University of Glasgow | School of Computing Science

Master of Science(R)(SE) Dissertation

GaitTriViT and GaitVViT: Transformer-based methods emphasizing spatial or temporal aspects in Gait Recognition

Hongyun Sheng

December 31, 2023

Abstract

In image recognition tasks, subjects with long distance and low resolution remains a challenge, whereas Gait Recognition, identifying subjects by walking patterns, is considered one of the most promising biometric technologies due to the stability and efficiency. Previous Gait Recognition methods mostly focused on constructing a sophisticated model structure to better extract spatial and temporal features from frame sequences, aiming to increase the distinctiveness between different feature representations for better model performance during evaluation. Moreover, these methods primarily based on traditional Convolutional Neural Networks (CNNs) due to the dominance of CNNs in Computer Vision.

However, since the alternative form of Transformer, named Vision Transformer, which originally has a wide application in Natural Language Processing (NLP), has introduced into Computer Vision field, the Vision Transformer has gained a strong attention by the outstanding performance in various tasks. Thus, unlike previous methods mainly based on Convolutional Neural Networks (CNNs), this project introduces two Transformer-based method: a completely Vision Transformer-based gait recognition method GaitTriViT and a Video Vision Transformer-based method GaitVViT. The GaitTriViT leveraging Vision Transformer to gain more fine-grained spatial features, while GaitVViT enhances the capacity of temporal extraction. This work evaluates their performances on two of the most popular benchmarks. The results show the still-existing gaps, and several encouraging outperforms compared with current State-of-the-Art (SOTA), demonstrating the difficulties and challenges these Transformer-based methods will encounter continuously. But I still believe in the promising future of Vision Transformers in the field of Gait Recognition.

Education Use Consent

I hereby grant my permission for this project to be stored, distributed and shown to other University of Glasgow students and staff for educational purposes. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Signature: Hongyun Sheng Date: 31 December 2023

Contents

1 Introduction

Gait is defined as the physical and behavioral biological characteristics exhibited by a human when walking upright. It can be used to describe an individual's walking pattern, and gait recognition is a technology that identifies individuals based on their distinct walking patterns. There are other biometric features including faces, fingerprints and irises, etc., but the superiority of gait lies in its ability to be easily captured from a long distance and the complete unobtrusiveness without any subject cooperation or contact for data acquisition. This makes gait recognition highly promising in real-world applications (Nixon and Carter 2006).

The attractiveness of gait recognition for identification purposes is high. For example, many video surveillance systems can only capture a low-resolution video with bad lighting conditions. When recognizing bank robbers, they may wear masks so faces are invisible, they may wear gloves so fingerprints are unavailable, and they may also wear hats where hairiness with DNA is no place to find, but they always need to walk or run, where gait can be easily captured. In these cases above, gait recognition might be the only possible choice for automatic recognition (Makihara et al. 2020).

Gait recognition research is currently under transition from evaluation stage to application stage. It could be used in applications including forensics, social security, immigration control, and video surveillance. In several criminal cases, gait recognition has been adopted as evidence for conviction. Back to 2011, one forensic study has already used gait features to provide evidence for identification (Bouchrika et al. 2011). There are also court believes the gait analysis could be a very valuable tool (Larsen et al. 2008). In Japan, a gait verification system for criminal investigation has been developed. The system is under a trial phase by the National Research Institute of Police Science (Iwama et al. 2013). Biometric tunnel proposed by Seely et al. (2008) led to the first live demonstration of gait as a biometric and maybe still could be the most promising future route of gait recognition in deployment like access control. And the first commercial software of gait recognition has been released by Watrix in Oct. 2018, which was developed by the Institute of Automation, Chinese Academy of Sciencses (CASIA). Users can present two videos, one as gallery and one as probe, to the software, then the software will output the match results.

As a task of recognition, obtaining real, effective, and distinctive representations from target data is the primary goal. However, gait recognition faces many challenges in practical applications due to several factors e.g. self-occlusion, viewing angles, walking status and carrying conditions like bringing a bag (Fan et al. 2023; Sepas-Moghaddam and Etemad 2022; Wan et al. 2018). As a task with extensive application prospects, these challenges urgently need to be addressed.

Currently, there are various gait recognition methods e.g. Gait Energy Image (GEI) by Han and Bhanu (2005), GaitSet (Chao et al. 2018), GaitPart (Fan et al. 2020) and GaitGL (Lin et al. 2022). They all tend to make improvements on traditional Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures (Fan et al. 2023; Sepas-Moghaddam and Etemad 2022). They use more sophisticated structures and deeper neural network layers to obtain an improved performance in extracting representative features. This choice is popular as CNN-based methods currently dominate the Computer Vision field, achieving remarkable results in image and video tasks that were beyond the reach of previous deep neural networks (Dosovitskiy et al. 2020).

However, Vision Transformer (ViT) methods introduced by Dosovitskiy et al. (2020) have recently made astonishing progress in tasks e.g. object detection (Carion et al. 2020), image segmentation (Chen et al. 2021) and image classification (Hong et al. 2022), and researchers are continually enhancing the performance, proposing many advanced novel architecture e.g. Swin Transformer by Liu et al. (2021) and VideoMAE by Tong et al. (2022), to endow ViT with more capabilities and potential. Moreover, the unique multi-head attention mechanism and the absence of down-sampling operations allow it to obtain finer-grained spatial features at the frame level in video-based recognition tasks. In contrast, these crucial fine-grained features are often blurred and lost in CNN-based methods due to multiple pooling and convolution operations (Alsehaim and Breckon 2022).

The combination of patch division and multi-head attention mechanism in ViT not only retains the ability to extract features from small regions but also possesses long-range dependencies. This aligns perfectly with the requirements of gait recognition tasks that focus on both local and global features simultaneously (Hou et al. 2022). Furthermore, viewing frames as patches with changes in scale allows the basic ViT structure to achieve sequence-level temporal attention, e.g. Video Vision Transformer(VViT) (Neimark et al. 2021), which can also be advantageous for the gait recognition task (Neimark et al. 2021; Arnab et al. 2021; Liu et al. 2021). Thus, two novel methods leveraging Vision Transformer technology are presented in this paper to tackle the gait recognition tasks.

Our work introduced two Transformer-based Gait Recognition model: GaitTriViT and GaitVViT. For GaitTriViT, it consists of a backbone for frame-level feature extraction, following by two parallel Transformer-based branch. One is Local Part Spatial Branch built for extraction of fine-grained set-level features in local regions, another is Global Temporal Branch built for extraction and aggregation of global features with temporal attention (Rao et al. 2018; Zhang et al. 2020; Fu et al. 2019). The final part of the model is multiple heads for classification, then a fusion loss function is used to optimize the model. The technology employed including Vision Transformer (Dosovitskiy et al. 2020), Temporal Clip Shift and Shuffle (TCSS) by Alsehaim and Breckon (2022). For GaitVViT, it adopts the technology from GaitGL by Lin et al. (2022) to construct the backbone structure, and the backbone is connected to a Video Vision Transformer Network (Arnab et al. 2021; Neimark et al. 2021) to build the final structure. After the backbone generate the first-step spatial frame-level features, the Video Vision Transformer models the features along the temporal dimension and generate the final features and predicted labels. Ideas of part-dependent (Fan et al. 2020) and frame set (Chao et al. 2018) are also introduced. The proposed methods are tested on two popular benchmarks: CASIA-B (Yu et al. 2006) and OUMVLP (Takemura et al. 2018).

In this work, several contributions are made as shown below:

- A novel method GaitTriViT based on Vision Transformer rather than the traditional CNN-based method. The global features and local features are both emphasized, along with the combination of spatial and temporal attention.
- In GaitTriViT, the camera angle and walking status of subjects from different frame sequences are labeled and incorporated into the Vision Transformer Block in Backbone during patch embedding phase, along with original position embedding, which are intended to enhance the robustness of Gait Recognition when facing challenges e.g. cross-view and multiple walking status.
- A novel method GaitVViT use Video Vision Transformer as temporal aggregator. Emphasizing on the temporal feature extraction, the method dedicates to enhance the temporal modeling performance of common framework.
- The evaluation of proposed methods on two popular benchmarks and the comparison to state-of-the-art indicate the challenges and potential for Transformer-based model in gait recognition tasks.

2 Related Works

2.1 Gait Recognition

In recent research, gait recognition methods can be broadly categorized into two main classes: model-based and appearance-based (Hou et al. 2022; Fan et al. 2023; Santos et al. 2023; Fan et al. 2022). Model-based methods estimate the underlying human body structures from the raw data and use them as input, e.g. 2D/3D poses (Cao et al. 2016; Roy et al. 2012; Martinez et al. 2017; Liao et al. 2020) and the SMPL model (Loper et al. 2015). In contrast, appearance-based methods favor directly extracting feature representations of human walking patterns from gait silhouettes. Due to the challenges of gait recognition tasks, which often involve long distances and low resolutions (Nixon and Carter 2006), recent studies have emphasized the practicality of appearance-based methods for their robustness (Fan et al. 2022). Among the various appearance-based approaches, feature extraction can be discussed from three perspectives: Spatial, Temporal, and Transformer.

2.1.1 Spatial Feature Extraction

In gait recognition research, the introduction of deep convolutional neural networks was pioneered by Wu et al. (2017), they studied an approach to gait based human identification via similarity learning by deep CNNs, aiming to recognize the most discriminative changes of gait patterns which suggest the change of human identity with a pretty small group of labeled multi-view human walking videos.

Subsequently, GaitSet, proposed by Chao et al. (2018), they adopted the strategy of dividing feature maps into strips from prior person re-identification researches, enhancing the description of the human body. It has been adopted by many following researchers ever since; GaitPart introduced by Fan et al. (2020) pushes forward the part-based concept further, presenting a part-dependent approach, they argued that the part-based schemas applied in gait recognition should be part-dependent rather than part-independent, because despite there are significant differences among human body parts in terms of appearance and moving patterns in gait cycle, it is highly possible that different parts of human body share the common attributes, e.g., color and texture. Thus, the parameters are designed part-dependent in FConv (Focal Convolution) layers to generate the fine-grained spatio-temporal representations; GaitGL developed by Lin et al. (2020; 2022) elaborated the disadvantage of extraction from either global appearances or local regions of humans only. They argued the representations based on global information often neglect the details of the gait frame, while local region based descriptors cannot capture the relations among neighboring regions, thus reducing their discriminativeness. Thus, they effectively combined global visual features and local region details, demonstrating the necessity to address both aspects simultaneously; SMPLGait by Zheng et al. (2022) aims to explore dense 3D representations for gait recognition in the wild. Leveraged the human body mesh to acquire three-dimensional geometric information, they proposed a novel framework to explore the 3D Skinned Multi-Person Linear (SMPL) model of the human body for gait recognition; MetaGait designed by Dou et al. (2023) argued that there are still conflicts between the limited binary silhouette and numerous covariates with diverse scales. Their model can learn an omni sample adaptive representation by the injected meta-knowledge in a calibration network of the attention mechanism, which

could guide the model to perceive sample-specific properties; also Fan et al. (2022), in their code repository OpenGait, drew insights from previous state-of-the-art methods and introduced GaitBase, which achieved excellent results. These studies often stack deeper convolutional layers or complex architecture to capture fine-grained, more robust, and discriminative features, to meet the various challenges of gait recognition tasks.

2.1.2 Temporal Feature Aggregation

The temporal modeling has consistently remained a significant focus in gait recognition tasks due to the inherent periodicity of walking patterns in the time dimension, i.e., gaits are repeating loops. Presently, there are three popular directions in existing research: 3DCNN-based, Set-based, and LSTM-based approaches.

Among 3DCNN-based methods Wolf et al. (2016) and Tran et al. (2015) directly employ 3D Convolutional Neural Networks to extract spatio-temporal features from sequential data. They indicated that 3D Convolutional Networks are more suitable for spatio-temporal feature learning compared to 2D and a homogeneous architecture with small $3 \times 3 \times 3$ convolution kernels in all layers is among the best performing architectures for 3D Convolutional Networks. However, this approach often encounters training difficulties and yields suboptimal performance; Set-based methods view frames within a cycle as an unordered set since human can easily identify a subject from a shuffled gait sequence. Furthermore, due to the short duration of each gait cycle, long-range dependencies and duplicate gait cycles are considered redundant. Take GaitSet (Chao et al. 2018) for example, in contrast to prior gait recognition methods which utilize the frames either a gait template or a gait sequence, they argued that the temporal information is hard to preserve in template, while the sequence keeps extra unnecessary sequential constraints and thus has low flexibility. So, they present a novel perspective regarding gait as a set consisting of independent frames. Their method is immune to permutation of frames and can naturally integrate frames from different videos under different scenarios. These Set-based methods typically learn spatial features frame by frame and then perform temporal aggregation at the set-level. On the other hand, LSTM-based methods like GaitNet by Zhang et al. (2019) argued that for each video frame, the current feature only contains the walking pose of the person in a specific instance, which can share similarity with another specific instance of a very different person. Therefore, modeling its temporal change is critical. That is where temporal modeling architectures like the recurrent neural network or long short-term memory (LSTM) work best. They use a three-layer LSTM network to extract ordered sequence features. These LSTM-based methods are capable of capturing features between consecutive frames, often yielding slightly better performance. However, they lack efficiency and robustness to noise, therefore, many researchers still prefer set-based approaches.

2.1.3 Attempts with Transformer

Multiple works had tried to tackle gait recognition task by introducing the Vision Transformer (Dosovitskiy et al. 2020). For example, Gait-ViT by Mogan et al. (2022) emphasized the lack of attention mechanism in Convolutional Neural Networks despite their well performance in image recognition tasks. The attention mechanism encodes information in the image patches, which facilitates the model to learn the substantial features in the specific regions. Thus, this work employs the Vision Transformer (ViT) integrated an attention mechanism naturally. However, they used the gait energy image (GEI) to model the temporal dimension by averaging the images over the gait cycle; Pinić et al. (2022) proposed a self-supervised learning (SSL) approach to pre-train the feature extractor, which is a Vision transformer architecture using the DINO model (Caron et al. 2021) to automatically learn useful gait features; Cui and Kang (2022) proposed GaitTransformer. They used a Multiple-Temporal-Scale Transformer (MTST), which consists of multiple transformer encoders with multi-scale position embedding, to model various long-term

temporal information of the sequence. Furthermore, e.g. Yang et al. (2023) and Zhu et al. (2023) also explore the vision transformer in gait recognition. However, the transformer-based methods have not outperformed CNN-based methods on the popular testing benchmarks and other challenging in-the-wild gait datasets (Fan et al. 2023).

2.2 Person Re-Identification

Person Re-Identification (Re-ID) is another research field similar to Gait Recognition. By definition, in Person Re-identification tasks, when being presented with a person-of-interest (query), person re-ID tells whether this person has been observed in another place (time) by another camera. From the perspective of computer vision, the most challenging problem in re-ID is how to correctly match two images of the same person under intensive appearance changes, such as lighting, pose, and viewpoint (Zheng et al. 2016). In another words, the Gait Recognition can be regard as a subset of person re-ID leveraging gait as inputs. There are many Transformer-based methods have been proposed, e.g. VID-Trans-ReID by Alsehaim and Breckon (2022) and TransReID by He et al. (2021). The shuffle operation on feature map and the choice of loss functions introduced in these method also served as inspiration in this paper.

3 Proposed Method

3.1 Introduction

Having studied the previous works, two novel transformer-based Gait Recognition methods are proposed: GaitTriViT and GaitVViT. GaitTriViT integrates the strengths of Vision Transformer (ViT) and incorporates excellent ideas from previous works and recent advancements in related fields. It places emphasis on both global and local regions, considering both temporal and spatial dimensions. Furthermore, several modifications are also made in this work differing from the common Gait Recognition frameworks. GaitVViT enhances the temporal modeling ability of common framework leveraging Video Vision Transformer (VViT), which works as a novel Temporal Pooling (TP) module.

3.2 Common Framework

Recent studies indicate a common framework in various Gait Recognition tasks (Fan et al. 2023), as shown in Figure 3.1. This framework abstracts complex structures into multiple modules, omitting internal details. The backbone maps the input gait sequence to features, typically used to extract frame-level spatial information. The Temporal Pooling (TP) module then aggregates feature maps along the temporal dimension, with operations e.g. Max Pooling (Fan et al. 2020), Recurrent Neural Networks (Jianhua et al. n.d.; Tran et al. 2021).

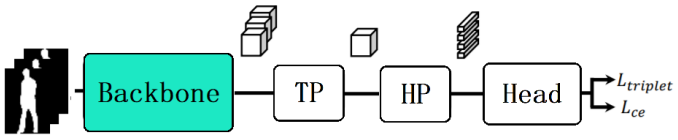

Figure 3.1: *From left to right: Inputs are silhouette sequence; Backbone Network maps inputs to feature embeddings; TP stands for Temporal Pooling to aggregate temporal dimension; HP stands for Horizontal Pooling to treat feature map as divided parts; the last part is Classification Head and loss function.*

Subsequently, the Horizontal Pooling (HP) module divides the feature map into several different parts in the horizontal direction, in line with the part-dependent concept introduced by Fan et al. (2020), and processes them independently. The Head may include several fully connected layers to obtain predicted labels, and it may also have a BNNneck (batch normalization neck, Luo et al. (2020)) to map the features to different spaces before calculating the loss. Finally, both triplet loss (Hoffer and Ailon 2015; Hermans et al. 2017) and cross-entropy loss (Rubinstein and Kroese 2004; De Boer et al. 2005) are used to optimize the model simultaneously.

3.3 GaitTriViT

3.3.1 Pipeline

Differing from the common framework of gait recognition, our method treats the original serial Temporal Pooling (TP) and Horizontal Pooling (HP) modules as two separate and parallel branches, as shown in Figure 3.2. The overall structure can be divided into several modules, including the backbone, local part spatial branch, global temporal branch, BNNeck head (Luo et al. 2020), and optimizer. Aligned silhouettes are fed into the Transformer-based backbone first, after the feature extraction, the feature maps go separately into two parallel branches. The local part spatial branch works to extract the spatial features focused in different local parts, and the global temporal branch works to model the spatio-temporal features with proper temporal informations. Those local and global features are delivered to classification heads with BNNeck (Luo et al. 2020) to generate the discriminative final feature representations and predicted labels, which will be used to calculate the losses for model optimization.

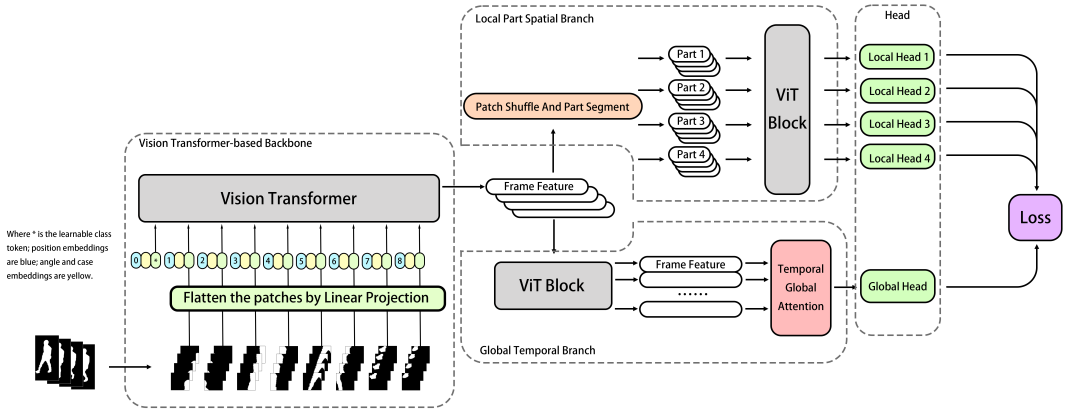

Figure 3.2: *From left to right: Inputs are cut into patches and fed into Vision Transformer-based Backbone; the feature generated go to two parallel branches, the upper one is Local Part Spatial Branch for detailed local feature extraction; the branch below are Global Temporal Branch to obtain feature in frame-bundle-level; then multiple Heads will map them and send to conduct Loss.*

3.3.2 Backbone

In this work, I use a Vision Transformer (Dosovitskiy et al. 2020) rather than any traditional Convolutional Neural Networks to build the backbone to extract frame-level spatial features. Because the Vision Transformer is more compact in contrast to a multi-layers CNN when they need to achieve similar performance, and Vision Transformer is full of potentiality in computer vision field. The original gait silhouette is in the form of a frame sequence $V_i = \{F_0, F_1, ..., F_t\}$, where each frame, after data-rearrangement and pre-processing, is in the form of $F_j \in \mathbb{R}^{H \times W \times C}$, with H, W, and C representing the height, width, and channels of the frame image, respectively. Each frame is divided into multiple patches of the same size as the original paper does (Dosovitskiy et al. 2020), i.e., $F_j = \{P_0, P_1, ..., P_n\}$. However, drawing inspiration from works by He et al. (2021) and Wang et al. (2022), this work also adopts their patch embedding strategy of allowing patches to overlap with each other. This approach helps the model to focus on local information while strengthening the connections between adjacent patches and reducing feature loss at the patch edges, which meets the need of this task to focus on body parts while not ignoring the constraints among each body parts.

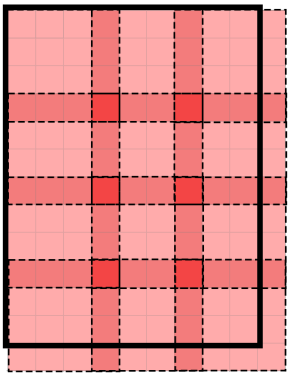

Figure 3.3: *When cutting images into patches, different from traditional method, the overlap strategy is adopted for more robust performance.*

$$N = \frac{H + d - s}{s} \times \frac{W + d - s}{s} \tag{3.1}$$

In (3.1), N is the number of divided patches, d is the patch size, and s is the stride length. After the patches are generated, we need to flatten them into tensors of 1-D dimensions using linear projection ℓ. Moreover, a learnable class token P^j_{cls} is inserted at the head position to represent the overall features of this frame.

$$F_j = \left[P^j_{cls}; \ell(P^j_0); \ell(P^j_1); ...; \ell(P^j_N)\right] \tag{3.2}$$

$$E_j = F_j + \lambda_1 E_{pos} + \lambda_2 E_{angle} + \lambda_3 E_{case} \tag{3.3}$$

Following the Vision Transformer original paper (Dosovitskiy et al. 2020), I add a learnable position embedding $E_{pos} \in \mathbb{R}^{N+1 \times D}$ representing spatial position of the patch to each patch. Furthermore, due to the challenges posed by cross-view and different walking status in appearance-based gait recognition tasks, we manually incorporate information that represents different subject appearances and various camera angles into the patch embedding. Many studies have demonstrated the effectiveness of this operation e.g. researches by He et al. (2021) and Alsehaim and Breckon (2022), they indicated these lightweight learnable embeddings perform well tackling cross-view and cross-status tasks. For example, the current frame sequence is selected from a video where a subject is captured by a camera at front while carrying a bag, which means the camera angle is 0 and walking status is bag carrying. Similar to position embedding, we introduce case embedding $E_{case} \in \mathbb{R}^{c \times D}$ and angle embedding $E_{angle} \in \mathbb{R}^{a \times D}$, c is the total number of existing walking situations, a stands for the total number of different camera angles. Then, we add these four altogether in proportions denoted by λ_1, λ_2, and λ_3, where we generate the final patch embedding E_j.

3.3.3 Local Part Spatial Branch

For each frame sequence representing a unique subject ID with unique camera angle and walking status, only several frames are selected in one batch, which is regarded as a frame bundle. For each frame bundle B that has undergone processing by the backbone, it now exists as follows:

$$B = \left[F_0\left\{P^0_{cls}, P^0_0, ..., P^0_N\right\}; ...; F_T\left\{P^T_{cls}, P^T_0, ..., P^T_N\right\}\right] \tag{3.4}$$

The bundle is sent to two branches, one of which is the local part spatial branch that will be discussed in this section. It corresponds to the Horizontal Pooling (HP) module in the gait recognition common framework and is used to extract spatial features at the frame set level within the bundle. Chao et al. (2018) and their proposed method GaitSet suggests that gait recognition does not require long-term dependencies and treating gait frames as a set can improve model robustness. As our method has two separate branches to process on the same feature representations in parallel, each branch can go further on their own specialized work and has no worries to affect another branch. In local part branch, the work is all about spatial local features. Therefore, we merge patches belonging to different frames but at the same position within the bundle to a group G as follows:

$$\hat{B} = \left[G_{cls}\left\{P^{F_0}_{cls}, ..., P^{F_T}_{cls}\right\}; G_0\left\{P^{F_0}_0, ..., P^{F_T}_0\right\}; ...; G_N\left\{P^{F_0}_1, ..., P^{F_T}_N\right\}\right] \tag{3.5}$$

Then, we shift and shuffle these patch groups (excluding class token group G_{cls} as it will always appears in the head position) using TCSS proposed by Alsehaim and Breckon (2022) to achieve more fine-grained feature extraction performance (see left part of Figure 3.4), which has been indicated by Zhang et al. (2018) and Huang et al. (2021). Briefly, the first few patch groups (in the order of position) are cut off and shifted to the end of patch groups, then, these patch groups are shuffled as shown in the middle part of Figure 3.4.

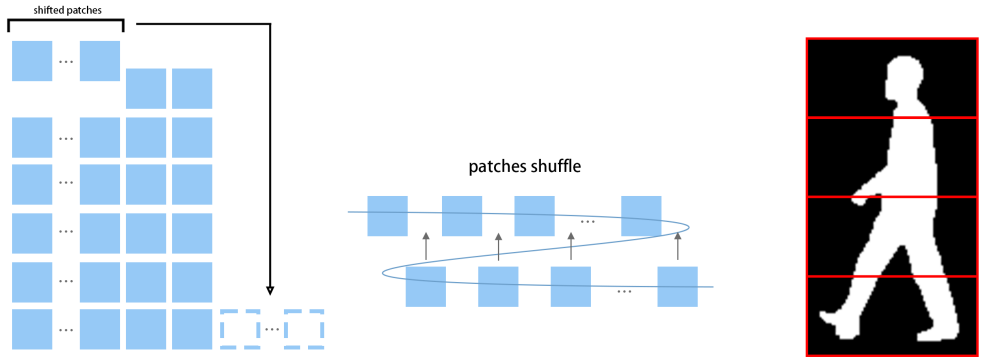

Figure 3.4: *In local part branch, as the left part of figure shows, each feature map needs to be undergone shift operation by a given amount, following by a shuffle operation shown in the middle of the figure, then each feature map will be divided into four strips from top to bottom for separate treatment.*

Subsequently, the frame bundle that has undergone shuffling is sent to part-dependent feature extraction (Fan et al. 2020). We divide image patches within each frame into multiple horizontal strips independently based on morphological characteristics (from top to bottom) as shown in the right part of Figure 3.4. In this work, the number of strips is set to 4 due to the balance between performance and computing complexity. The features of these strips are then sent to a shared Vision Transformer Block, different from the Vision Transformer in Backbone which extract at frame level, the block here sees the frame bundle in a unique way, like parts of aggregated frames stack. The shared Vision Transformer Block generate their corresponding local part features, named as $local_1$, $local_2$, $local_3$, and $local_4$.

3.3.4 Global Temporal Branch

The other branch after the backbone is Global Temporal Branch. This branch is built specifically for the spatio-temporal features extraction in the global level. It plays the similar role as Temporal Pooling (TP) module in the common framework. Initially, at the frame level, global features $Global = [global_0; ...; global_T]$ are obtained using a Vision Transformer Block which works similar as the Vision Transformer in Backbone. Then, a spatio-temporal attention which contains a two convolutional layers and a final Softmax function is applied to map the embedding dimension to 1 and generate the scores along temporal dimension, i.e., different gait frame (Rao et al. 2018), it looks like this, $Score = [score_0; ...; score_T]$. The final global-temporal feature $\hat{GLobal}$ is generated as follows:

$$\begin{aligned} Score &= attention\left(Global\right), \\ \hat{Global} &= \sum_{i=1}^{T} score_i \odot global_i \end{aligned} \tag{3.6}$$

3.3.5 BNNeck and Classification Head

After the extraction of global features and local part features from strips, since cross-entropy loss and triplet loss are simultaneously implemented, we add BNNeck proposed by Luo et al. (2020) to separate the features in embedding space. BNNeck adds a Batch Normalization layer after the generated features and before classifier Full Connection layers. They argued that many state-of-the-art methods combined ID loss and triplet loss to constrain the same feature which leads to better performance. However, the better performance let researchers ignore the inconsistency between the targets of these two losses in the embedding space. Thus, one global and four local bottlenecks, along with their linear classifiers are employed to generate ID_{global}, ID_1, ID_2, ID_3 and ID_4. These predicted ID labels are then sent to the optimizer along with the final features $\hat{Global}$, $local_1$, $local_2$, $local_3$ and $local_4$.

3.3.6 Loss

Inspired by Alsehaim and Breckon (2022), we jointly use label smoothing cross-entropy loss $\mathscr{L}_{ce}$, triplet loss $\mathscr{L}_{triple}$, attention loss $\mathscr{L}_{att}$, and center loss $\mathscr{L}_{center}$ all together.

$$\begin{aligned} \mathscr{L} =& \mathscr{L}_{ce}\left(ID_{global}\right) + \mathscr{L}_{triple}\left(\hat{Global}\right) + \beta \times \mathscr{L}_{center}\left(ID_{global}\right) + \mathscr{L}_{att} \\ &+ \frac{1}{parts} \sum_{i=1}^{parts} \left(\mathscr{L}_{ce}\left(ID_i\right) + \mathscr{L}_{triple}\left(local_i\right) + \mathscr{L}_{center}\left(ID_i\right)\right) \end{aligned} \tag{3.7}$$

Where *parts* is the number of parts we split within Local Part Spatial Branch in Figure 3.2 and $\beta = 5.0 \times 10^{-5}$. Within the loss formulation (3.7), not only the popular gait recognition losses e.g. label smoothing cross entropy loss (Szegedy et al. 2016) and triplet loss (Hermans et al. 2017) in the pipeline are used, an alternative attention loss by Pathak et al. (2020) is also added for cropping out noisy frames. We also include center loss introduced by Wen et al. (2016) with the aim of learning more robust discriminative features with the two key objectives, inter-class dispersion and intra-class compactness as much as possible.

3.4 GaitVViT

3.4.1 pipeline

For the proposed method GaitVViT, the model structure is shown in Figure 3.5. The development of this model stems from the demand to enhance the capability of temporal information extraction within a common framework. As current researches indicate, gait recognition methods emphasize temporal aspects typically by employing a complex attention mechanisms (Dou et al. 2023) or Recurrent Neural Networks (Zhang et al. 2019). In the current state-of-the-art methods, their temporal pooling modules often consist of a single layer of Max Pooling on the temporal dimension (Fan et al. 2023). From my perspective, these Temporal Pooling (TP) modules can be improved. Given the potential Vision Transformer and the variants of it, I argued that the Video Vision Transformer can be well-suited for this task. Thus, GaitVViT is introduced.

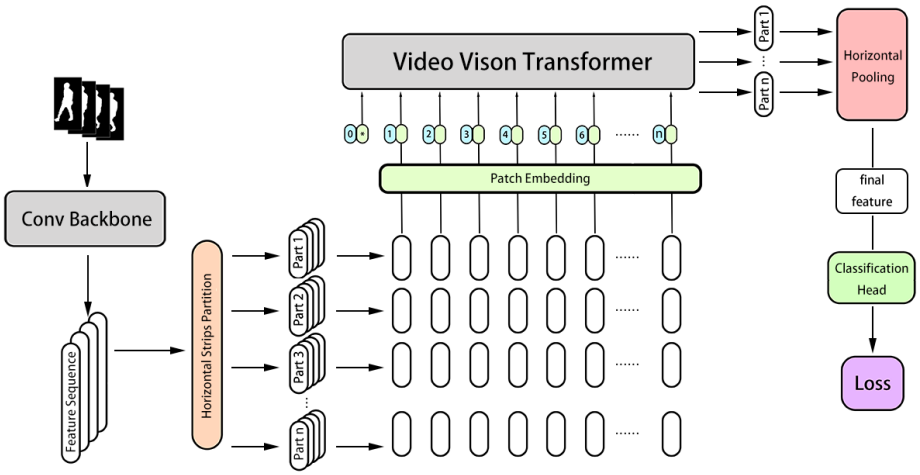

Figure 3.5: *The structure of GaitVViT. From left to right: Inputs are first fed into a CNN Backbone adopted from GaitGL; the features generated are cut into horizontal parts and fed into a shared Video Vision Transformer separately; the aggregated part features are then pooled by Horizontal Pooling module to obtain the final features; the head conduct batch normalization and predict the labels; the losses are calculated with both triplet loss and cross-entropy loss.*

GaitVViT adopts the Local Temporal Aggregation module and Global-Local Convolution module from GaitGL (Lin et al. 2020) as the backbone. So, in contrast to GaitTriViT, GaitVViT utilizes a traditional Convolutional Neural Networks as the backbone, GaitVViT takes a sequence of gait silhouette frames $Sils \in \mathbb{R}^{B\times C\times S\times H\times W}$ as inputs, where B is the batch size, C is the channel size, S is the number of frames, and $H \times W$ are the height and width of the pre-processed gait frames.

After the extraction of the CNN-based backbone, the inputs $Sils$ are mapped to a group of features $F \in \mathbb{R}^{B\times C'\times S'\times H'\times W'}$ where C' is the channel size after convolution, S' is temporal length after Local Temporal Aggregation, $H' \times W'$ is the shape of each feature map. Similar to GaitTriViT, strips partition and part-dependent ideas are adopted, GaitVViT segments the feature maps generated by the backbone into multiple horizontal parts, shown as $F = \{P_1, P_2, ..., P_n\}$, where n equals to the number of strips. For each $P_i \in \mathbb{R}^{B\times C'\times S'\times \frac{H'}{n}\times W'}$, feature map height become $\frac{H'}{n}$ by partition. These part features are then processed by a modified Video Vision Transformer,

generating the n part features $F_{VViT} = \{P_1^{VViT}, P_2^{VViT}, ..., P_n^{VViT}\}$. For each $P_i^{VViT} \in \mathbb{R}^{B\times C'\times \frac{H'}{n}\times W'}$, the temporal dimension is reduced by aggregation.

Subsequently, Horizontal Pooling pools each P_i^{VViT} to $P_i^{HP} \in \mathbb{R}^{B\times C'\times 1}$, then model concatenates these n part of P_i^{HP} together at the last dimension. The final feature is shown as $P_{final} \in \mathbb{R}^{B\times C'\times n}$. After passing through the classification head, each part will calculates its loss individually.

3.4.2 backbone

In GaitVViT, a traditional Convolutional Neural Network is implemented as the backbone. I adopted the Local Temporal Aggregation (LTA) and Global-Local Convolutional layer (GLConv) proposed by Lin et al. (2022). The overview of the backbone structure is shown in Figure 3.6. The CNN-based backbone consists of multiple convolutional layers. At first, each inputs will be

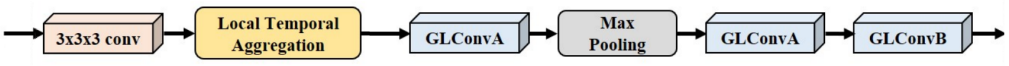

Figure 3.6: *The structure of backbone. From left to right: 3DCNN layer, Local Temporal Aggregation (LTA), Global and Local Extractor consists of GLConvA0, Max Pooling layer, GLConvA1 and GLConvB0.*

extracted by a 3DCNN layer with kernel size of $[3 \times 3 \times 3]$ to obtain shallow features. Next, the Local Temporal Aggregation (LTA) operation is employed to aggregate the temporal information and preserve more spatial information for trade off. After that, Global and Local Feature Extractor layers are implemented, which consists of GLConvA0 layer, Max Pooling layer, GLConvA1 layer and GLConvB0 layer. The Max Pooling operation is implemented to down-sample the feature size at last two dimension for computing complexity trade off. After extractor, the combined feature assembling both global and local information is generated.

The details of Global and Local Convolutional layer (GLConv) is shown in Figure 3.7. It basically consists two parallel path: one for local feature extraction and one for global feature extraction, which can take advantage of both global and local information.

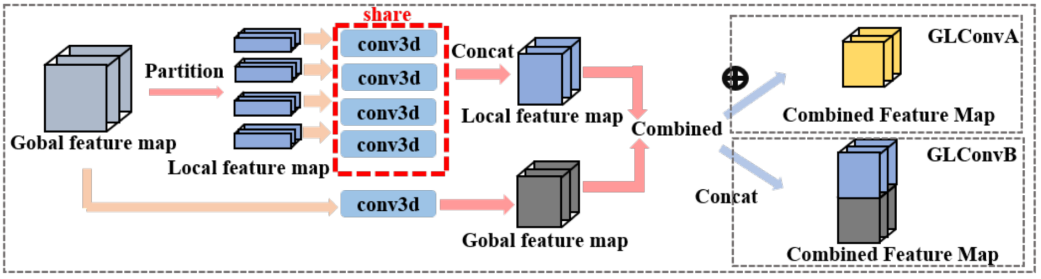

Figure 3.7: *The structure of GLConv layer. The feature map will go through two branch. The branch upper is local extraction where feature map need partition before 3D convolution, the branch below is global extraction takes whole feature map as input. And there two combination method: element-wise addition and concatenation.*

The global branch implements a basic 3DCNN layer. It extracts the whole gait information and pay attention to the relations among local regions. The local branch is basically a 3D version of

Focal Convolutional layer proposed by Fan et al. (2020). It implements a 3DCNN layer with shared kernel, the feature map will be split into several part before 3DCNN layer. They extract the local features and then combine them, which contain more detailed information than the global gait features. The GLConv has two different structures due to different combinations between global and local features, GLConvA uses element-wise addition and GLConvB uses concatenation.

3.4.3 Video Vision Transformer Encoder

After the extraction of backbone. Feature maps exist in the form of $F = \{P_1, P_2, ..., P_n\}$, where n equals to the number of strips. For each $P_i \in \mathbb{R}^{B\times C'\times S'\times \frac{H'}{n}\times W'}$, GaitVViT conducts the temporal aggregation individually.

Original Vision Transformer (ViT) (Dosovitskiy et al. 2020) regards an image as grid of non-overlapping patches, thus, the transformer extracts the features of each patch and constrains the spatial connection inter-patch. The modified Video Vision Transformer (VViT) regards each frame in a sequence as independent patch, and the multi-head self-attention among spatial patches in original ViT can be smoothly transferred to a temporal attention learning the connection among each frame. Researchers have implemented VViT-based methods in many video-based recognition tasks, e.g. ViViT by Arnab et al. (2021), Video Transformer Network by Neimark et al. (2021) and Video Swin Transformer by Liu et al. (2021).

GaitVViT adopted a modified LongFormer (Beltagy et al. 2020) as the specific Video Vision Transformer Encoder. The LongFormer in the encoder leverages sliding window to preserve the edge information between adjacent frames and strengthen the inter-frame connections. Before the Video Vision Transformer Encoder, part feature P_i will rearrange to $P_i^{pre} \in \mathbb{R}^{B\times S'\times(\frac{H'}{n}\times W'\times C')}$, then, all part features will be concatenated at the first dimension to form the $F_{pre} \in \mathbb{R}^{(B\times n)\times S'\times(\frac{H'}{n}\times W'\times C')}$. After the temporal extraction of encoder, the second dimension of F_{pre} is reduced, the aggregated feature $F_{post} \in \mathbb{R}^{(B\times n)\times(\frac{H'}{n}\times W'\times C')}$ will be rearranged back to $F_{VViT} = \{P_1^{VViT}, P_2^{VViT}, ..., P_n^{VViT}\}$, where $P_i^{VViT} \in \mathbb{R}^{B\times C'\times\frac{H'}{n}\times W'}$.

3.4.4 Classification Head and Loss

Similar to GaitTriViT, the final feature will be fed into a Batch Normalization layer followed by a full connected layer to generate the predicted labels. Both triplet loss and cross-entropy loss are employed to optimize the model. The triplet losses are calculated between feature anchors, and the cross-entropy losses are calculated on the predicted label matrix.

3.5 Summary

In this chapter, two Transformer-based architecture is proposed for Gait Recognition, GaitTriViT and GaitVViT. For GaitTriViT, the Vision Transformer is used as the frame-level backbone while incorporating case embedding and angle embedding to enhance frame-level feature extraction performance. Taking into account the similarity between Gait Recognition tasks and Person Re-Identification tasks, this work draw inspiration from several papers on ReID tasks and introduce Temporal Clip Shift and Shuffle (TCSS) by Alsehaim and Breckon (2022), as well as the combination of part-dependent strategy (Fan et al. 2020), dividing frame-level features into different strips before another Vision Transformer Block. These components above collectively build the local part spatial branch after the backbone, dedicated to extracting local spatial features. Another branch after the backbone employs another Vision Transformer Block to extract global features,where temporal attention (Rao et al. 2018; Zhang et al. 2020; Fu et al. 2019) is used to

jointly learn global temporal features. The features from both branches are combined to generate the final features, and the predicted labels are generated by classification heads. And for the optimizer, multiple loss functions are introduced to optimize the model together.

For GaitVViT, given the gait recognition common framework (Fan et al. 2023), I argue that the wildly implemented Temporal Pooling (TP) module often consists of a single Max Pooling layer. It will waste the sequence information and needs more attention for a complete improvement. The Video Vision Transformer (VViT) is the variant of original Vision Transformer (Dosovitskiy et al. 2020). VViT is created from the idea that regarding every frame in video as a patch. In the traditional transformer structure, every patch is a non-overlapping square region of an image, so when we change the scale and arrangement, the transformer is capable to conduct the extraction on a whole sequence and run the self-attention on temporal dimension. Adopting the Local Temporal Aggregation and Global and Local Convolutional layers from GaitGL proposed by Lin et al. (2022), this work connects the extracted feature representations to a Video Vision Transformer Encoder. The encoder implemented LongFormer (Beltagy et al. 2020) will conduct temporal extraction and aggregate the inputs to obtain the final features.

4 Implementation

4.1 Introduction

In this chapter, I discuss the datasets and implementation details. I chose two popular benchmarks, CASIA-B (Yu et al. 2006) and OUMVLP (Takemura et al. 2018), and introduce them briefly. Moreover, several details during training and evaluation phase are explained.

4.2 Datasets

CASIA-B is provided by Yu et al. (2006) to gait recognition and related researchers in order to promote the research. CASIA-B is a large multi-view gait database, which is created in January 2005. It has 124 subjects, and the gait data was captured from 11 views. Three variations, namely view angle, clothing and carrying condition changes, are separately considered. In this paper, we use the human silhouettes extracted from video files as benchmark. The format of the filenames in CASIA-B is 'xxx-mm-nn-ttt.png', where 'xxx' is subject id, 'mm' stands for walking status, including 'nm' (normal), 'cl' (in a coat) or 'bg' (with a bag), 'nn' is sequence number for each walking status, normal walking has six sequences, wearing coat and carrying bag have two sequences each; 'ttt' is view angle can be '000', '018', ..., '180'. Examples of CASIA-B are shown in Figure 4.1.

Each subject has a maximum of 110 sequences. We use subjects with ID from 1 to 74 as the training set, and subjects with ID from 75 to 124 as the test set. During the testing phase, we use the first four sequences from 'nm' (nm-1, nm-2, nm-3, nm-4) as the gallery set, and the remaining six sequences are divided into three query sets based on their respective situations: 'nm' query includes 'nm-5' and 'nm-6', 'bg' query includes 'bg-1' and 'bg-2', and 'cl' query includes 'cl-1' and 'cl-2' (Chao et al. 2018; Fan et al. 2020; Lin et al. 2022).

OUMVLP is part of the OU-ISIR Gait Database, stands for Multi-View Large Population Dataset, provided by Takemura et al. (2018). OUMVLP is meant to aid research efforts in the general area of developing, testing and evaluating algorithms for cross-view gait recognition. The Institute of Scientific and Industrial Research (ISIR), Osaka University (OU) has copyright in the collection of gait video and associated data and serves as a distributor of the OU-ISIR Gait Database.The data was collected in conjunction with an experience-based long-run exhibition of video-based gait analysis at a science museum. The dataset consists of 10,307 subjects (5,114 males and 5,193 females with various ages, ranging from 2 to 87 years) from 14 view angles, ranging 0°-90°, 180°-270°. Gait images of $1,280 \times 980$ pixels at 25 fps are captured by seven network cameras (Cam1-7) placed at intervals of 15-degree azimuth angles along a quarter of a circle whose center coincides with the center of the walking course. The illustration is shown in Figure 4.2.

Each subject has two sequences, 00 for probe and 01 for gallery. We select 5,153 subjects with odd-numbered IDs as the training set, and the remaining 5,154 subjects as the test set (Chao et al. 2018; Fan et al. 2020; Lin et al. 2022).

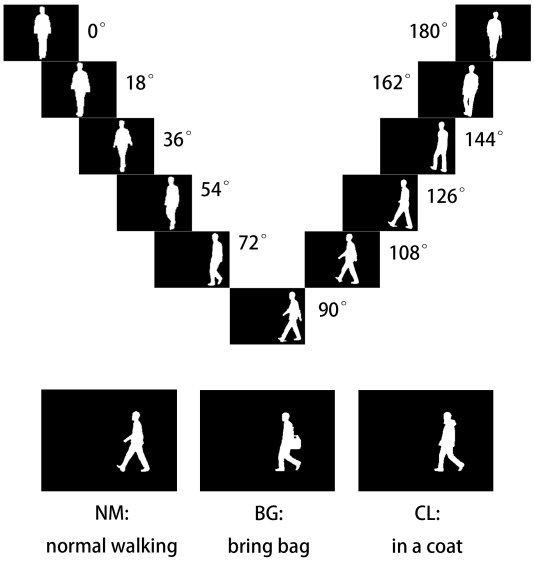

Figure 4.1: *Silhouettes from dataset CASIA-B, subjects are shot from 11 camera angles, 'BG', 'CL' and 'NM' stand for three different walking status: Bring Bag, Wearing Coat and Normal Walking.*

4.3 Implementation Details

We follow the way of Fan et al. (2022) to pre-process the silhouette data from CASIA-B and OUMVLP datasets. This pre-processing involved removing invalid frames, arranging files in a more structured index, aligning and cropping silhouette images to ensure the subject's body is in the center of the image and removing irrelevant backgrounds. After pre-processing, each frame image's size is 64×44. For GaitTriViT specifically, since we initialized the Vision Transformer backbone with parameters pre-trained on ImageNet-21K (Deng et al. 2009; Wightman 2019), the input of ViT requires RGB-like images with three channels and a size of 256×128. But our silhouettes are single-channel binary images. Therefore, we inserted a fully connected layer in the head of backbone with an input dimension of 1 and an output dimension of 3 to map the silhouette from $Sil \in \mathbb{R}^{H \times W \times 1}$ to $\hat{Sil} \in \mathbb{R}^{H \times W \times 3}$ pseudo-RGB images. And in data augmentation phase, we resized the images to the required size.

CASIA-B and OUMVLP datasets differ in camera angles and walking scenarios. CASIA-B has 11 camera angles and a total of 10 walking sequences. Therefore, in equation (3.3), E_{angle} has the shape of $[a \times D]$, where $a = 11$, and E_{case} is in shape of $[c \times D]$, where $c = 10$. D is the embedding dimension set to 768. In contrast, OUMVLP has 14 camera angles and no distinction in walking scenarios, so only $E'_{angle} \in \mathbb{R}^{a' \times D}$, where $a' = 14$.

In this work, excluding the ViT in backbone of GaitTriViT, most parameters are initialized using the Kaiming initialization (He et al. 2015). For GaitTriViT, the number of frames T in a frame bundle is set to 4. The selection strategy of frames in every bundle when training is dividing the whole sequence into T parts and randomly selecting one frame from each part, creating a frame bundle where each frame can be selected again. During testing, T frames are sequentially selected from the whole sequence. The batch size is set to 52, the optimizer is Stochastic Gradient Descent (SGD), and the scheduler is using Cosine Learning Rate Decay with Warming Up (Loshchilov and Hutter 2017). For GaitVViT, during training phase, the number of frames T in each batch is set to 30, the selection strategy is randomly choosing 30 frames in order among the sequence. During test phase, the model uses all frames within one sequence in order to generate

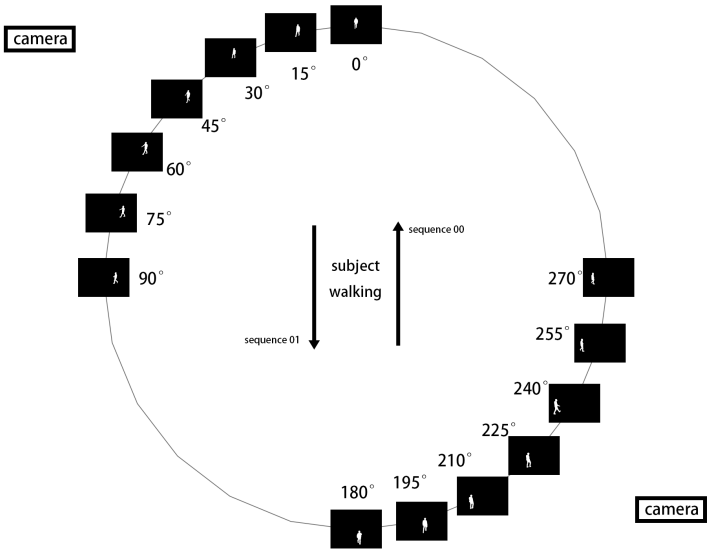

Figure 4.2: *Silhouettes from dataset OUMVLP, which don't have different walking status between sequences, but OUMVLP has the largest subjects quantity.*

the final feature. The batch size is set to 36, the optimizer is Adam, and the scheduler is Multi Step Learning Rate.

4.4 Summary

In this chapter, the choices of datasets and implementation details are explained. Two popular datasets are chosen in this work: CASIA-B and OUMVLP. CASIA-B is a classic dataset for Gait Recognition research specifically, it has 124 subjects, each subject has multiple sequences various in 11 camera angle and three walking status. OUMVLP is a new dataset compared to CASIA-B, it has the most subjects among the gait datasets so far, which consists of 5114 males and 5193 females captured from 14 camera angles, each subject has two sequences. Furthermore, several implementation details are explained including pre-processing, different setting on each benchmark and details in hyper-parameters.

5 Evaluation

5.1 State-of-the-Art Comparison

In Gait Recognition task, many researchers has made a lot of contributions. The GEINet by Shiraga et al. (2016) leverages the gait energy images (GEI) as the representations of gait features, open the researches towards gait recognition. The GaitSet by Chao et al. (2018) led the new era of appearance-based Gait Recognition, then plenty of novel models came out e.g. GaitPart by Fan et al. (2020), GaitGL by Lin et al. (2020), GaitBase by Fan et al. (2022), SRN by Hou et al. (2021), GLN by Hou et al. (2020) and DeepGait-3D by Fan et al. (2023). They all regarded as the state-of-the-art models by now. The evaluation results of two proposed methods GaitTriViT and GaitVViT are presented below, as shown in Table 5.1 and Table 5.2. The data of the state-of-the-art methods are collected from their own papers.

The 'Single' and 'Cross' marks indicate the different evaluation protocols. The 'Single' stands for the single-view-gallery evaluation which is the regular evaluation method for former state-of-the-art models, where the probe sequences under each walking conditions and the gallery sequences are divided into multiple views and the evaluation is conducted between each probe-gallery pairs with different views respectively, and the pairs with the same view angle are excluded from calculating the results. For example, the CASIA-B dataset (Yu et al. 2006) has 3 walking status and 11 camera angle, so, for probe sequences whose walking status is 'NM' and view angle is '090', they needed to compare with 10 galleries with different view angle excluding the gallery having the same view, the average of 10 results become the final result of this specific probe. The 'Cross' stands for cross-view-gallery evaluation. Particularly, for each probe view, the sequences of all gallery views and walking conditions are adopted for the comparison with the identical-view cases excluded. The accuracy under cross-view-gallery evaluation is quite higher than single-view-gallery, because subjects in some views may experience significant silhouette changes, bringing difficulty and less discriminativeness for recognition (Hou et al. 2023).

The experiments show that GaitTriViT faces huge difficulties on the two popular benchmarks. The regular single-view-gallery accuracy can only surpass the GEINet, indicating the bad generalization of GaitTriViT. Even the cross-view-gallery performances are dropped when the walking status is bring bag, or especially, wearing coat. The GaitTriViT has bad robustness towards appearance noises.

The GaitVViT performs better, on CASIA-B, when the walking status is normal walking, the performance of GaitVViT can slightly surpass the GaitGL at probe view of 0°, 18°, 36°, 54° and 126°, making the average accuracy slightly better too. But it doesn't perform well enough for a proposed transformer-enhanced method when the walking status is bag bring or wearing coat. Maybe due to the sensitiveness of transformer-based structure for appearance information, i.e., the method is less robust to appearance noises.

I compare between two proposed methods: GaitTriViT focus more on spatial feature extraction by employing three individual vision transformer in each extraction phase (one for frame-level features in backbone, one for set-level local features in local branch, and one for frame-level extraction in global branch before the attention module), while the specific temporal modeling task is assigned to a spatio-temporal attention module; in contrast, GaitVViT adopted the video vision

Table 5.1: *State-of-the-art Comparison on OUMVLP. Rank-1 Accuracy in 14 probe view angle, excluding identical-view cases. 'Single' and 'Cross' are two different evaluation protocols. GaitTriViT meets great challenge while GaitVViT reaches the level of GaitSet.*

Evaluation	Method	Probe View														Mean
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
Single	GaitSet	79.30	87.90	90.00	90.10	88.00	88.70	87.70	81.80	86.50	89.00	89.20	87.20	87.60	86.20	87.10
	GaitPart	82.60	88.90	90.80	91.00	89.70	89.90	89.50	85.20	88.10	90.00	90.10	89.00	89.10	88.20	88.70
	GLN	83.81	90.00	91.02	91.21	90.25	89.99	89.43	85.28	89.09	90.47	90.59	89.60	89.31	88.47	89.18
	GaitGL	84.90	90.20	91.10	91.50	91.10	90.80	90.30	88.50	88.60	90.30	90.40	89.60	89.50	88.80	89.70
	GaitBase	-	-	-	-	-	-	-	-	-	-	-	-	-	-	90.80
	DeepGait-3D	-	-	-	-	-	-	-	-	-	-	-	-	-	-	92.00
	GaitTriViT	58.32	72.90	80.30	82.15	76.27	75.85	73.36	59.53	73.45	79.62	81.32	75.70	75.37	72.16	74.02
	GaitVViT	81.20	88.95	90.26	90.54	89.02	89.27	88.40	85.34	87.42	88.98	89.26	87.54	87.86	86.51	87.90
Cross	GaitTriViT	84.52	97.71	98.73	98.67	98.43	99.50	99.52	88.84	98.09	98.68	98.83	98.65	99.32	99.45	97.07

transformer to replace the common Temporal Pooling (TP) module and enhance its functionality, which focus more in temporal aspect apparently. Given the situation that current transformer-based methods have not achieved astonishing outcomes in the field of gait recognition, current vision transformer structure may not be a good upstream backbone for inputs like gait silhouette. According to the argument by Fan et al. (2023), many patches on a gait silhouette are all-white (all 1) or all-black (all 0), where neither posture nor appearance information are provided. They call them dumb patches. Since all values from a dumb patch are all 0 or all 1, These all-1 or all-0 dumb patches can make backward gradients significantly ineffective or even computationally invalid for the parameters optimization of downstream ViT layers. So, as for GaitVViT, it meets the basic line of current state-of-the-art methods. A traditional CNN backbone make sure the performance away from too bad, despite the augmentation on Temporal Pooling (TP) gains no astonishing improvement.

5.2 Ablation Study

In this work, multiple technologies are employed on two methods. For GaitTriViT, there are Temporal Clips Shift and Shuffle (TCSS) and angle embedding (also case embedding on CASIA-B dataset). But it achieves not a promising performance on two popular benchmarks. To deep dive into the contribution of each technology and try to improve the performance by introducing extra mechanism, I carried several ablation experiments e.g. excluding specific module, changing selection method, rearranging the order of strips segmentation and introducing part embeddings. For GaitVViT, I also carry a ablation study by excluding certain modules or modification, to explore the individual contribution of each mechanism and potential of the model.

If not mentioned, the results below are obtained following the cross-view-gallery evaluation, as it is closer to real-world application scenarios.

5.2.1 Analysis of Excluding Specific Module

For GaitTriViT, given the utilization of multiple techniques in the proposed method and the observed insufficient model performance, understanding the individual contributions or potential hindrance of each technology becomes essential. Thus, I set pairs of tests with different situation on OUMVLP and CASIA-B, e.g. no TCSS (Alsehaim and Breckon 2022), no angle embedding or neither. The evaluation results are shown in Table 5.3 and Table 5.4.

The test results in Table 5.3 show that removing the Temporal Clips Shift and Shuffle (TCSS) module during inference could slightly improve the performance when the camera angle is not

Table 5.2: *State-of-the-art Comparison on CASIA-B. Rank-1 Accuracy in three walking status and 11 view angle, excluding identical-view cases. 'Single' and 'Cross' are two evaluation protocols. Results show GaitVViT reaches the same level of SOTA model and even surpass GaitGL in 'NM', while GaitTriViT meets a huge challenge.*

Evaluation	Status	Model	Probe View											Mean
			0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
Single	NM	GEINet	56.10	69.10	76.20	74.80	68.50	65.60	70.80	78.00	75.60	68.40	57.50	69.15
		GaitSet	90.80	97.90	99.40	96.90	93.60	91.70	95.00	97.80	98.90	96.80	85.80	95.00
		GaitPart	94.10	98.60	99.30	98.50	94.00	92.30	95.90	98.40	99.20	97.80	90.40	96.20
		GLN	93.20	99.30	99.50	98.70	96.10	95.60	97.20	98.10	99.30	98.60	90.10	96.88
		GaitGL	96.00	98.30	99.00	97.90	96.90	95.40	97.00	98.90	99.30	98.80	94.00	97.40
		GaitBase	-	-	-	-	-	-	-	-	-	-	-	97.60
		GaitTriViT	78.40	84.20	91.10	86.70	78.30	77.80	81.50	87.00	91.00	85.50	76.50	83.45
		GaitVViT	**96.50**	**99.20**	**99.40**	**98.20**	**96.90**	94.00	96.70	**99.40**	**99.30**	98.30	93.60	97.41
	BG	GEINet	44.80	53.64	54.55	51.73	49.40	46.60	47.30	56.50	58.20	49.90	45.10	50.70
		GaitSet	83.80	91.20	91.80	88.80	83.30	81.00	84.10	90.00	92.20	94.40	79.00	87.20
		GaitPart	89.10	94.80	96.70	95.10	88.30	84.90	89.00	93.50	96.10	93.80	85.80	91.50
		GLN	91.10	97.68	97.78	95.20	92.50	91.20	92.40	96.00	97.50	94.95	88.10	94.04
		GaitGL	92.60	96.60	96.80	95.50	93.50	89.30	92.20	96.50	98.20	96.90	91.50	94.50
		GaitBase	-	-	-	-	-	-	-	-	-	-	-	94.00
		GaitTriViT	71.00	74.50	78.80	76.26	67.70	65.20	68.80	77.20	79.90	77.07	66.70	73.01
		GaitVViT	90.50	95.60	95.90	93.64	89.30	82.40	88.20	94.30	96.30	94.04	90.80	91.91
	CL	GEINet	21.80	30.90	36.30	34.40	35.90	30.20	31.10	32.10	28.90	23.80	25.90	30.12
		GaitSet	61.40	75.40	80.70	77.30	72.10	70.10	71.50	73.50	73.50	68.40	50.00	70.40
		GaitPart	70.70	85.50	86.90	83.30	77.10	72.50	76.90	82.20	83.80	80.20	66.50	78.70
		GLN	70.60	82.40	85.20	82.70	79.20	76.40	76.20	78.90	77.90	78.70	64.30	77.50
		GaitGL	76.60	90.00	90.30	87.10	84.50	79.00	84.10	87.00	87.30	84.40	69.50	83.60
		GaitBase	-	-	-	-	-	-	-	-	-	-	-	77.40
		GaitTriViT	27.10	32.20	40.50	46.50	45.60	40.90	42.20	42.50	40.60	28.60	25.50	37.47
		GaitVViT	67.20	81.70	86.20	82.30	76.90	70.50	75.30	80.50	84.30	80.20	62.50	77.05
Cross	NM	GEINet	92.00	94.00	96.00	90.00	100.0	98.00	100.0	93.88	89.80	83.67	81.63	92.63
		GaitSet	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		SRN	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		GaitTriViT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.00	99.91
	BG	GEINet	80.00	94.00	90.00	87.76	94.00	94.00	92.00	94.00	90.00	84.00	82.00	89.25
		GaitSet	100.0	98.00	98.00	97.96	98.00	98.00	98.00	100.0	100.0	100.0	100.0	98.91
		SRN	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		GaitTriViT	95.00	89.00	94.00	93.94	97.00	93.00	93.00	97.00	94.00	89.90	86.00	92.89
	CL	GEINet	84.00	94.00	94.00	88.00	94.00	96.00	98.00	100.0	92.00	88.00	86.00	92.18
		GaitSet	98.00	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.82
		SRN	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		GaitTriViT	37.00	41.00	49.00	60.00	69.00	67.00	60.00	59.00	49.00	30.00	32.00	50.27

Table 5.3: *GaitTriViT's Ablation Study on OUMVLP. Rank-1 Accuracy divided in 14 probe view angle, excluding identical-view cases. Results show in dataset with only one walking status, removing TCSS can slightly improve the performance.*

Method	Probe View														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
no TCSS	**84.29**	97.65	**98.73**	**98.71**	98.43	99.48	**99.52**	**88.62**	**98.07**	**98.66**	**98.83**	**98.62**	99.30	99.45	**97.03**
no emb	41.87	79.27	89.87	86.25	88.24	98.29	98.89	49.63	77.84	89.77	87.91	93.29	98.16	98.88	84.15
baseline	82.68	97.67	98.63	98.63	98.63	99.54	99.52	87.85	97.69	98.46	98.79	98.54	99.34	99.47	96.82

Table 5.4: *GaitTriViT's Ablation Study on CASIA-B. Rank-1 Accuracy divided in 11 probe view angle and 3 walking status, excluding identical-view cases. Results show excluding single module in test influence the performance very slightly.*

Status	Method	Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	no TCSS	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.00	99.82
	no emb	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.00	99.82
	both neither	100.0	100.0	100.0	99.00	100.0	100.0	100.0	100.0	100.0	97.00	95.00	99.18
	baseline	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.00	99.91
BG	no TCSS	95.00	89.00	94.00	93.94	96.00	93.00	93.00	97.00	94.00	89.90	86.00	92.80
	no emb	95.00	89.00	**95.00**	92.93	96.00	93.00	93.00	97.00	94.00	**90.91**	85.00	92.80
	both neither	88.00	87.00	89.00	88.89	94.00	93.00	93.00	94.00	90.00	82.83	81.00	89.16
	baseline	95.00	89.00	94.00	93.94	97.00	93.00	93.00	97.00	94.00	89.90	86.00	92.89
CL	no TCSS	37.00	35.00	44.00	61.00	68.00	63.00	60.00	57.00	**50.00**	**32.00**	30.00	48.82
	no emb	37.00	35.00	44.00	**62.00**	68.00	64.00	60.00	57.00	**50.00**	31.00	30.00	48.91
	both neither	21.00	21.00	23.00	28.00	46.00	43.00	43.00	32.00	21.00	20.00	23.00	29.18
	baseline	37.00	41.00	49.00	60.00	69.00	67.00	60.00	59.00	49.00	30.00	32.00	50.27

near 90° and 270°. Maybe because the TCSS module shuffle the images, which introducing extra noises to appearance information, making the gait sequences less discriminative. But gait sequences near 90° and 270° show the side of subjects on silhouette, which usually contain more gait pattern information than appearance information in ratio, so the model will be more robust towards appearance noises and put more attention on generating distinct gait features. Also in Table 5.3, when angle embeddings are removed, more the probe view close to 0° and 180°, more significantly the test accuracy is dropped, while the accuracy drop from probe view near 90° and 270° can be almost ignored. The results also indicate the side silhouettes contain more discriminative information for a better inference performance, while the probe sequences away from side angle need angle embeddings to augment the feature representations.

For the test on CASIA-B, as shown in Table 5.4, the cross-view-gallery accuracy observe no changes when the modules are removed individually. Only when both modules are removed, significant accuracy decrease appears. Maybe because the amount of subjects in CASIA-B is much smaller than OUMVLP, so the model faces less challenges when modules are removed.

For GaitVViT, I conduct an ablation study excluding BNNeck (Luo et al. 2020) or baseline backbone. 'no BN' means the original BNNeck is replaced by Layer Normalization, 'ResNet' means the original backbone adopted from GaitGL (Lin et al. 2022) is replaced by a 4 layers ResNet backbone. Results are shown in Table 5.5. The data are obtained using single-view-

Table 5.5: *GaitVViT's Ablation Study on CASIA-B. Rank-1 Accuracy divided in 11 probe view angle and 3 walking status, excluding identical-view cases. Results show excluding BNNeck may increase the robustness towards appearance noises while the results in 'NM' drop slightly. The data in this table are obtained using single-view-gallery evaluation.*

Status	Method	Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	no BN	95.30	99.00	99.40	97.90	94.80	93.20	96.50	**99.50**	99.00	98.20	**93.80**	96.96
	ResNet	91.60	98.40	**99.70**	**98.30**	93.90	91.90	95.60	98.30	98.60	97.10	91.70	95.92
	baseline	96.50	99.20	99.40	98.20	96.90	94.00	96.70	99.40	99.30	98.30	93.60	97.41
BG	no BN	90.50	**95.90**	95.20	92.63	**90.80**	**82.60**	**89.50**	**95.30**	96.30	**95.76**	88.90	**92.13**
	ResNet	89.00	95.70	95.60	**93.94**	88.20	82.00	87.00	94.50	95.20	94.45	85.40	91.00
	baseline	90.50	95.60	95.90	93.64	89.30	82.40	88.20	94.30	96.30	94.04	90.80	91.91
CL	no BN	67.10	**85.40**	**87.40**	**84.10**	**77.50**	**73.80**	**78.10**	**81.70**	**86.10**	**83.30**	**68.40**	**79.35**
	ResNet	62.80	79.00	81.00	79.80	75.40	70.90	74.40	74.30	76.80	73.00	57.50	73.17
	baseline	67.20	81.70	86.20	82.30	76.90	70.50	75.30	80.50	84.30	80.20	62.50	77.05

gallery evaluation. The results show that the introducing of BNNeck will indeed increase the accuracy in normal walking status, but it appears slightly sensitive to appearance noises as probe sequence changing to 'BG' or 'CL'. Maybe the reason is that in training, Batch Normalization is carried when the batches are a mixture of three status, but it is not in evaluation, so the features are shifted. The results also show the contribution of original backbone in generating fine-grained global and local features.

5.2.2 Analysis of Different Selection Methods

In section State-of-the-art Comparison, the frame select strategy of GaitTriViT in test phase is picking the frames in query sequences serially, i.e. in the order of the frames are shot. It is different from the selection strategy when training, where we pick the required number of frames from the corresponding sub-sequences by cutting the whole sequence. One reason is obvious, the test selection strategy used is more likely to the real world scenarios, we obtain the silhouettes from the subject sequentially and we can process the task in real-time. The selection strategy in training is named 'intelligent' and the other is named 'dense'. The question is, if the selection strategy in evaluation was the same as training, will the model perform better or not. Several test with different frame selection methods are conducted and the results are shown in Table 5.6.

In Table 5.6, the 'intell 4' means 4 frames are selected through 'intelligent' strategy (i.e. same way when training) and run the inference independently; 'intell full' uses all probe sequence frames to conduct the inference; 'dense 28' using 28 frames for inference while not changing the original test selection strategy. The results show that in inference, more frames selected means higher evaluation accuracy. But when frames amount are small, using 'intelligent' strategy will improve the performance.

5.2.3 Analysis of Part Embeddings

Given the inspiration of position embedding in original Vision Transformer (Dosovitskiy et al. 2020) and implementation of angle embedding and case embedding in GaitTriViT, these additional learnable embeddings show their value. Prior works also indicate the effectiveness of the lightweight learnable embedding for learning invariant non-visual features (He et al. 2021; Peng

Table 5.6: *GaitTriViT's Analysis of different selection methods on CASIA-B. Rank-1 Accuracy divided in 11 probe view angle and 3 walking status, excluding identical-view cases. Results show the necessary of enough frame amount.*

Status	Method	Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	intell 4	98.34	98.28	98.45	97.70	98.39	98.92	99.27	99.34	98.72	96.90	97.00	98.30
	intell full	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.00	99.91
	dense 28	97.74	96.67	95.81	95.48	97.94	96.98	97.30	97.59	96.21	95.31	96.24	96.66
	baseline	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.00	99.91
BG	intell 4	89.43	85.41	89.45	87.03	92.02	87.55	89.14	89.84	87.14	82.85	81.39	87.39
	intell full	95.00	89.00	94.00	93.94	97.00	93.00	93.00	97.00	94.00	89.90	86.00	92.89
	dense 28	87.43	83.21	87.37	88.49	92.20	84.88	85.87	88.89	87.72	82.62	82.74	86.49
	baseline	95.00	89.00	94.00	93.94	97.00	93.00	93.00	97.00	94.00	89.90	86.00	92.89
CL	intell 4	34.13	32.48	40.16	48.26	60.36	61.73	55.39	51.38	43.15	29.99	29.92	44.27
	intell full	37.00	41.00	49.00	60.00	69.00	67.00	60.00	59.00	49.00	30.00	32.00	50.27
	dense 28	33.07	35.31	41.69	51.36	59.72	60.08	52.67	48.16	44.88	32.04	28.78	44.34
	baseline	37.00	41.00	49.00	60.00	69.00	67.00	60.00	59.00	49.00	30.00	32.00	50.27

et al. 2023). So it may also help non-context manual intervention like feature map partition and improve the model performance. The parameters are initialized with pre-trained GaitTriViT baseline checkpoint and fine-tuned with 80 epochs. The results are shown in Table 5.7.

The results show that adding the part embedding slightly increase the accuracy of GaitTriViT when subjects wearing coat. Because in proposed GaitTriViT, the feature map are divided into 4 strips, which may roughly corresponding to head and chest, waist and arms, crotch and thigh, as well as lower legs and feet. So, the part embedding will learn which part to emphasize. For gait sequences wearing a coat, the top three body parts are all self-occluded or blurred, so less discriminative representations can be extracted from the feature maps. Thus, the model will tend to put more attention on the bottom part which only has lower legs and feet. For wearing coat status, this change is beneficial, but for the normal walking probe and bag carrying probe, this change cause less attention on their information-dense top three parts. So they may facing a

Table 5.7: *GaitTriViT's Analysis of part embedding on CASIA-B. Rank-1 Accuracy divided in 11 probe view angle and 3 walking status, excluding identical-view cases. The performances are improved in 'CL' status but dropped in 'BG'.*

Status	Method	Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	part emb	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.00	99.00	99.82
	baseline	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.00	99.91
BG	part emb	90.00	86.00	89.00	88.89	96.00	88.00	89.00	96.00	93.00	**90.91**	82.00	89.89
	baseline	95.00	89.00	94.00	93.94	97.00	93.00	93.00	97.00	94.00	89.90	86.00	92.89
CL	part emb	**37.00**	**43.00**	**55.00**	**62.00**	**69.00**	65.00	**62.00**	53.00	44.00	**35.00**	**32.00**	**50.64**
	baseline	37.00	41.00	49.00	60.00	69.00	67.00	60.00	59.00	49.00	30.00	32.00	50.27

Table 5.8: *GaitTriViT's Analysis of different TCSS order on CASIA-B. Rank-1 Accuracy divided in 11 probe view angle and 3 walking status, excluding identical-view cases. The results show the order made almost no contribute to the performance.*

Status	Method	Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	TCSS after	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	**100.0**
	baseline	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.00	99.91
BG	TCSS after	87.00	82.00	88.00	90.91	95.00	85.00	82.00	92.00	85.00	84.85	80.00	86.52
	baseline	95.00	89.00	94.00	93.94	97.00	93.00	93.00	97.00	94.00	89.90	86.00	92.89
CL	TCSS after	33.00	34.00	44.00	48.00	65.00	65.00	**62.00**	52.00	43.00	**32.00**	28.00	46.00
	baseline	37.00	41.00	49.00	60.00	69.00	67.00	60.00	59.00	49.00	30.00	32.00	50.27

accuracy drop-down.

5.2.4 Analysis of Order between Shuffle and Partition

In GaitTriViT baseline, the TCSS with partition operation arise a question about the correct order of two operations. The part-dependent idea is like a manual operation to tell the model where belongs to a independent region that has different features from other region, i.e. different morphology characteristics between the limbs. So the original operation order where shift and shuffle are previous than partition will first shuffle within the whole feature map, which may make the partition meaningless. Therefore, this section try to move the TCSS module after the partition, wondering if the different order of TCSS and partition operations will make some changes. The model with different modules order is re-trained for 50 epochs. The results and comparison are shown in Table 5.8.

The results in the Table 5.8 show almost no improvement with the employment of TCSS after strategy. At every probe view of 'BG' status and almost every probe view of 'CL', the results encounter a certain percentage of decline. Results show the order change will not increase the feature discriminativeness but lose its robustness.

6 Conclusions

6.1 Conclusion

This paper proposes two novel Transformer-based Gait Recognition model, GaitTriViT and GaitVViT, to extract fine-grained features representing human walking patterns. For GaitTriViT, this work utilizes the rapidly evolving Vision Transformer instead of traditional Convolutional Neural Networks to build the model, in contrast to the Gait Recognition Pipeline, a strategy is employed that makes Temporal Pooling module and Horizontal Pooling module in parallel. By incorporating Vision Transformer and Spatio-temporal Attention mechanism, the temporal-global features are obtained in Global Temporal Branch. the model also utilizes part-dependent and shuffle strategies to extract spatial-local features in Local Spatial Branch, resulting in fine-grained features that emphasize in both global and local regions as well as temporal and spatial dimensions simultaneously without down-sampling. For GaitVViT, dissatisfied with the design of Temporal Pooling module in gait recognition common framework, the Video Vision Transformer is introduced for enhancement. The proposed Video Vision Transformer Encoder will take the output of GaitVViT backbone as sequence of patches, thus, encoder extracts the spatial feature at temporal dimension and generates the final spatio-temporal feature.

Evaluation results demonstrate that the proposed method GaitTriViT meets quite a challenge on both the popular benchmarks: CASIA-B and OUMVLP, while the other proposed method GaitVViT reach the line of state-of-the-art models based on traditional convolutional neural networks. I compare between two proposed methods and argue that current vision transformer structure may not be a good upstream backbone for binary inputs like gait silhouette. Modification and improvement are compulsory to tackle this challenge. And I still believe in the potential of Transformer-based structure in Gait Recognition as well as other video-based recognition tasks.

6.2 Limitations

In this work, two novel Transformer-based methods GaitTriViT and GaitVViT are proposed to tackle the Gait Recognition task. On two popular benchmarks: CASIA-B and OUMVLP, GaitTriViT meets huge difficulties, the results only surpass the GEINet method leveraging gait energy images (GEI) for temporal modeling. Among its own results, GaitTriViT also has a lot limitations. On the smaller CASIA-B dataset, based on different walking status, there are three scenarios: normal walking, bag carrying and wearing coat. Compared to probe status of normal walking, the evaluation in bag carrying status encounters a reasonable drop-down relatively. However, in the case of subjects wearing coats, there was a significant performance drop during evaluation, highlighting a lack of robustness in our method when silhouette appearances have significant changes. For GaitVViT, although the performance has meet the acceptable level on both popular benchmarks. It still a little far away from cutting-edge methods. It is not enough for a temporal augmented method.

6.3 Future Works

Given the lack of robustness in GaitTriViT when silhouette appearances have significant changes. If given the opportunity to work on this project again, I will focus on the robustness to minimize the noises of appearance by clothes change. For example, I may insert a module to separate the appearance and gait features. For GaitVViT, the generalization needs no worry, I could combine the tricks and technology of GaitTriViT (e.g. angel embedding) and the base structure of GaitVViT together, the fusion of two methods may helps to pursue better performance.

If I could take this further, I would take the influence of subjects' walking frequency on gait patterns into count. Moreover, I would extend beyond the constraints of popular silhouette datasets, I would obtain the wild datasets in real-world instead, integrating target detection and image segmentation modules. The focus would be enhancing the robustness of model when facing variations in subject appearances. Ultimately, I would test the model in real-world to evaluate the capabilities under diverse conditions.

Bibliography

Alsehaim, A. and Breckon, T. P. (2022), Vid-trans-reid: Enhanced video transformers for person re-identification, *in* '33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022', BMVA Press.
URL: *https://bmvc2022.mpi-inf.mpg.de/0342.pdf*

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M. and Schmid, C. (2021), 'Vivit: A video vision transformer'.
URL: *http://arxiv.org/abs/2103.15691*

Beltagy, I., Peters, M. E. and Cohan, A. (2020), 'Longformer: The long-document transformer'.
URL: *http://arxiv.org/abs/2004.05150*

Bouchrika, I., Goffredo, M., Carter, J. and Nixon, M. (2011), 'On using gait in forensic biometrics', *Journal of forensic sciences* **56**(4), 882–889.

Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y. (2016), 'Realtime multi-person 2d pose estimation using part affinity fields', *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 1302–1310.
URL: *https://api.semanticscholar.org/CorpusID:16224674*

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

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A. (2021), Emerging properties in self-supervised vision transformers, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)', pp. 9650–9660.

Chao, H., He, Y., Zhang, J. and Feng, J. (2018), 'Gaitset: Regarding gait as a set for cross-view gait recognition', *ArXiv* **abs/1811.06186**.
URL: *https://api.semanticscholar.org/CorpusID:53424263*

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L. and Zhou, Y. (2021), 'Transunet: Transformers make strong encoders for medical image segmentation'.

Cui, Y. and Kang, Y. (2022), 'Gaittransformer: Multiple-temporal-scale transformer for cross-view gait recognition', *2022 IEEE International Conference on Multimedia and Expo (ICME)* pp. 1–6.
URL: *https://api.semanticscholar.org/CorpusID:251846951*

De Boer, P.-T., Kroese, D. P., Mannor, S. and Rubinstein, R. Y. (2005), 'A tutorial on the cross-entropy method', *Annals of operations research* **134**, 19–67.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009), Imagenet: A large-scale hierarchical image database, *in* 'CVPR09'.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2020), 'An image is worth 16x16 words: Transformers for image recognition at scale', *ArXiv* **abs/2010.11929**.
URL: *https://api.semanticscholar.org/CorpusID:225039882*

Dou, H., Zhang, P., Su, W., Yu, Y. and Li, X. (2023), 'Metagait: Learning to learn an omni sample adaptive representation for gait recognition', *ArXiv* **abs/2306.03445**.
URL: *https://api.semanticscholar.org/CorpusID:253448299*

Fan, C., Hou, S., Huang, Y. and Yu, S. (2023), 'Exploring deep models for practical gait recognition'.
URL: *http://arxiv.org/abs/2303.03301*

Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y. and Yu, S. (2022), 'Opengait: Revisiting gait recognition toward better practicality', *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 9707–9716.
URL: *https://api.semanticscholar.org/CorpusID:253510828*

Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q. and He, Z. (2020), 'Gaitpart: Temporal part-based model for gait recognition', *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 14213–14221.
URL: *https://api.semanticscholar.org/CorpusID:219634265*

Fu, Y., Wang, X., Wei, Y. and Huang, T. (2019), 'Sta: Spatial-temporal attention for large-scale video-based person re-identification', **33**(01), 8287–8294.

Han, J. and Bhanu, B. (2005), 'Individual recognition using gait energy image', *IEEE transactions on pattern analysis and machine intelligence* **28**(2), 316–322.

He, K., Zhang, X., Ren, S. and Sun, J. (2015), Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *in* 'Proceedings of the IEEE International Conference on Computer Vision (ICCV)'.

He, S., Luo, H., Wang, P., Wang, F., Li, H. and Jiang, W. (2021), Transreid: Transformer-based object re-identification, *in* 'Proceedings of the IEEE/CVF international conference on computer vision', pp. 15013–15022.

Hermans, A., Beyer, L. and Leibe, B. (2017), 'In defense of the triplet loss for person re-identification', *ArXiv* **abs/1703.07737**.
URL: *https://api.semanticscholar.org/CorpusID:1396647*

Hoffer, E. and Ailon, N. (2015), Deep metric learning using triplet network, *in* 'Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3', Springer, pp. 84–92.

Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A. and Chanussot, J. (2022), 'Spectralformer: Rethinking hyperspectral image classification with transformers', *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–15.

Hou, S., Cao, C., Liu, X. and Huang, Y. (2020), Gait lateral network: Learning discriminative and compact representations for gait recognition, *in* 'European conference on computer vision', Springer, pp. 382–398.

Hou, S., Fan, C., Cao, C., Liu, X. and Huang, Y. (2022), 'A comprehensive study on the evaluation of silhouette-based gait recognition', *IEEE Transactions on Biometrics, Behavior, and Identity Science* pp. 1–1.
URL: *https://ieeexplore.ieee.org/document/9928336/*

Hou, S., Fan, C., Cao, C., Liu, X. and Huang, Y. (2023), 'A comprehensive study on the evaluation of silhouette-based gait recognition', *IEEE Transactions on Biometrics, Behavior, and Identity Science* **5**(2), 196–208.

Hou, S., Liu, X., Cao, C. and Huang, Y. (2021), 'Set residual network for silhouette-based gait recognition', *IEEE Transactions on Biometrics, Behavior, and Identity Science* **3**(3), 384–393.

Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G. and Fu, B. (2021), 'Shuffle transformer: Rethinking spatial shuffle for vision transformer'.

Iwama, H., Muramatsu, D., Makihara, Y. and Yagi, Y. (2013), 'Gait verification system for criminal investigation', *Information and Media Technologies* **8**(4), 1187–1199.

Jianhua, M., Lijun, Y. and Xiaopeng, Z. (n.d.), 'Infrared human gait recognition method based on long and short term memory network'.

Larsen, P. K., Simonsen, E. B. and Lynnerup, N. (2008), 'Gait analysis in forensic medicine', *Journal of forensic sciences* **53**(5), 1149–1153.

Liao, R., Yu, S., An, W. and Huang, Y. (2020), 'A model-based gait recognition method with body pose and human prior knowledge', *Pattern Recognit.* **98**.
URL: *https://api.semanticscholar.org/CorpusID:207757561*

Lin, B., Zhang, S., Wang, M., Li, L. and Yu, X. (2022), 'Gaitgl: Learning discriminative global-local feature representations for gait recognition'.
URL: *http://arxiv.org/abs/2208.01380*

Lin, B., Zhang, S. and Yu, X. (2020), 'Gait recognition via effective global-local feature representation and local temporal aggregation', *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* pp. 14628–14636.
URL: *https://api.semanticscholar.org/CorpusID:237108355*

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S. and Hu, H. (2021), 'Video swin transformer', *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 3192–3201.
URL: *https://api.semanticscholar.org/CorpusID:235624247*

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. and Black, M. J. (2015), 'Smpl: A skinned multi-person linear model', *ACM Trans. Graph.* **34**(6).
URL: *https://doi.org/10.1145/2816795.2818013*

Loshchilov, I. and Hutter, F. (2017), 'Sgdr: Stochastic gradient descent with warm restarts'.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S. and Gu, J. (2020), 'A strong baseline and batch normalization neck for deep person re-identification', *IEEE Transactions on Multimedia* **22**(10), 2597–2609.

Makihara, Y., Nixon, M. S. and Yagi, Y. (2020), 'Gait recognition: Databases, representations, and applications', *Computer Vision: A Reference Guide* pp. 1–13.

Martinez, J., Hossain, R., Romero, J. and Little, J. (2017), 'A simple yet effective baseline for 3d human pose estimation', *2017 IEEE International Conference on Computer Vision (ICCV)* pp. 2659–2668.
URL: *https://api.semanticscholar.org/CorpusID:206771080*

Mogan, J. N., Lee, C. P., Lim, K. M. and Anbananthen, K. S. M. (2022), 'Gait-vit: Gait recognition with vision transformer', *Sensors (Basel, Switzerland)* **22**.
URL: *https://api.semanticscholar.org/CorpusID:252633940*

Neimark, D., Bar, O., Zohar, M. and Asselmann, D. (2021), 'Video transformer network'. **URL:** *http://arxiv.org/abs/2102.00719*

Nixon, M. S. and Carter, J. N. (2006), 'Automatic recognition by gait', *Proceedings of the IEEE* **94**(11), 2013–2024.

Pathak, P., Eshratifar, A. E. and Gormish, M. (2020), 'Video person re-id: Fantastic techniques and where to find them (student abstract)', **34**(10), 13893–13894.

Peng, L., Chen, Z., Fu, Z., Liang, P. and Cheng, E. (2023), Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs, *in* 'Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)', pp. 5935–5943.

Pinić, D., Suanj, D. and Lenac, K. (2022), 'Gait recognition with self-supervised learning of gait features based on vision transformers', *Sensors (Basel, Switzerland)* **22**. **URL:** *https://api.semanticscholar.org/CorpusID:252502855*

Rao, S., Rahman, T., Rochan, M. and Wang, Y. (2018), 'Video-based person re-identification using spatial-temporal attention networks', *arXiv preprint arXiv:1810.11261* .

Roy, A., Sural, S. and Mukherjee, J. (2012), 'Gait recognition using pose kinematics and pose energy image', *Signal Processing* **92**(3), 780–792.

Rubinstein, R. Y. and Kroese, D. P. (2004), *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, Vol. 133, Springer.

Santos, C. F. G. D., Oliveira, D. D. S., Passos, L. A., Pires, R. G., Santos, D. F. S., Valem, L. P., Moreira, T. P., Santana, M. C. S., Roder, M., Papa, J. P. and Colombo, D. (2023), 'Gait recognition based on deep learning: A survey'.

Seely, R. D., Samangooei, S., Lee, M., Carter, J. N. and Nixon, M. S. (2008), The university of southampton multi-biometric tunnel and introducing a novel 3d gait dataset, *in* '2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems', IEEE, pp. 1–6.

Sepas-Moghaddam, A. and Etemad, A. (2022), 'Deep gait recognition: A survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T. and Yagi, Y. (2016), Geinet: View-invariant gait recognition using a convolutional neural network, *in* '2016 international conference on biometrics (ICB)', IEEE, pp. 1–8.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016), Rethinking the inception architecture for computer vision, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2818–2826.

Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T. and Yagi, Y. (2018), 'Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition', *IPSJ transactions on Computer Vision and Applications* **10**, 1–14.

Tong, Z., Song, Y., Wang, J. and Wang, L. (2022), 'Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training'.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015), Learning spatiotemporal features with 3d convolutional networks, *in* 'Proceedings of the IEEE International Conference on Computer Vision (ICCV)'.

Tran, L., Hoang, T., Nguyen, T., Kim, H. and Choi, D. (2021), 'Multi-model long short-term memory network for gait recognition using window-based data segment', *IEEE Access* **9**, 23826–23839. lstm but input is IMU.

Wan, C., Wang, L. and Phoha, V. V. (2018), 'A survey on gait recognition'.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P. and Shao, L. (2022), 'Pvt v2: Improved baselines with pyramid vision transformer', *Computational Visual Media* .

Wen, Y., Zhang, K., Li, Z. and Qiao, Y. (2016), A discriminative feature learning approach for deep face recognition, *in* 'Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14', Springer, pp. 499–515.

Wightman, R. (2019), 'Pytorch image models', `https://github.com/rwightman/pytorch-image-models`.

Wolf, T., Babaee, M. and Rigoll, G. (2016), 'Multi-view gait recognition using 3d convolutional neural networks', *2016 IEEE International Conference on Image Processing (ICIP)* pp. 4165–4169. **URL:** *https://api.semanticscholar.org/CorpusID:2489166*

Wu, Z., Huang, Y., Wang, L., Wang, X. and Tan, T. (2017), 'A comprehensive study on cross-view gait based human identification with deep cnns', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(2), 209–226.

Yang, Y., Yun, L., Li, R., Cheng, F. and Wang, K. (2023), 'Multi-view gait recognition based on a siamese vision transformer', *Applied Sciences* **13**(4), 2273.

Yu, S., Tan, D. and Tan, T. (2006), A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, *in* '18th International Conference on Pattern Recognition (ICPR'06)', Vol. 4, pp. 441–444.

Zhang, W., He, X., Yu, X., Lu, W., Zha, Z. and Tian, Q. (2020), 'A multi-scale spatial-temporal attention model for person re-identification in videos', *Trans. Img. Proc.* **29**, 3365–3373. **URL:** *https://doi.org/10.1109/TIP.2019.2959653*

Zhang, X., Zhou, X., Lin, M. and Sun, J. (2018), Shufflenet: An extremely efficient convolutional neural network for mobile devices, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'.

Zhang, Z., Tran, L., Yin, X., Atoum, Y., Liu, X., Wan, J. and Wang, N. (2019), Gait recognition via disentangled representation learning, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 4710–4719.

Zheng, J., Liu, X., Liu, W., He, L., Yan, C. and Mei, T. (2022), Gait recognition in the wild with dense 3d representations and a benchmark, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 20228–20237.

Zheng, L., Yang, Y. and Hauptmann, A. G. (2016), 'Person re-identification: Past, present and future'.

Zhu, D., Huang, X., Wang, X., Yang, B., He, B., Liu, W. and Feng, B. (2023), 'Multi-scale context-aware network with transformer for gait recognition'.