch (Paszke et al., 2019) with the SGD solver, a Nesterov momentum of 0.9 and a weight decay of 0.0005. The learning rate is set to 0.01 initially, with a decay rate of 0.5 every two epochs. The maximum epoch number is set to 15. During the training process, aligned faces are randomly cropped into 176 × 176 and horizontally

Table 4.2: Comparisons of AU recognition for 8 AUs on DISFA in terms of F1-frame score (in %). CLP† is a semi-supervised method. * means the method employed a pre-trained model on the additional dataset, such as ImageNet (Deng et al., 2009) and VGGFace2 (Cao et al., 2018), *etc.*

Method	AU Index								**Avg.**
	1	2	4	6	9	12	25	26	
DSIN	42.4	39.0	68.4	28.6	46.8	70.8	90.4	42.2	53.6
JAA	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
LP-Net	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
ARL	43.9	42.1	63.6	41.8	40.0	**76.2**	**95.2**	66.8	58.7
SRERL	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
JÂANet	**62.4**	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
JAA-DGCN	61.8	51.7	64.5	46.0	54.2	63.6	85.5	69.4	62.0
CLP†	42.4	38.7	63.5	59.7	38.9	73.0	85.0	58.1	57.4
MMA-Net	63.8	54.8	73.6	39.2	**61.5**	73.1	92.3	70.5	66.0
MGRR-Net	61.3	**62.9**	**75.8**	**48.7**	53.8	75.5	94.3	**73.1**	**68.2**
UGN-B*	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
HMP-PS*	21.8	48.5	53.6	56.0	58.7	57.4	55.9	56.9	61.0
DML*	62.9	65.8	71.3	51.4	45.9	76.0	92.1	50.2	64.4
PIAP*	50.2	51.8	71.9	50.6	54.5	79.7	94.1	57.2	63.8
TransAU*	46.1	48.6	72.8	56.7	50.0	72.1	90.8	55.4	61.5
Bio-AU*	41.5	44.9	60.3	51.5	50.3	70.4	91.3	55.3	58.2
MGRR-Net	61.3	62.9	**75.8**	48.7	53.8	75.5	94.3	**73.1**	**68.2**

flipped. Regarding the face alignment network and stem network, we set the value of the general parameters to be the same with Shao et al. (2021). The iteration layer number K is set to 2 except otherwise noted. The dimensionality of O_G is $(64, 44, 44)$ and D is 1024. We employ L=8 parallel attention layers in GATs. In this chapter, all the mapping Conv2D operations used 1×1 convolutional filters with a stride one and a padding 1. We use a 3×3 Conv2D operation with a stride two and padding one before learning the channel-level feature to reduce the parameters. λ is empirically set to 0.5 for the joint optimisation of face alignment and facial AU detection on two benchmarks. Following the settings in W. Li et al. (2018); Shao et al. (2021); K. Zhao, Chu, and Zhang (2016), our MGRR-Net initializes the parameters of the well-trained model trained on BP4D when training on DISFA. This initialization greatly alleviates the poor performance issue on DISFA due to data volume and AU category imbalance. The training time is approximately 1.5 hours per epoch. In addition, we average the predicted probability of the local information and the integrated information as the final predicted activation probability for each AU rather than simply using the integrated information of all the AUs.

4.4.2 Evaluation Metrics

For all methods, the frame-based F1 score (F1-frame, %) is reported, which is the harmonic mean of the Precision P and Recall R and calculated by $\mathrm{F1} = 2\mathrm{P} * \mathrm{R}/(\mathrm{P} + \mathrm{R})$. To conduct a more comprehensive comparison with other methods, we also evaluate the performance with

Table 4.3: Comparisons of AU recognition for 8 AUs on DISFA in terms of Accuracy and AUC (in %). * means the method employed pretrained model on additional dataset.

AU	Accuracy						AUC				
	JAA	ARL	JÂANet	MMA-Net	UGN-B*	**MGRR-Net**	DRML	SRERL	DML*	DAR-GCN	**MGRR-Net**
1	93.4	92.1	**97.0**	96.8	95.1	96.8	53.3	76.2	**90.5**	84.5	89.5
2	96.1	92.7	97.3	96.5	93.2	**97.4**	53.2	80.9	92.7	92.5	**93.0**
4	86.9	88.5	88.0	91.6	88.5	**92.7**	60.0	79.1	**93.8**	72.2	93.6
6	91.4	91.6	92.1	91.5	**93.2**	92.1	54.9	80.4	90.3	48.3	**91.1**
9	95.8	95.9	95.6	96.5	96.8	**96.9**	51.5	76.5	84.4	78.3	**91.9**
12	91.2	**93.9**	92.3	92.3	93.4	93.4	54.6	87.9	95.7	37.8	**95.9**
25	93.4	**97.3**	94.9	95.5	94.8	96.8	45.6	90.9	98.2	50.3	**99.0**
26	93.2	94.3	94.8	95.0	93.8	**95.6**	45.3	73.4	87.4	74.3	**94.4**
Avg.	92.7	93.3	94.0	94.5	93.4	**95.2**	52.3	80.7	91.6	67.3	**93.6**

AUC (%) refers to the area under the ROC curve and accuracy (%). In addition, the average results over all AUs (denoted as **Avg.**) are computed with "%" omitted.

4.4.3 Comparison with State-of-the-art Methods

We compare our proposed MGRR-Net with several frame-based AU detection baselines and the latest state-of-the-art methods, including Deep Structure Inference Network (DSIN) (Corneanu et al., 2018), Joint AU Detection and Face Alignment (JAA) (Shao et al., 2018), Multi-Label Co-Regularization (MLCR) (X. Niu, Han, Shan, & Chen, 2019), Local relationship learning with Person-specific shape regularization (LP-Net) (X. Niu, Han, Yang, et al., 2019), Attention and Relation Learning (ARL) (Shao et al., 2019), Semantic Relationships Embedded Representation Learning (SRERL) (G. Li et al., 2019), Joint AU detection and face alignment via Adaptive Attention Network (JÂANet) (Shao et al., 2021), Data-Aware Relation Graph Convolutional Neural network (DAR-GCN) (X. Jia, Zhou, Li, Li, & Yin, 2022), Dual-channel Graph Convolutional Neural Network (JAA-DGCN) (X. Jia, Xu, Zhou, Wang, & Li, 2023), a semi-supervised Contrastively Learning the Person-independent representations method (CLP) (Y. Li & Shan, 2023) and a Multiview Mixed Attention based Network (MMA-Net) (Shang, Du, Li, Yan, & Yu, 2023). To ensure reliable and fair comparisons, we directly use the results of these methods reported. Note that, the best and second-best results are shown using bold and underline, respectively. The experimental results of our MGRR-Net are shown with a grey background.

For a more comprehensive display, we present methods (marked with $*$) (Y. Chen, Chen, Wang, Wang, & Liang, 2022; Y. Chen, Song, et al., 2022; Cui, Kuang, Gao, Talamadupula, & Ji, 2023; Jacob & Stenger, 2021; T. Song, Chen, et al., 2021; T. Song, Cui, Zheng, & Ji, 2021; Y. Tang, Zeng, Zhao, & Zhang, 2021; S. Wang et al., 2021) that use additional data, such as ImageNet (Deng et al., 2009) and VGGFace2 (Cao et al., 2018), for pre-training their complex feature extraction stem network firstly, such as ResNet (K. He et al., 2016) *etc*. From Jacob

Table 4.4: Comparisons with state-of-the-art methods for 12 AUs on BP4D in terms of F1-frame (in %). * means the method employed pretrained model on additional dataset.

AU	F1-frame															
	MLCR	JAA	LP-Net	ARL	SRERL	JÂANet	CLP	MMA-Net	**Ours**	R-CNN*	UGN-B*	HMP-PS*	DML*	TransAU*	Bio-AU*	**Ours**
1	42.4	47.2	43.3	45.8	46.9	**53.8**	47.7	52.5	[52.6]	50.2	54.2	53.1	52.6	51.7	57.4	[52.6]
2	36.9	44.0	38.0	39.8	45.3	47.8	**50.9**	**50.9**	[47.9]	43.7	46.4	46.1	44.9	49.3	52.6	[47.9]
4	48.1	54.9	54.2	55.1	55.6	58.2	49.5	**58.3**	[57.3]	57.0	56.8	56.0	56.2	61.0	64.6	[57.3]
6	77.5	77.5	77.1	75.7	77.1	78.5	75.8	76.3	[**78.5**]	78.5	76.2	76.5	79.8	77.8	79.3	[78.5]
7	77.6	74.6	76.7	77.2	78.4	75.8	**78.7**	75.7	[77.6]	78.5	76.7	76.9	80.4	79.5	81.5	[77.6]
10	83.6	84.0	83.8	82.3	83.5	82.7	80.2	83.8	[**84.9**]	82.6	82.4	82.1	85.2	82.9	82.7	[84.9]
12	85.8	86.5	87.2	86.6	87.6	88.2	84.1	87.9	[**88.4**]	87.0	86.1	86.4	88.3	86.3	85.6	[**88.4**]
14	61.0	61.9	63.6	58.8	63.9	63.7	67.1	63.8	[**67.8**]	67.7	64.7	64.8	65.6	67.6	67.8	[**67.8**]
15	43.7	43.6	45.3	47.6	**52.2**	43.3	52.0	48.7	[47.6]	49.1	51.2	51.5	51.7	51.9	47.3	[47.6]
17	63.2	60.3	60.5	62.1	**63.9**	61.8	62.7	61.7	[63.3]	62.4	63.1	63.0	59.4	63.0	58.0	[**63.3**]
23	42.1	42.7	48.1	47.4	47.1	45.6	45.7	46.5	[**47.4**]	50.4	48.5	49.9	47.3	43.7	47.0	[47.4]
24	55.6	41.9	54.2	**55.4**	53.3	49.9	54.8	54.4	[51.3]	49.3	53.6	54.5	49.2	56.3	44.9	[51.3]
Avg.	59.8	60.0	61.0	61.1	62.9	62.4	62.4	63.4	[**63.7**]	62.6	63.3	63.4	63.4	64.2	64.1	[63.7]

and Stenger (2021); X. Niu, Han, Yang, et al. (2019), the pre-trained feature extractor improved the average F1-score by at least 1.2% on BP4D. Due to the fact that our stem network only consists of a few simple convolutional layers, even if we pre-trained on additional datasets, it is unsuitable compared to pre-training on deeper feature extraction networks, such as ResNet50 (K. He et al., 2016), ResNet101 (K. He et al., 2016) and Swin Transformer-Base (Z. Liu et al., 2021). To this end, we have grouped them together to facilitate comparison with our proposed MGRR-Net. Notably, our results show excellence, affirming the superiority and efficacy of our proposed learning methodology. To provide a fair comparison, we omit the need for additional modality inputs and non-frame-based models (P. Liu et al., 2019; Tallec, Dapogny, & Bailly, 2022; H. Yang et al., 2020; H. Yang, Yin, Zhou, & Gu, 2021).

Quantitative Comparison on DISFA

We compare our proposed method with its counterpart in Table 4.2 and Table 4.3. It has been shown that our MGRR-Net outperforms all its competitors with impressive margins. Compared with the existing end-to-end feature learning and multi-label classification methods DSIN (Corneanu et al., 2018) and ARL (Shao et al., 2019), our MGRR-Net shows significant improvements on all AUs. These results demonstrate the effectiveness of accurate muscle region localization for AU detection. Although ARL (Shao et al., 2019) also performs sequential multiple attention explorations on global features, we believe that the sequential mechanism may destroy the diversity of different attention-aware features and slow down the training time. JÂANet (Shao et al., 2021) is the latest state-of-the-art method which also joint AU detection and face alignment into an end-to-end multi-label multi-branch network. Compared with the baseline JÂANet (Shao et al., 2021), our MGRR-Net increases the average F1-frame and average accu-

Table 4.5: Comparisons with state-of-the-art methods for 12 AUs on BP4D in terms of Accuracy and AUC respectively (in %). * means the method employed pretrained model on additional dataset, such as ImageNet (Deng et al., 2009), *etc*. So we do not directly compare.

AU	Accuracy					AUC			
	UGN-B*	JAA	ARL	JÂANet	**MGRR-Net**	DRML	SRERL	DML*	**MGRR-Net**
1	78.6	74.7	73.9	75.2	**78.7**	55.7	67.6	**78.5**	78.1
2	80.2	80.8	76.7	80.2	**82.1**	54.5	70.0	75.9	**77.2**
4	80.0	80.4	80.9	**82.9**	81.6	58.8	73.4	**84.4**	83.8
6	76.6	78.9	78.2	**79.8**	78.7	56.6	78.4	**88.6**	88.4
7	72.3	71.0	**74.4**	72.3	73.7	61.0	76.1	**84.8**	82.3
10	77.8	80.2	79.1	78.2	**81.2**	53.6	80.0	**87.3**	86.3
12	84.2	85.4	85.5	86.6	**86.9**	60.8	85.9	**93.9**	93.6
14	63.8	64.8	62.8	65.1	**67.0**	57.0	64.4	71.8	**72.9**
15	84.0	83.1	84.7	81.0	**84.2**	56.2	75.1	80.7	**80.8**
17	72.8	73.5	**74.1**	72.8	72.2	50.0	71.7	75.0	**78.2**
23	82.8	82.3	82.9	82.9	**84.1**	53.9	71.6	78.7	**79.3**
24	86.4	85.4	85.7	**86.3**	86.0	53.9	74.6	84.3	**87.8**
Avg.	78.2	78.4	78.2	78.6	**79.7**	56.0	74.1	82.0	**82.4**

racy scores by large margins of 4.7% and 1.2% and shows clear improvements for most annotated AU categories. The main reason lies in JÂANet (Shao et al., 2021) completely ignores the correlation between branches and the individual modelling of each AU. Compared with JAA-DGCN (X. Jia et al., 2023) that also applies the graph relationship model, our MGRR-Net still performs better on most metrics because we model local relationships while supplementing a variety of information from the global face. Moreover, compared with the latest state-of-the-art MMA-Net (Shang et al., 2023), MGRR-Net achieves a 2.2% lead in the average F1-frame metric. In addition, compared with the current state-of-the-art AU detection methods based on pre-trained models, such as UGN-B (T. Song, Chen, et al., 2021), HMP-PS (T. Song, Cui, Zheng, & Ji, 2021), DML (S. Wang et al., 2021), PIAP (Y. Tang et al., 2021) and Bio-AU (Cui et al., 2023) *etc.*, we also achieve the best performance in terms of the average F1-frame.

Furthermore, the results of the Accuracy and AUC evaluations provide further evidence of the effectiveness of our method compared to other state-of-the-art methods. In particular, our MGRR-Net obtains improvement on the average of Accuray, *i.e.* 95.2 % *vs.* 94.5%, compared with MMA-Net (Shang et al., 2023). And on AUC metric, our MGRR-Net also achieves higher results on most metrics and increases 2.0% compared to DML* (S. Wang et al., 2021).

Quantitative Comparison on BP4D

Table 4.4 and 4.5 show the AU detection results of different methods in terms of F1-frame, Accuracy and AUC on BP4D dataset, where the method in the left of Table 4.4 uses a feature extractor without pre-training and the method with * is based on the pre-trained feature extrac-

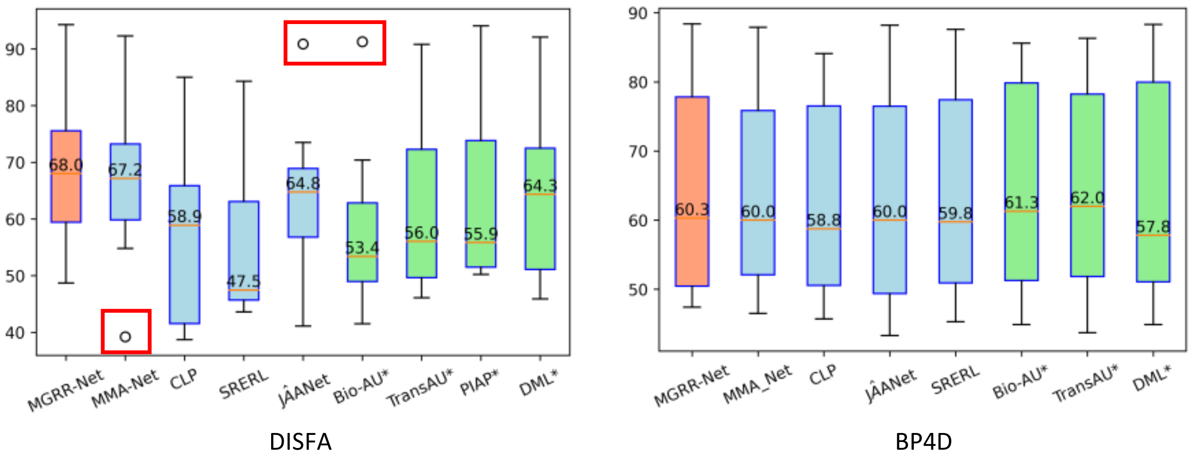


Figure 4.3: Box Plots of the distribution of performances on all AU categories (the labeled values are medians). (a) on DISFA 3-flod test set and (b) on BP4D 3-flod test set.

tor (our method is trained on BP4D only). Compared with the multi-branch combination-based JÂANet (Shao et al., 2021), the average F1 frame score and average accuracy score of MGRR-Net get 1.3% and 1.1% higher, respectively. Furthermore, compared with the latest graph-based relational modelling method SRERL (G. Li et al., 2019), MGRR-Net increases the average F1-frame and average AUC by large margins of 0.8% and 8.3%. This is mainly due to the fact that the proposed method models the semantic relationships among AUs while also gaining complementary features from multiple global perspectives to increase the distinguishability of each AU. In addition, our MGRR-Net achieves the best or second-best AU detection performance in terms of F1-frame, Accuracy and AUC for most of the 12 AUs annotated in BP4D compared with the state-of-the-art methods. For example, compared with the latest method MMA-Net (Shang et al., 2023), which simultaneously modelled the deep feature learning and the structured AU relationship in a unified framework, ours greatly outperforms it by 0.3% in terms of the average of F1-frame. In addition, compared with the advanced models pre-trained with additional data (marked with $*$ in Table 4.4 and Table 4.5), our MGRR-Net still has strong competitiveness.

Experimental results of MGRR-Net demonstrate its effectiveness in improving AU detection accuracy on DISFA and BP4D, as well as good robustness and generalization ability. Note that the main reason why some AUs are clearly less accurate than others is due to data imbalance, as shown in Figure 4.3, this is a phenomenon that exists in all existing methods (Cui et al., 2023; Jacob & Stenger, 2021; Y. Li & Shan, 2023; Shang et al., 2023; Shao et al., 2021; Y. Tang et al., 2021). In BP4D, where the data distribution is relatively reasonable, the results' distribution of each method is close. But in DISFA, where the data distribution is more extreme, the result distribution of our MGRR-Net can perform better, *i.e.* lower variance and no outliers. We infer that two aspects promote this improvement. On one hand, we use a weighted multi-label cross-entropy loss function as Eq.(4.9) to solve the data imbalance problem to a certain extent. On the other hand, our multi-level fused representation can complement each AU representation, as

Table 4.6: Effectiveness of key components of MGRR-Net evaluated on DISFA in terms of F1-frame score (in %).

Method		1	2	3	4	5	6	MGRR-Net
Setting	D_G	-	√	√	√	√	√	√
	O_G	-	-	√	-	√	√	√
	C_G	-	-	-	√	√	-	√
	P_G	-	-	-	√	-	√	√
AU Index	1	47.1	52.5	58.4	60.0	65.4	61.0	[61.3]
	2	61.1	58.1	63.0	65.7	64.5	67.3	[62.9]
	4	66.3	73.3	70.9	67.4	72.5	76.8	[75.8]
	6	44.7	44.4	46.2	43.8	42.6	40.9	[48.7]
	9	52.2	52.5	47.7	57.1	52.9	58.0	[53.8]
	12	74.9	73.2	72.1	75.4	75.3	74.8	[75.5]
	25	92.2	94.7	93.4	93.3	94.3	93.7	[94.3]
	26	66.2	71.2	71.8	64.7	71.4	65.8	[73.1]
Avg.		63.1	65.0	65.4	65.9	67.4	67.3	[68.2]

well as combine with other AU areas, to further improve AU classification.

4.4.4 Ablation Studies

We perform detailed ablation studies on DISFA to investigate the effectiveness of each part of our proposed MGRR-Net. Due to space limitations, we do not show the ablation results for BP4D, but it is consistent with DISFA. To assess the effect of different components, we run the experiments with same parameter setting (*e.g.* layer K=2) for variations of the proposed network in Table 4.6.

Effects of Region-level Dynamic Graph

In Table 4.6, we can see that learning by the dynamic graph initialized with prior knowledge (indicated by D_G) outperforms baseline with an improvement of average F1-frame from 63.1% to 65.0%, indicating that the dynamic graph could get richer features from other correlated AU regions to improve robustness. Furthermore, to cancel out the initialization of prior knowledge, we randomly initialize the dynamic graph, which decreases F1-frame to 64.7%. These observations suggest that the relationship reasoning in the dynamic graph can significantly boost the performance of AU detection, while prior knowledge makes a great contribution but not predominantly.

Effects of Multi-level Global Features

We test the contributions of multiple important global feature components of the model in Table 4.6, namely, original global feature (O_G) from stem network, channel-level global feature (C_G) from channel-level MH-GAT and pixel-level global feature (P_G) from pixel-level MH-GAT. After we supplemented original global feature (O_G) for each target AU, the average

Table 4.7: Performance comparison of MGRR-Net with different iteration step number K on DISFA in terms of F1-frame score (in %).

Layers	AU Index								**Avg.**
	1	2	4	6	9	12	25	26	
K=1	64.5	58.3	74.9	46.1	**54.4**	75.4	92.3	73.1	67.4
K=2	61.3	62.9	75.8	**48.7**	53.8	**75.5**	**94.3**	**73.1**	**68.2**
K=3	**65.5**	**67.0**	**77.6**	40.0	44.9	75.1	94.0	68.8	66.6

F1-frame score has been improved from 65.0% to 65.4%, demonstrating the effectiveness of global detail supplementation. The fusion of channel- and pixel-level global features (C_G and P_G) results in a 0.9% increase, indicating that they make the AU more discriminative than only using the original global features. Comparing the results of the fifth test (with C_G) and the sixth test (with P_G) in Table 4.6 with the third test, one of the channel-level and pixel-level global features can boost the performance by roughly the same amount. It suggests that by supplementing and training different levels of global features for each AU branch, more global details can be provided to detect AUs in terms of different expressions and individuals.

Finally, the hierarchical gated fusion of multi-level global and local features leads to a significant performance improvement to 68.2% in terms of F1-frame score. It validates that the dynamic relationship of multiple related face regions provides more robustness, while the supplementation of multi-level global features makes the AU more discriminative.

Effects of Layer Number

We evaluate the impact of layer number of our proposed iterative reasoning network. As shown in Table 4.7, MGRR-Net achieves the averaged F1-frame score of 67.4%, 68.2% and 66.6% on DISFA when the reasoning layer number K is set to 1, 2 and 3 respectively. The averaged F1-frame scores on BP4D dataset are 63.5%, 63.7%, and 63.1% respectively. It achieves the best performance when K=2, and is overfitted when K>2. Finally, the optimal number of layers is 2 for our MGRR-Net on DISFA and BP4D datasets.

4.4.5 Visualization of Results

To better understand the effectiveness of our proposed model, we visualize the learned class activation maps of MGRR-Net corresponding to different AUs in terms of different expressions, postures and individuals, as shown in Figure 4.4. Three examples are from DISFA and three are from BP4D (Two bad examples of abnormal offsets happening are shown at the bottom of Figure 4.4.), containing visualization results of different genders and different poses with different AU categories. Through the learning of MGRR-Net, not only the concerned AU regions can be accurately located, but also the positive correlation with other AU areas can be established and other details of the global face can be supplemented. The different activation maps of the same AU on different individuals show that our MGRR-Net can dynamically adjust according to the

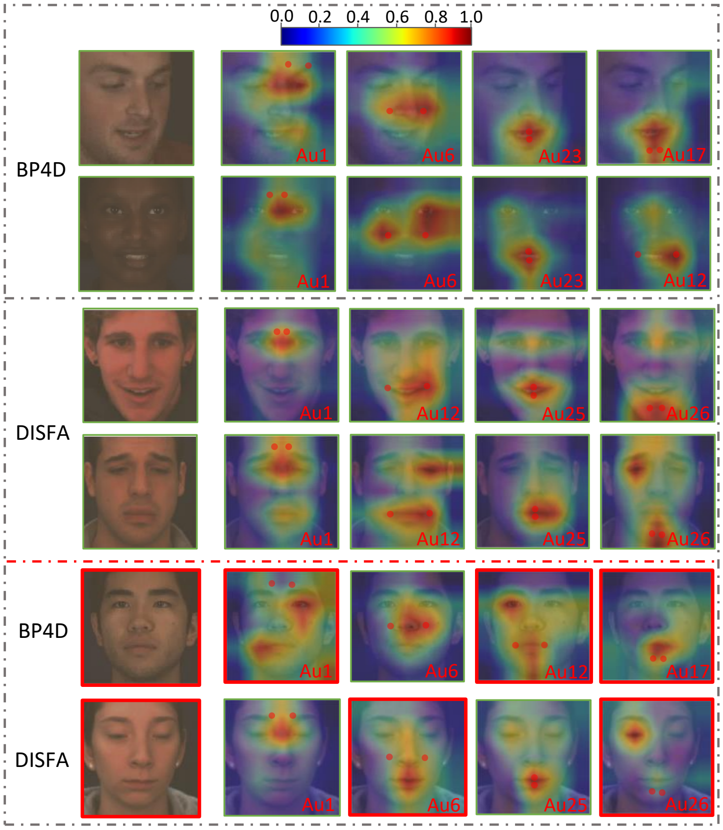


Figure 4.4: Class activation maps that show the discriminative regions for different AUs in terms of different expressions and individuals on DISFA and BP4D datasets. We show the region center positions defined by the detected landmarks for the corresponding AUs. Abnormally shifted AU activation maps are marked with red boxes.

differences of expression, posture, and individual. Some activation maps are inconsistent with the predefined AU areas, which may be caused by the insensitivity to the target predefined areas after the introduction of multi-level global supplementation. Furthermore, the supplementation of global features with multiple perspectives allows different AUs to access a lot of information outside the defined areas, as shown in Figure 4.4, which is helpful for adaptive changes in terms of different individuals and their expressions.

4.5 Conclusion

In this chapter, we have proposed a novel multi-level graph relational reasoning network (termed MGRR-Net) for facial AU detection. Each layer of MGRR-Net can encode the dynamic relationships among AUs via a region-level relationship graph and multiple complementary levels of global information covering expression and subject diversities. The multi-layer iterative feature refinement finally obtains robust and discriminative features for each AU. Extensive experimental evaluations on DISFA and BP4D show that our MGRR-Net outperforms state-of-the-art AU detection methods with impressive margins.

4.6 Limitation

Although MGRR-Net effectively addresses the modeling of relationships between facial AUs and captures local and global contextual relationships from multiple perspectives, the overall complexity of the model remains a significant limitation. Its intricate architecture and numerous interactions make it challenging to interpret the feature representations underlying each AU prediction. Consequently, the predicted AU states may lack transparency, making it difficult to understand how specific muscular features influence the final predictions. This complexity could hinder the practical application of the model in real-world scenarios, where interpretability is crucial for user trust and decision-making. Future work should focus on enhancing the interpretability of MGRR-Net to provide clearer insights into its predictions and the underlying mechanisms of AU recognition.

Chapter 5

Towards End-to-End Explainable Facial Action Unit Recognition via Vision-Language Joint Learning

Improving the individual AU representation becomes the key to boosting the FAU recognition task. The main challenge is how to maintain discriminative inter-AU features while ensuring rich semantic representation capabilities within AUs. Existing works (Ge, Jose, et al., 2023, 2024; Ge, Wan, et al., 2021; Shao et al., 2021) have made great progress, but two prevalent defects remain unexplored: (i) improving the representation capability of individual AUs through inter-AU relational connections may compromise the distinctiveness of features among AUs, and (ii) directly obtaining classification results lacks an explainable basis for judgments. The main reason for the former is that inter-AU information-based relationship reasoning networks, such as GCN (G. Li et al., 2019; T. Song, Chen, et al., 2021) or GNN (Ge, Jose, et al., 2024) etc., are prone to the over-smoothing problem (Rusch, Bronstein, & Mishra, 2023) among the different AU nodes, especially on small-scale face datasets. This makes differentiating the AU states difficult. The latter issue is caused by the current mainstream paradigm as shown in Figure 5.1, which only focuses on the classification results and ignores the corresponding explanations.

5.1 Introduction

In this work, we bring fresh insights toward explainable facial AU recognition via the integration of language generation supervisors within an end-to-end FAU recognition network. We argue that detailed AU language descriptions, such as “The corners of the lips are markedly raised and angled up obliquely. The nasolabial furrow has deepened ..." for activated AU12 (Lip Corner Puller) in Figure 5.1, can provide more linguistic semantics and relations to corresponding AU appearance features than mainstream relational reasoning methods (Ge, Jose, et al., 2024; G. Li et al., 2019; Luo et al., 2022). Furthermore, different AUs correspond to different descriptions,

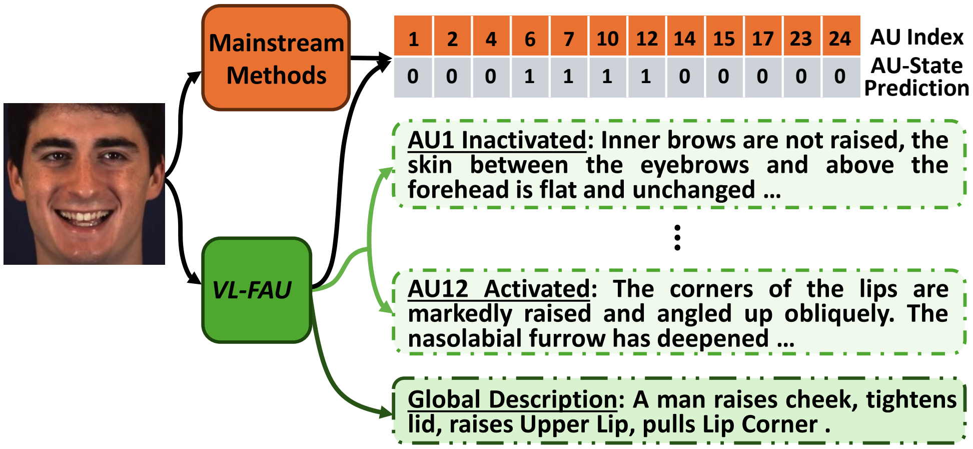

Mainstream Methods
VL-FAU
1 2 4 6 7 10 12 14 15 17 23 24
AU Index
0 0 0 1 1 1 1 0 0 0 0 0
AU-State Prediction
AU1 Inactivated: Inner brows are not raised, the skin between the eyebrows and above the forehead is flat and unchanged ...
AU12 Activated: The corners of the lips are markedly raised and angled up obliquely. The nasolabial furrow has deepened ...
Global Description: A man raises cheek, tightens lid, raises Upper Lip, pulls Lip Corner .

Figure 5.1: Comparative analysis of FAU recognition paradigms is shown between conventional methods and our *VL-FAU* . While the mainstream methods provide direct predictions of AU activation states (orange stream), the VL-FAU model not only offers activation predictions but also provides detailed local and global descriptions of the corresponding AUs in natural language.

keeping them better distinguishable. Different from encoding pre-provided activated AU language descriptions into AU classification in H. Yang et al. (2021), we introduce the language generation model as language semantic supervision for the classification of different AU states. It can provide AU prediction with the explainable language descriptions as Figure 5.1 while providing sufficient semantics to enrich AU representations and supervising the AU detection pertinently to distinguish AU.

5.1.1 Motivation

Motivated by the above insights, we propose a novel end-to-end vision-language joint learning model (termed *VL-FAU*) for FAU detection with explainable language generation. Compared with the mainstream methods in Figure 5.1, our main innovations lie in two aspects: (i) exploring the potential of joint language generation auxiliary training for intra-AU semantic enhancement and inter-AU semantic feature differentiation and (ii) providing interpretability for end-to-end FAU recognition. Specifically, we first introduce a new dual-level AU feature refinement based on multi-scale combined representations from a vision backbone. This way, we provide each AU branch with a unique attention-aware representation ability. Secondly, to improve the semantics and guarantee the distinguishability of AU features as well as their interpretability, we integrate the end-to-end multi-branch framework with the local and global language generations for explicit semantic guidance. On one hand, each local branch of AU recognition is simultaneously supervised by a corresponding local language generator via fine-grained semantic-supervised optimization. Such schema constrains the semantics and relationships of AU features by local language modeling and improves the inter-branch distinguishability for each face image. On the other hand, global facial features within and between subjects are difficult to distinguish because muscle changes are mostly subtle under different states. To this end, we introduce a global language model to generate a description focusing on all activated muscles as global semantic supervision. It provides better distinction between different whole-face representations within and between subjects via multiple facial state foci as shown in Figure 5.1. Finally, vision AU recognition and language description generation are jointly optimized.

5.1.2 Contributions

The contributions of this work are as follows:

- We propose a novel end-to-end vision-language joint learning scheme for explainable FAU recognition (VL-FAU), which leverages auxiliary training of language generators to improve discrimination and explanation for FAU recognition.
- We design a new dual-level AU representation learning method based on the multi-scaled facial representation for AU branches, which provides stronger attention-aware AU representation ability;
- We design a novel joint supervision method with local and global language generations for FAU recognition. In such schema, the local language generation provides explicit semantic supervision of each independent AU branch, thereby improving inter-AU discriminability and intra-AU detailed semantics, while the global language model maintains the global distinguishability of different facial states within and between subjects.
- We extend FAU datasets with new local and global language descriptions for different facial muscle states to facilitate language-interpretable FAU recognition.

We conduct extensive experiments on two widely used benchmarks, *i.e.,* BP4D and DISFA, to evaluate the proposed VL-FAU model. VL-FAU outperforms state-of-the-art approaches in AU recognition and provides detailed language descriptions of the individual AU decision and global face state for explanations.

5.2 Related Work of Multi-task Joint Learning

Recently, multi-task joint learning (Ge, Wan, et al., 2021; Huang et al., 2018; L. Qin et al., 2023; F. Zhu, Zhu, Chang, & Liang, 2020) has been an emerging topic in deep learning research, which enables the same target to efficiently obtain multiple representation capabilities in a unified model. For example, Shao et al. (2021) joined face alignment into a facial AU recognition framework to assist in locating the muscles corresponding to AUs. L. Qin et al. (2023) combined four face tasks into a unified framework and assigned an attention module for each face analysis subnet. However, these methods only focus on computer vision tasks and ignore the modality complementary advantages of multiple modality-task joint learning, *i.e.* interpretability complementarity and semantic complementarity. For example, Ju et al. (2021) proposed a joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection, which can provide image sentiment predictions with explicit language explanations. However, most visual-language joint learning models (Fu et al., 2024; Huang et al., 2018; Ju et al., 2021; F. Zhu et al., 2020) rely on the joint input of both modalities during the inference process.

In this chapter, we argue that explicit language can facilitate the modeling of inter-AU relevance and diversity. Each AU description can include muscle details and relationship descriptions identified by AU, which not only replace the visual relational reasoning among AUs, but also ensure the independence of AUs to improve distinguishability. The most relevant research to ours is SEV-Net (H. Yang et al., 2021), which introduced AU language descriptions as a prior embedding into AU representations. However, significant drawbacks are that SEV-Net relied on the human pre-given language descriptions of activated AUs for inference and followed the mainstream AU recognition paradigm in Figure 5.1 for only AU prediction. In contrast to SEV-Net (H. Yang et al., 2021), our VL-FAU introduces language generation auxiliary models for local AU prediction and global face refinement. It focuses on end-to-end language-explainable facial-image AU recognition without any pre-given reference for inference.

5.3 The proposed Method – VL-FAU

As shown in Fig. 5.2, the proposed approach –*VL-FAU*– consists of two key components, *i.e.*, multi-level AU representation learning, and local and global auxiliary language generation. The former contains dual-level AU individual refinement for multi-branch FAU recognition based on multi-scale feature combinations from Swin-Transformer. The latter consists of local AU language generation, which facilitates semantic supervision of each AU and global facial language generation for whole-face semantic supervision. Both of these can help AU representations become more robust and distinguishable, thus improving the performance. Thus, we create an end-to-end framework with a multi-label classifier for explainable FAU recognition, supervised with local and global AU language generation auxiliaries.

5.3.1 Multi-level AU Representation Learning

Global Facial Representation Extraction

Given a face image I, we adapt the widely-used Swin-Transformer as the stem network (Z. Liu et al., 2021) to extract the global feature V by combining the multi-scale representations from different stages. Multi-scale feature learning and combination (G. Li et al., 2019; Prudviraj, Vishnu, & Mohan, 2022; J. Qin, Huang, & Wen, 2020; L. Qin et al., 2023) is a popular image

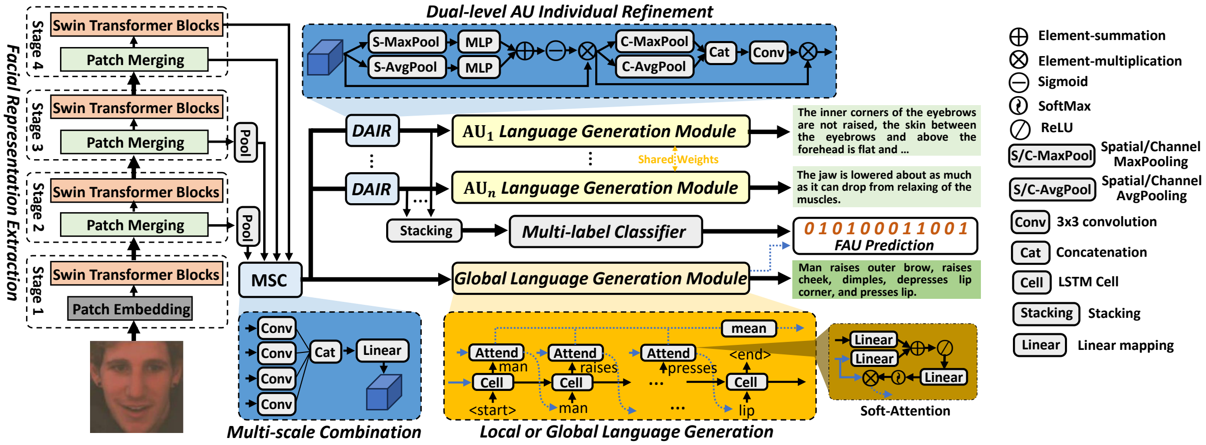


Figure 5.2: The overall end-to-end architecture of the proposed VL-FAU for explainable facial AU recognition. Given one face image, the multi-scale combined facial representation is extracted based on a pre-trained Swin-Transformer. VL-FAU is based on the multi-branch network containing multiple independent AU recognition branches as well as a global language generation branch. Each independent AU recognition branch owns a dual-level AU individual refinement module (DAIR) for individual AU attention-aware mining and a local AU language generation module for explicit semantic auxiliary supervision to improve the inter-AU distinguishability. A global language generation based on the multi-scale facial representation is leveraged to preserve shared stem feature diversity via multiple facial state foci. Finally, the multi-branch AU refined representations are stacked for multi-label classification with local and global language auxiliary supervisions (best viewed in color).

representation approach to leverage different levels of semantics from the backbone, where the low-level features from shallow blocks contain more texture information and the deeper blocks contain high-level semantics. In the work, we follow this strategy (*multi-scale combination – MSC*) to represent the global face image, which contains 4 independent 3×3 convolutions to learn and reshape the representations (V_1^s, V_2^s, V_3^s, V_4^s) from different stem stages. After that, four-level feature maps are combined as the global facial representation V by a learnable Linear layer. The above process can be formulated as:

$$V = W^m([\mathrm{Conv}(V_1^s) : \mathrm{Conv}(V_2^s) : \mathrm{Conv}(V_3^s) : \mathrm{Conv}(V_4^s)]), \tag{5.1}$$

where $[:,:]$ means the concatenation operation, $W^m \in \mathbb{R}^{D \times d}$ is mapping parameter and Conv means 3×3 convolution. For simplicity, we will not repeat the detailed structure of the stem Swin-Transformer network (Z. Liu et al., 2021) here.

Dual-level AU Individual Refinement

Multi-branch network is a good way to learn the rich and fine-grained individual facial AU representation for the final classification. In this chapter, we propose a new multi-branch dual-level AU individual refinement (termed *DAIR*) for the final attention-aware AU representations. As shown at the top of Figure 5.2, each DAIR in each branch contains two levels of attention learning, i.e. channel-level and spatial-level, which employs two pooling strategies (M. Lin, Chen, & Yan, 2014; Sudholt & Fink, 2016). While existing research from both perspectives is extensive (Ge, Jose, et al., 2024; Lu & Hu, 2022; L. Qin et al., 2023; Woo, Park, Lee, & Kweon, 2018), this study represents the first adaptation within independent FAU branches for enhancing the independent fine-grained attention-aware AU representations rather than coarse-grained global representation. Specifically, following Woo et al. (2018), we extract the max-pooled and average-pooled channel-level vectors (F^c_{max} and F^c_{avg}) along the spatial axis for channel-wise attention mining. After that, two shared learnable multi-layer perceptrons (MLP) are used to obtain the mapping of channel-wise vectors and then a sigmoid function is applied to get the *i*-th individual AU-channel attention. Similarly, we later extract the max-pooled and average-pooled spatial-level vectors (F^s_{max} and F^s_{avg}) along the channel axis for spatial-wise attention mining. We utilise a 3×3 convolution to generate a spatial attention map, which focuses on highly responsive muscle areas associated with the current AU. Finally, the *i*-th individual AU attention-aware representation $\hat{V}_i$ can be obtained as below:

$$\bar{V}_i = \sigma(\mathrm{MLP}(F^c_{max}) + \mathrm{MLP}(F^c_{avg}))V, \tag{5.2}$$

$$\hat{V}_i = \mathrm{Conv}([F^s_{max} : F^s_{avg}])\bar{V}_i, \tag{5.3}$$

where σ is the sigmoid function.

5.3.2 Auxiliary Supervision with Local and Global Language Generation

In addition to improving AU individual representations in multiple AU branches, our main innovation is to provide a new approach and inspiration for explainable FAU recognition, *i.e.*, joint auxiliary language generation for FAU recognition, rather than a new complex language model. Language generation provides explicit semantic supervision while giving linguistic interpretability of the corresponding identified AUs, rather than just simply recognising the AUs. It contains two aspects: (i) global face language generation for explicit semantic auxiliary supervision of the whole face image representation, and (ii) local AU language generation for individual AU semantic auxiliary supervision.

Global Language Generation

Different from the mainstream facial AU recognition (G. Li et al., 2019; Shao et al., 2018), we introduce a new global language auxiliary model to generate the language description for activated AUs of the whole face. It brings benefits for subsequent multi-branch FAU recognition, i.e., it maintains the accuracy and variability of intra- and inter-subject stem features by explicitly focusing on multiple different AU muscles. Thus, it somewhat overcomes the data imbalance, i.e., inactive AUs are far more numerous than activated AUs in benchmarks.

Specifically, as shown in Figure 5.2, we treat the multi-scale stem-feature V as the encoded image feature and input it into an attention-aware language decoder similar to image captioning (Ge et al., 2019; Xu et al., 2015), which contains a soft-attention module to mine the attention-aware visual representation based on the past generated word $s^g_{0:i-1}$ for new word s^g_i. For example, when generating current word s^g_i, we first calculate the soft-attention α_i between the image feature V and the last hidden state h_{i-1} of generated word s^g_{i-1} by linear-based mapping operations with SoftMax function (LeCun, Bengio, & Hinton, 2015). And then the i-th attention-aware visual feature $\alpha_i V$ is combined with the last hidden state h_{i-1}, which is fed into the i-th LSTM cell (Hochreiter & Schmidhuber, 1997) with the previous cell state c_{i-1} to generate the new hidden state h_i and cell state c_i of new word. The above process can be formulated as:

$$\alpha_i = \mathrm{SoftMax}(W^a(\mathrm{ReLU}(W^v V + W^h h_{i-1}))), \tag{5.4}$$

$$(h_i, c_i) = \mathrm{Cell}([\alpha_i V : h_{i-1}], c_{i-1}), \tag{5.5}$$

where $W^v \in \mathbb{R}^{d\times d}$, $W^h \in \mathbb{R}^{h\times d}$, and $W^a \in \mathbb{R}^{1\times d}$ are the parameters of mapping function. During the training process, we use a shared learnable parameter $W^s \in \mathbb{R}^{h\times voc}$ to obtain vocabulary-length predicted vector and obtain the max-score index as the predicted word, as follows:

$$s^g_i = \arg\max(W^s h_i), \tag{5.6}$$

During the inference process, the beam search (Klein & Manning, 2003) can be used to obtain the most optimal global description $S^g = [s^g_1, ..., s^g_T]$.

Local Language Generation

We introduce an individual local language generation model for each AU branch as an auxiliary semantic supervision, which can generate the corresponding fine-grained language description for each AU determination. Compared with the global facial description, the local AU language description contains more facial muscle details described for corresponding AUs. Besides, different AUs have different local language descriptions, which are more fine-grained and diverse. The main motivation of the joint language model in each AU branch is not only to improve the distinguishability between AUs using fine-grained semantic auxiliary supervision, but also to provide specific language interpretation for each AU prediction.

In particular, different from global language generation, each local language model utilizes the proposed DAIR to further refine the encoded face feature V, obtaining the distinguishable attention-aware AU representation $\hat{V}_i$ for the subsequent language generation. The decoder architecture of each local language generation model is the same as the global language model. To save space, we use $\oint_{\mathrm{i}}$ to represent the i-th local language model for the i-th AU branch and omit the model details. Finally, we can obtain the i-th local AU description $S_i^l = [s_{1;i}^l, ..., s_{T;i}^l]$ with length T, as follows:

$$S_i^l = \oint_{\mathrm{i}} (\mathrm{DAIR_i(V)}), \tag{5.7}$$

Note that, the local language models among the multiple AU branches are shared to maintain efficiency.

5.3.3 Vision-Language Joint Learning

VL-FAU joints facial AU recognition (vision) and description generation (language) into an end-to-end multi-branch network. Among them, FAU recognition is predominant due to the explicit AU annotations, while the language models are used for auxiliary semantic supervision and to improve feature diversity and distinguishability of visual stem and sub-branches. For facial AU recognition, one fully connected layer with SoftMax function is used as a multi-label binary classifier to classify the AU activation state, which adopts a weighted multi-label cross-entropy loss function as follows,

$$\mathscr{L}_{Fau} = -\frac{1}{N}\sum\nolimits_{i=1}^{N} \gamma_i [y_i \log(p(y_i)) + (1-y_i)\log(1-p(y_i))], \tag{5.8}$$

$$\gamma_i = \frac{1/\varepsilon_i}{\sum_{i=1}^{N}(1/\varepsilon_i)} \tag{5.9}$$

where N is AU number, y_i and $p(y_i)$ denote the ground-truth and predicted probability for the i-th AU occurrence, respectively. γ_i is a balancing weight of the i-th AU calculated by the i-th AU occurrence rate ε_i in the training set.

Local AU language generation auxiliary training provides detailed semantic supervision for AU predictions and is optimised by commonly used negative log-likelihood loss, as follows:

$$\mathscr{L}_{Lgen} = -\frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{1}{T}\sum\nolimits_{t=1}^{T} (\log p(s_{t;i}^l | \hat{V}_i, s_{[0:t-1];i}^l)) \tag{5.10}$$

where N is the number of AU branches and T is the length of each description. The t-th word $s_{t;i}^l$ of AU$_i$ description is generated based on the previous words $s_{[0:t-1];i}^l$ and the AU$_i$ refined visual feature $\hat{V}_i$.

Moreover, the introduced global language generation is also optimized by a similar objective function in Eq. (5.10) with the global generation $S^g = [s_1^g, ..., s_T^g]$ as $\mathscr{L}_{Ggen}$:

$$\mathscr{L}_{Ggen} = -\frac{1}{T}\sum\nolimits_{t=1}^{T} (\log p(s_t^g | V, s_{[0:t-1]}^g)) \tag{5.11}$$

Different from the local language generation for specific AU branches, the global language generation faces the challenge of linguistic semantic diversity due to facial state changes. To this end, we add a global AU classification loss $\mathscr{L}_{Gau}$ together with $\mathscr{L}_{Ggen}$ as a constraint, where

AU states are predicted by a shared linear classifier based on the average attention-aware visual feature from language model.

Finally, the joint loss of our VL-FAU model can be optimized by maximizing the following lower bound:

$$\mathscr{L} = \mathscr{L}_{Fau} + \mathscr{L}_{Lgen} + \mathscr{L}_{Ggen} + \mathscr{L}_{Gau}. \tag{5.12}$$

5.4 Experiments

5.4.1 Data Processing

We choose the same datasets with ALGRNet (Ge, Jose, et al., 2023), including BP4D (X. Zhang et al., 2014) and DISFA (Mavadati et al., 2013) datasets to evaluate our proposed model. During training, language descriptions of different AU states are hand-crafted, containing descriptions of both activated and inactivated AU states. Each local AU description contains multiple muscle details with potential associations according to FACS (Ekman & Rosenberg, 1997). The global face description is generated from the AU annotations and FACS (Ekman & Rosenberg, 1997), which only focus on the activated AU muscles. The examples are shown in Figure 5.5. We evaluated the model using the common 3-fold subject-exclusive cross-validation protocol (Ge, Wan, et al., 2021; X. Li, Zhang, Zhang, et al., 2023; Shao et al., 2018, 2021).

5.4.2 Training strategy.

The whole end-to-end network is implemented with PyTorch on a single NVIDIA RTX 3090Ti GPU using AdamW solver with β_1 = 0.9, β_2 = 0.999, and a weight decay of 0.0005. Maximum epochs are set to 15 with a batch size of 64. During training process, aligned faces are randomly cropped into 224×224 and horizontally flipped. We randomly employ the cutout augmentation (DeVries & Taylor, 2017) to overcome overfitting and improve the robustness during training. The stem extractor (Swin-Transformer-base) is pre-trained on ImageNet. The dimensionality of the global feature is mapped from 1024 to 512 (d=512), matching the hidden state dimension (h=512).

5.4.3 State-of-the-art Comparisons

We perform extensive experiments to compare our VL-FAU with mainstream FAU recognition studies and the latest state-of-the-art methods on two widely used FAU benchmarks in Table 5.1 and Table 5.2, which mainly includes JAA-Net (Shao et al., 2018), LGRNet (Ge, Wan, et al., 2021), UGN-B (T. Song, Chen, et al., 2021), HMP-PS (T. Song, Cui, Zheng, & Ji, 2021), FAU-Trans (Jacob & Stenger, 2021), ME-GraphAU (Luo et al., 2022), KDSRL (Chang & Wang, 2022), KS(X. Li, Zhang, Wang, & Yin, 2023), AAR (Shao, Zhou, Cai, Zhu, & Yao, 2023), SMA-ViT(X. Li, Zhang, Zhang, et al., 2023) and SEV-Net (H. Yang et al., 2021). The best and second-best results are bold and underlined.

Table 5.1: Comparisons of AU recognition for 8 AUs on DISFA in terms of F1-frame score (in %).

Method	AU Index								**Avg.**
	1	2	4	6	9	12	25	26	
JAA-Net$_{(ECCV2019)}$	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
UGN-B$_{(AAAI2021)}$	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
HMP-PS$_{(CVPR2021)}$	21.8	48.5	53.6	56.0	58.7	57.4	55.9	56.9	61.0
FAU-Trans$_{(CVPR2021)}$	46.1	48.6	72.8	**56.7**	50.0	72.1	90.8	55.4	61.5
ME-GraphAU$_{(IJCAI2021)}$	54.6	47.1	72.9	54.0	55.7	76.7	91.1	53.0	63.1
KDSRL$_{(CVPR2022)}$	60.4	59.2	67.5	52.7	51.5	76.1	91.3	57.7	64.5
KS$_{(ICCV2023)}$	53.8	**59.9**	69.2	54.2	50.8	75.8	92.2	46.8	62.8
AAR$_{(TIP2023)}$	**62.4**	53.6	71.5	39.0	48.8	76.1	91.3	**70.6**	64.2
SMA-ViT$_{(TAC2023)}$	51.2	49.3	64.7	48.3	50.6	**87.6**	85.1	61.2	62.2
VL-FAU(ours)	60.9	56.4	**74.0**	46.3	**60.8**	72.4	**94.3**	66.5	**66.5**
SEV-Net$_{(CVPR2021)}$	55.3	53.1	61.5	**53.6**	38.2	71.6	95.7	41.5	58.8
VL-FAU(ours)	**60.9**	**56.4**	**74.0**	46.3	**60.8**	**72.4**	**94.3**	**66.5**	**66.5**

Table 5.2: Comparisons with state-of-the-art methods for 12 AUs on BP4D in terms of F1-frame(in %).

Method	12 AUs												**Avg.**
	1	2	4	6	7	10	12	14	15	17	23	24	
JAA-Net$_{(ECCV2019)}$	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
LGRNet $_{(FG2021)}$	50.8	47.1	57.8	77.6	77.4	84.9	88.2	66.4	49.8	61.5	46.8	52.3	63.4
UGN-B $_{(AAAI2021)}$	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
HMP-PS$_{(CVPR2021)}$	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	49.9	54.5	63.4
FAU-Trans$_{(CVPR2021)}$	51.7	49.3	61.0	77.8	79.5	82.9	86.3	67.6	51.9	63.0	43.7	56.3	64.2
ME-GraphAU$_{(IJCAI2022)}$	53.7	46.9	59.0	78.5	80.0	84.4	87.8	67.3	52.5	63.2	50.6	52.4	64.7
KDSRL$_{(CVPR2022)}$	53.3	47.4	56.2	79.4	80.7	**85.1**	**89.0**	**67.4**	**55.9**	61.9	48.5	49.0	64.5
KS$_{(ICCV2023)}$	55.3	48.6	57.1	77.5	**81.8**	83.3	86.4	62.8	52.3	61.3	51.6	**58.3**	64.7
AAR$_{(TIP2023)}$	53.2	47.7	56.7	75.9	79.1	82.9	88.6	60.5	51.5	61.9	51.0	56.8	63.8
SMA-ViT$_{(TAC2023)}$	52.7	45.6	59.8	**83.8**	79.2	83.5	87.2	64.0	54.1	61.2	**52.6**	**58.3**	65.2
VL-FAU(ours)	**56.3**	**49.9**	**62.6**	79.5	80.1	82.6	88.6	66.8	51.3	**63.5**	51.3	57.1	**65.8**
SEV-Net$_{(CVPR2021)}$	58.2	50.4	58.3	**81.9**	73.9	**87.8**	87.5	61.6	**52.6**	62.2	44.6	47.6	63.9
VL-FAU(ours)	**56.3**	**49.9**	**62.6**	79.5	**80.1**	82.6	88.6	**66.8**	51.3	**63.5**	**51.3**	**57.1**	**65.8**

Quantitative comparison on DISFA: We compare our proposed VL-FAU with its counterpart in Table 5.1 and Table 5.3. Our VL-FAU outperforms mainstream studies with impressive margins. In particular, compared with the state-of-the-art AAR (Shao et al., 2023), which joint surprised local attention maps to obtain the multi-branch attention-aware AU representation, our VL-FAU increases the average F1-frame by 2.3% and shows clear improvements for most annotated AU categories. Compared with the best model KDSRL (Chang & Wang, 2022) on DISFA,

Table 5.3: Comparisons with state-of-the-art methods on DISFA and BP4D in terms of Accuracy(in %).

Method	8 AUs (DISFA)								**Avg.**	12 AUs (BP4D)												**Avg.**
	1	2	4	6	9	12	25	26		1	2	4	6	7	10	12	14	15	17	23	24	
JAA-Net	93.4	96.1	86.9	91.4	95.8	91.2	93.4	93.2	92.7	74.7	80.8	80.4	78.9	71.0	**80.2**	85.4	64.8	83.1	**73.5**	82.3	85.4	78.4
JÂ	**97.0**	**97.3**	88.0	92.1	95.6	92.3	94.9	**94.8**	94.0	75.2	80.2	82.9	79.8	72.3	78.2	86.6	**65.1**	81.0	72.8	**82.9**	86.3	78.6
UGN-B	95.1	93.2	88.5	**93.2**	**96.8**	**93.4**	94.8	93.8	93.4	78.6	80.2	80.0	76.6	72.3	77.8	84.2	63.8	**84.0**	72.8	82.8	**86.4**	78.2
VL-FAU(ours)	96.5	96.9	**92.0**	91.0	96.3	91.8	**96.7**	93.0	**94.3**	**79.1**	**82.3**	**83.0**	**80.5**	**77.4**	78.7	**86.8**	64.9	82.9	73.4	82.6	86.3	**79.8**

the average F1-frame score of our VL-FAU is also improved from 64.5% to 66.5%. Furthermore, we achieve the best performance in terms of average accuracy in Table 5.3, compared with all methods.

Quantitative comparison on BP4D: FAU recognition results by different methods on BP4D are shown in Table 5.2 and Table 5.3, where the proposed VL-FAU model achieves a new state of the art compared with all methods in terms of average F1-frame score. Our VL-FAU outperforms the baseline JAA-Net (Shao et al., 2018), which integrates AU detection and face alignment, in terms of average F1-frame and accuracy by 5.8% and 1.4%, respectively. This is mainly because JAA-Net focuses on the local AU regions based on the detected landmarks, resulting in poor distinguishability between different AUs, especially for the same individual with subtly different face states, which can be improved by our method. Moreover, VL-FAU achieves the best or second-best F1 and accuracy scores in recognizing most of the 12 AUs in BP4D, outperforming other state-of-the-art methods.

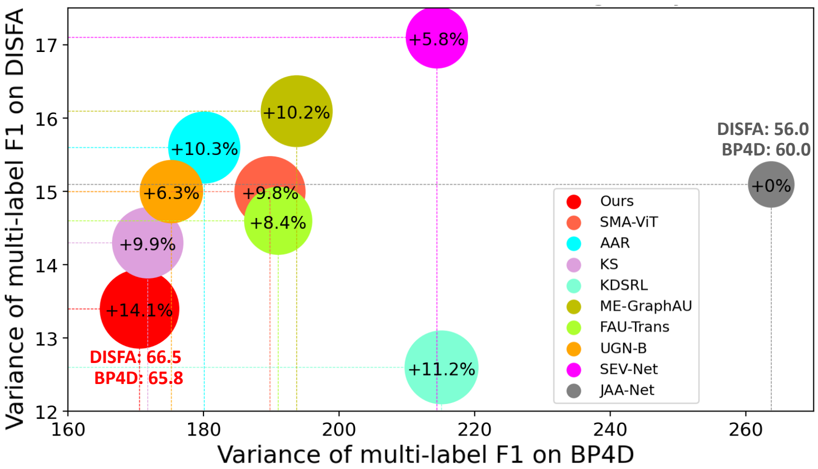


Figure 5.3: Multi-Label Performance Balancing Analysis. X-axis and Y-axis denote the variances of multi-label F1 scores on BP4D and DISFA, respectively. Circle size indicates the total relative performance improvement (%) compared with JAA-Net on BP4D and DISFA.

In addition, compared with SEV-Net (H. Yang et al., 2021), which prior encodes the pre-provided linguistic descriptions into image features, our VL-FAU achieves 7.7% and 1.9% higher average F1-frame scores on DISFA and BP4D, respectively. Experimental results demon-

Table 5.4: Effectiveness of key components of VL-FAU evaluated on BP4D in terms of F1-frame score (in %) of FAU recognition and top-5 accuracy (in %) of local and global language generation models.

Model		Setting			AU Index												Avg.	LLGA	GLGA
		MARL	LLGA	GLGA	1	2	4	6	7	10	12	14	15	17	23	24		Acc.	Acc.
VL-FAU	①	-	-	-	50.0	46.3	60.3	78.0	80.0	83.8	88.6	64.1	50.2	64.0	50.1	56.5	64.3	-	-
	②	√	-	-	50.0	45.5	60.0	79.6	80.1	83.1	88.7	66.5	51.0	63.3	53.6	55.3	64.7	-	-
	③	√	√	-	51.4	48.2	60.2	79.1	80.8	83.4	88.6	65.0	52.4	65.6	52.1	57.0	65.3	86.3	-
	④	√	-	√	54.8	47.4	61.2	79.2	79.4	84.1	88.8	63.8	52.2	65.2	50.6	55.6	65.2	-	64.4
	⑤	√	√	√	56.3	49.9	62.6	79.5	80.1	82.6	88.6	66.8	51.3	63.5	51.3	57.1	**65.8**	86.6	64.7

strate the effectiveness of VL-FAU in improving AU recognition accuracy on DISFA and BP4D by our proposed joint learning with language generation. Besides, as shown in Figure 5.3, we also provide the multi-label performance balancing analysis on DISFA and BP4D. Although the performance varies greatly between different AUs due to the inherent category imbalance in datasets, our VL-FAU still achieves a better performance-balance than existing methods and maintains the best overall results.

5.4.4 Ablation Studies

We perform extensive ablation studies on BP4D to investigate how each component affects the overall performance of the proposed VL-FAU. Due to space limitations, we do not show the ablation results for DISFA, but it is consistent with BP4D. Table 5.4 presents component ablation studies focusing on the various modules within VL-FAU, including (1) multi-level AU representation learning (MARL), (2) local language generation auxiliary (LLGA), and (3) global language generation auxiliary (GLGA). In addition, Figure 5.4 gives a qualitative analysis of local and global language generations.

(1) Multi-level AU Representation Learning (MARL). Compared with the baseline model ①, the result has an improvement of average F1-frame from 64.3% to 64.7% when considering the proposed multi-level AU representation learning (indicated variant ② with MARL). It indicates that the dual-level individual AU refinement based on the multi-scale stem feature combination could get richer and more fine-grained AU features and hence improve recognition performance.

(2) Local Language Generation Auxiliary (LLGA). In Table 5.4, we test the contributions of local language generation auxiliary (LLGA) of our VL-FAU model. Compared with variant ② and variant ③, after we provide local language generation auxiliary for each AU branch, the average F1-frame score has been improved from 64.7% to 65.3%. In addition, most of the 12 AUs annotated in BP4D achieve significant improvements. These observations demonstrate that by providing each AU branch with local language generation as an explicit auxiliary semantic supervision, the discriminative ability between AUs becomes stronger due to the gain of language expressiveness rather than single visual appearance features.

(3) Global Language Generation Auxiliary (GLGA). Similarly, we conduct a comparison between variant ② and variant ④ to verify the effectiveness of the proposed global language generation auxiliary (GLGA) for whole-face representation learning. Results in Table 5.4 show that the proposed VL-FAU facilitates the target FAU recognition task when using GLGA (indicated variant ④). In particular, the averaged F1-frame score increases from 64.7 % to 65.2 % and most AUs achieve significant improvements. These validate the effectiveness of the proposed GLGA which provides better discriminability of global representations by focusing on the language semantics of activated facial AUs, especially for the same subject with subtly different AU states.

Finally, when both local and global language generation auxiliaries (indicated model ⑤) are considered, the proposed VL-FAU achieves the best recognition performance in terms of average F1, significantly better than the single variants (variant ③ and ④) and variant ② in Table 5.4. In addition, the local and global language generation performances also have certain improvements, achieving 86.6% and 64.7% on top-5 accuracy of word generation in LLGA and GLGA. These quantitative comparisons experimentally demonstrate that exploring explicit semantic-auxiliary supervisions for facial AU recognition is a beneficial way for discriminating different AU states under intra- and inter-subject.

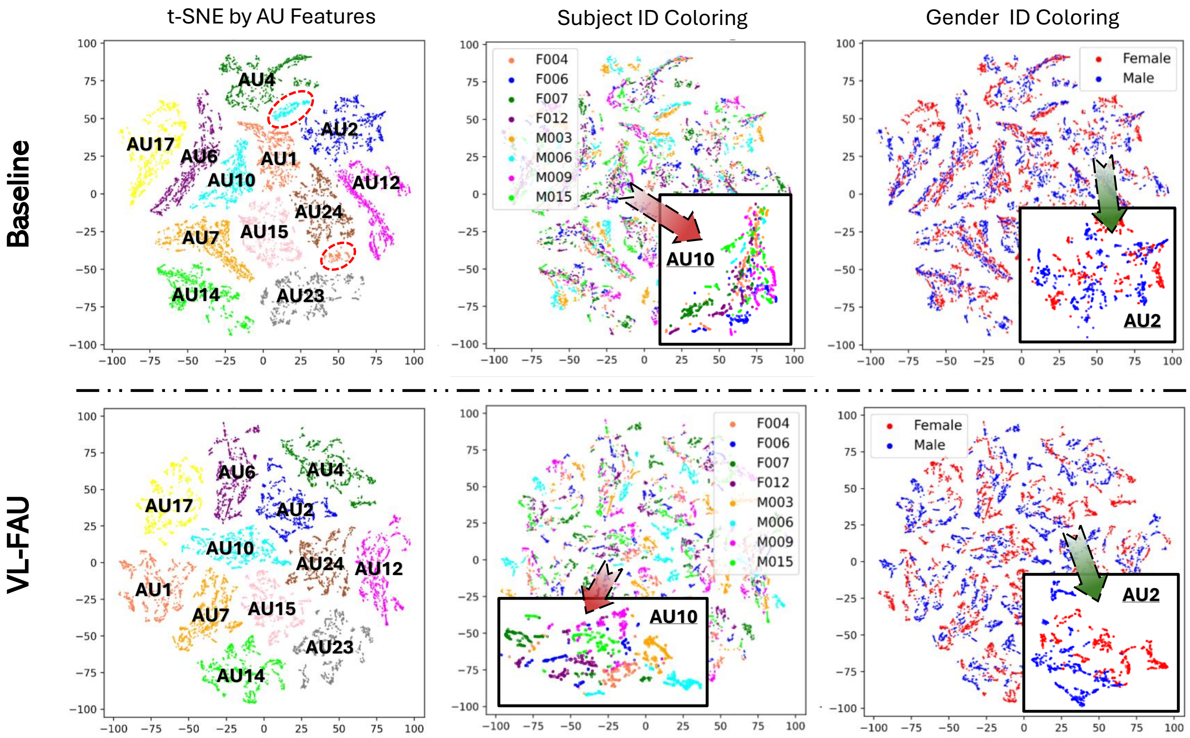
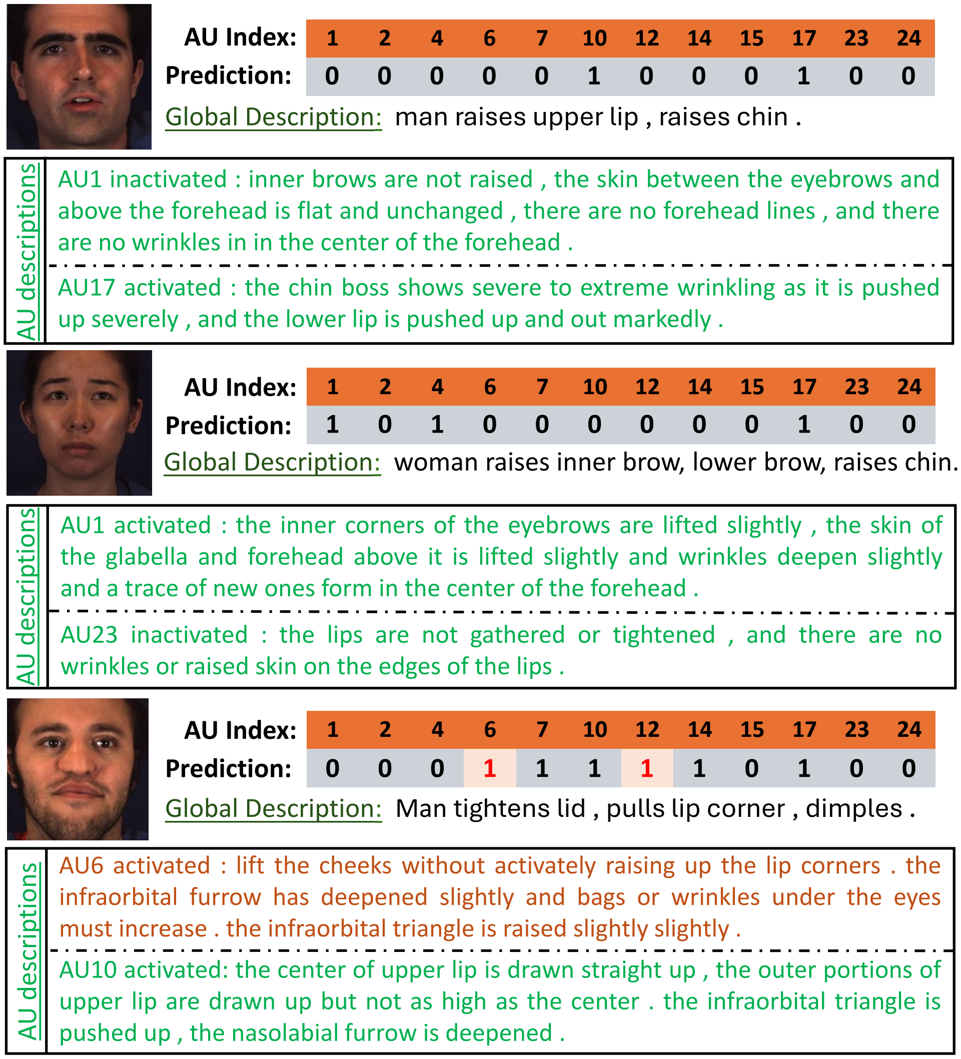


Figure 5.4: t-SNE visualization of the baseline model (w/o local and global language generation auxiliary) and full VL-FAU model on BP4D.

(4) Qualitative Analysis of Local and Global Generation Auxiliary. Besides the above quantitative comparisons, we further proved the detailed qualitative analysis of our main innovations – local and global generation auxiliaries for facial AU recognition. As shown in Figure 5.4, we use t-SNE (Van der Maaten & Hinton, 2008) to visualize the AU features learned by the proposed VL-FAU and the corresponding baseline without local and global generation auxiliaries (baseline ②). Note that, both models are used the same multi-branch networks with MARL. Specifically, we extract the AU features of 8 random gender-balanced subjects for clearer visualization from multiple AU branches before the final classification and then visualize them by t-SNE. We provide different colouring schemes in Figure 5.4 to analyze the impacts of VL-FAU on different aspects, including AU category, subject ID, and gender. (1) Comparisons of clustered AU features in the first column, our VL-FAU can better divide different AU features into different clusters, while the baseline is not sufficiently discriminative on some AU features (marked with red circles). Besides, we notice that AU features optimized from baseline are mapped in a narrow space compared with our VL-FAU. These observations indicate that by joining language generation auxiliary, our VL-FAU can maintain higher discriminability between multiple AU representations. (2) The second column shows the subject ID coloring results. Our VL-FAU distinguishes different subject AU features more widely within the same AU cluster, indicating that our VL-FAU provides higher discriminability between different subjects for facial AU recognition. (3) Due to the explicit language supervision of gender information in GLGA, VL-FAU can achieve clearer gender colorization results, as shown in the last column of Figure 5.4.

5.4.5 Visualization of Results

To further understand the quality of the proposed vision-language joint learning for FAU recognition and description generation, we visualize the predicted AU states and their corresponding local and global-level descriptions, as shown in Figure 5.5. Two positive examples from BP4D contain visualizations of different genders with different AU states. Compared with the mainstream paradigm, our VL-FAU provides explainable FAU recognition with language generations. In detail, local descriptions contain multiple detailed muscle changes with natural connections, improving intra-AU semantics and inter-AU distinguishability. In addition, the global descriptions contain diverse activated AU states with gender information, which can improve the inter-face distinguishability within and between subjects. Besides, we provide a bad case, which makes wrong predictions in two AUs. However, the global description ignores the misrecognition. Although AU6 is incorrectly predicted, the detailed description matches the facial expression, possibly due to labeling ambiguity. Overall, our VL-FAU can give better explainable facial AU recognition with explicit local and global language descriptions.



AU Index:	1	2	4	6	7	10	12	14	15	17	23	24
Prediction:	0	0	0	0	0	1	0	0	0	1	0	0

Global Description: man raises upper lip , raises chin .

AU1 inactivated : inner brows are not raised , the skin between the eyebrows and above the forehead is flat and unchanged , there are no forehead lines , and there are no wrinkles in in the center of the forehead .

AU17 activated : the chin boss shows severe to extreme wrinkling as it is pushed up severely , and the lower lip is pushed up and out markedly .

AU Index:	1	2	4	6	7	10	12	14	15	17	23	24
Prediction:	1	0	1	0	0	0	0	0	0	1	0	0

Global Description: woman raises inner brow, lower brow, raises chin.

AU1 activated : the inner corners of the eyebrows are lifted slightly , the skin of the glabella and forehead above it is lifted slightly and wrinkles deepen slightly and a trace of new ones form in the center of the forehead .

AU23 inactivated : the lips are not gathered or tightened , and there are no wrinkles or raised skin on the edges of the lips .

AU Index:	1	2	4	6	7	10	12	14	15	17	23	24
Prediction:	0	0	0	1	1	1	1	1	0	1	0	0

Global Description: Man tightens lid , pulls lip corner , dimples .

AU6 activated : lift the cheeks without activately raising up the lip corners . the infraorbital furrow has deepened slightly and bags or wrinkles under the eyes must increase . the infraorbital triangle is raised slightly slightly .

AU10 activated: the center of upper lip is drawn straight up , the outer portions of upper lip are drawn up but not as high as the center . the infraorbital triangle is pushed up , the nasolabial furrow is deepened .

Figure 5.5: Visualizations of our proposed explainable facial AU recognition (VL-FAU) with explicit local and global language descriptions on BP4D.

5.5 Conclusion and future work

In this chapter, we proposed a novel end-to-end vision-language joint learning (VL-FAU) for explainable FAU recognition along with language generations as explanations. As auxiliary supervisions, local and global language generations are joined into a multi-branch AU recognition network with multi-level AU representation learning. Local AU language generation provides explicit fine-grained semantic supervision for each AU classification with detailed language descriptions, improving the discrimination of inter-AU representations. Global language generation, employing multi-scale combined stem features, offers diverse semantic supervision for the whole facial feature to maintain diversity and distinction across intra- and inter-subject representation changes. Our VL-FAU finally provides predictions of AU states as well as interpretable language descriptions for individual AUs and global faces. Extensive experimental evaluations on DISFA and BP4D show that our VL-FAU outperforms state-of-the-art AU recognition methods with impressive margins.

VL-FAU introduces a traditional language model for explainable FAU recognition considering computational power and efficiency limitations. We believe our attempts provide new inspiration for multimodal multi-task joint training for explainable FAU recognition. In the future, we would like to investigate the further combination of FAU recognition with popular LLMs for more diverse and fine-grained explainable generations.

II. Visual Relational Reasoning and Embedding for Image-Sentence Retrieval

In the second part of our research, we aim to further investigate the importance of visual relational reasoning and embedding in multimodal tasks. To this end, we extend our study to the multimodal task of image-text retrieval, exploring the effectiveness of various novel multimodal relational reasoning and embedding approaches in enhancing cross-modal understanding and alignment performance. Further exploration of cross-modal task (image-sentence retrieval), together with the unimodal task (facial action unit recognition) in the first part, form a complete and comprehensive verification of the effectiveness of modality relational reasoning and embedding. Specifically, we propose a variety of novel relational reasoning and encoding structures for image-sentence retrieval, such as Structured Multi-modal Feature Embedding and Alignment (SMFEA (Ge, Chen, et al., 2021)) in Chapter 6, Hybrid-modal Interaction with Multiple Relational Enhancements (*Hire* (Ge, Chen, et al., 2024)) in Chapter 7 and Visual Semantic-Spatial Self-Highlighting Model (3SHNet (Ge, Xu, et al., 2024)) in Chapter 8. These methods contribute to a more nuanced cross-modal alignment, enhancing the image-text retrieval performance by better capturing the complex context-aware relationships between visual and textual data.

Chapter 6

Structured Multi-modal Feature Embedding and Alignment

The current state-of-the-art image-sentence retrieval methods implicitly align the visual-textual fragments, like regions in images and words in sentences, and adopt attention modules to highlight the relevance of cross-modal semantic correspondences. However, the retrieval performance remains unsatisfactory due to a lack of consistent representation in both semantics and structural spaces. In this work, we propose to address the above issue from two aspects: (i) constructing intrinsic structure (along with relations) among the fragments of respective modalities, *e.g.*, *"dog $\rightarrow$ play $\rightarrow$ ball"* in semantic structure for an image, and (ii) seeking explicit inter-modal structural and semantic correspondence between the visual and textual modalities.

In this paper, we propose a novel **S**tructured **M**ulti-modal **F**eature **E**mbedding and **A**lignment (SMFEA) model for image-sentence retrieval. In order to jointly and explicitly learn the visual-textual embedding and the cross-modal alignment, SMFEA creates a novel multi-modal structured module with a shared context-aware referral tree. In particular, the relations of the visual and textual fragments are modeled by constructing Visual Context-aware Structured Tree encoder (VCS-Tree) and Textual Context-aware Structured Tree encoder (TCS-Tree) with shared labels, from which visual and textual features can be jointly learned and optimized. We utilize the multi-modal tree structure to explicitly align the heterogeneous image-sentence data by maximizing the semantic and structural similarity between corresponding inter-modal tree nodes. Extensive experiments on Microsoft COCO and Flickr30K benchmarks demonstrate the superiority of the proposed model in comparison to the state-of-the-art methods.

6.1 Introduction

Cross-modal retrieval, *a.k.a* image-sentence retrieval, plays an important role in real-world multimedia applications, *e.g.*, queries by images in recommendation systems, or image-sentence retrieval in search engines. Image-sentence retrieval aims at retrieving the most relevant images

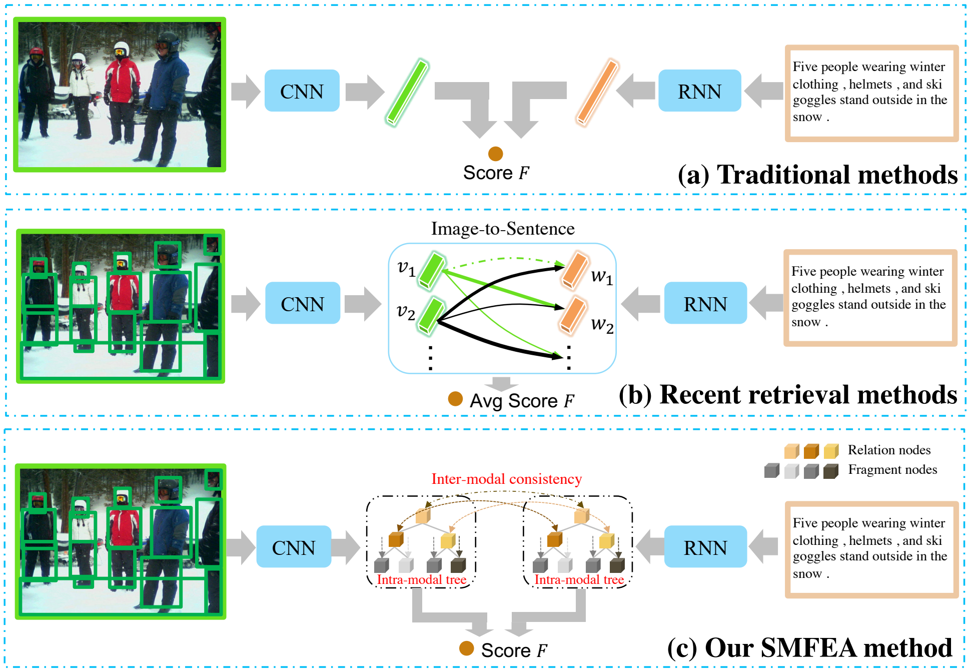


Figure 6.1: Illustration of the different schemes: (a) the traditional instance-level alignment methods, (b) the recent fragment-level alignment methods, and (c) our SMFEA method. Compared with (a) and (b), our SMFEA in (c) exploits intra-modal relations of visual/textual fragments via a tree encoder and aligns them explicitly in the corresponding nodes in two modal trees.

(or sentences) given a query sentence (or image), and has attracted increasing research attention recently (Faghri et al., 2017; Frome et al., 2013; Huang et al., 2018; K.-H. Lee et al., 2018; C. Liu et al., 2020; H. Liu et al., 2018; H. Wang et al., 2020; L. Wang et al., 2016). Its main challenge lies in capturing the effective alignment (both in semantics and structural spaces) between the visual and textual modalities.

Typically, traditional approaches (Faghri et al., 2017; Frome et al., 2013; L. Wang et al., 2016) model the cross-modal alignment on an instance level by directly extracting the global instance-level features of the visual and the textual modalities via Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) respectively, and estimate the visual-textual similarities based on the global features, as shown in Figure 6.1 (a). However, as argued in Frome et al. (2013), cross-modal semantic gap is harder to bridge with solely the global characteristics of images and sentences. To address this issue, recent works (Huang et al., 2018; K.-H. Lee et al., 2018; C. Liu et al., 2019) extract the features of the visual and textual fragments, i.e., object regions in images and words in sentences, and align the visual and the textual fragment features via a soft attention mechanism, as shown in Figure 6.1 (b). However, there are two key defects with the above fragment-level alignment approaches. On one hand, these approaches neglect the intra-modal contextual semantic and structural relations of the fragments, thus failing to capture the semantics of the images or the sentences effectively. On the other hand, these approaches make the inter-modal fragment alignment implicitly with the many-to-many matching across the visual and textual modalities and with this, it is difficult to improve the consistency of semantic and structural representation between modalities.

6.1.1 Motivation

In this paper, we argue that the key issues in image-sentence retrieval can be addressed by: (i) constructing the intra-modal context relations of the visual/textual fragments with a structured embedding module; and (ii) aligning the inter-modal fragments and their relations explicitly using a shared semantic structure, as shown in Figure 6.1 (c). We propose a novel structured multi-modal feature embedding and alignment model with visual and textual context-aware tree encoders (VCS-Tree and TCS-Tree) for image-sentence retrieval, termed SMFEA. On one hand, the context-aware structured tree encoders are created for both modalities in order to capture the intrinsic structured relation among the fragments of visual/textual modalities (which we call context-aware structure information). We use a shared referral tree as a supervisor for both modalities, which contains rich semantic content and structure information in an in-order traversal way (which we call semantics and structural spaces). On the other hand, the shared referral tree can also improve the inter-modal alignment in semantic correspondence between nodes in two tree encoders of both modalities. Moreover, we use the KL-divergence between two spaces to optimize the unified joint embedding space by aligning semantic distributions of tree nodes between modalities, which improves the robustness and fault tolerance of multi-modal feature representations.

6.1.2 Contribution

The contributions of this paper are as follows:

- We propose two context-aware structured tree encoders (VCS-Tree and TCS-Tree) to parse the intrinsic (within modality) relations among the fragments of respective modalities. Thus this leads to effective semantic representation for pair-wise alignment of image and sentence.
- We mine the explicit semantic and structural consistency of inter-modality corresponding tree nodes in visual and textual tree structures to align the heterogeneous cross-modality features.
- The proposed SMFEA outperforms the state-of-the-art approaches for image-sentence retrieval on two benchmarks, *i.e.*, Flickr30K and Microsoft COCO.

6.2 Related Work

6.2.1 Image-Sentence Retrieval

The key issue in image-sentence retrieval task is to measure the visual-textual similarity between an image and a sentence. From this perspective, most existing image-sentence retrieval methods can be roughly categorized into two groups: global semantic embedding alignment-based methods (Frome et al., 2013; Mao et al., 2014; Vendrov et al., 2015; L. Wang et al., 2016; S. Wang, Chen, Chen, & Shi, 2018); and local semantic embedding alignment-based methods (Karpathy & Fei-Fei, 2015; Karpathy et al., 2014; K.-H. Lee et al., 2018; Z. Niu, Zhou, Wang, Gao, & Hua, 2017). As for the global embedding, Frome et al. (2013) utilized a linear mapping network to unify the whole image and the full-text features. Further the distance between any mismatched pair was increased than that between a matched pair using a ranking loss function. For local semantic embedding alignment, DVSA (Karpathy & Fei-Fei, 2015) first adopted R-CNN to detect salient objects and inferred latent alignments between word-level textual features in sentences and region-level visual features in images. Moreover, an attention mechanism (K.-H. Lee et al., 2018; Nam et al., 2017; Y. Wang et al., 2019) has been applied to capture the fine-grained interplay between images and sentences for the image-sentence retrieval task. However, all the above methods fail to take into consideration the high-level representation of semantics and structure, such as concepts extracted from images or sentences and their structural relationships, thus only allowing implicit inference of correspondence between the concepts. K. Li et al. (2019) proposed a visual semantic reasoning network with graph convolutional network (GCN) to generate a visual representation that captures key concepts of a scene. H. Wang et al. (2020) proposed to integrate commonsense knowledge into the multi-modal representation learning for visual-textual embedding. To create consensus-aware concept (CAC) representations that are concepts without any ambiguity in both modalities, they used a co-occurrence concept correlation graph. However, we argue that merely predicting the consensus concept to align the visual and textual embedding space is not enough. Ignoring the intrinsic semantic structure and inter-modal structure alignment are detrimental to the performance of the model. Hence there is a need for a consistent multi-modal explicit structure embedding, such as a multi-modal structured semantic tree.

6.2.2 Structured Feature Embedding

In terms of structured feature embedding, exiting works for multimedia data (F. Chen et al., 2019; F. Chen, Ji, Su, Wu, & Wu, 2017; T. Chen & Luo, 2020) employed different structures, e.g., chain, tree, and graph. F. Chen et al. (2019, 2017) proposed to enhance the visual representation for image captioning task by a linear-based structured tree model. However, because of the simple linear-based tree model in these schemes (F. Chen et al., 2019, 2017), limited contextual information is transferred between different layers and without using any attention mechanism. T. Chen and Luo (2020) applied a chain structure model using an RNN for visual embeddings, which unfortunately ignores the underlying structure. Besides, the single-modal structured embedding models failed to capture the interaction between the modalities. Recently, GCN is employed in K. Li et al. (2019); C. Liu et al. (2020) to improve the interaction and integrate different item representations by a learned graph. For instance, C. Liu et al. (2020) proposed to learn the correspondence of objects and relations between modalities by two different visual and textual structure reasoning graphs, however, fails to unify the precise pairing of the two modal structures.

In contrast to previous studies, SMFEA models the relation structure of intra-modal fragments/words by the use of a fixed contextual structure and aligns two modalities into a joint embedding space in terms of semantics and structure. The most relevant existing work to ours is H. Wang et al. (2020), which aligns the visual and textual representations through measuring the consistency of the corresponding concepts in each modality. However unlike H. Wang et al. (2020), SMFEA approaches this in a novel way by exploiting the learned multi-modal semantic trees to enhance the structured embedding of the visual and textual modalities. By aligning the inter-modal semantics and structure consistently, the joint embedding space is obtained to reduce the heterogeneous (inter-modality) semantic gap. Doing so allows us to provide more robustness than H. Wang et al. (2020), which also improves the interpretability of the model.

6.3 The Proposed Method – SMFEA

The overview of SMFEA is illustrated in Figure 6.2. We will first describe the multi-modal feature extractors (① in Figure 6.2) in our work in Section 6.3.1. Then, the context-aware representation module is introduced in detail in Section 6.3.2 with context-aware structured tree encoders (② in Figure 6.2) and the consensus-aware concept (CAC) representation learning module (③ in Figure 6.2). Finally, the objective function is discussed in Section 6.3.3. For clarity, the main notations and their definitions throughout the paper are shown in Table 6.1.

6.3.1 Multi-modal Feature Extractors

Our multi-modal feature extractors include two components to encode the region-level visual representations and word-level textual representations into the instance-level multi-modal features.

Visual Representations

To better represent the salient entities and attributes in images, we take advantage of the bottom-up-attention network (Anderson et al., 2018) to embed the extracted sub-regions in an image.

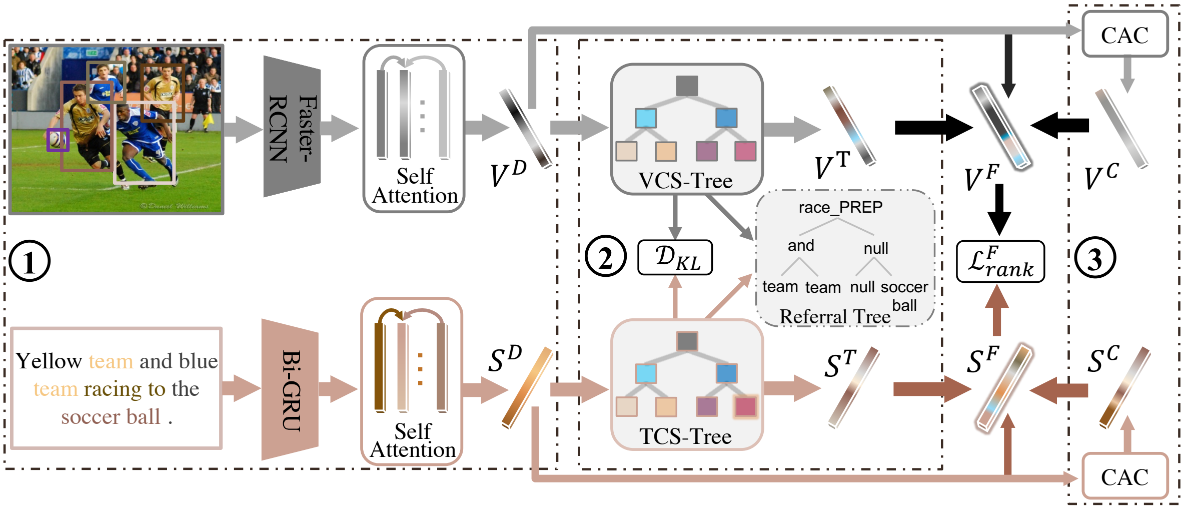


Figure 6.2: An illustration of our **S**tructured **M**ulti-modal **F**eature **E**mbedding and **A**lignment (SMFEA) for image-sentence retrieval (best viewed in color).

Specifically, given an image I, we extract a set of image fragment-level sub-region features $V = \{v_1, \cdots, v_K\}$, $v_j \in \mathbb{R}^{2048}$, where K is the number of selected sub-regions, from the average pooling layer in Faster-RCNN Ren et al. (2015).

Furthermore, we employ the self-attention mechanism (Vaswani et al., 2017) to refine the instance-level latent embeddings of sub-region features for each image, thus concentrating on the salient information exploited by the fragment-level features. In particular, following Vaswani et al. (2017), the fragment visual features $V = \{v_1, \cdots, v_K\}$ are used as the key and value items. And the initialization of instance-level features $\bar{V}$, embedded by the mean of region features, serves as the query item to fuse the important fragment features with different learning weights α as new instance-level visual representation V^D. These can be formulated as:

$$\bar{V} = \frac{1}{K} \sum\nolimits_{i=1}^{K} v_i \tag{6.1}$$

$$\alpha_i = \frac{exp(\bar{V} v_i)}{\sum_{i=1}^{K} exp(\bar{V} v_i)} \tag{6.2}$$

$$V^D = \sum\nolimits_{i=1}^{K} \alpha_i v_i \tag{6.3}$$

Word-level Textual Representations

For sentences, word-level textual representations are encoded by a bi-directional GRU network (Schuster & Paliwal, 1997). In particular, we first represent each word w_j in sentence $S = [w_1, \cdots, w_N]$ with length N as a one-hot vector being the cardinality of the D_v-length vocabulary dictionary. The one-hot vector of w_j is projected into a fixed dimensional space $e_j = W_f w_j$ (W_f

Table 6.1: Main notations and their definitions.

Notation	Definition
I	an image
S	a sentence
K	the number of selected sub-regions in an image
N	the number of words in a sentence
v_i	the feature of i-th sub-region in an image
w_j^f	the feature of j-th word in a sentence
α	the learning weights of query and key items for two modalities
$\bar{V},\bar{S}$	the query representations of image and sentence
V^D,S^D	the instance-level visual and textual representations
$\hat{V}^D,\hat{S}^D$	the mapped instance-level visual and textual representations
M	the number of VCS-Tree/TCS-Tree nodes
$T(t)$	the set of children of tree node t
i_t,f_t,o_t	the input gate, forget gate and output gate in tree model of $T(t)$
$\tilde{c}_t,c_t,h_t$	the candidate cell values, cell state and hidden state of tree node t
h^V,h^S	the set of hidden states of visual and textual tree nodes
y^V,y^S	the predicted scores of the fragment categories for VCS-Tree and TCS-Tree
z^V,z^S	the ground truth of the fragment categories in the shared referral tree for VCS-Tree and TCS-Tree
V^T,S^T	the context-aware structured enhancement embeddings of two modalities
V^C,S^C	the CAC representations of two modalities
β_d,β_t,β_c	the tuning parameters to balance three types of features of final integrated visual and textual features
V^F,S^F	the final embeddings of two modalities
P^V,P^S	the probability distributions on visual and textual predicted fragment/relation category

denotes the mapping parameter) and then sequentially fed into the bi-directional GRU. The final hidden representation for each word is the average of the hidden vectors in both directions as follows:

$$w_j^f = \frac{\overrightarrow{GRU}(e_j) + \overleftarrow{GRU}(e_j)}{2} \tag{6.4}$$

where $j \in [1,N]$. Similar to the procedure in the visual branch, we finally get the refined instance-level textual representation S^D of a sentence based on the word-level textual features.

6.3.2 Context-Aware Representation

Our aim is to construct the intrinsic relations among the fragments of the visual/textual modality. Hence, we construct two novel context-aware structured trees from instance-level visual and textual features, with the help of a shared referral tree. To facilitate the inter-modal semantics and structure correspondence, with the aim to bridge the heterogeneous (i.e., between modalities) semantic gap, our model aligns semantic categories of the corresponding modality nodes.

Shared Referral Tree Encoder

During the training we construct, for each of the modalities, context-aware structured trees of three-layer tree structures, supervised by shared labels (called shared referral tree). The shared referral tree is constructed by Stanford Parser (Socher, Lin, Ng, & Manning, 2011) from sen-

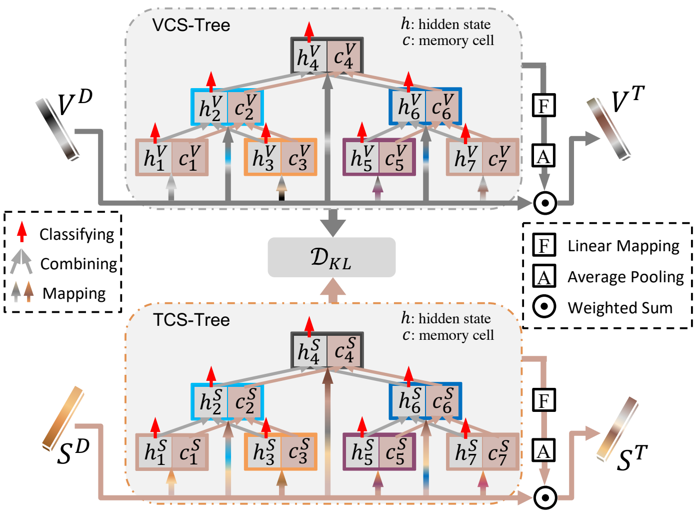


Figure 6.3: The architecture of VCS-Tree and TCS-Tree in two branches. The fundamental operations for both tree encoders include mapping, combining, and classifying. For each corresponding parsing node in multi-modal tree encoders, we utilize the same semantic category to guarantee semantic correctness. We employ the KL-divergence to guarantee consistency between cross-modal node structures. Nodes with indexes 1∼7 in each modal tree encoder (best viewed in color).

tence, and the pos-tag tool and lemmatizer tool in NLTK (Loper & Bird, 2002) are applied to whiten the source sentences to reduce the irrelevant words and noise configurations. As shown in the middle of Figure 6.2, it is a fixed-structure, three-layer binary tree, which only contains nouns (or noun pair, adjective-noun pair), verbs, coverbs, prepositions, and conjunctions. “Null" in the referral tree means the ignorable node or the unknown category (not in the entity or relation dictionaries). Only nouns are regarded as fragments and used as leaf nodes in the subsequent training. Correct semantic content can be represented by the shared referral tree in in-order traversal way. A referral tree is created for each sentence and the corresponding image pair.

Context-aware Structured Tree Encoders

We construct a visual context-aware structured tree (VCS-Tree) and a textual context-aware structured tree (TCS-Tree) to parse the intra-modal structural relations of the respective fragments/words. Moreover, the VCS-Tree and TCS-Tree are utilized to align the inter-modal nodes between the images and sentences. As shown in Figure 6.3, the tree structure of two modalities is the same, where each modality tree parses the instance-level features V^D/S^D into a three-layer architecture with seven nodes (same as referral tree), of which four leaf nodes are used to parse fragments in the 1^{st} layer and three parent nodes to parse relations in the 2^{nd} and 3^{rd} layers, to organise the semantic and structural relations of an image or a sentence. There are two main reasons why we adopt this fixed structure: (i) inspired by F. Chen et al. (2019, 2017), the tree with seven nodes can express the main semantic content of each image-sentence pair; and (ii) it is suitable for improving the consistency of coarse semantics and structural representation between modalities, thereby improving the robustness and interpretability of the model. For simplicity, we will only introduce the detailed structure of VCS-Tree and do not repeat the details for the TCS-Tree.

As shown in the top branch of Figure 6.3, instance-level visual feature V^D is first mapped into different semantic spaces by a linear mapping function with the parameter $W_i^o \in \mathbb{R}^{2048\times D_v}$, which serve as the inputs to different layers in VCS-Tree:

$$\hat{V}_i^D = V^D W_i^o, i \in \{1,2,...,7\}, \tag{6.5}$$

For simplicity, we do not explicitly represent the bias terms in our paper.

For the VCS-Tree, we broadcast the context information between different layer nodes in a novel $LSTM$-based ternary tree encoder with a fixed structure. It can get final structured tree embedding of the image supervised by the shared referral tree. In particular, we describe the updating of a parent node t in VCS-Tree, where the detailed computation is described in Eq.(6.6 - 6.11). $T(t)$ denotes the set of children of node t. The process can be formulated as:

$$i_t = \sigma(W^i \hat{V}_t^D + U^f \tilde{h}_t), \tag{6.6}$$

$$f_t = \sigma(W^f \hat{V}_t^D + U^i \tilde{h}_t), \tag{6.7}$$

$$o_t = \sigma(W^o \hat{V}_t^D + U^o \tilde{h}_t), \tag{6.8}$$

$$\tilde{c}_t = \tanh(W^u \hat{V}_t^D + U^u \tilde{h}_t), \tag{6.9}$$

$$c_t = i_t \odot \tilde{c}_t + f_t \odot \sum (f_k \odot c_k), \tag{6.10}$$

$$h_t = o_t \odot \tanh(c_t), \tag{6.11}$$

where i_t, f_t, o_t denote the input gate, forget gate and output gate, $\tilde{c}_t, c_t, h_t$ are the candidate cell value, cell state and hidden state of tree node t, σ is the sigmoid function, $\odot$ is the element-wise multiplication, all W^* and U^* are learning weight matrices, $\tilde{h}_t$ is the summing of hidden states of children nodes $T(t)$, and $T(k)$ are sub-trees of $T(t)$ in Eq.(6.10). In this way, the features of the parent nodes in higher layers can contain the rich context-aware semantic information by the $LSTM$-based attention mechanism, which combines the children nodes information as well as the leaf nodes. Finally, each node is classified into the fragment/relation category by the Softmax classifier. And the sum of all node hidden states in the tree in an in-order traversal manner, which are mapped into the same dimension with original visual features as the structured tree enhancement embedding V^T as follows:

$$y_{t^1}^V = \text{Softmax}(W_e h_{t^1}^V), t^1 \in \{1,3,5,7\}, \tag{6.12}$$

$$y_{t^{2,3}}^V = \text{Softmax}(W_r h_{t^{2,3}}^V), t^2 \in \{2,6\}, t^3 \in \{4\}, \tag{6.13}$$

$$V^T = \sum (W_i h_i^V), i \in \{1,2,...,7\}, \tag{6.14}$$

where y_{t1}^{V} and $y_{t2,3}^{V}$ denote the predicted scores of the fragment categories for 1^{st} layer and relation categories for 2^{nd} or 3^{rd} layers. W_e and W_r denote the mapping parameters for the fragment and relation categories according to these dictionaries F. Chen et al. (2017), respectively. W_i denotes the mapping parameters.

Likewise, our TCS-Tree with seven node structures takes the mapped instance-level textual feature $\hat{S}^{D}$ as the inputs. The structure of TCS-Tree is same with VCS-Tree and the final structured textual embedding S^{T} is obtained by the sum of the hidden states h^{S} of the TCS-Tree and the original instance-level feature S^{D}. Furthermore, the predicted probability vectors for seven nodes in different layers of textual fragment categories y_{t1}^{S} and relation categories $y_{t2,3}^{S}$ are obtained.

We capture the intra-modal context relations of the visual/textual fragments by minimizing the loss of the category classification. It can guarantee the correct semantic representation of the content of the image and corresponding sentence. Furthermore, we narrow the inter-modal distance of images and sentences by minimizing the loss of the Kullback Leibler(KL) divergence for both modality tree nodes probability distributions. Details are given in the Section 6.3.3.

CAC Representation Learning Module

Following H. Wang et al. (2020), we also exploit the commonsense knowledge to capture the underlying interactions among various semantic concepts by learning the dual modalities consensus-aware concept (CAC) representations V^{C}/S^{C}, which can improve the fine-grained semantic information of our context-aware representations to a certain extent. Due to space restrictions, we are not repeating the process in H. Wang et al. (2020).

Multiple Representations Fusing Module

To comprehensively characterize the semantic and structured expression for both the modalities, we combine the instance-level representations V^{D}/S^{D}, the context-aware structured enhancement features V^{T}/S^{T} and CAC representations V^{C}/S^{C} into fusing modalities representations V^{F}/S^{F} with simple weighted sum operation, as following:

$$V^{F} = \beta_d V^{D} + \beta_t V^{T} + \beta_c V^{C} \tag{6.15}$$

$$S^{F} = \beta_d S^{D} + \beta_t S^{T} + \beta_c S^{C} \tag{6.16}$$

where $\beta_d, \beta_t, \beta_c$ are the tuning parameters for balancing. This allows the SMFEA model to get rich semantic and structure representation for each modality and also keep cross-modal consistency of structure and semantics between the modalities.

6.3.3 Objective Function

In the above training process, all the parameters can be simultaneously optimized by minimizing a bidirectional triplet ranking loss (Faghri et al., 2017), where we exploit positive and negative samples and as follows:

$$\mathscr{L}_{rank}(I,S) = \sum_{(I,S)} [\nabla - \mathrm{Cos}(I,S) + \mathrm{Cos}(I,\bar{S})]_{+} + \sum_{(I,S)} [\nabla - \mathrm{Cos}(I,S) + \mathrm{Cos}(\bar{I},S)]_{+} \tag{6.17}$$

where ∇ is a margin constraint, $\mathrm{Cos}(\cdot,\cdot)$ indicates cosine similarity function, and $[\cdot]_{+} = \max(0,\cdot)$. Note that, (I,S) denotes the given matched image-sentence pair and its corresponding negative samples are denoted as $\bar{I}$ and $\bar{S}$, respectively.

Moreover, we minimize the loss of the node category classification on both visual and textual context-aware structured tree encoders to improve the structured semantic referring ability, using a cross-entropy loss as follows:

$$\mathscr{L}_{CE}(V^{D},S^{D}) = -\textstyle\sum_{i=1}^{M} (\mathrm{CE}(y_i^V, z_i^V) + \mathrm{CE}(y_i^S, z_i^S)) \tag{6.18}$$

where y_i^V and y_i^S indicate the predicted fragment/relation categories of the i-th node in three layers of VCS-Tree and TCS-Tree with M nodes, respectively. z^V and z^S are category labels of the nodes, as detailed in Section 6.3.2. And to further narrow the semantic gap between modalities, we employ the Kullback Leibler (KL) divergence to regularize the probability distributions on visual and textual predicted fragment/relation category scores, which is defined as:

$$\mathscr{D}_{KL}(P^V \parallel P^S) = \textstyle\sum_{i=1}^{M} P_i^V log(P_i^V / P_i^S) \tag{6.19}$$

where P_i^V and P_i^S denote the predicted probability distributions of cross-modal corresponding tree nodes.

In this way, we utilize a shared referral tree to modal the intra-modal embedding explicitly and employ the fixed cross-modal tree alignment to guarantee the inter-modal consistency of the structure and semantics between images and sentences. Finally, the joint loss of the SMFEA model is defined as:

$$\mathscr{L} = \mathscr{L}_{rank}^{F}(V^{F},S^{F}) + \mathscr{L}_{CE}(V^{D},S^{D}) + \mathscr{D}_{KL}(P^V \parallel P^S) \tag{6.20}$$

Note that, we use the final fusing features V^F and S^F to calculate the similarity scores during the inference process.

6.4 Experiments

In this section, we report the results of experiments to evaluate the proposed approach, SMFEA. We will introduce experimental settings first. Then, SMFEA is compared with the state-of-the-art image-sentence retrieval approaches quantitatively. Finally, we qualitatively analyze the results in detail.

6.4.1 Implementation Details

Our model is trained on a single NVIDIA 2080Ti GPU with 11 GB memory. The whole network except the Faster-RCNN model is trained from scratch with the default initializer of PyTorch using ADAM optimizer (Kingma & Ba, 2014). The learning rate is set to 0.0002 initially with a decay rate of 0.1 every 25 epochs. The maximum epoch number is set to 50. The margin of triplet ranking loss ∇ is set to 0.2. The cardinality of our dictionary is 8481 for Flickr30K and 11353 for MS-COCO. The cardinalities of our fragment category and relation category are 1440 and 247, respectively. The dimensionality of word embedding space is set to 300, which is transformed to 1024-dimensional by a bi-directional GRU to get the word representation. For the region-level visual feature, 36 regions are selected with the highest class detection confidence scores. And then a full-connect layer is applied to transform these region features from 2048-dimensional to a 1024-dimensional (*i.e.*, D_v=1024). The dimension of the hidden states of nodes are set 128 in both VCS-Tree and TCS-Tree. Regarding CAC learning process, we set the value of the general parameters to be the same with H. Wang et al. (2020). We empirically set $\beta_d, \beta_t, \beta_c = 0.6, 0.2, 0.2$ in Eq.(6.15) and Eq.(6.16).

6.4.2 Comparison with State-of-the-art Methods

As in MIR literature, we follow the standard protocols for running the evaluation on the Flickr30K and MS-COCO datasets and hence for comparison purposes report the results of the baseline methods in Table 6.2 and Table 6.3, including (1) early works, *i.e.*, VSE++ (Faghri et al., 2017), SCO (Huang et al., 2018), SCAN* (K.-H. Lee et al., 2018), and (2) state-of-the-art methods, *i.e.*, PFAN (Y. Wang et al., 2019), VSRN* (K. Li et al., 2019), IMRAM* (H. Chen et al., 2020), MMCA (Wei et al., 2020), CAAN (Q. Zhang et al., 2020) and CVSE (H. Wang et al., 2020). Note that, the ensemble models with “*" are further improved due to the complementarity between multiple models. The best and second-best results are shown using bold and underline, respectively.

Quantitative Comparison on Flickr30K

Quantitative results on the Flickr30K 1K test set are shown in Table 6.2, where the proposed approach SMFEA outperforms the state-of-the-art methods with impressive margins for rSum.

Table 6.2: Comparisons of experimental results on Flickr30K 1K test set. $*$ indicates the performance of an ensemble model.

Method	Sentence Retrieval			Image Retrieval			rSum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
VSE++	52.9	79.1	87.2	39.6	69.6	79.5	407.9
SCAN*	67.4	90.3	95.8	48.6	77.7	85.2	465.0
PFAN	70.0	91.8	95.0	50.4	78.7	86.1	472.0
VSRN*	71.3	90.6	96.0	54.7	81.8	88.2	482.6
CAAN	70.1	91.6	**97.2**	52.8	79.0	87.9	478.6
CVSE	73.5	92.1	95.8	52.9	80.4	87.8	482.4
SMFEA(ours)	**73.7**	**92.5**	96.1	**54.7**	**82.1**	**88.4**	**487.5**

Table 6.3: Comparisons of experimental results on MS-COCO 1K test set. $*$ indicates the performance of an ensemble model.

Method	Sentence Retrieval			Image Retrieval			rSum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
VSE++	64.7	-	95.9	52.0	-	92.0	304.6
SCO	69.9	92.9	97.5	56.7	87.5	94.8	499.3
SCAN*	72.7	94.8	98.4	58.8	88.4	94.8	507.9
VSRN*	76.2	94.8	98.2	**62.8**	89.7	95.1	516.8
MMCA	74.8	**95.6**	97.7	61.6	89.8	95.2	514.7
IMRAM*	**76.7**	**95.6**	**98.5**	61.7	89.1	95.0	516.6
CAAN	75.5	95.4	**98.5**	61.3	89.7	95.2	515.6
CVSE	74.8	95.1	98.3	59.9	89.4	95.2	512.7
SMFEA(ours)	75.1	95.4	98.3	62.5	**90.1**	**96.2**	**517.6**

Though for a few recall metrics slight variations in performance exist, overall SMFEA shows steady improvements over all baselines. SMFEA achieves 3.6%, 1.9%, and 8.9% improvements in terms of Recall@1 on sentence retrieval, Recall@1 on image retrieval, and rSum, respectively, compared with the state-of-the-art method CAAN (Q. Zhang et al., 2020). Furthermore, compared with some ensemble methods, e.g. VSRN (K. Li et al., 2019), our SMFEA achieves the best performance on most evaluation metrics.

Quantitative Comparison on MS-COCO

Quantitative results on MS-COCO 1K test set are shown at the top of Table 6.3. Specifically, compared with the baseline CVSE (H. Wang et al., 2020), SMFEA achieves 0.3% and 2.6% improvements in terms of Recall@1 on both image and sentence retrieval, respectively. SMFEA also achieves 4.9% improvements in terms of rSum compared with CVSE (H. Wang et al., 2020). Furthermore, on the larger image-sentence retrieval test data (MS-COCO 5K test set), including 5000 images and 25000 sentences, our SMFEA outperforms recent methods with a large gap of Recall@1 as shown in Table 6.4. Following the common protocol (H. Chen et al., 2020; Wei et al., 2020; Q. Zhang et al., 2020), SMFEA achieves 4.2%, 8.8%, and 8.9% improvements

Table 6.4: Comparisons of experimental results on MS-COCO 5K test set.

Method	Sentence Retrieval		Image Retrieval		rSum
	Recall@1	Recall@10	Recall@1	Recall@10	
VSE++	41.3	81.2	30.3	72.4	353.5
SCAN*	50.4	90.0	38.6	80.4	410.9
VSRN*	53.0	89.4	40.5	81.1	415.7
IMRAM*	53.7	**91.0**	39.7	79.8	416.5
MMCA	54.0	90.7	38.7	80.8	416.4
CAAN	52.5	90.9	41.2	82.9	421.1
SMFEA(ours)	**54.2**	89.9	**41.9**	**83.7**	**425.3**

in terms of rSum compared with the state-of-the-art methods CAAN (Q. Zhang et al., 2020), IMRAM (H. Chen et al., 2020) and MMCA (Wei et al., 2020), respectively. Especially on the larger test set, the proposed SMFEA model clearly demonstrates its strong effectiveness with the huge improvements.

Table 6.5: Comparison results on cross-dataset generalization from MS-COCO to Flickr30k.

Method	Sentence Retrieval		Image Retrieval		
	Recall@1	Recall@10	Recall@1	Recall@10	rSum
VSE++	40.5	77.7	28.4	66.6	213.2
LVSE	46.5	82.2	34.9	73.5	237.1
SCAN	49.8	86.0	38.4	74.4	248.6
CVSE	56.4	**89.0**	39.9	77.2	262.5
SMFEA(ours)	**57.1**	88.4	**41.0**	**80.4**	**266.9**

Generalization Ability for Domain Adaptation

In order to further verify the generalization of our proposed SMFEA, we conduct the challenging cross-dataset generalization ability experiments which are meaningful for evaluating the cross-modal retrieval performance in real-scenario. Particularly, similar to CVSE (H. Wang et al., 2020), we transfer our model trained on MS-COCO to Flickr30K dataset. As shown in Table 6.5, our SMFEA achieves significantly outperforms the baseline CVSE (H. Wang et al., 2020), especially in terms of Recall@1 for both modalities retrieval. It reflects that SMFEA is highly effective and robust for image-sentence retrieval with excellent capability of generalization.

6.4.3 Ablation Studies

We perform detailed ablation studies on Flickr30K to investigate the effectiveness of each component of our SMFEA.

Table 6.6: Ablation studies on Flickr30K 1K test set.

Method	Sentence Retrieval			Image Retrieval			
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	rSum
w/o trees	71.7	91.5	94.7	51.3	78.9	87.2	475.3
w/o $\mathscr{D}_{KL}$	72.1	90.9	94.3	53.2	81.1	86.6	478.2
w/o $\mathscr{L}_{CE}$	72.4	91.4	94.9	53.8	81.5	87.0	481.0
SMFEA	**73.7**	**92.5**	**96.1**	**54.7**	**82.1**	**88.4**	**487.5**

Effects of Different Configurations of Context-aware Tree Encoders

Table 6.6 shows the comparing between SMFEA and its corresponding baselines. SMFEA decreases absolutely by 2.0% and 3.4% in terms of Recall@1 for sentence and image retrieval on Flickr30K when removing the multi-modal context-aware structure tree encoders (indicated by w/o trees in Table 6.6). More detailed, comparison shows that removing $\mathscr{D}_{KL}$ or $\mathscr{L}_{CE}$ makes absolute 3.1% and 2.2% drop in terms of Recall@1-Sum (summing Recall@1 for image retrieval and sentence retrieval) on Flickr30K, respectively. It has shown that the context-aware structure tree encoders with joint $\mathscr{D}_{KL}$ or $\mathscr{L}_{CE}$ objectives can slightly improve the effectiveness. Please note that our SMFEA without tree encoders (indicated by w/o trees) is reproduced by using the official codes of CVSE (Q. Zhang et al., 2020) with slightly different parameters, which may result in different performances compared with Q. Zhang et al. (2020). In addition, to better understand how the proposed SMFEA model learns the cross-modal fragments/relations, we visualize the learned relation and fragment categories of nodes in VCS-Tree and TCS-Tree in Figure 6.7. The proposed VCS-Tree and TCS-Tree capture the intrinsic context semantic relation among the fragments in images and sentences in the in-order traversal manner. Also, the explicit consistency of the inter-modal corresponding tree nodes is fully excavated.

Table 6.7: Effects of different encoding structures of SMFEA on Flickr30K 1K test set.

Module	Sentence Retrieval			Image Retrieval			
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	rSum
chain-based	70.7	91.4	95.6	51.2	80.4	86.7	476.0
linear-based	71.2	91.8	95.3	51.7	81.0	87.2	478.2
SMFEA	**73.7**	**92.5**	**96.1**	**54.7**	**82.1**	**88.4**	**487.5**

Effects of Different Embedding Structures of SMFEA

As shown in Table 6.7, SMFEA decreases absolutely 1.92% in terms of the average of all metrics on Flickr30k when replacing context-aware tree structure by a chain-based approach (Hochreiter & Schmidhuber, 1997). In addition, the linear-based tree (F. Chen et al., 2017) degrades the average score by 1.55% compared with our SMFEA. These observations suggest that our context-aware tree encoders can improve the semantic and structural context consistency mining effectiveness between visual and textual features.

Table 6.8: Effects of different configurations of hyperparameters β_* on Flickr30K 1K test set.

$[\beta_d, \beta_t, \beta_c]$	Sentence Retrieval			Image Retrieval			
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	rSum
$[1.0, 0.0, 0.0]$	64.1	88.6	92.6	47.3	75.2	84.1	451.9
$[0.6, 0.0, 0.4]$	66.5	89.9	93.7	49.1	76.9	85.0	461.1
$[0.6, 0.4, 0.0]$	70.9	90.8	93.7	52.1	79.3	86.9	473.7
SMFEA	**73.7**	**92.5**	**96.1**	**54.7**	**82.1**	**88.4**	**487.5**

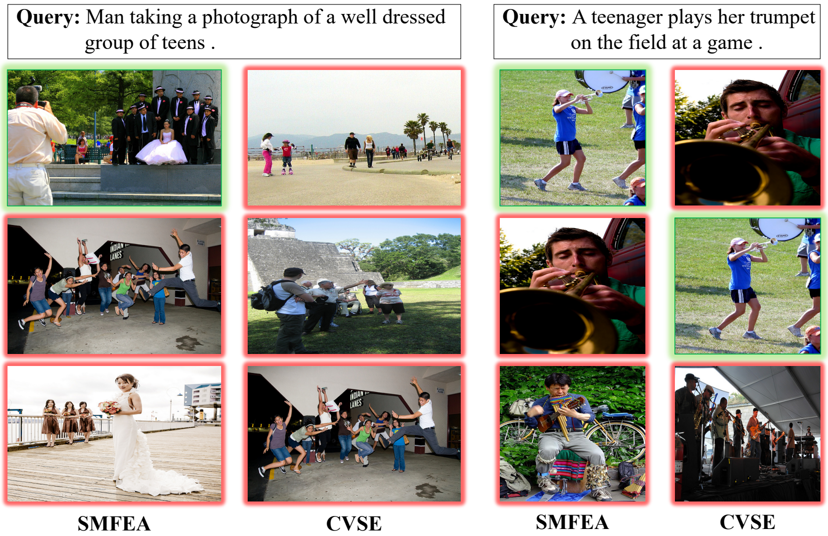
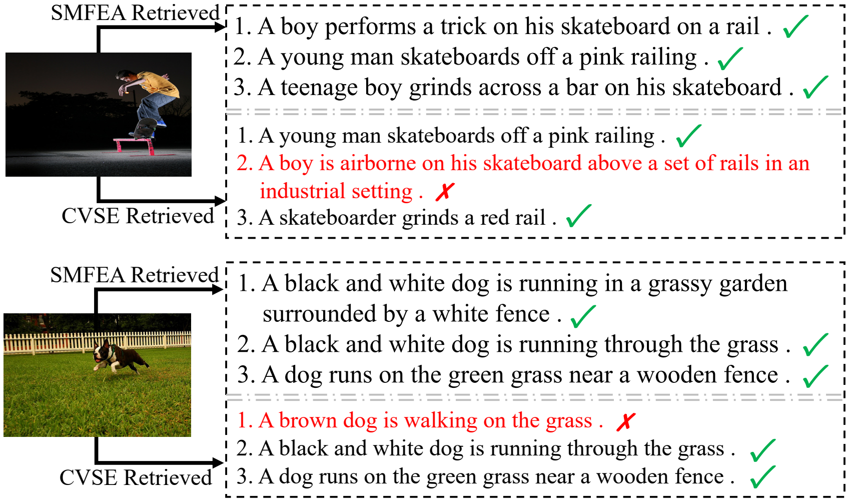

Query: Man taking a photograph of a well dressed group of teens .
Query: A teenager plays her trumpet on the field at a game .
SMFEA
CVSE
SMFEA
CVSE

Figure 6.4: Visual comparisons of image retrieval between our SMFEA and CVSE (H. Wang et al., 2020) on Flickr30K (best viewed in color).

Effects of Different Configurations of Hyperparameters β_*

We evaluate the impact of different multi-modal representations in Eq.(6.15) and Eq. (6.16), including the instance-level features (V^D/S^D), consensus-aware concepts representations (V^C/S^C) and context-aware structured tree embedding and aligning features (V^T $/S^T$), for image-sentence retrieval. As shown in Table 6.8, $[\beta_d, \beta_t, \ \beta_c]$ denotes different balance parameters in Eq.(6.15). For instance, β_d denotes the proportion of SMFEA employing the instance-level multi-modal features. Combining all three representations ($[\beta_d, \beta_c, \ \beta_t] = [0.6, 0.2, 0.2]$) in SMFEA achieves the best performance over all metrics. Moreover, compared with combining the CAC ($[\beta_d, \beta_t, \beta_c] = [0.6, 0.0, 0.4]$), combining the multi-modal context-aware structured tree features with alignment model ($[\beta_d, \beta_t, \beta_c] = [0.6, 0.4, 0.0]$) achieves 12.6% improvement in terms of rSum on Flickr30. It is obvious that our multi-modal context-aware structured tree embedding and alignment model improves the larger performance boost for both modalities retrieval, which validates the importance of learning the intra-modal relations and inter-modal consistency of tree node correspondences.



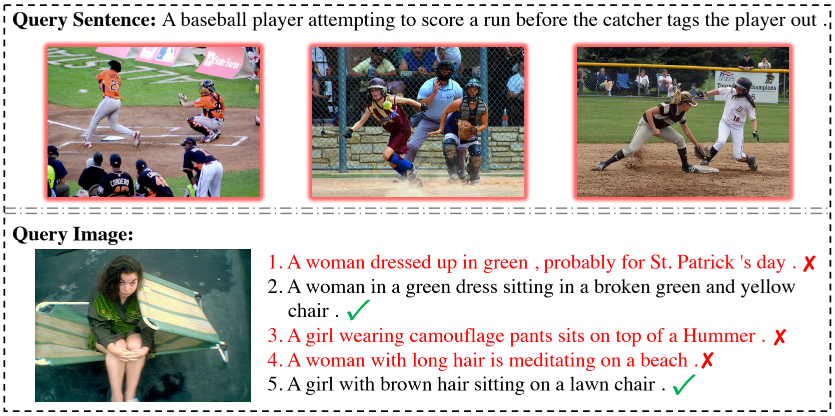
1. A boy performs a trick on his skateboard on a rail . ✓
2. A young man skateboards off a pink railing . ✓
3. A teenage boy grinds across a bar on his skateboard . ✓

1. A young man skateboards off a pink railing . ✓
2. A boy is airborne on his skateboard above a set of rails in an industrial setting . ✗
3. A skateboarder grinds a red rail . ✓

1. A black and white dog is running in a grassy garden surrounded by a white fence . ✓
2. A black and white dog is running through the grass . ✓
3. A dog runs on the green grass near a wooden fence . ✓

1. A brown dog is walking on the grass . ✗
2. A black and white dog is running through the grass . ✓
3. A dog runs on the green grass near a wooden fence . ✓

Figure 6.5: Visual comparisons of sentence retrieval examples between SMFEA and CVSE (H. Wang et al., 2020) on Flickr30K (best viewed in color).


Query Sentence: A baseball player attempting to score a run before the catcher tags the player out .

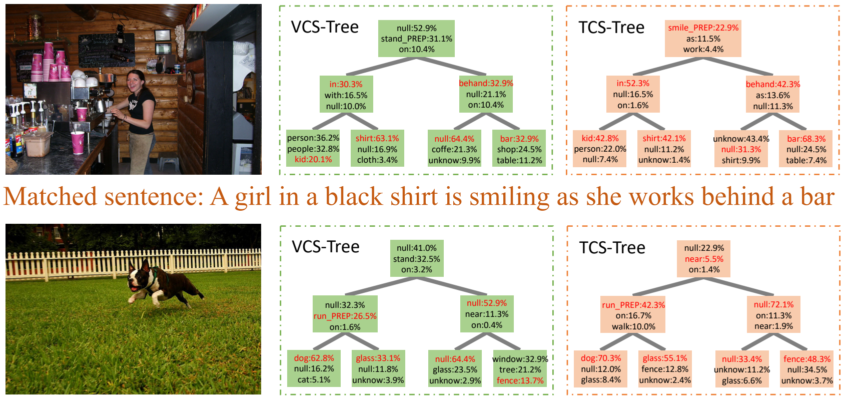
Query Image:

1. A woman dressed up in green , probably for St. Patrick 's day . ✗
2. A woman in a green dress sitting in a broken green and yellow chair . ✓
3. A girl wearing camouflage pants sits on top of a Hummer . ✗
4. A woman with long hair is meditating on a beach .✗
5. A girl with brown hair sitting on a lawn chair . ✓

Figure 6.6: Visualization of the failed image retrieval and sentence retrieval examples on Flickr30K by SMFEA (best viewed in color).

6.4.4 Visualization of Results

To better understand the effectiveness of our proposed model, we visualize matching results of the sentence retrieval and image retrieval on Flickr30K in Figure 6.4 and Figure 6.5. For image retrieval shown in Figure 6.4, we show the top 3 ranked images for each text query matched by our proposed SMFEA in the first column, and followed by CVSE (H. Wang et al., 2020) in the second column. The true matches are outlined in green boxes and the false matches are in red. Furthermore, as shown in Figure 6.5, we visualize the sentence retrieval results (top-3 retrieved sentences) predicted by SMFEA and CVSE (H. Wang et al., 2020), where the mismatches are highlighted in red. Examples of failed image retrieval and sentence retrieval are shown in Figure 6.6. However, in this case, the wrong images/sentences have similar semantic or structural content to true matches. We argue that the reason for this phenomenon may be that our current tree structure model is to unify the coarse-grained semantics and structural consistency between the two modalities. It has a good ability to improve the robustness of the model. But there are certain shortcomings in the distinction of similar sets. We will build fine-grained vocabularies



Matched sentence: A girl in a black shirt is smiling as she works behind a bar .

Matched sentence: A dog runs on the green grass near a wooden fence .

Figure 6.7: Visualization of learned VCS-Tree and TCS-Tree in our SMFEA on Flickr30K. The red font means the correct semantic items according to the referral tree (best viewed in color).

to improve future work.

6.5 Conclusion

In this paper, we exploit image-sentence retrieval with structured multi-modal feature embedding and cross-modal alignment. Our work serves as the first to narrow the cross-modal heterogeneous gap by aligning the explicitly inter-modal semantic and structure correspondence between images and sentences with the visual/textual inner context-aware structured tree encoder (VCS-Tree/TCS-Tree) capturing. We proposed a novel structured multi-modal feature embedding and alignment (SMFEA) model, which contains a VCS-Tree and a TCS-Tree to enhance the intrinsic context-aware structured semantic information for image and sentence, respectively. Furthermore, the consistency estimation of the corresponding inter-modal tree nodes is maximized to narrow the cross-modal pair-wise distance. Extensive quantitative comparisons demonstrate that our SMFEA can achieve state-of-the-art performance across popular standard benchmarks, MS-COCO and Flickr30K, under various evaluation metrics.

6.6 Limitation

Despite the advancements achieved by the SMFEA model in image-sentence retrieval, there are notable limitations that warrant discussion. (1) Loss of Detailed Semantics: While the explicit semantic and structural constraints implemented in SMFEA do enhance the accuracy of visual representations to some extent, the reliance on a fixed tree structure can lead to significant semantic loss. This rigidity may hinder the model's ability to capture the full diversity of semantic relationships and contextual nuances present in the data, resulting in a narrowed representation that fails to encapsulate the richness of visual content. (2) Lack of Interaction Between Modalities: Another limitation is the insufficient interaction between the visual and textual modalities. The design of SMFEA primarily focuses on aligning features based on their respective structures, but this lack of direct inter-modal interaction may constrain the overall alignment performance. Without dynamic exchanges of information between images and sentences, the model may struggle to fully leverage the complementary strengths of each modality, thereby reducing the effectiveness of cross-modal understanding. Future research should focus on addressing these limitations to develop more robust models that better capture the complexities of multimodal data and enhance performance in image-sentence retrieval task.

Chapter 7

Hybrid-modal Interaction with Multiple Relational Enhancements

The key issue lies in jointly learning the visual and textual representation to estimate their similarity accurately. Most existing methods focus on feature enhancement within modality or feature interaction across modalities, which, however, neglects the contextual information of the object representation based on the inter-object relationships that match the corresponding sentences with rich contextual semantics. Different from the Chapter 6, we want to explore the ability of flexible intra-modal modelling approaches and inter-modal interactive reasoning methods for modality feature representation enhancements. In this chapter, we propose a Hybrid-modal Interaction with multiple Relational Enhancements (termed *Hire*) for image-sentence retrieval, which correlates the intra- and inter-modal semantics between objects and words with implicit and explicit relationship modelling. In particular, the explicit intra-modal spatial-semantic graph-based reasoning network is designed to improve the contextual representation of visual objects with salient spatial and semantic relational connectivities, guided by the explicit relationships of the objects' spatial positions and their scene graph. We use implicit relationship modelling for potential relationship interactions before explicit modelling to improve the fault tolerance of explicit relationship detection. Then the visual and textual semantic representations are refined jointly via inter-modal interactive attention and cross-modal alignment. To correlate the context of objects with the textual context, we further refine the visual semantic representation via cross-level object-sentence and word-image-based interactive attention. Extensive experiments validate that the proposed hybrid-modal interaction with implicit and explicit modelling is more beneficial for image-sentence retrieval. And the proposed *Hire* obtains new state-of-the-art results on MS-COCO and Flickr30K benchmarks.

7.1 Introduction

To accurately measure the semantic similarity of two modalities and establish the association between two modalities, numerous methods (Faghri et al., 2017; Frome et al., 2013; Ge, Chen, et al., 2021; Huang et al., 2018; K.-H. Lee et al., 2018; C. Liu et al., 2019; S. Long et al., 2022; Wen et al., 2020) have been proposed to bridge the semantic gap between visual and textual representations. Typically, earlier approaches (Faghri et al., 2017; Frome et al., 2013; L. Wang et al., 2016) estimated the image-texts similarities based on the projected global visual and textual representations, which are directly extracted from the whole image and the full sentence via Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) respectively. However, these rough representations are difficult to accurately identify and fully utilize high-level semantic concepts, especially those of images.

Recently, many methods (Cheng et al., 2022; Ge, Chen, et al., 2021; S. Long et al., 2022; Qu et al., 2021; Wen et al., 2020) further take advantage of fine-grained region-level visual features from object detectors (Ren et al., 2015) with salient semantic content to enhance the high-level semantic representation of images, and align them with the word-level features of sentences. These methods can be divided into two main kinds, intra-modal feature interactions (Cheng et al., 2022; Ge, Chen, et al., 2021; L. Wang et al., 2016; Wen et al., 2020) and inter-modal feature interactions (Huang et al., 2018; K.-H. Lee et al., 2018; C. Liu et al., 2019; Qu et al., 2021), to obtain a better multi-modal joint embedding space. Intra-modal representation learning has been widely studied in many multi-modal tasks, such as image captioning (F. Chen et al., 2019), video caption retrieval (X. Yang, Feng, Ji, Wang, & Chua, 2021), and so on. Similarly, for image-text matching, intra-modal representation learning is also important to improve the visual or textual semantic representation via the implicit and explicit semantic relationships reasoning methods within each modality, such as the graph convolution networks (GCNs) (C. Liu et al., 2020; S. Long et al., 2022; Wen et al., 2020), self-attention mechanism (SA) (Qu et al., 2020; Y. Wu et al., 2019) and tree encoder (Ge, Chen, et al., 2021; X. Yang et al., 2020), etc. For instance, (Y. Wu et al., 2019) proposed intra-modal self-attention embeddings to enhance the representations of images or texts by self-attention mechanism, which can exploit subtle and fine-grained fragment relations in image and text, respectively. K. Li et al. (2019) proposed an implicit relationship reasoning modal based on Graph Convolutional Networks to build up connections between image regions and then generate the global visual features with semantic relationships. Ge, Chen, et al. (2021) developed a structured tree encoder within each modality to enhance the semantic and structural consistency representation of matched images and texts for cross-modal matching. Intra-modal independent representation learning can adequately model relationships between entities within each modality via implicit or explicit reasoning approaches, which, however, fails to capture the fine-grained semantic correspondence interactions among the two modalities.

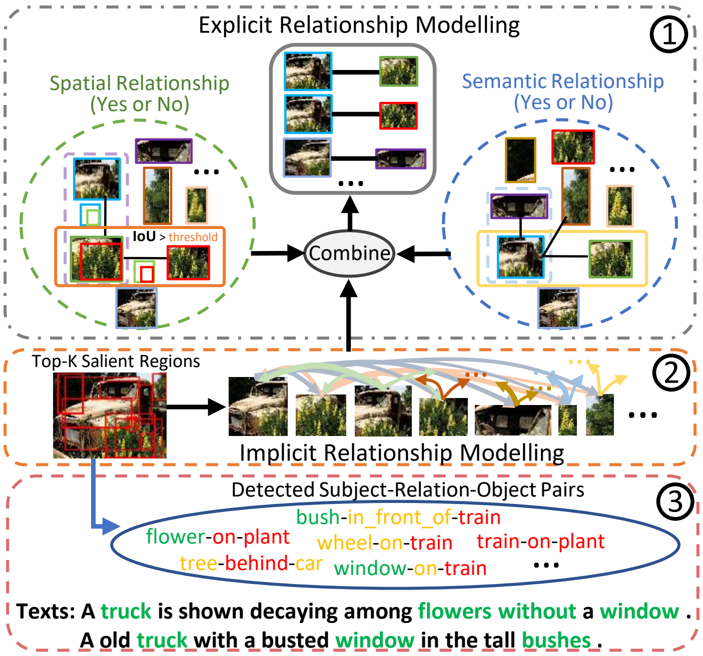

Explicit Relationship Modelling
①
Spatial Relationship
(Yes or No)
IoU > threshold
Semantic Relationship
(Yes or No)
Combine
Top-K Salient Regions
②
Implicit Relationship Modelling
Detected Subject-Relation-Object Pairs
③
bush-in_front_of-train
flower-on-plant
wheel-on-train
train-on-plant
tree-behind-car
window-on-train
Texts: A truck is shown decaying among flowers without a window .
A old truck with a busted window in the tall bushes .

Figure 7.1: Illustration of the explicit and implicit intra-modal modelling schemas for the semantic relationship. ① the explicit spatial-semantic relationship modelling schema: objects along with their spatial and semantic relationships are jointly modelled based on the relative position and the detected scene-graphs. However, the subject-relation-object pairs (③) in detected scene graphs of each image usually have some errors or do not match the text. For example, in *window-on-train*, the word labels of relation "on" and object "train" are hard to accurately represent the corresponding semantic content, or even wrong (in red). To this end, the relational connectivity (relationship exists or not) rather than the object/attribute label is encoded into the object features. In addition, some relation pairs are even missing due to the limitation on the label range of the offline detector, *e.g. truck-with-window*. Fortunately, it can be relieved by the implicit relationship modelling (②) due to its construction of the general relationship among object regions. ② the implicit relationship modelling schema: object relationships are constructed by fully connecting the object regions, where the information can be propagated and aggregated among objects according to their potential relationships. However, it is hard to maintain strong inter-object relationships in a multi-layer network. To deal with the above issues, it's intuitive to combine both implicit and explicit relationship modelling to cooperate visual semantic representation with the inter-object relationship.

To address the above problem, many studies (Huang et al., 2018; K.-H. Lee et al., 2018; C. Liu et al., 2019; Z. Wang et al., 2019) based on inter-modal interaction operations are proposed to further narrow the semantic gaps between multiple modalities, which improve the retrieval performance by learning the accurate fine-grained visual-textual semantic correspondences between the fragments of image and text. For instance, SCAN (K.-H. Lee et al., 2018) attended object regions to each word to generate the text-aware visual features for text-to-image matching and, conversely, for image-to-text matching. IMRAM (H. Chen et al., 2020) further proposed an iterative matching scheme with a cross-modal attention unit and a memory distillation unit to explore such fine-grained correspondence and refine knowledge alignments progressively.

Moreover, recent methods (C. Liu et al., 2020; Qu et al., 2020; Wei et al., 2020; Q. Zhang et al., 2020) combined intra- and inter-modal interactions to jointly improve semantic relation representation within each modality and accurate visual-textual semantic correspondence between the two modalities, further boosting retrieval performance. For instance, MMCA (Wei et al., 2020) integrated intra-modal and inter-modal interactions in a parallel pattern, in which both interactions employ implicit transformer-based self-attention mechanism (Vaswani et al., 2017), but inter-modal interaction concatenates cross-modal region-word features for attention calculation. DIME (Qu et al., 2020) introduced a multi-layer modality interaction framework with different intra- and inter-modal interaction cells, stacked in width and depth. However, the hand-crafted multi-interaction combining methods (C. Liu et al., 2020; Wei et al., 2020; Q. Zhang et al., 2020) lack exploration on the impact of different combinations of intra- and inter-modal interactions on matching performance and DIME (Qu et al., 2021), relying on soft links and multiple interaction cell stacking, increases model complexity. Additionally, these methods, despite notable improvements, overlook the limited representation of inter-object relationships compared to the strong textual context, resulting in a weakened role of visual semantics in image-text matching. The basic intuition of our work lies in two aspects to deal with the above problems. On the one hand, the intra-modal feature interactions, whether implicit or explicit, are crucial to enhance the visual/textual representation with the semantic relationships among fragments, especially among the visual region features that lack contextual representation. However, either implicit or explicit intra-modal interactions have their own defects. Notably, providing the fully-connected information flows among objects, through the implicit intra-modal interaction (Cheng et al., 2022; Q. Zhang et al., 2020), usually leaves the relationship information weak and ambiguous due to the redundant information, which affects the object discrimination as shown in Figure 7.1 (2). Additionally, the effect of implicit intra-modal interaction on the structured correlation among the objects and their relationships will be weakened when the object features pass the multi-layer network without further supervision. Explicit intra-modal interaction heavily relies on the off-the-shelf detector (Anderson et al., 2018; Yao, Pan, Li, Qiu, & Mei, 2017) to concatenate the object region features with the features of the detected inter-object relationships via the graph-based modelling, which, however, introduces additional recognition error from object and attribute labels. Moreover, it also neglects the spatially relative positions. For instance, in S. Wang et al. (2020), objects and their corresponding relationships are detected guided by the scene graph, and their label-based embeddings are aggregated with the object region features to feed the Graph Convolution Networks (GCNs). However, due to the heterogeneous training data, the detected object and relation labels (e.g. '*train-on-plant*') are usually inconsistent with the expressions of the corresponding sentences as shown in Figure 7.1 (3). To address the above issues, it's natural for us to integrate both the implicit and the explicit intra-modal interactions to enhance the object representation, which tackles the limitations of the structured information in implicit interactions and provides flexibility in explicit interactions. To enhance object discrimination, we consider an integrated structured model to capture the explicit information of the inter-object relationships, including the semantic and spatial considerations. As manifested in Figure 7.1 (1), by explicitly constructing the inter-object relationships, the semantic relationship modelling provides a strong semantic correlation between objects while the spatial relationship modelling reduces the feature redundancy of spatially overlapping. Notably, we do not use the additional detected labels to mitigate the error interference from the detection and facilitate the end-to-end representation learning.

On the other hand, the effects of different combinations of the intra- and inter-modal interactions on matching results are different, which, however, are not sufficiently discussed in the existing literature (Cheng et al., 2022; C. Liu et al., 2020; Wei et al., 2020; Q. Zhang et al., 2020). Most of the existing hand-crafted methods combining intra-modal and inter-modal interactions directly use simple serial-pattern (C. Liu et al., 2020), or parallel-pattern (Q. Zhang et al., 2020) combinations, which lack the discussion and exploration of different combinations. Although DIME (Qu et al., 2021) proposed a dynamic route exploration approach in multiple layers with multi-interaction, it relies on a huge serial and parallel network, which contains three layers and each layer contains four interactions. In this work, we will explore, in detail, the impact of different combinations on retrieval performance, including multiple intra- and inter-modal interactions among images and sentences with explicit and implicit modelling, and discuss the potential reasons.

7.1.1 Motivation

Driven by the above considerations, we present a novel hybrid-modal interaction method for image-text matching via multiple relational reasoning modules within and across modalities (termed ***Hire***), which better correlates the intra- and inter-modal semantics between objects and words. For the intra-modal semantic correlation, the inter-object relationships are explicitly reflected on the spatially relative positions, and the scene graph guided potential semantic relationships among the object regions. We then propose a relationship-aware GCNs model (termed *R-GCNs*) to enhance the object region representations with their relationships, where the graph nodes are object region features and the graph structures are determined by the inter-object relationships, i.e. each edge connection in the graph adjacency matrices relies on whether there is a relationship with high confidence. In addition, to mitigate the impact of relation omission by the off-the-shelf detector and adequately keep structured correlations among the objects and their relationships in a multi-layer network, we perform implicit relational reasoning between objects before explicitly modelling them. Experiments also prove that this information supplement effectively improves the effect of retrieval. For the inter-modal semantic correlation, the implicit and explicit semantic enhanced representations of object regions, as well as the enhanced semantic representations of words that undergo a fully-connected self-attention model, are attended alternatively in the inter-modal interactive attention, where the object region features are attended to each word to refine its feature and conversely, the word feature are attended to each object region to refine its feature. To correlate the context of objects with textual context, we further refine the representations of object regions and words via cross-level object-sentence and word-image-based interactive attention. The intra-modal semantic correlation, inter-modal semantic correlation, and similarity-based cross-modal alignment are jointly executed to enhance the cross-modal semantic interaction further.

7.1.2 Contribution

The contributions of this chapter are as follows:

- We propose an intuitive intra-model interaction model that combines implicit and explicit relationship modelling to guarantee a structured correlation among the objects and their relationships with continuous correlation guidance in a multi-layer network, overcoming the relationship omissions and erroneous via the self-attention mechanism.

- We explore an explicit intra-modal semantic enhanced correlation to utilize the inter-object spatially relative positions and inter-object semantic relationships guided by a scene graph, and propose a relationship-aware GCNs model (R-GCNs) to enhance the object region features with their relationships. This module mitigates the error interference from the detection and enables end-to-end representation learning.

- We conduct exhaustive experiments on a variety of cross-modal interaction methods. Then we propose a comprehensive method (*Hire*) to unite the intra-modal semantic correlation, inter-modal semantic correlation, and the similarity-based cross-modal alignment to simultaneously model the semantic correlations on three grain levels, *i.e.* intra-fragment, inter-fragment, inter-instance. Especially, cross-level interactive attention is proposed to model the correlations between fragments and instances.

The proposed *Hire* is sufficiently evaluated with extensive experiments on MS-COCO and Flickr30K benchmarks and achieves a new state-of-the-art for image-text matching.

7.2 Related Work of Hybrid-modal Interactive Enhanced Retrieval

Recently, some studies (Fu et al., 2024; C. Liu et al., 2020; Qu et al., 2021; Wei et al., 2020; Q. Zhang et al., 2020) try to combine the intra- and cross-modal interactions to further improve the fine-grained inter-modal object-word correspondence with intra-modal interaction enhancement. For instance, MMCA (Wei et al., 2020) proposed a hybrid-modal relational interaction method to exploit the fine-grained relationships among the fragments via a parallel pattern of self-attention and cross-attention approaches. However, the above hybrid-modal interaction methods employed implicit relationship modelling within a modality, which makes it hard to keep a structured correlation among the objects and their relationships in a multi-layer network without continuous correlation guidance. The most relevant existing work to ours is DIME (Qu et al., 2021), which dynamically learns interaction patterns through soft-path decisions in a 4-layer network, where each layer contains two intra-modal and two inter-modal interaction strategies, respectively. However, DIME relies on a large and complex network, which contains 12 units in 4 types, to assign weights to the output features of different interaction units. This makes its path selection challenging to interpret. And it still suffers from the aforementioned issue of hard maintaining strong inter-object relationships in the multi-layer network.

In contrast to previous studies, *e.g.*, MMCA (Wei et al., 2020), DIME (Qu et al., 2021), etc., our *Hire* approaches the inter-object modelling in a novel way by exploiting the spatial and semantic graph to enhance the structured relationship embedding based on implicit reasoning. The joint embedding space is obtained by aligning the fine-grained inter-modal semantic fragments further to reduce the heterogeneous (inter-modality) semantic gap. Doing so allows us to provide more robustness than DIME (Qu et al., 2021), which also improves the interpretability of the model.

7.3 Problem Formulation

Image-sentence retrieval aims at matching the most relevant images in the image database (or texts in the sentence database) given a text query (or image query). Here, assume we have an image database $\mathscr{I} = \{I_1, I_2, \ldots, I_N\}$ and a text database $\mathscr{S} = \{S_1, S_2, \ldots, S_M\}$, which contain N images and M sentences, respectively. This chapter aims to facilitate efficient image-text matching via fine-grained intra-modal relationship utilization and cross-modal semantic correspondence.

To this end, we first take advantage of the bottom-up-attention model (Anderson et al., 2018) to extract top-K fine-grained sub-region features $\hat{V} = [\hat{v}_1, \ldots, \hat{v}_K]$, $\hat{v}_i \in \mathbb{R}^{2048}$, for each image I, based on the category confidence score in an image, which can better represent the salient objects and attributes. Afterwards, a fully connected (FC) layer with the parameter $W^o \in \mathbb{R}^{2048 \times D_v}$ is used to project these feature vectors into a D_v-dimensional space. Finally, these projected object region features $V = [v_1, \cdots, v_K]$, $v_i \in \mathbb{R}^{D_v}$, are taken as initial visual representations without semantic relationship enhancement. For sentence texts, we follow the recent trends in the community of Natural Language Processing and utilize the pre-trained BERT (Devlin et al., 2018) model to extract word-level textual representations. Similar to visual features processing, we also utilize FC layers to project the extracted word features into a D_t-dimensional space for sentence S, denoted as $T = [t_1, t_2, \cdots, t_m]$, $t_j \in \mathbb{R}^{D_t}$, with length m. To facilitate cross-modal interaction and embedding space consistency, we project the visual and textual representations into the same dimension (D_v=D_t). For subsequent local-global inter-modal interaction and final

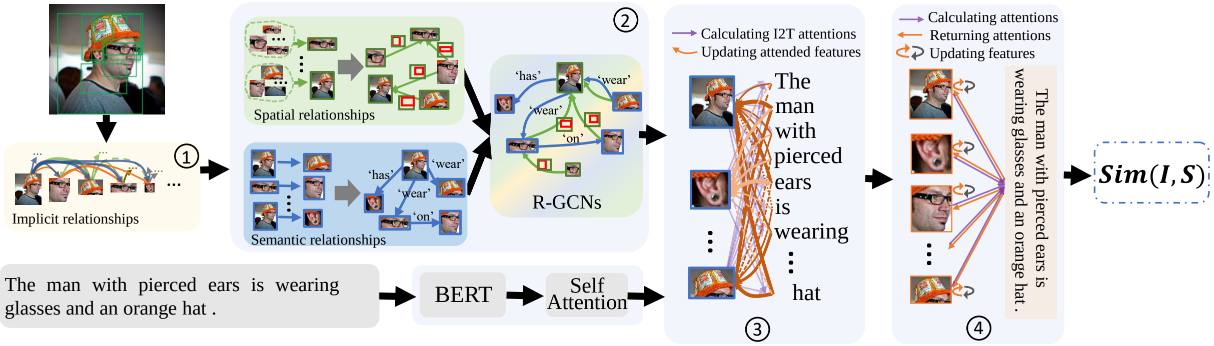


Figure 7.2: The overall framework (image-to-text version) of *Hire*. In intra-modal semantic correlation (① and ②), an implicit relationship reasoning is first used to obtain the potential semantic connections among all candidate regions, similarly for high-level textual word embeddings from pre-trained BERT. And then, a relationship-aware GCNs (R-GCNs) is constructed to integrate the explicit spatial and semantic relationships between every two objects into their region representations by changing the relationship-determined graph adjacency matrix. In inter-modal semantic correlation (③ and ④), the visual and textual semantic features are further enhanced via object-word interactive attention and the visual semantic representation is refined via the cross-level object-sentence and word-image-based interactive attention. Visual and textual semantic similarity is finally estimated for the cross-modal alignment.

cross-modal similarity calculation, we use the average-pooling operation to obtain the global image feature $\bar{V}$ for text-to-image and the global text feature $\bar{T}$ for image-to-text.

Next, we leverage multiple intra-modal interactions to enhance the semantic representation within modalities and inter-modal interactions to narrow the semantic gap between heterogeneous visual-textual modalities. Notably, we sufficiently explore the impact of different combinations of interactions and ultimately construct our proposed *Hire*, which unite the intra-modal semantic correlation, inter-modal semantic correlation and the similarity-based cross-modal alignment together to model the semantic correlations on three levels, i.e. intra-fragment (especially for inter-object within visual modality), inter-fragment between two modalities, and inter-instance from one modality to another modality. Firstly, the visual representation V and textual representation T are independently enhanced by an implicit relationship interaction based on a self-attention mechanism within each modality, and an explicit spatial-semantic relationship interaction based on relationship-aware GCNs is further used to improve the visual context information among the detected salient objects in images. Then, a local-local inter-modal interaction is leveraged to improve the micro consistency of the embedding space of multi-modal features via fine-grained inter-modal fragment (object-word/word-object) correlations, and a local-global inter-modal interaction is used to keep the macro consistency via similarity-based inter-instance (image-word/sentence-object) alignment. Finally, the visual and textual semantic similarity is measured for the cross-modal alignment.

7.4 The Proposed Method – *Hire*

Figure 7.2 shows the overall pipeline of our proposed *Hire*, which includes two intra-modal interactions and two inter-modal interaction modules for image-text matching. For a clear presentation, we mainly describe image-to-text direction, and the text-to-image version is in a similar pattern. We will first describe the intra-modal interactions for the relationship reasoning within each modality in Section 7.4.1. Afterwards, two inter-modal interaction methods are described in Section 7.4.2 on calculating micro and macro fragment correlations from another modality. Finally, the objective function is discussed in Section 7.4.3.

7.4.1 Intra-modal Relationship Interactions

Due to little inter-object relationships reflected in object representations compared to the strong context of the textual structure, we combine implicit and explicit relationship modelling approaches to improve the visual semantic representing ability. The main motivation is that explicit relational graph reasoning based on the detected scene graphs maintains the inter-object relationship structure well, but suffers from relationship omission. To this end, we employ implicit inter-object relationship modelling to improve the robustness of visual representation.

Implicit relationship modelling. To refine the object-level latent embeddings of sub-region features for each image, we employ the self-attention mechanism (Vaswani et al., 2017) to concentrate on the salient information with potential correlations. In particular, following Vaswani et al. (2017), the projected object visual features $V = [v_1, \cdots, v_K]$ are used as the key and value items, and each target object v_i serves as the query item. Each attention weight for each query object is calculated as follows:

$$\alpha_{ij} = Att(W^q v_i, W^k v_j) = W^q v_i (W^k v_j)^T / \sqrt{D}, \tag{7.1}$$

$$A_{ij} = softmax(\alpha_{ij}) = \frac{exp(\alpha_{ij})}{\sum_{j=1}^{K} exp(\alpha_{ij})}, \tag{7.2}$$

where W_q, W_k are the parameters of mapping from D_v to D, and $\sqrt{D}$ acts as a normalization factor. Following Vaswani et al. (2017), we also employ multi-head dot product by L parallel attention layers to speed up the calculation efficiency, and a feed-forward network (FFN) based on two FC layers (with ReLU activation function) is followed to obtain the final reasoning representation v_i^A for the *i-th* target object. The overall working flow is formulated as:

$$v_i^A = FFN(W^h ||_{l=1}^{L} (head_1, \ldots, head_L)), \tag{7.3}$$

$$head_l = \sum_{j=1}^{K} (A_{ij}^l W^{vl} v_j), \tag{7.4}$$

$$A_{ij}^l = softmax(Att(W^{ql} v_i, W^{kl} v_j)) = softmax(W^{ql} v_i (W^{kl} v_j)^T / \sqrt{D/L}), \tag{7.5}$$

where W^h is the mapping parameter, W^{ql}, W^{kl}, W^{vl} map the feature dimension to $1/L$ of the original, $||$ means concatenation. Finally, the implicit relationship enhanced visual representation $V^A = [v_1^A, \ldots, v_K^A]$ is obtained. Similar to the above procedure, we also get the concentrated textual representation $T^A = [t_1^A, \ldots, t_m^A]$ for the sentence.

Explicit visual relationship modelling. To further improve the maintenance of contextual relationships among the salient objects in images, we construct a spatial-semantic graph for each image and enhance the object region features with their relationships via a relationship-aware GCNs model. On the one hand, different from existing approaches (Cheng et al., 2022; K. Li et al., 2019) based on implicit relationship graph reasoning, scene graphs have well-defined object relationships, which can overcome the disadvantage of fusing redundant information. And unlike approaches (S. Long et al., 2022; S. Wang et al., 2020) based on scene-graph enhancement, we do not encode the word labels predicted by the pre-trained visual scene-graph generator, like Zellers, Yatskar, Thomson, and Choi (2018). We consider that word labels from visual scene graphs of external models have errors and are semantically different from the words in the corresponding texts. This tends to introduce noise that corrupts the cross-modal semantic alignment. On the other hand, since features from the top-K candidate object regions are used for representing the image information, this leads to some regions with semantic overlap but with minor positional bias. Study (Cheng et al., 2022) also indicated that the regions with larger Intersection over Union (IoU) as potentially more closely.

Different from Cheng et al. (2022); Ge, Chen, et al. (2023), combining spatial and semantic relationships in one graph further increases the diversity of semantic correlations, e.g. different high IoU regions with similar content can connect with some related objects which usually miss connections in the original scene graph due to confidence settings. In particular, we construct an explicit spatial-semantic non-fully connected graph $\mathscr{G} = (V^A, E)$ for each image. The spatial IoUs and semantic correlations between sub-regions are combined to construct the adjacency matrix $E \in \mathbb{R}^{K \times K}$ as edges for the graph. Of which, if the IoU_{ij} of the i-th region and the j-th region exceeds the threshold μ, it indicates that there is a relationship edge between the two object regions. Otherwise, it is 0. Likewise, if p-th object is associated with j-th object in the semantic relations extracted by a pre-trained visual scene-graph generator, there is a relationship edge between the two object regions and 0 otherwise. In this way, if the j-th object region has a high IoU score with i-th object region and semantic relationship with p-th object, then all three objects have associated edges with improving the robustness of relationship modelling. The values of edges are learning and updating based on the semantic similarities between the correlated objects, where the pairwise semantic similarity of i-th and j-th objects is calculated as:

$$E_{ij} = (W^{\varphi} v_i^A)^T (W^{\phi} v_j^A), \tag{7.6}$$

where W^{φ} and W^{ϕ} denote the mapping parameters. For simplicity, we do not explicitly represent the bias term in our chapter.

For the final object region features V^G, the currently popular Graph Convolutional Networks (GCNS) (K. Li et al., 2019) with residuals are used, which can enhance the object representations by updating and embedding of spatial and semantic relationship graphs, named relationship-aware GCNs (R-GCNs), as shown in Figure 8.2. Formally,

$$V^G = (EV^A W^g)W^r + V^A, \tag{7.7}$$

where $W^g \in \mathbb{R}^{D_v \times D_v}$ is the weight matrix of the GCN layer, W^r is the residual weights.

7.4.2 Inter-modal Semantic Relationship Interactions

After image objects and text words are reinforced with semantic relationships within each modality, we apply two mainstream inter-modal interaction mechanisms to further enhance the feature representation of the target modality with attention-ware information from another modality. For a clearer presentation, we describe the process as an example of image-to-text.

Local-local inter-modal interaction. Similar to literature (K.-H. Lee et al., 2018; Qu et al., 2021), we mine attention between image objects and text words to narrow the semantic gap between the two modalities. As shown in Figure 8.2 ③, taking the image-to-text example (Due to a clearer presentation), we first calculate the cosine similarities for all object-word pairs and calculate the attention weights by a per-dimension λ-smoothed Softmax function (Chorowski et al., 2015), as follows:

$$c_{ij} = \frac{(v_i^G)^T t_j^A}{||v_i^G||\,||t_j^A||}, i \in [1,K], j \in [1,m], \tag{7.8}$$

$$\beta_{ij} = \frac{exp(\lambda c_{ij})}{\sum_{j=1}^{N} exp(\lambda c_{ij})}, \tag{7.9}$$

Finally, we obtain the attended object representation $v_i^F \in V^F$ via a conditional fusion strategy (Qu et al., 2021) from correspondence attention-aware textual vector q_i^t (q_i^t=$\sum_{j=1}^{m} \beta_{ij} t^A$), as follows,

$$v_i^F = \text{ReLU}(W_1^f(v_i^A \odot \text{Tanh}(W_2^f q_i^t) + W_3^f q_i^t)) + v_i^A, \tag{7.10}$$

where W_*^f are the mapping parameters, ReLU and Tanh are activation functions. To fully explore fine-grained cross-modal interactions, we perform the above process twice. Similar, we can also obtain the word-object interaction enhancement textual features T^F for the text-to-image version.

Local-global inter-modal interaction. As shown in Figure 7.2 ④, we further discover the salience of the fragments in one modality guided by the global contextual information of the other modality, which makes each fragment contain more contextual features. Specifically, for image-to-text, we first calculate the semantic similarity between the objects of image $V^F = \{v_1^F, \cdots, v_K^F\}$ and global textual feature $\bar{T}$. Then, we can obtain the relative importance of each object via a sigmoid function. Finally, we add residual connections between the attention-aware object features and the enhanced object features V^F, as well as the original features V. The above process can be formulated as:

$$r_i = \sigma(W^r v_i^F \odot \bar{T}), \tag{7.11}$$

$$v_i^O = r_i v_i^F + v_i^F + \mathrm{ReLU}(v_i), \tag{7.12}$$

where W^r denotes the mapping parameter. Similarly, for text-to-image, we enhance the word features by calculating the relative importance of each word between the words of the sentence and the global image feature $\bar{V}$.

To obtain the final match score between the image and sentence, we average and normalize the final object features of the image and calculate the cosine similarity with the global text features.

7.4.3 Objective Function

In the above training process, we use a bidirectional triplet ranking loss (Faghri et al., 2017) to lead the distances between correlated image-text pairs closer than distances for uncorrelated pairs after the hybrid-modal interactions when aligning the image and sentence as follows:

$$\mathscr{L}_{rank}(I,S) = \sum_{(I,\hat{S})} [\nabla - \cos(I,S) + \cos(I,\hat{S})]_+ + \sum_{(\hat{I},S)} [\nabla - \cos(I,S) + \cos(\hat{I},S)]_+ \tag{7.13}$$

where ∇ serves as a margin constraint, $\cos(\cdot,\cdot)$ indicates cosine similarity function, and $[\cdot]_+ = \max(0,\cdot)$. Note that (I,S) denotes the given matched image-text pair, and its corresponding negative samples are denoted as $\hat{I}$ and $\hat{S}$, respectively. For image-to-text direction, $cos(I,S) = cos(V^O,\bar{T})$, and $cos(I,S) = cos(\bar{V},T^O)$ is for text-to-image direction. In addition, to preserve the semantic relevance of heterogeneous modalities in a cascaded approach consisting of multiple modules, we optimize an additional triplet ranking loss $\mathscr{L}_{add}$ for the enhanced visual and textual embeddings after the intra-modal interactions. Finally, all parameters can be simultaneously optimized by minimizing the joint bidirectional triplet ranking loss $\mathscr{L} = \mathscr{L}_{rank} + \mathscr{L}_{add}$.

7.5 Experimental Setup

7.5.1 Implementation Details

Our model is trained on a single TITAN RTX GPU with 24 GB memory. The whole network except the Faster-RCNN model (Ren et al., 2015) is trained from scratch with the default initializer of PyTorch. The ADAM optimizer (Kingma & Ba, 2014) is used with a mini-batch size of 80. Similar to Qu et al. (2021), during the training process, we also add some negative samples

Table 7.1: Comparisons of experimental results on MS-COCO 5-folds 1K test set. * indicates the performance of an ensemble model. † denotes the significant improvements on Recall@1 (paired t-test, $p < 0.01$) compared with the best baseline (*i.e.* AME*). Red numbers denote the improvements compared with state-of-the-arts.

Method	Image-to-Text			Text-to-Image			rSum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
IMRAM$^*_{CVPR'20}$	76.7	95.6	98.5	61.7	89.1	95.0	516.6
CAAN$_{CVPR'20}$	75.5	95.4	98.5	61.3	89.7	95.2	515.6
GSMN$^*_{CVPR'20}$	78.4	96.4	98.6	63.3	90.1	95.7	522.5
SMFEA$_{ACMMM'21}$	75.1	95.4	98.3	62.5	90.1	96.2	517.6
SGRAF$^*_{AAAI'21}$	79.6	96.2	98.5	63.2	90.7	96.1	524.3
VSE$\infty_{CVPR'21}$	79.7	96.4	98.9	64.8	91.4	96.3	527.5
DIME$^*_{SIGIR'21}$	78.8	96.3	98.7	64.8	91.5	96.5	526.6
VSRN++$^*_{TPAMI'22}$	77.9	96.0	98.5	64.1	91.0	96.1	523.6
GraDual$^*_{WACV'22}$	77.0	96.4	98.6	65.3	91.9	96.4	525.6
NAAF$^*_{CVPR'22}$	80.5	96.5	98.8	64.1	90.7	96.5	527.2
AME$^*_{AAAI'22}$	79.4	**96.7**	98.9	65.4	91.2	96.1	527.7
RCTRN$^*_{ACMMM'23}$	79.4	96.6	98.3	**66.9**	92.2	96.8	530.2
KIDRR$^*_{IP\&M'23}$	80.9	96.5	**99.0**	65.0	91.1	96.1	528.6
CMSEI*	81.4	96.6	98.8	65.8	91.8	96.8	531.1
Hire* (ours)	**81.6**$^\dagger{}_{+0.7}$	96.6$_{-0.1}$	**99.0**$_{+0.0}$	66.4$_{-0.5}$	**92.3**$_{+0.1}$	**96.8**$_{+0.0}$	**532.6**$^\dagger{}_{+2.4}$

from another modality for each query with the same number as the batch size. The learning rate is set to 0.0002 initially, with a decay rate of 0.1 every 15 epochs. The maximum epoch number is set to 30. The margin of triplet ranking loss ∇ is set to 0.2. The threshold μ is set to 0.4. For the visual object features, Top-K (K=36) object regions are selected with the highest class detection confidence scores. The visual scene graphs are generated by Neural Motifs (Zellers et al., 2018), and we use the maximum IoU to find the corresponding regions in the original Top-K salient regions. The textual features are extracted by a basic version of the pre-trained 12-layer BERT with a hidden size of 768. The initial dimensions of visual and textual embedding space are set to 2048 and 768, respectively, which are transformed to the same 1024-dimensional (i.e., $D_v = D_s$=1024). Most dimensions of mapping parameters are set to 256-dimensional (D=256) for the joint embedding space. We use 16 (L=16) parallel attention layers in multi-head operations. Similar to Qu et al. (2021), the λ is set to 4 in the image-to-text direction and nine in the text-to-image direction. During the training process, we randomly mask 10% words of each sentence.

7.5.2 Comparison with State-of-the-art Methods

We compare our proposed *Hire* with three kinds of image-text matching methods, including (1) intra-modal interaction-based, inter-modal interaction-based and hybrid-modal interaction-based methods.

- Intra-modal interaction-based methods: SGRAF (Diao, Zhang, Ma, & Lu, 2021b), VSRN

Table 7.2: Comparisons of experimental results on MS-COCO 5K test set. * indicates the performance of an ensemble model. † denotes the statistical significance for p < 0.01 over Recall@1 compared with the best baseline (*i.e.* AME*). Red numbers denote the improvements compared with state-of-the-arts.

Method	Image-to-Text			Text-to-Image			rSum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
$\text{VSRN}^*_{\text{ICCV'19}}$	53.0	81.1	89.4	40.5	70.6	81.1	415.7
$\text{IMRAM}^*_{\text{CVPR'20}}$	53.7	83.2	91.0	39.7	69.1	79.8	416.5
$\text{CAAN}_{\text{CVPR'020}}$	52.5	83.3	90.9	41.2	70.3	82.9	421.1
$\text{VSE}\infty_{\text{CVPR'21}}$	58.3	85.3	92.3	42.4	72.7	$\underline{83.2}$	434.3
$\text{DIME}_{\text{SIGIR'21}}$	59.3	85.4	91.9	43.1	$\underline{73.0}$	83.1	435.8
$\text{VSRN++}^*_{\text{TPAMI'22}}$	54.7	82.9	90.9	42.0	72.2	82.7	425.4
$\text{NAAF}^*_{\text{CVPR'22}}$	58.9	85.2	92.0	42.5	70.9	81.4	430.9
$\text{AME}^*_{\text{AAAI'22}}$	59.9	85.2	92.3	$\underline{43.6}$	72.6	82.7	436.3
$\text{RCTRN}^*_{\text{ACMMM'23}}$	57.1	83.4	91.9	43.6	71.9	83.7	431.6
$\text{KIDRR}^*_{\text{IP\&M'23}}$	$\underline{60.3}$	$\underline{86.1}$	$\underline{92.5}$	43.5	72.8	82.8	$\underline{438.0}$
CMSEI*	61.5	86.3	92.7	44.0	73.4	83.4	441.2
***Hire** (ours)**	**61.7**$^{\dagger}_{+1.4}$	**86.7**$_{+0.6}$	**92.8**$_{+0.3}$	**45.2**$^{\dagger}_{+1.6}$	**74.5**$_{+1.5}$	**84.2**$_{+1.0}$	**445.0**$^{\dagger}_{+7.0}$

(K. Li et al., 2019), VSE∞ (J. Chen et al., 2021) (the reported version with same object inputs), SMFEA(Ge, Chen, et al., 2021), VSRN++ (K. Li et al., 2022), AME (J. Li, Niu, & Zhang, 2022), and CHAN (Pan, Wu, & Zhang, 2023) *etc*. These methods focus on feature enhancement via relationship reasoning within an independent modality.

- Inter-modal interaction-based methods: SGRAF (Diao et al., 2021b), CAAN (Q. Zhang et al., 2020), IMRAM (H. Chen et al., 2020), NAAF (K. Zhang et al., 2022), and RCTRN*(W. Li, Ma, et al., 2023). These methods focus on the multi-modal attention mechanism to explore the cross-modal fine-grained semantic correspondences.

- Hybrid-modal interaction-based methods: CAAN (Q. Zhang et al., 2020), GraDual (S. Long et al., 2022), and DIME (Qu et al., 2021). These methods combine intra- and inter-modal interactions to enhance the visual and textual representations via intra-modal relationship modelling and inter-modal fragment attention modelling.

7.6 Experimental Results

In this section, we report the results of our experiments to evaluate the proposed approach, *Hire*. Note that some ensemble models with “*" are further improved due to the complementarity between multiple models. For a fair comparison, we also provide the ensemble results in Table 7.1, Table 7.2, and Table 7.3, which are averaged similarity scores of image-to-text version and text-to-image version.

Table 7.3: Comparisons of experimental results on Flickr30K 1K test set. '*' indicates the performance of an ensemble model. † denotes the statistical significance for p < 0.01 over Recall@1 compared with the best baseline (*i.e.* AME*)

Method	Image-to-Text			Text-to-Image			rSum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
CAAN$_{CVPR'20}$	70.1	91.6	97.2	52.8	79.0	87.9	478.6
GSMN$^*_{CVPR'20}$	76.4	94.3	97.3	57.4	82.3	89.0	496.8
SMFEA$_{ACMMM'21}$	73.7	92.5	96.1	54.7	82.1	88.4	487.5
SGRAF$^*_{AAAI'21}$	77.8	94.1	97.4	58.5	83.0	88.8	499.6
DIME$^*_{SIGIR'21}$	81.0	$\underline{95.9}$	98.4	63.6	88.1	93.0	520.0
VSRN++$^*_{TPAMI'22}$	79.2	94.6	97.5	60.6	85.6	91.4	508.9
GraDual$^*_{WACV'22}$	78.3	96.0	98.0	64.0	86.7	92.0	511.4
NAAF$^*_{CVPR'22}$	81.9	96.1	98.3	61.0	85.3	90.6	513.2
AME$^*_{AAAI'22}$	$\underline{81.9}$	95.9	$\underline{98.5}$	$\underline{64.6}$	$\underline{88.7}$	$\underline{93.2}$	$\underline{522.8}$
CHAN$_{CVPR'23}$	80.6	96.1	97.8	63.9	87.5	92.6	518.5
RCTRN$^*_{ACMMM'23}$	78.4	95.4	96.8	60.4	84.9	93.7	509.6
KIDRR$^*_{IP\&M'23}$	80.2	94.9	98.0	61.5	84.5	90.1	509.2
CMSEI*	82.3	96.4	98.6	64.1	87.3	92.6	521.3
***Hire** (ours)**	**83.0**† $_{+1.1}$	**97.0**$_{+1.1}$	**98.8**$_{+0.3}$	**65.9**† $_{+1.3}$	**89.1**$_{+0.4}$	**93.4**$_{+0.2}$	**527.1**† $_{+4.3}$

7.6.1 Quantitative Comparison on MS-COCO.

On 5-folds 1K dataset. Table 7.1 presents the experimental results compared with the previous methods on MS-COCO 5-folds 1K. Specifically, compared with the best intra-modal interaction-based method KIDRR* (X. Xie, Li, Tang, Yao, & Ma, 2023), our *Hire* obtains a significant improvement on most metrics, e.g., 81.6% *vs.* 80.9% and 66.4% *vs.* 65.0% on Recall@1 for image-to-text and text-to-image, respectively. Compared with the best inter-model interaction model RCTRN* (W. Li, Ma, et al., 2023) on MS-COCO 1K test set, our *Hire* achieves 2.4% improvements in terms of rSum. Compared with the best hybrid-modal interaction method DIME (Qu et al., 2021), which also combines multiple intra- and inter-model interactions in a multi-layer network, our *Hire* achieves higher results on all metrics, e.g., 81.6% *vs.* 78.8% and 66.4% *vs.* 64.8% in terms of Recall@1 for text retrieval and image retrieval, respectively. And *Hire* clearly outperforms the methods GraDual (S. Long et al., 2022) and KIDRR* (X. Xie et al., 2023), which also employ graph networks, by 7.0% and 4.0% in terms of *rSum*, respectively.

On Full 5K dataset. On the larger image-text matching test data (MS-COCO Full 5K test set), including 5000 images and 25000 sentences, *Hire* obtains a significant improvement on all metrics compared with recent methods as shown in Table 7.2. Compared with the latest state-of-the-arts AME (J. Li et al., 2022), RCTRN* (W. Li, Ma, et al., 2023) and KIDRR* (X. Xie et al., 2023) , our *Hire* achieves 8.7%, 13.4% and 7% improvements in terms of *rSum* via the common protocol (J. Li et al., 2022; K. Li et al., 2022), respectively. And compared with the best hybrid-modal interaction method DIME (Qu et al., 2021), *Hire* also demonstrates superiority (e.g., 61.7% *vs.* 59.3% on Recall@1 of text retrieval and 45.2% *vs.* 43.1% on Recall@1 of image

Table 7.4: Comparison results on cross-dataset generalization from MS-COCO to Flickr30k. $^\natural$ means the results are obtained from their published pre-trained model. † denotes the statistical significance for $p < 0.01$ over Recall@1 compared with the best baseline (*i.e.* DIME*)

Method	Image-to-Text			Text-to-Image			rSum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
VSE++$_{\text{BMVC'18}}$	40.5	67.3	77.7	28.4	55.4	66.6	335.9
LVSE$_{\text{CVPR'18}}$	46.5	72.0	82.2	34.9	62.4	73.5	371.5
SCAN$^*_{\text{ECCV'18}}$	49.8	77.8	86.0	38.4	65.0	74.4	391.4
CVSE$_{\text{ECCV'20}}$	56.4	83.0	89.0	39.9	68.6	77.2	414.1
VSE$\infty^\natural_{\text{CVPR'21}}$	68.0	89.2	93.7	50.0	77.0	84.9	462.8
DIME$^{*\natural}_{\text{SIGIR'21}}$	67.4	90.1	94.5	53.7	79.2	86.5	471.4
CMSEI*	69.6	89.2	95.2	53.7	79.5	87.2	474.4
***Hire** (ours)**	**71.6**$^\dagger_{+3.6}$	**90.5**$_{+0.4}$	**95.2**$_{+0.7}$	**55.0**$^\dagger_{+1.3}$	**80.1**$_{+0.9}$	**87.4**$_{+0.9}$	**479.8**$^\dagger_{+8.4}$

retrieval). It clearly demonstrates the powerful effectiveness of the proposed *Hire* model with the huge improvements.

7.6.2 Quantitative Comparison on Flickr30K

The experimental results on the Flickr30k dataset are shown in Table 7.3. We can observe that our *Hire* outperforms all its competitors with impressive margins on all metrics. In particular, compared with the state-of-the-art method AME (J. Li et al., 2022), *Hire* achieves higher results on all metrics (over 1.1% and 1.3% on Recall@1 for text retrieval and image retrieval, and higher 4.3% in terms of *rSum*). In addition, compared with the most relevant existing work DIME (Qu et al., 2021), *Hire* achieves 2.0%, 2.3% and 7.1% improvements of Recall@1 on image-to-text, Recall@1 on text-to-image and *rSum*, respectively.

7.6.3 Generalization Ability for Domain Adaptation

We further validate the generalization ability of the proposed *Hire* on challenging cross-datasets (It means training the model on one dataset and testing the model on another), which is meaningful for evaluating the cross-modal retrieval performance in real-scenario. Specifically, similar to CVSE (H. Wang et al., 2020), we transfer our model trained on MS-COCO to Flickr30K dataset. As shown in Table 7.4, the proposed *Hire* has an impressive advantage in cross-modal retrieval compared with its competitors. For instance, compared with the best method DIME (Qu et al., 2021), *Hire* achieves significantly outperforms on Recall@1 of text retrieval, Recall@1 of image retrieval, and *rSum* with 4.2%, 1.3% and 8.4% improvements, respectively. It reflects that *Hire* has excellent generalisation capability for cross-dataset image-text matching.

Table 7.5: Ablation studies on MS-COCO 1K test set. All values are ensemble results by averaging two models' (I-T and T-I) similarity. CMSEI*(w/o) means that the spatial-semantic graph is split into two separate graphs, as well as lacking textual semantic enhancement.

Method	Image-to-Text			Text-to-Image			rSum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
Hire	**81.6**	**96.6**	**98.9**	**66.4**	**92.3**	**96.8**	**532.6**
w/o VSA	81.3	96.2	98.4	65.3	91.6	96.3	529.1
w/o TSA	81.5	96.3	98.6	66.2	92.3	96.5	531.4
w/o SA	81.1	96.5	98.7	66.0	92.2	96.7	531.2
CMSEI*(w/o)	80.9	96.0	98.2	65.1	91.5	96.4	528.1
w/o VSSG	80.1	96.2	98.1	64.1	91.5	96.4	526.4
w/o LLII	79.2	95.7	97.6	64.2	91.0	95.5	523.2
w/o LGII	81.1	96.6	98.8	66.0	92.2	96.5	531.2

Table 7.6: Performance comparison of component orders on MS-COCO 1K test set. All values are ensemble results by averaging two models' (I-T and T-I) similarity.

Combination	Image-to-Text			Text-to-Image			rSum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
Hire $\mathscr{A}$(①②) $\mathscr{B}$(③④)	**81.6**	**96.6**	**98.9**	**66.4**	**92.3**	**96.8**	**532.6**
$\mathscr{B}$(③④) $\mathscr{A}$(①②)	71.4	90.8	92.7	64.4	91.1	96.3	506.7
$\mathscr{A}$(②①) $\mathscr{B}$(③④)	81.1	96.0	98.7	66.0	91.8	96.2	529.8
$\mathscr{A}$(①②) $\mathscr{B}$(④③)	81.4	96.6	98.8	66.1	92.2	96.7	531.8

7.6.4 Ablation Studies

In this subsection, we perform detailed ablation studies in Table 7.5 on the MS-COCO 5-folds 1K test set to evaluate the effectiveness of each component in our proposed *Hire*. And we also explore and discuss the impact of different combinations of multiple intra- and inter-modal interactions on the effectiveness of cross-modal retrieval.

Effects of visual-textual implicit reasoning. In Table 7.5, the performance of *Hire* drops from 532.6% to 529.1% and to 531.4%, when removing the visual and textual implicit reasoning model (indicated by w/o VSA or w/o TSA), respectively. When removing the self-attention-based implicit reasoning model, it degrades the Recall@1 score by 0.5% and 0.4% on image-to-text and text-to-image, and reduces 1.4 % in terms of rSum. These observations suggest that implicit attention can slightly improve the information concentration between the fragments within each modality.

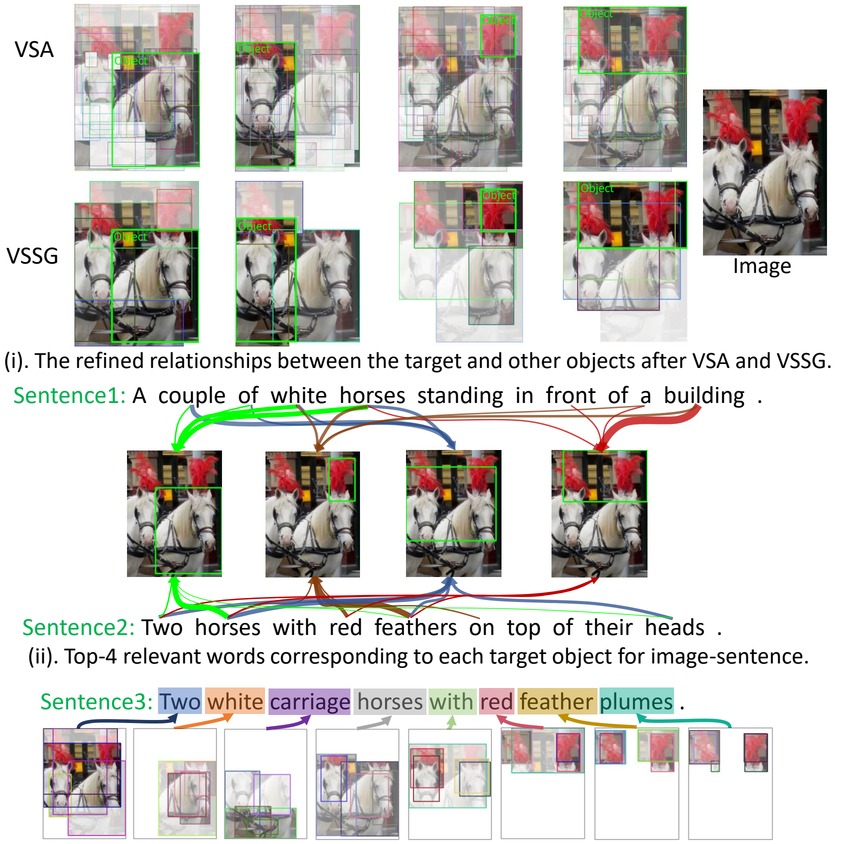
Effects of visual spatial-semantic graph reasoning. In Table 7.5, *Hire* decreases absolutely by 6.2% on MS-COCO 5-fold 1K test set in terms of *rSum* when removing the visual spatial-semantic graph (w/o VSSG). It suggests that spatial-semantic graph reasoning plays an important role in concentrating on relevant regional fragment features, both spatially and semantically. In addition, compared with CMSEI (Ge, Chen, et al., 2023), which split the spatial and semantic relationships into two separate graphs, our *Hire* increases 4.5% in terms of *rSum* on

MS-COCO. It demonstrates that the integration of spatial and semantic relationships can further improve the effective construction of fragment relationships and improve the robustness of the model.

Effects of explicit textual graph reasoning. We also model explicit relationships existing in the text to explore their effects. Specifically, we apply the Stanford enhanced dependency parser (D. Chen & Manning, 2014) following C. Liu et al. (2020) to extract the explicit textual scene graph and use the same R-GCN module as the vision component to model its relationship. However, when adding the textual R-GCN into our model, the matching performance drops from 532.6 to 529.5 in terms of rSum. We speculate that the main reason is that the original sentence already provides richer contextual information than the parsed textual scene graph, where the parsed textual scene graph is incomplete due to the lack of some attributes during the parsing process.

Effects of local-local and local-global inter-modal interactions. We evaluate the impact of the local-local and local-global inter-modal interaction (LLII and LGII) for *Hire*. As shown in Table 7.5, the absence of LLII and the absence of LGII reduce 9.4% and 1.4% in terms of rSum on MS-COCO 5-folds 1K test set, respectively. It is obvious that the multiple inter-modal interactions play a vital role in image-text matching process, which also suggests that cross-modal interactions effectively narrow the semantic gap between the two modalities.

Effects of different combinations. In Table 7.6, we explore the effect of different combinatorial orders of intra- ($\mathscr{A}$: ① implicit intra-modal fragment interaction and ② explicit intra-modal fragment interaction) and inter-modal ($\mathscr{B}$: ③ local-local inter-modal interaction and ④ local-global inter-modal interaction) interactions on cross-modal retrieval. Our *Hire* firstly concentrates the relevant information on each target fragment within modality based on the implicit and explicit relationships and then refines the local features based on the cross-level local-local and local-instance attentions, which can improve the semantic representation of each local fragment and further improve later inter-modal interactions with these contextual relationship enhancements. Specifically, when the inter-modal feature interactions are used first and then the intra-modal feature enhancements are used, the retrieval performance drops from 532.6% to 506.7% in terms of rSum. It suggests that intra-modal interactions integrating potential relationships between the correlated objects into regional features can help the later inter-modal feature interactions obtain more contextual information. Once the order of interactions is reversed, each fragment that obtains contextual information from another modality may be corrupted by subsequent intra-modal interactions, and the original intra-modal relationships will not be accurate based on new contextual object features. Furthermore, we change the order of implicit and explicit relationship reasoning module within intra-modal interaction ($\mathscr{A}$: ①②→②①) and the order of local-local and local-global cross-modal interactions ($\mathscr{B}$: ③④→④③) to evaluate the effectiveness of different combinations of intra-modal interaction and inter-modal interaction, respectively. When the order of implicit and explicit relational reasoning within the modali-


VSA
VSSG
Object
Image
Sentence1: A couple of white horses standing in front of a building .
Sentence2: Two horses with red feathers on top of their heads .
Sentence3: Two white carriage horses with red feather plumes .

(i). The refined relationships between the target and other objects after VSA and VSSG.

(ii). Top-4 relevant words corresponding to each target object for image-sentence.

(iii). Top-5 relevant object regions corresponding to each target word for sentence-image.

Figure 7.3: Visualization of main modules: (i) the refined relationships between the target object (in green box) and other correlated object regions after implicit visual relationship reasoning (VSA) and explicit visual spatial-semantic graph reasoning (VSSG), (ii) results on top-4 region-words pair correspondences of each target object (in green box) for image-to-text, (iii) results on top-5 word-regions pair correspondences of each target word for text-to-image. The degree of white coverage of regions and the thickness of lines indicate different learning weights (best viewed in color).

ties is changed, *Hire* decreases its rSum score to 529.8% on MS-COCO. It suggests that the implicit relational reasoning makes up for the omission of the explicit relationship modelling caused by the scene graph model, thereby improving the fault tolerance of relationship reasoning and model robustness. when changing the order of local-local and local-global inter-modal interactions, the effect of the model does not fluctuate much.

7.6.5 Visualization

In Figure 7.3, to better understand the process of intra- and inter-modal interactions of *Hire*, we visualize (i) the refined relationships between each target object and other objects via the implicit and explicit visual object relationship reasoning modules (VSA and VSSG), (ii) the top-4 relevant words corresponding to each object region for image-to-text, and (iii) the top-5 relevant object regions corresponding to each word for text-to-image after local-local inter-modal interaction. As shown in Figure 7.3 (i), we have observed that the implicit VSA facilitates the information flow between different regions, but it cannot accurately capture object relationships. The proposed explicit VSSG provides more precise spatial and semantic correlations between the object regions, which can concentrate relevant regional information on the target object in both spatial and semantic levels. The combination of implicit and explicit relationship reasoning contributes to the more comprehensive interaction of cross-modal information in multiple levels. In addition, we also visualize the detailed results of the local-local inter-modal interaction for the relevant pairs on the region-words level (Figure 7.3 (ii)) and the word-regions level (Figure 7.3 (iii)) guided by VSSG on image-to-text and text-to-image directions, respectively. The results show that the inter-modal interactions accurately calculate the micro fragment correlations of one modality from the other modality, which reflects its ability on effectively narrowing the semantic gap between different modalities.

7.7 Conclusion

In this chapter, we propose *Hire*, a novel semantic enhanced hybrid-modal interaction method for image-text matching. *Hire* engages in (i) enhancing the visual semantic representation with the implicit and explicit inter-object relationships and (ii) enhancing the visual and textual semantic representation with multi-level joint semantic correlations on intra-fragment, inter-fragment, and inter-instance. To this end, we propose the hybrid-modal (intra-modal and inter-modal) semantic correlations and advance the integrated structured model with cross-modal semantic alignment in an end-to-end representation learning way. Extensive quantitative comparisons demonstrate that our *Hire* achieves state-of-the-art performance on most of the standard evaluation metrics across MS-COCO and Flickr30K benchmarks.

Chapter 8

Visual Semantic-Spatial Self-Highlighting Model

In Chapter 7, while complex intra-modal relational reasoning and inter-modal interactions enhanced the feature representation capabilities of each modality - especially for complex image features - these extensive interactions also significantly impacted retrieval efficiency due to the high computational cost of complex cross-modal exchanges. In contrast, the SMFEA model in Chapter 6 utilises tree-structure alignment supervision in the textual modalities to achieve semantic and structural consistency, thereby improving the semantic representational power of each modality while maintaining a more efficient processing framework. However, SMFEA lacks direct cross-modal interaction, which limits its ability to comprehensively enhance complex visual representations, especially in recognizing and representing salient objects in images.

In this chapter, we propose a novel visual **S**emantic-**S**patial **S**elf-**H**ighlighting **Net**work (termed ***3SHNet***) for high-precision, high-efficiency and high-generalization image-sentence retrieval. 3SHNet highlights the salient identification of prominent objects and their spatial locations within the visual modality, thus allowing the integration of visual semantics-spatial interactions and maintaining independence between two modalities. This integration effectively combines object regions with the corresponding semantic and position layouts derived from segmentation to enhance the visual representation. And the modality-independence guarantees efficiency and generalization. Additionally, 3SHNet utilizes the structured contextual visual scene information from segmentation to conduct the local (region-based) or global (grid-based) guidance and achieve accurate hybrid-level retrieval. Extensive experiments conducted on MS-COCO and Flickr30K benchmarks substantiate the superior performances, inference efficiency and generalization of the proposed 3SHNet when juxtaposed with contemporary state-of-the-art methodologies. Specifically, on the larger MS-COCO 5K test set, we achieve 16.3%, 24.8%, and 18.3% improvements in terms of rSum score, respectively, compared with the state-of-the-art methods using different image representations, while maintaining optimal retrieval efficiency. Moreover, our performance on cross-dataset generalization improves by 18.6%.

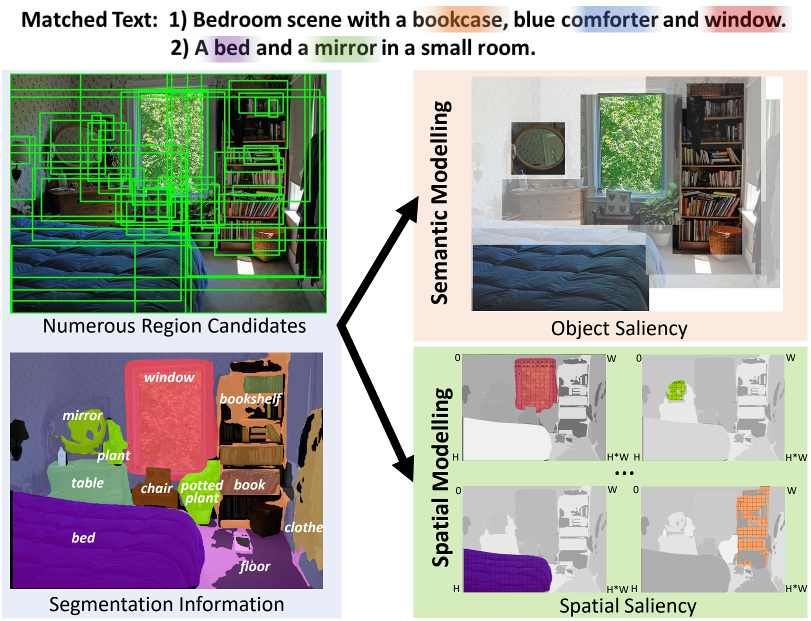


Figure 8.1: Segmentation is combined with the mass object regions to highlight the prominent objects and their locations.

8.1 Introduction

Image-sentence retrieval is a critical task that poses significant challenges. One of the primary difficulties is the inherent semantic gap when measuring the precise semantic similarities between the visual and textual modalities. Especially the textual semantics are more specific than the visual semantics since the sentence contains well-structured and explicitly-semantical instances, while the image involves rich visual semantic objects and a complex context. The semantic gap is widening with the enormous development of the recent powerful language models, such as BERT (Devlin et al., 2018).

Two popular schemes are developed to enhance the visual semantic features for image-sentence retrieval. One is text-dependent visual representation learning (H. Chen et al., 2020; Ge, Chen, et al., 2023; Pan et al., 2023; Qu et al., 2021) while the other is hybrid-level visual representation enhancing (Guo et al., 2023; H. Zhang, Mao, Zhang, & Zhang, 2022; Y. Zhang, Zhou, Wang, Tian, & Li, 2020). On the one hand, text-dependent visual representation learning methods provide fine-grained correspondence across two modalities, where the highly relevant words are encoded into each object region by combination or attention. For instance, in SCAN (K.-H. Lee et al., 2018), text-aware visual features for image-to-sentence retrieval were crafted through an attention-based cross-modal interaction, which involved the model selectively attending to object regions associated with each word, amplifying the emphasis on visual semantics within the generated features. However, due to the joint embedding and deep interaction of visual and textual semantic features during both training and inference, these methods have two significant defects: (i) a massive redundant computation reduces the inference speed of retrieval since the visual feature of an image always needs to be recomputed once its similarity is estimated with a new sentence, and (ii) due to the data differences and the differential effects of the different-dataset sentences on the visual representation, the trained model on one dataset cannot directly test well on another, which impedes cross-domain generalization.

On the other hand, to gain a deeper understanding of images that includes global contextual information and fine-grained local representation, the latest studies (J. Chen et al., 2021; D. Wu et al., 2022; H. Zhang et al., 2022; Y. Zhang et al., 2020) have proposed hybrid-level visual representations, which combine both local and global features. This approach allows for the complementary advantages of local fragments and global image features to be fully realized. These methods improve retrieval performance by capturing all possible semantics in global and local features without relying on textual guidance. For example, D. Wu et al. (2022) fused local-level region and global-level features through a two-stage interaction strategy for image-text retrieval to extract a more holistic image representation. However, in human interaction, people tend to focus on prominent objects and their spatial locations (Madani et al., 2018; Walther, 2006), which is ignored in D. Wu et al. (2022). In image-sentence retrieval, the alignment between visual and textual data involves the consistencies of the prominent objects and their spatial locations, as the textual annotations inherently reflect the annotator's attention. Consequently, it is logical to incorporate human-like attention modelling into a visual representation, especially when the visual and textual modalities remain independent.

8.1.1 Motivation

Motivated by the aforementioned insights, the main research objectives of this study lie in two aspects: (i) to overcome the deep textual dependence for visual representation learning; and (ii) to explore the human-like attention from two aspects, *i.e.* object semantic- and spatial-level, via a visual salient interaction schema in an end-to-end framework for cross-modal alignment. To this end, we propose a visual semantic-spatial self-highlighting network (3SHNet) towards high-precision, high-efficiency, high-generalization image-sentence retrieval. In particular, 3SHNet highlights the prominent objects and their spatial locations via the visual multi-modal interaction between the object regions and the segmentation results. This approach emphasises the semantic and spatial saliencies arising from the intersection of the scattered regions and the structured visual scene information. It unifies them based on the correspondence between the semantic and position layouts derived from segmentation, as illustrated in Figure 8.1. In fact, most of the current segmentation and multi-level object representation schemas are inspired by the human's visual attention perception, as revealed by Anderson et al. (2018); Vacher, Launay, Mamassian, and Coen-Cagli (2023); H. Zhu, Meng, Cai, and Lu (2016), where the segmentation schema is deemed to simulate the process of capturing the meaningful cues for human-like attention while the object representation schema simulates the process of formulating human-like attention. Thus, we argue that the fusion of segmentation and object-region features more comprehensively coincides with the information process of human-like attention since both of them are unique and indispensable parts of bio-inspired attention. Additionally, 3SHNet takes full advantage of the structured contextual visual scene information from segmentation to conduct the local (region-based) or global (grid-based) guidance to achieve accurate hybrid-level retrieval. Thus, 3SHNet self-highlights the semantic-spatial saliencies to reduce the visual-textual gap while keeping the visual-textual modality independent. This allows 3SHNet to pre-extract and retain the visual features for each image before inference, even though there are new linguistic queries or new linguistic candidate sets and makes 3SHNet insensitive to sentence differences across different datasets. 3SHNet can provide an efficient and effective paradigm to inspire existing real-world application scenarios, such as building a multi-modal recommendation system to help users quickly find the products or building a smart photo album to help users quickly find the images they need, etc.

8.1.2 Contribution

Outlined below are the contributions made by this chapter:

- We explore the high-precision, high-efficiency, and high-generalisation image-sentence retrieval when the visual modality is independent of the textual modality. To achieve this, human-like attention is forged within the visual modality.

- We introduce a novel visual semantic-spatial self-highlighting network (3SHNet), where the segmentation is first utilized in image-sentence retrieval and interacted with the global and local visual features as the structured contextual guidance for semantic-spatial saliencies. This allows for a unified interpretation of semantic-spatial saliencies over the segmentation maps.

- The proposed 3SHNet is verified to attain the state-of-the-art (SOTA) retrieval performances on two standard image-sentence retrieval benchmarks, *i.e.* MS-COCO and Flickr30K. Especially, 3SHNet has proven superior on the local-level, global-level, and hybrid-level visual features compared to the SOTAs, demonstrating its robustness. Furthermore, the high-efficiency of 3SHNet is verified with higher inference speed compared to the SOTA, while the high-generalization is demonstrated on the cross-dataset training-testing setting.

8.2 Related Work

Current studies to image-sentence retrieval can be categorized based on whether they incorporate visual-textual interaction or not: (i) cross-modal interaction retrieval (Ge, Chen, et al., 2023; K.-H. Lee et al., 2018; W.-H. Li et al., 2021; Qu et al., 2021), and (ii) modality-independent representation retrieval (J. Chen et al., 2021; Ge, Chen, et al., 2021; H. Liu et al., 2018; Y. Ma et al., 2023; S. Wang, Chen, et al., 2018). Especially, most works explored the fine-grained correspondence of the cross-modality local-level representations, *i.e.* employing cross-modal attention mechanisms to establish connections between image regions and sentence words. For instance, DIME (Qu et al., 2021) adopted a multi-layer multiple cross-modality interaction framework by cross-modal attention-aware region/word aggregating and region-sentence/word-image correspondence learning. Besides, more successes were also witnessed in recent large-scale pre-training visual-language models (Y.-C. Chen et al., 2020; J. Li et al., 2021; X. Li et al., 2020), which rely heavily on large-scale data-driven pattern and the powerful computation facilities, *e.g.*, 400M image-text pairs and 256 V100 GPUs are used in Radford et al. (2021). However, since they rely on the deep cross-modal interaction and need a new traversal computation cost once there is a new query of image or sentence, the retrieval takes a long inference time, which is hardly applied to real-life scenarios. To address this issue, many recent modality-independent representation learning methods (J. Chen et al., 2021; Ge, Chen, et al., 2021; Z. Li, Guo, Feng, Hwang, & Xue, 2022) are proposed to encode visual and language information from both modalities into a joint embedding space without using any cross-attention interactions.

However, due to the lack of textual guidance, most modality-independent representation retrieval methods (J. Chen et al., 2021; D. Wu et al., 2022) proceed from the essence of modality-independent representation learning to elaborate visual-intrinsic feature enhancement models that can substitute textual guidance to reduce the semantic gap. On the one hand, some methods (Cheng et al., 2022; S. Long et al., 2022) found the problem of missing relationships between visual objects and improved visual contextual representation by detecting object associations in scene graphs. On the other hand, some latest works (D. Wu et al., 2022; Y. Zhang et al., 2020) started from various image representations to enhance the visual distinguishability via combining multiple levels of visual representation, *i.e.,* local- and global-level image features. Indeed, they boost retrieval performance; however, they still neglect how humans pay attention to the prominent objects and their locations on the multiple levels of visual representation.

SAN (Ji et al., 2019) introduced a global saliency detector to generate salience-weighted maps for images with additional supervision in a cross-modal interaction retrieval framework. However, these saliency maps lack semantic and spatial information about objects and background information, and retrieval speed is still limited due to text dependency. Visual semantic segmentation (Hu, Chen, et al., 2023; Hu, Huang, et al., 2023; M. Wu et al., 2022) can represent coarse-grained image semantics and precise spatial locations, which is usually used in many studies (Mousavian, Košecká, & Lien, 2015; Y. Zhao, Yu, Gao, & Shen, 2022). For example, DIFNet ((M. Wu et al., 2022)) took segmentation feature as another visual information flow to improve image captioning performance, where the segmentation features are used independently of the original visual features. However, these methods fuse the segmentation information and weaken the most important characteristics of semantic segmentation results, such as accurate high-level category semantic information and its explicit spatial locations, which are also absent

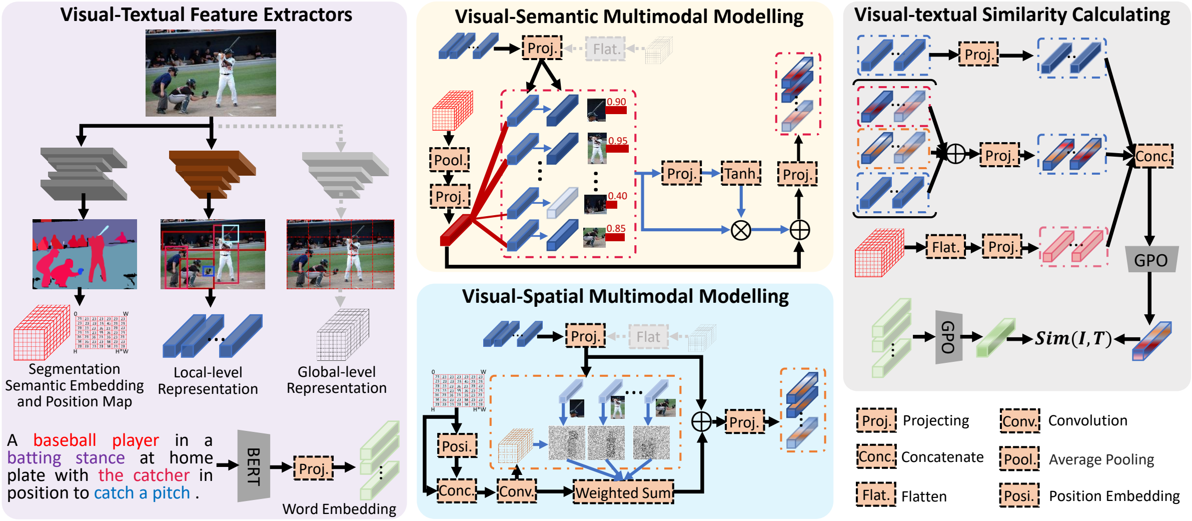

Visual-Textual Feature Extractors
Segmentation Semantic Embedding and Position Map
Local-level Representation
Global-level Representation
A baseball player in a batting stance at home plate with the catcher in position to catch a pitch .
BERT
Proj.
Word Embedding
Visual-Semantic Multimodal Modelling
Proj.
Flat.
Pool.
Proj.
0.90
0.95
0.40
0.85
Proj.
Tanh
Proj.
Visual-Spatial Multimodal Modelling
Proj.
Flat
Posi.
Conc.
Conv.
Weighted Sum
Proj.
Visual-textual Similarity Calculating
Proj.
Proj.
Conc.
Flat.
Proj.
GPO
GPO
Sim(I,T)
Proj. Projecting
Conv. Convolution
Conc. Concatenate
Pool. Average Pooling
Flat. Flatten
Posi. Position Embedding

Figure 8.2: Illustration of the proposed 3SHNet. It mainly consists of visual-semantic modelling module (VSeM) and visual-spatial modelling module (VSpM), where the semantic feature and the position map of the segmentation are respectively imposed to guide the local- and global-level visual features in visual multimodal interactions.

in salient object detection (Borji, Cheng, Jiang, & Li, 2015; Ji et al., 2019; Pang, Zhao, Zhang, & Lu, 2020).

8.3 The Proposed Method – 3SHNet

Figure 8.2 illustrates the framework of our proposed 3SHNet. 3SHNet mainly consists of two innovative visual-semantic multimodal modelling (VSeM) and visual-spatial multimodal modelling (VSpM) modules to respectively highlight the prominent objects and their spatial locations via visual semantic-spatial salient interactions over segmentation maps. They allow to self-highlight within the visual modality, effectively substituting attention cues from sentences, while maintaining high-efficiency and high-generalization. For clarity, the main notations and their definitions throughout the chapter are shown in Table 8.1.

8.3.1 Visual-Textual Feature Extractors

For readability, we first introduce the feature extraction process of 3SHNet. 3SHNet use both the local- and global-level image representations to fully capture the comprehensive visual semantics. For fine-grained local-level image representation, we use bottom-up-attention network (Anderson et al., 2018) to extract K sub-region features V^l=$\{v_1^l, \cdots, v_K^l\}$ to cover the main semantic and represent the whole image I∈$\mathbb{R}^{H^I \times W^I \times 3}$. For contextual global-level image representation, we use ResNeXt (S. Xie, Girshick, Dollár, Tu, & He, 2017) to extract the grid-based features $V^g \in \mathbb{R}^{H \times W \times 2048}$ of the whole image after an AdaptiveAvgPool2d pooling operation

Table 8.1: Main notations and their definitions.

Notation	Definition
I	an image in database
T	a sentence in database
V^l	the sub-region features of the image
K	the number of detected sub-regions for each image
V^g	the grid-based image features of the image
V^s	the semantic segmentation feature of the image
V^m	the segmentation map for the image
E	the word features for each sentence
α_i	the salience weights of i-th sub-region in the image
$\{\dot{v}^l\}$	the attention-aware sub-region features
$\{\ddot{v}^l\}$	the fine-grained salience object representations after visual-semantic multimodal modelling
$PE()$	the position encoding function
p	each pixel index in the image
$\ddot{V}^m$	the refined feature of segmentation map
β_{ij}	the position correspondence coefficient between the i-th region and j-th position index
$\ddot{V}^l$	the visual-spatial representations after Visual-spatial multimodal modelling

(M. Lin et al., 2014). To enhance visual representation of the prominent objects and their locations, we introduce the segmentation feature $V^s \in \mathbb{R}^{H \times W \times C^s}$ and segmentation map $V^m \in \mathbb{R}^{H^I \times W^I}$ for image I, where H, W, H^I, W^I, C^s are respectively feature height, feature width, image height, image width and the semantic categories. These variables are extracted from an FPN-based network (Y. Xiong et al., 2019) containing high-level object semantics and their corresponding spatial information.

For the sentences, we employ a pre-trained language model, specifically BERT (Devlin et al., 2018), for the extraction of word-level textual representations, aligning with the contemporary approach in contemporary natural language processing research. Specifically, we extract the textual features $E = \{e_1, e_2, ..., e_N\}$ ($e_i \in \mathbb{R}^{768}$) via the pre-trained BERT (Devlin et al., 2018) for each sentence, where N is the number of words. To get a joint embedding space measured with visual representation, we utilize a fully connected (FC) layer to map the extracted word features into a D-dimensional space for each sentence T.

8.3.2 Visual Semantic-Spatial Multimodal Modelling

We aim to reconstruct image representations from two perspectives: visual-semantic multimodal modelling and visual-spatial multimodal modelling. Figure 8.2 illustrates the training process of 3SHNet, where the image is transformed into either fine-grained local-level object region features or global-level grid-based features for semantic-spatial modelling. We use object region features as an example.

Visual-semantic multimodal modelling.

As shown in the middle of Figure 8.2, guided by the semantic segmentation features, the salience of the object region is highlighted. And they further interact together. Specifically, the semantic segmentation feature V^s is first projected into a coarse-grained D-dimensional semantic space by an FC layer after a global average pooling operation. To distinguish the differentiation not only between center and marginal regions but also within marginal regions, we calculate the cosine similarities for all segmentation-region pairs and then obtain the salience weights $\{\alpha_i\}$ of all object regions guided by the segmentation features via a Sigmoid function (McCulloch & Pitts, 1943) (Softmax function (Chorowski et al., 2015) is also available as discussed in detail in Section 8.4.5). These can be formulated as:

$$\ddot{V}^s = \mathrm{FC}(\mathrm{AvgPool}(V^s)),\ \ddot{V}^s \in \mathbb{R}^D, \tag{8.1}$$

$$\alpha_i = \mathrm{Sigmoid}(\frac{(\ddot{V}^s)^T (W^l v_i^l)}{\sqrt{D}||\ddot{V}^s||\ ||W^l v_i^l||}), i \in [1, K], \tag{8.2}$$

where α_i is the salience weight of i-th region and W^l denotes the linear projection parameter shared for the i-th region feature mapping as key and value elements. After these, we can obtain the salience regions $\{\dot{v}_i^l\}$, where $\dot{v}_i^l = \alpha_i W^l v_i^l$. Then we conditionally fuse the weighted fine-grained object features with the segmentation features to further enhance their semantic representation as follows:

$$\ddot{v}_i^l = \ddot{W}^l(\mathrm{Tanh}(\dot{W}^l \dot{v}_i^l)\dot{v}_i^l + \ddot{V}^s), \tag{8.3}$$

where $\dot{W}^l, \ddot{W}^l$ denote the linear projection parameters. Finally, we obtain the fine-grained salience object representations $\ddot{V}^l = \{\ddot{v}_i^l\}$ combined with the semantically definitive segmentation features.

Visual-spatial multimodal modelling.

Different from approaches (Ge, Chen, et al., 2023; K. Li et al., 2022) that model spatial relationships among objects, we take advantage of explicit and salient object spatial segmentation boundaries in semantic segmentation maps to explore the positional relevance of visual local semantic and structured semantics. It also differs from Transformer (Vaswani et al., 2017), which simply embeds the position information and attends among multiple components. In particular, a positional encoding function (Vaswani et al., 2017) based on a trigonometric function is applied to embed each pixel index $p \in [1, H^I \times W^I]$ of the segmentation map $V^m \in \mathbb{R}^{H^I \times W^I}$ for an image I in a dense vector, as follows:

$$PE_j(p) = \begin{cases} \mathrm{Sin}(p/10000^{j/d}), & if\ j\ is\ even, \\ \mathrm{Cos}(p/10000^{j/d}), & if\ j\ is\ odd, \end{cases} \tag{8.4}$$

where $j \in [1,d]$ and d is dimensions of the positional embedding. As shown in the middle of Figure 8.2, we concatenate the positional embedding and the normalized segmentation map into a new dense vector $\dot{V}^m \in \mathbb{R}^{H^I \times W^I \times (d+1)}$. Then a convolutional layer is used to down-sample the vector, which can further refine the positional embedding with certain semantics. The above process can be formulated as:

$$\dot{V}^m = \text{Concat}(\text{PE}(V^m_p), V^m), \tag{8.5}$$

$$\ddot{V}^m = \text{Conv2d}(\dot{V}^m), \tag{8.6}$$

where $\ddot{V}^m \in \mathbb{R}^{H^p \times W^p \times C^p}$ ($C^p \ll C^s$ to keep model efficient) serve as the key and value for next visual-spatial attention modelling. Similarly, each region feature v^l_i is projected into the same dimension with $\ddot{V}^m$ as a query and then calculates the position correspondence coefficient β_i with the refined positional embedding $\ddot{V}^m$ by a per-dimension λ-smoothed Softmax (Chorowski et al., 2015). A spatially concentrated feature will be obtained for each object, as follows:

$$c_{ij} = \frac{(U^l v^l_i)(\ddot{V}^m_j)^T}{||U^l v^l_i||\ ||\ddot{V}^m_j||}, i \in [1,K], j \in [1, H^p \times W^p], \tag{8.7}$$

$$\beta_{ij} = \frac{exp(\lambda c_{ij})}{\sum_{j=1}^{H^p \times W^p} exp(\lambda c_{ij})}, \tag{8.8}$$

$$\dddot{v}^m_i = \sum_{j=1}^{H^p \times W^p} \beta_{ij} \ddot{v}^m_j, \tag{8.9}$$

where U^l denotes the linear projection parameter. Finally, we combine the spatial embeddings and the corresponding region features with a mapping parameter $\dddot{U}^l$ as visual-spatial representations $\dddot{V}^l = \{\dddot{v}^l_i\}$.

$$\dddot{v}^l_i = \dddot{U}^l(\dddot{v}^m_i + U^l v^l_i) \tag{8.10}$$

Note that we focus on estimating the position relevance of the high-level local object semantics and the segmentation semantics to acquire the local positional representation. Additionally, the local positional representation is associated with local object semantic features in Eq.(10) to enhance the semantic-spatial representation.

8.3.3 Feature Aggregation and Objective Function

To calculate the visual-textual similarities, as shown in the right of Figure 8.2, we aggregate multiple representations into a measurably uniform embedding space. For the visual aggregation, we first combine the semantically enhanced representations, the spatially enhanced representations and the original region features and project them as the visual semantic-spatial representations. Then, we aggregate these fine-grained local-level visual features, visual semantic-spatial features, and semantic segmentation features for each image by a popular generalized pooling operator (GPO) (J. Chen et al., 2021), since GPO automatically seeks the best pooling function compared to the traditional pooling strategy, *e.g.* max-pooling. Similarly, we use GPO to obtain the final textual embedding space for textual aggregation. Finally, the similarity scores are calculated between two-modality representations. During training, a bidirectional triplet ranking loss with hard negative mining (Faghri et al., 2017) is adopted as the optimization objective, as follows:

$$\begin{aligned}\mathscr{L}_{rank} = \sum_{(I,T)} \{&\max[0, \gamma - \mathrm{Cos}(I,T) + \mathrm{Cos}(I,\bar{T})] \\ &+ \max[0, \gamma - \mathrm{Cos}(I,T) + \mathrm{Cos}(\bar{I},T)]\}\end{aligned} \tag{8.11}$$

where γ is a margin constraint.

8.4 Experiments

8.4.1 Implementation details.

The whole network except the offline visual extractors is implemented with PyTorch on a single TITAN RTX GPU using AdamW optimizer with weight decay factor 10e-4, where the learning rate is set to 5e-4 initially. The maximum epoch number is configured at 25, accompanied by a mini-batch size of 256. The joint embedding space possesses a dimensionality of 1024. The margin parameter γ is specified as 0.2. We used the same pre-extracted features as the compared methods to guarantee the fairness. Specifically, we used the pre-extracted local-level region features, global-level grid features and segmentation results from Faster-RCNN (Ren et al., 2015), ResNext-101 (S. Xie et al., 2017), and UPSNet (Y. Xiong et al., 2019), respectively. The single-thread feature extraction speeds of these visual encoders are 2.1 FPS for Faster-RCNN, 2.1 FPS for ResNext-101 and 10.5 FPS for UPSNet, respectively. These models are pre-trained on small-scale datasets, such as ImageNet (Russakovsky et al., 2015) and Visual Genome (Krishna et al., 2017), *etc.* The main fine-grained pre-extracted local- and global-level visual features are the same as the compared methods to guarantee fairness. For the local-level visual feature, the strategy entails choosing 36 regions (K=36) characterized by the highest confidence scores in object detection (Ren et al., 2015), containing some redundant and useless regions. For the global-level visual feature, the size of grid features is 7×7×2048. Following (M. Wu et al., 2022), the semantic segmentation feature size is 7×7×133, in which the dimension of 133 is logit corresponding to object categories. The size of the segmentation map is resized to 64×64 to reduce calculations. Note that although we use the extra segmentation features, the off-line speed testing is not influenced since it excludes the feature extraction process in a practical way. The segmentation features without our deep feature interactions have inapparent gains on the retrieval performance according to the comparisons in later ablation studies, and Table 8.2 and Table 8.3 also show the comparisons between ours and the embeddable SOTA method (VSE$\infty^{\text{w/ Seg.}}$ in same mini-batch) with such extra features.

Table 8.2: Comparisons of performances on MS-COCO 1K (5-folds) test set. The best results are highlighted in bold typeface. $*$ indicates the performance metrics attributed to the ensemble model. For clearer comparison, the ensemble model is shown with a blue background and the improvements of the best contrasting method with underline are marked.

Type	Method	Image-to-Sentence			Sentence-to-Image			rSum
		Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
Region	IMRAM*	76.7	95.6	98.5	61.7	89.1	95.0	516.6
	VSE∞	79.7	96.4	98.9	64.8	91.4	96.3	527.5
	DIME*	78.8	96.3	98.7	64.8	91.5	96.5	526.6
	VSRN++*	77.9	96.0	98.5	64.1	91.0	96.1	523.6
	NAAF*	80.5	96.5	98.8	64.1	90.7	96.5	527.2
	AME*	79.4	96.7	98.9	65.4	91.2	96.1	527.7
	CMSEI*	81.4	96.6	98.8	65.8	91.8	96.8	531.1
	CHAN	81.4	96.9	98.9	66.5	92.1	96.7	532.6
	RCTRN*	79.4	96.6	98.3	66.9	92.2	96.8	530.2
	KIDRR*	80.9	96.5	99.0	65.0	91.1	96.1	528.6
	DCIN*	81.4	96.8	99.0	66.1	92.1	96.6	532.0
	EKDM*	81.4	96.7	**99.4**	68.5	**93.5**	**97.6**	537.1
	DCIN*	81.4	96.8	99.0	66.1	92.1	96.6	532.0
	MKTLON*	81.8	96.6	98.8	66.1	91.6	96.6	531.5
	3SHNet	83.1	97.2	99.3	68.7	92.4	96.6	537.3
	3SHNet*	**$84.3_{+2.9}$**	**$97.3_{+0.6}$**	$99.1_{-0.3}$	**$69.7_{+1.2}$**	$93.1_{-0.4}$	$97.0_{-0.6}$	**$540.5_{+3.4}$**
Grid	SCO	69.9	92.9	97.5	56.7	87.5	94.8	499.3
	VSE∞	80.4	96.8	99.1	66.4	91.1	95.5	531.6
	3SHNet	83.1	97.5	99.1	69.8	92.7	96.8	538.9
	3SHNet*	**$84.5_{+4.1}$**	**$97.7_{+0.9}$**	**$99.3_{+0.2}$**	**$70.6_{+4.2}$**	**$93.3_{+2.2}$**	**$97.1_{+1.0}$**	**$542.5_{+10.9}$**
Region+Grid	CRGN	73.8	95.6	98.5	60.1	88.9	94.5	511.4
	MLSL	77.1	96.3	98.6	63.8	90.1	95.9	521.8
	VSE∞	82.2	97.5	**99.5**	68.1	92.9	97.2	537.4
	CMCAN	81.2	96.8	98.7	65.4	91.0	96.2	529.3
	Imp.*	83.7	97.7	99.1	68.4	92.8	**97.5**	539.2
	RAAN*	76.8	96.4	98.3	61.8	89.5	95.8	518.6
	HGAN	81.1	96.9	99.0	67.4	92.2	96.6	533.2
	VSE∞$^{w/\ Seg.}$	83.1	97.6	99.5	68.9	93.0	97.2	539.3
	3SHNet	85.0	**97.7**	99.2	71.2	93.5	97.2	543.7
	3SHNet*	**$85.8_{+2.1}$**	**$97.7_{+0.0}$**	$99.3_{+0.2}$	**$71.8_{+3.4}$**	**$93.7_{+0.9}$**	$97.4_{-0.1}$	**$545.7_{+6.5}$**

8.4.2 Quantitative Comparison

We report the performances of 3SHNet on MS-COCO and Flick30K with the local-level region-based image features, the global-level grid-based image features and the hybrid-level (region+grid) image features, respectively in Table 8.2, Table 8.3 and Table 8.4, compared with the corresponding state-of-the-art studies, including (1) region-based methods, *i.e.,* IMRAM* (H. Chen et al., 2020), SGRAF (Diao et al., 2021a), VSE∞ (J. Chen et al., 2021), DIME* (Qu et al., 2021), NAAF* (K. Zhang et al., 2022), CHAN (Pan et al., 2023), CMSEI* (Ge, Chen, et al., 2023), DCIN* (W. Li, Su, et al., 2023), RCTRN* (W. Li, Ma, et al., 2023), KIDRR* (X. Xie et al., 2023) and MKTLON* (X. Qin, Li, Hao, Ge, & Pang, 2024) *etc.,* (2) grid-based methods, *i.e.,* SCO* (Huang et al., 2018) and VSE∞ (J. Chen et al., 2021), and (3) region-grid-based methods, *i.e.,* CRGN (Y. Zhang et al., 2020), MLSL (W.-H. Li et al., 2021), CMCAN (H. Zhang et al.,

Table 8.3: Comparisons of performances on larger 5K test set. The best results are highlighted in bold typeface. $*$ indicates the performance of the ensemble model. For clearer comparison, the ensemble model is shown with a blue background and the improvement of the best contrasting method (with underline) is marked.

Type	Method	Image-to-Sentence			Sentence-to-Image			rSum
		Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
Region	IMRAM*	53.7	83.2	91.0	39.7	69.1	79.8	416.5
	VSE∞	58.3	85.3	92.3	42.4	72.7	83.2	434.3
	DIME*	59.3	85.4	91.9	43.1	73.0	83.1	435.8
	VSRN++*	54.7	82.9	90.9	42.0	72.2	82.7	425.4
	NAAF*	58.9	85.2	92.0	42.5	70.9	81.4	430.9
	AME*	59.9	85.2	92.3	43.6	72.6	82.7	436.3
	CMSEI*	61.5	86.3	92.7	44.0	73.4	83.4	441.2
	CHAN	59.8	87.2	93.3	44.9	74.5	84.2	443.9
	RCTRN*	57.1	83.4	91.9	43.6	71.9	83.7	431.6
	KIDRR*	60.3	86.1	92.5	43.5	72.8	82.8	438.0
	DCIN*	60.8	86.3	93.0	44.0	74.6	84.3	443.0
	MKTLON*	61.4	86.7	92.8	44.3	73.9	83.7	442.8
	3SHNet	63.8	88.1	94.0	47.0	76.6	85.4	454.9
	3SHNet*	**65.3**$_{+5.5}$	**88.8**$_{+1.6}$	**94.1**$_{+0.8}$	**48.2**$_{+3.3}$	**77.5**$_{+3.0}$	**86.3**$_{+2.1}$	**460.2**$_{+16.3}$
Grid	SCO	42.8	72.3	83.0	33.1	62.9	75.5	369.6
	VSE∞	59.1	85.9	92.8	44.1	74.1	84.0	440.0
	3SHNet	64.1	88.9	94.3	48.0	77.4	86.3	459.0
	3SHNet*	**66.2**$_{+7.1}$	**89.8**$_{+3.9}$	**94.7**$_{+1.9}$	**49.0**$_{+4.9}$	**78.3**$_{+4.2}$	**86.8**$_{+2.8}$	**464.8**$_{+24.8}$
Region+Grid	CRGN	51.2	80.6	89.7	37.4	68.0	79.5	406.4
	VSE∞	62.5	87.8	94.0	46.0	75.8	85.7	451.8
	CMCAN	61.5	-	92.9	44.0	-	82.6	-
	Imp.*	63.5	87.9	93.5	46.8	76.1	85.1	452.9
	HGAN	60.0	85.8	92.8	45.4	75.3	85.1	444.4
	VSE$\infty^{\text{w/ Seg.}}$	63.9	88.3	94.2	47.8	76.9	86.0	457.1
	3SHNet	67.1	89.8	95.2	49.9	78.8	87.2	468.0
	3SHNet*	**67.9**$_{+4.4}$	**90.5**$_{+2.6}$	**95.4**$_{+1.9}$	**50.3**$_{+3.5}$	**79.3**$_{+3.2}$	**87.7**$_{+2.6}$	**471.2**$_{+18.3}$

2022), Imp.* (D. Wu et al., 2022), RAAN* (Y. Wang et al., 2023) and HGAN* (Guo et al., 2023). Our 3SHNet achieves the best on the above three different image features. Following (J. Chen et al., 2021), we calculate the average of the ranking results of local- and global-level inputs as the final hybrid-level ranking results of ensemble patterns.

Quantitative comparison on MS-COCO.

Table 8.2 and Table 8.3 present the quantitative results on two distinct MS-COCO test sets, 5-folds 1K and full 5K (the latter is a larger retrieval test set comprising 5000 images and 25000 sentences). Our 3SHNet significantly exceeds existing state-of-the-art methods on all recall metrics with different visual features. Specifically, for region-based features, compared to the best text-dependent method CHAN (Pan et al., 2023) on MS-COCO 1K test sets, our single-model 3SHNet achieves improvements of 1.7% and 2.2% on Recall@1 of image-to-sentence retrieval and sentence-to-image retrieval, respectively. The ensemble model 3SHNet* also gets improvements of 3.4% on rSum compared to the text-dependent ensemble method EKDM* (S. Yang et al., 2023). On the larger 5K test set, both our 3SHNet and 3SHNet* achieve significant improvements compared to CHAN (Pan et al., 2023)and DCIN* (W. Li, Su, et al., 2023) with 454.9(+11.0) and 460.2(+17.2) on rSum, respectively. Notably, our single-model 3SHNet out-

Table 8.4: Comparisons of performances on Flickr30K 1K test set. ∗ indicates the performance metrics attributed to the ensemble model. The best results are indicated in bold. For clearer comparison, our ensemble model is shown with a blue background and the improvement of the best contrasting method (with underline) is marked.

Method	Image-to-Sentence			Sentence-to-Image			rSum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
With region-based image representation							
IMRAM*	74.1	93.0	96.6	53.9	79.4	87.2	484.2
VSE∞	81.7	95.4	97.6	61.4	85.9	91.5	513.5
DIME*	81.0	95.9	98.4	63.6	88.1	93.0	520.0
VSRN++*	79.2	94.6	97.5	60.6	85.6	91.4	508.9
NAAF*	81.9	96.1	98.3	61.0	85.3	90.6	513.2
AME*	81.9	95.9	98.5	64.6	88.7	93.2	522.8
CMSEI*	82.3	96.4	**98.6**	64.1	87.3	92.6	521.3
CHAN	80.6	96.1	97.8	63.9	87.5	92.6	518.5
RCTRN*	78.4	95.4	96.8	60.4	84.9	93.7	509.6
KIDRR*	80.2	94.9	98.0	61.5	84.5	90.1	509.2
EKDM*	82.3	96.5	98.5	61.5	86.0	90.9	515.7
3SHNet	82.0	96.2	98.3	64.8	87.3	92.8	521.4
3SHNet*	**84.7**$_{+2.8}$	**96.8**$_{+0.9}$	98.0$_{-0.5}$	**66.1**$_{+1.5}$	**88.7**$_{+0.0}$	**93.4**$_{+0.2}$	**527.8**$_{+5.0}$
With grid-based image representation							
SCO*	55.5	82.0	89.3	41.1	70.5	80.1	418.5
VSE∞	81.5	**97.1**	98.5	63.7	88.3	93.2	522.3
3SHNet	83.9	96.7	97.9	65.1	88.6	93.3	525.5
3SHNet*	**84.9**$_{+3.4}$	97.0$_{-0.1}$	**98.5**$_{+0.0}$	**67.2**$_{+3.5}$	**89.6**$_{+1.3}$	**94.0**$_{+0.8}$	**531.3**$_{+9.0}$
With region- and grid-based image representation							
CRGN	70.5	91.2	94.9	50.3	77.7	85.2	469.8
MLSL	72.2	92.4	98.2	56.8	83.3	91.3	494.2
VSE∞	85.3	97.2	98.9	66.7	89.9	94.0	532.0
CMCAN	79.5	95.6	97.6	60.9	84.3	89.9	507.8
Imp.*	84.5	97.3	99.0	66.8	89.7	94.3	531.6
RAAN*	77.1	93.6	97.3	56.0	82.4	89.1	495.5
HGAN	80.3	96.5	98.3	62.3	87.8	93.1	518.3
3SHNet	86.1	97.6	98.8	68.6	90.1	94.4	535.6
3SHNet*	**87.1**$_{+2.6}$	**98.2**$_{+0.9}$	**99.2**$_{+0.2}$	**69.5**$_{+2.7}$	**91.0**$_{+1.3}$	**94.7**$_{+0.4}$	**539.7**$_{+8.1}$

performs the best grid-based retrieval model VSE∞ (J. Chen et al., 2021) on all metrics, achieving the highest rSum scores of 538.9(+7.3) and 459.0(+19.0) on MS-COCO 1K (5-folds) and full 5K test sets, respectively. By combining multi-level visual features, our 3SHNet significantly boosts the retrieval performance compared to the state-of-the-art Imp.*. For example, it improves 2.1% on Recall@1 of image-to-sentence retrieval and 3.4% on Recall@1 of sentence-to-image retrieval on the 1K test set, and 4.4% and 3.5% on the larger 5K test set, respectively. We also provide comprehensive comparisons on different batch sizes in Section 8.4.7 to fully demonstrate our superiority.

Quantitative comparison on Flickr30K.

Table 8.4 shows the quantitative results on a different dataset, the Flickr30K test set, where the proposed 3SHNet outperforms the state-of-the-art studies on three types of visual representations with the impressive gains of rSum. Specifically, when using region-based image representation to align images and sentences, our method respectively improves the state-of-the-art AME* (J. Li et al., 2022) by 2.8%, 1.5% in terms of Recall@1 on image-to-sentence and sentence-to-image retrieval directions, and by 5.0% on rSum. When using grid-based image representation, our 3SHNet still markedly exceeds other models on all metrics, where it outperforms the second-best VSE∞ (J. Chen et al., 2021) by 9.0% in terms of rSum. By combining local and global image representation, the retrieval performances are significantly improved and our 3SHNet significantly outperforms all competing methods, e.g., improving 8.1% on rSum compared with the second-best Imp.* model (D. Wu et al., 2022). These observations on the Flickr30K benchmark serve as additional evidence of the robustness and superiority of our retrieval model.

Table 8.5: Results on generalizability across datasets from MS-COCO to Flickr30k. $*$ indicates the performance of the ensemble model. Results marked with $\natural$ indicate that they are derived from the released pre-trained model in their published works.

Method	Image-to-Sentence			Sentence-to-Image			Rsum
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	
CVSE	56.4	83.0	89.0	39.9	68.6	77.2	414.1
SGR	51.4	79.2	87.2	40.5	68.6	77.7	404.6
SGRAF*	65.7	87.2	93.4	48.1	73.9	81.9	450.2
VSE∞$^\natural$	68.0	89.2	93.7	50.0	77.0	84.9	462.8
DIME$^\natural$	63.5	86.9	93.1	49.7	76.1	83.9	453.2
DIME$^{*\natural}$	67.4	90.1	94.5	53.7	79.2	86.5	471.4
ESA	69.5	89.1	93.8	51.5	77.9	85.7	467.4
3SHNet* (Region)	**72.7**	**90.9**	94.2	**54.5**	**79.5**	**86.8**	**478.6**
3SHNet (Grid)	**70.9**	**91.6**	**94.9**	**53.7**	**79.6**	**87.0**	**477.8**
3SHNet (Region+Grid)	**74.9**	**93.2**	**96.2**	**55.8**	**81.4**	**88.5**	**490.0**

8.4.3 Generalization Capability for Cross-dataset Adaptation

Generalization is one crucial and practical capability for cross-modal retrieval. Due to modality independence, 3SHNet is expected to acquire better generalization. To evaluate the generalization capability of our proposed visual semantic-spatial self-highlighting network, we create a cross-dataset transferring evaluation by pre-training methods on MS-COCO and validating them on Flickr30K test set in Table 8.5, which turns out to be more abundant than the existing experimental settings. MS-COCO (T.-Y. Lin et al., 2014) and Flickr30K (Young et al., 2014) have certain inherent differences in textual-annotation quality and inherent homogeneity of images, although they are both real-world datasets. As revealed by Guan, Liu, Ma, Qian, and Ji (2018); M. Yu and Sun (2023), Flickr30K sentence descriptions are more diverse while MS-COCO pays more attention to the consistency of image content. Therefore, training on MS-COCO with testing on Flickr30K can reveal the zero-shot generalization capability of our proposed model when we change different textual-description domains. Specifically, Table 8.5 shows that 3SHNet alternatives accomplish the best. Especially among transferring models, single-model 3SHNet outperforms SGR (Qi, Zhang, Qi, & Lu, 2021), CVSE (H. Wang et al., 2020), VSE∞ (J. Chen et al., 2021), DIME (Qu et al., 2021) and ESA (H. Zhu, Zhang, Wei, Huang, & Zhao, 2023), while ensemble-model 3SHNet* surpasses SGRAF* (Diao et al., 2021a) and DIME* significantly. These performance gains reflect not only the conspicuous generalization ability of 3SHNet but also the superiority of our visual inner multi-modal interaction.

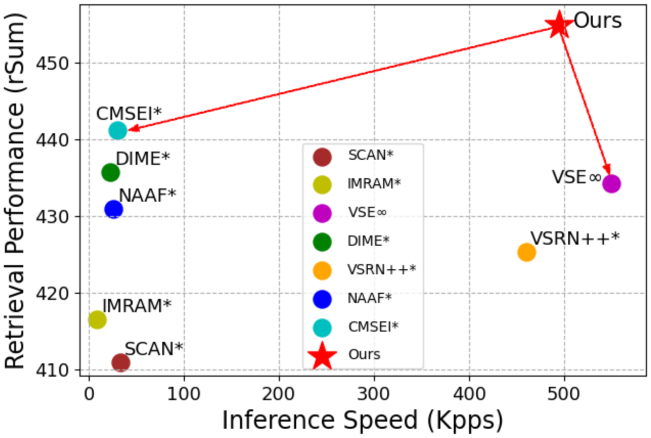


Figure 8.3: Inference speed (Kpps (H. Wang et al., 2022) means the number of image/sentence queries completed per second) and performance on MS-COCO 5K test set for image-text retrieval on single GPU (upper right is better).

8.4.4 Inference Speed

To evaluate the efficiency of 3SHNet, we report both retrieval performances and off-line inference speeds in Figure 8.3. Following the existing methods (H. Chen et al., 2020; Ge, Chen, et al., 2023; K.-H. Lee et al., 2018; Qu et al., 2021; K. Zhang et al., 2022), we employ the caching strategy to exclude the non-real-time calculation of pre-stored features since reducing the repeated interactive calculation of images with texts to improve the real-time inference speed is one of our focuses. In our 3SHNet, the feature computing in our entire visual branch and textual branch can be taken as a feature pre-extraction process that benefited from modality independence. Thus, the ultimate inference speed will be found in the practical operation. Indeed, this is also an urgent demand in practical applications, such as multi-modal recommendation systems (Y. Chen, Liu, Zhao, & Zhu, 2020; Karedla, Love, & Wherry, 1994), since the feature pre-extraction can be processed offline. Compared with the text-dependent methods, SCAN*, IMRAM*, DIME*, NAAF* and CMSEI* and modality-independent methods, VSRN++*, VSE∞, our 3SHRNet achieves comprehensive advantage on both performance and efficiency. For example, 3SHRNet is nearly 10 times faster than CMSEI* with the improvement of 15 points in performance. These superior results reflect the effective inner multi-modal interaction and modality guidance in 3SHRNet under the modality independence.

Table 8.6: Ablation studies on MS-COCO 1K (5-folds) and full-5K test sets. *R@* is the abbreviation of *Recall@*.

Method					MS-COCO 1K (5-folds)							MS-COCO 5K						
NO.	Reg.	VSpM	VSeM	Seg.	Image-to-Sentence			Sentence-to-Image			rSum	Image-to-Sentence			Sentence-to-Image			rSum
					R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
1	√	√	√	√	**83.1**	**97.2**	**99.3**	**68.7**	**92.4**	**96.6**	**537.3**	**63.8**	**88.1**	**94.0**	**47.0**	**76.6**	**85.4**	**454.9**
2	√	√	-	√	82.7	97.3	99.0	68.3	92.4	96.7	536.4	63.3	87.8	93.3	46.6	75.9	85.4	452.4
3	√	-	√	√	81.8	97.0	99.0	68.6	92.4	96.5	535.4	62.6	87.1	93.6	47.1	75.9	85.3	451.6
4	√	√	√	-	82.1	97.1	99.2	67.8	92.3	96.5	535.0	61.7	87.3	93.7	46.1	75.8	85.0	449.6
5	√	-	-	√	81.7	96.9	99.0	67.1	92.2	96.6	533.4	61.3	87.2	93.7	46.2	75.7	85.2	449.2
6	√	-	-	-	81.0	96.6	98.9	65.9	91.3	96.0	529.8	61.7	87.0	93.1	44.1	73.6	83.5	442.9
7	-	-	-	√	55.8	82.3	90.2	44.9	77.4	87.3	437.9	31.0	58.8	70.7	24.1	51.7	64.4	300.6

8.4.5 Ablation Studies

The effectiveness of semantic-spatial self-highlighting.

We conduct different feature combinations to observe the performances of different feature combinations and evaluate the superior of our semantic-spatial self-highlighting method. In Table 8.6, the comparison of No.6 and No.5 shows that segmentation features (Seg.) do contribute to the retrieval performance. But when simultaneously comparing No.7 and No.5, we find that the contribution of segmentation features is far below the one of the regular region-based local-level features. Besides, by comparing No.4 and No.5 that fuse region and segmentation features respectively by feature concatenation and by our semantic-spatial self-highlighting, we find that the visual multimodal interactive features from VSpM and VSeM are superior to simple fusion features, *e.g.* No.4 *vs.* No.5 on rSum gets 535.0 *vs.* 533.4 and 449.6 *vs.* 449.2 on MS-COCO 1K and 5K test sets respectively. These reflect that: (1) segmentation features play an indecisive role in semantic richness, and (2) our semantic-spatial self-highlighting method does promote the effective embedding of segmentation features under the multimodal interaction.

Effects of visual-semantic modelling and visual-spatial modelling.

To evaluate the impact of proposed visual-semantic modelling (VSeM) and visual-spatial modelling (VSpM) on image-sentence retrieval, we remove the visual-semantic salience embedding from VSeM and the visual-spatial embedding from VSpM, respectively. As shown in Table 8.6, the proposed approach makes absolute 1.9% and 3.3% drops in terms of rSum on MS-COCO 1K and 5K test sets when removing the visual-semantic salience embedding of visual-semantic modelling (indicated in NO. 3). And it decreases absolutely 0.9% and 2.5% on rSum when removing the visual-spatial modelling (indicated in NO. 2). These results manifest that our proposed visual semantic-spatial self-highlighting network can improve the distinguishability of image representations via the visual semantic and spatial interactions with corresponding segmentations.

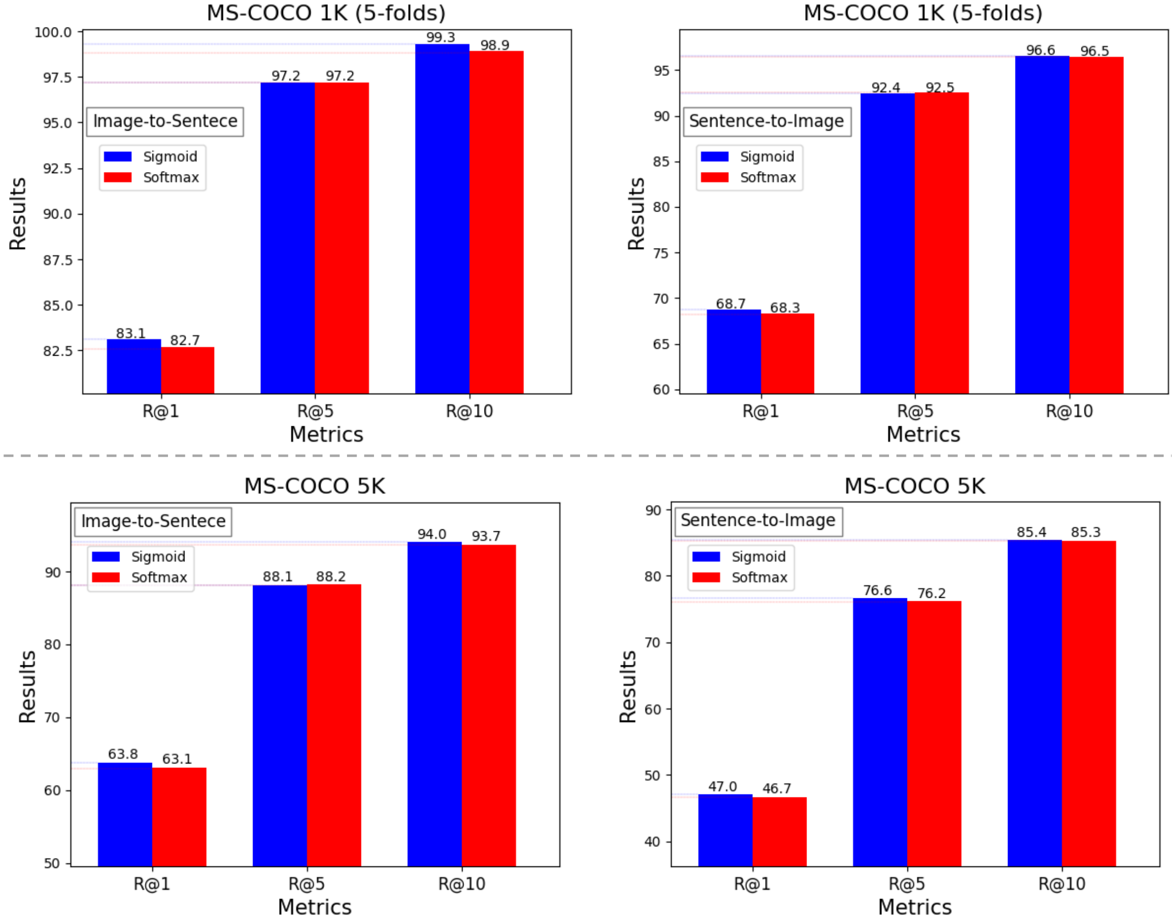


Figure 8.4: Comparisons of our proposed 3SHNet with different activate functions (Sigmoid (McCulloch & Pitts, 1943) *VS.* Softmax (Chorowski et al., 2015)) in visual-semantic modelling. *R@* is the abbreviation of *Recall@*.

Effects of different activation functions.

Different activation functions have different meanings in Eq. (2) for visual-semantic modelling (VSeM). We chose Sigmoid (McCulloch & Pitts, 1943) because there may be several equally important objects in an image, which should all be attended to and should not be given less attention due to the weight limitation of Softmax (Chorowski et al., 2015). Additionally, the goal of the cosine similarity used in Sigmoid function is to enhance the differentiation among the marginal region features for more complete saliency. Since all the region proposals, including the marginal region proposals are related to the main semantic. It's unreasonable to take the center object region proposals as the same and with high semantic relation and take the marginal region proposals as the same and with little semantic relation. Indeed, to relieve the problem that the similarity value range is [-1,1] instead of (-∞,+∞), we follow a generic operation to divide similarity by $\sqrt{D}$ before the Sigmoid operation in practice. Figure 8.4 shows experimental results using different activation functions on the MS-COCO 1K and 5K test sets. These observations demonstrate two aspects: (i) Both activation functions can motivate the effectiveness of the proposed visual-semantic modelling, which can obtain better results than the current state-of-the-art methods. (ii) As mentioned above, the Sigmoid function can activate objects with equal saliency as much as possible and enhance the differentiation among the marginal region features for more complete saliency via cosine similarity, allowing it to achieve slightly better experimental results than the Softmax function used in VSeM.

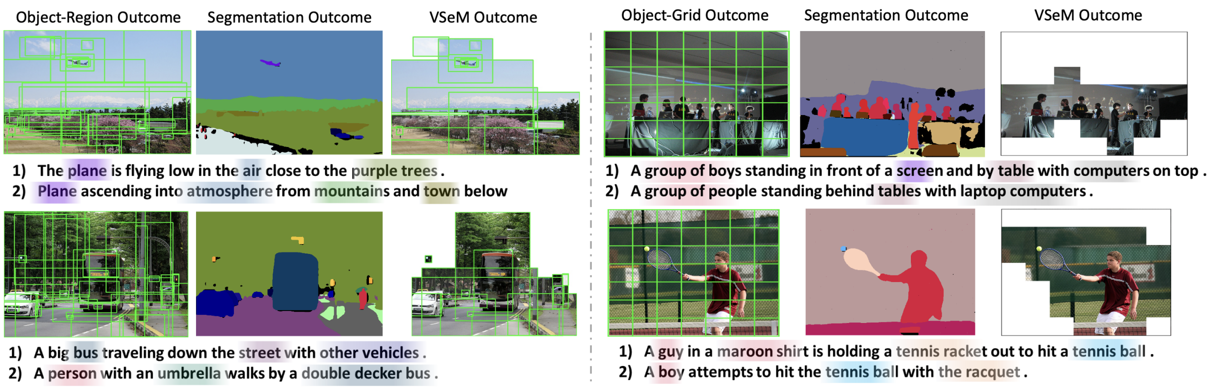

Object-Region Outcome
Segmentation Outcome
VSeM Outcome
Object-Grid Outcome
Segmentation Outcome
VSeM Outcome
1) The plane is flying low in the air close to the purple trees .
2) Plane ascending into atmosphere from mountains and town below
1) A group of boys standing in front of a screen and by table with computers on top .
2) A group of people standing behind tables with laptop computers .
1) A big bus traveling down the street with other vehicles .
2) A person with an umbrella walks by a double decker bus .
1) A guy in a maroon shirt is holding a tennis racket out to hit a tennis ball .
2) A boy attempts to hit the tennis ball with the racquet .

Figure 8.5: Salient region-level (on the left) and grid-level (on the right) object visualizations from visual-semantic multimodal modelling (VSeM) guided by segmentations on MS-COCO dataset. Each visualization contains a visual image containing the original object outcome, its segmentation outcome and the corresponding VSeM outcome, and two random matching sentences with object highlights. The greater the salience of objects, the greater the transparency (best viewed in color).

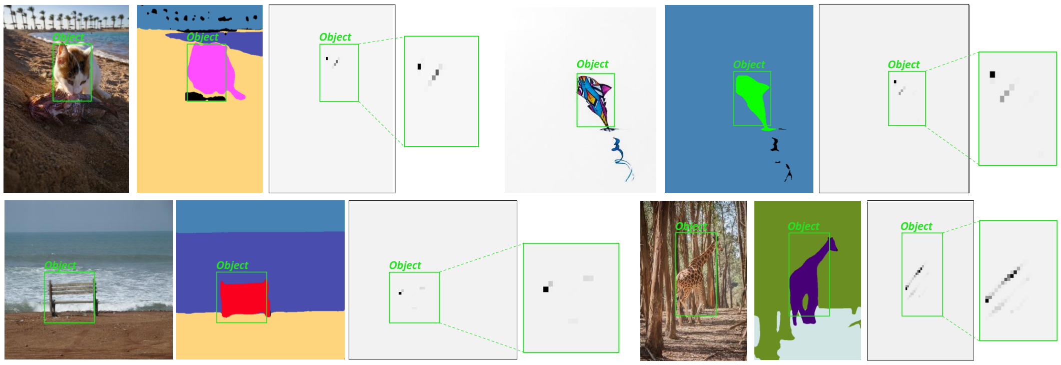
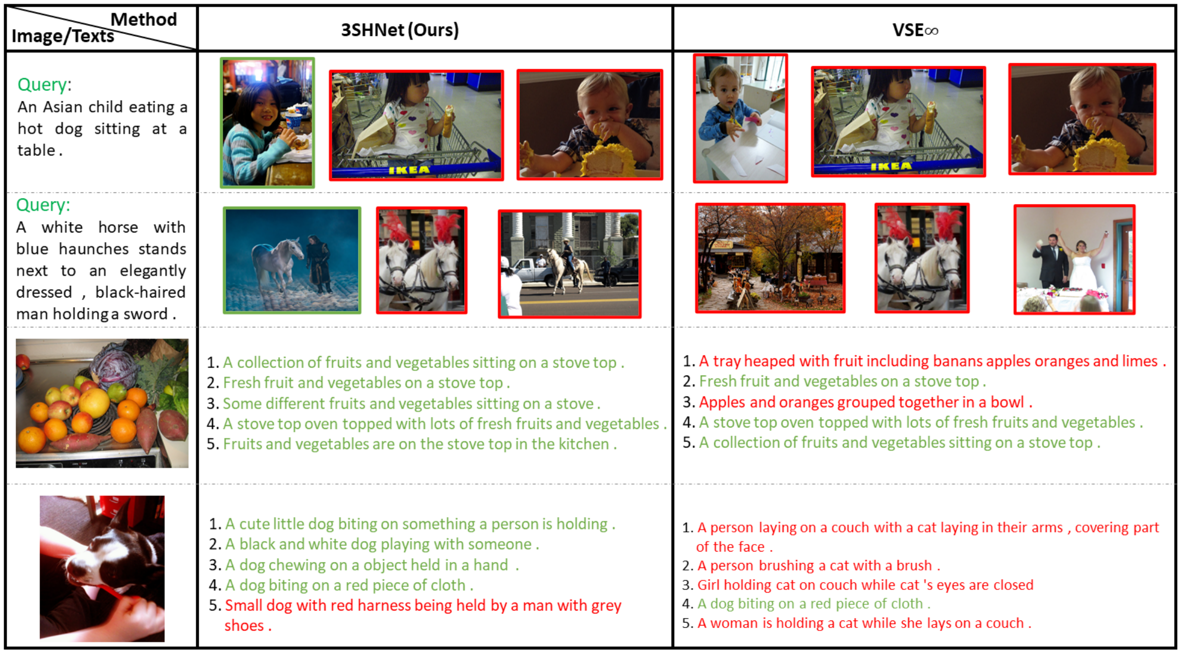


Figure 8.6: Visualization of salient spatial locations of corresponding salient objects on MS-COCO by visual-spatial multimodal modelling (VSpM). The greater the salience, the more pronounced the black colour (best viewed in color).

8.4.6 Qualitative Analysis

To further understand the contribution of visual semantic and spatial salience embedding from our proposed VSeM and VSpM, we visualize the salient regions and girds used to represent the main content of images in Figure 8.5 and visualize the spatial embedding weights of the most salient corresponding regions in Figure 8.6. It is clear from Figure 8.5 that our visual-semantic


Method
Image/Texts
3SHNet (Ours)
VSE∞
Query:
An Asian child eating a hot dog sitting at a table .
Query:
A white horse with blue haunches stands next to an elegantly dressed , black-haired man holding a sword .

3SHNet (Ours):

1. A collection of fruits and vegetables sitting on a stove top .
2. Fresh fruit and vegetables on a stove top .
3. Some different fruits and vegetables sitting on a stove .
4. A stove top oven topped with lots of fresh fruits and vegetables .
5. Fruits and vegetables are on the stove top in the kitchen .

VSE∞:

1. A tray heaped with fruit including banans apples oranges and limes .
2. Fresh fruit and vegetables on a stove top .
3. Apples and oranges grouped together in a bowl .
4. A stove top oven topped with lots of fresh fruits and vegetables .
5. A collection of fruits and vegetables sitting on a stove top .

3SHNet (Ours):

1. A cute little dog biting on something a person is holding .
2. A black and white dog playing with someone .
3. A dog chewing on a object held in a hand .
4. A dog biting on a red piece of cloth .
5. Small dog with red harness being held by a man with grey shoes .

VSE∞:

1. A person laying on a couch with a cat laying in their arms , covering part of the face .
2. A person brushing a cat with a brush .
3. Girl holding cat on couch while cat 's eyes are closed
4. A dog biting on a red piece of cloth .
5. A woman is holding a cat while she lays on a couch .

Figure 8.7: Comparisons of sentence-to-image and image-to-sentence retrieval between our 3SHNet and VSE∞ (J. Chen et al., 2021) on MS-COCO test set. For image retrieval, we showcase the foremost three ranked images, ordered in a left-to-right ranking fashion. Ranked images that align correctly are denoted in green, while any discordant matches are denoted in red. For image-to-sentence retrieval, we display the top five sentences retrieved for each image query, with any mismatches distinctly accentuated in red (best viewed in color).

modelling can concentrate on the main regions and grids containing salient objects guided by corresponding segmentations to improve image representation capabilities. In addition, we also visualize some examples in Figure 8.6 to help understand the effectiveness and interpretability of our visual-spatial modelling. Accurate visual-spatial embeddings can further enhance the ability and distinguishability of salient object representations in images.

Furthermore, more visualizations and analyses of two-modality retrieval cases are available for comprehensive comparisons in Figure 8.7. For image retrieval, we exhibit the top three ranked images corresponding to each sentence query by our proposed 3SHNet and the latest comparison VSE∞ (J. Chen et al., 2021), respectively. True matches are highlighted within green-bordered boxes, while incorrect matches are indicated by red borders. In addition, we also show the image-to-sentence retrieval results (top-3 retrieved sentences) forecasted by our 3SHNet and VSE∞ (J. Chen et al., 2021), where instances of discrepancy are highlighted in red. These observations demonstrate that our approach can obtain more accurate search results.

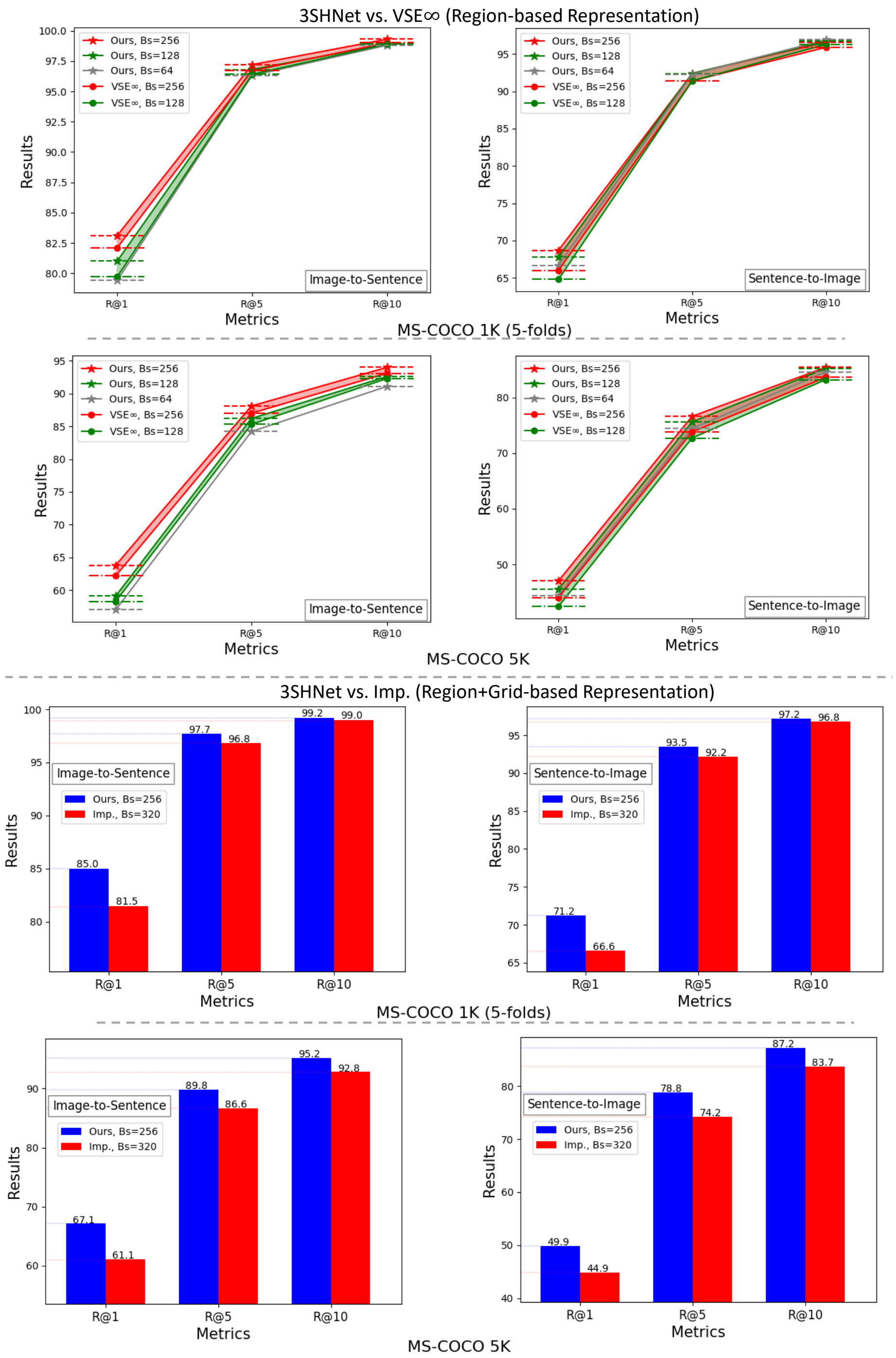


Figure 8.8: Comparisons of our proposed 3SHNet and two state-of-the-art methods (VSE∞ (J. Chen et al., 2021) and Imp. (D. Wu et al., 2022)) in different batch sizes used the same visual representations, respectively. *R@* is short for *Recall@*.

8.4.7 Discussion

Discussion on different batch sizes.

The ISR literature shows that a larger batch size may improve retrieval performance. The main reason is that during the training process, the two-way triple ranking loss of the hard negative mining strategy (Faghri et al., 2017) is used, so a larger batch can have a greater probability of containing high-quality negative samples, which better optimizes our objective function. Since our model is free from the dependence on textual guidance, we can use a relatively high throughput under the same computing facilities. As shown in Figure 8.8, we report more results from our proposed 3SHNet on MS-COCO benchmark compared with the latest methods, *i.e.,* VSE∞ (J. Chen et al., 2021) and Imp. (D. Wu et al., 2022) used the same visual representations in different batch sizes. These observations suggest that a larger mini-batch can somewhat improve the performance of cross-modal retrieval. In addition, in contrast to the state-of-the-art methodologies in the same batch size, our 3SHNet outperforms them by a large margin in Figure 8.8 on all six recall metrics. Our proposed visual semantic-spatial self-highlighting network can boost the efficacy of image-sentence retrieval.

Table 8.7: Comparisons between the proposed 3SHNet and some large-scale pre-trained visual-language methods on MS-COCO full-5K test set. z means zero-shot cross-modal retrieval results, where the model is pre-trained on the large-scale image-sentence pairs. Bs means the mini-batch size.

Method	Pretrain	GPUs	Bs	Sentence Retrieval		Image Retrieval		
				Recall@1	Recall@5	Recall@1	Recall@5	rSum
ViLBERT	3.3M	8 TitanX	512	57.5	84.0	41.8	71.5	254.8
UNITER	9.6M	16 V100	192	63.3	87.0	48.4	76.7	275.4
Unicoder	3.8M	4 V100	-	62.3	87.1	46.7	76.0	272.1
OSCAR	6.5M	16 V100	1,024	70.0	91.1	54.0	80.8	295.9
CLIP$_z$	400M	592 V100	32,768	58.4	81.5	37.8	62.4	240.1
ALIGN$_z$	1.8B	1024 TPUv3	16384	58.6	83.0	45.6	69.8	257.0
ALIGN	1.8B	1024 TPUv3	16384	77.0	93.5	59.9	83.3	313.7
ALBEF	4.0M	8 A100	512	73.1	91.4	56.8	81.5	302.8
3SHNet	N/A	1 TitanX	256	67.9	90.5	50.3	79.3	288.0

Discussion on large-scale pre-trained models.

Pre-trained visual language representations on large-scale datasets are becoming increasingly popular, especially in companies with large-scale parallel computing power. However, due to the limitation of computation facility requirements, it is difficult to carry out large-scale pre-training in universities or research institutions. For example, the pre-training of UNITER-base and UNITER-large in (G. Li, Duan, Fang, Gong, & Jiang, 2020) involved the utilization of 882 and 3645 V100 GPU hours, respectively. Additionally, most of the excellent large-scale pre-training methods (Y.-C. Chen et al., 2020; C. Jia et al., 2021; Radford et al., 2021) also depend on extensive cross-modal interactions within vast collections of image-text pairs. The text-dependent visual representation learning approach leads to a long inference retrieval time, which is hardly applied to real-life scenarios. In this section, we report the results of our proposed approach compared to some popular methods, such as ViLBERT (B. Zhang, Hu, Jain, Ie, &

Sha, 2020), UNITER (Y.-C. Chen et al., 2020), Unicoder (G. Li et al., 2020), OSCAR (X. Li et al., 2020), CLIP (Radford et al., 2021), ALIGN (C. Jia et al., 2021) and ALBEF (J. Li et al., 2021) that pre-trained on large-scale datasets. As shown in Table 8.7, compared to the large-scale visual-language pre-trained methods, our approach achieves competitive results at a smaller computation facility requirement without large-scale visual-language pre-training. For example, the performances of our 3SHNet are better than UNITER (Y.-C. Chen et al., 2020), requiring 16 V100 GPUs in six evaluation metrics. In addition, as mentioned in J. Chen et al. (2021), VSE (J. Chen et al., 2021; Faghri et al., 2017) methods (the same framework as our 3SHNet) exhibit significantly enhanced speed in large-scale multi-modal retrieval due to the expeditious pre-computed computation or indexing of holistic embeddings (Johnson, Douze, & Jégou, 2019).

Although our approach is relatively lower than ALIGN (C. Jia et al., 2021), which has stronger computation facilities and a larger number of image-text pairs at larger batch sizes, we propose a scheme that can apply our proposed method on large-scale datasets, which we will explore further in future work. Specifically, recently, the introduction of Kirillov et al. (2023) makes it easy and fast to obtain semantic segmentation results in arbitrary scenarios, which will facilitate the application of our proposed VSeM and VSpM to large-scale datasets.

8.5 Theoretical and Practical Implications

We propose 3SHNet for image-sentence retrieval that introduces a novel segmentation-based visual semantic-spatial self-highlighting schema into an end-to-end modality-independence modelling cross-modal alignment framework. The mainly visual-semantic and visual-spatial multimodal interactions mine the semantic saliency and spatial saliency of visual objects respectively, thereby improving the discriminability of image representations during the cross-modality alignment process. Guidance information from semantic segmentation overcomes the lack of textual dependence and maintains modality independence, thereby ensuring retrieval efficiency.

For theoretical implications, the proposed 3SHNet overcomes the obvious shortcomings of the two existing mainstream methods, i.e., low efficiency and low generalization due to the deep textual dependence in textual-guidance-based visual representation learning methods (H. Chen et al., 2020; Ge, Chen, et al., 2023; Pan et al., 2023; Qu et al., 2021) and the ignoring human-like attention on the prominent objects and their locations in the visual hybrid-level representation enhancing methods (J. Chen et al., 2021; Ge, Chen, et al., 2021; H. Liu et al., 2018; Y. Ma et al., 2023). In particular, our 3SHNet introduces the segmentation information to highlight the semantic saliency and spatial saliency of objects within the visual modality, which can replace the complex visual-textual interaction operations to keep the retrieval high-efficiency and improve the salience of prominent objects and their locations in the visual hybrid-level representations to improve the retrieval performance.

For practical implications, our 3SHNet aims to construct a high-precision, high-efficiency, and high-generalisation image-sentence retrieval model. This ensures retrieval efficiency while ensuring retrieval performance, thus providing the possibility for practical applications, such as multi-modal retrieval within search engines (R. He, Xiong, Yang, & Park, 2011). It can be applied to both large websites and private systems, such as library multimedia systems (M.-H. Lee et al., 2003). Furthermore, 3SHNet does not rely on large-scale computing resources, thus ensuring its portability to new private data.

8.6 Conclusion

In this chapter, we enhance visual representation under the modality-independent pattern for high-precision, high-efficiency, and high-generalization image-sentence retrieval, where the visual semantic salience and spatial locations are highlighted based on visual segmentations. Specially, the visual-semantic and visual-spatial multimodal interactions are designed in our 3SHNet based on a two-tower modality-independent framework, which involves the hybrid-level visual representations, *i.e.,* local-level region-base feature and global-level grid-based feature. The superiority of our 3SHNet is evident in extensive quantitative comparisons, showcasing its state-of-the-art performance and efficiency across popular benchmarks such as MS-COCO and Flickr30K under various evaluation metrics.

Chapter 9

Conclusion

In this thesis, we argue that novel modality relational reasoning and embedding approaches can effectively improve context-aware image semantic representation and thus boost the performance of related tasks. We explore this on different tasks with different modalities, including unimodal facial action unit (FAU) recognition and multimodal image-sentence retrieval, thereby fully verifying the importance of modality-based relational reasoning and encoding for learning and enhancing image context representations.

On the one hand, we explore the contributions of modality relational reasoning and embedding of context-aware semantic image representation learning for the unimodal facial action unit recognition task. We developed a series of modality relational reasoning and embedding approaches for facial image representation learning, which can enhance the contextual connections of facial representations, such as the natural linkages between local muscle regions. Specifically, we want to model the latent relationships among local regions of face images, enabling better correlation between multiple regional lesion muscles and texture changes. This motivates the proposed ALGRNet (Ge, Jose, et al., 2023; Ge, Wan, et al., 2021) in Chapter 3 and MGRR-Net (Ge, Jose, et al., 2024) in Chapter 4. Firstly, ALGRNet capitalizes on the precision and adaptability of muscle region localization and leverages the comprehensive facial semantic feature representation offered by AU detection models. By harnessing the interactive relationships and interplay between adaptive and symmetrical muscle regions, ALGRNet effectively captures the dynamic nature of these regions across various expressions and individual characteristics. In particular, ALGRNet employs a novel relational reasoning and embedding mechanism (called *skip-BiLSTM*) to facilitate efficient information exchange, allowing for seamless transfer of local muscle features while modelling the potential assistance and exclusion relationships among AU branches. Secondly, a novel multi-level graph relational reasoning network (termed MGRR-Net) is proposed to explore the region-level facial image representation learning and local-global face feature interactions. In particular, each layer of MGRR-Net can encode the dynamic relationships among AUs via a region-level relationship graph and multiple complementary levels of global information covering expression and subject diversities. The multi-layer iterative feature refinement finally obtains robust and discriminative features for each AU. Finally, different from the implicit relational reasoning and embedding methods of ALGRNet and MGRR-Net, we further use an explicit language relation model, termed VL-FAU (Ge, Fu, et al., 2024) in Chapter 5, to guide the relational connections of different visual AU features, thereby improving the contextual representation ability of images. We conduct extensive experiments on two widely used benchmarks, i.e., BP4D and DISFA, to evaluate the proposed implicit and explicit relational reasoning and embedding methods, i.e., ALGRNet, MGRR-Net, and VL-FAU model. Competitive experimental performances demonstrate the effectiveness of our approaches, demonstrating the contribution of relational reasoning and embedding to context-aware image representation learning for FAU recognition.

Limitations. While the proposed approaches —ALGRNet, MGRR-Net, and VL-FAU— demonstrate significant advancements in context-aware image semantic representation for facial action unit (FAU) recognition, there are notable limitations to our work. One major limitation is the lack of consideration for the sequential information present in video data. Although our methods focus on enhancing the contextual connections among static facial images, they do not explicitly account for the temporal dynamics that characterize facial expressions over time. This omission may lead to insufficient modeling of the contextual information that evolves across frames in a video sequence. Facial expressions are inherently dynamic, with muscle movements and expressions transitioning smoothly over time, and our methods primarily operate on individual frames without leveraging the rich sequential information available in video data. As a result, this could limit the models' ability to capture important temporal relationships and patterns that are crucial for a comprehensive understanding of facial actions.

Future Work. In our future work, we will build a powerful facial visual representation backbone based on modality relational reasoning and embedding approach with a language model for supervision. We envision that the backbone model has language interpretability and can use relational reasoning and embedding to improve the representation and distinguishability of face features. Exp_Blip (Yuan, Zeng, & Shan, 2023) employed a BLIP-2-based architecture to pre-train a vision-to-language model, where text descriptions of faces are generated by GPT-3.5 with rule-based captions from FACS (Ekman & Rosenberg, 1997). Inspired by Exp_Blip (Yuan et al., 2023) and our new VL-FAU (Ge, Fu, et al., 2024), we will construct a new multimodal face benchmark and new face fundamental representation backbone. The new benchmark will contain the annotations of AU states, the corresponding language descriptions and even emotion annotations. The new face fundamental representation backbone will introduce the modality relational reasoning and embedding approaches to improve the representation ability of facial images. In this way, it can provide powerful support for a wide range of face image tasks, such as FAU recognition, emotion recognition, face alignment, etc. In addition, future work should explore integrating temporal modeling techniques to better utilize video sequences, allowing for more robust contextual representations that reflect both spatial and temporal dimensions of facial expressions. This enhancement could potentially improve the performance of FAU recognition tasks and contribute to a more holistic understanding of facial dynamics.

On the other hand, we explore the contributions of modality relational reasoning and embedding of context-aware image semantic representation learning for the multimodal image-sentence retrieval task. Different from unimodal vision tasks, accurate visual representation of contextual semantics is more challenging in multimodal tasks such as image-text retrieval. This is because image semantic representation is richer and more ambiguous than text with explicit semantics. To improve the context-aware image representation ability in multimodal tasks, in this thesis, we proposed a series of novel modality relational reasoning and embedding approaches for multimodal image-sentence retrieval, including the tree-based SMFEA (Ge, Chen, et al., 2021) in Chapter 6, the GCN-based CMSEI (Ge, Chen, et al., 2023) and *Hire* (Ge, Chen, et al., 2024) in Chapter 7 and a novel visual self-highlight model 3SHNet (Ge, Xu, et al., 2024) in Chapter 8. We first explore the role of different structured relational reasoning and embedding structures in promoting visual feature learning. Specifically, SMFEA in Chapter 6 creates a novel multi-modal structured module with a shared context-aware referral tree, improving the semantic and structural consistency between images and sentences. In this work, the semantic and structural relationship expressions of image representation are constrained by explicit textual representation, thus reducing the ambiguity of image semantic features. Moreover, we propose a hybrid-modal interaction with multiple relational enhancements for image-sentence retrieval, termed *Hire* in Chapter 7, mainly leveraging novel relation-aware graph convolutional networks (GCNs) to reason about the relationships of important objects in images and embed them with relational connections. In addition, in Chapter 8, we introduce another visual modality (segmentation information) that interacts with visual appearance features further to improve visual representation capabilities through cross-modal semantic alignment optimization, highlighting the salient recognition of prominent objects in the visual modality and their spatial locations. In the multimodal task (image-text retrieval), under the optimization of multimodal semantic alignment, the effectiveness of modality relational reasoning and embedding for contextual semantic representation, especially images, is further verified.

Limitations. While our research on modality relational reasoning and embedding approaches for context-aware image semantic representation learning in multimodal image-sentence retrieval tasks shows promise, there are notable limitations. A significant limitation is that we have not fully leveraged the capabilities of Multimodal Large Language Models (MLLMs). MLLMs can handle both image and text data, providing the potential for deeper semantic understanding and reasoning in multimodal contexts. By incorporating MLLMs into our framework for joint learning, we could further enhance the alignment and understanding between modalities, improving the interrelationships between images and text.

Future Work. Pre-trained visual language representations on large-scale datasets (J. Li et al., 2021; Radford et al., 2021; B. Zhang et al., 2020) are becoming increasingly popular, especially in companies with large-scale parallel computing power. They simply and crudely used data-driven to mine the semantic alignment between images and texts, thereby improving the semantic expression of images. For instance, CLIP (Contrastive Language-Image Pre-training) model (Radford et al., 2021) leveraged a large dataset of images and their corresponding captions to train separate neural networks for images and text, learning to map both modalities into a shared embedding space where matching image-caption pairs are closer together and non-matching pairs are farther apart, thus enabling powerful zero-shot learning capabilities. However, there are still flaws in this. They fail to model the relationships between objects within an image, which limits their ability to understand and reason about complex scenes and the interactions between different entities. And training and deploying these large-scale models, such as CLIP (Radford et al., 2021), require substantial computational resources, making it less accessible for smaller organizations. In our future work, we will explore efficient cross-modal retrieval methods with modality relational reasoning and embedding based on the pre-trained large-scale visual-language models. In fact, we have tried some explorations to improve the efficiency of feature representation (Fu et al., 2024) and cross-modal retrieval (Z. Long, Ge, McCreadie, & Jose, 2024). We will further explore the effectiveness of relational reasoning and embedding in the future.

Overall, we comprehensively validate the facilitation of modality relational reasoning and embedding for representation learning, especially to enhance the feature representation of complex images, in multiple tasks, including unimodal FAU recognition and multimodal image-sentence retrieval. Therefore, context-aware image semantic representation via modality relational reasoning and embedding is a promising and meaningful research direction.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Ieee conference on computer vision and pattern recognition* (pp. 6077–6086).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460.

Barrios Dell'Olio, G., & Sra, M. (2021). Farapy: An augmented reality feedback system for facial paralysis using action unit intensity estimation. In *The 34th annual acm symposium on user interface software and technology* (pp. 1027–1038).

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2015). Salient object detection: A benchmark. *IEEE Trans. Image Process.*, *24*(12), 5706–5722.

Butt, M. N., & Iqbal, M. (2011). Teachers' perception regarding facial expressions as an effective teaching tool. *Contemporary Issues in Education Research*, *4*(2), 11–14.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *Ieee international conference on automatic face & gesture recognition* (pp. 67–74).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229).

Chang, Y., & Wang, S. (2022). Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Ieee conference on computer vision and pattern recognition* (pp. 20417–20426).

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 740–750).

Chen, F., Ji, R., Ji, J., Sun, X., Zhang, B., Ge, X., . . . Wang, Y. (2019). Variational structured semantic inference for diverse image captioning. *Advances in Neural Information*

Processing Systems, *32*.

Chen, F., Ji, R., Su, J., Wu, Y., & Wu, Y. (2017). Structcap: Structured semantic embedding for image captioning. In *Acm multimedia* (pp. 46–54).

Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., & Han, J. (2020). Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Ieee conference on computer vision and pattern recognition* (pp. 12655–12663).

Chen, J., Hu, H., Wu, H., Jiang, Y., & Wang, C. (2021). Learning the best pooling strategy for visual semantic embedding. In *Ieee conference on computer vision and pattern recognition* (pp. 15789–15798).

Chen, T., & Luo, J. (2020). Expressing objects just like words: Recurrent visual embedding for image-text matching. *arXiv preprint arXiv:2002.08510*.

Chen, Y., Chen, D., Wang, T., Wang, Y., & Liang, Y. (2022). Causal intervention for subject-deconfounded facial action unit recognition. In *Aaai conference on artificial intelligence* (Vol. 36, pp. 374–382).

Chen, Y., Chen, D., Wang, Y., Wang, T., & Liang, Y. (2021). Cafgraph: Context-aware facial multi-graph representation for facial action unit recognition. In *Proceedings of the 29th acm international conference on multimedia* (pp. 1029–1037).

Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018). Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3339–3348).

Chen, Y., Liu, Y., Zhao, J., & Zhu, Q. (2020). Mobile edge cache strategy based on neural collaborative filtering. *IEEE Access*, *8*, 18475–18482.

Chen, Y., Song, G., Shao, Z., Cai, J., Cham, T.-J., & Zheng, J. (2022). Geoconv: Geodesic guided convolution for facial action unit recognition. *Pattern Recognit.*, *122*, 108355.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., … Liu, J. (2020). Uniter: Universal image-text representation learning. In *European conference on computer vision* (pp. 104–120).

Chen, Z., Wu, Z., Lin, Z., Wang, S., Plant, C., & Guo, W. (2023). Agnn: Alternating graph-regularized neural networks to alleviate over-smoothing. *IEEE Transactions on Neural Networks and Learning Systems*.

Chen, Z.-M., Wei, X.-S., Wang, P., & Guo, Y. (2019). Multi-label image recognition with graph convolutional networks. In *Ieee conference on computer vision and pattern recognition* (pp. 5177–5186).

Cheng, Y., Zhu, X., Qian, J., Wen, F., & Liu, P. (2022). Cross-modal graph matching network for image-text retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, *18*(4), 1–23.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in Neural Information Processing Systems*, *28*.

Corneanu, C., Madadi, M., & Escalera, S. (2018). Deep structure inference network for facial action unit recognition. In *European conference on computer vision* (pp. 298–313).

Cui, Z., Kuang, C., Gao, T., Talamadupula, K., & Ji, Q. (2023). Biomechanics-guided facial action unit detection through force modeling. In *Ieee conference on computer vision and pattern recognition* (pp. 8694–8703).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Ieee conference on computer vision and pattern recognition* (pp. 248–255).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Diao, H., Zhang, Y., Ma, L., & Lu, H. (2021a). Similarity reasoning and filtration for image-text matching. In *Aaai conference on artificial intelligence* (Vol. 35, pp. 1218–1226).

Diao, H., Zhang, Y., Ma, L., & Lu, H. (2021b). Similarity reasoning and filtration for image-text matching. In *Aaai conference on artificial intelligence* (Vol. 35, pp. 1218–1226).

Dong, J., Ma, L., Li, Q., Wang, S., Liu, L.-a., Lin, Y., & Jian, M. (2008). An approach for quantitative evaluation of the degree of facial paralysis based on salient point detection. In *International symposium on intelligent information technology application workshops* (pp. 483–486).

Dong, X., Zhang, H., Zhu, L., Nie, L., & Liu, L. (2022). Hierarchical feature aggregation based on transformer for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(9), 6437–6447.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 3rd international conference on learning representations.*

Ekman, P., & Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (facs)*. Oxford University Press, USA.

Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., … others (2015). From captions to visual concepts and back. In *Ieee conference on computer vision and pattern recognition* (pp. 1473–1482).

Feldman, R. S., Jenkins, L., & Popoola, O. (1979). Detection of deception in adults and children via facial expressions. *Child Development*, 350–355.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems* (pp. 2121–2129).

Fu, J., Ge, X., Xin, X., Karatzoglou, A., Arapakis, I., Wang, J., & Jose, J. M. (2024). Iisan: Efficiently adapting multimodal representation for sequential recommendation with decoupled peft. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval* (pp. 687–697).

Gaber, A., Taher, M. F., Abdel Wahed, M., Shalaby, N. M., & Gaber, S. (2022). Comprehensive assessment of facial paralysis based on facial animation units. *Plos One*, *17*(12), e0277297.

Ge, X., Chen, F., Jose, J. M., Ji, Z., Wu, Z., & Liu, X. (2021). Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Acm international conference on multimedia* (pp. 5185–5193).

Ge, X., Chen, F., Shen, C., & Ji, R. (2019). Colloquial image captioning. In *Ieee international conference on multimedia and expo* (pp. 356–361).

Ge, X., Chen, F., Xu, S., Tao, F., & Jose, J. M. (2023). Cross-modal semantic enhanced interaction for image-sentence retrieval. In *Proceedings of the ieee winter conference on applications of computer vision* (pp. 1022–1031).

Ge, X., Chen, F., Xu, S., Tao, F., Wang, J., & Jose, J. M. (2024). Hire: Hybrid-modal interaction with multiple relational enhancements for image-text matching. *arXiv preprint arXiv:2406.18579*.

Ge, X., Fu, J., Chen, F., An, S., Sebe, N., & Jose, J. M. (2024). Towards end-to-end explainable facial action unit recognition via vision-language joint learning. In *Acm multimedia.*

Ge, X., Jose, J. M., Wang, P., Iyer, A., Liu, X., & Han, H. (2023). Algrnet: Multi-relational adaptive facial action unit modelling for face representation and relevant recognitions. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Ge, X., Jose, J. M., Xu, S., Liu, X., & Han, H. (2024). Mgrr-net: Multi-level graph relational reasoning network for facial action unit detection. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–20.

Ge, X., Wan, P., Han, H., Jose, J. M., Ji, Z., Wu, Z., & Liu, X. (2021). Local global relational network for facial action units recognition. In *Ieee international conference on automatic face and gesture recognition* (pp. 01–08).

Ge, X., Xu, S., Chen, F., Wang, J., Wang, G., An, S., & Jose, J. M. (2024). 3shnet: Boosting image–sentence retrieval via visual semantic–spatial self-highlighting. *Information Processing & Management*, *61*(4), 103716.

Goncalves, L., & Busso, C. (2022). Auxformer: Robust approach to audiovisual emotion recognition. In *Ieee international conference on acoustics, speech and signal processing* (pp. 7357–7361).

Gong, Y., Liu, A. H., Rouditchenko, A., & Glass, J. (2022). Uavm: Towards unifying audio and visual models. *IEEE Signal Processing Letters*, *29*, 2437–2441.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, *18*(5-6), 602–610.

Guan, Z., Liu, K., Ma, Y., Qian, X., & Ji, T. (2018). Sequential dual attention: coarse-to-fine-grained hierarchical generation for image captioning. *Symmetry*, *10*(11), 626.

Guo, J., Wang, M., Zhou, Y., Song, B., Chi, Y., Fan, W., & Chang, J. (2023). Hgan: Hierarchical graph alignment network for image-text retrieval. *IEEE Trans. Multimedia*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Ieee conference on computer vision and pattern recognition* (pp. 770–778).

He, R., Xiong, N., Yang, L. T., & Park, J. H. (2011). Using multi-modal semantic association rules to fuse keywords and visual features automatically for web image retrieval. *Information Fusion*, *12*(3), 223–230.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hossain, S. M., Jamal, Z., Noshin, A. A., & Khan, M. M. (2022). Comparative study of deep learning algorithms for the detection of facial paralysis. In *2022 ieee 13th annual information technology, electronics and mobile communication conference (iemcon)* (pp. 0368–0377).

House, W. (1985). Facial nerve grading system. *Otolaryngol Head Neck Surg*, *93*, 184–193.

Hsu, G.-S. J., Huang, W.-F., & Kang, J.-H. (2018). Hierarchical network for facial palsy detection. In *Ieee conference on computer vision and pattern recognition workshops* (pp. 580–586).

Hsu, G.-S. J., Kang, J.-H., & Huang, W.-F. (2018). Deep hierarchical network with line segment learning for quantitative analysis of facial palsy. *IEEE Access*, *7*, 4833–4842.

Hu, J., Chen, C., Cao, L., Zhang, S., Shu, A., Jiang, G., & Ji, R. (2023). Pseudo-label alignment for semi-supervised instance segmentation. In *Ieee international conference on computer vision* (pp. 16337–16347).

Hu, J., Huang, L., Ren, T., Zhang, S., Ji, R., & Cao, L. (2023). You only segment once: Towards real-time panoptic segmentation. In *Ieee conference on computer vision and pattern recognition* (pp. 17819–17829).

Huang, Y., Wu, Q., Song, C., & Wang, L. (2018). Learning semantic concepts and order for image and sentence matching. In *Ieee conference on computer vision and pattern recognition* (pp. 6163–6171).

Hwang, S. J., Ravi, S. N., Tao, Z., Kim, H. J., Collins, M. D., & Singh, V. (2018). Tensorize, factorize and regularize: Robust visual relationship learning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1014–1023).

Jacob, G. M., & Stenger, B. (2021). Facial action unit detection with transformers. In *Ieee conference on computer vision and pattern recognition* (pp. 7680–7689).

Jaiswal, S., & Valstar, M. (2016). Deep learning the dynamic appearance and shape of facial action units. In *Ieee winter conference on applications of computer vision* (pp. 1–8).

Ji, Z., Wang, H., Han, J., & Pang, Y. (2019). Saliency-guided attention network for image-sentence matching. In *Ieee international conference on computer vision* (pp. 5754–5763).

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., … Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning* (pp. 4904–4916).

Jia, X., Xu, S., Zhou, Y., Wang, L., & Li, W. (2023). A novel dual-channel graph convolutional neural network for facial action unit recognition. *Pattern Recogn. Lett.*, *166*, 61–68.

Jia, X., Zhou, Y., Li, W., Li, J., & Yin, B. (2022). Data-aware relation learning-based graph convolution neural network for facial action unit recognition. *Pattern Recognit. Lett.*, *155*, 100–106.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, *7*(3), 535–547.

Ju, X., Zhang, D., Xiao, R., Li, J., Li, S., Zhang, M., & Zhou, G. (2021). Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Conference on empirical methods in natural language processing* (pp. 4395–4405).

Karedla, R., Love, J. S., & Wherry, B. G. (1994). Caching strategies to improve disk system performance. *Computer*, *27*(3), 38–46.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Ieee conference on computer vision and pattern recognition* (pp. 3128–3137).

Karpathy, A., Joulin, A., & Fei-Fei, L. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems* (pp. 1889–1897).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., … others (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2015). Unifying visual-semantic embeddings with multimodal neural language models. *Trans. Assoc. Comput. Linguist.*

Klein, D., & Manning, C. D. (2003). A* parsing: Fast exact viterbi parse selection. In *Naacl* (pp. 119–126).

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., … others (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32–73.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

LeCun, Y., & Cortes, C. (n.d.). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Lee, K.-H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In *European conference on computer vision* (pp. 201–216).

Lee, M.-H., Kang, J.-H., Myaeng, S. H., Hyun, S. J., Yoo, J.-M., Ko, E.-J., ... Lim, J.-H. (2003). A multimedia digital library system based on mpeg-7 and xquery. In *Icadl* (pp. 193–205).

Li, G., Duan, N., Fang, Y., Gong, M., & Jiang, D. (2020). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Aaai conference on artificial intelligence* (Vol. 34, pp. 11336–11344).

Li, G., Zhu, X., Zeng, Y., Wang, Q., & Lin, L. (2019). Semantic relationships guided representation learning for facial action unit recognition. In *Aaai conference on artificial intelligence* (Vol. 33, pp. 8594–8601).

Li, J., Niu, L., & Zhang, L. (2022). Action-aware embedding enhancement for image-text retrieval. In *Aaai conference on artificial intelligence* (Vol. 36, pp. 1323–1331).

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, *34*, 9694–9705.

Li, K., Zhang, Y., Li, K., Li, Y., & Fu, Y. (2019). Visual semantic reasoning for image-text matching. In *Ieee international conference on computer vision* (pp. 4654–4662).

Li, K., Zhang, Y., Li, K., Li, Y., & Fu, Y. (2022). Image-text embedding learning via visual and textual semantic reasoning. *IEEE Trans. Pattern Anal. Mach. Intell.*, *45*(1), 641–656.

Li, W., Abtahi, F., & Zhu, Z. (2017). Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Ieee conference on computer vision and pattern recognition* (pp. 1841–1850).

Li, W., Abtahi, F., Zhu, Z., & Yin, L. (2018). Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(11), 2583–2596.

Li, W., Ma, Z., Deng, L.-J., Wang, P., Shi, J., & Fan, X. (2023). Reservoir computing transformer for image-text retrieval. In *Acm multimedia* (pp. 5605–5613).

Li, W., Su, X., Song, D., Wang, L., Zhang, K., & Liu, A.-A. (2023). Towards deconfounded image-text matching with causal inference. In *Acm multimedia* (pp. 6264–6273).

Li, W.-H., Yang, S., Wang, Y., Song, D., & Li, X.-Y. (2021). Multi-level similarity learning for image-text retrieval. *Inf. Process. & Manag.*, *58*(1), 102432.

Li, X., Komulainen, J., Zhao, G., Yuen, P.-C., & Pietikäinen, M. (2016). Generalized face antispoofing by detecting pulse from face videos. In *Ieee international conference on pattern recognition* (pp. 4244–4249).

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... others (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision* (pp. 121–137).

Li, X., Zhang, X., Wang, T., & Yin, L. (2023). Knowledge-spreader: Learning semi-supervised facial action dynamics by consistifying knowledge granularity. In *Ieee international conference on computer vision* (pp. 20979–20989).

Li, X., Zhang, Z., Zhang, X., Wang, T., Li, Z., Yang, H., ... Yin, L. (2023). Disagreement matters: Exploring internal diversification for redundant attention in generic facial action analysis. *IEEE Transactions on Affective Computing*.

Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 280–287).

Li, Y., Huang, X., & Zhao, G. (2021). Micro-expression action unit detection with spatial and channel attention. *Neurocomputing*, *436*, 221–231.

Li, Y., & Shan, S. (2023). Contrastive learning of person-independent representations for facial action unit detection. *IEEE Trans. Image Process.*.

Li, Y., Wang, S., Zhao, Y., & Ji, Q. (2013). Simultaneous facial feature tracking and facial expression recognition. *IEEE Transactions on Image Processing*, *22*(7), 2559–2573.

Li, Z., Guo, C., Feng, Z., Hwang, J.-N., & Xue, X. (2022). Multi-view visual semantic embedding. In *International joint conference on artificial intelligence* (Vol. 2, p. 7).

Liang, X., Shen, X., Feng, J., Lin, L., & Yan, S. (2016). Semantic object parsing with graph lstm. In *Computer vision–eccv 2016: 14th european conference, amsterdam, the netherlands, october 11–14, 2016, proceedings, part i 14* (pp. 125–143).

Lin, M., Chen, Q., & Yan, S. (2014). Network in network. *International Conference on Learning Representations*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Liu, C., Mao, Z., Liu, A.-A., Zhang, T., Wang, B., & Zhang, Y. (2019). Focus your attention: A bidirectional focal attention network for image-text matching. In *Acm multimedia* (pp. 3–11).

Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., & Zhang, Y. (2020). Graph structured network for image-text matching. In *Ieee conference on computer vision and pattern recognition* (pp. 10921–10930).

Liu, H., Lin, M., Zhang, S., Wu, Y., Huang, F., & Ji, R. (2018). Dense auto-encoder hashing for robust cross-modality retrieval. In *Acm multimedia* (pp. 1589–1597).

Liu, P., Han, S., Meng, Z., & Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In *Ieee conference on computer vision and pattern recognition* (pp. 1805–1812).

Liu, P., Zhang, Z., Yang, H., & Yin, L. (2019). Multi-modality empowered network for facial action unit detection. In *Ieee winter conference on applications of computer vision* (pp. 2175–2184).

Liu, P., Zhou, J. T., Tsang, I. W.-H., Meng, Z., Han, S., & Tong, Y. (2014). Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In *European conference on computer vision* (pp. 151–166).

Liu, X., He, Y., Cheung, Y.-M., Xu, X., & Wang, N. (2022). Learning relationship-enhanced semantic graph for fine-grained image–text matching. *IEEE transactions on cybernetics*, *54*(2), 948–961.

Liu, X., Xia, Y., Yu, H., Dong, J., Jian, M., & Pham, T. D. (2020). Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(10), 2325–2332.

Liu, Z., Dong, J., Zhang, C., Wang, L., & Dang, J. (2020). Relation modeling with graph convolutional networks for facial action unit detection. In *Mmm* (pp. 489–501).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 10012–10022).

Long, S., Han, S. C., Wan, X., & Poon, J. (2022). Gradual: Graph-based dual-modal representation for image-text matching. In *Ieee winter conference on applications of computer vision* (pp. 3459–3468).

Long, Z., Ge, X., McCreadie, R., & Jose, J. M. (2024). Cfir: Fast and effective long-text to image retrieval for large corpora. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval* (pp. 2188–2198).

Loper, E., & Bird, S. (2002). Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Lu, E., & Hu, X. (2022). Image super-resolution via channel attention and spatial attention. *Applied Intelligence*, *52*(2), 2260–2268.

Luo, C., Song, S., Xie, W., Shen, L., & Gunes, H. (2022). Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *International joint conference on artificial intelligence* (pp. 1239–1246).

Ma, C., Chen, L., & Yong, J. (2019). Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. *Neurocomputing*, *355*, 35–47.

Ma, Y., Sun, X., Ji, J., Jiang, G., Zhuang, W., & Ji, R. (2023). Beat: Bi-directional one-to-many embedding alignment for text-based person retrieval. In *Acm multimedia* (pp. 4157–4168).

Madani, K., Kachurka, V., Sabourin, C., Amarger, V., Golovko, V., & Rossi, L. (2018). A human-like visual-attention-based artificial vision system for wildland firefighting assistance. *Applied Intelligence*, *48*, 2157–2179.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., & Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, *4*(2), 151–160.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*, 115–133.

Messina, N., Amato, G., Carrara, F., Falchi, F., & Gennaro, C. (2018). Learning relationship-aware visual features. In *Proceedings of the european conference on computer vision (eccv) workshops* (pp. 0–0).

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *The fourth international conference on 3d vision* (pp. 565–571).

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 1359–1367).

Mousavian, A., Košecká, J., & Lien, J.-M. (2015). Semantically guided location recognition for outdoors scenes. In *Ieee international conference on robotics and automation* (pp. 4882–4889).

Nam, H., Ha, J.-W., & Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In *Ieee conference on computer vision and pattern recognition* (pp. 299–307).

Niu, X., Han, H., Shan, S., & Chen, X. (2019). Multi-label co-regularization for semi-supervised facial action unit recognition. In *Advances in neural information processing systems* (pp. 909–919).

Niu, X., Han, H., Yang, S., Huang, Y., & Shan, S. (2019). Local relationship learning with person-specific shape regularization for facial action unit detection. In *Ieee conference on computer vision and pattern recognition* (pp. 11917–11926).

Niu, X., Han, H., Zeng, J., Sun, X., Shan, S., Huang, Y., ... Chen, X. (2018). Automatic engagement prediction with gap feature. In *Acm international conference on multimodal interaction* (pp. 599–603).

Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2017). Hierarchical multimodal lstm for dense visual-semantic embedding. In *Ieee international conference on computer vision* (pp. 1881–1889).

O'Reilly, B. F., Soraghan, J. J., McGrenary, S., & He, S. (2010). Objective method of assessing and presenting the house-brackmann and regional grades of facial palsy by production of a facogram. *Otology & Neurotology*, *31*(3), 486–491.

Pan, Z., Wu, F., & Zhang, B. (2023). Fine-grained image-text matching by cross-modal hard aligning network. In *Ieee conference on computer vision and pattern recognition* (pp. 19275–19284).

Pang, Y., Zhao, X., Zhang, L., & Lu, H. (2020). Multi-scale interactive network for salient object detection. In *Ieee conference on computer vision and pattern recognition* (pp. 9413–9422).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (pp. 8026–8037).

Pei, J., Zhong, K., Yu, Z., Wang, L., & Lakshmanna, K. (2023). Scene graph semantic inference for image and text matching. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(5), 1–23.

Prudviraj, J., Vishnu, C., & Mohan, C. K. (2022). M-ffn: multi-scale feature fusion network for image captioning. *Applied Intelligence*, *52*(13), 14711–14723.

Qi, X., Zhang, Y., Qi, J., & Lu, H. (2021). Self-attention guided representation learning for image-text matching. *Neurocomputing*, *450*, 143–155.

Qin, J., Huang, Y., & Wen, W. (2020). Multi-scale feature fusion residual network for single image super-resolution. *Neurocomputing*, *379*, 334–342.

Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., & Deng, W. (2023). Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Qin, X., Li, L., Hao, F., Ge, M., & Pang, G. (2024). Multi-level knowledge-driven feature representation and triplet loss optimization network for image–text retrieval. *Information Processing & Management*, *61*(1), 103575.

Qu, L., Liu, M., Cao, D., Nie, L., & Tian, Q. (2020). Context-aware multi-view summarization network for image-text matching. In *Acm multimedia* (pp. 1047–1055).

Qu, L., Liu, M., Wu, J., Gao, Z., & Nie, L. (2021). Dynamic modality interaction modeling for image-text retrieval. In *International acm sigir conference on research and development in information retrieval* (pp. 1104–1113).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. *OpenAI blog*.

Reed, L. I., Sayette, M. A., & Cohn, J. F. (2007). Impact of depression on response to comedy: a dynamic facial coding analysis. *Journal of Abnormal Psychology*, *116*(4), 804.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

Rubinow, D. R., & Post, R. M. (1992). Impaired recognition of affect in facial expression in depressed patients. *Biological Psychiatry*, *31*(9), 947–953.

Rusch, T. K., Bronstein, M. M., & Mishra, S. (2023). A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211–252.

Sankaran, N., Mohan, D. D., Lakshminarayana, N. N., Setlur, S., & Govindaraju, V. (2020). Domain adaptive representation learning for facial action unit recognition. *Pattern Recognit.*, *102*, 107127.

Sankaran, N., Mohan, D. D., Setlur, S., Govindaraju, V., & Fedorishin, D. (2019). Representation learning through cross-modality supervision. In *Ieee international conference on automatic face & gesture recognition* (pp. 1–8).

Sathik, M., & Jonathan, S. G. (2013). Effect of facial expressions on student's comprehension recognition in virtual educational environments. *SpringerPlus*, *2*(1), 1–9.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pretraining for speech recognition. *INTERSPEECH*.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, *45*(11), 2673–2681.

Shang, Z., Du, C., Li, B., Yan, Z., & Yu, L. (2023). Mma-net: Multi-view mixed attention mechanism for facial action unit detection. *Pattern Recogn. Lett.*.

Shao, Z., Liu, Z., Cai, J., & Ma, L. (2018). Deep adaptive attention for joint facial action unit detection and face alignment. In *European conference on computer vision* (pp. 705–720).

Shao, Z., Liu, Z., Cai, J., & Ma, L. (2021). Jaa-net: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, *129*(2), 321–340.

Shao, Z., Liu, Z., Cai, J., Wu, Y., & Ma, L. (2019). Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing*.

Shao, Z., Zhou, Y., Cai, J., Zhu, H., & Yao, R. (2023). Facial action unit detection via adaptive attention and relation. *IEEE Transactions on Image Processing*.

Shi, B., Ji, L., Lu, P., Niu, Z., & Duan, N. (2019). Knowledge aware semantic concept expansion for image-text matching. In *International joint conference on artificial intelligence* (Vol. 1, p. 2).

Shi, J., Alikhani, I., Li, X., Yu, Z., Seppänen, T., & Zhao, G. (2019). Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Trans. Circuits Syst. Video Technol.*, *30*(8), 2781–2795.

Shi, T., Ge, X., Jose, J. M., Pugeault, N., & Henderson, P. (2024). Detail-enhanced intra-and inter-modal interaction for audio-visual emotion recognition. *arXiv preprint arXiv:2405.16701*.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 802–810.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd international conference on learning representations.*

Socher, R., Lin, C. C.-Y., Ng, A. Y., & Manning, C. D. (2011). Parsing natural scenes and natural language with recursive neural networks. In *International conference on machine learning.*

Song, A., Wu, Z., Ding, X., Hu, Q., & Di, X. (2018). Neurologist standard classification of facial nerve paralysis with deep neural networks. *Future Internet*, *10*(11), 111.

Song, T., Chen, L., Zheng, W., & Ji, Q. (2021). Uncertain graph neural networks for facial action unit detection. In *Aaai conference on artificial intelligence* (p. 5993–6001).

Song, T., Cui, Z., Wang, Y., Zheng, W., & Ji, Q. (2021). Dynamic probabilistic graph convolution for facial action unit intensity estimation. In *Ieee conference on computer vision and pattern recognition* (pp. 4845–4854).

Song, T., Cui, Z., Zheng, W., & Ji, Q. (2021). Hybrid message passing with performance-driven structures for facial action unit detection. In *Ieee conference on computer vision and pattern recognition* (pp. 6267–6276).

Song, T., Zheng, W., Song, P., & Cui, Z. (2018). Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, *11*(3), 532–541.

Storey, G., Jiang, R., Keogh, S., Bouridane, A., & Li, C.-T. (2019). 3dpalsynet: a facial palsy grading and motion recognition framework using fully 3d convolutional neural networks. *IEEE Access*, *7*, 121655–121664.

Sudholt, S., & Fink, G. A. (2016). Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In *International conference on frontiers in handwriting recognition* (pp. 277–282).

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning* (pp. 1139–1147).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision* *and pattern recognition* (pp. 2818–2826).

Tallec, G., Dapogny, A., & Bailly, K. (2022). Multi-order networks for action unit detection. *IEEE Trans. Affect. Comput.*.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web* (pp. 1067–1077).

Tang, Y., Zeng, W., Zhao, D., & Zhang, H. (2021). Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Ieee international conference on computer vision* (pp. 12899–12908).

Tao, F., Ge, X., Ma, W., Esposito, A., & Vinciarelli, A. (2023). Multi-local attention for speech-based depression detection. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Tong, Y., & Ji, Q. (2008). Learning bayesian networks with qualitative constraints. In *Ieee conference on computer vision and pattern recognition* (pp. 1–8).

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347–10357).

Vacher, J., Launay, C., Mamassian, P., & Coen-Cagli, R. (2023). Measuring uncertainty in human visual segmentation. *arXiv preprint arXiv:2301.07807*.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 36–46).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *International conference on learning representations* (p. 1-12).

Vendrov, I., Kiros, R., Fidler, S., & Urtasun, R. (2015). Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

Walther, D. (2006). *Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics*. California Institute of Technology.

Wang, H., He, D., Wu, W., Xia, B., Yang, M., Li, F., ... Wang, J. (2022). Coder: Coupled diversity-sensitive momentum contrastive learning for image-text retrieval. In *European conference on computer vision* (pp. 700–716).

Wang, H., Zhang, Y., Ji, Z., Pang, Y., & Ma, L. (2020). Consensus-aware visual-semantic embedding for image-text matching. In *European conference on computer vision* (pp.

18–34).

Wang, L., Li, Y., Huang, J., & Lazebnik, S. (2018). Learning two-branch neural networks for image-text matching tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, *41*(2), 394–407.

Wang, L., Li, Y., & Lazebnik, S. (2016). Learning deep structure-preserving image-text embeddings. In *Ieee conference on computer vision and pattern recognition* (pp. 5005–5013).

Wang, S., Chang, Y., & Wang, C. (2021). Dual learning for joint facial landmark detection and action unit recognition. *IEEE Transactions on Affective Computing*.

Wang, S., Chen, S., Chen, T., & Shi, X. (2018). Learning with privileged information for multi-label classification. *Pattern Recognit.*, *81*, 60–70.

Wang, S., Peng, G., & Ji, Q. (2018). Exploring domain knowledge for facial expression-assisted action unit activation recognition. *IEEE Transactions on Affective Computing*, *11*(4), 640–652.

Wang, S., Wang, R., Yao, Z., Shan, S., & Chen, X. (2020). Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Ieee winter conference on applications of computer vision* (pp. 1508–1517).

Wang, Y., Su, Y., Li, W., Sun, Z., Wei, Z., Nie, J., ... Liu, A.-A. (2023). Rare-aware attention network for image–text matching. *Inf. Process. & Manag.*, *60*(3), 103280.

Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., & Fan, X. (2019). Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*.

Wang, Z., Chen, T., Ren, J., Yu, W., Cheng, H., & Lin, L. (2018). Deep reasoning with knowledge graph for social relationship understanding. *arXiv preprint arXiv:1807.00504*.

Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., & Shao, J. (2019). Camp: Cross-modal adaptive message passing for text-image retrieval. In *Ieee conference on computer vision and pattern recognition* (pp. 5764–5773).

Wang, Z., Song, S., Luo, C., Deng, S., Xie, W., & Shen, L. (2024). Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition. In *Ieee conference on computer vision and pattern recognition* (pp. 1270–1280).

Wei, X., Zhang, T., Li, Y., Zhang, Y., & Wu, F. (2020). Multi-modality cross attention network for image and sentence matching. In *Ieee conference on computer vision and pattern recognition* (pp. 10941–10950).

Wen, K., Gu, X., & Cheng, Q. (2020). Learning dual semantic relations with graph attention for image-text matching. *IEEE Trans. Circuits Syst. Video Technol.*, *31*(7), 2866–2879.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *European conference on computer vision* (pp. 3–19).

Wu, D., Li, H., Gu, C., Guo, L., & Liu, H. (2022). Improving fusion of region features and grid features via two-step interaction for image-text retrieval. In *Acm multimedia* (pp. 5055–5064).

Wu, M., Zhang, X., Sun, X., Zhou, Y., Chen, C., Gu, J., ... Ji, R. (2022). Difnet: Boosting visual information flow for image captioning. In *Ieee conference on computer vision and pattern recognition* (pp. 18020–18029).

Wu, Y., & Ji, Q. (2016). Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *Ieee conference on computer vision and pattern recognition* (pp. 3400–3408).

Wu, Y., Wang, S., Song, G., & Huang, Q. (2019). Learning fragment self-attention embeddings for image-text matching. In *Acm multimedia* (pp. 2088–2096).

Wu, Y., Wei, Y., Wang, H., Liu, Y., Yang, S., & He, X. (2023). Grounded image text matching with mismatched relation reasoning. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 2976–2987).

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Ieee conference on computer vision and pattern recognition* (pp. 1492–1500).

Xie, X., Li, Z., Tang, Z., Yao, D., & Ma, H. (2023). Unifying knowledge iterative dissemination and relational reconstruction network for image–text matching. *Inf. Process. & Manag.*, *60*(1), 103154.

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Ieee conference on computer vision and pattern recognition* (pp. 532–539).

Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., & Urtasun, R. (2019). Upsnet: A unified panoptic segmentation network. In *Ieee conference on computer vision and pattern recognition* (pp. 8818–8826).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048–2057).

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Aaai conference on artificial intelligence.*

Yang, H., Wang, T., & Yin, L. (2020). Adaptive multimodal fusion for facial action units recognition. In *Acm international conference on multimedia* (pp. 2982–2990).

Yang, H., Yin, L., Zhou, Y., & Gu, J. (2021). Exploiting semantic embedding and visual feature for facial action unit detection. In *Ieee conference on computer vision and pattern recognition* (pp. 10482–10491).

Yang, S., Li, Q., Li, W., Liu, M., Li, X., & Liu, A. (2023). External knowledge dynamic modeling for image-text retrieval. In *Acm multimedia* (pp. 5330–5338).

Yang, X., Dong, J., Cao, Y., Wang, X., Wang, M., & Chua, T.-S. (2020). Tree-augmented cross-modal encoding for complex-query video retrieval. In *International acm sigir conference on research and development in information retrieval* (pp. 1339–1348).

Yang, X., Feng, F., Ji, W., Wang, M., & Chua, T.-S. (2021). Deconfounded video moment retrieval with causal intervention. In *International acm sigir conference on research and* *development in information retrieval* (pp. 1–10).

Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017). Boosting image captioning with attributes. In *Ieee international conference on computer vision* (pp. 4894–4902).

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.*, *2*, 67–78.

Yu, J., Zhang, W., Lu, Y., Qin, Z., Hu, Y., Tan, J., & Wu, Q. (2020). Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval. *IEEE Transactions on Multimedia*, *22*(12), 3196–3209.

Yu, M., & Sun, A. (2023). Dataset versus reality: Understanding model performance from the perspective of information need. *J. Assoc. Inf. Sci. Technol.*, *74*(11), 1293–1306.

Yuan, Y., Zeng, J., & Shan, S. (2023). Describe your facial expressions by linking image encoders and large language models. In *Bmvc* (p. 377).

Zellers, R., Yatskar, M., Thomson, S., & Choi, Y. (2018). Neural motifs: Scene graph parsing with global context. In *Ieee conference on computer vision and pattern recognition* (pp. 5831–5840).

Zhang, B., Hu, H., Jain, V., Ie, E., & Sha, F. (2020). Learning to represent image and text with denotation graph. In *Conference on empirical methods in natural language processing.*

Zhang, H., Mao, Z., Zhang, K., & Zhang, Y. (2022). Show your faith: Cross-modal confidence-aware network for image-text matching. In *Aaai conference on artificial intelligence* (Vol. 36, pp. 3262–3270).

Zhang, K., Mao, Z., Wang, Q., & Zhang, Y. (2022). Negative-aware attention framework for image-text matching. In *Ieee conference on computer vision and pattern recognition* (pp. 15661–15670).

Zhang, Q., Lei, Z., Zhang, Z., & Li, S. Z. (2020). Context-aware attention network for image-text retrieval. In *Ieee conference on computer vision and pattern recognition* (pp. 3536–3545).

Zhang, X., Yang, H., Wang, T., Li, X., & Yin, L. (2024). Multimodal channel-mixing: Channel and spatial masked autoencoder on facial action unit detection. In *Ieee winter conference on applications of computer vision* (pp. 6077–6086).

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., … Girard, J. M. (2014). Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, *32*(10), 692–706.

Zhang, Y., Zhou, W., Wang, M., Tian, Q., & Li, H. (2020). Deep relation embedding for cross-modal retrieval. *IEEE Trans. Image Process.*, *30*, 617–627.

Zhao, K., Chu, W.-S., De la Torre, F., Cohn, J. F., & Zhang, H. (2016). Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Transactions on Image Processing*, *25*(8), 3931–3946.

Zhao, K., Chu, W.-S., & Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. In *Ieee conference on computer vision and pattern recognition* (pp. 3391–3399).

Zhao, R., Zhang, L., Fu, B., Hu, C., Su, J., & Chen, Y. (2024). Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 38, pp. 19643–19651).

Zhao, Y., Yu, X., Gao, Y., & Shen, C. (2022). Learning discriminative region representation for person retrieval. *Pattern Recognit.*, *121*, 108229.

Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., & Shen, Y.-D. (2020). Dual-path convolutional image-text embeddings with instance loss. *ACM Trans. Multimedia Comput. Commun. Appl.*, *16*(2), 1–23.

Zhou, M., Zhou, L., Wang, S., Cheng, Y., Li, L., Yu, Z., & Liu, J. (2021). Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Ieee conference on computer vision and pattern recognition* (pp. 4155–4165).

Zhu, F., Zhu, Y., Chang, X., & Liang, X. (2020). Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Ieee conference on computer vision and pattern recognition* (pp. 10012–10022).

Zhu, H., Meng, F., Cai, J., & Lu, S. (2016). Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation*, *34*, 12–27.

Zhu, H., Zhang, C., Wei, Y., Huang, S., & Zhao, Y. (2023). Esa: External space attention aggregation for image-text retrieval. *IEEE Trans. Circuits Syst. Video Technol.*.