Interpreting and Synthesising Human Faces and Articulated Animals from Video Data

Tong Shi

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

Computer Vision and Autonomous Systems
School of Computing Science
College of Science and Engineering
University of Glasgow

October 2025

Abstract

This thesis studies Human Emotion Recognition, 3D Human Head Reconstruction and Articulated Animal Reconstruction from in-the-wild videos. This is to enable interpretable analysis for human emotions, and facilitate controllable synthesis of human faces and reconstruction of articulated animals by jointly considering appearance (e.g., colour, opacity, and scale), audio, and dynamic motion cues. Our approaches to all three tasks share a common theme: they incorporate explicit representations of 2D and 3D motion and geometry.

When interpreting human emotions from a video, it is natural to do so from different modalities, that is, fusing features from 2D images and audio segments. Here lies our first contribution: understanding human emotions from talking videos by training a multi-modal neural network to predict various emotion categories in a principled fashion. Particularly, this is done by jointly model 2D visual features, optical flow feature, audio signals, and motion representations through an intra- and inter-modal interaction pipeline. We show it achieves state-of-the-art performance on multi-modality emotion recognition setting.

Beyond interpreting 2D images and audio features from a talking portrait video, we further estimate a 3D shape and learn how to reconstruct the 3D portrait and deform its shape so that it could talk, i.e. synthesising talking portrait videos. Our second contribution is a regression approach to synthesise talking portrait videos, which supports training purely from 2D images – without 3D supervision, and without using pre-defined 3D shapes from face specific priors such as 3D morphable models, landmarks and depth maps. Moreover, this model is generic, so it allows sampling new portrait and animating it condition on one arbitrary audio chunk.

Going beyond human heads, our third contribution addresses more complex articulated and deformable objects, articulated animals in particular, which are challenging to reason about in terms of motion and the articulated structure. In this task, we reconstruct a articulated 3D Animal model given an animal monocular video. It models jointly complex animal pose variation and canonical appearance, and optimises an implicit opacity–colour texture that is supported on a mesh scaffold.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I am truly fortunate to have had Professor Paul Henderson as my supervisor. His guidance and inspiration throughout my Phd have not only deepened my understanding of research but also profoundly influenced my attitude towards both work and life. Paul's dedication, meticulousness, and passion for research have been a constant source of motivation for me and has changed my research taste quite a lot. Paul has helped me in many aspects — from helping shape my research direction, improving my coding skills, to carefully reviewing my papers even in the final hours before submission deadlines. Beyond academia, I deeply admire his kindness, humility, and true gentlemanly character. These qualities have greatly shaped who I am today, both as a researcher and as a person. Besides, I was fortunate to have a desk close to my supervisor's office. He was always patient and generous with his time whenever I knocked on his door with questions. Those quiet moments of dedication and mutual encouragement will remain an unforgettable memory for me.

I am also deeply grateful to my second supervisor, Professor Nicolas Pugeault, for his continuous encouragement and insightful discussions throughout my PhD. He has provided me with many valuable ideas and helped me explore different research directions. His support has eased my stress and given me confidence in my work. I also greatly appreciate his patience in explaining complex concepts and revising my papers. His guidance has made my research journey much smoother and more enjoyable.

I would also like to thank my collaborators — Xuri Ge, Melonie de Almeida, and Daniela Ivanova. Xuri, now one faculty, has offered me invaluable guidance on evaluating my experimental results and improving my academic writing. Melonie has been kind and supportive, helping me with experimental baselines and helping me in coding and engineering. Daniela has always been patient and insightful, often coming up with creative ideas whenever I faced difficulties. I am truly glad to have been part of the CVAS group in the School of computing Science. It has been inspiring to witness its growth from a small team to a strong and dynamic research community. I have met many wonderful colleagues and friends there, and it has made my PhD experience both productive and fulfilling. And I would like to thank my parents, who have always supported me!

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Chapter 1

Introduction

This chapter introduces the motivation, challenge and objective of the thesis. In particular, we introduce the three main tasks: (i) human emotion understanding from talking-head videos, (ii) synthesising talking portraits from monocular videos, and (iii) synthesising articulated animals from monocular videos using a mesh scaffold. We then present the thesis tasks and contributions. Finally, we present the thesis statement and provide an overview of the thesis organisation.

1.1 Motivation and Challenges

Nowadays, computer vision systems have evolved rapidly to accurately understand diverse scenes. In today's information-driven era, the advancement of multimedia technologies (such as digital humans, avatars, virtual reality, and generative models) has rapidly developed under an enormous amount of diverse media data, including images, audio, videos, and text. By these technologies, we can identify objects, perceive and reason about them by either a deterministic or generative way. However, computer vision systems still struggle with the task of interpreting and synthesising diverse objects from the data in the wild [10]. Substantial progress has been made, over decades of research, understanding face dynamics and 3D animals, and these remain to be a central pursuit in computer vision and artificial intelligence more generally.

For the purposes of this thesis, we consider the task of perceiving and reasoning about more challenging articulated objects from 2D images, specifically Human faces and Animals. These entities exhibit rich dynamics but often appear under difficult conditions such as occlusions and limited viewpoints. The essential goal is to find human interpretable 3D representations of these articulated objects or their labels such as emotions. Ultimately, this research aims to understand, predict, generate, reconstruct, and rig these entities.

Interpretation of articulated objects requires predicting important attributes of these entities and has myriad applications; we now give a brief flavour of some. Interpreting the emotion category from humans is an important task in many fields, e.g. affective computing and human-robot-interaction, where machines need to understand human emotion correctly in order to respond appropriately. In most cases, emotion recognition from facial expressions relies on regression models trained on diverse facial data to improve generalization. However, faces within the same emotion category can vary widely in texture and appearance. Although the face data could be fetched widely, but one emotion class face might have various face images under different condition. Therefore, this variability requires regression models manner to capture facial representations which contains emotion information, and using more information for modelling these representations such as fusing the feature from other modalities.

Synthesising human faces and articulated animals requires generating, reconstructing and animating these entities. Reconstructing 3D heads and animals is difficult, particularly when only monocular data is available for training. Talking Head Generation aims to synthesize and controllably animate 3D human heads driven by modalities such as text and audio. This task typically requires higher-fidelity geometry and texture than general 3D object reconstruction. Analogously, articulated animal reconstruction demands similarly detailed modeling under non-rigid motion and occlusions. A common—and particularly challenging—setup is single-image reconstruction, where a full 3D representation must be recovered from a single view, in contrast to the more supervised multi-view setting. Unlike human faces, where well-established rigging techniques and parametric models enable accurate reconstruction, articulated animals such as dogs and cats present additional difficulties. We consider whole bodies of animals, and this is immediately more complex since there are lots of legs and complex articulation. Besides, obtaining accurate camera intrinsics and extrinsics from in-the-wild articulated animal data is difficult. This is because monocular videos of animals often suffer from occlusions, uncontrolled lighting, and unpredictable poses. Although recent advances have made it possible to reconstruct rigid scenes reliably from moving cameras, recovering non-rigid, deformable 3D shapes from monocular sequences remains a highly unconstrained and challenging problem.

The overall goals of this thesis are to predict human emotions, and to reconstruct and animate 3D representations of human faces and animals. We describe the specific tasks and contributions in detail in the next section.

1.2 Thesis Tasks and Contributions

This thesis tackles three related but self-contained tasks, touching on three different, yet related, aspects of human faces and animals, and makes contributions for each of them; In particular, we do not attempt to build a system incorporating all our contributions, although we do present suggestions on how they can be combined in future work. We now describe each of our contributions.

1.2.1 Human Emotion Recognition

Task. Understanding human emotions from multimodal data is a key problem in affective computing, human–computer interaction, and social robotics. In particular, Audio-Visual Emotion Recognition (AVER) aims to infer emotional states by jointly modelling the dynamic cues present in both the visual (facial expressions) and auditory (speech) modalities. While recent progress in multimodal learning has improved overall recognition accuracy, two major challenges remain largely unsolved. First, it is difficult to effectively capture fine-grained temporal variations within each modality, such as subtle facial texture changes or micro-expressions between consecutive frames, which are often lost after global pooling or high-level feature abstraction. Second, modelling **complex inter-modal dependencies** between audio and video streams remains challenging, as the two modalities have inherently different feature distributions and temporal dynamics. Existing AVER methods often fuse modalities through simple concatenation or shared attention layers, which can lead to information loss and insufficient exploitation of cross-modal complementarity. Consequently, they struggle to represent subtle yet discriminative emotion-related details, particularly under unconstrained and diverse emotional expressions. We tackle multimodal human emotion recognition under a difficult setting: given one image and one audio chunk, the proposed model fuses visual and acoustic cues to achieve efficient and accurate emotion classification.

Contributions. To address these challenges, we propose a new framework called *Detail-Enhanced Intra- and Inter-modal Interaction (DE-III)* for audio-visual emotion recognition. Our method explicitly models local temporal details in video and promotes deeper, bidirectional interaction between modalities. The key contributions of this work are summarized as follows:

- We leverage **optical flow** as an explicit representation of fine-grained facial motion and texture variations between video frames. To integrate this information, we design a **pairwise optical flow–video attention fusion module**, which adaptively weights and fuses flow and frame features, yielding a detail-aware visual embedding that better reflects emotional cues.

- We develop novel **intra- and inter-modal feature enhancement modules** that allow audio and video modalities to attentively exchange information in both directions. These modules incorporate residual connections to preserve modality-specific characteristics while effectively capturing complementary information from the other modality.
- We propose a **Cross-attention Classification** with independent prediction heads for audio cross to video, video cross to audio, and fused representations. This structure encourages robust learning across modalities and provides a more comprehensive supervision signal during optimization.
- Extensive experiments on three widely used benchmarks—**CREMA-D**, **MSP-IMPROV**, and **RAVDESS**—demonstrate the effectiveness of our method. The proposed DE-III achieves new state-of-the-art performance on both categorical and continuous emotion recognition tasks, surpassing recent transformer-based fusion approaches and highlighting the benefit of detailed intra- and inter-modal modelling.

1.2.2 Talking Head Generation

Task. Talking Head Generation (THG) aims to synthesize realistic and temporally coherent videos of a talking portrait, conditioned on a portrait image and an audio sequence. This task is fundamental for applications such as digital avatars, movie generation, and virtual content creation. Despite recent progress in neural rendering and implicit 3D representations, generating photorealistic talking heads remains an open challenge, especially when only monocular videos are available for training. The key difficulty lies in jointly modelling **3D head geometry, dynamic motion, and appearance consistency** under natural, unconstrained conditions. Existing methods often rely on parametric priors such as 3D Morphable Models (3DMM) or deformation fields that encode predefined facial motion spaces. Although such priors help stabilize optimization, they inevitably constrain the expressiveness of generated motion and fail to generalize across diverse identities, expressions, and head poses. Moreover, implicit NeRF-based methods tightly couple static appearance and dynamic motion within volumetric fields, which makes it difficult to achieve accurate geometry and consistent rendering across different viewpoints. These limitations motivate a new formulation that explicitly separates the static 3D structure of the head from the dynamic, audio-driven motion while remaining free from handcrafted priors and multi-view supervision. Our task is that, from one portrait image and a short audio segment, our method reconstructs a personalized 3D head and animates it to produce a temporally coherent 4D talking sequence.

Contributions. To address these challenges, we propose *Splat-Portrait*, a novel framework

for audio-driven talking head generation built upon the principles of **3D Gaussian Splatting (3DGS)**. Our approach leverages the explicit, differentiable nature of 3D Gaussians to learn realistic head geometry and motion directly from monocular video data. By decomposing the synthesis process into static and dynamic components, Splat-Portrait achieves both structural accuracy and temporal expressiveness. The main contributions of our work can be summarized as follows:

- **3D Gaussian-based static reconstruction.** We introduce a static branch that reconstructs the subject's canonical 3D head representation from a single portrait image using Gaussian splats. This branch simultaneously estimates the background through an inpainting-based refinement network, allowing seamless head–background compositing without requiring segmentation masks or multi-view alignment.

- **Audio-conditioned dynamic animation.** We design an audio-driven decoder that directly animates Gaussian splats by predicting per-point temporal offsets modulated by speech features. Unlike deformation-based methods, our model avoids complex motion fields and instead learns fine-grained correlations between phonetic content and geometric displacement, leading to synchronized and expressive facial motion.

- **Two-stage self-supervised training.** We adopt a progressive training strategy: a large-scale pretraining phase focuses on stable static reconstruction, followed by an audio-conditioning stage where temporal dynamics are introduced. Additionally, we integrate **Score Distillation Sampling (SDS)** to enhance photo-realism and geometric fidelity under unseen viewpoints.

- **Generalization and Portrait in a content.** Our method requires no 3D ground-truth supervision, yet generalizes effectively across identities and expressions. The Gaussian representation ensures faster rendering and lower memory cost compared to volumetric NeRFs [11], while maintaining high-quality details. Our method remain a portrait in a content, i.e. a reconstructed portrait with background. Extensive experiments has showed that our method are better than pervious methods.

In summary, Splat-Portrait provides a unified and efficient framework for monocular 3D talking head generation, bridging the gap between explicit 3D representations and dynamic motion learning. It demonstrates that Gaussian Splatting, when combined with audio-conditioned modelling, can serve as a powerful foundation for building expressive and generalizable digital humans.

1.2.3 Articulated Animal Reconstruction

Task. Reconstructing animatable 3D animal avatars from monocular videos is a long-standing challenge in computer vision and graphics. Unlike humans, animals exhibit highly non-rigid and often unpredictable deformations, possess complex surface appearances (e.g., fur, colour patterns, tails), and typically lack well-calibrated camera parameters or annotated 3D datasets. The task requires jointly recovering both the articulated 3D structure and the detailed appearance of the animal over time, using only casual, in-the-wild videos. Existing works either rely on single-frame image-based shape regression or multi-view video inputs with strong supervision, which limit their generalization and robustness in real-world scenarios. Moreover, prior template-based approaches using the SMAL parametric model struggle with ambiguous or unseen viewpoints, since they depend on sparse keypoint correspondences that mainly cover front-facing body parts. Template-free methods, while flexible, often fail to maintain geometric consistency or produce temporally coherent motion. Hence, Our task is how to leverage both the geometric prior of an animal template and the dense visual cues available in videos to achieve faithful, temporally consistent 3D reconstructions of animals from monocular inputs.

Contributions. We adopt, build on, and evaluate more data on an existing method called AnimalAvatar [9], an approach that reconstructs fully animatable and textured 3D animal models from monocular articulated animal videos. It jointly estimates the animal's pose, shape, and appearance by combining the SMAL [12] parametric template with dense image-to-surface correspondences and implicit neural representations. Our method current has majority components are learned from AnimalAvatar [9], the main contributions of our method build on AnimalAvatar are as follows:

- **Dense tracking loss** Apart from AnimalAvatar [9], we propose a dense tracking method that provides 3D-aware pixel motion information to further supervise the poses and shapes of articulated animals. This method introduces a pixel tracking mechanism that captures how each pixel of the animal moves across subsequent frames, and by combining this loss with the existing losses we further improves the accuracy of articulated animal reconstruction.
- **More evaluation, and state-of-the-art performance.** Apart from the data used from AnimalAvatar, we perform a much more thorough evaluation of AA than the original authors, on more diverse video data under challenging scenarios e.g. light, occlusion, complex articulation and complex animal texture. We conduct extensive experiments on these data and find that the original AnimalAvatar method does not generalize well to these challenging cases.

Our method that is built on AnimalAvatar, apply dense tracking loss function to enhance the 3D reconstruction and analyse a larger amount of data across various articulation states.

1.3 Thesis Statement

The central claim of this thesis is that interpreting and synthesizing rigged entities—human faces and articulated animals—from visual observations can be effectively achieved by learning structured and controllable representations through understanding, generation, and reconstruction. These tasks share a common objective: to recover and exploit structured representations of rigged entities from the monocular observations. By leveraging 2D images, monocular video data, motion cues, as well as geometric and semantic consistency across modalities, the approaches proposed in this thesis aim to build understandable regression model for classification and animatable, high-fidelity, and generalizable 3D representations from monocular data for generation. For instance, in human emotion recognition, relational understanding across audio cues and 2D images enhances the emotion recognition robustness. In talking head generation, disentangling static and dynamic structures through explicit 3D representations, such as Gaussian Splatting, facilitates photorealistic reconstruction from limited viewpoints. Furthermore, in articulated animal reconstruction, embedding dense surface correspondences and implicit appearance representations promotes geometric and textural coherence, even under unconstrained camera motion and complex non-rigid deformations. Overall, leveraging 2D images, prior knowledge, and motion cues from monocular videos through supervised learning enables effective interpretation of human emotion recognition and reconstruction of 3D human faces and animals

1.4 Thesis Outline

We devote one section to describing each of our main contributions for every chapter, and we list the related publications and paper works for each chapter.

Chapter 2. We present related work and background on our three sub-tasks in Chapter 2. This includes the baselines, related technical approaches, datasets, metrics used in related works and highlights the unique challenges of each topics.

Chapter 3 focuses on understanding human emotions from multimodal signals. We introduce a detail-enhanced framework for audio-visual emotion recognition (DE-III) method, which captures fine-grained temporal dynamics within and across modalities. This chapter presents a comprehensive evaluation on benchmark datasets and demonstrates that incorporating optical-flow–based motion cues and cross-modal attention substantially improves

emotion recognition accuracy and robustness.

Chapter 4 introduces talking head generation. We propose *Splat-Portrait*, a novel audio-driven framework built upon Gaussian Splatting. The chapter describes how disentangling static and dynamic components allows realistic 3D head reconstruction and expressive motion synthesis from audio input. The model provides both static 3D Portrait in a content and 4D talking Portrait given corresponding audio chunk. Extensive experiments show that the proposed method achieves superior rendering quality, and generation to unseen head parts.

Chapter 5 addresses the reconstruction of articulated animal avatars from monocular videos. It presents our method which builds upon AnimalAvatar [9], and various scientific experiments on more animal videos. AnimalAvatar consists of a dense correspondence–guided method combining the SMAL template model with implicit texture representations, and camera optimization process. Our idea is to extend it with dense tracking loss. This chapter demonstrates how dense tracking supervision and temporally coherent optimization enable faithful 3D animal reconstructions under challenging real-world conditions, achieving state-of-the-art results on multiple animal entities.

Chapter 6 concludes the thesis by synthesizing the findings across the three sub-tasks of this thesis. It discusses the shared challenges of interpreting and synthesizing human faces and articulated animals, highlights the importance of and contributions of this thesis, and finally it outlines potential future directions toward unified, generalizable talking head generation frameworks for both humans and animals.

1.4.1 Publications and Paper Works

The following is a list of publications included in this thesis:

Chapter 3: Shi, T., Ge, X., Jose, J. M., Pugeault, N., & Henderson, P. (2024). *Detail-enhanced intra- and inter-modal interaction for audio-visual emotion recognition.* In *International Conference on Pattern Recognition (ICPR)*. Springer Nature Switzerland.

Chapter 4: Shi, T., de Almeida, M., Ivanova, D., Pugeault, N., & Henderson, P. (2025). *SplatPortrait: Generalizing Talking Heads with Gaussian Splatting.* In *International Conference on Multimedia Modeling (MMM)*.

Chapter 2

Background and Related Work

In the chapter, we introduce the related work for each of our three sub-tasks, We organise the related works into two subsections: Human Faces and Animals. In this first Section, we review the research on understanding and generating human faces, highlighting widely adopted approaches and commonly used datasets. The second subsection covers the reconstruction of articulated animals, with an emphasis on techniques that handle non-rigid motion and occlusion.

2.1 Understanding and Reconstructing Human Heads

In this section, we introduce the related work for our first two sub-tasks, Human Emotion Recognition and Talking Head Generation. These tasks are both relative to human face images, and human audio. We review key research in the field of understanding human expressions, the generation of expressive human faces, their condition on audio, and the datasets used in the face domain. Research on Emotion Recognition has been attempts broadly many modalities, such as text, image, audio, and psychological Signals. Our focuses are face images and audio, we discuss in following order: 1) Emotion recognition from face images, 2) Emotion recognition from Speech and 3) Audio–Visual Emotion Recognition (AVER). Research on talking head generation broadly falls into two streams. (1) 2D-based approaches operate purely in the image plane—manipulating pixels or learned 2D features—without explicit 3D reasoning. (2) 3D-based approaches explicitly reconstruct a head using a 3D representation (e.g., meshes, implicit fields, Gaussians) and then deform this representation over time into a 4D dynamic model to drive animation.

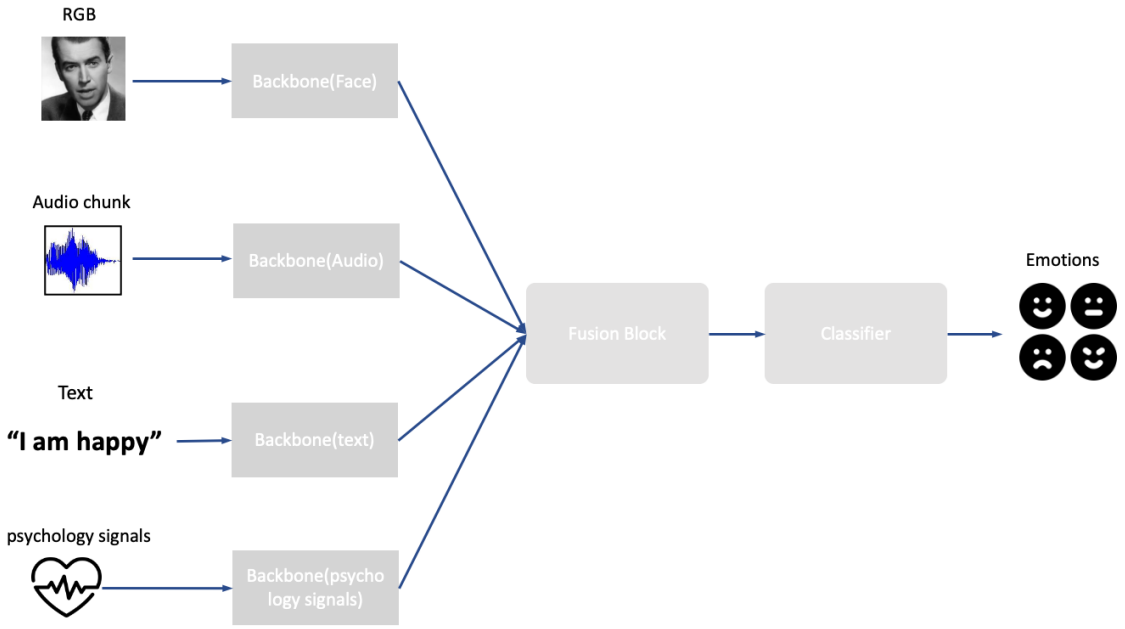

Figure 2.1: Overview for multi–modal ER systems: multiple streams fuse together for emotion classification

2.1.1 Emotion Recognition from Face images

Research on emotion recognition spans single–modality pipelines, such as faces alone, speech alone, or text alone. And multi–modal systems that couple the multiple streams including RGBs, audio, text or psychology signals, through some form of fusion as shown in Figure 2.1. Bringing multiple streams together is general a big research. We specifically focus on the facial (RGB) and acoustic (speech) modalities. Below we review the pieces most relevant to our setting: facial expression recognition (FER), speech emotion recognition (SER), and audio–visual emotion recognition (AVER) with decision– or feature–level fusion. Early FER systems relied on hand-crafted geometric or appearance cues (e.g., landmarks) with shallow classifiers such as SVM/HMM, and required careful preprocessing to cope with pose and illumination changes [13, 14]. Recent efforts inject temporal context attention to focus on salient facial regions, and intensity–aware objectives to handle subtle or occluded expressions [15, 16, 17, 18]. In constrained settings (e.g., lab videos), these models achieve strong accuracy, yet their performance can degrade in the wild due to motion blur, head rotations, and identity/expression entanglement, motivating complementary audio cues and cross–modal priors.

A recurring observation is that discriminative evidence for many expressions lies in *subtle*, short-lived motions (eyebrow raises, nasolabial changes, micro-smiles). Intensity-aware formulations—where labels reflect graded expression strength rather than binary presence—also help models avoid over-fitting to exaggerated prototypes and generalize to everyday,

low-intensity displays [16, 17]. In practice, combining fine temporal windows with region-focused attention has proven more reliable than scaling networks depth-wise on single frames [18]. In-the-wild scenarios introduce frequent head rotations, partial occlusions (hands, hair, glasses), and motion blur. Classical pipelines mitigated these with alignment and illumination normalization, but deep systems increasingly rely on built-in mechanisms: spatial attention suppresses noisy regions and highlights stable facial parts (eyes/eyebrows/mouth) [15, 18]; multi-branch designs fuse global and local crops to retain identity-invariant evidence; and short-term temporal modeling compensates for occasional frame corruption by integrating adjacent, higher-quality frames [17]. Nevertheless, when occlusions are persistent or the head is far from frontal, identity/expression entanglement resurfaces, and performance drops relative to controlled lab conditions reported by surveys [19, 20].

Another practical bottleneck is dataset bias and label ambiguity, especially for classes with overlapping semantics (e.g., "neutral" vs. low-intensity "sad"). Recent works address this with stronger augmentation schedules, class-rebalancing objectives, and curriculum-style intensity modeling [16, 17]. Surveys consistently note that improvements in FER accuracy often come from procedural choices—temporal sampling, region selection, and calibration of attention—rather than adding more parameters alone [21, 19, 20]. These insights motivate designs that (i) explicitly encode motion, (ii) localize attention to stable facial parts, and (iii) regularize for intensity, which together yield stronger out-of-domain behavior without excessive model complexity [15, 18].

Consistent with these trends, we adopt simple off-the-shell backbones, i.e. EfficientNet [22], and Dinov2 [23] augmented with lightweight temporal cues and region-sensitive attention. Rather than relying on deeper static encoders, we preserve fine motion patterns over short horizons and balance global context with part-level focus; this mirrors the empirically validated recipe across recent FER literature for handling subtle expressions and mild occlusions in the wild [15, 16, 18, 17].

2.1.2 Emotion Recognition from Speech

Compared with facial streams, speech naturally carries strong sequential regularities at multiple time scales (phoneme, syllable, word). SER has a parallel trajectory: classical pipelines transform audio into prosodic and spectral descriptors (e.g., MFCC, pitch, ZCR) and train SVM/HMM–style classifiers [24, 25, 26]. Modern SER replaces manual features or augments them with learned representations from recurrent models (RNN/LSTM), often with attention to emphasize emotionally informative segments [27, 28, 29, 30, 31]. Practical systems also address data scarcity and long silences through augmentation and silence removal; when benchmarks such as SAVEE, MSP-IMPROVE, and RAVDESS are used, bidirectional LSTM variants with moderate model size can reach high accuracy while keeping deployment

affordable [32, 33, 34]. These results underscore a key point for multi-modal design: while audio alone can be highly predictive, its error modes (e.g., lexical ambiguity, background noise) are complementary to visual ones. Log–Mel spectrogram containes time-frequency patterns, while BiLSTMs pool information over longer contexts to capture prosodic contours and rhythm; attention layers could further highlight emotionally salient spans (e.g., sustained pitch elevations) [29, 27, 28]. End–to–end systems often combine short windows for local detail with utterance-level pooling for global decisions, which improves robustness to speaking rate variation and hesitations [30, 31]. Classical HMMs and modern RNNs thus form a continuum: both model temporal dependencies, but deep encoders reduce reliance on handcrafted features while maintaining sequence awareness [25, 24].

In-the-wild recordings suffer from channel mismatch, background noise, and variable loudness. Common practice includes voice activity detection and silence removal, dynamic range normalization, and targeted augmentation (time/frequency masking, pitch shifting) to mitigate overfitting [24, 27]. Because emotion corpora are typically small and imbalanced, systems favor compact backbones with attention in lieu of very deep models, trading a slight drop in peak accuracy for improved generalization and latency—a pattern repeatedly observed on SAVEE, MSP-IMPROV, and RAVDESS [32, 35, 36]. Utterance segmentation also matters: aggregating decisions over overlapping chunks stabilizes predictions under long pauses and coarticulation, while calibrated thresholds curb false positives in neutral segments [28, 31]. Overall, survey evidence suggests that careful preprocessing and temporal pooling strategies are as impactful as architectural scale for SER in realistic settings [24, 27]. Guided by these findings, audio branch prioritizes a compact spectrogram front end followed by BiLSTM with local attention, rather than deep stacks. This choice reduces computational overhead and latency while preserving the prosodic cues that complement the visual stream.

2.1.3 Audio–Visual Emotion Recognition (AVER)

Fusing facial and speech streams usually boosts robustness and cross-domain generalization [37, 38]. A straightforward AVER pipeline is Early AVER: train shallow models per modality and fuse at the score or feature level (e.g., GMM/SVM) [39, 40, 41]. With multimodality modelling, a common approach is using two separate backbones for video alone and for audio alone, then late fusion using weighted averaging or a small MLP; more advanced variants add cross attention for visual dynamics, temporal relation to align the two streams over time [42, 43, 44]. However, bigger fusion stacks often increase latency and compute without reliable gains in the wild [45]. In practice, lightweight designs that (i) cap backbone complexity per modality, (ii) use simple but well-calibrated decision-level fusion, and (iii) model operation on various datasets tend to offer a better accuracy–complexity

trade-off [41, 44].

A recurring theme across AVER is the handling of temporal asynchrony and modality imbalance. Audio often carries higher short–term variability than video; naïve concatenation can suppress subtle facial cues or, conversely, over–weight noisy speech segments. To mitigate this, studies adopt temporal alignment (e.g., fixed hop sizes, windowed pooling) and normalization strategies that stabilize per–stream statistics before fusion [42, 44]. Cross–attention mechanisms explicitly modulate one stream by the other to focus on complementary evidence, though the added parameters and memory footprint can be substantial on long clips [43, 45]. Decision–level fusion remains attractive when deployment constraints dominate: calibrated weighting or simple meta–classifiers on per–modality posteriors yield competitive accuracy while keeping inference predictable and easy to harden against dropouts [41, 44]. Evaluation choices matter, too. On small– to mid–scale benchmarks like SAVEE, MSP-IMPROVE, and RAVDESS, broad ablations show that careful per-modality regularization, moderate backbone depth, and conservative fusion beat heavier stacks once latency and robustness are considered [41, 44]. Recent studies also stress uncertainty-aware fusion and test-time calibration to cope with real-world noise, supporting the view that scaling up the architecture alone does not ensure generalization [37, 38, 45]. Overall, a practical recipe emerges: keep encoders compact yet expressive, align and normalize signals before fusion, and favor calibrated decision-level aggregation when efficiency and reliability are the priorities [41, 44].

2.1.4 2D based Talking Head Generation

Understanding how human emotions arise—and how they map across audio and visual cues—is crucial for generating truly expressive human faces. We now introduce Talking Head Generation. Research on 2D talking-head generation spans two intertwined threads: (i) self-/weakly-supervised image animation that transfers motion from a driving signal to a static portrait via learned 2D correspondences, and (ii) speech-driven synthesis that conditions the animation on audio while preserving the identity and visual fidelity of the input face. Below we outline related approaches, focusing strictly on 2D formulations without requiring explicit 3D reconstruction.

Early systems established the idea of learning to warp a source portrait using motion cues estimated from a driving frame or video, e.g. Figure 2.2, or to animate a single portrait image based on one audio chunk, e.g. Figure 2.3. Both methods learn to manipulate a face via disentangled codes and a learned warping field [46]. First-Order Motion Model (FOMM) generalizes this idea with keypoint-based motion and learned local affine transforms, enabling one-shot reenactment and robust transfer across identities [47]. Subsequent work improves motion expressivity and temporal stability: Motion Transformer leverages

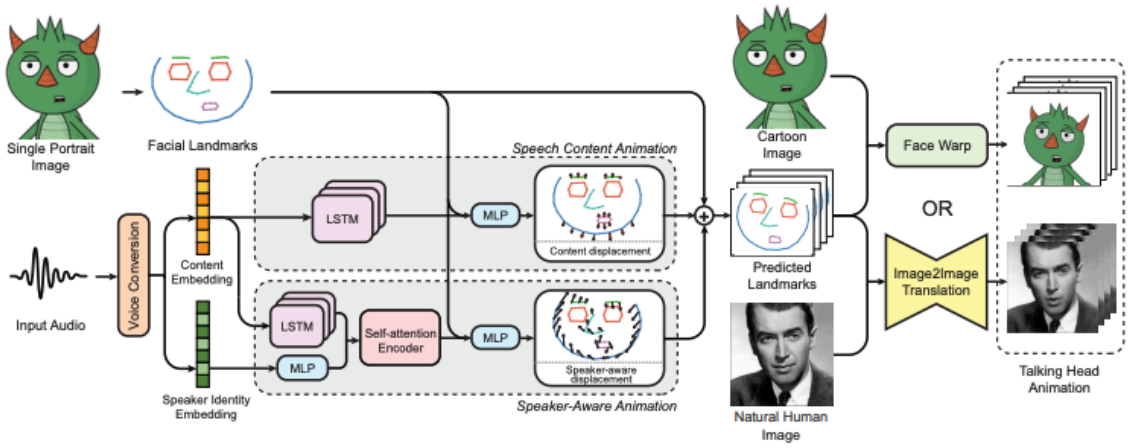

Figure 2.2: Example for 2D talking Head Generation based on frame driven method and audio driven method reproduced from MakeitTalk [4]

transformer-based temporal aggregation for unsupervised animation [48], other 2D method [5] navigates a latent space with a warping renderer to produce temporally coherent edits and reenactment [49]. Methods such as sadtalker [50] further decouple pose and expression within a 2D editing pipeline to reduce jitter and identity leakage during reenactment [51]. One-shot free-view talking-head synthesis also appears in the 2D literature through strong warping and rendering modules that preserve identity across poses [52]. Collectively, these approaches show that 2D motion fields, learned keypoints, and latent warping can deliver convincing facial dynamics without explicit geometry, though challenges remain in handling large pose, occlusion, and long-range temporal consistency.

For audio-driven 2D talking head generation, early work like Anitalker [5], as shown in fig:anitalker generates mouth motion and coarse head dynamics from speech with identity preservation. Wav2Lip ("a lip-sync expert") [53] demonstrates that precise lip synchronization can be obtained by focusing on an audio-visual sync discriminator and a 2D generator, improving robustness "in the wild", it also conclude that audio feature has high correlations to lip region than other facial parts. Subsequent systems emphasize controllability and pose sensitivity in purely 2D setups: PC-AVS factorizes pose control from the audio-visual representation for pose-consistent animation [54], and Audio2Head targets one-shot generation with natural head motion while remaining in the 2D warping-and-rendering regime [**?**]. With the advent of diffusion, many works replace adversarial renderers with audio-conditioned diffusion processes to enhance visual fidelity and reduce artifacts while staying 2D at inference time (e.g., DiffTalk [55], Diffused Heads [56], Diff2Lip for lip-synchronization [57], and EMO for expressive audio-to-video diffusion [58]). These diffusion-based 2D generators tend to improve image realism and diversity, though at increased computational cost.

Across both reenactment and speech-driven settings, a central 2D question is how to

Figure 2.3: Example for 2D talking Head Generation based on audio driven method reproduced from Anitalker [5]

encode motion in a way that is expressive yet identity-agnostic. Purely explicit 2D control points (e.g., landmarks) are highly interpretable but can be too rigid for subtle dynamics, motivating learned motion latents and stronger warping renderers. An interesting single-identity study by Luo et al. [59] transfers human talking-face motions to animal faces, achieving cross-domain facial reenactment. Another persistent issue is that identity leakage during cross-identity driving, which 2D pipelines address by better factorization of appearance/motion and by training objectives that regularize identity consistency. Evaluation typically relies on 2D reconstruction and perceptual metrics (PSNR/SSIM/LPIPS), identity similarity (CSIM), and audio-visual sync scores (e.g., LSE-D) on public talking head monocular datasets such as VoxCeleb [60], HDTF [61], and VFHQ [62] that provide unconstrained faces and in-the-wild conditions for 2D face learning. Overall, 2D talking head generation has converged on (i) learned warping fields guided by keypoints or motion latents, (ii) audio-conditioned 2D generative models (increasingly diffusion-based) for lip-sync and prosody-aligned motion, and (iii) lighter identity and pose controls that keep pipelines efficient while remaining robust under real-world variability.

2.1.5 3D Head Reconstruction

Purely 2D pipelines animate a reference image via warping and inpainting [63, 64, 65, 66, 67]; they are fast and data-friendly but struggle with disocclusions and identity drift under large view changes. 3D head reconstruction is closely linked to digital avatar methods that

support diverse inputs (multi-view images, monocular video, or a single portrait image) and a range of representations, from classical meshes to volumetric/radiance fields and recent popular Gaussian Splatting representation. Current systems build animatable head avatars with (i) textured meshes [68, 69, 70, 71, 72], (ii) volumetric or NeRF-like models [73, 74, 75], and (iii) point-based methods [76]. Meshes are efficient, easy to rig, and align well with 3DMM controls [77], but can miss fine details (e.g., hair/teeth). Volumetric/radiance-field methods capture rich appearance and view-dependent effects at higher rendering cost [11], often guided by 3DMM parameters or learned latents [78, 79, 80, 74]. A complementary line uses 3D Gaussian Splatting [81], achieving photoreal quality with efficient optimization and real-time rendering; head-specific variants show high fidelity and explicit rigging [82, 83, 84, 85, 86].

Single-image avatars rely on large-scale priors to (i) directly regress meshes [87, 88, 89], (ii) predict feature-grid/tri-plane-style representations and use feed-forward neural renderers [90, 91, 92, 93, 94], or (iii) adapt NeRF-like fields for one-shot settings [95, 96]. Expression/pose control is typically driven by 3DMM conditioning, latent controls, or learned deformation fields [97, 98, 99, 100]. Across different 3D representations, the main trade-offs are rendering speed vs. photorealism, explicit control vs. expressivity, and robustness to challenging domains (e.g., stylized or low-light data).

Across different 3D representations, neural scene representations have become a compelling alternative for modeling facial geometry and view-dependent appearance from sparse captures. Implicit radiance fields (NeRF) parameterize colour and density as continuous functions, achieving high-fidelity novel-view synthesis under multi-view supervision [101]. Deformable extensions such as Nerfies and D-NeRF model nonrigid motion via per-frame warps or canonical fields, improving temporal coherence for expressive faces [102, 103]. Head-specific variants condition the radiance field on audio to animate lips and jaw directly from speech (e.g., AD-NeRF), enabling talking-head synthesis without explicit blendshape rigs [104]. To increase efficiency, tri-plane feature representations (EG3D) replace full volumetric grids with three orthogonal feature planes, delivering fast rendering and GAN-based inversion/editing while preserving fine identity cues [105]. More recently, explicit point-based encodings with 3D Gaussian Splatting (3DGS) trade volumetric integration for differentiable rasterization of anisotropic Gaussians, yielding order-of-magnitude speedups and sharper details from limited views [106]. While these neural 3D families excel at photorealistic view synthesis and can capture subtle, view-dependent facial reflectance, they typically rely on short-baseline multi-view or carefully calibrated monocular video, and require additional canonicalization.

2.1.6 Face priors

Face specific priors are very important for 3D head reconstruction. The classic 3D Morphable Model (3DMM) is one typical face specific prior with a linear shape–appearance subspace learned from scans, fitting via analysis–by–synthesis [107]. Its public successor, the Basel Face Model (BFM), broadened access to a research community and catalyzed a decade of work on 3D face analysis [108]. Subsequent efforts increased population coverage and statistical power by collecting larger, more diverse scan corpora to learn neutral–expression identity spaces [109, 110]. These linear identity spaces are compact and easy to fit, but their expressiveness is limited by the training cohorts (age, ethnicity, expression neutrality) and by linearity assumptions. Beyond pure identity spaces, expression variability has been modeled additively by learning a residual PCA space on top of the neutral identity subspace [111], or jointly via multilinear (tensor) models that disentangle identity and expression factors [112]. Practical facial reenactment pipelines often combine linear identity, linear expression, and albedo components to enable real–time performance capture and editing from RGB video [113]. Expression–specific PCA banks [114] offer another alternative. While effective, these approaches can struggle to represent the continuum of natural facial motion when training data cover only a few discrete poses; multilinear models, in particular, presuppose a limited set of expression modes shared across subjects.

High-end capture pipelines combine dense scanning and semi-automatic tracking to produce personalized rigs and photoreal animation (e.g., Digital Emily) [115]. Online or real-time variants learn corrective spaces or example-based rigs to adapt to new performances without full re-scans [116, 117, 118]. At the same time, internet photo collections and monocular videos have been exploited for person-specific reconstruction using shape-from-shading, 3D flow, and texture synthesis [119, 120, 121], demonstrating feasibility outside controlled studios but with remaining challenges in geometry accuracy, temporal coherence, and illumination disentanglement. Building expressive models from many subjects and expressions hinges on robust nonrigid registration. Variants of nonrigid ICP and learned correspondence have been developed for static faces [122, 123], while dynamic capture pipelines reduce drift using anchor frames and repeated expressions [124]. Co-registration methods align entire datasets while simultaneously learning a shared template and deformation priors [125], enabling consistent mesh topology across subjects and frames—a prerequisite for learning statistical spaces from 4D sequences. Practical pipelines also benefit from explicit eye modeling to avoid attributing eyeball geometry to eyelids during fitting, which otherwise yields artifacts and photometric residuals in the periocular region [126]. Recent model families integrate the above ingredients—large, diverse identity scans; sequence-driven expression spaces; localized parameterizations; and robust co-registration—to balance genericity with per-subject fidelity. Compared to earlier tensor or residual schemes bound to discrete

expression sets, sequence-trained expression spaces can better capture the continuous nature of facial motion observed in 4D scans, while still retaining a compact parameterization amenable to optimization and animation.

2.1.7 Face probabilistic model

Single-view 3D head reconstruction [127] is an ambiguous problem due to the fact that training data usually have limited variation in poses, particularly in face monocular videos. Recently, diffusion models have been employed for conditional novel view synthesis [128] and also multi-view synthesis [129]. Since the results usually have ambiguous geometry, the output rendered results can exhibit noticeable artifacts, particularly a lack of texture details in unseen views. This can be mitigated by distilling prior knowledge from a 2D model [7, 130]. Diffusion further expands the toolbox: latent diffusion for images/videos provides strong priors and scalable decoders [131, 132, 133, 134]; 2D-diffusion–based 3D/4D synthesis recovers shape/dynamics via score distillation and trajectory or hierarchical priors [135, 136, 137, 138]. Multi-view diffusion generates consistent view sets for reconstruction [139], enabling "generate-and-reconstruct" workflows such as CAT3D [140]. For single-image heads, Morphable Diffusion conditions on 3DMM signals to yield controllable, 3D-consistent multi-view imagery [141].Existing 3D reconstruction works found that distilling knowledge from 2D images could help to make the 3D representation much more controllable by reconstructing a geometry at every step of the denoising process [142]. Other works pre-train a robust reconstructor [143] and use a 3D prior [144] which can be used in an image-conditioned auto-decoding framework. However, their work is complex and computationally heavy to train. We also leverage a pretrained 2D generative prior when training for 3D reconstruction; this helps our method with extreme-view 3D head reconstruction, but avoids expensive iterative sampling.

2.1.8 Face Datasets

One of the biggest limitation for face datasets, is that in the wild data usually don't cover full head viewpoints, but mostly are frontal viewpoints, therefore existing 3D face reconstruction models usually fail to reconstruct the back view of the head. We introduce the commonly used datasets in face domain in this section. Most of face datasets are face-centric and downloaded from YouTube or staged recordings, offering many identities with emotion labels; many include audio–video (VoxCeleb1/2 [60, 145], LRW, LRS2 [146, 147], CREMA-D [148], MEAD [149], VFHQ [150], TalkingHead-1KH [151], Obama [152], while others are images only, e.g. CelebA [153], CelebAMask-HQ [154]. They differ mainly in amount of identities, modalities, labels, resolution and quality. CelebV-HQ [155]

Dataset	Condition/Modality	# of identities	# of male Speakers	# of videos	# of clips	Video resolution	# of hours	# of utterances	Attributes	License	Note	Acquisition
VoxCeleb1	Audio/ Youtube Video (is not publicly available but could manually download (Official one only process with Ilp region))	1251	690	22,496	N/A	N/A	352	153,516	N/A	Yes	N/A	Only Audio part
VoxCeleb2 *	Audio/ Youtube Video	6112	3761	150,480	N/A	N/A	2442	1,128,246	Nationality Gender	Yes	N/A	Yes
LRW *	Audio/ Video/ Text(words in sentences)	1000+	N/A	N/A	N/A	N/A	N/A	1000+	Words	Yes	N/A	requested
LRS2 *	Audio/ Video/ Text (words in sentences)	N/A	N/A	N/A	N/A	N/A	195	8M	Words	Yes	N/A	requested
CREMA-D *	Audio/ Video	91	48	N/A	7442	N/A	N/A	4715	6 emotions: Anger, Disgust, Fear, Happy, Neutral, Sad (four emotion level: Low Medium, High and Unspecified)	Yes	races and ethnicities	Yes Git Ifs Downloaded
MEAD	Audio/ Video high-quality	60 actors	N/A	N/A	281,400	1920 × 1080	Avg 38min 57s	N/A	8 emotions: angry disgust contempt fear happy sad surprise neutral (three emotion level: Weak, Medium, Strong)	No, but a published paper ECCV	Only 48 actor part 1 released	Baidu Drive Google Drive
CelebA	Face images	10,177	N/A	N/A	N/A	N/A	N/A	N/A	202,599 face images 5 landmark locations, 40 binary attributes annotations per image	No, Has an agreement, Seems have an unoffical license.	http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	Website
CelebAMask-HQ	face images	N/A	N/A	N/A	N/A	N/A	N/A	N/A	With size of 512 x 512 and 19 classes including all facial components and accessories such as skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat, eyeglass, earring, necklace, neck, and cloth	No, Has an agreement, Seems have an unoffical license.	For more versions http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	Downloaded
CelebV-HQ	Video	15,653	N/A	N/A	35,666	512x512 at least	N/A	N/A	40 appearance manually labeled(narrow eyes, Big lips, etc) 35 action(eat, drink, etc) 8 emotions(Netural, Happy, Sadness, Anger, Fear, Surprise, Disgust, Contempt)	No, ECCV 2022 An agreement in github.	Image and video quality is better than Voxceleb Action unit smoothness is better than Voxceleb	Official script Downing /Baidu drive(github issue) Google drive(github issue)
VFHQ	Youtube Video	N/A	N/A	7228	over 16,000 high-fidelity clips	at least 700 x ?	N/A	N/A	Nationality 54% Female and 46% male Motion diversity (pixel displacement) large motion, middle motion and slow motion	Yes In an agreement https://liangbinxie.github.io/projects/vfhq/license.txt	Large dataset with 4T Resized 512x512	Baidu Drive One Drive
TalkingHead-1KH	Video	N/A	N/A	180K videos	N/A	N/A	about 1000 hours	N/A	N/A	Yes In the paper it says yes, no public download page. The github 404. So cannot find the lisence.	Collected large-scale talkinghead video dataset. It contains 180K videos, which are often with higher quality and larger resolution than those in VoxCeleb2.	Need to email the author. No public link found
Synthesizing Obama: Learning Lip Sync from Audio	Audio/ Video	1	1	15	N/A	N/A	N/A	N/A	N/A	NO SIGGRAPH 2017	N/A	Manually Download

Figure 2.4: Conclusion about face dataset [5]

and CelebAMask-HQ [154] are high quality datasets. High-quality videos or images, e.g., HDTF [61], CelebV-HQ, and CelebAMask-HQ [155, 154] has shown more geometric detail and mouth teeth fidelity. MEAD [149], RAVDESS [36], MSP-IMPROV [35], CREMA-D [148] contain more emotion annotations. LRW, LRS2, HDTF, TalkingHead-1K and VoxCeleb show reliable alignment of audio–video sequences, thus they are better for lip Synchronization. Semantic masks in CelebAMask-HQ [154]) are quite precise, and there are diverse poses and motions are shown in Voxceleb, LRS, VFHQ and TalkingHead-1KH [60, 145, 147, 150, 151].

All these datasets various in many aspects, they commonly offer many identities and in-the-wild captures, which helps train more generic generative models. For 2D talking head modelling, LRW, LRS2, Talkinghead-1KH, and voxceleb2 [146, 147, 151, 145] are preferable, they typically do not include emotion labels, their alignment quality benefits speech-conditioned generation. In this thesis, for the human emotion recognition task, we employ RAVDESS, MSP-IMPROV, and CREMA-D because they provide well-annotated, acted emotions collected in controlled lab environments, leading to high label reliability. For the talking-head generation task, we use monocular datasets, i.e. HDTF and TalkingHead-1K since they present a more challenging, limited-viewpoint setting with a larger number of identities captured in the wild.

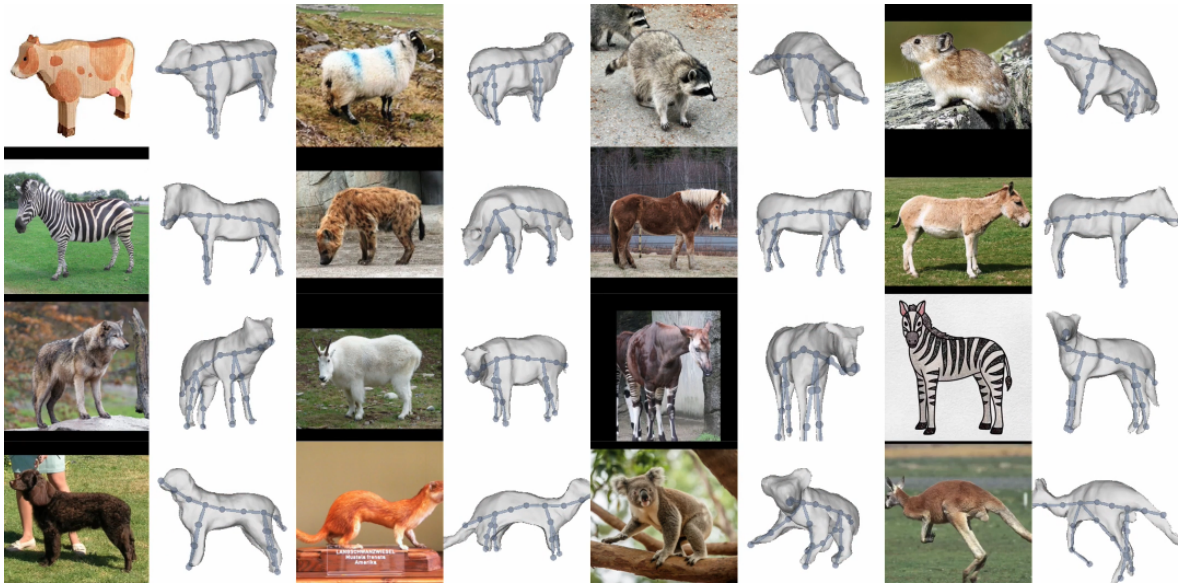

Figure 2.5: Example for single-mage 3D animal reconstruction reproduced from 3D-Fauna [6]

2.2 Reconstructing Articulated Animals

2.2.1 3D Reconstruction of Animals and priors

Both 3D head reconstruction and articulated animal reconstruction benefit from strong priors for accurate geometry. Unlike the human-face domain, animals exhibit complex, varied body topologies, which makes learning substantially harder. As shown in Figure 2.5, body topologies vary across animal species. In practice, 3D priors for animals are typically provided by SMAL [156] rather than FLAME [157], which helps for generate articulated animals. Early approaches therefore fit pre-defined parametric SMAL [156] model to single images using 2D keypoints and silhouettes, and later extend this paradigm to multi-view settings [158]. A complementary line of work directly optimises category shapes from small image or video collections, augmenting mask supervision with additional cues such as keypoints [159, 160], self-supervised semantic correspondences [161, 3], optical flow, surface normals [162] , and category-specific templates [163].

Practically, these methods rely on differentiable rendering and carefully designed objectives that balance data terms and priors. Typical losses include keypoint reprojection and silhouette consistency, mask IoU terms, flow- and normal-consistency across frames or views, as well as geometry regularizers such as Laplacian/ARAP smoothness, symmetry constraints, and articulation consistency around a kinematic skeleton [68–71, 74–76]. Many pipelines also estimate or refine camera parameters jointly with shape and pose, enforce multi-view cycle consistency, and use temporal smoothness to stabilize video-based optimization [68–71]. When available, category templates (or learned morphable families)

constrain the search space and improve identifiability under heavy occlusion and background clutter [6, 58, 80, 81]. Despite impressive progress, persistent challenges include large intra-class variability, extreme articulation, and scale–pose ambiguity from single views. Robust correspondence learning [74–76], stronger motion cues from long videos [68–71], and richer anatomical priors continue to close the gap, showing that careful integration of weak 2D signals with lightweight shape priors can mitigate the scarcity of ground-truth 3D annotations for animals.

2.2.2 Animal Rigging

Research on animal rigging includes controllable skeletal–skin models for animation has evolved along two threads and is increasingly hybrid. Template-based methods build on parametric LBS rigs such as SMAL [156], transferring human-pose practices to quadrupeds and fitting shape/pose to images or videos; they benefit from dense pixel–mesh correspondences (e.g., CSE) [8] to overcome the front-view bias and sparsity of keypoints and to stabilize non-frontal views and temporal consistency. To further constrain optimization in casual videos, recent systems couple SMAL with Structure-from-Motion for camera separation and temporal regularization, and replace sparse joints with CSE-driven dense supervision to improve coverage of side/rear views. Meanwhile, template-free pipelines (e.g., LASR/ViSER/BANMo/RAC) learn canonical spaces and deformation fields from monocular or multi-video inputs, often achieving compelling view synthesis but sometimes sacrificing anatomical plausibility and mesh fidelity relative to rig-constrained models. Contemporary approaches increasingly hybridize these lines: SMAL-like kinematic structure anchors articulation; CSE fields provide dense correspondences for analysis-by-synthesis fitting; and implicit appearance modules (e.g., duplex/NRF-style textures) are deformed consistently with the rig to capture fine surface detail, yielding animatable avatars that track motion and texture in casual videos and outperform prior template-based (BARC/BITE) [1, 2] and template-free (RAC) baselines on COP3D [162] and APTv2 [164].

2.2.3 Learnable Deformable 3D Representations for Animals

Learning deformable 3D animal categories requires a representation that is both expressive enough to capture non-rigid shape variation and structured enough to remain identifiable from weak supervision. Early works rely on explicit triangular meshes with per-vertex offsets, regularized by geometric priors such as ARAP, and often paired with linear blend skinning (LBS) to model articulation [165, 166, 10, 167, 168]. Parametric animal models like SMAL constrain deformation to a low-dimensional manifold and provide rig-aware control, at the cost of requiring curated scans and limiting topology change [12]. To increase

flexibility while retaining control, cage- or bone-driven models and other low-dimensional controls have been explored [169, 168, 10]. In contrast, implicit neural fields (e.g., SDFs and NeRF) represent geometry volumetrically and have shown strong capacity for detail and topology variation [170, 171, 101, 172], and their dynamic counterparts (D-NeRF, Nerfies, A-NeRF) capture motion via time- or pose-conditioned radiance fields [173, 102, 174]. However, purely implicit approaches face practical issues when coupling articulation with rendering, notably the need to invert deformations from posed/world space back to a canonical field, which is harder than forward skinning [175]. Recent category-from-video systems (LASR, ViSER, BANMo) therefore combine optical flow/masks and engineered objectives to optimize animatable shapes; while effective, they can be brittle and computationally heavy [160, 161, 176]. Hybrid designs that marry implicit fields with explicit meshes offer a promising middle ground: an SDF in canonical space is converted to a mesh (e.g., via DMTet) for efficient posing and differentiable rendering, keeping strong priors on articulation while retaining implicit expressivity [177]. MagicPony [163] exemplifies this trend by learning a category-level implicit–explicit prior and articulating an extracted mesh for analysis-by-synthesis training from single images, showing that hybrid SDF–mesh representations alleviate optimization instabilities and reduce supervision compared to purely mesh- or NeRF-based pipelines [163]. *Animal Avatars* similarly demonstrates that, in casual videos, SMAL-based LBS with dense correspondence and photometric losses yields riggable avatars, highlighting the benefit of rig-consistent deformation spaces when supervision is limited [178]. Overall, the field is converging on learnable deformable representations that integrate (i) explicit articulation priors (rigs, skeletons), (ii) implicit fields for geometry/appearance, and (iii) correspondence- or feature-based supervision, achieving animatable, re-targetable reconstructions with reduced reliance on 3D ground truth [163, 176].

Chapter 3

Detail-Enhanced Intra- and Inter-modal Interaction for Audio-Visual Emotion Recognition

We now introduce our method for understanding human expression, corresponding to our first task, human emotion recognition. In this task, emotion recognition is approached through apparent emotion, namely the observable affective signals conveyed by facial expression and vocal cues. Since our model only has access to visual and audio cues, the task is to infer emotion labels from these observable signals rather than to directly access the subject's latent internal state. The setting that given a face image and its corresponding audio segment as input and outputs the predicted emotion class. We describe the task in detail in the following sections.

3.1 Introduction

Capturing complex temporal relationships between video and audio modalities is vital for Audio-Visual Emotion Recognition (AVER). However, existing methods lack attention to local details, such as facial state changes between video frames, which can reduce the discriminability of features and thus lower recognition accuracy. In this paper, we propose a Detail-Enhanced Intra- and Inter-modal Interaction network (DE-III) for AVER, incorporating several novel aspects. We introduce optical flow information to enrich video representations with texture details that better capture facial state changes. A fusion module integrates the optical flow estimation with the corresponding video frames to enhance the representation of facial texture variations. We also design attentive intra- and inter-modal feature enhancement modules to further improve the richness and discriminability of video

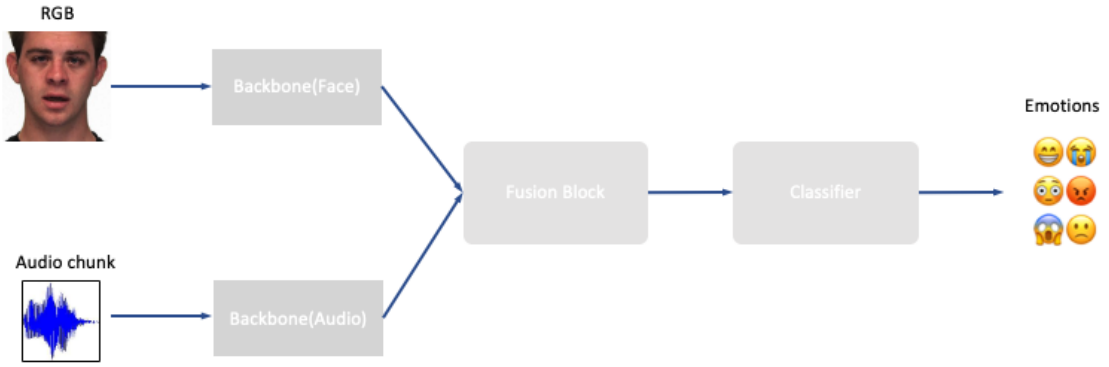

Figure 3.1: Setting for our task(AVER): Given one face image and corresponding audio chunk, emotion is classified based on fusion method

and audio representations. A detailed quantitative evaluation shows that our proposed model outperforms all existing methods on three benchmark datasets for both concrete and continuous emotion recognition. To encourage further research and ensure replicability, our project code is public avaliable at https://github.com/stonewalking/DE-III.

The basic setting of our task is that giving one face image and one audio chunk, we fuse the feature extracted from them, and regress the features for final classification, as shown in Figure 3.1. Emotion perception is attracting ever-increasing research attention due to its wide range of applications, such as affective computing [179], human-computer interaction [35], and social robotics [180]. Multi-modal emotion recognition, especially integrating audio and video (i.e. AVER), is particularly important since it makes use of the information present in two modalities that are vital to human communication. Unlike single-modal emotion recognition, multi-modal emotion recognition has access to different representations of the same emotion from different modalities. This improves feature representation capabilities and distinguishability, leading to improved recognition accuracy [181, 182].

However, there are still two challenges that are the focus of ongoing research in AVER: (i) how to enhance the representation of fine details within modalities, such as tiny details of facial motion (e.g. due to micro-expressions), and (ii) how to better leverage inter-modal associations to fully exploit the complementary information from different modalities. Solving both will enable learning better feature representations, and improve emotion recognition accuracy.

When learning features from one modality, intra-modal temporal relationship mining [183, 184, 185] and feature detail enhancement [186] are important ways to make features more discriminative. For instance, [185] proposed an adaptive graph attention network to explore the relationship between frames of videos for micro-expression recognition, while [186] introduced optical flow to replace face images for micro-expression recognition based on a multi-scale feature representation. However, these methods focus on the single-modal

setting, and cannot exploit information from multiple modalities. [183] used self-attention [187] within each modality to enhance their representation and then fused them by a linear-based function to classify; however this cannot fully account for the complex, nonlinear relationships between audio and video.

Multi-modal approaches have recently become mainstream [188, 189, 190] since considering both audio and video further improves representations, by fusing information in associated video frames and audio fragments. For example, [182] explored the effectiveness of different variants of transformer-based inter-modal attention mechanisms for AVER and showed inter-modal interaction can significantly improve performance. However, although inter-modal interaction improves recognition, these methods do not investigate modeling temporal relationships within each modality. [191] adopts a multi-branch joint auxiliary training method, designing independent audio and video branches and multi-modal fusion to enhance feature relationships, which greatly improves recognition performance. [192, 193] used a network shared across modalities to encourage consistency of the multi-modal feature space. However, since different modalities have different feature distributions and properties, a shared network may not fully capture the unique characteristics of each modality, resulting in information loss.

Most of the relationship modeling strategies mentioned above [194, 195, 192, 193] model temporal relationships based on implicit appearance representation of video frames and audio fragments, but ignore an inherent challenge of AVER – that in video features the frame-to-frame variations of faces are much weaker than in audio. For example, there may be significant changes in content and intonation between two audio fragments, while there is little difference between video frames. It is clear that these missing explicit details, especially state changes between face frames of videos, may lead to reduced discriminability of feature representations during the relationship modeling process, thereby affecting the accuracy of AVER.

We address these issues by introducing a multi-modal interaction network (Figure 3.2) that incorporates an explicit representation of visual detail changes between frames, and which can better fuse the complementary information from video and audio. Different from methods that directly model relationships between local regions of a facial sequence [196, 197, 198, 199], optical flow is a simple and effective way to represent the state changes between the facial frames. Optical flow can enhance the discriminability of visual representations by directly highlighting significant detail differences between frames, especially those texture changes that can express facial emotions [186]. To this end, we propose a novel detail-enhanced intra- and inter-modal interactions network (called DE-III) for AVER, which integrates explicit optical flow information into an end-to-end multi-modal interaction framework. In addition, two independent multi-modal interaction fusion mechanisms and multiple residual connections further alleviate the information loss problem in existing

shared interaction strategies [192, 193]. Our main contributions are as follows:

- we explicitly capture detail changes between video frames using optical flow, and integrate this information using a lightweight attentive fusion module;
- we design novel detail-enhanced intra- and inter-modal interaction modules for the video and audio modalities, which can effectively fuse associated information of one modality into the other modality and reduce information loss by residual connections.

We evaluate the resulting model and several variants on three widely used benchmarks and obtain highly competitive results including a new state-of-the-art on multiple metrics, e.g. 83.7% F1-Micro score on CREMA-D, 82.7% accuracy on RAVDESS and the highest scores on MSP-IMPROV with 89.3%, 88.7% and 85.8% for valence, arousal and dominance.

3.2 Related Work

Emotion recognition has received a significant amount of attention in the computer vision community. Numerous methods [200, 201, 202, 203] have been proposed to solve this task by using different data modalities, such as images, speech and text. These methods can be divided into two main kinds: unimodal methods (that input just one modality), and multimodal methods (that input two or more modalities). Our proposed DE-III belongs to the latter category, combining audio and video modalities to improve the performance of emotion recognition.

3.2.1 Unimodal Emotion Recognition

Unimodal emotion recognition methods [200, 204, 201, 202, 203] focus on application scenarios where only one kind of data is available; they design feature enhancement and interaction methods based on the inherent properties of the corresponding modality. The most common methods are text-based [205, 206, 204] and image-based [200, 201, 202]. For example for text, [204] present a BERT-based model to explore the importance of context extraction in texts for emotion recognition. One work by [207] proposed one sequence-based convolutional neural network to detect human emotion from big data. However, it is harder to to accurately predict human emotions from a text transcription compared to using richer modalities such as images or videos. For image data, [202] proposed feature decomposition and reconstruction learning for effective facial image expression recognition. [208] introduced the image depth information to improve the context information of images, which improved the representation capability and thus recognition accuracy. Moving to video, [209]

introduced facial micro-expression analysis methods that can improve emotion recognition by capturing richer contextual sequence information than static images. Although unimodal emotion recognition has achieved substantial progress and delivers promising results, it is inherently limited by having less information available than multi-modal approaches.

3.2.2 Multi-modal Emotion Recognition

Recently, multi-modal emotion recognition has become mainstream [210, 211, 191, 195, 193, 192, 182, 212, 181] due to its ability to fully exploit the complementary information present in different modalities. For instance, [182] explored the effectiveness of different variants of transformer-based inter-modal attention mechanisms for audio-video emotion recognition and showed that inter-modal interaction can significantly improve performance. [213] showed that combining audio and with a corresponding text transcription improves the representation ability of features, since audio captures details of intonation, while text captures semantics more explicitly. Moreover, [211] fused three modalities (audio, text and vision), further improving recognition accuracy. The above works indicate that combining multiple modalities can significantly enhance the discrimination ability of fused representations and thus the recognition performance. In this work, we study multimodal-based emotion recognition, specifically for audio-video emotion recognition (AVER). The most similar works to ours are [192, 193], both of which used a transformer-based architecture that is shared across video and audio modalities to encourage consistency of the multi-modal feature space. However, their proposed shared network cannot fully capture the unique feature distributions of each modality, such as explicit facial state changes between face video frames, resulting in the loss of information during the multimodal relationship modeling process. Unlike [191, 212, 181], which adopt attention-based neural network to effectively process and integrate audio modalities, our model not only learns the intra-relationships within video feature representations but also models the inter-relationships when attentively fuses the audio representation. Our proposed model augments video features with optical flow information before fusing with the audio features. Unlike traditional methods [214] that directly combine the optical flow features with visual representations, we use Conformer [195] networks to extract context-aware features, and design a novel pairwise O-V attention fusion module to combine them.

3.3 Method

The overall framework of our proposed model DE-III is shown in Figure 3.2. We first extract video and audio features, then enhance their representative power through temporal relationship modelling within their respective modalities, also fusing optical flow information with

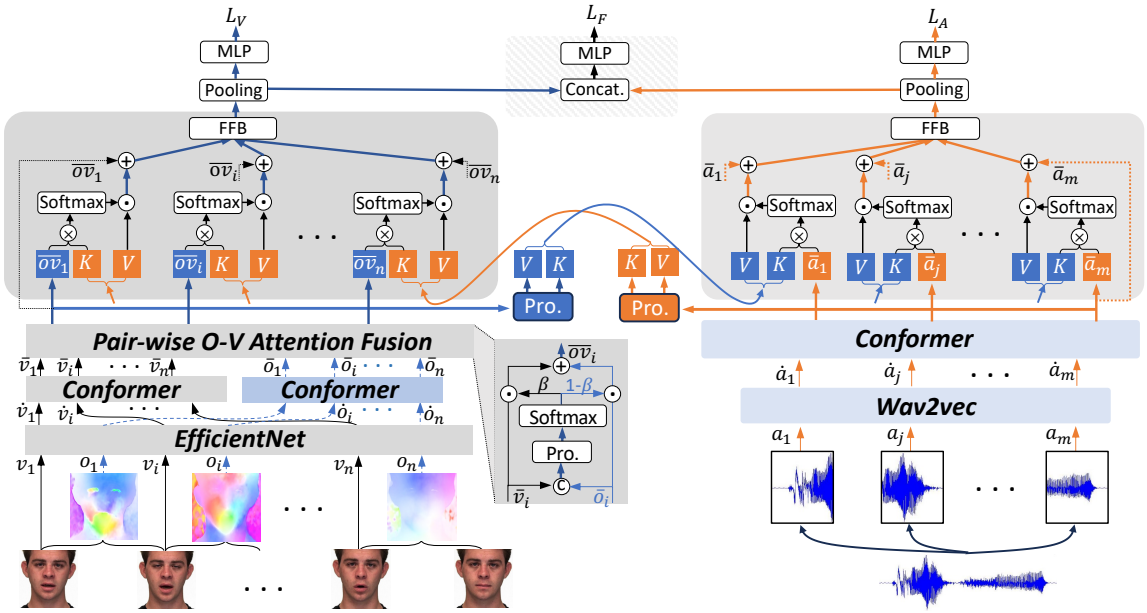

Figure 3.2: Overview of our proposed method *DE-III*. Given video frames v_i and audio fragments a_i, we extract features and pass these through separate Conformer encoders. We introduce explicit information about facial motions – captured by optical flow o_i – to enhance video feature representations, with a new pair-wise O-V attention fusion module that effectively integrates the information from optical flow and video frames. We propose an inter-modal feature enhancement module (large boxes near top) to attentively fuse the associated audio and video representations in both directions, i.e. audio-to-video and video-to-audio. During training, the final emotion predictions are calculated independently from three sets of features: the video features albeit with audio information fused (i.e. without the model components in the chequered box); the converse using the audio features; and finally using both sets of features after a further fusion stage. During inference, we use the prediction head that performed best on validation data.

the video features to better capture detail changes. Then, the inter-modal feature enhancement module performs attention-weighted fusion of each modality's information with the other modality.

3.3.1 Audio Self-enhancement Module

To represent the information in audio, we use a pre-trained wav2vec model [215] to embed the extracted audio fragments.The original speech audio is resampled at 16 kHz. Specifically, we split a given audio clip into a sequence of m fragments $A = \{a_1, a_2, \ldots, a_m\}$ using a sliding window. Then we use the wav2vec-large-robust model to extract corresponding fragment-level representations $\dot{A} = \{\dot{a}_1, \dot{a}_2, \ldots, \dot{a}_m\}$. Next, a Conformer encoder [195] (a transformer-based model with convolutions to improve temporally-local information processing) is used to obtain enhanced audio-fragment representations $\bar{A} = \{\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_m\}$ that account for (intra-modal) local and global temporal relationships.

3.3.2 Video Pairwise Attention Enhancement Module

Different from the audio features where contextual semantics are clear, i.e. there is clear semantic content and significant intonation changes, in the video, subtle yet important changes in facial texture tend to be lost during feature extraction. We therefore use a pre-trained optical flow model [216] to extract the flow o_i between adjacent pairs of video frames $\{v_{i-1},\, v_i\}$, where $i \in \{1,\, \ldots,\, n\}$ and n is the number of video frames; this can explicitly represent fine-grained changes of facial texture such as micro-expressions. In our implementation, the optical flow is estimated as a 2D vector field and then converted into a three-channel RGB image using an HSV-based encoding. Specifically, motion direction is mapped to hue, normalized motion magnitude to saturation, and the value channel is fixed. Then, we employ the widely-used EfficientNet-B2 model [179], which has been fine-tuned on VGGface2 [217] dataset, to extract representations for video frames and their corresponding optical flow maps; we denote these features by $\dot{V} = \{\dot{v_1},\, \dot{v_2},\, \ldots,\, \dot{v_n}\}$ and $\dot{O} = \{\dot{o_1},\, \dot{o_2},\, \ldots,\, \dot{o_n}\}$ respectively. To further enhance the representational ability of these visual features, we use two independent Conformer encoders [195] to embed them into the same dimensional space as the audio modality. This also allows for subsequent inter-modal interaction. We next propose a simple and efficient pairwise O-V attention fusion module to combine the features of frames and optical flow into a joint embedding space. Specifically, we use a fully-connected (FC) layer to map the features at each time-point to two channels, then apply a softmax function [218] and interpret these values as weights for the frame and flow features respectively. We finally obtain the detailed-enhanced video representation $\overline{ov}_i$ by a weighted sum of linearly-projected frame features and corresponding flow features. Thus, we set

$$[\bar{o}_i : \bar{v}_i] = [\mathrm{Conformer}(\dot{o}_i) : \mathrm{Conformer}(\dot{v}_i)], \tag{3.1}$$

$$(\beta_o, \beta_v) = \mathrm{softmax}(FC([\bar{o}_i : \bar{v}_i])), \tag{3.2}$$

$$\overline{ov}_i = \beta_o W_o \bar{o}_i + \beta_v W_v \bar{o}_i, \tag{3.3}$$

where $[:]$ denotes concatenation along the channel dimension, W_o and W_v are the linear projection parameters, and $\beta_o + \beta_v = 1$. We refer to the two conformers followed by the OV-fusion as the pair-wise attention enhancement (PAE) module. Compared with simple concatenation, our pair-wise O-V attention fusion adaptively selects the more informative channel. Since the softmax weights β_o and β_v are likely to often be close to 0 and 1, and the larger value is frequently pushed close to 1 while the smaller one is pushed close to 0, meaning that the fusion often primarily attends to either the optical-flow or the RGB feature rather than averaging them equally.

3.3.3 Inter-modal Feature Enhancement Module

Inspired by the attention mechanisms [190, 191], we next design an inter-modal feature enhancement module (IFE) that allows each modality to attend to the other and integrate relevant information. For simplicity we describe only the audio-to-video fusion (IFE-Video); however a similar approach is used for video-to-audio. We want to allow the enhanced video frame features $\overline{ov}_i$ to attend to features of relevant audio fragments $\bar{A} = \{\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_m\}$. Different from traditional self-attention [187] and cross-attention [181], we take the target video frame $\overline{ov}_i$ as the query to calculate the attention weights, with the audio fragment defining the keys and values after the linear projections. Attentive fusion from another modality allows relevant modality information to be extracted and integrated, thereby improving the distinguishability of target modality representation. Finally, we obtain the video representations $\ddot{OV} = \{\ddot{ov}_i\}$ after IFE by adding a residual connection, and passing through a feed-forward block (FFB) which contains two linear layers. In summary, we set

$$s_{ij} = \frac{(W_{ov}\overline{ov}_i)(W_a\bar{a}_j)^T}{||W_{ov}\overline{ov}_i||\ ||W_a\bar{a}_j||} \quad \forall\ i \in \{1, \ldots, n\},\ j \in \{1, \ldots, m\} \tag{3.4}$$

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{j=1}^{m} \exp(s_{ij})} \tag{3.5}$$

$$\ddot{ov}_i = \sum_{j=1}^{m} \alpha_{ij}\bar{W}_a\bar{a}_j + \overline{ov}_i, \tag{3.6}$$

where W_{ov}, W_a and $\bar{W}_a$ are linear projection parameters. Similarly, we obtain the attention-aware video fragment representations of each audio fragment and combine them with an audio residual operation to give the final audio representations $\ddot{A} = \{\ddot{a}_1, \ddot{a}_2, ..., \ddot{a}_m\}$.

3.3.4 Feature Aggregation and Objective Function

Since we want to make a single prediction for an entire video, we max-pool the features along the temporal axis, yielding a video-centric feature vector $\ddot{ov}^*$ from $\ddot{OV}$, and audio-centric feature vector $\ddot{a}^*$ from $\ddot{A}$ (note that $\ddot{ov}^*$ still incorporates information fused from the audio modality as described in Section 3.3, and vice-versa). Since the supervision is provided at the utterance level, pooling allows information from all frames and audio fragments to contribute to the final representation, rather than assigning the training signal to only a single time step. We use three independent emotion prediction heads (each a multi-layer perceptron) with corresponding losses to jointly optimize different branches the model – the video-cross loss L_V (using $\ddot{ov}^*$ as input to the MLP), audio-cross loss L_A (using $\ddot{a}^*$) and audio-visual fusion loss L_F (using $\ddot{ov}^*$ concatenated with $\ddot{a}^*$). The overall objective function is the sum of the three losses. We use multi-class cross-entropy for datasets with discrete emotion class labels, and concordance correlation coefficient (CCC) for datasets with continuous labels.

Table 3.1: Comparisons with state-of-the-art methods for AVER on CREMA-D, MSP-IMPROV and RAVDESS (in %). The best results are bold and second-best underlined.

Method	CREMA-D		MSP-IMPROV			RAVDESS
	F1-Macro	F1-Micro	Val.	Aro.	Dom.	Acc.
Multi. [181]	64.4	69.2	<u>77.5</u>	<u>76.1</u>	77.8	78.5
MMER [182]	–	–	–	–	–	<u>81.6</u>
UAVM [192]	74.9	76.9	47.1	54.4	68.7	–
AuxFormer [191]	69.8	76.3	67.2	65.2	<u>82.0</u>	–
LADDER [212]	**80.2**	<u>80.3</u>	–	–	–	–
DE-III (ours)	<u>79.5</u>	**83.7**	**89.3**	**88.7**	**85.8**	**82.7**

Specifically, CCC is given by

$$\mathcal{L}_{\text{CCC}} = 1 - \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3.7}$$

where μ_x and μ_y are the mean of the predicted result $\hat{y}$ and the label y, respectively, σ_x and σ_y are their standard deviations, and ρ is their Pearson correlation coefficient (a ρ value close to ± 1 suggests a strong linear relationship, while a value of 0 signifies the absence of any linear correlation). During inference we can use predictions from any of the three heads; for our main experiments we use the prediction head that performed best on the validation data.

3.4 Experiments

3.4.1 Experimental Setup

Datasets and Metrics.

To verify the effectiveness of our proposed approach, we evaluate it on three popular AVER datasets: CREMA-D [219], MSP-IMPROV [35] and RAVDESS [36]. CREMA-D consists of 7,442 facial videos with corresponding audio from 96 participants (48 male, 48 female). Each audio-video clip is labeled with one of 6 concrete emotion classes – anger, disgust, fear, happiness, sadness, and neutrality. RAVDESS consists of 2,880 videos from 24 actors, each enacting eight concrete emotional states. MSP-IMPROV consists of 8,385 audio-video clips from 12 participants (6 male, 6 female) with each clip labeled by both concrete emotional states and continuous emotional states – valence, arousal and dominance; following previous works [181, 182, 192, 212] we use only the continuous labels. We adhered to the protocol in [182, 191], with 5 separate folds where each fold divides the data into training, validation, and test sets with non-overlapping actor identities. We evaluate based on the most commonly-used metrics for each dataset – F1-Macro and F1-Micro for CREMA-D [219],

Accuracy for RAVDESS [36] and CCC for MSP-IMPROV [35].

Implementation Details.

All models were trained for up to 20 epochs using early stopping on the validation set, and we report our results on the test set. We choose hyper-parameters based on validation set performance. We use AdamW for optimization with a learning rate of 5×10^{-6} and weight decay of 5×10^{-2}. The face images are extracted from each frame of every video clip and resized to 224×224 pixels. We generate optical flow maps using [216] and normalize their magnitude by a standard deviation calculated from the local optical flow magnitude at every pixel position within an entire video clip. We use the pre-trained EfficientNet-B2 from [179] to extract features from the video frames and optical flow maps. The audio features are extracted using wav2vec2-large-robust [215]. Separate Conformer encoders for video and audio map the extracted features to vectors of 1408-dimension each. Each Conformer block has a hidden dimensionality of 512, with 8 attention heads. The number of blocks in the acoustic, visual, and optical flow Conformers were set to 3, 3, and 2, respectively. For the prediction heads, we use MLPs with hidden dimensionality of 512. Our IFE module (Section 3.3) uses single-head attention [187] with the linear feed-forward block and the highlighted fusion feature dimensions remain unchanged. All models are trained and evaluated across five folds on each dataset with five different random seeds, and the reported results are averaged over all folds and seeds.Our model was implemented in PyTorch and trained on 2 NVIDIA RTX A5000 GPUs, taking 1 hour.

3.4.2 Quantitative Comparison

In Table 3.1 we present quantitative results for our method and several existing works: 1) Multi [181], a transformer-based cross-modal attention fusion method; 2) MMER [182], with multiple self-attention fusion mechanisms; 3) UAVM [192], a transformer-based feature enhancement model with a shared audio-visual encoder; 4) AuxFormer [191], a transformer framework with two independent auxiliary branches; 5) LADDER [212], a transformer-based cross-attention framework with auxiliary reconstruction tasks. We see that compared with the previous best method LADDER [212] on CREMA-D, our DE-III achieves higher performance in terms of F1-Micro score, 83.7% vs. 80.3%. On MSP-IMPROV, our DE-III attains excellent CCC values of 89.3% for valence (Val.), 88.7% for arousal (Aro.), and 85.8% for dominance (Dom.), establishing a new state-of-the-art for this dataset. Moreover, we also achieve a better accuracy (Acc.) score on RAVDESS compared with the SOTA method, 82.7% for DE-III vs. 81.6% for MMER.

Table 3.2: Effectiveness of our inter-modal feature enhancement module (IFE), evaluated on CREMA-D.

Method	Cross attention		Accuracy	
	A-cross	V-cross	F1-Macro	F1-Micro
IFE-Fusion	✓	✓	77.2	82.2
IFE-Audio	✓	✗	78.3	82.2
IFE-Video	✗	✓	**79.5**	**83.7**
None-IFE	✗	✗	75.8	78.6

3.4.3 Ablation Studies

In this section, we evaluate the performance benefit due to various components and design decisions in our model.

Effects of Inter-modal Feature Enhancement (IFE).

In the IFE block, we define video attending to audio as V-cross and audio attending to video as A-cross. We first experiment with removing the IFE module (i.e. without any inter-modality fusion, only RGB images and flow maps, denoted None-IFE). In Table 3.2, we see a large performance drop in this setting – compared with the best output (from IFE-Video), the F1-Macro and F1-Micro scores decrease by 3.7% and 5.1% on the CREMA-D test set, respectively. This suggests that inter-modality fusion plays an important role in improving AVER capabilities. Recall that our model has three prediction heads: IFE-Audio (i.e. using features $\ddot{a}^*$), IFE-Video (i.e. using $\ddot{o}\ddot{v}^*$) and IFE-Fusion (i.e. using their concatenation). While the main results use IFE-Video at inference time, we also report results from the others in Table 3.2. IFE-Video achieves the best AVER performance, 79.5% F1-Macro and 83.7% F1-Micro. The other prediction heads achieve slightly lower though still competitive results.

Effects of Video Pairwise Attention Enhancement (PAE) Module.

To demonstrate our ablations on pair-wise attention enhancement (PAE) Module, we categorize different settings as ”Fuse when?”, ”Visual input”, ”sequential model”, and ”Fuse how?”. Results on CREMA-D are given in Table 3.3, all using the IFE-Video prediction head. We first present results when trained with only one part of the video information, i.e. RGB images only (IFE-V-F), or optical flow maps only (IFE-V-O). We see that IFE-V-O achieves 55.4% F1-macro and 64.9% F1-micro. The result shows optical flow information present low capability to distinguish emotions, and it is much weaker than using RGB images only. When combining optical flow maps with RGB images in the full model (IFE-Video), there is a remarkable performance improvement vs. IFE-V-F. It indicates that the

Table 3.3: Effectiveness of different approaches to inter-modal fusion within our model, evaluated on CREMA-D.

Model	Fuse when?		Visual input		Seq. model		Fuse how?			Accuracy	
	Early	Late	Flow	RGB	Conf.	Transf.	Concat	Sum	PAE	F1-Macro	F1-Micro
IFE-V-O	n/a		✓		✓			n/a		55.4	64.9
IFE-V-F	n/a			✓	✓			n/a		76.7	81.4
IFE-V-FOSC		✓	✓	✓	✓		✓			78.5	81.7
IFE-V-FODC		✓	✓	✓	✓		✓			77.8	82.6
IFE-V-FODS		✓	✓	✓	✓			✓		78.0	81.8
IFE-V-Early	✓		✓	✓	✓				✓	79.2	83.0
IFE-V-Trans		✓	✓	✓		✓			✓	77.9	82.6
IFE-Video		✓	✓	✓	✓				✓	**79.5**	**83.7**

flow maps indeed augment the video feature representations. Next, we replace our PAE with one single conformer followed by one OV-fusion block. To pass the image and optical flow features together into the conformer, we attempt several alternative operations– temporal concatenation (IFE-V-FOSC), channelwise concatenation (IFE-V-FODC), and summation (IFE-V-FODS). We see (Table 3.3) that our PAE module achieves the highest recognition performance, with 1.0% improvement over IFE-V-FOSC on F1-macro and 1.1% improvement over IFE-V-FODC on F1-Micro. These observations indicate that our PAE module is a more effective fusion method for combining visual features and optical flow features. Finally, we explore early fusion and late fusion strategies. We find that by moving OV-fusion block before the Conformer (IFE-V-Early), accuracy decreases slightly vs. having OV-fusion after the Conformer (IFE-Video), by 0.3% F1-Macro and 0.7% F1-Micro. We hypothesise that this is because the additional computation performed beforehand by the Conformer is beneficial in helping the OV-fusion module to determine whether to focus on image or flow information for each time-point. Additionally, we compare our method by replacing the conformer to the vanilla transformer [187], the accuracy decreases slightly by 1.6% and 1.1%, this demonstrates that the conformer is superior to the vanilla transformer at the image level in capturing changes in facial details from feature representations.

Effects of optical-flow extraction variants.

We next experiment with using different sliding window lengths and strides when extracting the optical flow from the videos. Firstly, we vary the window length while keeping the stride fixed to 1 (i.e. moving frame by frame). Secondly, we vary both the window length and the stride together (i.e. non-overlapping windows). The results in Table 3.4 show that using a window length of 1 with a stride of 1 performs best. Increased window lengths, with fixed or increasing strides, show consistent drops in performance, with the worst-performing variant having window length of 5 and stride of 1 (achieving 74.3% F1-Macro, versus 79.5% for

Table 3.4: Effectiveness of different feature extractors and frame-selection strategies for optical-flow, evaluated on CREMA-D for our IFE-Video model variant.

Feature extractor	Window	Stride	Accuracy	
			F1-Macro	F1-Micro
EfficientNet-B2 [179]	1	1	**79.5**	**83.7**
	3	1	76.1	81.4
	5	1	74.3	80.8
	7	1	75.2	81.6
	3	3	77.2	82.4
	5	5	76.2	80.7
	7	7	78.5	82.8
DINOv2 [23]	1	1	76.8	82.6

window length and stride of 1). This indicates that temporally-fine-grained information is valuable in increasing the accuracy of emotion recognition. We also experiment with using a different backbone feature extractor for the optical flow, since face images and flow-maps are quite different domains. We choose DINOv2 [23], which has been shown to be robust across many image domains, and fix the window length and stride to 1 (i.e. the best-performing setting). However, we find it performs worse than using EffcientNet pre-trained on a large face images dataset, dropping from 79.5% to 76.8% F1-Macro and from 83.7% to 82.6% F1-Micro.

3.4.4 Qualitative Analysis

To better understand the behavior of our model, we visualize the inter-modal fusion weights α_{ij} for IFE-Audio and IFE-Video (see Section 3.3) in Figure 4.10. The brightness of each location in the heatmap represents the strength with which the modality on the horizontal axis is attending to that on the vertical axis, at that particular time-point. The pattern of attention varies considerably for different points along the horizontal axis, showing that the model does not attend to fixed, specific points in the other modality, but adapts depending on the current features, and presumably the varying emotional states depicted in the video. Notably, the heatmaps do not exhibit a bright diagonal line; this indicates that time-points generally attend not to the corresponding time-point in the other modality, but to other (presumably relevant or informative) time-points. Overall these results suggest that our inter-modal feature enhancement module can selectively fuse the useful information from each modality into the other.

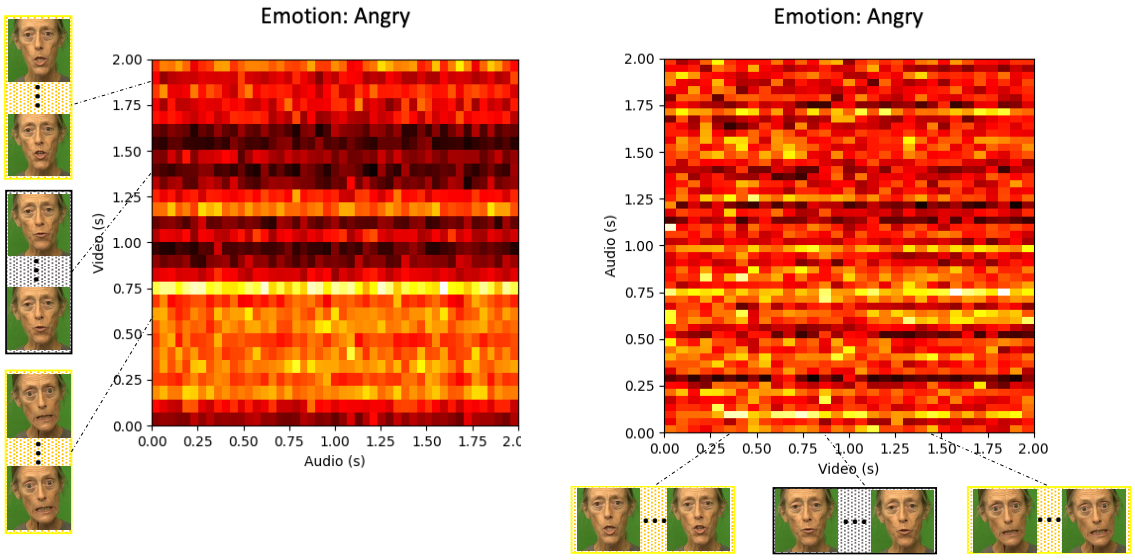

Figure 3.3: Heatmaps showing inter-modality attention weights calculated by IFE-Audio (left) and IFE-Video (right), for an example sequence with emotion 'angry'. The horizontal axis corresponds to time-points in one modality, which is fusing in information from the other modality on the vertical axis. Brighter colors indicate stronger attention to the time-point on the vertical axis, from the time-point on the horizontal axis.

3.5 Conclusion

We have presented a new model, DE-III, for audio-visual emotion recognition, which combines intra- and inter-model feature enhancement in a unified framework. DE-III introduces a pair-wise attention fusion method that integrates explicit facial detail changes between video frames, captured by optical flow. It not only improves the distinguishability of features within each visual modality, but also further increases the effectiveness of subsequent inter-modal feature interactions. Our results demonstrate that DE-III enhances emotion recognition by optimally fusing the information available in different modalities. Indeed, our model achieves state-of-the-art performance on three popular datasets, for both concrete and continuous emotion labels. This chapter addresses the task of interpreting human expressions, while the next chapter focuses on talking head generation, and facilitate reconstructing expressive human faces.

3.6 Future Work

In this section, we summarise the limitations of this work and propose future work to enhance understanding of human emotion states. In this task, we focus on multimodal modeling to better understand human emotional states through a deterministic model.

Despite being trained and evaluated on three benchmark datasets, our model still faces generalization challenges due to the limited number of training identities, which may lead to reduced accuracy on in-the-wild data. Further improvement could be achieved by training on larger-scale datasets, such as LRW [146] and LRS2 [147], that include more diverse identities and well-synchronized audio-visual data. We extract features from 2D images and audio cues, combine with optical flow feature for fusion [23, 22]. These features are implicit features, more explicit feature injected into the pipeline might benefit the overall accuracy, such as landmarks, action units and face correspondences feature. To make the fused representation more robust and semantically aligned, audio–visual contrastive pretraining [220, 221] can establish soft cross-time anchors before fine-tuning, and max-pooling over time can be upgraded to attention-based or multi-scale temporal aggregation for longer clips. Besides, it is also possible to inject 3D facial priors (3DMM/action units, head pose, lip dynamics) to capture micro-expressions [222]. Human emotion recognition has largely overlooked 3D representations. Moving forward, 3D-aware models should be optimized end-to-end with geometry-aware losses, validated on in-the-wild datasets.

Chapter 4

Splat-Portrait: Generalizing Talking Heads with Gaussian Splatting

In this Chapter, we tackle our second task, talking head generation. Unlike building a regression model for human expression classification, in this task, we aim to generate expressive and natural talking heads. The setting of it is that it takes one portrait image and its corresponding audio segment as input and generate 3D head and 4D talking sequences. We describe the tasks in detail in the following sections.

4.1 Introduction

Talking Head Generation (THG) aims to synthesize natural-looking talking videos from conditioning information such as driving speech [223, 50, 224, 225] or driving videos [226, 225, 92]. Previous 3D talking head generation methods have relied on domain-specific heuristics such as warping-based facial motion representation priors to animate talking motions, yet still produce inaccurate 3D avatar reconstructions, thus undermining the realism of generated animations. We introduce Splat-Portrait, a Gaussian-splatting-based method that addresses the challenges of 3D head reconstruction and lip motion synthesis. Our approach automatically learns to disentangle a single portrait image into a static 3D reconstruction represented as static Gaussian Splatting, and a predicted whole-image 2D background. It then generates natural lip motion conditioned on input audio, without any motion driven priors. Training is driven purely by 2D reconstruction and score-distillation losses, without 3D supervision nor landmarks. Experimental results demonstrate that Splat-Portrait exhibits superior performance on talking head generation and novel view synthesis, achieving better visual quality compared to previous works.

The generation of talking heads has received increasing attention due to its importance

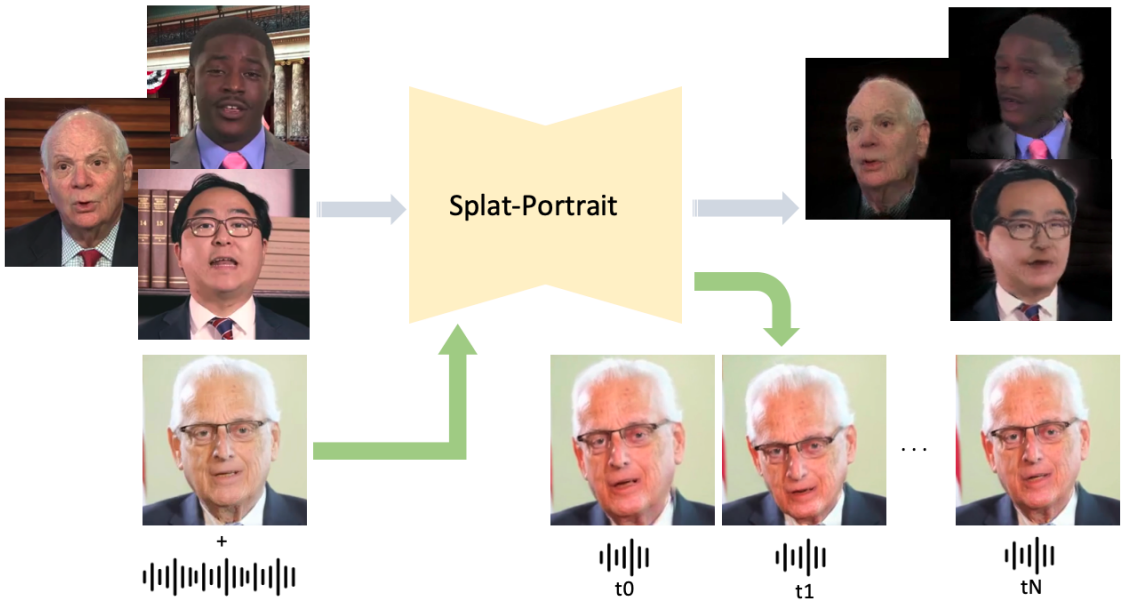

Figure 4.1: Settings of Splat-Portrait: given a single portrait image and a corresponding audio segment, we reconstruct a 3D head model and predict 4D splatting offsets the ensuing facial expression trajectory.

in various applications, including digital humans [227], virtual video conferencing [228] and visual dubbing [229]. We study the single-image setting: given a single portrait image and a corresponding audio segment, we reconstruct a 3D head model and predict 4D splatting offsets ensuing facial expression trajectory, which we render and composite to produce a talking-head video (see Fig. 4.1).

Recent 2D methods [50, 230, 231] have achieved significant improvements in video quality and achieve expressive animation results. However, such 2D methods struggle to generate views with large head pose variations, and are not guaranteed to output 3D-consistent renderings from different poses.

3D THG methods [225, 226, 92, 92] have attracted increasing attention in the past two years, since they simultaneously reconstruct accurate 3D geometry and generate expressive facial motions, allowing realistic and 3D-consistent portrait rendering from arbitrary user-controllable viewpoints. The majority of such methods focus on personal talking head generation [232, 233, 234, 235, 236], where they overfit a single person's head; they maintain realistic 3D geometry and preserve rich texture details. However, in this work, we consider the more challenging setting where we synthesise a 3D talking head given just a single 2D image, learning a model that generalizes across identities even without 3D supervision.

To represent 3D or 4D faces in THG, Neural Radiance Fields (NeRF) [11] or 3D Gaussian Splatting (3DGS) [237] are commonly used. NeRF-based methods often exhibit problems such as visual jitters, unsynchronized lip movements, and rendering artifacts; this is

because the implicit definition of NeRF entangles static facial geometry with dynamic motion, complicating simultaneous control of lip motions and 3D geometry reconstruction.

Other works [129, 238, 239] have explored 3D Gaussian Splatting (3DGS) [237] for 3D avatar generation in the person-generic setting. Compared to NeRF, 3DGS not only improves inference speed and visual quality, but is also more controllable due to its explicit point-cloud-based representation; this makes it possible to animate facial movements more directly and intuitively. For example, [233] drives Gaussian point clouds for facial motion using parametric 3D facial models [157], but it is still challenging to generalize to a person-generic setting. In addition, significant efforts have been made to design and improve 4D animation conditioned on driving information [225, 236, 238, 233]; the model reconstructs 3D geometry from a single portrait image and learns the corresponding facial motions. These motions are predicted from either an audio sequence or a video sequence driven by motion representation priors, e.g., PNCC, SECC and FLAME [240, 226, 157, 241]. These methods relax the difficulty of model training by injecting domain priors, e.g. 3D distillation [242] and motion driven priors, which can lead to unnatural results with limited 3D texture details.

Overall, existing works either reconstruct accurate 3D geometry but require multi-view inputs; or they learn it from monocular videos but yield inaccurate geometry.

In this work, we introduce Splat-Portrait, a novel audio-driven THG method based on 3DGS (see Fig. 4.5). Our method operates in the single-view setting—it reconstructs the 3D shape of the head directly from one image, outputting pixel-aligned Gaussian splats. To enable lip motion during speech, our model learns to directly animate these splats conditioned on an audio sequence.

Our approach is self-supervised from monocular videos only, and does not rely on 3D morphable models such as FLAME to represent facial shape and expression. We first train our model for static splat reconstruction on a large dataset without audio, then fine-tune on a smaller dataset of portrait videos to learn the correct splat dynamics. This strategy avoids 3D supervision, with the exception of easily-obtained approximate camera intrinsics and extrinsics. During the fine-tuning stage, to further improve the realism of extreme viewpoints that are rare in the training data, we adopt score distillation sampling (SDS) [7, 130], to extract knowledge from a powerful 2D diffusion prior [243].

Existing works [91, 226] typically model only the head region, or model the head and torso regions as a whole, while disregarding the background. This results in a video of a 'floating head', rather than a realistic video of the talking head in context. To address this, our model also predict a static RGB background image, and alpha-blend the rasterized splats over this. Driven only by the unsupervised frame-prediction loss, our model automatically learns to reduce the opacity of splats in the background region, and to inpaint the background even behind the head, resulting in realistic disocclusions when the head rotates.

In summary, our main contributions are as follows:

- A novel model architecture that disentangles a single portrait image into an accurate 3D splat representation of the head over an inpainted 2D background.
- Given audio sequences and corresponding time deltas, we show how to directly animate the 3D splats by predicting and adding dynamic offsets, without any complex motion representation such as a deformation model.
- A self-supervised training recipe that uses only monocular videos, without 3D supervision, and integrates knowledge from a strong 2D face prior, distilling its knowledge to improve reconstruction of extreme views.

Experimental results on the HDTF [61] and TH-1KH [244] datasets demonstrate that our approach yields higher video fidelity and quality compared with OTAvatar [92], HiDe-NeRF [226], Real3D-Portrait[225], and NeRFFaceSpeech [245] and GAGavatar[239].

4.2 Related Work

4.2.1 3D Head Reconstruction.

3D Gaussian Splatting (3DGS) [237] has emerged as a popular method for 3D head reconstruction due to its efficient rendering speed and superior reconstruction quality [233, 129, 246, 238]. Neural Radiance Fields (NeRF) [11] have been widely adopted for 3D talking head generation; NeRF represents scenes through volumetric radiance fields encoded by neural networks, enabling photorealistic renderings from novel viewpoints. NeRF-based methods have naturally extended into talking-head synthesis [232, 247]. Early NeRF-driven approaches for talking-head reconstruction [248, 92, 249, 247] often require subject-specific training, limiting their scalability. Recent methods leverage 3DGS to address these limitations by significantly improving rendering speed and depth estimation [237, 233, 234]. 3DGS represents scenes explicitly with discrete geometric primitives (3D Gaussians), enabling efficient optimization and real-time rendering. Notably, Rivero et al. [250] introduced a dynamic head reconstruction framework using 3DGS, and GaussianHead [246] further advanced these capabilities. By binding the Gaussians to an underlying geometric model, dynamic talking heads can be generated. However, these works for directly regressing 3D representations require prediction in a canonical space, which often fails to handle extreme head poses or significant appearance variations, such as non-photorealistic or animated scenarios. Current techniques still exhibit overfitting issues and rely heavily on domain priors during training, such as the parametric FLAME model [157]. Our method builds upon 3DGS to reconstruct dynamic talking heads directly from a single image.

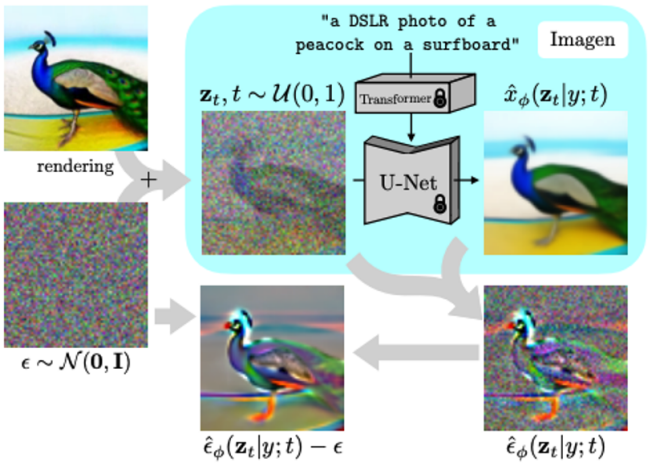

Figure 4.2: The pipeline of score distillation sampling from DreamFusion [7]

4.2.2 Probabilistic 3D Reconstruction.

Single-view 3D head reconstruction [127] is an ambiguous problem due to the fact that training data usually have limited variation in poses, particularly in face monocular videos. Recently, diffusion models have been employed for conditional novel view synthesis [128] and also multi-view synthesis [129]. Since the results usually have ambiguous geometry, the output rendered results can exhibit noticeable artifacts, particularly a lack of texture details in unseen views. This can be mitigated by distilling prior knowledge from a 2D model [7, 130]. This method is so called score distillation sampling, the idea is originally shared by [7], the typical pipeline of it is shown as Figure 4.2.

Existing 3D reconstruction works found that distilling knowledge from 2D images could help to make the 3D representation much more controllable by reconstructing a geometry at every step of the denoising process [142]. Other works pre-train a robust reconstructor [143] and use a 3D prior [144] which can be used in an image-conditioned auto-decoding framework. However, their work is complex and computationally heavy to train. We also leverage a pretrained 2D generative prior when training for 3D reconstruction; this helps our method with extreme-view 3D head reconstruction, but avoids expensive iterative sampling.

4.2.3 Face Animation.

Initial efforts for talking head animation utilized 2D approaches, employing generative adversarial networks, image-to-image translation [251] or diffusion models [252], to generate facial animations. Most 2D talking head generation methods design a mapping relationship between face images and audio feature. These methods [50, 253] often underestimate detailed individual differences. Recently, 3D facial animation methods [226, 225] became popular, however they adopt PNCC SECC as driving features, leading to unnatural expres-

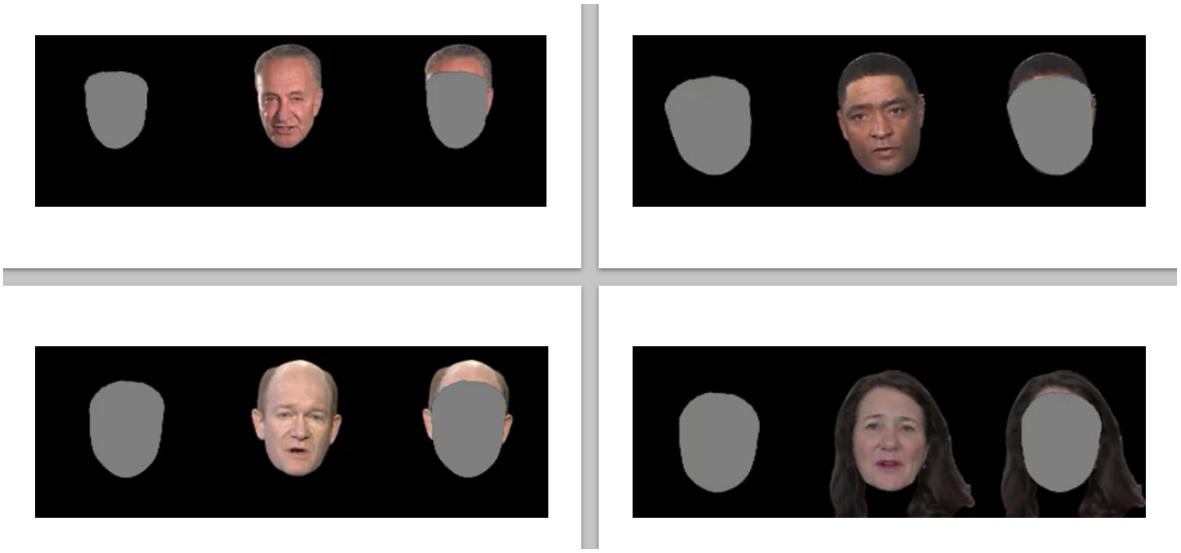

Figure 4.3: Visualization of 3DMM shapes and poses obtained from sample identities in the HDTF dataset.

sions and lip motion. Some warping-based methods [254, 254, 223] employ 3D morphable models(3DMM), or face blend shapes, which support animation via disentangled representation of shape, expression and pose. The visualized 3DMM example are shown in Figure 4.3. However, these approaches can fall short of accurately reproducing a talking face due to limited amplitude, leading to shortcomings in identity preservation and pose controllability. Our approach is designed to directly edit the 3D representation to animate lip motion over time.

4.3 Methodology

The overall architecture of our method Splat-Portrait is illustrated in Fig. 4.5. Splat-Portrait consists of two main stages: (1) pre-training to reconstruct 3D static splats (Sec. 4.3.1); (2) fine-tuning with an audio-conditioned dynamic decoder (Sec. 4.3.2), while also using score distillation (Sec. 4.3.3) to refine appearance from extreme viewpoints.

4.3.1 Static Splat Generation

3D Gaussian Splatting (3DGS) [237] uses anisotropic 3D Gaussians as geometric primitives to explicitly represent 3D scenes. For our 3D head reconstruction, we first pre-train a static generator (SG) that outputs pixel-aligned splats, as shown in Fig. 4.5. The design of SG is based on Splatter-Image [255]. However, unlike [255] we do not have access to wide-baseline multi-view images for training; instead we use more challenging monocular video data. We also predict an inpainted 2D RGB background as well the per-pixel 3D splat

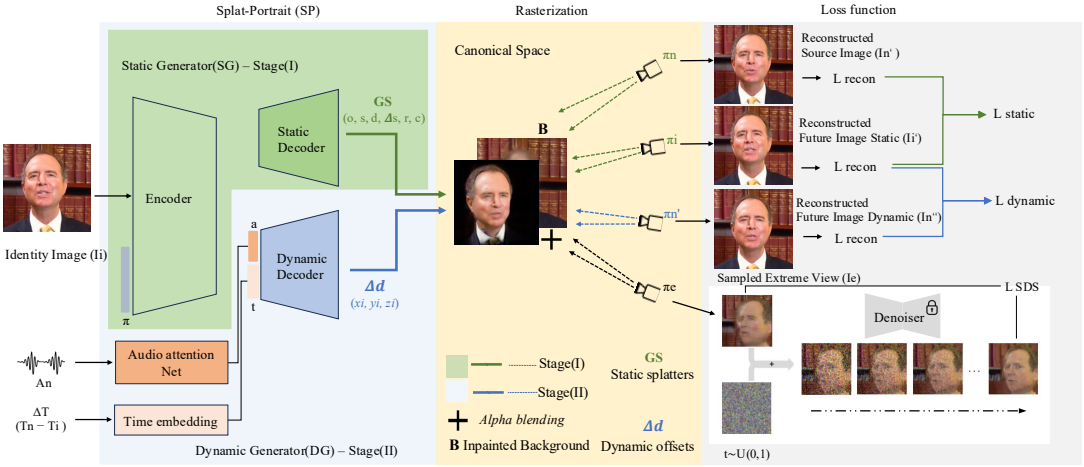

Figure 4.4: Overview of Splat-Portrait. The identity image I_i is passed through a U-Net Static Generator(SG) to reconstruct static 3D Gaussian Splats, alpha-blended over a predicted 2D background. The dynamic decoder estimates splat offsets at timestep T_n using audio features A_n and time embedding ΔT. The training procedure consists of two stages, stage(I): an initial pre-training phase, where the static components are trained on a large-scale dataset using a static reconstruction loss $\mathcal{L}_{\text{static}}$, and stage(II): a fine-tuning phase on a smaller dataset incorporating an additional dynamic reconstruction loss $\mathcal{L}_{\text{dynamic}}$. And a score distillation loss $\mathcal{L}_{\text{SDS}}$ on extreme viewpoints applied during both stages.

attributes, and alpha-blend the rasterized splats over this. During training, we randomly choose pairs of frames from a video, denoted source image I_i and future image I_n. Given I_i, the network predicts a set of Gaussian Splatting parameters GS, at each pixel: opacity o, scale s, depth d, static offset Δ_s, rotation r, splat colour c (encoding per-pixel 3D Gaussian attributes), and the 2D background colour RGB. The view-space 3D position p of the Gaussian at a pixel with ray direction $\mathbf{r}$ is then given by $p = \mathbf{r}\, d + \Delta_s$.

During training, we feed the network with I_i at time step Ti. Additionally, we inject the approximate camera-to-world translation and focal length π. We do so by encoding each entry via a sinusoidal positional embedding of order 9, resulting in 60 dimensions in total. These are applied to the U-Net blocks via FiLM [256] conditioning. During our experiments, we found this helps with convergence of depth predictions.

Given the Gaussian attributes described above, we use the differentiable rasterizer $\mathcal{R}$ from [237] to render the splats at canonical space with static offset enabled to reconstruct images I_i^* and I_n^* at the camera poses of both I_i and I_n respectively. We compute a combined L2 and LPIPS reconstruction loss $\mathcal{L}_{\text{static}}^{\text{rec}}$ between corresponding rendered and ground-truth images, i.e.

$$\mathcal{L}\text{static} = ||I_i - I_i^*||_2 + ||I_n - I_n^*||_2 + \lambda_{\text{LPIPS}} \left[\mathcal{L}_{\text{LPIPS}}(I_i,\, I_i^*) + \mathcal{L}_{\text{LPIPS}}(I_n,\, I_n^*)\right] \tag{4.1}$$

Here the LPIPS term combines VGGface and VGG19 features, and the weight λ is empirically set to 0.01. For each image, we render (and calculate the loss) twice, once with a random coloured background and once with our predicted 2D background alpha-blended behind the splats. We found that this helps to improve the colour and opacity for both background and foreground regions, without incorporating any mask supervision.

4.3.2 Audio-Conditioned Dynamic Splats

For predicting audio-conditioned dynamics representing lip movements, we designed a dynamic decoder with skip connections from the SG decoder. We only use this during the fine-tuning stage, after a good static reconstruction model has been learnt during pre-training. It predicts time-dependent offsets for every splat, conditioned on the audio signal and a time delta indicating what instant in the audio we want the splat offsets for. In our experiments we found that including this time delta improves convergence of the dynamic decoder.

For a given input frame I_i, and future frame I_n plus its contemporaneous audio segment, we first extract audio features using `Wav2Vec2-XLSR 53` [257]. Our model employs dedicated networks to fuse audio and temporal information effectively. Specifically, audio features are first encoded through an audio feature extraction module (AudioNet), which comprises several 1D convolutional layers followed by fully connected layers to yield compact audio embeddings. These embeddings are further refined through an attention-based network (AudioAttNet); this consists of a series of convolutional layers with decreasing channel sizes (from 16 to 1) interleaved with LeakyReLU activations. The output from these convolutional layers is then reshaped and passed through a linear layer followed by a softmax operation to calculate attention weights across the audio sequence. The weighted audio embeddings are summed to produce a refined audio representation capturing temporal dependencies across audio frames. For temporal embeddings, positional encoding or Fourier-based embeddings are utilized to encode timestep information; then audio and temporal embeddings are combined to form the conditioning feature. This combined embedding is injected into the dynamic decoder using FiLM conditioning, allowing the audio and time delta to control the generated motion. During training, when the input image shows a closed or only slightly open mouth, the teeth and parts of the inner mouth can be fully or partially occluded. When the mouth opens at a future time step, the model must still generate plausible teeth and mouth appearance. This means the task is not only to deform visible geometry, but also to predict the occluded content from the source image. In practice, this requires the dynamic splats to model the teeth and inner mouth part.

Our dynamic decoder outputs a dynamic offset Δ_d for the splat at each pixel, conditioned on time T. Hence the splat position at time T is $p_T = p + \Delta_d$. To effectively train our model and maintain the static reconstruction ability learnt during pre-training, we adopt both

$\mathcal{L}_{\text{static}}$ loss and the SDS loss introduced in Sec. 4.3.3. We render the source frame I_i^* with dynamic offsets fixed to zero (as in Sec. 4.3.1), but now render the future frame I_n^{**} using the predicted offsets. Our dynamic reconstruction loss in the fine-tuning stage is:

$$\mathcal{L}\text{dynamic} = ||I_i - I_i^*||_2 + ||I_n - I_n^{**}||_2 + \lambda_{\text{LPIPS}} \left[\mathcal{L}_{\text{LPIPS}}(I_i,\, I_i^*) + \mathcal{L}_{\text{LPIPS}}(I_n,\, I_n^{**})\right] \quad (4.2)$$

4.3.3 Distillation from a 2D diffusion prior

In the fine-tuning stage, we also use score distillation [7] to extract knowledge from a 2D diffusion model [243] to improve the appearance of extreme poses. We first render our predicted reconstruction at a randomly sampled extreme pose, then crop and align the image following [258] to match the distribution learnt by the 2D diffusion model. Given this aligned image x_{clean}, we then add a random amount of noise then run the reverse diffusion process. This score distillation sampling (SDS) is important for our talking head generation setting. Unlike original DreamFusion [7] SDS, our goal is not to synthesize a new identity or alter the underlying facial motion, but to use the 2D diffusion model as a portrait realism prior for the supervision of rendered head images. we apply this prior only to extreme viewpoints, where monocular talking head reconstruction is typically most under-constrained due to limited viewpoints and missing texture evidence. In this way, the diffusion prior mainly improves realism and texture completion for challenging poses while preserving the geometry, identity, and expression predicted by our model.

Specifically, we define a sequence of noise levels σ as follows:

$$\sigma_i = \left[\sigma_{\text{max}}^{\frac{1}{\rho}} + \frac{i}{N-1}\left(\sigma_{\text{min}}^{\frac{1}{\rho}} - \sigma_{\text{max}}^{\frac{1}{\rho}}\right)\right]^{\rho}, \quad (4.3)$$

where σ_{max} and σ_{min} denote maximum and minimum noise levels, ρ is a hyper parameter controlling the distribution of timesteps, and N is the total number of discretized steps. We choose the noise level from 60%–80% of the original range used in training the diffusion model, since we found this range effectively preserves the portrait's overall appearance while significantly improving texture inpainting for extreme viewpoints.

The noised image at the initial timestep t_0 is generated by adding Gaussian noise to the normalized input image, i.e. $x_{\text{noised}} = x_{\text{clean}} + \sigma_0 \cdot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. For each subsequent timestep, we perform an Euler integration step to progressively denoise the image. Specifically, given the current timestep t_{cur} and next timestep t_{next}, the Euler step is computed as:

$$x_{\text{next}} = x_{\text{cur}} + (t_{\text{next}} - t_{\text{cur}}) \cdot d_{\text{cur}}, \quad d_{\text{cur}} = \frac{x_{\text{cur}} - \text{net}(x_{\text{cur}}, t_{\text{cur}})}{t_{\text{cur}}}, \quad (4.4)$$

where net represents the pre-trained denoiser model.

Figure 4.5: Visualization results for SDS loss during training, with top row: sampled extrem viewpooints, middle row: with random noise added, bottom row: denoised rgb

This sampling procedure yields a denoised face image (the final x_{next}) that is similar to the original rendered one, but more realistic according to the diffusion prior. We define a loss $\mathcal{L}_{\text{SDS}}$ as the L2 reconstruction loss between the rendered frame x_{clean} and the denoised frame, back-propagating only into the former. This guides our rendered frames to look more like similar realistic samples from the diffusion model. During training, we randomly sample extreme viewpoints following a bullet-effect trajectory, with pitch variations up to $\pm 12.5^{\circ}$ and yaw variations up to $\pm 45^{\circ}$ from the canonical view, and apply the SDS loss between I_i and x_{clean}. Note that unlike [7] and common practice, we apply the SDS loss during model training, not during inference, meaning the latter remains very fast.

4.3.4 Overall Loss

Our total losses are defined as follows. For stage one (static pretraining):

$$\mathcal{L}_{\text{total\_static}} = \mathcal{L}_{\text{static}}(I_i,\, I_n) + \mathcal{L}_{\text{SDS}}. \tag{4.5}$$

For stage two (audio-conditioned fine-tuning):

$$\mathcal{L}_{\text{total\_dynamic}} = \mathcal{L}_{\text{Ldynamic}}(I_i,\, I_n) + \mathcal{L}_{\text{SDS}}. \tag{4.6}$$

In both cases, we use AdamW for optimization, with a learning rate of 2.5×10^{-5} and weight decay of 10^{-5}.

4.4 Experiments

Datasets and Implementation Details. We evaluate our approach on two widely used datasets of monocular talking portrait videos – HDTF [61] and TalkingHead-1KH [244]. HDTF consists of over 400 samples of talking videos from over 350 subjects. For TalkingHead-1KH, we manually select 1100 identity videos following a similar distribution as HDTF, such that there is no occlusion over the torso and mouth, and with static background. Each identity video contains minimum 300 frames and maximum 10000 frames. We extract frames at 25Hz, and the audio sampling rate is 16kHz. We resize the image frames to 256×256. Following the steps in [232], we follow [247] to use 3DMM optimization to extract approximate intrinsic and extrinsic camera parameters. We use the complete video clips (often with substantial camera motion) for training. For evaluation, we randomly sample 50 identity videos as test sets, and use the first 5s of each. We adopt the SongUNet [259] architecture for our static encoder and dynamic decoder.

Metrics. We measure the quality of synthetic images using structural similarity (SSIM), peak signal-to-noise ratio (PSNR), Learned Perceptual Image Patch Similarity LPIPS [260], and Fréchet Inception Distance (FID); we use Cosine similarity (CSIM) for measuring identity preservation, and SyncNet [261] to measure lip synchronization scores (LipSync).

Baselines. We compare our approach to several existing 3D talking head generation works. **OTAvatar** [92] is a video-driven method that uses a pre-trained 3D GAN to obtain a 3D talking portrait video; **HiDe-NeRF** [226], a 3D talking face model that uses a motion prior and deformation field for face animation; **Real3D-Portrait** [225], a nerf-based method that uses images generated by EG3D to train a 3D model; **NeRFFaceSpeech** [245] one nerf-based audio driven method for synthesising talking head video, and the state-of-the-art **GAGavatar** [239]. Additionally in the audio-driven setting, we extend GAGavatar with ARtalker [262]. Note OTAvatar and HiDe-NeRF are video-driven methods, they are not directly driven by audio; for fair comparison, we use the same identity video as driving video for evaluation. We set the input image size as 256×256 to enable fair comparison, upsampling for methods that require this. We compare the baselines using their preferred masking and cropping settings.

4.4.1 Quantitative Evaluation

We compare with the baselines in *same identity* and *cross-identity* settings. During testing, the driving motion condition and head pose are obtained from a reference video. Under the

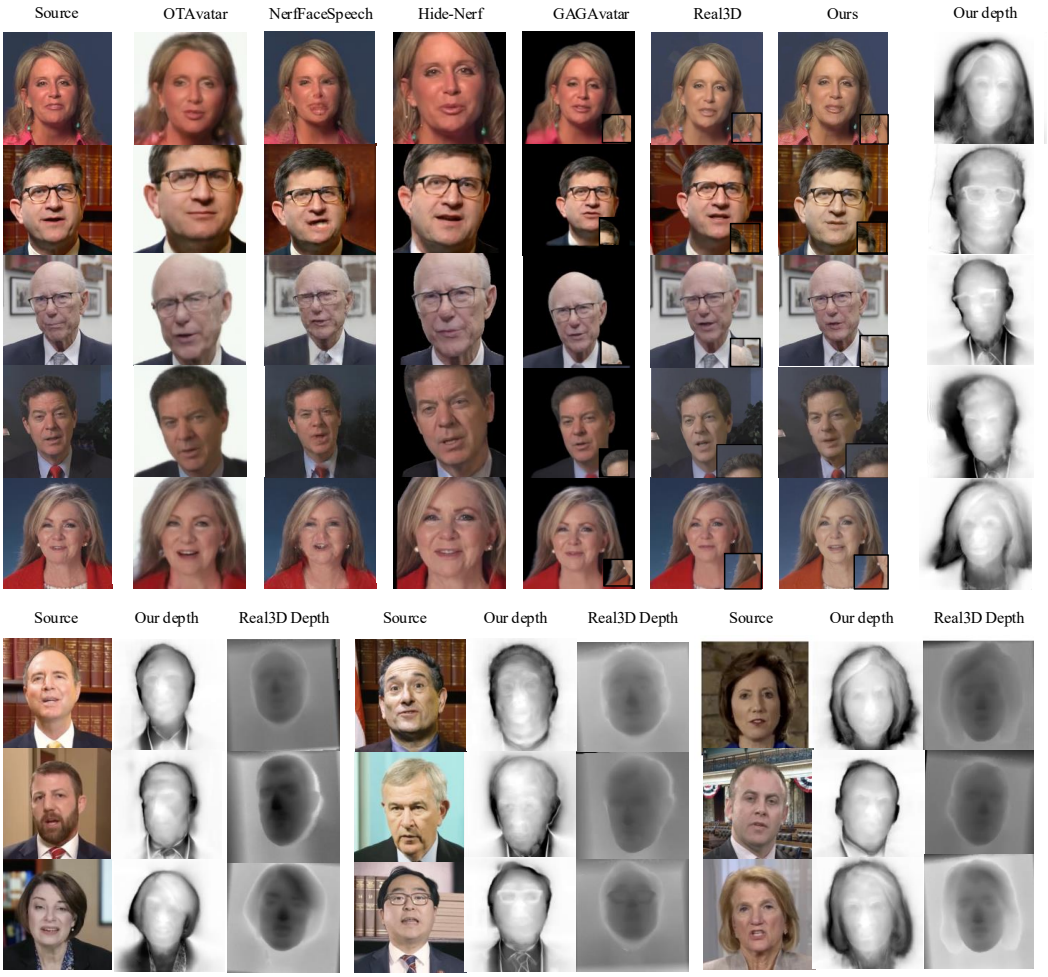

Figure 4.6: Qualitative results. **Top:** We show source frames from five videos, future predicted frames from ours and baselines, and future depths from ours. **Bottom:** Additional examples of 3D reconstruction, for our method and Real3D-Portrait, displaying the input frame, and the reconstructed depth-map from each method.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	FID ↓	LipSync ↑
OTAvatar	13.85	0.488	0.432	0.559	78.98	5.908
NeRFFaceSpeech	13.90	0.520	0.480	0.580	64.6	4.880
HiDe-NeRF	21.44	0.685	0.221	0.716	28.63	5.552
Real3D-Portrait	22.40	0.758	0.191	0.761	35.69	**6.681**
GAGAvatar + ARtalker	23.08	0.786	0.182	0.753	37.89	6.580
Ours	**23.87**	**0.814**	**0.128**	**0.811**	**25.58**	6.328

Table 4.1: Quantitative evaluation of our method and baselines in the same identity setting.

Method	CSIM ↑	FID ↓	LipSync ↑
NeRFFaceSpeech	0.450	50.80	4.423
OTAvatar	0.521	79.32	5.032
HiDe-NeRF	0.628	31.23	5.652
Real3D-Portrait	0.691	40.82	6.521
GAGAvatar + ARtalker	0.687	35.82	**6.503**
Ours	**0.726**	**28.62**	6.218

Table 4.2: Quantitative evaluation of our method and baselines in the cross-identity setting.

Method	Background?	Torso?	Parametric 3D facial model Prior?
Real3d-Portrait	Y	Y	Y
GAGAvatar	N	N	Y
Hide-Nerf	N	N	Y
Ours	Y	Y	N

Figure 4.7: Setting comparison for baselines and our method

same-identity setting, we use the first frame of the reference video as the source image; otherwise, the source image is of a different identity. For the cross-identity setting, for Real3D-Portrait, we only compare with its audio-driven setting. Quantitative results concerning the quality and fidelity for the same-identity setting are listed in Tab. 4.1. These show that our method outperforms other state-of-the-art approaches on almost all fidelity metrics. This is despite our method being trained without 3D supervision, using only a dataset of monocular videos. Splat-Portrait achieves the best overall video quality, as well as higher LipSync score, demonstrating that our 3D deformable model without any motion representation could sync well on lip motions. In Figure 4.7, we show different settings for the baselines and our method. Moreover, our model achieves the highest performance on CSIM, meaning it has a strong ability to preserve subject identity in different views. We also compare the baselines with cross-identity evaluation, where the driving videos are obtained from a reference video, and we use another identity for target. Since there is no ground truth for this setting, we evaluate the results only on CSIM, FID and Lip sync. The results are given in Tab. 4.2. We see that our method performs best on FID and CSIM, which indicates our model still yields high video generation quality even in this more challenging setting.

4.4.2 Qualitative evaluation

In this section we provide visual comparisons of all tested methods (see Figure 4.6). We find that our method preserves face texture details, such as hair and wrinkles well, yielding

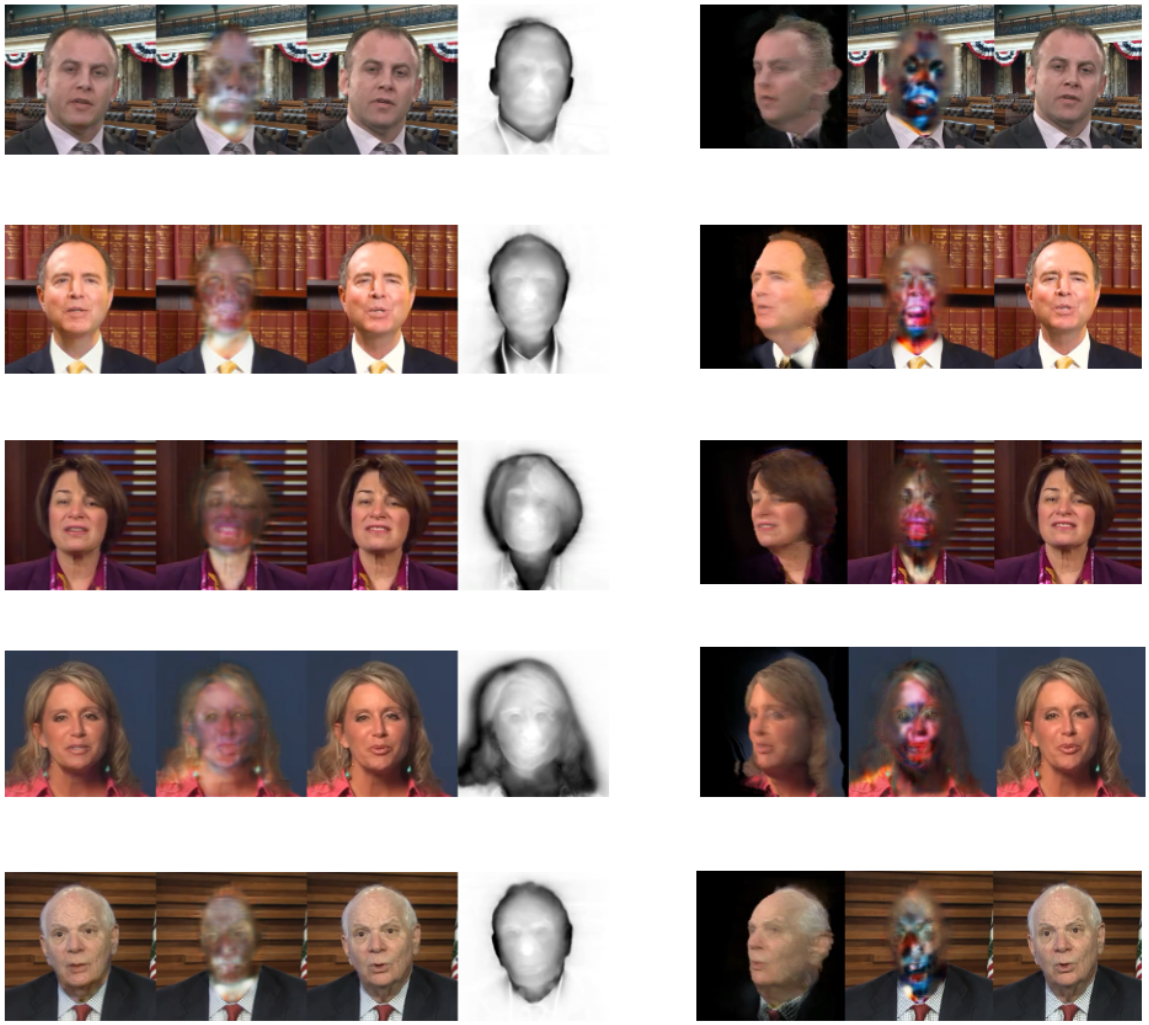

Figure 4.8: More results of Splat-Portrait. From left to right: predicted frame, predicted background(second and last second columns), depth, extreme-view renderings, and ground truth (third and last columns).

high-quality novel views. In particular, our method preserves details such as earrings which move during the video. Since we do not require the head to be pre-segmented, our method handles fine details at the silhouette edges well, and effectively blends the rendered portrait over the estimated background. Fig 4.6 also compares depth-maps rendered by our model with those from Real3D-Portrait, to better visualise the quality of the 3D shape. We show more of our results in Figure 4.8. Compared with Real3D-Portrait, it is clear that our method preserves much more detailed geometry information.

4.4.3 Ablation Study

We test four ablations of our model: (1) w/o time delta, which does not inject the time embedding (see Sec. 4.3.2); (2) w/o pre-training, which does not pre-train the static generator

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o time delta	22.68	0.768	0.146
w/o pre-training	23.30	0.758	0.149
w/o SDS	23.58	0.788	0.147
w/o static offset	23.30	0.791	0.145
only future l2 loss	23.41	0.772	0.138
Full (SP)	**23.87**	**0.814**	**0.128**

Figure 4.9: Ablation study showing the benefit of different components of our model.

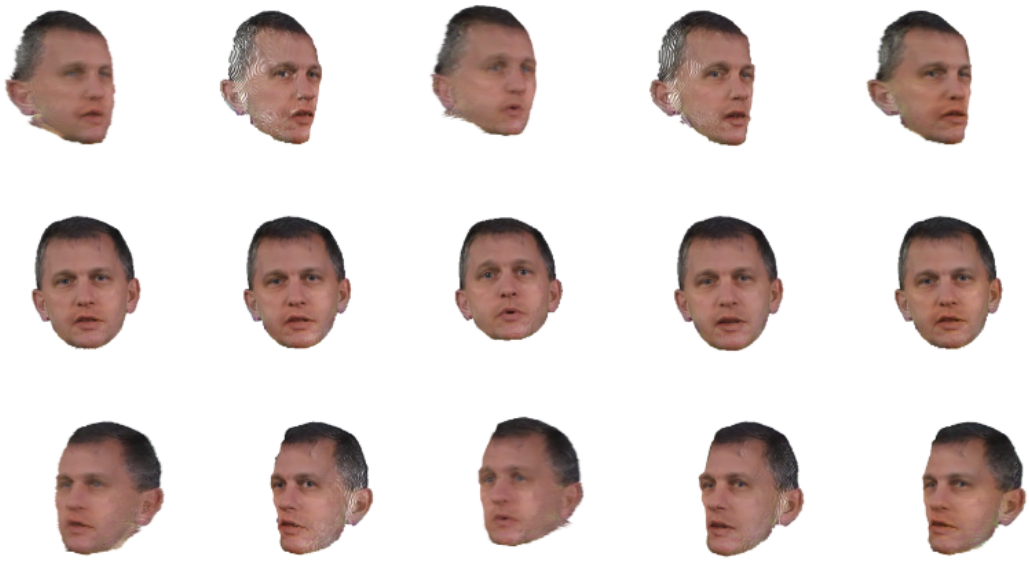

Figure 4.10: Ablation results under three head yaw angles: -35° (top row), 0° (middle row), and +35° (bottom row). The columns from left to right correspond to the following settings: without pre-training, without SDS loss, without static offset, using only the future L2 loss, and our full model.

(see Sec. 4.3.1); (3) w/o SDS, which omits the score distillation loss during the fine-tuning stage (see Sec. 4.3.3); (4) without static offsets during fine-tuning stage; (5) without the initial-frame reconstruction loss, only the future reconstruction loss. We show the results in Fig. 5.6. Without the pre-training process, we see the 3D geometry accuracy drops significantly, the reconstructed 3D head exhibits flattened geometry, with reduced 3D structure. Training with only one frame for supervision instead of two randomly selected frames, it is hard to reconstruct depths and static offsets (and thus the static shape of the face) well, as some structural information is instead represented in the dynamic offsets. As shown in Fig. 5.6, when enabling static splat offsets, the visualized 3D representation shows a smooth, realistically curved geometry. Lastly our SDS loss greatly enhances the realism of extreme poses.

4.5 Conclusion

We proposed Splat-Portrait for Talking Head Generation. Our method is trained on monocular videos without 3D supervision, yet can synthesize accurate 3D geometry and plausible lip movements directly from a single portrait image, yielding state-of-the-art results. By effectively disentangling static and dynamic attributes and using a score-distillation loss, Splat-Portrait significantly enhances realism, particularly from extreme viewpoints rarely encountered during training. Additionally, the simplicity and efficiency of our model structure allow it to animate 3D splats effectively without complex deformation models, making it lightweight and practical for real-world applications. In this chapter, we focus on reconstructing human faces. In the next chapter, we introduce methods and insights for the reconstruction of animals.

4.6 Future work

In this section, we discuss the future directions of Splat-Portrait. Our current splat-portrait model is trained purely with 2D supervision. While it yields strong reconstruction quality, it still fails to recover a complete head under large viewpoint changes, primarily because the training data lack full head coverage. A promising direction is to adopt a powerful pre-trained generative prior—e.g., CogVideoX-2B [263], to improve generalization to unseen regions via prior-driven completion. In addition, our approach does not create any stochastic content; the reconstructed expressions are sometimes imperfect. In particular, eye blinking and other fine-grained facial nuances are rarely captured. Incorporating a stochastic expression prior (e.g., diffusion- or VAE-based) could better model the nuances of facial dynamics. Moreover, audio features are strongly correlated with the mouth region but are less correlated with other facial areas, making audio-conditioned expression synthesis inherently ambiguous. Future work should therefore explore uncertainty-aware objectives and latent-variable conditioning to disambiguate audio-to-expression mapping. Finally, although our method reconstructs portrait content reliably, it struggles under large torso motions because we only use camera extrinsics associated with the head. We plan to jointly estimate head and torso extrinsics without relying on any "ground-truth" camera poses, enabling end-to-end, pose-free optimization that accounts for coupled rigid motions.

Chapter 5

Articulated Animal Reconstruction

In this chapter, we focus on our final task, Articulated Animal Reconstruction, Unlike the human-face domain, animals exhibit complex, varied body topologies, which makes reconstructing a 3D animal substantially much harder. In contrast to building a generic talking human face model, our goal in this task is to overfit a single monocular animal video in order to reconstruct the animal's 3D geometry and 4D motions.

5.1 Introduction

We consider the task of creating animatable 3D dog avatars from single-view videos. This problem is difficult because animals often move in unpredictable, non-rigid ways and show large variations in appearance such as fur patterns and tails. Based on the original work from AnimalAvatar [178], we analyse their performance, and make some extensions. The overall setting of this approach is given a monocular video of a dog, AnimalAvatar builds a template-based method to reconstruct the shape, time-dependent motion and texture, as shown in Figure 5.1.

AnimalAvatar builds a 4D representation that jointly captures pose changes and canonical appearance across frames. To improve shape fitting, AnimalAvatar extends the SMAL model with Continuous Surface Embeddings (CSE), providing dense image-to-mesh constraints instead of relying on sparse keypoints. For appearance modelling, AnimalAvatar designs an implicit duplex-mesh texture defined in the canonical pose, which deforms with SMAL pose parameters and ensures photometric consistency with the input frames. However, in AnimalAvatar we often observe misalignment between the SMAL model and the ground-truth animal videos. To further enhance the mesh optimization, we propose a dense tracker loss based on AnimalAvatar to further track better 3D geometry. We also evaluate more animal videos than original AnimalAvatar did. Experiments on these videos show that

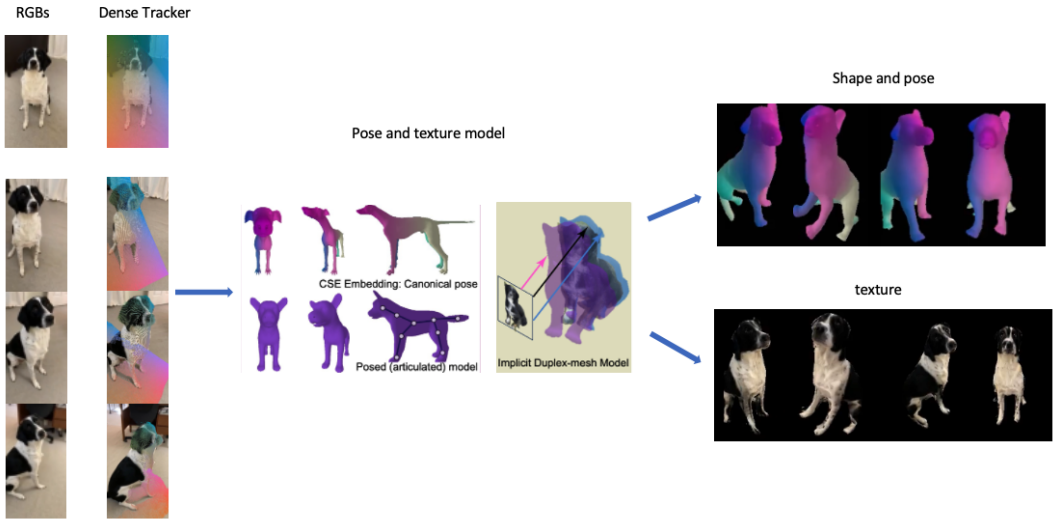

Figure 5.1: Setting of the extended AnimalAvatar method: given a monocular video of a dog, we extract their dense tracker feature, and then AnimalAvatar builds a template-based method to reconstruct the shape, time-dependent motion and texture.

our extension on AnimalAvatar outperforms both template-based and template-free baselines in pose estimation and appearance reconstruction.

Reconstructing realistic and animatable 3D animal models from everyday videos is an emerging and challenging topic in computer vision. Unlike rigid objects or static scenes, animals exhibit a wide range of complex, non-rigid deformations and diverse surface appearances such as fur, colour variations, and tails. Building 3D models that can represent both geometry and motion of animals from casual, monocular videos would enable numerous applications in augmented and virtual reality, video editing, and behaviour analysis.

While modern approaches have achieved impressive progress in reconstructing humans from videos [168, 264, 265], extending these techniques to animals remains difficult. Human reconstruction benefits from large 3D scan datasets and dense annotations [266, 267], but similar resources are rarely available for animals. Consequently, existing methods either rely on synthetic data or limited annotated image collections such as StanfordExtra [268, 269].

Template-based methods address this by using the SMAL model [156], a parametric representation similar to the SMPL model for humans. SMAL provides a deformable mesh template with pose and shape parameters, enabling single-view or multi-view 3D reconstruction of quadruped animals [1, 2]. However, these approaches often depend on sparse 2D keypoints that primarily cover frontal body parts, making reconstructions from side or rear views unreliable.Recent progress in differentiable rendering [270] and neural implicit representations [101, 105] has opened new possibilities for learning 3D geometry and texture from images. At the same time, works such as BANMo [176] and RAC [3] demonstrated the po-

tential of learning canonical 3D embeddings from videos of deformable objects. However, these methods are either template-free and prone to geometric inconsistency, or template-based but lack dense image-to-surface correspondences.

The key idea of AnimalAvatar to integrate the SMAL parametric model with *Continuous Surface Embeddings (CSE)* [271], enabling dense pixel-to-vertex correspondence across diverse viewpoints. Furthermore, AnimalAvatar adopt a duplex-mesh implicit texture representation that maintains appearance consistency under articulated motion, and a temporal optimization strategy that decouples camera and subject motion using Structure-from-Motion (SfM) [272]. In this work, we extend the AnimalAvatar method, and do a more thorough evaluation on Cop3d data. For accurately reconstructing animatable and textured 3D animal models directly from monocular videos, we propose a new dense tracker loss. Through these design choices, our extended model produces geometrically accurate, temporally consistent, and photo-realistic reconstructions without requiring multi-view data or 3D supervision. We evaluate the overall approach on the CoP3D [162], achieving superior performance compared to both template-based (BARC, BITE) and template-free (RAC) baselines. In summary, our main novel contributions of the extended AnimalAvatar [178] from the original AnimalAvatar [178] method are twofold. First, we introduce a dense tracking loss, beyond AnimalAvatar [178], that provides 3D-aware pixel motion supervision for articulated animal reconstruction. By modelling pixel movements across consecutive frames, it adds stronger constraints on pose and shape estimation and improves reconstruction accuracy. Second, we perform a substantially more comprehensive evaluation than the original AnimalAvatar work, using not only its original data but also more diverse and challenging videos with difficult lighting, occlusions, complex articulations, and complex animal textures. These experiments show that the original method generalizes poorly in such scenarios, while our approach achieves more robust performance.

5.2 Related Work

5.2.1 Template-based Animal Reconstruction

Early progress in 3D animal reconstruction has relied on parametric template models derived from limited 3D scans. The SMAL model [156] introduced a skinned linear shape representation similar to SMPL [168], enabling optimization of animal pose and shape from 2D keypoints. Subsequent works improved realism by integrating breed-specific shape priors and learning-based regressors. We show Flame, SMPL and SMAL model in Figure 5.2 For instance, BARC [1] leveraged breed information to infer dog shape and pose directly from single images, while BITE [2] enhanced SMAL-based estimation with learned pose

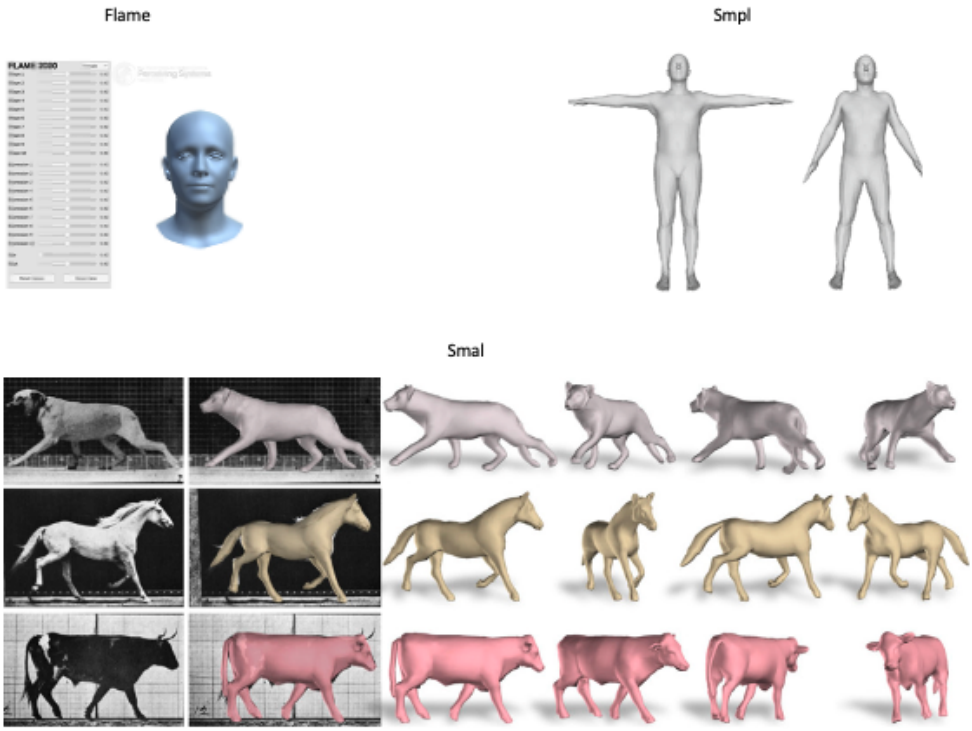

Figure 5.2: Three Parametric model: Flame, SMPL and SMAL.

priors and multi-view consistency. Despite their success, these template-based approaches heavily rely on sparse semantic keypoints and annotated silhouettes, which restrict their generalization to unseen poses and camera viewpoints. Furthermore, they often fail to capture fine-scale geometric or appearance details beyond the coarse mesh topology defined by the template.

5.2.2 Template-free and Implicit Representations

Recent advances in neural implicit modeling and differentiable rendering have enabled learning 3D geometry and appearance directly from image supervision. Methods such as SoftRas [270] and NeRF [101] have shown that dense photometric consistency can replace explicit 3D supervision. These techniques were later extended to dynamic or deformable scenes, as in BANMo [176] and RAC [3], which reconstruct canonical 3D models of humans and animals from monocular videos. However, fully template-free models typically struggle with non-rigid objects like animals, suffering from pose ambiguity, drift, and lack of temporal coherence. Hybrid methods combining implicit fields with deformable templates [273, 274] attempt to overcome these issues but still depend on strong priors or controlled multi-view data.

Figure 5.3: Continuous Surface Embeddings for in the wild animals, reproduced from [8]

5.2.3 Dense Correspondence and Feature Supervision

Dense pixel-to-surface correspondences offer an alternative to sparse landmark supervision for articulated reconstruction. Continuous Surface Embeddings (CSE) [271] introduced a descriptor space that provides dense correspondences between 2D pixels and 3D surface points, enabling stronger geometric constraints. This approach has been successfully applied in works such as CoP3D [162], which reconstructs deformable pets from casual videos, and APTv2 [164], a large-scale dataset for animal pose estimation and tracking. CSE has various animal categories and provide rich dense correspondences, examples are shown in Figure 5.4. Combining such dense supervision with parametric models like SMAL allows for more accurate alignment across varied viewpoints. Nevertheless, most existing studies focus either on static pose estimation or rely on controlled laboratory settings. AnimalAvatar builds upon these insights and introduces a unified framework that jointly optimises animal pose, shape, and texture from monocular videos by combining SMAL priors with dense correspondence learning and implicit appearance modelling.

5.2.4 Dense Tracking for Animal Reconstruction

While dense correspondence descriptors such as CSE [271] offer powerful frame-wise pixel-to-surface mappings, they do not explicitly model temporal coherence across frames, which is crucial for reconstructing dynamic animals. Recent advances in dense point tracking have introduced architectures that jointly reason over spatial and temporal correlations to establish long-range correspondences between arbitrary frames in a video.

The most notable of these is AllTracker [275], a state-of-the-art model that unifies optical flow estimation and point tracking under a multi-frame dense correspondence framework.

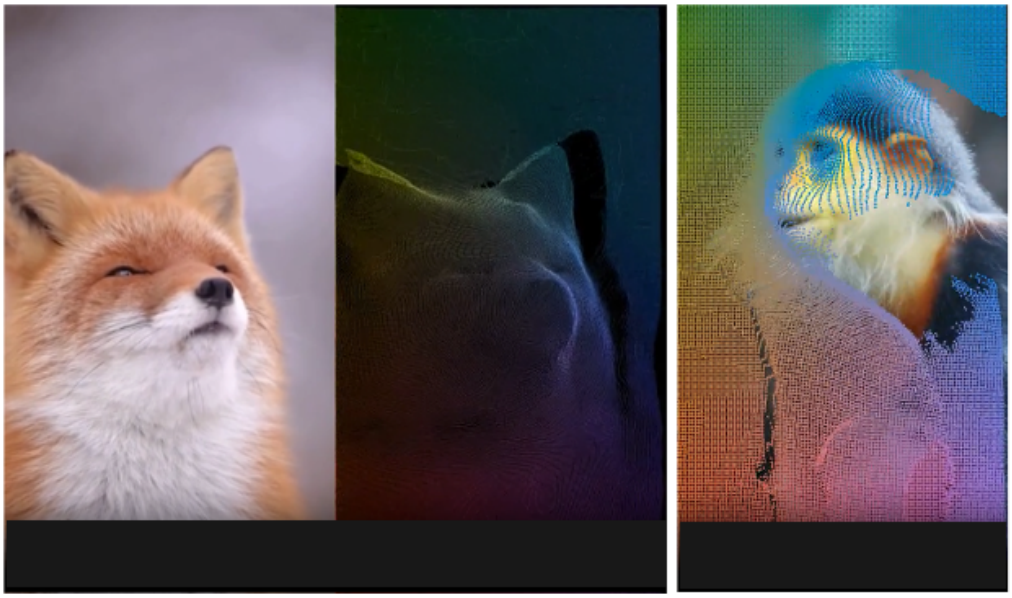

Figure 5.4: Qualitative results of AllTracker for pixel tracking on animal images. From left to right: input image, tracking representation, and intermediate pixel tracking visualization result.

Unlike conventional optical flow models [276, 277], which estimate instantaneous motion between adjacent frames, or sparse point trackers such as CoTracker3 [278], AllTracker estimates the flow between a reference "query" frame and all subsequent frames simultaneously, producing full-resolution, all-pixel correspondence maps. The dataset it trained on contains animal data, one example of AllTracker is shown in Figure 5.4. It operates in a sliding-window fashion across long video sequences, using a recurrent module with spatial 2D convolutions and pixel-aligned temporal attention to iteratively refine motion estimates. This design enables AllTracker to maintain spatial precision while capturing long-range temporal dependencies, achieving state-of-the-art tracking accuracy on diverse datasets, including those with non-rigid animal motion such as BADJA [279] and Horse10 [279].

In the context of 3D animal reconstruction, dense temporal tracking provides complementary supervision to geometric fitting. In our method, we integrate AllTracker as a *motion-aware supervision signal*, guiding the optimization of articulated animal poses across frames. Specifically, the dense motion field predicted by AllTracker offers per-pixel trajectory cues that regularize the deformation of the SMAL-based [156] mesh, improving temporal coherence and mitigating pose jitter under camera movement or occlusion. This allows our system to align both geometric and photometric consistency over time, outperforming keypoint-only or frame-independent correspondence supervision.

By combining dense tracking from AllTracker with spatial surface embeddings from CSE, our extended approach from AnimalAvatar benefits from both high-frequency temporal cues and stable spatial correspondences, effectively bridging the gap between optical flow-

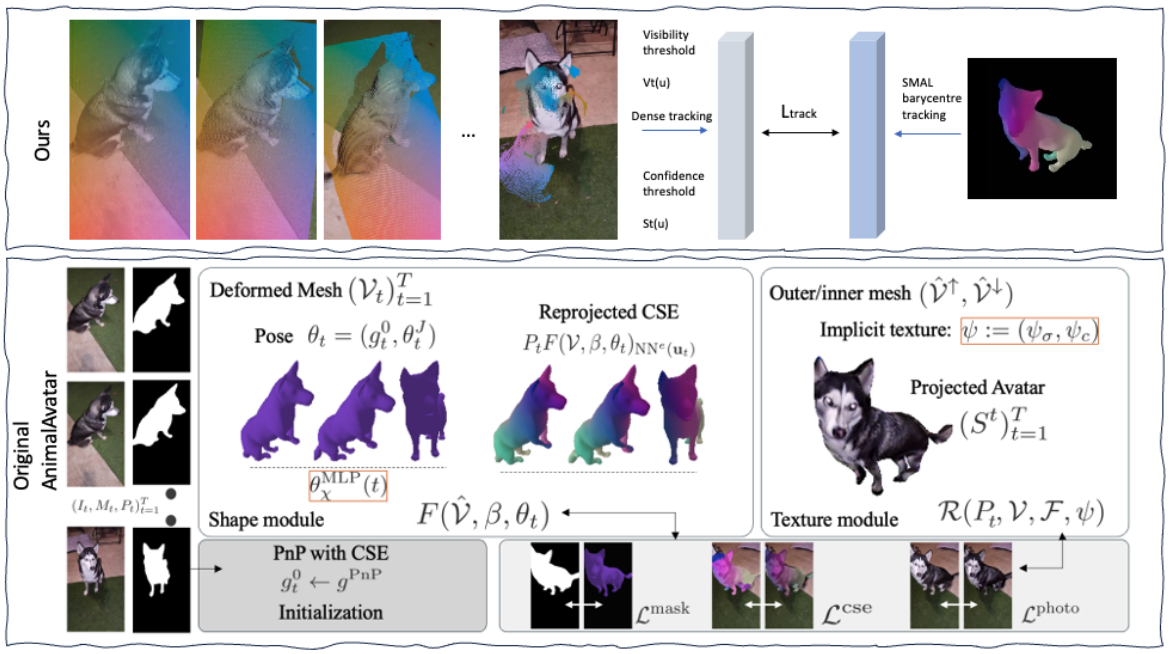

Figure 5.5: Pipeline: The system follows a two-stage optimization process. In the first stage, it initialize the root pose $\mathbf{g}_t^0$ using a PnP-RANSAC procedure, guided by the CSE-based mesh–pixel correspondences. In the second stage, it jointly optimize the shape parameters $\boldsymbol{\beta}$, the time-varying pose parameters$\boldsymbol{\theta}_t$, and the implicit texture representation $\boldsymbol{\psi}$ in an analysis-by-synthesis manner. This joint refinement is driven by multiple complementary supervision signals, including the silhouette consistency loss $\mathcal{L}_{\text{mask}}$, and the photometric reconstruction loss $\mathcal{L}_{\text{photo}}$, and our extended part the dense correspondence loss $\mathcal{L}_{\text{track}}$

based motion estimation and template-based reconstruction. Such synergy enables more robust and temporally consistent animal reconstruction from casual monocular videos.

5.3 Extending AnimalAvatar

As illustrated at the beginning of the chapter, we extent the AnimalAvatar with dense tracking supervision, and We conduct extensive experiments on more challenging animal videos. The aim of original Animal avatar is to reconstruct a fully animatable and textured 3D model of an articulated animal from a single monocular video. AnimalAvatar recovers the underlying 3D shape, time-varying pose, and consistent appearance without relying on multi-view supervision or 3D ground truth. It is a unified optimization framework that combines parametric modelling, dense image-to-surface correspondences, and implicit appearance representation. Figure 5.5 provides an overview of the system with extended part from our idea and original AnimalAvatar idea.

5.3.1 Overview

Given a monocular video sequence $\{I_t\}_{t=1}^{T}$, the system jointly estimates three components: (i) the animal's 3D pose and shape via the SMAL model [156], (ii) the camera motion trajectory, and (iii) a deformable implicit texture representation that captures photometric consistency over time. To ensure geometric accuracy and temporal coherence, in this thesis, we introduce dense tracking supervision signals derived from AllTracker [275] and optimize all components under an analysis-by-synthesis paradigm.

5.3.2 Parametric Shape and Pose Representation

The system adopt the SMAL model [156], a skinned linear blend model for quadruped animals defined by shape parameters β, pose parameters θ_t, and a set of joint transformations. Each frame's mesh $M_t(\beta, \theta_t)$ is obtained by applying blend-skinning to a canonical template mesh M_0. The model provides a low-dimensional space capturing inter-breed variations [1] while maintaining articulation constraints. This parametric prior stabilizes optimization, especially under occlusion or ambiguous motion. We further refine the mesh vertices with a non-rigid offset field Δv_t, allowing the model to capture individual-specific and fine-grained deformations beyond the linear shape space.

The system model each dog's 3D geometry (parameters $\boldsymbol{\theta}_t, \boldsymbol{\beta}$ in Eq. (1)) using the SMAL canonical template. SMAL has a deformable template mesh $(\hat{\mathbf{V}}, \mathbf{F})$ with $\hat{\mathbf{V}} \in \mathbb{R}^{3889\times 3}$ vertices and $\mathbf{F} \in \mathbb{N}^{7774\times 3}$ triangular faces. The deformation of $\hat{\mathbf{V}}$ is

$$\mathcal{F}\Big(\hat{\mathbf{V}},\, \boldsymbol{\beta},\, \boldsymbol{\theta}_t\Big) := \mathbf{V} \in \mathbb{R}^{3889\times 3}, \tag{5.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{d_\beta}$ are shape parameters (PCA coefficients and bone lengths) that control non-rigid shape in a canonical pose, and $\boldsymbol{\theta}_t := (\mathbf{g}_0,\, \boldsymbol{\theta}_J)$ are pose parameters: (i) $\mathbf{g}_0 \in \mathrm{SE}(3)$ is the global rigid transform; (ii) $\boldsymbol{\theta}_J \in \mathbb{R}^{d_J}$ are joint angles that deform the limbs.

To model time-varying deformation, the system estimates a tuple $(\boldsymbol{\theta}_t)_{t=1}^{T}$ of SMAL pose coefficients $\boldsymbol{\theta}_t$ for the T frames, and a single vector $\boldsymbol{\beta}$, since intrinsic deformation is typically time-invariant. Because animal motion is usually smooth, we define $\boldsymbol{\theta}_t$ from a smooth temporal basis $\gamma(\tau(t))$ via

$$\boldsymbol{\theta}_t := \boldsymbol{\theta}^{\text{MLP}}(\tau(t)), \tag{3}$$

where $\boldsymbol{\theta}^{\text{MLP}}$ is a shallow multi-layer perceptron that takes the positional encoding γ of the timestamp $\tau(t) \in \mathbb{R}_+$ of frame I_t.

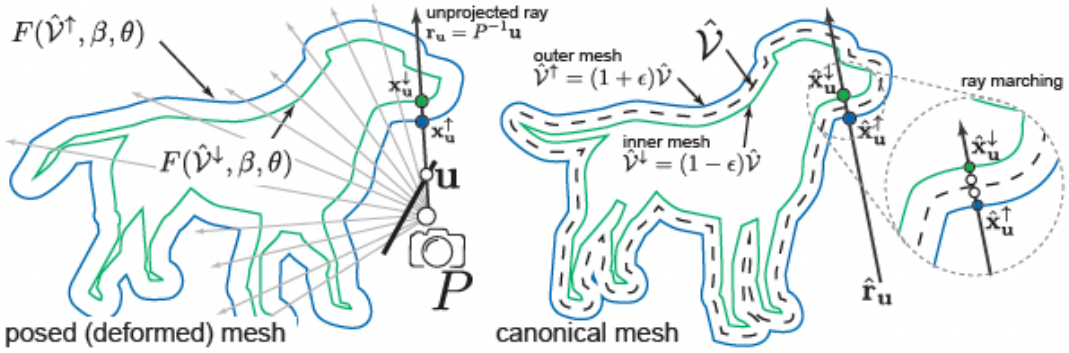

Figure 5.6: **Implicit duplex-mesh model of AnimalAvatar** It is defined that radiance ψ_c and opacity ψ_σ inside an $\mathbb{R}^3$ band bounded by the canonical duplex meshes with vertices $\hat{V}^{\uparrow}$ and $\hat{V}^{\downarrow}$. Given a view ray r_u, AnimalAvatar intersect the posed boundaries $F(\hat{V}^{\uparrow}, \beta, \theta)$ and $F(\hat{V}^{\downarrow}, \beta, \theta)$, map the hits to canonical space to form $\hat{r}_u$, and render colour via EA raymarching.

5.3.3 Camera Motion Factorization

Accurate recovery of articulated motion requires separating animal motion from camera motion. We first estimate the camera trajectory and intrinsics via Structure-from-Motion (SfM) [272], which provides initial camera poses $\{R_t, T_t\}$. During optimization, we decouple the global motion by fixing the SfM trajectory and optimizing only the local pose parameters θ_t of the SMAL model. This factorization prevents entanglement between subject and camera motion, yielding temporally smooth 3D reconstructions. We further enforce a temporal regularization loss $\mathcal{L}_{\text{temp}} = \sum_t \|\theta_t - \theta_{t-1}\|_2^2$ to ensure stable motion transitions.

5.3.4 Dense Correspondence Supervision

Sparse 2D keypoints are insufficient to constrain complex body parts like tails, legs, and ears under severe self-occlusion. To overcome this limitation, the system employ dense correspondence supervision using the CSE descriptor space [271]. Each pixel p in frame I_t is associated with a learned descriptor $\mathbf{c}_p$, and each vertex v_i on the mesh is embedded into the same feature space. By aligning $\mathbf{c}_p$ and $\mathbf{c}_{v_i}$, we obtain dense pixel-to-vertex correspondences that cover the entire animal body, including rear and side views. We use a cosine-similarity loss to enforce descriptor alignment:

$$\mathcal{L}_{\text{CSE}} = 1 - \frac{\mathbf{c}_p \cdot \mathbf{c}_{v_i}}{\|\mathbf{c}_p\| \|\mathbf{c}_{v_i}\|}. \tag{5.2}$$

This dense constraint complements the sparse keypoint loss $\mathcal{L}_{\text{kpt}}$ and improves geometric coverage, especially for unconstrained video data [162, 164].

5.3.5 Implicit Duplex-Mesh Texture Modeling

To capture view-dependent appearance and subtle fur texture, we define an implicit radiance field around the mesh surface following a duplex-mesh representation [274]. For each surface point $\mathbf{x}$ within a narrow shell around the mesh, we predict colour and opacity using an MLP $f_\psi(\mathbf{x}, \mathbf{n}, \mathbf{d})$, conditioned on surface normal $\mathbf{n}$ and viewing direction $\mathbf{d}$:

$$f_\psi : (\mathbf{x}, \mathbf{n}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma). \tag{5.3}$$

This implicit texture is differentiably rendered via a soft rasterizer [270] that accumulates per-pixel radiance using volumetric compositing.The duplex structure allows the texture to deform coherently with articulated motion, avoiding texture stretching common in UV-based mapping.

Given the animal shape $(\mathcal{V}, \mathcal{F})$ and implicit texture ψ, we render from a camera P using a differentiable renderer $\mathcal{R}$:

$$\mathcal{R}(P, \mathcal{V}, \mathcal{F}, \psi) := \bar{I}, \tag{4}$$

which outputs $\bar{I} \in [0,1]^{3\times H\times W}$ for a camera with projection matrix $P \subset \mathbb{R}^{3\times 4}$.

To obtain $\bar{I}$, we iterate over each pixel $u \in [1..H] \times [1..W]$ and perform Emission-Absorption (EA) marching along the canonical ray $\hat{r}_u$, defined from the camera-space ray $r_u = P^{-1}u$ in rest-pose coordinates. Concretely, we first intersect r_u with the posed *outer* boundary mesh $\mathcal{F}(\hat{\mathcal{V}}^\uparrow, \beta, \theta)$, then use the intersection's barycentric coordinates on the canonical mesh $\hat{\mathcal{V}}^\uparrow$ to obtain a 3D point $\hat{x}_u^\uparrow \in \hat{\mathcal{N}}_\epsilon$. Repeating the same for the *inner* mesh $\mathcal{F}(\hat{\mathcal{V}}^\uparrow, \beta, \theta)$ yields a second point $\hat{x}_u^\downarrow \in \hat{\mathcal{N}}_\epsilon$. The two points $\hat{x}_u^\uparrow$ and $\hat{x}_u^\downarrow$ define $\hat{r}_u$, along which we apply EA and accumulate the outputs of ψ_c and ψ_σ to produce the final colour at pixel u (see supplementary for details). Note that this EA rendering differs from duplex radiance fields, which directly use an MLP to map positional encodings of the two intersections $\hat{x}_u^\downarrow, \hat{x}_u^\uparrow$ to a surface colour.

5.3.6 Dense Tracking Supervision

Dense frame-to-frame tracking provides complementary temporal constraints that are not captured by per-frame correspondences. We employ AllTracker [275] to obtain dense, long-range pixel trajectories and visibility/confidence maps for each frame t w.r.t. a reference frame (we use the first frame, t=0). Concretely, AllTracker predicts for every pixel u a 2D trajectory $\mathbf{T}_t(u) \in \mathbb{R}^2$, a visibility score $\nu_t(u) \in [0,1]$, and a confidence score $s_t(u) \in [0,1]$. Let $M_t(u) \in \{0,1\}$ denote the foreground mask, and $\tau_{\text{vis}}, \tau_{\text{conf}}$ be thresholds for visibility

and confidence respectively. We define the valid set

$$\mathcal{V}_t \;=\; \big\{u \;\big|\; M_t(u)=1,\; \nu_t(u)>\tau_{\text{vis}},\; s_t(u)>\tau_{\text{conf}}\big\}. \tag{5.4}$$

To compare AllTracker's trajectories with our geometry as the dense loss, we render appearance-based correspondences $\hat{\mathbf{U}}_t(u) \in \mathbb{R}^2$ for each pixel u: given the reference-frame surface point associated with u (via the posed SMAL mesh and differentiable rendering [270]), we project its 3D location under the current pose (β, θ_t) and camera to obtain its expected image location at frame t. This yields a dense, geometry-consistent correspondence field aligned with our model (Sec. 5.6). We name this process as SMAL barycentre tracking.

We supervise the trajectories using an ℓ_1 objective normalized by the number of valid pixels and coordinate dimensions:

$$\mathcal{L}_{\text{track}}^{(t)} \;=\; \frac{1}{2\,|\mathcal{V}_t| + \varepsilon} \sum_{u\in\mathcal{V}_t} \big\|\mathbf{T}_t(u) - \hat{\mathbf{U}}_t(u)\big\|_1, \tag{5.5}$$

and sum it over time, $\mathcal{L}_{\text{track}} = \sum_t \mathcal{L}_{\text{track}}^{(t)}$. This loss encourages the projected motion of mesh-attached surface points to agree with dense video evidence, thereby reducing pose drift and improving temporal coherence under non-rigid motion and occlusions. In practice, we compute $\hat{\mathbf{U}}_t$ by ray–mesh intersection with the posed duplex boundaries $F(\hat{V}^{\uparrow}, \beta, \theta_t)$ and $F(\hat{V}^{\downarrow}, \beta, \theta_t)$ in view space, mapping intersections to the canonical band to obtain the canonical ray $\hat{r}_u$ and back to frame t via EA rendering [270] (see Fig. 5.6). Compared with adjacent-frame optical flow [277] or sparse point trackers [278], AllTracker's all-pixel, multi-frame design provides long-range, high-resolution cues that are particularly beneficial for articulated animals with self-occlusions and thin structures.

5.3.7 Joint Optimization and total loss

We jointly optimize β, $\{\theta_t\}$, $\{\Delta v_t\}$, and ψ with the following objective:

$$\mathcal{L}_{\text{total}} \;=\; \lambda_{\text{photo}}\mathcal{L}_{\text{photo}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{CSE}}\mathcal{L}_{\text{CSE}} + \lambda_{\text{kpt}}\mathcal{L}_{\text{kpt}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{track}}\mathcal{L}_{\text{track}}. \tag{5.6}$$

Here, $\mathcal{L}_{\text{track}}$ is the dense tracking supervision in Eq. (5.5), which complements the dense spatial constraints from CSE with long-range temporal cues from AllTracker [275], yielding smoother and more stable motion estimates. The combination of parametric priors, dense supervision, and implicit texture modelling provides a strong inductive bias for reconstructing animals from real-world videos. Unlike previous template-only methods [1, 2], AnimalAvatar supports temporally coherent animation and fine-grained appearance recovery.

Compared to fully template-free pipelines [176, 3], our formulation maintains geometric consistency and reduces pose drift. In summary, AnimalAvatar combined with our extended method achieves high-fidelity reconstruction from unconstrained inputs by bridging structured animal models with dense, differentiable supervision.

5.4 Experiments

5.4.1 Dataset

Dataset. AnimalAvatar curated a small set of relatively easy 50 dog videos from **CoP3D** [162] for evaluation. CoP3D is a collection of crowd-sourced "turntable" videos of dogs and cats annotated with cameras and foreground masks. Apart from the dog videos from AnimalAvatar, we further incorporate 16 randomly chosen identities from the CoP3D dataset and manually add 5 additional challenging dog videos. These videos are particularly challenging, from drastic motion changes (e.g., rolling on the floor) to intricate textures (e.g., rich facial hair), which small templates might struggle to model. The added videos are selected to test the robustness of the method to AnimalAvatar, seeing how performance drops beyond the curated validation set in the original paper. Unless stated otherwise, videos are resized to 256×256 and each video contains around 200 frames. We name original 50 dog videos from AnimalAvatar as **OriginalSet**, we name our randomly selected 16 identities as **RandomSet**, and final we name our manually selected challenging dog videos as **DifficultSet**.

Same as AnimalAvatar, we adopt an interleaved split: contiguous blocks of 15 frames are used for training, interleaved by 5-frame blocks for testing (abbreviated as `15/5`), similar to [162]. All methods are evaluated with the same splits. AnimalAvatar additionally report results with two stricter protocols, `15/10` and `15/15`, to stress-test view extrapolation under fewer supervision frames.

5.4.2 Implementation Details

In our extended version of AnimalAvatar [9], SMAL [156] parameters (β, θ_t) are initialized from Bite model [159], except for g_t^0. We use the public AllTracker implementation [275] to generate per-pixel trajectories and visibility/confidence maps. Only pixels with visibility $> \tau_{\text{vis}}$ and confidence $> \tau_{\text{conf}}$ (empirically 0.5) contribute to $\mathcal{L}_{\text{track}}$. Trajectories are downsampled to the original input images resolution.

5.4.3 Metrics

Same as AnimalAvatar, we report both reconstruction fidelity and temporal stability. Silhouette accuracy is measured by **IoU** between the predicted silhouette $\hat{M}_t$ and ground-truth mask M_t:

$$\text{IoU} = \frac{|\hat{M}_t \cap M_t|}{|\hat{M}_t \cup M_t|}.$$

Appearance fidelity is assessed via **PSNR** and perceptual distance **LPIPS** [280] between the rendered image $\hat{I}_t$ and the frame I_t. This metric is particularly useful as it takes into account human visual perception and the structural information of the images, providing a more accurate measure of visual similarity compared to PSNR. To capture worst-case behavior (e.g., extreme poses), we also report the worst 5th-percentile variants $\textbf{IoU}_{\textbf{w5}}$ and $\textbf{PSNR}_{\textbf{w5}}$ computed per sequence by sorting frame-wise scores and averaging the bottom 5%.

5.4.4 Baselines and Evaluation Protocols

We first report the reproduced results of AnimalAvatar using the official code and settings, with baselines: (i) **BARC** [1], a template-based single-image regressor leveraging breed priors; (ii) **BITE** [2], which refines SMAL fits [156] with stronger priors and a test-time loop; and (iii) the template-free **RAC** [3], trained on animal videos to learn animatable category models. BARC and BITE regress pose/shape only; RAC also predicts texture. For fairness, we run official code where available and keep resolution, masks, and splits identical. The quality of our textures can only be directly compared to RAC's model, which also includes texture, but we cannot compare them to BITE or BARC because they only output 3D shapes.

We then evaluate our extension of AnimalAvatar and compare it with the original version. AnimalAvatar provides, for each evaluation identity, a global shape and a set of sparse keypoints. Besides, AnimalAvatar further refines the sparse keypoints by running the process with multiple random seeds and selecting the best result. One example of sparse keypoints is shown in Figure 5.7. These global shape parameters and sparse keypoints are extracted from BITE [2] and are used as initial model parameters and supervision signals during model training. When comparing with AnimalAvatar, especially on our selected set, we consider three evaluation protocols.

Protocol One we consider replicate AnimalAvatar results, we just use the global shape parameters and sparse keypoints extracted from BITE [2] for comparison.

Protocol Two we average the global shape parameters across all 50 identities and use the averaged results as the initial parameters for training on the OriginalSet, RandomSet, and DifficultSet by using the original AnimalAvatar method and compare against it. In the fol-

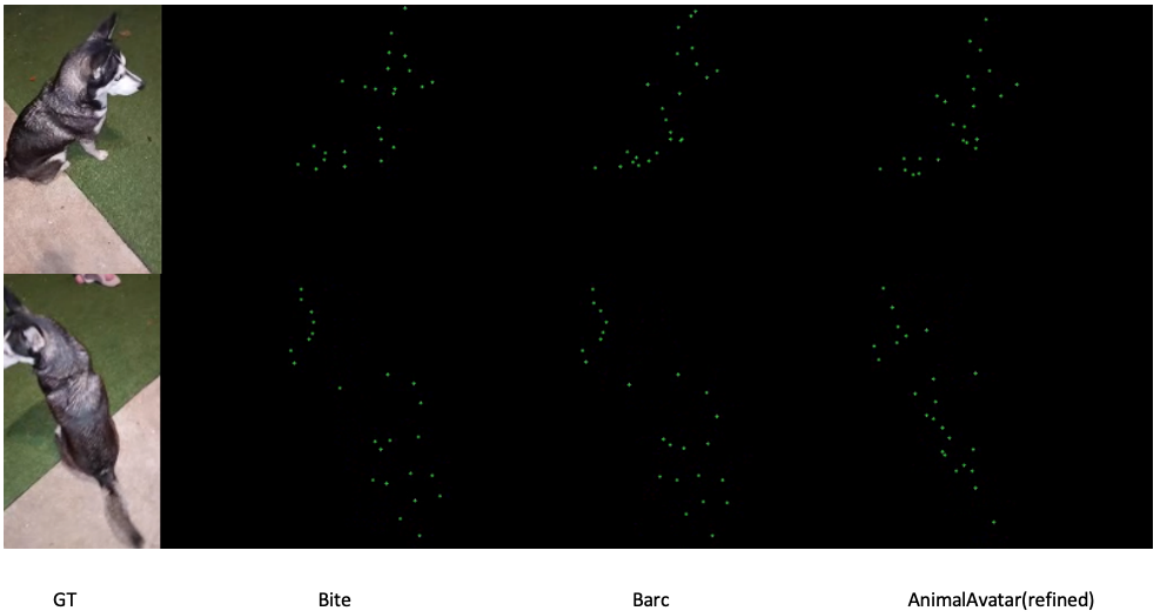

Figure 5.7: Sparse keypoint visulazation: from Bite [2], Barc [1] and AnimalAvatar [9]

lowing section, we our replicated results from original AnimalAvatar and the results from our new evaluation protocol.

Protocol Three we add dense tracking loss based on original AnimalAvatar method, and we experiment it with different visibility and confidence scores. We compare it against original AnimalAvatar method on the OriginalSet.

5.4.5 Experimental Results

In **Protocol One**, Table 5.1 reports the original AnimalAvatar results. In Figure 5.8, we show the replicated AnimalAvatar qualitative comparison results. On the evaluation, our method again surpasses all three baselines. We observe that RAC performs significantly worse than the other methods.AnimalAvatar method achieves the best perceptual quality (**LPIPS**) and strong **PSNR**, indicating that the duplex texture recovers high-frequency appearance details while maintaining geometric accuracy. In terms of silhouettes, AnimalAvatar match or surpass the best template baseline in mean **IoU** and substantially improve $\mathbf{IoU_{w5}}$, highlighting the benefit of dense CSE constraints for uncommon viewpoints. RAC [3] benefits from learned category priors but exhibits shape drift on thin structures; BARC/BITE [1, 2] produce sharp silhouettes on easy views but degrade under large out-of-plane rotations due to frame-independent fitting.

In **Protocol Two** , we average the global shape parameters across all 50 identities and use the resulting mean shape as the initialization for training. As illustrated in Figure 5.9 and summarized in Table 5.2, the global shape parameters play a crucial role in guiding pose

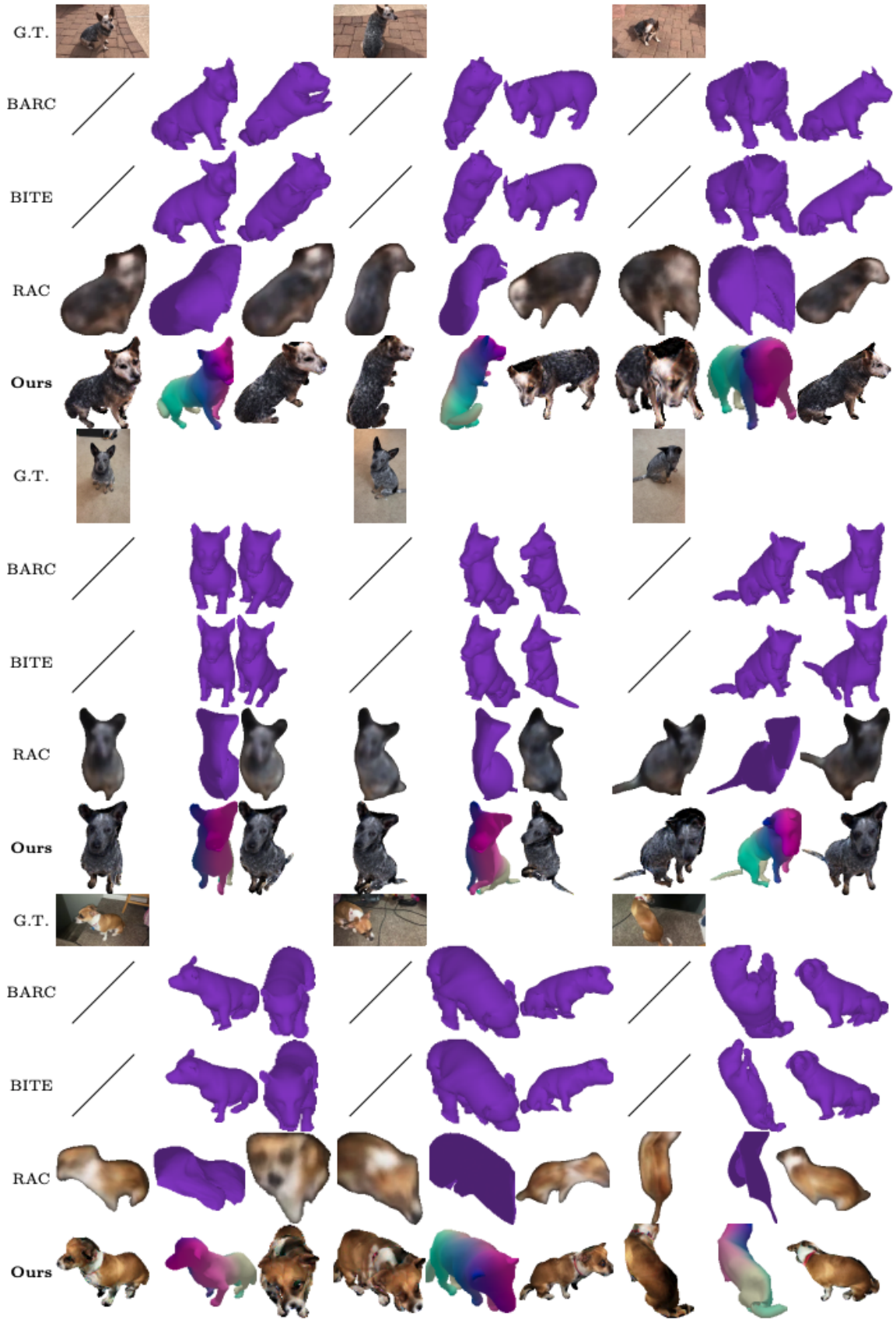

Figure 5.8: **AnimalAvatar Qualitative Comparison Results**, It is worth to notice that, unlike template-based approaches, the reconstructed meshes from RAC are very far from the actual shape of a dog

Table 5.1: AnimalAvatar comparison results against BARC [1], BITE [2], and RAC [3]

Method	Dataset	IoU↑	IoU_{w5}↑	PSNR↑	$PSNR_{w5}$↑	LPIPS↓
BARC [1]	OriginalSet	0.74	0.47	–	–	–
BITE [2]	OriginalSet	0.81	0.59	–	–	–
RAC [3]	OriginalSet	0.76	0.52	21.86	17.51	0.164
AnimalAvatar	OriginalSet	0.82	0.78	22.2	19.50	0.040

Table 5.2: AnimalAvatar with and without Averaged Globe Shape cross OriginalSet, DifficultSet, and RandomSet

Method	Dataset	IoU↑	IoU_{w5}↑	PSNR↑	$PSNR_{w5}$↑	LPIPS↓
AA + Averaged Shape	OriginalSet	0.78	0.72	20.2	17.50	0.058
AA	OriginalSet	0.82	0.78	22.2	19.50	0.040
AA + Averaged Shape	DifficultSet	0.72	0.66	18.7	16.47	0.056
AA	DifficultSet	0.74	0.68	19.8	17.50	0.052
AA + Averaged Shape	RandomSet	0.75	0.70	21.0	18.80	0.058
AA	RandomSet	0.78	0.75	22.0	19.00	0.048

reconstruction. By providing a stable initial shape prior, they help the optimization process converge toward plausible body structures. However, while this averaged initialization offers a consistent starting point, it limits the model's ability to capture individual geometric variations across subjects. Consequently, we observe that although the OriginalSet maintains reasonable reconstruction quality, the results on DifficultSet and RandomSet degrade notably. In particular, complex poses such as folding, twisting, or lying down lead to distorted or incomplete geometry.

Quantitatively, the metrics in Table 5.2 confirm this observation: using averaged shape parameters consistently underperforms compared to the full AnimalAvatar model across IoU, PSNR, and LPIPS metrics. This suggests that individualized global shape estimation is essential for accurately modeling fine-grained animal geometry, especially under challenging poses. Overall, the AnimalAvatar framework benefits significantly from personalized shape priors, which enable better alignment between pose and geometry during optimization.

In **Protocol Three**, we investigate different configurations of the dense tracking loss when integrated with the AnimalAvatar framework, specifically examining the effects of varying the confidence and visibility scores. As shown in Table 5.3, the dense tracking loss has a significant influence on reconstruction quality. Using the default AllTracker setting, with both confidence and visibility scores set to 0.5, leads to noticeably improved geometric accuracy.

From Figure 5.11, we observe that overly strong or weak tracking supervision tends to harm reconstruction quality. In these cases, the recovered geometry often exhibits distorted

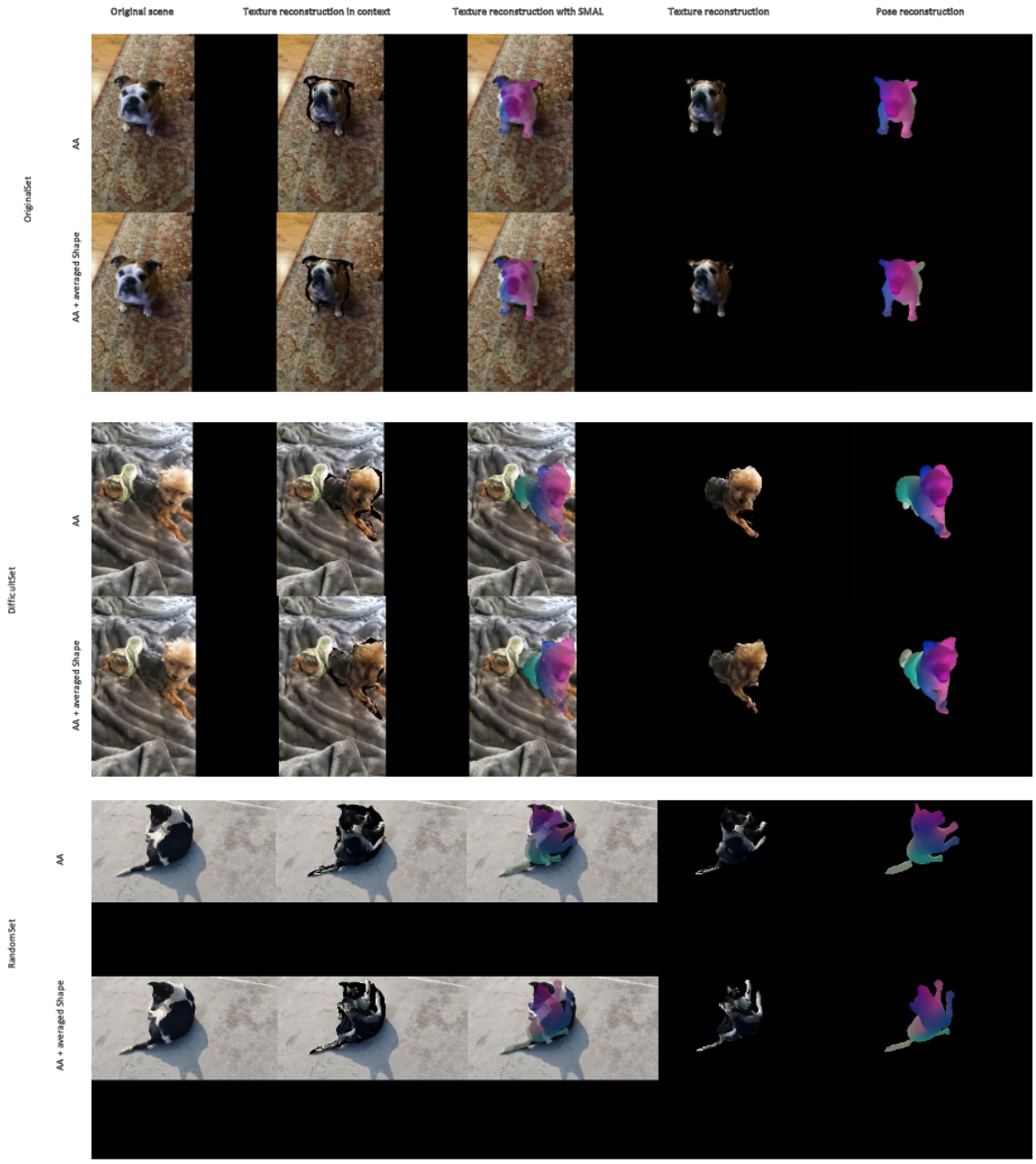

Figure 5.9: **AnimalAvatar with and without Averaged Globe Shape**

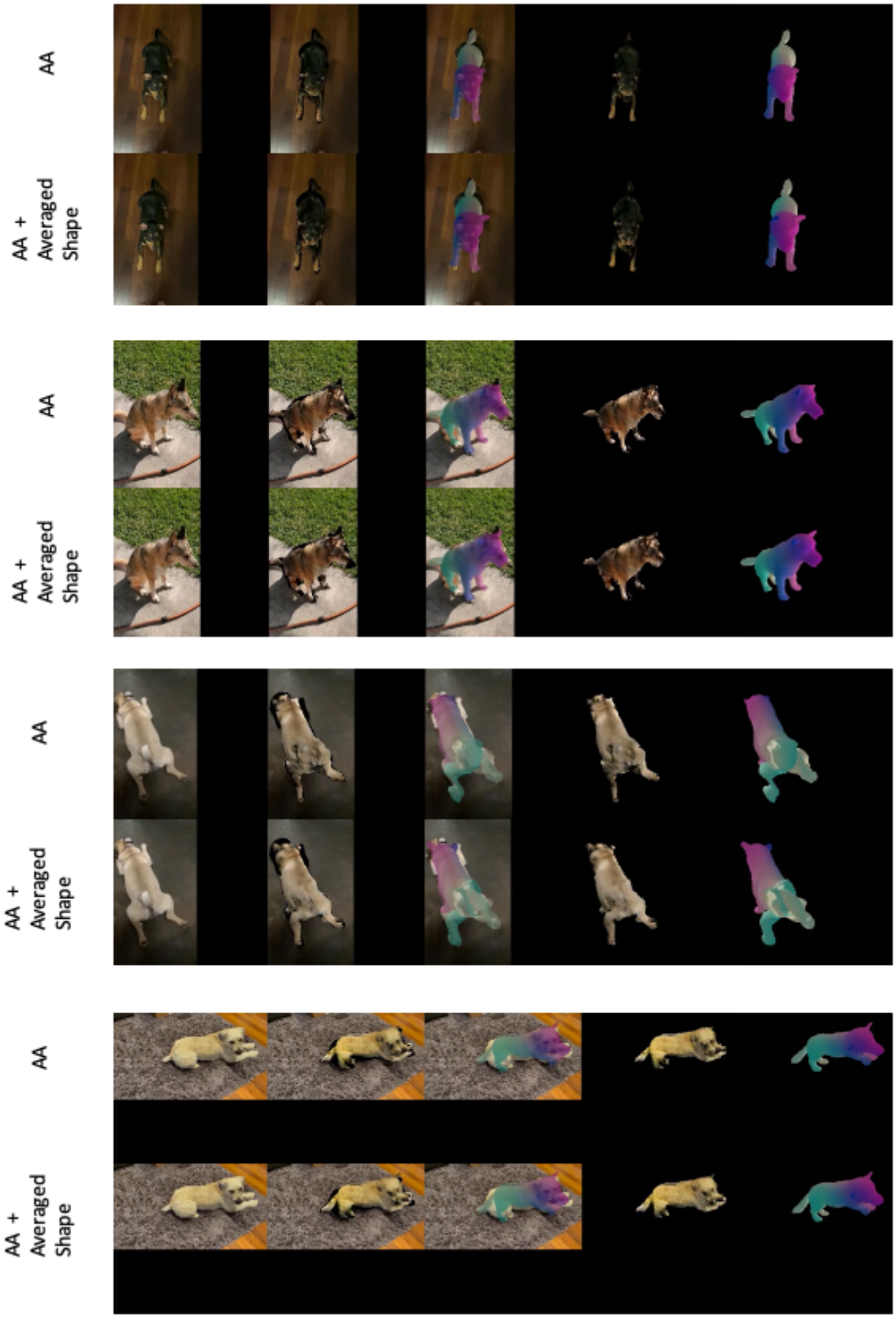

Figure 5.10: **RandomSet only: AnimalAvatar with and without Averaged Globe Shape**

Table 5.3: AnimalAvatar with and without Averaged Globe Shape cross OriginalSet, DifficultSet, and RandomSet

Method	confidence	visitability	IoU↑	IoU_{w5}↑	PSNR↑	PSNR_{w5}↑	LPIPS↓
AA + $\mathcal{L}_{\text{track}}$	0.80	0.50	0.70	0.69	18.60	17.80	0.058
AA + $\mathcal{L}_{\text{track}}$	0.20	0.50	0.73	0.70	19.30	18.30	0.068
AA + $\mathcal{L}_{\text{track}}$	0.50	0.50	0.82	0.80	23.8	21.28	0.038
AA	–	–	0.82	0.78	22.2	19.50	0.040

Table 5.4: Ablation on CoP3D reporting performance with various loss terms removed and without camera motion factorization ($g_t^{\text{cam}} = g_t^0$).

	w/o $\mathcal{L}_{\text{chamfer}}$	$\mathcal{L}_{\text{cse}}$	$\mathcal{L}_{\text{keypoint}}$	$\mathcal{L}_{\text{color}}$	$\mathcal{L}_{\text{arap}}$	$\mathcal{L}_{\text{edge}}$	$g_t^{\text{cam}} = g_t^0$	AA	AA+ $\mathcal{L}_{\text{track}}$
IoU ↑	0.70	0.81	0.80	0.81	0.81	0.83	0.72	0.82	**0.82**
PSNR ↑	20.65	20.89	21.54	21.62	21.61	21.88	19.12	22.20	**23.8**
LPIPS ↓	0.060	0.051	0.048	0.047	0.047	0.045	0.067	0.04	**0.040**

structures, such as folded limbs or duplicated facial regions, resulting in unrealistic textures (e.g., "one and a half" dog faces). In contrast, applying a balanced dense tracking loss (0.5 for both confidence and visibility) yields more stable and coherent results. This balanced weighting effectively guides small facial vertices toward correct spatial positions, enabling the model to better capture detailed shape and pose variations during optimization.

5.4.6 Ablation Studies For Original AnimalAvatar

To validate the overall design choices of the original AnimalAvatar and our dense tracking loss, we first replicated the ablation results of AnimalAvatar on the OriginalSet to evaluate the contribution of its individual components (subtractive ablation). We then introduce the dense tracking loss as an additional additive ablation—following the same "protocol three"—to assess how much improvement it brings over the subtractive ablation results and the original AnimalAvatar results.

Motion factorization We further evaluate the effectiveness of our rigid motion factorization, which separates the measured motion into the motion of the camera and the motion of the shape (Sec. 3.3). To this end, we conduct an experiment where the rigid motion g_t of each rendering camera P_t is replaced by the root rigid component g_t^0 of the SMAL deformation coefficients θ_t, i.e., $\forall t \in [1..T],\ g_t^{\text{cam}} = g_t^0$, effectively discarding the structure-from-motion (SfM) estimate g_t^{SfM}. As shown in Tab. 2, this simplification leads to a notable drop in performance across all metrics, thereby validating the necessity of our proposed rigid motion factorization.

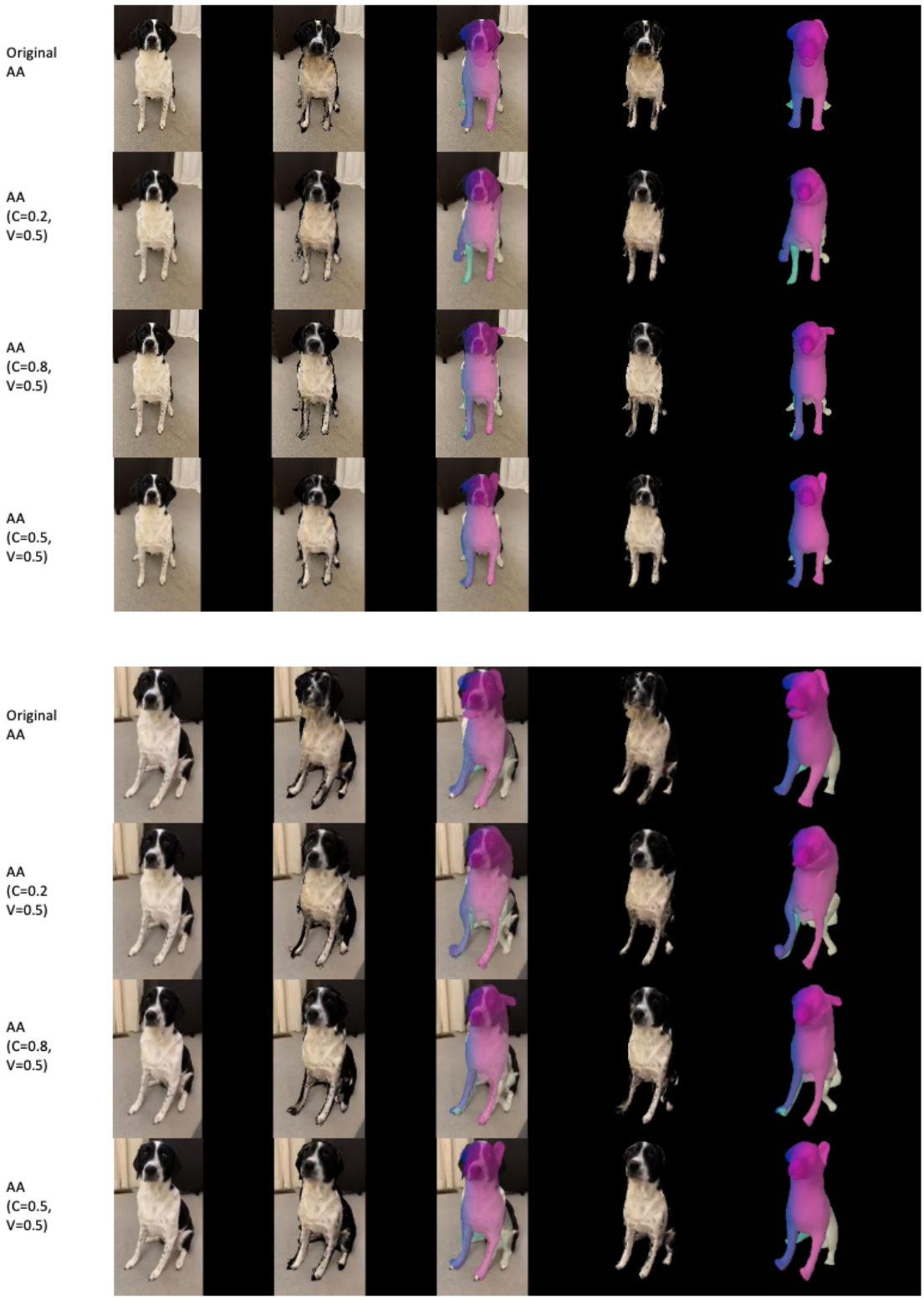

Figure 5.11: **AnimalAvatar with Dense Tracking Loss: C indicates confidence score and V indicates advisability**

Effect of dense CSE supervision Removing $\mathcal{L}_{\text{CSE}}$ degrades $\mathbf{IoU}_{\mathbf{w5}}$ more than **IoU**, indicating that dense correspondences mainly help hard poses (rear/side views). Perceptual quality also drops, suggesting that better geometry aids texture fitting.

Effect of dense tracking loss Disabling $\mathcal{L}_{\text{track}}$ enhances both gemotry and texture reconstruction, while keeping mean **IoU** nearly unchanged, showing that AllTracker [275] chiefly improves temporal coherence beyond what per-frame CSE [271] can provide.

5.5 Conclusion

In this chapter, we examine and extend AnimalAvatar by introducing a dense tracking loss and conducting more comprehensive, unbiased evaluation experiments. The goal of this work is not simply to synthesize articulated animals; rather, it aims to analyze how the method still depends on pre-defined global shape parameters, SMAL-CSE model, and camera factorization. This reliance comes from the nature of our task — overfitting a single animal video instead of building a general model.

To conclude, we extend and evaluate more data based on AnimalAvatar [9], which presents a method for recovering a textured, fully animatable 3D dog model from casually captured monocular videos. The original AnimalAvatar optimize an implicit opacity–colour texture defined in the canonical pose (duplex-mesh band), which deforms coherently with articulation and preserves fine-scale appearance. Building upon this design, our designed dense tracking loss strengthens template-based fitting by coupling the SMAL prior with dense *Continuous Surface Embeddings* (CSE), yielding reliable pixel-to-surface correspondences even under challenging viewpoints. For our dense tracking supervision, we integrate all-pixel, long-range trajectories from a dense tracker to regularize the per-frame pose, reduce temporal drift and the global shape. This complementary temporal cue, combined with original CSE's spatial constraints leads to more stable motion and sharper reconstructions. We test original AnimalAvatar with more data from CoP3D: OriginalSet, RandomSet, and DifficultSet, by using dense tracking loss, the overall approach surpasses representative template-free and template-based baselines in both geometric fidelity and photometric quality, while remaining practical for in-the-wild footage, without multi-view capture or 3D supervision. We test our dense tracking loss with scientific experiments, i.e. different confidence and vistability score settings. Future work includes handling tracking under heavy occlusions, and pretraining correspondence/texture modules on larger, diverse video datasets. In the following chapter, we conclude all three tasks and discuss future work.

5.6 Future Work

One key direction for future work is to develop a generic model capable of understanding and representing articulated animals across different species. This requires training on diverse animal species, as the SMAL model may not generalize well to unseen species due to its limited pose and shape space. A promising approach would be to build a 3D morphable model with latent variables that can enrich the pose and shape coefficients, which in turn requires training a latent-based 3D morphable model encoder.

Furthermore, the ablation studies show that the keypoint loss has limited contribution to the final performance while introducing unnecessary computational overhead. As illustrated in Figure 5.7, the visual results suggest that sparse keypoints are not sufficiently robust to capture the true skeletal structure of animals, often leading to geometric inaccuracies. These findings indicate that dense correspondence plays a more critical role, and that additional feature-based constraints are necessary for accurately learning animal geometry.

Besides, according to our work in Chapter 4, which reconstructs human faces without relying on a 3D morphable model, humans and animals share the property of inherent articulation. Building on this insight, we believe that developing a unified framework that connects both human and animal faces could be a promising direction for future work. Since both human and animal faces can be represented using parametric models for reconstruction and reenactment, such a framework could enable motion or expression transfer across species, leading to more flexible animation and a deeper understanding of articulated motion in faces in general. In the following chapter, we discuss the key insights and implications drawn from our studies on human faces and animals, and how they can inform future research.

Chapter 6

Conclusions

This chapter summarizes the key contributions of this thesis, which focuses on exploring explicit 2D and 3D motion and geometry representations of human faces, as well as synthesizing human faces and articulated animals from in-the-wild videos. Our research advances the understanding of human emotions and the synthesis of dynamic facial and animal models through the development of a multi-modal emotion recognition framework, a Splat-Portrait reconstruction model, and an extended version of the AnimalAvatar [178] method.

Despite these advancements, challenges such as full view reconstruction and expressive animation remain a problem for us to solve. In the chapter, we first summarise the contributions and insights of this thesis. Then, we discuss potential avenues for future work that incorporate ideas spanning several of the earlier chapters.

6.1 Summary of insights and contributions

We summarizes the three major works presented in this thesis, each addressing a different aspect of either human face or articulated animals. The first work focus on developing a deterministic model for classifying human emotions. We achieve it by fusing two modalities together, i.e. face images and audio segments. We design a Intra- and Inter-modal Interactive model for fusing facial and audio feature. This is more powerful than simple approaches of using single modality, and we show it achieves comparable performance on multi-modality emotion recognition task. Key contributions of this work are as follows: We introduce optical flow as an explicit representation of fine-grained facial motion and texture, and design a flow–video attention fusion module that adaptively combines flow and frame features for more detail-aware visual embeddings. We develop intra- and inter-modal enhancement modules to enable deeper bidirectional interaction between audio and visual modalities while preserving their modality-specific information. We propose a cross-attention classification

framework with independent prediction heads for each modality and their fusion, improving cross-modal consistency and supervision.

The second work focused on developing a robust method for 3D human head reconstruction and motion analysis from monocular videos. By leveraging only 2D supervision, and optimization strategies, it enabled accurate estimation of head pose, facial geometry, and appearance under audio conditions. We build a regression approach to synthesis talking portrait videos, which supports training purely from 2D images – without 3D supervision, and without using pre-defined 3D shapes from face specific prior such as 3D morphable model, landmarks and depth maps. Moreover, this model is generic, so it allows sampling new portrait and animating it condition on one arbitrary audio chunk. Main contributions of this work are summarized as follows: We present Splat-Portrait, an audio-driven talking-head framework based on 3D Gaussian Splatting (3DGS), enabling realistic geometry and motion learning directly from monocular videos. A static reconstruction branch builds canonical 3D head representations from a single portrait using Gaussian splats, with inpainting-based background refinement for seamless head–background compositing. An audio-conditioned animation module predicts per-point temporal offsets from speech features, achieving expressive and synchronized facial motion without explicit deformation fields. A two-stage self-supervised training strategy—static pretraining followed by audio conditioning—combined with Score Distillation Sampling (SDS) enhances realism and geometric fidelity under novel views. Our method requires no 3D supervision, generalizes across identities and expressions, and achieves faster rendering with higher detail than NeRF-based approaches.

Our third work explored the reconstruction and animation of articulated animals through the extension of the AnimalAvatar [9] framework and extensive evaluation experiments. This study proposed a dense tracking loss and additional evaluation on more dag videos, it should that dense tracking loss demonstrates superior performance over original AnimalAvatar method, and combine with it, it beats existing baselines in pose estimation and appearance reconstruction in a overfitting settings. More importantly, it revealed key limitations of current template-based models, such as their dependence on sparse keypoints and predefined global parameters. The insights from this work highlight the importance of developing a more generic and unified model that can handle diverse species and non-rigid motions. Our evaluation contribute new insights into how parametric modeling, dense correspondence learning, and temporal optimization can be integrated to achieve realistic, animatable 3D animal avatars from in the wild monocular videos.

Overall, these three works together advance the field of human face and animals related research, by developing from human-centric models to more general human face representation and complex articulated animal representations. The next section discusses the broader implications and future research directions inspired by these findings.

6.2 Future work

In the conclusion of the preceding chapters, we outlined potential extensions and future research for each work. We now discuss some directions for future work that draw together ideas presented in the different chapters.

6.2.1 Towards Human Emotion Recognition in 3D-aware Condition

From the work presented in Chapter 3, we explored how to fuse audio features with 2D facial images to enhance emotion recognition. As discussed in Section 2.1.8, the training datasets mainly consist of monocular videos containing front-facing human faces captured in the wild. While this setting works well for frontal views, it limits the understanding of facial dynamics under more complex 3D poses. In Chapter 3, we used optical flow to capture local motion information between consecutive frames and integrated it into our full model. Although this approach helps represent short-term facial dynamics from 2D images, it relies heavily on closely neighboring frame pairs to provide accurate motion cues. When the temporal gap between frames increases or when the viewpoint changes drastically, the estimated dynamics can become unreliable. Understanding facial dynamics in 3D space would not only provide a more complete representation of expressions under extreme viewpoints but also offer more stable cues about emotional states compared to using optical flow alone. Therefore, it would be worthwhile to investigate how facial expressions evolve across different spatial orientations in 3D space, toward building a 3D-aware Human Emotion Recognition system. This requires collecting and labelling a dataset with multi-view facial images and corresponding emotion annotations. With such data, a model trained on it, or one that optimises a 3DMM based on such data, could accurately present emotions in 3D space. Besides, exploring and incorporating 3D geometric information, such as depth maps or reconstructed 3D face meshes, could help build this system..

6.2.2 Towards Emotion-Aware 3D Head Generation

One interesting future direction for 3D head generation is to develop an emotion-aware model capable of reconstructing a static 3D head that appropriately reflects an inferred emotional state from text or audio input. For example, given an abstract auditory input such as a piece of music or a speech segment, the system could first infer the underlying emotion state (e.g., happiness, sadness, anger) using an emotion recognition module trained on multimodal emotion datasets (e.g., RAVDESS[281], CREMA-D [219]). Then, conditioned on

the predicted emotion, the model could synthesize a corresponding 3D head that expresses this emotion through facial geometry and texture adjustments.

Similarly, in a text-driven scenario, a simple sentence such as "I am happy" or "I feel nervous" could serve as input to a language-based emotion encoder (e.g., leveraging large language models or pretrained multimodal encoders like CLIP [282] or EmotiCLIP [283]) to extract the emotional intent. This emotional embedding could then be used to guide a conditional 3D head generation model, implicit neural representations, or 3D Gaussian Splatting frameworks, allowing for controllable emotional expression in the synthesized 3D head.

Such a system would bridge affective computing and 3D generative modeling, enabling applications in digital avatars, virtual humans, and emotionally responsive human–AI interaction. Achieving this goal would likely require (1) constructing or curating a large-scale dataset that aligns 3D facial geometry with emotion labels, (2) developing a disentangled latent space separating identity, pose, and emotion, and (3) designing a multimodal conditioning framework that robustly maps emotion cues from text or audio to expressive 3D head geometry.

6.2.3 Towards a Generalized Talking Head Model Across Humans and Animals

In Chapter 4, we described how to train a generic model capable of performing 3D reconstruction from human portrait images. The model is trained exclusively on monocular human videos, which enables it to reconstruct talking human faces effectively. When applied to movie production, for instance, animating cartoons or creating animated films, this model cannot be directly extended to cases such as talking animals, even though their visual appearance and motion share certain similarities with humans (e.g., facial symmetry, eye and mouth regions). This limitation stems from the lack of suitable datasets in the wild, such as "talking animals" or "talking cartoons" datasets.

As discussed in Section 2.1.4, a single-identity method [59] has attempted to perform human-to-animal face reenactment; however, it fails to generalize to diverse human–animal face mappings due to its reliance on overfitting during training. To build a generic model that can animate both talking human and animal faces, several future directions can be explored. In 2D base animation, a cross-domain representation learning framework could be introduced, for example in [284], where a shared latent space encodes motion and expression across species. By learning disentangled representations of identity, expression, and pose, the model could transfer expressive motion patterns—such as lip and jaw movements—from human speech videos to animal faces.

Besides, geometry-aware priors can be integrated to handle species-specific differences.

For example, human and animal faces could share a canonical 3D structure with corresponding regions (eyes, nose, mouth), but have species-dependent shape deformations modeled through parametric templates or deformation networks. Using weak 3D supervision from animal datasets or synthetic renderings could help bridge the gap between these morphologies and understanding their kinematic facial skeletons. Since both humans and animals produce correlated mouth dynamics with speech or vocalization, audio features could serve as a universal cue to guide lip and jaw movements across species. A cross-species motion transfer model could be trained to predict plausible mouth motion for animals by leveraging large-scale human audio-visual data and a smaller set of animal talking clips. Domain adaptation techniques such as adversarial learning or feature alignment could be applied to make the model robust to appearance variations in fur, texture, or lighting, which often differ drastically from human data. Learning a cross-domain 3D morphable model under a unified framework, where kinematic keypoints are automatically aligned across human and animal faces, could further regularize the training process and enable consistent geometry reconstruction across domains.

In summary, creating a unified framework for animating both human and animal faces would require learning shared representations of expression and motion, while respecting the unique geometric structures of different species. Such a model would enable new possibilities in cross-species animation, digital storytelling, and behavioral analysis, bridging the gap between human and animal visual communication.

Bibliography

[1] N. Rueegg, S. Zuffi, K. Schindler, and M. J. Black, "Barc: Learning to regress 3d dog shape from images by exploiting breed information," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[2] N. Rueegg, S. Tripathi, K. Schindler, M. J. Black, and S. Zuffi, "Bite: Beyond priors for improved three-dimensional dog pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[3] G. Yang, C. Wang, N. Reddy, and D. Ramanan, "Reconstructing animatable categories from videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[4] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makeittalk: Speaker-aware talking-head animation," *ACM Transactions on Graphics*, 2020.

[5] T. Liu, F. Chen, S. Fan, C. Du, Q. Chen, X. Chen, and K. Yu, "Anitalker: animate vivid and diverse talking faces through identity-decoupled facial motion encoding," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 6696–6705.

[6] Z. Li, D. Litvak, R. Li, Y. Zhang, T. Jakab, C. Rupprecht, S. Wu, A. Vedaldi, and J. Wu, "Learning the 3d fauna of the web," in *CVPR*, 2024.

[7] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[8] N. Neverova, D. Novotny, M. Szafraniec, V. Khalidov, P. Labatut, and A. Vedaldi, "Continuous surface embeddings," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 258–17 270, 2020.

[9] R. Sabathier, N. J. Mitra, and D. Novotny, "Animal avatars: Reconstructing animatable 3d animals from casual videos," in *European Conference on Computer Vision*. Springer, 2024, pp. 270–287.

[10] S. Wu, T. Jakab, C. Rupprecht, and A. Vedaldi, "Dove: Learning deformable 3d objects by watching videos," *arXiv:2107.10844*, 2021.

[11] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[12] S. Zuffi, A. Kanazawa, D. W. Jacobs, and M. J. Black, "3d menagerie: Modeling the 3d shape and pose of animals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6365–6373.

[13] G. Varma and V. S. Dhaka, "A comprehensive study of svm and hmm for facial expression recognition," *International Journal of Computer Applications*, 2019.

[14] M. H. Siddiqi, R. Ali, and Y. T. Park, "Facial expression recognition using contourlet transform and svm," *Applied Intelligence*, 2015.

[15] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention," *IEEE Trans. Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[16] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective spatio-temporal feature learning robust to expression intensity," *IEEE Trans. Affective Computing*, vol. 10, no. 2, pp. 223–236, 2019.

[17] N. Samadiani *et al.*, "A review on automatic facial expression recognition systems and datasets," *Pattern Recognition*, 2021.

[18] M. Said *et al.*, "High-resolution facial expression recognition with regional attention," in *ICPR*, 2021.

[19] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, 2018.

[20] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: a review," *Procedia Computer Science*, vol. 175, pp. 689–694, 2020.

[21] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53 040–53 065, 2019.

[22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*. PMLR, 2019, pp. 6105–6114.

[23] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[24] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing, modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[25] Y. Lin *et al.*, "Audio-visual hmm for emotion recognition," in *ICME*, 2012.

[26] J. Fernandes *et al.*, "Speech emotion recognition using svm," *Procedia Computer Science*, vol. 143, pp. 494–501, 2018.

[27] R. Khalil *et al.*, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[28] M. Jahangir *et al.*, "Deep learning driven ser: Challenges and advances," *IEEE Access*, vol. 9, pp. 167 544–167 580, 2021.

[29] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using rnns with local attention," in *ICASSP*, 2017, pp. 2227–2231.

[30] J. Zhao, X. Mao, and L. Chen, "Merged deep cnns for speech emotion recognition," *IET Signal Processing*, vol. 12, pp. 713–721, 2018.

[31] M. B. Er, "Hybrid deep and acoustic features for speech emotion classification," *IEEE Access*, vol. 8, pp. 221 640–221 653, 2020.

[32] S. Haq, P. J. B. Jackson, and J. Edge, "Human speech emotion recognition using savee," in *LLH*, 2010.

[33] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[34] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *ICDEW*, 2006, pp. 8–8.

[35] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, 2016.

[36] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[37] H. Lee *et al.*, "Audio-visual emotion recognition using bert-based fusion," *IEEE Access*, vol. 9, pp. 48 734–48 744, 2021.

[38] S. Yoon *et al.*, "Cross-modal alignment for audio-visual emotion recognition," *Pattern Recognition Letters*, 2022.

[39] S. Dobrišek *et al.*, "Towards efficient multi-modal emotion recognition," *International Journal of Advanced Robotic Systems*, vol. 10, no. 53, pp. 1–10, 2013.

[40] R. Sara *et al.*, "Peak-picking and score-level fusion for aver," *Multimedia Tools and Applications*, 2016.

[41] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in the wild," *Machine Vision and Applications*, vol. 30, pp. 975–985, 2019.

[42] Z. Farhoudi and S. Setayeshi, "Fusion of deep features with bel and temporal conv for aver," *Speech Communication*, vol. 127, pp. 92–123, 2020.

[43] G. P. Rajasekar *et al.*, "A joint cross-attention model for audio–visual fusion in dimensional er," *arXiv:2203.14779*, 2022.

[44] T. Hussain, W. Wang, N. Bouaynaya *et al.*, "Deep learning for audio-visual emotion recognition," in *International Conference on Information Fusion*, 2022, pp. 1–8.

[45] F. Noroozi *et al.*, "Survey on emotional ai: Robustness, generalization, and efficiency," *IEEE Transactions on Affective Computing*, 2019.

[46] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson, "Synsin: End-to-end view synthesis from a single image," in *CVPR*, 2020.

[47] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[48] J. Tao, B. Wang, T. Ge, Y. Jiang, W. Li, and L. Duan, "Motion transformer for unsupervised image animation," in *European Conference on Computer Vision (ECCV)*, 2022.

[49] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, "Latent image animator: Learning to animate images via latent space navigation," in *International Conference on Learning Representations (ICLR)*, 2022.

[50] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single

image talking face animation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8652–8661.

[51] Y. Pang, Y. Zhang, W. Quan, Y. Fan, X. Cun, Y. Shan, and D. Yan, "Dpe: Disentanglement of pose and expression for general video portrait editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[52] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," *arXiv preprint arXiv:2107.09293*, 2021.

[53] K. R. Prajwal, T. Afouras, J. S. Chung, and A. Zisserman, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[54] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[55] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu, "Difftalk: Crafting diffusion models for generalized audio-driven portraits animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[56] M. Stypułkowski, K. Vougioukas, S. He, M. Zieba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat gans on talking-face generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[57] S. Mukhopadhyay, S. Suri, R. T. Gadde, and A. Shrivastava, "Diff2lip: Audio-conditioned diffusion models for lip-synchronization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[58] L. Tian, Q. Wang, B. Zhang, and L. Bo, "Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions," in *European Conference on Computer Vision*. Springer, 2024, pp. 244–260.

[59] J. Luo, C. Wang, M. Vasilkovsky, V. Shakhrai, D. Liu, P. Zhuang, S. Tulyakov, P. Wonka, H.-Y. Lee, J. Davis *et al.*, "T2bs: Text-to-character blendshapes via video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 13 625–13 637.

[60] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[61] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3661–3670.

[62] L. Xie, X. Wang, H. Zhang, C. Dong, and Y. Shan, "Vfhq: A high-quality dataset and benchmark for video face super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[63] E. Burkov, I. Pasechnik, A. Grigorev, and V. Lempitsky, "Neural head reenactment with latent pose descriptors," in *CVPR*, 2020.

[64] M. C. Doukas, S. Zafeiriou, and V. Sharmanska, "Headgan: One-shot neural head synthesis and editing," in *ICCV*, 2021.

[65] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu, "Pirenderer: Controllable portrait image generation via semantic neural rendering," in *ICCV*, 2021.

[66] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *NeurIPS*, 2019.

[67] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *ICCV*, 2019.

[68] S. Athar, S. Saito, Z. Yang, S. Pidhorsky, and C. Cao, "Bridging the gap: Studio-like avatar creation from a monocular phone capture," in *ECCV*, 2024.

[69] C. Cao, T. Simon, J. K. Kim, G. Schwartz, M. Zollhoefer, S. Saito, S. Lombardi, S. Wei, D. Belko, S. Yu *et al.*, "Authentic volumetric avatars from a phone scan," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–19, 2022.

[70] P. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Nießner, and J. Thies, "Neural head avatars from monocular rgb videos," in *CVPR*, 2022.

[71] S. Ma, T. Simon, J. Saragih, D. Wang, Y. Li, F. D. la Torre, and Y. Sheikh, "Pixel codec avatars," in *CVPR*, 2021.

[72] Y. Zheng, V. F. Abrevaya, M. C. Bühler, X. Chen, M. J. Black, and O. Hilliges, "Im avatar: Implicit morphable head avatars from videos," in *CVPR*, 2022.

[73] Z. Bai, F. Tan, Z. Huang, K. Sarkar, D. Tang, D. Qiu, A. Meka, R. Du, M. Dou, S. Orts-Escolano *et al.*, "Learning personalized high quality volumetric head avatars from monocular rgb videos," in *CVPR*, 2023.

[74] S. Giebenhain, T. Kirschstein, M. Georgopoulos, M. Rünz, L. Agapito, and M. Nießner, "Mononphm: Dynamic head reconstruction from monocular videos," in *CVPR*, 2024.

[75] W. Zielonka, T. Bolkart, and J. Thies, "Instant volumetric head avatars," in *CVPR*, 2022.

[76] Y. Zheng, W. Yifan, G. Wetzstein, M. J. Black, and O. Hilliges, "Pointavatar: Deformable point-based head avatars from videos," in *CVPR*, 2023.

[77] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH*, 1999.

[78] S. Athar, Z. Xu, K. Sunkavalli, E. Shechtman, and Z. Shu, "Rignerf: Fully controllable neural 3d portraits," in *CVPR*, 2022.

[79] G. Gafni, J. Thies, M. Zollhöfer, and M. Nießner, "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction," in *CVPR*, 2021.

[80] X. Zhao, L. Wang, J. Sun, H. Zhang, J. Suo, and Y. Liu, "Havatar: High-fidelity head avatar via facial model conditioned neural radiance field," *ACM Trans. Graph.*, vol. 43, no. 1, pp. 1–16, 2023.

[81] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, 2023.

[82] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang, and L. Nie, "Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians," in *CVPR*, 2024.

[83] Z. Shao, Z. Wang, Z. Li, D. Wang, X. Lin, Y. Zhang, M. Fan, and Z. Wang, "Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting," in *CVPR*, 2024.

[84] J. Xiang, X. Gao, Y. Guo, and J. Zhang, "Flashavatar: High-fidelity head avatar with efficient gaussian embedding," in *CVPR*, 2024.

[85] Y. Xu, B. Chen, Z. Li, H. Zhang, L. Wang, Z. Zheng, and Y. Liu, "Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians," in *CVPR*, 2024.

[86] S. Qian, T. Kirschstein, L. Schoneveld, D. Davoli, S. Giebenhain, and M. Nießner, "Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians," in *CVPR*, 2024.

[87] N. Drobyshev, T. Khakhulin, A. Ivakhnenko, V. Lempitsky, and E. Zakharov, "Realistic one-shot mesh-based head avatars," in *ECCV*, 2022.

[88] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *CVPR*, 2020.

[89] W. Zielonka, T. Bolkart, and J. Thies, "Towards metrical reconstruction of human faces," in *ECCV*, 2022.

[90] Y. Deng, D. Wang, X. Ren, X. Chen, and B. Wang, "Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data," in *CVPR*, 2024.

[91] X. Li, S. De Mello, S. Liu, K. Nagano, U. Iqbal, and J. Kautz, "Generalizable one-shot 3d neural head avatar," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[92] Z. Ma, X. Zhu, G.-J. Qi, Z. Lei, and L. Zhang, "Otavatar: One-shot talking face avatar with controllable tri-plane rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 901–16 910.

[93] P. Tran, E. Zakharov, L. Ho, A. T. Tran, L. Hu, and H. Li, "Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment," in *CVPR*, 2024.

[94] A. Trevithick, M. Chan, M. Stengel, E. Chan, C. Liu, Z. Yu, S. Khamis, M. Chandraker, R. Ramamoorthi, and K. Nagano, "Real-time radiance fields for single-image portrait view synthesis," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–17, 2023.

[95] D. Rebain, M. Matthews, K. M. Yi, D. Lagun, and A. Tagliasacchi, "Lolnerf: Learn from one look," in *CVPR*, 2022.

[96] Y. Zhuang, H. Zhu, X. Sun, and X. Cao, "Mofanerf: Morphable facial neural radiance field," in *ECCV*, 2022.

[97] N. Drobyshev, J. Chelishev, T. Khakhulin, A. Ivakhnenko, V. Lempitsky, and E. Zakharov, "Megaportraits: One-shot megapixel neural head avatars," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2663–2671.

[98] T. Wang, A. Mallya, and M. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *CVPR*, 2021.

[99] S. Xu, G. Chen, Y. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo, "Vasa-1: Lifelike audio-driven talking faces generated in real time," in *ICLR*, 2024.

[100] Z. Xu, J. Zhang, J. H. Liew, W. Zhang, S. Bai, J. Feng, and M. Z. Shou, "Pv3d: A 3d generative model for portrait video generation," in *ICLR*, 2023.

[101] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision*, 2020.

[102] K. Park, U. Sinha, P. Hedman, J. T. Barron, P. P. Srinivasan, B. Mildenhall, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *ICCV*, 2021.

[103] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *CVPR*, 2021.

[104] Y. Guo, K. Chen, S. Liang, Y. Liu, and H. Bao, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *ICCV*, 2021.

[105] E. R. Chan, C. Z. Lin, M. A. Chan *et al.*, "Efficient geometry-aware 3d gans," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[106] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," in *SIGGRAPH*, 2023.

[107] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH*, 1999, pp. 187–194.

[108] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *AVSS*, 2009, pp. 296–301.

[109] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," in *CVPR*, 2016, pp. 5543–5552.

[110] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3d morphable models," *International Journal of Computer Vision*, 2017.

[111] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3d face recognition with a morphable model," in *FG*, 2008, pp. 1–6.

[112] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 426–433, 2005.

[113] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 183:1–183:14, 2015.

[114] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas, "Expression flow for 3d-aware face component transfer," in *SIGGRAPH*, 2011, pp. 60:1–60:10.

[115] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec, "The digital emily project: Photoreal facial modeling and animation," in *SIGGRAPH Courses*, 2009, pp. 12:1–12:15.

[116] H. Li, T. Weise, and M. Pauly, "Example-based facial rigging," *ACM Transactions on Graphics*, vol. 29, no. 4, pp. 32:1–32:6, 2010.

[117] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," *ACM Transactions on Graphics*, vol. 30, no. 4, pp. 77:1–77:10, 2011.

[118] S. Bouaziz, Y. Wang, and M. Pauly, "Online modeling for realtime facial animation," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 40:1–40:10, 2013.

[119] I. Kemelmacher-Shlizerman and S. M. Seitz, "Face reconstruction in the wild," in *ICCV*, 2011, pp. 1746–1753.

[120] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Total moving face reconstruction," in *ECCV*, 2014, pp. 796–812.

[121] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "What makes tom hanks look like tom hanks?" in *ICCV*, 2015, pp. 3952–3960.

[122] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid icp algorithms for surface registration," in *CVPR*, 2007, pp. 1–8.

[123] A. Salazar, S. Wuhrer, C. Shu, and F. Prieto, "Fully automatic expression-invariant face correspondence," *Machine Vision and Applications*, vol. 25, no. 4, pp. 859–879, 2014.

[124] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross, "High-quality passive facial performance capture using anchor frames," *ACM Transactions on Graphics*, vol. 30, no. 4, pp. 75:1–75:10, 2011.

[125] D. A. Hirshberg, M. Loper, E. Rachlin, and M. J. Black, "Coregistration: Simultaneous alignment and modeling of articulated 3d shape," in *ECCV*, 2012, pp. 242–255.

[126] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "A 3d morphable eye region model for gaze estimation," in *ECCV*, 2016, pp. 297–313.

[127] H. Dhamo, Y. Nie, A. Moreau, J. Song, R. Shaw, Y. Zhou, and E. Pérez-Pellitero, "Headgas: Real-time animatable head avatars via 3d gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 459–476.

[128] D. Watson, W. Chan, R. Martin-Brualla, J. Ho, A. Tagliasacchi, and M. Norouzi, "Novel view synthesis with diffusion models," *arXiv preprint arXiv:2210.04628*, 2022.

[129] F. Taubner, R. Zhang, M. Tuli, and D. B. Lindell, "Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models," *arXiv preprint arXiv:2412.12093*, 2024.

[130] A. Haque, M. Tancik, A. A. Efros, A. Holynski, and A. Kanazawa, "Instruct-nerf2nerf: Editing 3d scenes with instructions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 740–19 750.

[131] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[132] J. Baldridge, J. Bauer, M. Bhutani, N. Brichtova, A. Bunner, K. Chan, Y. Chen, S. Dieleman, Y. Du, Z. Eaton-Rosen *et al.*, "Imagen 3," *arXiv preprint arXiv:2408.07009*, 2024.

[133] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *CVPR*, 2023.

[134] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *CVPR*, 2024.

[135] S. Bahmani, X. Liu, W. Yifan, I. Skorokhodov, V. Rong, Z. Liu, X. Liu, J. J. Park, S. Tulyakov, G. Wetzstein *et al.*, "Tc4d: Trajectory-conditioned text-to-4d generation," in *ECCV*, 2024.

[136] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell, "4d-fy: Text-to-4d generation using hybrid score distillation sampling," in *CVPR*, 2024.

[137] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," in *ICLR*, 2023.

[138] J. Sun, B. Zhang, R. Shao, L. Wang, W. Liu, Z. Xie, and Y. Liu, "Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior," in *ICLR*, 2024.

[139] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," in *ICLR*, 2023.

[140] R. Gao, A. Holynski, P. Henzler, A. Brussee, R. Martin-Brualla, P. P. Srinivasan, J. T. Barron, and B. Poole, "Cat3d: Create anything in 3d with multi-view diffusion models," in *NeurIPS*, 2024.

[141] X. Chen, M. Mihajlovic, S. Wang, S. Prokudin, and S. Tang, "Morphable diffusion: 3d-consistent diffusion for single-image avatar creation," in *CVPR*, 2024.

[142] A. Tewari, T. Yin, G. Cazenavette, S. Rezchikov, J. Tenenbaum, F. Durand, B. Freeman, and V. Sitzmann, "Diffusion with forward models: Solving stochastic inverse problems without direct supervision," *Advances in Neural Information Processing Systems*, vol. 36, pp. 12 349–12 362, 2023.

[143] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 22 226–22 246, 2023.

[144] N. Müller, Y. Siddiqui, L. Porzi, S. R. Bulo, P. Kontschieder, and M. Nießner, "Diffrf: Rendering-guided 3d radiance field diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4328–4338.

[145] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[146] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV*, 2016.

[147] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," in *Proc. CVPR*, 2018, lRS2/LRS3 .

[148] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, 2014.

[149] S. Wang *et al.*, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *Proc. ACM Multimedia*, 2020, .

[150] Anonymous, "Vhq dataset," $\langle\mu\rangle$, 2021, /.

[151] . Wang *et al.*, "Talkinghead-1kh," $\langle\mu\text{œ}\rangle$, 2023, .

[152] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," in *ACM SIGGRAPH*, 2017.

[153] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. ICCV*, 2015, celebA.

[154] Y. Lee, Z. Liu *et al.*, "Celebamask-hq," https://github.com/switchablenorms/CelebAMask-HQ, 2020, .

[155] Y. Yao *et al.*, "Celebv-hq: A large-scale video dataset for high-resolution face generation," in *Proc. ECCV*, 2022.

[156] S. Zuffi, A. Kanazawa, D. W. Jacobs, and M. J. Black, "3d menagerie: Modeling the 3d shape and pose of animals," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[157] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[158] N. Rüegg, S. Zuffi, K. Schindler, and M. J. Black, "Barc: Learning to regress 3d dog shape from images by exploiting breed information," in *CVPR*, 2022, pp. 3876–3884.

[159] N. Rüegg, S. Tripathi, K. Schindler, M. J. Black, and S. Zuffi, "Bite: Beyond priors for improved 3d dog pose estimation," in *CVPR*, 2023, pp. 8867–8876.

[160] G. Yang, D. Sun, V. Jampani *et al.*, "Lasr: Learning articulated shape reconstruction from a monocular video," in *CVPR*, 2021.

[161] G. Yang, D. Sun, V. Jampani, D. Vlasic, F. Cole, C. Liu, and D. Ramanan, "Viser: Video-specific surface embeddings for articulated 3d shape reconstruction," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[162] S. Sinha, R. Shapovalov, J. Reizenstein *et al.*, "Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[163] S. Wu, R. Li, T. Jakab, C. Rupprecht, and A. Vedaldi, "Magicpony: Learning articulated 3d animals in the wild," *arXiv:2211.12497v3*, 2023.

[164] Y. Yang, Y. Deng, Y. Xu, and J. Zhang, "Aptv2: Benchmarking animal pose estimation and tracking with a large-scale dataset," *arXiv preprint arXiv:2312.10308*, 2023.

[165] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik, "Learning category-specific mesh reconstruction from image collections," in *ECCV*, 2018.

[166] X. Li, S. Liu, K. Kim, S. D. Mello, V. Jampani, M.-H. Yang, and J. Kautz, "Self-supervised single-view 3d reconstruction via semantic consistency," in *ECCV*, 2020.

[167] O. Sorkine and M. Alexa, "As-rigid-as-possible surface modeling," in *Symposium on Geometry Processing*, 2007.

[168] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics*, 2015.

[169] Y. Wang, N. Aigerman, V. G. Kim, S. Chaudhuri, and O. Sorkine-Hornung, "Neural cages for detail-preserving 3d deformations," in *CVPR*, 2020.

[170] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *CVPR*, 2019.

[171] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *CVPR*, 2019.

[172] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," in *NeurIPS*, 2021.

[173] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *CVPR*, 2020.

[174] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin, "A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose," in *NeurIPS*, 2021.

[175] X. Chen, Y. Zheng, M. J. Black, O. Hilliges, and A. Geiger, "Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes," in *ICCV*, 2021.

[176] G. Yang, M. Vo, N. Neverova, D. Ramanan, A. Vedaldi, and H. Joo, "Banmo: Building animatable 3d neural models from many casual videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[177] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler, "Deep marching tetrahedra: A hybrid representation for high-resolution 3d shape synthesis," in *NeurIPS*, 2021.

[178] R. Sabathier, N. J. Mitra, and D. Novotny, "Animal avatars: Reconstructing animatable 3d animals from casual videos," in *arXiv preprint arXiv:2403.17103*, 2024.

[179] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," in *SISY*. IEEE, 2021, pp. 119–124.

[180] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers in Robotics and AI*, vol. 7, 2020.

[181] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *ACL*, vol. 2019. NIH Public Access, 2019, p. 6558.

[182] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Self-attention fusion for audiovisual emotion recognition with incomplete data," in *ICPR*. IEEE, 2022, pp. 2822–2828.

[183] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *ICMI*, 2019, pp. 562–566.

[184] X. Ge, J. M. Jose, S. Xu, X. Liu, and H. Han, "Mgrr-net: Multi-level graph relational reasoning network for facial action unit detection," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–20, 2024.

[185] Y. Zhang, H. Wang, Y. Xu, X. Mao, T. Xu, S. Zhao, and E. Chen, "Adaptive graph attention network with temporal fusion for micro-expressions recognition," in *ICME*. IEEE, 2023, pp. 1391–1396.

[186] M. Wang, "Micro-expression recognition based on multi-scale attention fusion," in *ICDSCA*. IEEE, 2021, pp. 853–861.

[187] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[188] F. Ma, W. Zhang, Y. Li, S.-L. Huang, and L. Zhang, "An end-to-end learning approach for multimodal emotion recognition: Extracting common and private information," in *ICME*, 2019, pp. 1144–1149.

[189] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-GCN: Correlation based graph convolutional network for audio-video emotion recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 3793–3804, 2020.

[190] Y. Yin, L. Jing, F. Huang, G. Yang, and Z. Wang, "Msa-gcn: Multiscale adaptive graph convolution network for gait emotion recognition," *Pattern Recognition*, p. 110117, 2023.

[191] L. Goncalves and C. Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in *ICASSP*. IEEE, 2022, pp. 7357–7361.

[192] Y. Gong, A. H. Liu, A. Rouditchenko, and J. Glass, "Uavm: Towards unifying audio and visual models," *IEEE Signal Processing Letters*, vol. 29, pp. 2437–2441, 2022.

[193] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, "Versatile audio-visual learning for handling single and multi modalities in emotion regression and classification tasks," *arXiv preprint arXiv:2305.07216*, 2023.

[194] L. Tarantino, P. N. Garner, A. Lazaridis *et al.*, "Self-attention for speech emotion recognition." in *Interspeech*, 2019, pp. 2578–2582.

[195] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *ICASSP*. IEEE, 2021, p. 5749–5753.

[196] N. Perveen, D. Roy, and K. M. Chalavadi, "Facial expression recognition in videos using dynamic kernels," *Trans. Image Process.*, vol. 29, pp. 8316–8325, 2020.

[197] X. Ge, P. Wan, H. Han, J. M. Jose, Z. Ji, Z. Wu, and X. Liu, "Local global relational network for facial action units recognition," in *FG*. IEEE, 2021, pp. 01–08.

[198] M. Hu, P. Ge, X. Wang, H. Lin, and F. Ren, "A spatio-temporal integrated model based on local and global features for video expression recognition," *The Visual Computer*, pp. 1–18, 2021.

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

[200] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6897–6906.

[201] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 616–17 627, 2021.

[202] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7660–7669.

[203] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen, "Eeg based emotion recognition: A tutorial and review," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–57, 2022.

[204] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of bert-based approaches," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021.

[205] X. Wang, L. Kou, V. Sugumaran, X. Luo, and H. Zhang, "Emotion correlation mining through deep learning models on natural language text," *IEEE transactions on cybernetics*, vol. 51, no. 9, pp. 4400–4413, 2020.

[206] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 49–67, 2021.

[207] K. Shrivastava, S. Kumar, and D. K. Jain, "An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network," *Multimedia tools and applications*, vol. 78, pp. 29 607–29 639, 2019.

[208] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 234–14 243.

[209] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5826–5846, 2021.

[210] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 02, 2020, pp. 1359–1367.

[211] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: Multi-modal fusion network for emotion recognition in conversation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4652–4661.

[212] L. Goncalves and C. Busso, "Learning cross-modal audiovisual representations with ladder networks for emotion recognition," in *ICASSP*. IEEE, 2023, pp. 1–5.

[213] B. Maji, M. Swain, R. Guha, and A. Routray, "Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[214] A. Kemmou, A. El Makrani, I. El Azami, and M. H. Aabidi, "Automatic facial expression recognition under partial occlusion based on motion reconstruction using a denoising autoencoder," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 1, pp. 276–289, 2024.

[215] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[216] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer *et al.*, "Perceiver IO: A general architecture for structured inputs & outputs," *arXiv preprint arXiv:2107.14795*, 2021.

[217] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *International Conference on Automatic Face and Gesture Recognition*, 2018.

[218] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *NeurIPS*, vol. 28, 2015.

[219] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, 2014.

[220] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017.

[221] P. Morgado, I. Misra, and N. Vasconcelos, "Audio-visual instance discrimination with cross-modal agreement," in *CVPR*, 2021.

[222] J. Li *et al.*, "Micro-expressions: An overview," *Information*, vol. 16, no. 10, p. 876, 2025.

[223] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, J. He, H. Liu, and Z. Fan, "Emotalk: Speech-driven emotional disentanglement for 3d face animation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 20 687–20 697.

[224] T. He, J. Guo, R. Yu, Y. Wang, J. Zhu, K. An, L. Li, X. Tan, C. Wang, H. Hu *et al.*, "Gaia: Zero-shot talking avatar generation," *arXiv preprint arXiv:2311.15230*, 2023.

[225] Z. Ye, T. Zhong, Y. Ren, J. Yang, W. Li, J. Huang, Z. Jiang, J. He, R. Huang, J. Liu *et al.*, "Real3d-portrait: One-shot realistic 3d talking portrait synthesis," *arXiv preprint arXiv:2401.08503*, 2024.

[226] W. Li, L. Zhang, D. Wang, B. Zhao, Z. Wang, M. Chen, B. Zhang, Z. Wang, L. Bo, and X. Li, "One-shot high-fidelity talking-head synthesis with deformable neural radiance field," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 969–17 978.

[227] C. Liu, "An analysis of the current and future state of 3d facial animation techniques and systems," 2009.

[228] T. Ye, Y. Zhang, M. Jiang, L. Chen, Y. Liu, S. Chen, and E. Chen, "Perceiving and modeling density for image dehazing," in *European conference on computer vision*. Springer, 2022, pp. 130–145.

[229] J. Saunders and V. Namboodiri, "Dubbing for everyone: Data-efficient visual dubbing using neural rendering priors," *arXiv preprint arXiv:2401.06126*, 2024.

[230] M. Xu, H. Li, Q. Su, H. Shang, L. Zhang, C. Liu, J. Wang, Y. Yao, and S. Zhu, "Hallo: Hierarchical audio-driven visual synthesis for portrait image animation," *arXiv preprint arXiv:2406.08801*, 2024.

[231] T. Liu, Z. Ma, Q. Chen, F. Chen, S. Fan, X. Chen, and K. Yu, "Vqtalker: Towards multilingual talking avatars through facial motion tokenization," *arXiv preprint arXiv:2412.09892*, 2024.

[232] Z. Peng, W. Hu, Y. Shi, X. Zhu, X. Zhang, H. Zhao, J. He, H. Liu, and Z. Fan, "Synctalk: The devil is in the synchronization for talking head synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 666–676.

[233] H. Yu, Z. Qu, Q. Yu, J. Chen, Z. Jiang, Z. Chen, S. Zhang, J. Xu, F. Wu, C. Lv *et al.*, "Gaussiantalker: Speaker-specific talking head synthesis via 3d gaussian splatting," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3548–3557.

[234] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu, "Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting," in *European Conference on Computer Vision*. Springer, 2025, pp. 127–145.

[235] K. Cho, J. Lee, H. Yoon, Y. Hong, J. Ko, S. Ahn, and S. Kim, "Gaussiantalker: Real-time high-fidelity talking head synthesis with audio-driven 3d gaussian splatting," *arXiv preprint arXiv:2404.16012*, 2024.

[236] J. Cha, S. Yoon, V. Strizhkova, F. Bremond, and S. Baek, "Emotalkinggaussian: Continuous emotion-conditioned talking head synthesis," *arXiv preprint arXiv:2502.00654*, 2025.

[237] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42.

[238] S. Aneja, A. Sevastopolsky, T. Kirschstein, J. Thies, A. Dai, and M. Nießner, "Gaussianspeech: Audio-driven gaussian avatars," *arXiv preprint arXiv:2411.18675*, 2024.

[239] X. Chu and T. Harada, "Generalizable and animatable gaussian head avatar," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=gVM2AZ5xA6

[240] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[241] F. Taubner, R. Zhang, M. Tuli, and D. B. Lindell, "Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2025, pp. 5318–5330.

[242] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 123–16 133.

[243] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in neural information processing systems*, vol. 35, pp. 26 565–26 577, 2022.

[244] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 039–10 049.

[245] G. Kim, K. Seo, S. Cha, and J. Noh, "Nerffacespeech: One-shot audio-driven 3d talking head synthesis via generative prior," *arXiv preprint arXiv:2405.05749*, 2024.

[246] J. Wang, J.-C. Xie, X. Li, F. Xu, C.-M. Pun, and H. Gao, "Gaussianhead: Impressive head avatars with learnable gaussian diffusion," *arXiv preprint arXiv:2312.01632*, 2023.

[247] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5784–5794.

[248] S. Yao, R. Zhong, Y. Yan, G. Zhai, and X. Yang, "Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering," *arXiv preprint arXiv:2201.00791*, 2022.

[249] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8649–8658.

[250] A. Rivero, S. Athar, Z. Shu, and D. Samaras, "Rig3dgs: Creating controllable portraits from casual monocular videos," *arXiv preprint arXiv:2402.03723*, 2024.

[251] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[252] M. Stypułkowski, K. Vougioukas, S. He, M. Zieba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat gans on talking-face generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5091–5100.

[253] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: speaker-aware talking-head animation," *ACM Transactions On Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[254] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[255] S. Szymanowicz, C. Rupprecht, and A. Vedaldi, "Splatter image: Ultra-fast single-view 3d reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 10 208–10 217.

[256] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[257] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[258] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[259] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[260] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[261] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[262] X. Chu, N. Goswami, Z. Cui, H. Wang, and T. Harada, "Artalk: Speech-driven 3d head animation via autoregressive model," 2025. [Online]. Available: https://arxiv.org/abs/2502.20323

[263] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[264] A. A. A. Osman, T. Bolkart, and M. J. Black, "Star: A sparse trained articulated human body regressor," in *European Conference on Computer Vision*, 2020.

[265] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4d: Reconstructing and tracking humans with transformers," in *International Conference on Computer Vision*, 2023.

[266] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[267] T.-Y. Lin, M. Maire, S. Belongie *et al.*, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[268] K. Ehsani, H. Bagherinezhad, J. Redmon, R. Mottaghi, and A. Farhadi, "Who let the dogs out? modeling dog behavior from visual data," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[269] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained visual categorization," in *CVPR Workshop on Fine-Grained Visual Categorization*, 2011.

[270] S. Liu, T. Li, W. Chen, and H. Li, "Soft rasterizer: A differentiable renderer for image-based 3d reasoning," in *International Conference on Computer Vision*, 2019.

[271] N. Neverova, D. Novotný, and A. Vedaldi, "Continuous surface embeddings," in *Advances in Neural Information Processing Systems*, 2020.

[272] J. L. Schoenberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[273] A. Noguchi, X. Sun, and T. Harada, "Conrf: Controllable neural radiance fields," in *European Conference on Computer Vision*, 2022.

[274] S. Peng, J. Dong, Q. Wang, S. Zhang, H. Bao, and X. Zhou, "Animatable neural radiance fields for modeling dynamic human performances," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[275] A. W. Harley, Y. You, X. Sun, Y. Zheng, N. Raghuraman, Y. Gu, S. Liang, W.-H. Chu, A. Dave, S. You *et al.*, "Alltracker: Efficient dense point tracking at high resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 5253–5262.

[276] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," 2010.

[277] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015.

[278] N. Karaev, I. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker3: Simpler and better point tracking by pseudo-labelling real videos," 2024.

[279] B. Biggs, T. Roddick, A. Fitzgibbon, and R. Cipolla, "Creatures great and smal: Recovering the shape and motion of animals from video," in *ACCV*, 2018.

[280] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[281] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[282] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[283] K. Zhou, B. Zhang, Y. Chen, Y. Xu, and X. Li, "Emoticlip: Multimodal emotion recognition using clip-based affective representations," *arXiv preprint arXiv:2302.00923*, 2023.

[284] Y. Ma, H. Liu, H. Wang, H. Pan, Y. He, J. Yuan, A. Zeng, C. Cai, H.-Y. Shum, W. Liu *et al.*, "Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–12.