Multi-modal Gait Recognition via Effective Spatial-Temporal Feature Fusion

Yufeng Cui Yimei Kang†
College of Software, Beihang University, Beijing, China
{cuiyufeng,kangyimei}@buaa.edu.cn

Abstract

Gait recognition is a biometric technology that identifies people by their walking patterns. The silhouettes-based method and the skeletons-based method are the two most popular approaches. However, the silhouette data are easily affected by clothing occlusion, and the skeleton data lack body shape information. To obtain a more robust and comprehensive gait representation for recognition, we propose a transformer-based gait recognition framework called MMGaitFormer, which effectively fuses and aggregates the spatial-temporal information from the skeletons and silhouettes. Specifically, a Spatial Fusion Module (SFM) and a Temporal Fusion Module (TFM) are proposed for effective spatial-level and temporal-level feature fusion, respectively. The SFM performs fine-grained body parts spatial fusion and guides the alignment of each part of the silhouette and each joint of the skeleton through the attention mechanism. The TFM performs temporal modeling through Cycle Position Embedding (CPE) and fuses temporal information of two modalities. Experiments demonstrate that our MMGaitFormer achieves state-of-the-art performance on popular gait datasets. For the most challenging "CL" (i.e., walking in different clothes) condition in CASIA-B, our method achieves a rank-1 accuracy of 94.8%, which outperforms the state-of-the-art single-modal methods by a large margin.

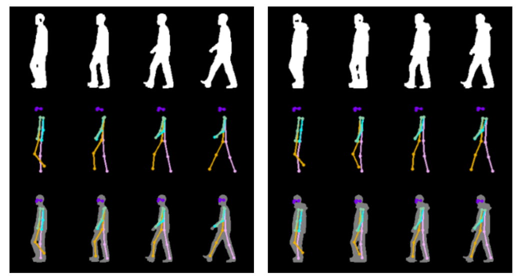

Figure 1. Comparison of different gait representations of a subject from the CASIA-B gait dataset at different timesteps of normal walks (a) and walking in different clothes (b). Each row depicts the same frames as silhouette image, and 2D skeleton pose, the combination of skeletons and silhouettes, respectively, from top-to-bottom. Combines the complementary strengths of silhouette and skeleton, it is expected to be a more comprehensive representation for gait.

1. Introduction

Gait recognition is a biometric technology that identifies people by their walking patterns, which is one of the most promising video-based biometric technologies in the long-distance recognition system. However, it is still challenging to perform reliable gait recognition, as its performance is severely affected by many complex factors, including clothing, carrying conditions, cross-view, *etc.*. To alleviate these issues, various methods have been proposed. The appearance-based and model-based methods are the two most popular approaches for video-based gait recognition. The appearance-based (*i.e.*, silhouettes-based) methods [5, 9, 14, 19, 27] rely on binary human silhouette images segmented from the original video frame to eliminate the influence of external factors. They utilized convolutional neural networks (CNN) to extract spatio-temporal features and achieved state-of-the-art performance. The model-based methods [2,16,17,23] consider the underlying physical structure of the body and express the gait in a more comprehensible model. The most recent model-based approaches are skeletons-based, in which they represent gait with the skeletons obtained from videos through pose estimation models. With clear and robust skeleton representation, recent skeletons-based methods could even show competitive results compared to appearance-based methods.

Although both silhouette-based and skeletons-based methods have their advantages, we argue that the incompleteness of both input representations of the gait information limits further improvement of these methods. As shown in Fig.1(a), although the silhouettes retain most body shape information, the self-obscuring problem occurs when body areas overlap. Moreover, when clothing condition changes, as shown in Fig.1(b), the external body shape is significantly changed by clothing obscuration. However,

†Corresponding Author.

skeletons only keep the internal body structure information which effectively solves the clothing-obscuring and self-obscuring problems, but completely ignoring the discriminative body shape information leads to poor performance. Thus, we could observe that the silhouette retains the external body shape information and omits some body-structure clues, and the skeleton preserves the internal body structure information. The two data modalities are complementary to each other, and their combination is expected to be a more comprehensive representation of gait.

Motivated by the observations above, to obtain robust and comprehensive gait representation for recognition, we propose a transformer-based gait recognition framework called MMGaitFormer, which effectively fuses and aggregates the spatial-temporal information from the skeletons and silhouettes. Precisely, the proposed framework consists of four main modules at three stages. Firstly, the silhouette sequence and skeleton sequence are extracted from the original RGB video by segmentation and pose estimation methods, respectively. After that, we feed the silhouettes and skeletons into independent encoding modules to extract unique spatio-temporal feature maps for each modal. Finally, we propose a Spatial Fusion Module (SFM) and a Temporal Fusion Module (TFM) for spatial and temporal feature fusion, respectively. As a video-based recognition task, how to effectively extract discriminative gait features from spatio-temporal information is the most critical issue. In this work, we consider both fine-grained fusion at the spatial level and fine-aligned fusion at the temporal level. In the SFM, we design a co-attention module to enable the interactions between the silhouettes and skeletons. Specifically, we construct strategies called Fine-grained Body Parts Fusion (FBPF) to guide SFM for fine-grained feature fusion learning based on prior positional relationships between joints in the skeleton and corresponding parts in the silhouette. In the TFM, we introduced an embedding modeling operation for fine-aligned temporal modeling, in which we design the Cycle Position Embedding (CPE) to efficiently capture gait cycle features and better model the temporal information for gait sequences.

The main contributions of the proposed method are summarized as follows: (1) We propose an effective and novel multi-modal gait recognition framework called MMGaitFormer, which utilizes a more comprehensive gait representation constructed from silhouettes and skeletons for better recognition. (2) A co-attention-based Spatial Fusion Module is proposed to perform a fine-grained body parts fusion (FBPF) of spatial gait features by using the prior positional relationships of each skeleton joint and each silhouette part. (3) We propose a novel Temporal Fusion Module for feature fusion at the temporal level, in which we design the Cycle Position Embedding (CPE) to model temporal relationships for gait sequences of arbitrary length. Experiments demonstrate that our MMGaitFormer achieves state-of-the-art performance on popular gait datasets. For the most challenging condition (*i.e.*, walking in different clothes) in CASIA-B [26], our method achieves a rank-1 accuracy of 94.8%, which outperforms the state-of-the-art Single-modal methods by a large margin (+11.2% accuracy improvement).

2. Related work

Appearance-based Methods rely on binary human silhouette images extracted from the original images. Most recent methods directly consider gait as a sequence of silhouettes. These methods [5, 6, 9, 18, 20] follow a similar pipeline, which extract spatial features using a well-designed network at the frame level and then use a spatio-temporal aggregate module to obtain the gait representation. For instance, GaitPart [9] designed a Micro-motion Capture Module (MCM) module to model the local micro-motion features. GaitGL [20] proposed a 3D CNN network to simultaneously aggregate local spatio-temporal information. GaitTransformer [6] proposed Multiple-Temporal-Scale Transformer (MTST) for gait temporal modeling. Although the silhouette-based approach achieved state-of-the-art performance, the silhouette data will inevitably meet the problem of clothing obscuring and self obscuring, limiting its further improvement.

Model-based Methods consider the underlying physical structure of the body and express the gait in a more comprehensible model [2, 16, 17]. The most recently model-based methods commonly take skeletons as raw input data extracted from the original videos with pose estimation models. PoseGait [17] utilizes human prior knowledge to design pose features and uses CNN to extract feature representations for recognition. GaitGraph [23] extracted the gait information from human 2D joints based on Graph Convolutional Network (GCN) and achieve competitive results. Although the skeleton-based methods are robust against view and appearance changes, the skeleton data contains less body shape information than the silhouette images.

Multi-modal Gait Recognition [4, 7, 13] approaches that integrate depth, multi-sensor and video data have shown improvements in recognition performance in early research. However, homogeneous multi-modal methods that solely rely on video data have not been fully explored, and existing methods [15, 21, 25] still suffer critical issues: (1) Simply concatenating the final global features of the two modalities could not effectively capture fine-grained spatial information. (2) The temporal information of the two modal sequences is not fully utilized, and how to effectively fuse their temporal features remains an open problem. Inspired by the remarkable success of Transformer [24] in multi-modal learning, we propose a transformer-based approach that leverages two complementary data modalities, *i.e.*, silhouette and skeleton, for comprehensive gait recognition.

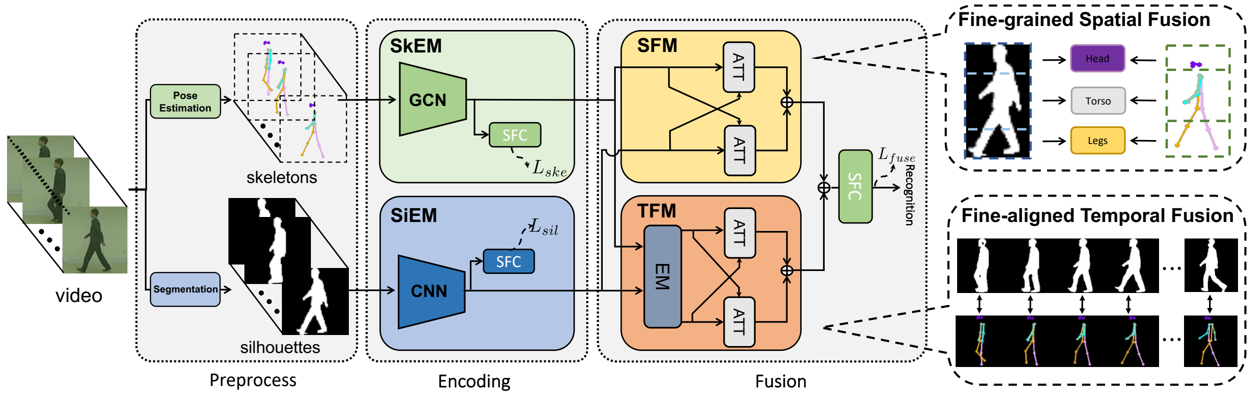

Figure 2. The pipeline of our MMGaitFormer. In the preprocessing stage, the silhouette sequence and skeleton sequence are extracted from the original RGB video by segmentation method and pose estimation method, respectively. In the Encoding stage, we feed the input silhouettes and skeletons into Silhouette Encoding Module (SiEM) and Skeleton Encoding Module (SkEM) to learn spatial-temporal feature maps, respectively. In the Fusion stage, a Spatial Fusion Module (SFM) and a Temporal Fusion Module (TFM) are proposed for effective fine-grained spatial and fine-aligned temporal feature fusion, respectively. ATT means cross-attention block, and two ATTs form a co-attention structure for feature fusion. Embedding Modeling (EM) in TFM is used for temporal modeling. Separate Fully Connected Layer (SFC) is used for the feature mapping in the Encoding and Fusion stage.

3. Method

In this section, we will describe the technical details of our MMGaitFormer. In Sec.3.1, we present an overview of our method. In Sec.3.2, we discuss the design motivation of SiEM and SkEM. In Sec.3.3, we introduce our proposed Spatial Fusion Module on how to integrate skeleton information and silhouette information by Fine-grained Body Parts Fusion (FBPF). In Sec.3.4, we elaborate on our proposed Temporal Fusion Module on how to use the Cycle Position Embedding (CPE) to model and fuse the temporal information of two modalities sequence.

3.1. Pipeline

To efficiently obtain, process, and fuse the gait representation of both modalities, we propose an effective and novelty framework called MMGaitFormer which effectively fuses the complementary spatio-temporal information of both modalities while preserving the unique discriminative features of each modality. The pipeline of the proposed multi-modal gait recognition framework is shown in Fig.2.

In the preprocess stage, two types of gait representations will be obtained offline from the original gait video. One is the silhouette sequence $S \in \mathbb{R}^{C_1^s \times T_1^s \times H_1 \times W_1}$ extracted by segmentation method, where C_1 is the number of channels, T_1^s is the length of the silhouette sequence and (H_1, W_1) is the image size of each frame. Another input is the skeleton sequence which is extracted by a pose estimation model [10, 22]. The skeleton sequence can be described by $A \in \mathbb{R}^{N_1 \times N_1}$ structurally and by $K \in \mathbb{R}^{C_1^k \times T_1^k \times N_1}$ feature-wise, where C_1^k is the number of channels, T_1^k is the length of the sequence and N_1 is the number of joints.

In the encoding stage, given the sequence of silhouette S and skeleton K, the feature maps $F_s \in \mathbb{R}^{C_2^s \times T_2^s \times H_2}$, and $F_k \in \mathbb{R}^{C_2^k \times T_2^k \times N_2}$ are then extracted from the Silhouette Encoding module (SiEM) and Skeleton Encoding module (SkEM), respectively, in order to learn the unique spatio-temporal information of each gait representation.

In the fusion stage, these feature maps are then fed into two branches: (1) The Spatial Fusion Module fuses each silhouette part and each skeleton node at a fine granularity using the co-attention structure and obtains the spatial feature representation Y_s. (2) The Temporal Fusion Module models the temporal relation by Embedding Modeling and fuses long-term feature information for each modal for temporal feature representation Y_t. We concatenate the Y_s and Y_t as the final feature representation Y for the gait sequence.

Finally, we choose a combined loss to train the proposed network, consisting the fusion loss L_{fuse}, the silhouette loss L_{sil} and the skeleton loss L_{ske}. The total loss is defined as $L = L_{fuse} + L_{sil} + L_{ske}$. We utilize the separate Batch All triplet loss [12] as the loss function.

3.2. Silhouette and Skeleton Encoding Module

Motivation. The data structures of the two modal representations are too different, so it is difficult to fuse them directly on the data-level. Therefore, we design independent encoding modules to capture the unique discriminative information of each modal and enhance the spatial-temporal feature representation for the subsequent fusion. To speed up the model convergence, we specially perform silhouette loss L_{sil} and skeleton loss L_{ske} to supervise the learning of each modal feature separately.

Operation. Inspired by GaitGL [19] and GaitGraph [23], we design our SiEM network and SkEM network. The SiEM network is composed of 3D CNN blocks [20], Max Pooling Layers and Micro-motion Capture Module (MCM) [9]. For the SkEM, we introduce the graph convolutional network (GCN) to extract spatio-temporal gait features from the sequence of skeleton graphs. The output channel of the last block is set to 128, which is the same as the output of the SiEM to facilitate subsequent fusion processing. The SiEM and SkEM in our framework can also be replaced by any gait recognition networks. The more complex architecture of the SiEM and SkEM may bring in more considerable performance gains, but that is not the priority of the proposed method. Therefore, SiEM and SkEM can be considered the baseline of our approach.

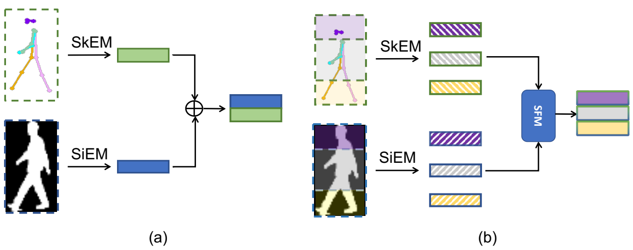

Figure 3. Comparison of different spatial fusion strategies. (a) illustrates global feature-level fusion, (b) illustrates our proposed co-attention based fine-grained feature fusion.

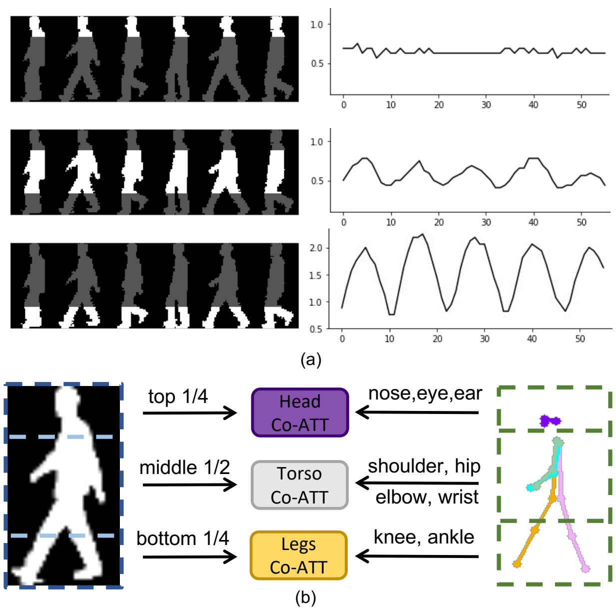

Figure 4. (a) The human body area can be divided into three parts: head, torso and legs, and different area of human gait possess evidently different shapes and moving patterns during walking. (shown by the images of the aspect ratio) (b) Fine-grained Body Parts Fusion (FBPF): The computation of co-attention is restricted between the corresponding regions of head, torso and legs.

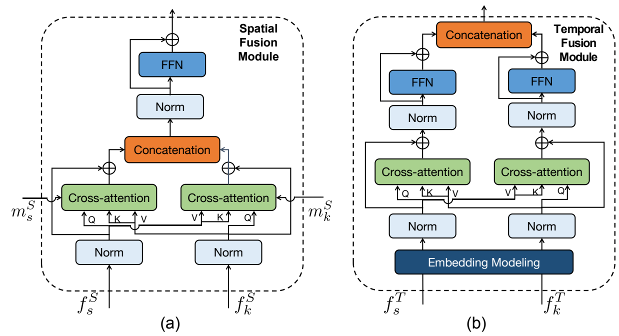

Figure 5. The network structure of our proposed Spatial Fusion Module (a) and Temporal Fusion Module (b), both of which contain a co-attention structure. Each co-attention structure consists of two interconnected cross-attention blocks. The input of SFM is the spatial feature embedding of silhouette f_s^S and skeleton f_k^S, and the pre-defined attention mask of cross-attention m_s^S and m_k^S for fine-grained body parts fusion to restrict skeleton and silhouette to corresponding regions for restricted attention computation. TFM's input is the temporal feature embedding of silhouette f_s^T and skeleton f_k^S.

3.3. Spatial Fusion Module

Motivation. Concurrently with this work, individual approaches [21, 25] are also beginning to explore more robust features through the fusion of multiple gait modalities. However, As shown in Fig.3 (a), these methods have a relatively simple means of fusion and focus on the fusion at the global feature level by a concatenation operation [25]. Such a fusion operation lacks interpretability and flexibility and also lacks the use of prior spatial information about the human body. Moreover, these methods usually rely on pre-trained models for each modal, which makes them more like ensemble models than multi-modal models. To address these issues, we propose a co-attention based fusion module shown in Fig.5 (a) which adopts the interpretive fusion of each body part's external shape (silhouette) and internal structure (skeleton) by the attention mechanism (*i.e.*, Fine-grained Body Parts Fusion), which is shown in Fig.3 (b). The attention-based learning structure also makes the method more flexible, allowing end-to-end training without relying on pre-trained models for each modal.

Fine-grained Body Parts Fusion. As shown in Fig.4 (a), the human body area can be divided into three parts: head, torso, and legs, and different body parts of human gait possess different shapes and moving patterns during walking. Motivated by the above observations, we argue that spatial feature fusion should be fine-grained and propose a simple but effective strategy to achieve a more comprehensive fine-grained spatial feature fusion by using human prior knowledge. We restrict the silhouette and skeleton features to compute cross-attention only with the corresponding body parts by constructing attention masks m_s^S and m_k^S, as shown in Fig.4(b). On the one hand, the fusion between each body part effectively utilizes the prior knowledge of the human body and is therefore more interpretable. On the order hand, the restricted attention computation can reduce the computational complexity by half and effectively reduce the risk of overfitting.

In this work, we established a simple mapping relationship between silhouette and skeleton to construct predefined attention masks shown in Fig.4 (b). The top quarter (0-1/4), middle half (1/4-3/4), and the bottom quarter (3/4-1) of the feature embedding f_s^S represent silhouette features of the head, torso, and legs respectively. Similarly, the skeleton node vector is also divided into the same three areas of the head (The node features of *nose, eye, ear* in f_k^S), torso (*shoulder, elbow, wrist, hip*), and legs (*knee, ankle*).m_s^S and m_k^S are transposes of each other.

Spatial Co-attention Aggregation. The co-attention fusion module enables the interactions between the silhouettes and skeletons, which establishes various spatial relationships between silhouette parts and skeleton joints to exploit complementary strengths of the two data modalities for a more robust and comprehensive gait feature representation for recognition. Compared to individual cross-attention modules, the co-attention structure can better integrate the complementary advantages of the skeleton and silhouette. And by constructing Attention mask for restricted attention computation, the risk of overfitting of Transforbased methods is reduced while improving interpretability.

Operation. As visualized in Fig.5 (a), the co-attention module includes interlaced multi-head cross-attention blocks. In this work, our cross-attention blocks follow the ViT's [8] multi-head attention structure. For the feature maps $F_s \in \mathbb{R}^{C_2^s \times T_2^s \times H_2}$ and $F_k \in \mathbb{R}^{C_2^k \times T_2^k \times N_2}$, max-pooling are used in the temporal axis to get the spatial feature embedding $f_s^S \in \mathbb{R}^{C_3 \times H_2}$ and $f_k^S \in \mathbb{R}^{C_3 \times N_2}$, respectively. These feature embeddings are then fed into co-attention structure for complementary information fusion, and subsequently followed by feed-forward network (FFN) layer to generate the spatial feature representation Y_s.

3.4. Temporal Fusion Module

Motivation. As a video-based recognition task, the temporal relationships between gait frames contain unique biological information which is critical for recognition. To better exploit the temporal information of the gait sequences of both modals, we propose an attention-based Temporal Fusion Module (TFM) to aggregate the temporal features of both modals. Moreover, as shown in Fig.4(a), gait is a cyclical and symmetric process. Therefore, we proposed the Cycle Position Embedding to better model and align the temporal information for the sequences of two modals.

Cycle Position Embedding. The attention mechanism cannot distinguish the position information of the input fea-

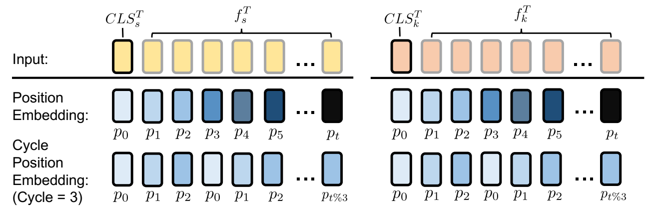

Figure 6. The Comparison of Embedding Modeling by Position Embedding and Embedding Modeling by our proposed Cycle Position Embedding (cycle = 3)

ture sequence. As shown in Fig.6, for existing vision transformer methods [8], Position Embedding of the same length as the input sequence is used to indicate the order of the input tokens. However, this approach limits the transformer only to extract spatial-temporal information from fixed-length gait sequences. To address this shortcoming, we proposed Cycle Position Embedding (CPE), expressed as $P_s = \{p_i | i = 1, ..., s\}$, the s is the cycle size of position embedding. We repeat the position embedding until it has the same length as the feature embeddings to process sequences of any size. On the one hand, the process of repeating for position embeddings simulates the gait cycle process which is a more efficient way to model the gait cycle in sequence. And the size of the cycle s can be set interpretably according to the number of frames in a gait cycle. On the other hand, the risk of overfitting is further reduced by limiting the number of learnable parameters, helping the proposed Transformer-based model to converge better.

Moreover, the same frames in both sequences are performed with the same position embedding for fine-aligned temporal modeling. We prepend a sequence of feature embeddings for each modality with a learnable class embedding (expressed as CLS_s^T and CLS_k^T), whose state at the output of the attention block serves as the temporal feature representation of the corresponding modality.

Temporal Co-attention Aggregation. The network structure of TFM is illustrated in Fig.5 (b). Similar to SFM, we also design a co-attention module to fuse and aggregate the temporal information of two modals. Specifically, the temporal features of the two modals differ significantly, so we employ two separate FFN layers to map the unique temporal features of the two modals separately.

Operation For the feature maps $F_s \in \mathbb{R}^{C_2^s \times T_2^s \times H_2}$, and $F_k \in \mathbb{R}^{C_2^k \times T_2^k \times N_2}$, mean pooling is used in the spatial axis to get the temporal feature embedding $f_s^T \in \mathbb{R}^{C_3 \times T_2^s}$ and $f_k^T \in \mathbb{R}^{C_3 \times T_2^k}$, respectively. The embedding modeling operation is applied to these feature embedding for temporal modeling. These feature embeddings are then fed into the co-attention structure for feature fusion and enhancement and a temporal fusion feature representation Y_t is obtained.

4. EXPERIMENTS

4.1. Datasets and Evaluation Protocol

CASIA-B [26] is the most popular dataset for the cross-view gait recognition task. It contains 124 subjects where six sequences are sampled in normal walking (NM), two sequences are in walking with a bag (BG), and the rest are in walking in coats (CL). Each walking has 11 views which are uniformly distributed in $[0°, 180°]$ at an interval of $18°$. In total, there are $(6 + 2 + 2) \times 11 = 110$ walking sequences per subject. Following the large-sample training (LT) settings in [5], our experiments take the first 74 subjects as the training set and the rest 50 as the test set. For evaluation, each subject's first four normal walking sequences are regarded as the gallery, and the rest are regarded as the probe.
OUMVLP is the largest public gait dataset which has released both the silhouette data [11] and the skeleton data [1], in which the skeleton data is extracted by AlphaPose [10] and Openpose [3]. It contains 10307 subjects, 14 views per subject, and 2 walking sequences (#00-#01) per view. For fair comparison with previous state-of-the-art (SOTA) methods, we conduct experiments following the same protocol as [5, 9, 20], the 10307 subjects are divided into two groups: 5153 training and 5154 testing subjects. For evaluation, sequences#01 are kept in the gallery, and sequences#00 are regarded as the probe.

4.2. Training Details

Input. We adopt the same preprocessing approach as [5] to obtain gait silhouettes for CASIA-B and OUMVLP. The silhouette image of each frame is normalized to the size 64 × 44. For the CASIA-B, in which skeleton is not available, we utilized HRNet [22] to extract skeleton data. For the OUMVLP, we directly use the skeleton data of AlphaPose [10] provided by the OUMVLP-Pose [1].
Setting.All experiments utilize AdamW optimizer with a weight decay of 1e-4. For CASIA-B dataset, the batch size $P \times K$ is set to 8×16. During training, input sequences are set to a length of 64. During testing, entire sequences are utilized for gait feature extraction. The iteration number is set to 12K. Specially, we found that the SkEM and SiEM required a higher learning rate (LR) than the SFM and TFM for faster convergence. Therefore, LR in fusion modules is set to $0.1\times$ as that in the encoding module. And encoding module's LR is first set to 1e-3 and reset to 1e-4 after 5K. For OUMVLP dataset, the batch size $P \times K$ is set to 32×8. The iteration number is set to 60K. The SkEM module based on skeleton data performs poorly on the OUMVLP dataset. Therefore, we downscale the skeleton features in the spatial dimension by mean operation before performing the concatenation operation in SFM module. The LR is first set to 1e-4, reset to 1e-5 after 50K. According to the statistics of the CASIA-B dataset, the average number of frames in a gait cycle is 28 and the Encoder modules downscale the temporal dimension by a factor of four. Therefore, the cycle size s of the CPE is set to $28/4 = 7$.

4.3. Comparison with State-of-the-Art Methods

Evaluation on CASIA-B. Tab.1 shows a comparison between the SOTA methods and the proposed MMGaitFormer framework. It can be seen that our method achieves the best average accuracy in all three conditions. Compared with SOTA silhouette-based gait recognition method GaitGL [19], our method improves by **+4.6%** on mean accuracy. Compared with the skeleton-based method GaitGraph [23], our method obtains an impressive improvement by **+20.1%**. As shown in Tab.1, the proposed MMGaitFormer meets a new state-of-the-art, and the mean rank-1 accuracy is **96.4%**, which outperforms our baseline methods SkEM (**+22.1%**) and SiEM (**+6.4%**) by a large margin. Moreover, we further explore the effect of different walking conditions (NM, BG, and CL). For our proposed MMGaitFormer, the recognition accuracy in these conditions is **98.4%**, **96.0%**, and **94.8%**, respectively. It can be observed that the proposed method has an excellent performance in both normal and complex conditions. Significantly, the performance of ours is much better than that of GaitGL [19] in CL conditions by **+11.2%**. The impressive experimental results prove that the complementary advantages of skeleton and silhouette are used to obtain the great potential of robustness to clothing changes in gait recognition.
Evaluation on OUMVLP. We further evaluate the performance of the proposed method on the OUMVLP dataset, which is the worldwide largest public gait dataset. As shown in Tab.2, MMGaitFormer meets a new state-of-the-art performance and the mean rank-1 accuracy is **90.1%** which increases by **2.5%** compared with our baseline method, *i.e.*, SiEM. The improvement is smaller compared to the improvements in CASIA-B. Considering that the main improvements in CASIA-B were made on CL condition, OUMVLP contains only normal walks, which may lead to fewer improvements. Moreover, for the skeleton-based methods, both the benchmark method CNN-Pose [1] provided by the OUMVLP-Pose dataset and our re-implementation baseline method GaitGraph [23] perform poorly on OUMVLP, which may be one of the possible reasons for limiting the performance of our method. Again, it is worth mentioning that our method outperforms all SOTA silhouette-based methods while training only 1/4 of the epoch. Furthermore, we anticipate the possibility of further improving the results by utilizing improved SkEM and SiEM modules, which will be explored in future research.

4.4. Ablation Study

Effectiveness of SFM and TFM. To validate the effectiveness of the proposed SFM and TFM, we conducted experiments to compare the performance of our single modal en-

Table 1. The rank-1 accuracy (%) on CASIA-B dataset under all view angles with different conditions, excluding identical-view case. * means our reimplementation for encoding module.

Gallery NM		$0° - 180°$											mean
Probe	Methods	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	PoseGait [17]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	GaitGraph [23]	85.3	88.5	91.0	92.5	87.2	86.5	88.4	89.2	87.9	85.9	81.9	87.7
	GaitGraph* (SkEM)	82.3	84.1	83.7	85.4	84.0	82.8	85.0	81.7	84.6	86.5	81.8	83.8
	GaitNet [27]	93.1	92.6	90.8	92.4	87.6	95.1	94.2	95.8	92.6	90.4	90.2	92.3
	GaitSet [5]	90.8	97.9	**99.4**	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart [9]	94.1	98.6	99.3	**98.5**	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GaitGL [19]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	**99.3**	98.8	94.0	97.4
	GaitGL* (SiEM)	95.1	98.6	99.0	97.4	94.9	93.5	96.2	98.6	99.0	97.5	90.9	96.4
	MMGaitFormer (ours)	**98.1**	**98.6**	99.0	98.1	**98.4**	**97.8**	**98.1**	**99.0**	99.2	**99.1**	**97.3**	**98.4**
BG	PoseGait [17]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	GaitGraph [23]	75.8	76.7	75.9	76.1	71.4	73.9	78.0	74.7	75.4	75.4	69.2	74.8
	GaitGraph*(SkEM)	67.5	72.4	72.7	71.2	72.4	72.3	73.1	73.4	70.6	69.8	65.5	71.0
	GaitNet [27]	88.8	88.7	88.7	94.3	85.4	92.7	91.1	92.6	84.9	84.4	86.7	88.9
	GaitSet [5]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart [9]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	GaitGL [19]	92.6	**96.6**	96.8	95.5	93.5	89.3	92.2	96.5	**98.2**	**96.9**	91.5	94.5
	GaitGL* (SiEM)	89.2	94.9	94.3	93.1	90.0	86.6	88.4	93.3	96.3	95.3	84.6	91.4
	MMGaitFormer (ours)	**97.1**	95.9	**97.1**	**95.7**	**96.1**	**95.2**	**95.2**	**97.1**	97.3	96.1	**93.5**	**96.0**
CL	PoseGait [17]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
	GaitGraph [23]	69.6	66.1	68.8	67.2	64.5	62.0	69.5	65.6	65.7	66.1	64.3	66.3
	GaitGraph*(SkEM)	65.2	66.8	65.7	64.8	70.9	64.9	72.1	68.9	69.9	70.3	69.1	68.1
	GaitNet [27]	50.1	60.7	72.4	72.1	74.6	78.4	70.3	68.2	53.5	44.1	40.8	62.3
	GaitSet [5]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart [9]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GaitGL [19]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GaitGL* (SiEM)	77.6	88.3	90.5	85.4	81.7	75.0	81.2	84.7	87.2	83.7	68.0	82.1
	MMGaitFormer (ours)	**93.9**	**98.0**	**96.9**	**96.0**	**93.7**	**91.6**	**93.5**	**96.4**	**96.5**	**95.7**	**90.2**	**94.8**

Table 2. The mean rank-1 accuracy (%) on OUMVLP excluding the identical-view cases. * means our reimplementation for encoding module.

Method	Input	Mean Acc
CNN-Pose [1]	skeleton	20.4
GaitGraph* [23](SkEM)		21.1
Gaitset [5]	silhouette	87.1
GaitPart [9]		88.7
GLN [14]		89.2
GaitGL [19]		89.7
GaitGL* (SiEM)		87.6
BiFusion [21]	fuse	89.9
MMGaitFormer(ours)		**90.1**

coding module with that of our multi-modal network. As shown in Tab.3, the first two rows show the averaged accuracies of SkEM and SiEM, which could be considered as our baseline. From the last three rows, we can observe that: (1) Great performance gains were achieved by using only SFM, especially in the CL condition, which demonstrates that the complementary fusion of two modalities in spatial can significantly improve gait recognition in the presence of occlusion. (2) While there is no significant performance gain from using only TFM, its performance is still better than using only the skeleton for recognition. (3) Our approach achieves the best performance when using both SFM and TFM for feature fusion, demonstrating that our two-branch fusion framework can aggregate both temporal and spatial features of the two modalities for more comprehensive gait recognition. We can observe that the improvement from SFM is much more significant than the improvement from TFM. Considering that vision contains much more information than temporal information in the task of video recognition, we can still regard TFM as a practical auxiliary fusion module.

Analysis of Spatial Fusion Module. (1) We first perform an ablation study on the co-attention structure in SFM. The first two rows in Tab.4 show the comparison between the cross-attention using only a single modal and the co-attention structure of two modals. As shown in Tab.4, any cross-attention blocks removal leads to performance degra-

Table 3. Ablation studies on the CASIA-B dataset. The results are rank-1 accuracies averaged on all 11 views, excluding identical-view cases.

SkEM	SiEM	TFM	SFM	NM	BG	CL	Mean
✓				83.8	71.0	68.1	74.3
	✓			96.4	91.4	82.1	90.0
✓	✓	✓		84.4	74.4	66.3	75.0
✓	✓		✓	98.0	94.6	92.7	95.1
✓	✓	✓	✓	**98.4**	**96.0**	**94.8**	**96.4**

dation, reduced average accuracy by **-17.7%** and **-2.7%** respectively. The co-attention structure using two cross-attention blocks achieves the best performance. We can conclude that each cross-attention block can effectively improve gait recognition performance. The co-attention structure can better integrate the complementary advantages of the skeleton and silhouette. (2) To validate the effectiveness of our proposed Fine-grained Body Parts Fusion (FBPF) strategy, we conduct the ablation experiments by removing the attention mask used for Fine-grained Body Parts Fusion in SFM. As shown in the last two rows of the Tab.4, the proposed fusion strategy improves the average rank-1 accuracy by **+2.4%**, which proves that our strategy can effectively guide the fusion of aligned local features, helping the model to converge better and achieve better performance.

Table 4. Analysis of Spatial Fusion Module. **w/o Sil-CA**: remove the cross attention block which query is silhouette feature from the co-attention of SFM. Similarly, **w/o Ske-CA**: remove the cross attention block which query is skeleton feature. **w/o mask**: remove the pre-defined attention mask for FBPF in SFM.

Method	NM	BG	CL	Mean
Ours w/o Sil-CA	87.6	78.2	70.4	78.7
Ours w/o Ske-CA	97.8	94.0	89.3	93.7
Ours w/o mask	97.0	92.8	92.4	94.0
Ours	**98.4**	**96.0**	**94.8**	**96.4**

Analysis of Temporal Fusion Module. In Tab.5, we show the effectiveness of our temporal embedding modelling in the Temporal Fusion Module. (1) When no Embedding Modeling is performed on the input sequence of TFM, the average accuracy decreased by **-0.8%**. In particular, the results without EM are essentially the same as those without TFM (The fourth row in Tab.3). The result demonstrates that temporal modelling can capture temporal relationship information for better fusion. (2) When using the vanilla Position Embedding (PE) for Embedding (shown in Fig.6), the accuracy reduced by **-2.3%**. Considering that PE does not fully consider the feature of gait cycle process, the direct introduction of too many training parameters may lead to poor model performance because of overfitting. The result also demonstrates that our proposed Cycle Position Embedding(CPE) model the temporal information of gait sequences more effectively.

Table 5. Analysis of Temporal Fusion Module. **w/o EM**: remove the Embedding Modeling in TFM, which means no temporal modeling. **w/ EM (PE)**: Embedding Modeling by vanilla Position Embedding [8]. **w/ EM(CPE)**: Embedding Modeling by our proposed Cycle Position Embedding.

Method	NM	BG	CL	Mean
Ours w/o EM	98.1	94.8	94.1	95.6
Ours w/ EM (PE)	97.3	93.5	91.5	94.1
Ours w/ EM (CPE)	**98.4**	**96.0**	**94.8**	**96.4**

Comparison with different fusion approaches. To ensure fair comparisons with single-modal-based approaches, we adopt a careful experimental design that compares our approach to different fusion strategies. We introduce two strategies of global feature fusion for comparison, as described in Section 3.3 and illustrated in Figure 3 (a). The results of the comparative experiments are presented in Table 6, which shows that our approach achieves a mean rank-1 accuracy improvement of **+2.0%** over the concatenation-based fusion approach. Furthermore, when compared to state-of-the-art multi-modal gait recognition methods, our proposed MMGaitFormer achieves a significant **2.7%** improvement in recognition accuracy, particularly in the challenging CL conditions. These results demonstrates the effectiveness of our proposed fine-grained fusion method, which is a more comprehensive approach to better exploit the complementary advantages of silhouette and skeleton.

Table 6. Comparison with different Fusion module. **add fusion**: global feature fusion with add operation, **cat fusion**: global feature fusion with concatenation operation.

Method	NM	BG	CL	Mean
add fusion	97.3	92.8	91.7	93.9
cat fusion	97.6	93.2	92.4	94.4
TransGait [15]	98.1	94.9	85.8	92.9
BiFusion [21]	**98.7**	96.0	92.1	95.6
Ours	98.4	**96.0**	**94.8**	**96.4**

5. Conclusion

Motivated by the complementary strengths of the silhouettes and skeletons for comprehensive gait representation for recognition, we propose a transformer-based multi-modal framework called MMGaitFormer. In this work, we propose a Spatial Fusion Module and a Temporal Fusion Module to perform fine-grained fusion at the spatial level and fine-aligned fusion at the temporal level. Extensive experiments have shown the effectiveness of our framework.

References

[1] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):421–430, 2020. 6, 7

[2] Imed Bouchrika and Mark S Nixon. Model-based feature extraction for gait analysis and recognition. In *International conference on computer vision/computer graphics collaboration techniques and applications*, pages 150–160. Springer, 2007. 1, 2

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 6

[4] Francisco M Castro, Manuel J Marin-Jimenez, Nicolás Guil, and Nicolás Pérez de la Blanca. Multimodal feature fusion for cnn-based gait recognition: an empirical comparison. *Neural Computing and Applications*, 32:14173–14193, 2020. 2

[5] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, volume 33, pages 8126–8133, 2019. 1, 2, 6, 7

[6] Yufeng Cui and Yimei Kang. Gaittransformer: Multiple-temporal-scale transformer for cross-view gait recognition. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 2

[7] Ruben Delgado-Escano, Francisco M Castro, Julián Ramos Cózar, Manuel J Marín-Jiménez, and Nicolas Guil. An end-to-end multi-task and fusion cnn for inertial-based gait recognition. *IEEE Access*, 7:1897–1908, 2018. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5, 8

[9] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020. 1, 2, 4, 6, 7

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017. 3, 6

[11] Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE Transactions on Information Forensics and Security*, 14(1):102–113, 2018. 6

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[13] Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll. The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation*, 25(1):195–206, 2014. 2

[14] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European Conference on Computer Vision*, pages 382–398. Springer, 2020. 1, 7

[15] Guodong Li, Lijun Guo, Rong Zhang, Jiangbo Qian, and Shangce Gao. Transgait: Multimodal-based gait recognition with set transformer. *Applied Intelligence*, pages 1–13, 2022. 2, 8

[16] Rijun Liao, Chunshui Cao, Edel B Garcia, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Chinese conference on biometric recognition*, pages 474–483. Springer, 2017. 1, 2

[17] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 1, 2, 7

[18] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3054–3062, 2020. 2

[19] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *CVPR*, pages 14648–14656, 2021. 1, 4, 6, 7

[20] Beibei Lin, Shunli Zhang, Xin Yu, Zedong Chu, and Haikun Zhang. Learning effective representations from global and local features for cross-view gait recognition. *arXiv preprint arXiv:2011.01461*, 4(6), 2020. 2, 4, 6

[21] Yunjie Peng, Saihui Hou, Kang Ma, Yang Zhang, Yongzhen Huang, and Zhiqiang He. Learning rich features for gait recognition by integrating skeletons and silhouettes. *arXiv preprint arXiv:2110.13408*, 2021. 2, 4, 7, 8

[22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 3, 6

[23] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: graph convolutional network for skeleton-based gait recognition. In *2021 ICIP*, pages 2314–2318. IEEE, 2021. 1, 2, 4, 6, 7

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[25] Likai Wang and Jinyan Chen. A two-branch neural network for gait recognition. *arXiv preprint arXiv:2202.10645*, 2022. 2, 4

[26] Shiqi Yu, Daoliang Tan, and Tieniu Tan. 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)*, volume 4, pages 441–444. IEEE, 2006. 2, 6

[27] Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu. On learning disentangled representations for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 7