Gait Pyramid Attention Network: Toward Silhouette Semantic Relation Learning for Gait Recognition

Jianyu Chen, Zhongyuan Wang, *Member, IEEE*, Peng Yi, *Member, IEEE*, Kangli Zeng, *Student Member, IEEE*, Zheng He, and Qin Zou

***Abstract*—Gait recognition refers to video-based biometric techniques for recognizing subjects by walking patterns under a long-distance situation. Despite the progress of existing gait recognition methods, the recognition ability of subjects in carrying situations is still limited (e.g., carrying bags or wearing coats or jackets). To address this issue, this paper proposes to extract human gait information from different ranges of receptive fields, providing richer internal features of the deep network. Moreover, we propose two attention mechanisms, Local Pyramid Attention and Global Attention Fusion Learning, to focus on the key features in human gait from different perspectives. Depending on the different attention mechanisms employed in the network, three network variants are derived, where the Gait Pyramid Attention Network (GPAN) contains two attention mechanisms, while GPAN-P and GPAN-L contain a single attention mechanism. We evaluated our method on two large datasets, CASIA-B and OUMVLP. Experiments show that the proposed network gives an average rank-1 accuracy of 97.8% on CASIA-B under normal walking conditions. We also achieve 94.2% and 81.8% accuracy on CASIA-B dataset under the complex bag-walking and coat-walking scenarios, which are dramatically superior to the state-of-the-art methods.**

***Index Terms*—Gait recognition, attention mechanism, deep learning, biometrics, feature representation.**

I. Introduction

GAIT refers to the walking style of the subject, which can be used to identify the subject. Unlike other biometric features such as the face, fingerprints, veins, iris, etc., gait features can be captured from a distance by ordinary or low-resolution cameras without requiring cooperation. Therefore, gait recognition enjoys huge potential in authentication, social security and crime prevention. In real-world scenarios, gait recognition will be disturbed by external factors, such as the dressing or backpack conditions of the subject, and viewpoint between the camera and the subject. Considering all these factors, gait recognition has always been a challengeable issue.

Recently, several gait recognition methods have been developed to tackle these concerns [1], [2], [3], [4]. Additionally, an apparent trend in gait recognition indicates migration from traditional methods to deep learning-based solutions. Generally, existing deep learning gait recognition methods can be roughly divided into two categories: (*i*) compression of all gait contours into a single image or a template containing gait information. Wu et al. [5] initially adopted convolutional networks to capture gait features from gait energy images (GEI). Then, He et al. [6] proposed a multi-task generative adversarial network (MGAN) and periodic energy images (PEI) for learning more perspective-specific gait features and simultaneously extracting temporal information. Dhiman and Vishwakarma [7], [8] proposed gait energy maps for abnormal human behavior recognition. Through a comprehensive and exhaustive survey of abnormal behavior recognition methods, they proposed a robust framework to perform video-based abnormal human behavior recognition. Vishwakarma et al. [9] exploited changes in silhouette orientation in key frames to identify anomalous human activities. Aggarwal and Vishwakarma [10] proposed to use shape descriptors based on Zernike moments to detect the presence of covariates, and then used segmentation approach to crop out the average ability silhouette parts infected by covariates, finally using Spatial Distribution of Oriented Gradients and Mean of Directional Pixels methods to perform feature extraction. (*ii*) extract temporal representation from the original gait silhouette sequence directly. Liao et al. [11] proposed a gait network that extracts features from gait silhouettes and uses long short-term memory to model the temporal information of gait sequences. As the discontinuous input may lead to severe degradation of recognition performance, Chao et al. [4] proposed the GaitSet, which directly treats gait silhouettes as a collection to extract temporal information instead of sequencing each frame like a video, and improves the flexibility of gait recognition. Fan et al. [12] pointed out that each part of the human body has its expression, where local micro-motion features are the most significant features of human gait, and thus proposed GaitPart, which has achieved remarkable results. Although these methods have contributed to the development of gait recognition, conditions like carrying bags and wearing

Manuscript received 8 November 2021; revised 16 January 2022 and 5 June 2022; accepted 6 October 2022. Date of publication 11 October 2022; date of current version 5 December 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFF0602102; in part by the National Natural Science Foundation of China under Grant U1903214, Grant 62071339, Grant 61872277, Grant 62171324, and Grant 62072347; and in part by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant GML-KF-22-16. This article was recommended for publication by Associate Editor I. Kakadiaris upon evaluation of the reviewers' comments. *(Corresponding author: Zhongyuan Wang.)*

Jianyu Chen, Zhongyuan Wang, Peng Yi, and Kangli Zeng are with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: chenjyu@whu.edu.cn; wzy_hope@163.com; yipeng@whu.edu.cn; kangli.zeng@whu.edu.cn).

Zheng He and Qin Zou are with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: hezheng@whu.edu.cn; qzou@whu.edu.cn).

Digital Object Identifier 10.1109/TBIOM.2022.3213545

coats can change the gait appearance in real scenarios, which poses a significant challenge to gait recognition. Moreover, existing methods [3], [4], [12] still fail to provide accurate recognition of appearance changes caused by covariates.

In response to the above issues, inspired from the Feature Pyramid Networks [13] concept in object detection, we propose a multi-scale pyramidal convolutions combination to extend the receptive field, aiming to extract multiple internal features of the human body in carrying-bag regions. Meanwhile, we apply the attention mechanism to learn semantic relations from local and global perspectives, prompting the network to pay more attention to interesting contextual information. Specifically, our work includes the following three main contributions.

- A novel gait pyramidal attention network (GPAN) is proposed, which uses multi-scale receptive fields to learn the relations of gait representations within different regions, thus extracting rich contextual information as the internal feature.
- Two different attention mechanisms (Local Pyramid Attention and Global Attention Fusion Learning) are suggested to explore the regions of attention from local and global perspectives. Moreover, three network variants are derived by the optimal combination of the different attention mechanisms.
- Extensive experiments confirm that our method achieves competitive results on CASIA-B [14] and OUMVLP [15], outperforming the state-of-the-art gait recognition methods. It indicates that GPAN is robust to both changes in view and human appearance silhouette.

In this paper, the rest of the sections are organized as follows. The subsequent section will describe recent progress in gait recognition, feature pyramid models, and attention mechanisms. Next, we present the proposed gait method, followed by a comparative analysis of the proposed method with other gait methods on the gait datasets. Then, we perform ablation experiments and a visual description of the proposed module. Finally, the entire paper is summarized.

II. Related Work

We introduce the developments of gait recognition methods, pyramid networks, and attention mechanisms.

A. Gait Recognition

Body-Based Representation: Methods in related domains are performed on the silhouette or skeleton of the subject. Among them, silhouettes are typical representations in gait missions and can be computed by bigamizing each image after excluding the background. Gait silhouettes have been shown to describe the state of an individual through a single frame, which approaches forces recognition methods to pay more attention to human gait than other non-gait factors. Certainly, gait silhouettes are also more sensitive to changes in the appearance of an individual (e.g., wearing different clothes and carrying conditions). Skeleton-based body representations can be captured by depth-sensing cameras or predicted by posture estimation algorithms. The static or dynamic features of the human body are obtained by the joint movements of the human skeleton [16]. Compared to silhouette-based methods, skeletons consider the location of joints and are generally robust to viewpoint changes while offering more accuracy for identifying changes in appearance. However, skeleton-based methods rely heavily on joint detection and generally use pose estimators which also increases the computational burden [17]. Considering these limitations, some excellent model-based methods [11], [18], [19] have broken the bottleneck of pose estimation and achieved the competitive performance. For example, Liao et al. proposed PoseGait [18] to construct specialized features based on 3D pose information for gait recognition, and they proposed a pose-based temporal-spatial network (PTSN) [11], which handles changes in clothing and carrying conditions based on the human pose. Li et al. [19] proposed an end-to-end model-based method that extracts gait features via an SMPL model. In addition, An et al. [20] proposed a large-scale human posture database, OUMVLP-Pose, to address the lack of data resulting in model-based approaches that cannot be fully explored.

Temporal-Based Representation: In this context, we will describe the methods for processing temporal information in gait tasks. Existing researches are generally classified into template-based and sequence-based approaches. The former compresses the silhouette sequences into a single image and aggregates temporal information by averaging the silhouettes over at least a gait period. Template-based methods include: (*i*) gait energy maps [21], which are the primary method for averaging gait silhouette maps over one cycle; (*ii*) chrono gait images, [22] where the silhouettes on each gait image are extracted and encoded into an individual map with the single-channel mapping function; (*iii*) frame-difference energy images [23], with a clustering and denoising algorithm retaining motion information, which is valid when the silhouettes are incomplete; (*iv*) gait entropy images [24], where the entropy is calculated first for each pixel in a single frame and then averaged over a single gait template; (*v*) period energy images [6], where a multi-channel mapping function based on frame amplitudes is used to obtain spatiotemporal information. In contrast, sequence-based methods extract a series of gait silhouettes directly as input, including: (*i*) 3D CNN methods [25], [26], which use 3D convolution to extract spatiotemporal information directly from sequences; (*ii*) LSTM methods [3], [27] that fuse gait sequence frames with LSTM units; (*iii*) graph convolution methods [28], [29] that use the spatiotemporal graph attention network to obtain graph relationships between gait frames. (*iv*) dynamic motion image methods [30], [31], [32] recognize human activity through motion rhythm information and postures in dynamic and RGB images.

Feature Representation: Regarding the range of feature representations, the description is offered from the global and local perspectives. Exploring a gait silhouette from the whole is called global representation learning, while studying the gait map after segmentation is called local representation learning, such as body components, vertical bins, and horizontal bins. Then local features of these segments are fed into the deep network for processing. An advantage of the global

representation is that it is more sensitive to occlusions, appearance changes, and the absence of body components [4], [33]. Local representation learning frequently contributes differently to the final recognition performance, thus improving the overall recognition performance of gait recognition methods [4], [12]. Meanwhile, the potential relationships between local features are worth to be explored, and such information is beneficial to enhance the robustness of gait recognition methods against perspective changes [2]. Some representative feature methods have achieved excellent performance. Verlekar et al. [34] proposed to use computed perceptual hashes on the leg regions of gait energy images to obtain walking directions and then compare them with the perceptual hashes of different walking directions in the database to identify subjects.

B. Feature Pyramid Networks

Pyramid features are extensively used in the object detection domain to overcome scale constraints. One challenging problem for object detection is scale variations. The intuitive idea is to use multi-scale image pyramids, but the high computational burden **cannot** be ignored. Hence image feature pyramids are proposed. The feature pyramid approach, which constructs and uses feature pyramids within the network, has frequently handled scale variations due to their low computational overhead. Feature pyramid networks (FPN) [13] can be enriched with different levels of semantic feature enhancement through top-down paths and lateral connections. FPN structure can fuse low-resolution feature maps with strongly semantic information and high-resolution feature maps with weak semantic information, which captures rich spatial information in less computational complexity. EfficientDet [35] reuses bidirectional paths repeatedly to enable higher-level feature fusion. NAS-FPN [36] constructs more robust pyramid structures by using the neural network search. Temporal pyramid network (TPN) [37] exploits pyramidal structures to model visual speed in video and extends it to action recognition.

C. Attention Mechanism

Attention mechanisms enhance the representation of informative features while suppressing features that are less useful, thus enabling to focus on salient regions among the contextual content. The squeeze-excitation (SE) attention module [38] operates against the scale of feature channels to capture the correlation of channels between features. The convolutional block attention module (CBAM) [39] enriches the attention maps by adding maximum pooling operations to large channel attention. Based on CBAM, global second-order pooling (GSoP) [40] proposes a second-order pooling method to extract rich feature fusion. Furthermore, Non-Local (NL) [41] construct dense spatial feature maps and capture long-range dependencies by non-local operations. Inspired by NL, double attention network (A2Net) [42] introduces a new relational function that integrates spatial attention into the feature maps. Selective kernel network (SKNet) [43] proposes a dynamic attention-seeking mechanism that allows the neurons to adaptively rescale the perceptual domain according to the multi-scale feature maps. Split-attention network (ResNeSt) [44] proposes splitting attention blocks to assure that attention can interact across groups of feature maps. In order to process channel attention in the frequency domain, new multi-spectral channel attention is provided by frequency channel attention network (FCANet) [45]. Global context network (GCNet) [46] introduces a spatial attention module and develops a long-range channel dependency. Dual attention network (DANet) [47] adaptively combines both local and global features by aggregating attention modules from different branches. The majority of the above approaches simplify the computational cost and short-distance channel communication, which are still limited in balancing the long-term and short-term channel interaction. In this paper, we propose the Local Pyramidal Attention (LPA) and Global Attention Fusion Learning (GAFL) to model channel dependencies from different perspectives, which achieves the integration of attention while simplifying the model complexity and improving the recognition performance simultaneously.

III. Method

In this section, we introduce GPAN in detail, which exploits larger receptive fields to learn feature information against the human body carrying status while focusing on partial differential representations of the human body, aiming to learn features in semantic information in a more targeted manner. For further understanding, the whole framework is shown in Fig. 1. Firstly, we give an overview of the entire pipeline of GPAN. Then, static partial feature extraction structure PFE is explained. Finally, two different types of attentional mechanisms are described from different attentional viewpoints.

A. Pipeline

The gait silhouettes of the sequence T frames are fed into the Pyramid Silhouette-level Feature Extractor (PSF), as shown in Fig. 2. This block extracts spatial information exclusively from a single frame, which consists of three different scales of pyramidal convolutions designed to provide multi-scale interior features, where the corresponding internal feature of each original gait frame s_i is S_i.

$$\begin{aligned} S_i &= PSF(s_i) \\ &= \sum_{t=1}^{3} Pyconv_{n\times n}(s_i), \ \ n = 2t-1 \end{aligned} \tag{1}$$

where Pyconv means pyramidal convolution, t means the level of Pyconvs, i indicates index of the frames in the silhouette sequence. Then, the sequence of features as $S = \{S_i \,|\, i = 1, 2, 3, \ldots, N\}$, where S_i is a tensor with respect to the channel, height, and width dimensions.

Immediately afterward, these internal features are fed into a Multilayer Silhouette-level Feature Extractor (MFE), which contains three parts. To learn fine-grained partial spatial information from the rich internal features while addressing the limitation that the receptive fields decreases as the number of layers deepens. We present to use Partial Feature Extraction

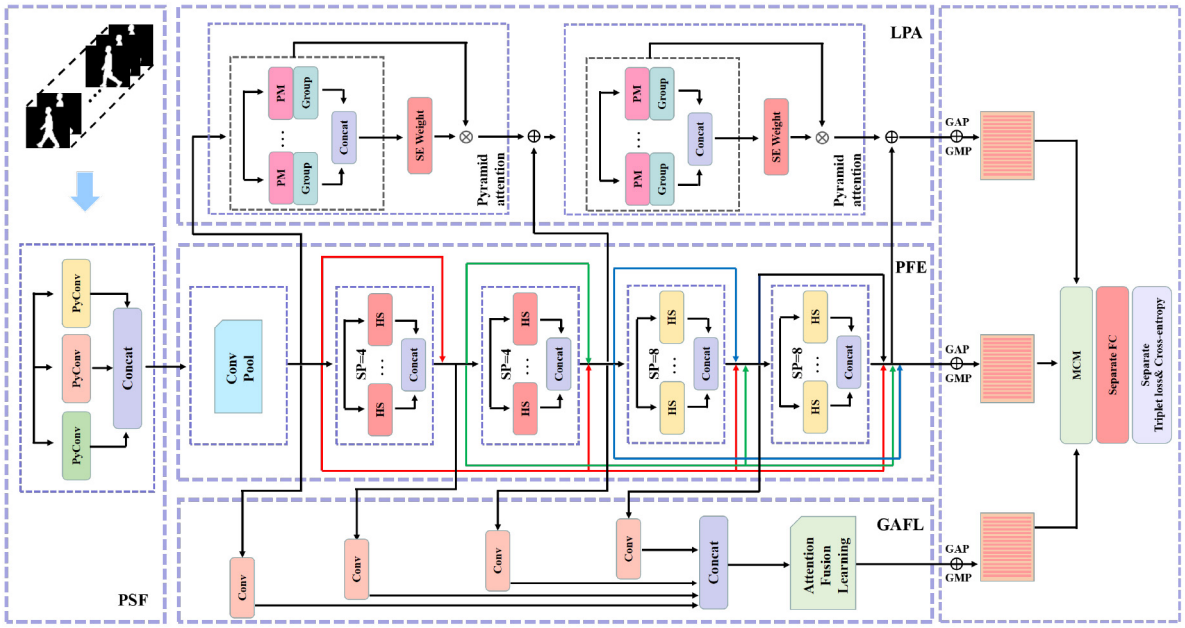


Fig. 1. The overall framework diagram of the proposed GPAN. It consists of five components, among which the former four components are used to extract spatial features, and the final component extracts temporal features and then performs classification. SP in the PFE indicates the number of horizontal split blocks. The red, green, and blue lines denote the dense residual connections from shallow to deep layers.

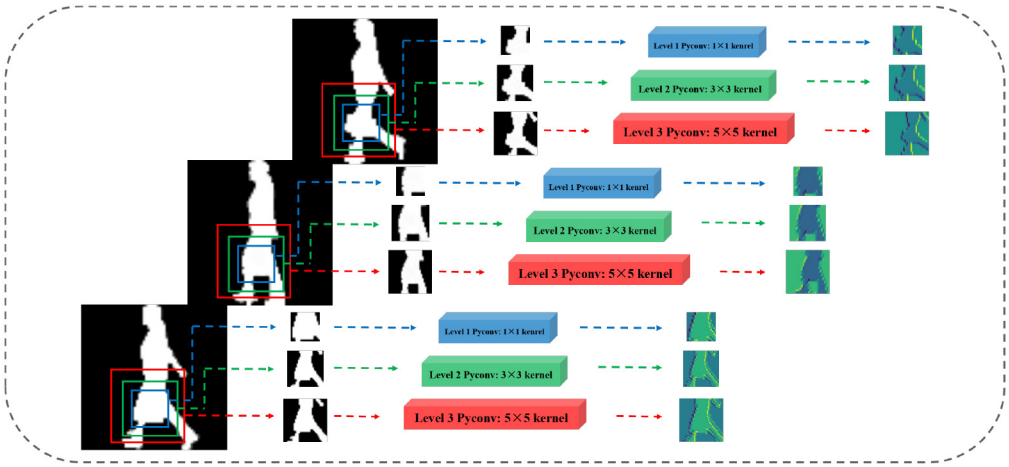


Fig. 2. The proposed pyramid silhouette-level feature extractor. Silhouettes are fed into Pyconvs with different perceptual fields for feature extraction, and then the obtained features are fused to provide richer interior features.

(PFE) to capture spatial features of partial information in each frame. Additionally, to ensure that more visual fields can be focused on the region of salient features, i.e., the silhouette differences caused by the subject in the walking with a bag (BG) and walking with a coat (CL) compared to walking normally (NM). Here, we propose two attentional mechanisms, including local pyramidal attention (LPA) and global attention fusion learning (GAFL). With the incorporation of attention, the network focuses on regions of interest more directly and thus extract more detailed features. Specifically, a sequence of detailed features denoted as $M = \{M_i \,|i = 1, 2, 3, \ldots, N\}$ can be obtained by

$$M_i = f_{MFE}(S_i), i = 1, 2, 3, \ldots, N \tag{2}$$

$$MFE = \{LPA; PFE; GAFL\} \tag{3}$$

Next, the Segmentation Pooling (SP) splits the features M_i horizontally to extract the distinguished part of information features in the human body. For the j-th part of M_i, the SP block transforms it into a vector $H_{j,i}$ using Global Average Pooling and Global Max Pooling.

$$H_{j,i} = Avgpool\big(M_{j,i}\big) + Maxpool\big(M_{j,i}\big) \tag{4}$$

Finally, each feature is split into the specified feature vectors that obtain a matrix. Each vector row corresponding to each matrix represents the j-th part of the gait variation. Thus, the spatiotemporal features of each part can be extracted by aggregating the vector rows in the matrix. In this regard, we draw on the excellent Micro-motion Capture Module (MCM) from existing work [12], expressed by

$$Vec = f_{MCM}\big(H_{j,i}\big) \tag{5}$$

where each feature vector will obtain the motion representation through MCM, extracting the motion rhythm information between features by aggregating these motion representations. Then, multiple independent FC layers will map the extracted vectors to the metric space to achieve the final individual recognition.

B. Pyramid Silhouette-Level Feature Extractor

Most networks use smaller kernels to extract features from the input since increasing the size entails high costs. In order to address the limitation that small kernels cannot cover large input regions, a series of small convolutions with pooling is used to reduce the size of the input and thus increase the receptive field of the network. However, existing work [48] has shown that this supposed empirical receptive field is much smaller than the theoretical receptive field. Secondly, down-sampling the input without extracting enough contextual information in advance can also affect the feature extraction process. For providing internal features with rich information to the deeper layers in the network, we learn the concept of pyramid network [13] from the object detection.

In the PSF, we propose to use pyramid convolutions combination (PConvs) to process the input silhouettes, which generates a diversity of complementary feature sequences. In the existing research [13], pyramidal networks are usually a hierarchical structure that stacks a certain number of kernels. Kernels have different spatial scales for each layer of the network, rising from bottom to top in the pyramid. As shown in Fig. 2, a parallel structure is used for our pyramidal convolutional combination, consisting of three kernels with different perceptual fields. As a result, the network can obtain richer multi-scale features, where kernels with smaller receptive fields emphasize minor targets and details, while kernels with larger receptive fields focus on larger targets and contextual information.

C. Multilayer Silhouette-Level Feature Extractor

The multilayer silhouette-level feature extractor consists of three components, including a Partial Feature Extraction Block and two Attention Extraction Blocks, designed to extract the appearance minutia features of each gait silhouette. Next, we describe the exact structure and function of each block separately in detail.

1) Partial Feature Extraction Block:

Description: Recently, some works [2], [4] have used partial representation methods to separate gait features into local regions. They show that partial representation learning has excellent potential for recognizing critical gait features. Therefore, as shown in Fig. 3, considering the different contributions of each partial representation of the human body to the final recognition performance, we use the Horizontal Split (HS) block to slice the feature map into several parts horizontally and convolve each part separately.

Motivation: Aiming to improve the fine-grained learning of partial spatial features, we propose to use partial convolution combination. The number of kernels in the combination is aligned with silhouette feature segments. Meanwhile, the perceptual fields of the kernels are constrained to ensure that more fine-grained feature information can be captured.

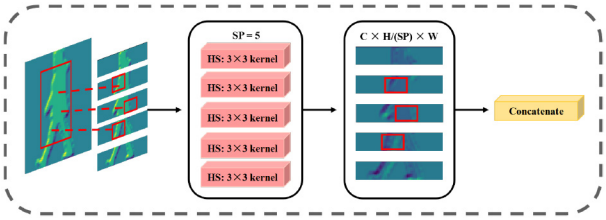


Fig. 3. The proposed PFE block. The pink kernel is the horizontal split block, where SP indicates the number of HS, which keeps consistent with the number of splits of silhouette features.

Operation: Initially, the input features are separated into preset parts along the height. Then, an equal number of kernels extract the features from each part separately. Finally, the extracted features of each part are concatenated again along the height into the original dimension. The specific structure of the PFE is given in Fig. 1, where the parameters of the predefined parts are based on the optimal results in subsequent ablation experiments.

2) Local Pyramid Attention Block:

Description: Attention mechanisms have been applied to many computer vision domains, such as image classification, object detection, semantic segmentation, scene analysis [49], [50]. Specifically, the attention mechanisms are classified into two types, i.e., channel attention and spatial attention. Recently, several works demonstrate that significant performance improvements have been achieved by using either channel attention, spatial attention, or a combination of both [42], [51]. Among them, the most commonly used approach for channel attention is the SE module [38], and the approach that considers both spatial and channel attention is the CBAM module [39]. However, there remain two issues to be addressed. The first one is how to efficiently exploit different scale appearance features to enrich the spatial feature. The second one is how to tackle the issue that channels or spatial attention cannot establish long-range channel dependence. Several research approaches [45], [46] have been proposed to resolve these two existing problems, however, they also bring higher model complexity and burden to the network.

Motivation: We intend to develop an attention module that strikes a balance between the cost and performance. In this work, a new efficient and effective block called Local Pyramid Attention is proposed. The block can process the input tensor at multiple scales and then integrates information from different scales in each channel feature map. Such a strategy allows more accurate fusion of adjacent contextual features to learn a richer multi-scale feature representation. The structure of the backbone combined with LPA is named GPAN-P.

Operation: As shown in Fig. 4, the input feature map X is divided into N parts along the channel dimension. For each part, the number of channels is $C' = C/N$. It is worth mentioning that C is divisible by N, which allows the input features to be extracted from multiple scales in parallel, resulting in a feature map from a single type of kernel. Accordingly, the corresponding spatial information on the feature map of

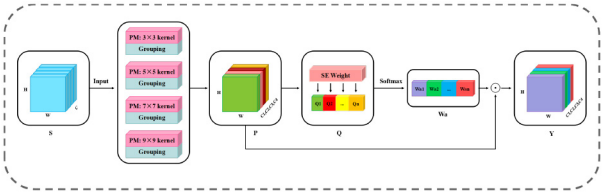


Fig. 4. The proposed LPA block. Left input features pass through the pyramid attention to generate the final feature maps. PM refers to the pyramidal convolution module, and S indicates the input features. P, Q, and Y correspond to Eqs. (8), (10), and (14), respectively. H, W, and C denote the height, width and channel of the feature, respectively.

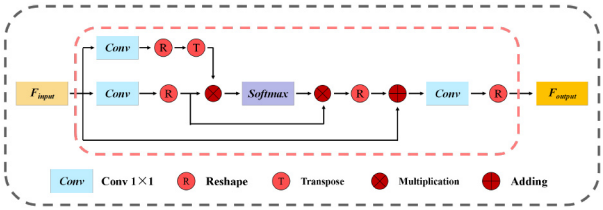


Fig. 5. The proposed GAFL block. Where the final feature tensor is obtained by adding initial features and attention-inducing features. The blue box refers to a convolution operation, R denotes a reshaping of the vector, and T indicates a transposition of the vector.

each channel can be obtained. To address the extra parameters brought by the increase in kernel scale, we incorporate group convolution [52] in the pyramid structure and develop a novel criterion to select the grouping. Among them, the correlation of multi-scale convolution with the number of groups is given by

$$N_G = 2^{\frac{K-1}{2}} \tag{6}$$

where K refers to the convolutional kernel size ($K = 3, 5, 7, 9$) and N_G means the number of groupings. Finally, the multi-scale feature after grouping is expressed by

$$P_i = PM(S_i)_{N_G}, \quad i = 1, 2, 3, \ldots, N \tag{7}$$

where PM represents group convolution and P_i stands for the different scales of the feature map. Therefore, all multi-scale features are obtained by fusion as follows

$$P = \Delta(P_1, P_2, P_3, \ldots, P_N) \tag{8}$$

where $\Delta(\cdot)$ denotes the concat operation, the vector of attention weights in different scales can be obtained by extracting the channel attention weight information from the multi-scale feature maps. Specifically, the SEWeight is used to obtain attention weights from different scales of the input feature maps, which is expressed as

$$Q_i = W_{SE}(P_i), i = 1, 2, 3, \ldots, N \tag{9}$$

where W_{SE} is the operation of the weight associated with attention, in this manner, we can merge the weight information from different scales and generate more targeted attention for feature maps. The implementation is detailed as

$$Q = \Delta(Q_1, Q_2, Q_3, \ldots, Q_N) \tag{10}$$

where Q denotes the multi-scale attentional weight vector, then different spatial scales are selected by using Softmax function across channel. The weights are expressed as

$$Wa_i = Soft\max(Q_i), \quad i = 1, 2, 3, \ldots, N \tag{11}$$

where Softmax is used to obtain the calibrated channel weights that contain both spatial and channel attention information. In this manner, channel attention interaction can be achieved from global and local perspectives. Thus the entire channel attention vector can be obtained as

$$Wa = \Delta(Wa_1, Wa_2, Wa_3, \ldots, Wa_N) \tag{12}$$

where Wa denotes the attention weight vector of the channel after the interaction, then the multi-scale channel attention weights are multiplied with the corresponding feature maps as

$$Y_i = \Phi(Wa_i, P_i), \quad i = 1, 2, 3, \ldots, N \tag{13}$$

where $\Phi(\cdot)$ denotes the multiplication in the channel dimension and Y_i denotes the feature map after acting with the channel attention weights. The final output can be expressed as

$$Y = \Delta(Y_1, Y_2, Y_3, \ldots, Y_N) \tag{14}$$

In summary, the LPA block can integrate multi-scale spatial information and cross-channel attention into each feature group. Therefore, a better interaction capability of channel information can be obtained with this attention block.

3) Global Attention Fusion Learning Block: Description: As shown in Fig. 1, we added dense residual connections [53] (red, green and blue lines) in the PFE block to increase the interaction between multiple layers of information. Although dense residual connections can directly transfer feature information from shallow layers to deeper layers, such connections do not take advantage of the interdependencies between layers, i.e., deeper information is not available to shallow layers. Given this, we envision treating feature maps in each layer as a specific class and associating specific classes between different layers. By capturing the dependencies between the features of different layers, the network can allocate varying attention weights to the features of different depth, thus enhancing the feature extracting ability of the model. Therefore, we propose a layer attention mechanism that can balance the feature information in each layer, aiming to learn the correlation between layers of the deep network and further improve the feature representation capability.

Motivation: The purpose of this attention is to address the inability of channel attention to weighing features in multi-scale layers, resulting in the continuous weakening of long-time information from shallow layers. Although shallow features can be delivered by skip connections, following long skip connections, they are treated equally across layers, which hinders the representational power of CNNs. Hence, we consider exploring the interrelationships between features at the hierarchy level and propose a GAFL block to address this issue, where the structure of the backbone combined with GAFL is named GPAN-L.

Operation: As shown in Fig. 5, the input of GAFL block is a feature group G extracted from multiple layers with

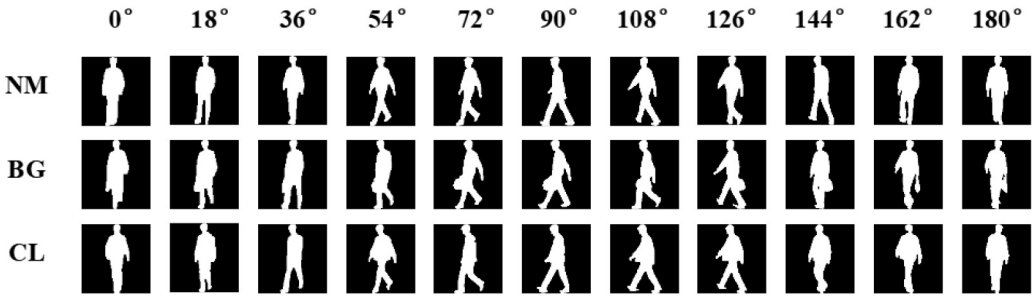


Fig. 6. Sampling on CASIA-B dataset. The first row is the sequence obtained from subjects walking normally (NM), the second row is the gait silhouettes of the subject walking with a bag (BG), and the third row is the motion trajectory captured in walking with a coat (CL).

TABLE I
THE EXACT STRUCTURE OF THE FEATURE EXTRACTION BASELINE. IN_CHANNEL, OUT_CHANNEL, KERNEL AND PADDING DENOTE THE INPUT CHANNEL, OUTPUT CHANNEL, KERNEL SIZE AND PADDING OF THE CONVOLUTION, RESPECTIVELY. AMONG THEM, SNUM MEANS THE NUMBER OF FILTERS SPLIT IN HSCONV

Feature Extract Backbone							
Block	Layer	Operation	In_Channel	Out_Channel	Kernel	Padding	SNum
PSF	Layer-1	PyConv	1	32	1	0	1
		PyConv	1	32	3	1	1
		PyConv	1	32	5	2	1
PFE	Layer-2	Conv	96	96	3	1	1
	Layer-3	MaxPool, Kernel = 2, Stride = 2					
	Layer-4	HSConv	96	128	3	1	4
	Layer-5	HSConv	128	128	3	1	4
	Layer-6	MaxPool, Kernel = 2, Stride = 2					
	Layer-7	HSConv	128	256	3	1	8
	Layer-8	HSConv	256	256	3	1	8

dimensionality $N \times H \times W \times C$. Initially, we convert the dimensionality of this feature group into a two-dimensional matrix of $N \times HWC$. Then, the correlation between the different layers is computed by multiplying the matrix with the corresponding transpose matrix. Finally, the transformed feature group is multiplied with the prediction correlation matrix and added with the input features to obtain the new feature group, as follows:

$$G_{out} = f_{GAFL}(G_{in}) \oplus G_{in} \tag{15}$$

where G_{in} represents the input and G_{out} represents the output after function calculation, $\oplus$ indicates accumulation operation. Such an approach allows the network to learn sufficient the feature interactions between multiple layers during feature extraction.

D. Implementation Details

We borrowed the GaitSet baseline and improved on the structure, as shown in Table I. The combination of two loss functions is widely applied [54], [55] in face recognition. To obtain a better performance, we train the network with a combination of cross-entropy loss and triplet loss, which are equally distributed. The batch size is (k, s) in the training process, where k refers to the number of subjects and s denotes the number of samples per subject. In the testing phase, gait videos are directly fed into the model for feature extraction. Significantly, to handle the uncertainty of gait video length in the training phase, we set a fixed-length input clip on the sampler for constraint. Operationally, the original video is first intercepted as a 30-40 frame clip, and then the sorted 30 frames are randomly sampled. It is important to emphasize that the clips less than 15 frames in the original video will not be sampled.

IV. EXPERIMENTS

We evaluate the proposed GPAN on the public gait datasets of CASIA-B [14] and OUMVLP [15]. In this section, we first present a detailed description of the two datasets. Then, the proposed method is compared with the existing state-of-the-art methods on publicly available datasets. Finally, a detailed ablation study is performed on CASIA-B for each component of the proposed method to verify the validity of each part in the model.

A. Datasets and Training Details

CASIA-B: CASIA-B [14] is a classical gait dataset contains 124 subjects as shown in Fig. 6. Every subjects contains 11 views, each of which consists of 10 sequences obtained by recording three different gait walking patterns of the subjects. The first 6 sequences were obtained under normal walking, the middle 2 sequences were obtained while the subject was carrying a bag (BG), and the remaining 2 sequences were obtained while the subject was wearing a coat or jacket (CL). For the sake of fairness, we strictly followed CASIA-B essential protocols [3]. In this case, the first 74 sequences are classified as the training set, and the other sequences are classified as the testing set. The first 4 sequences are defined as the gallery during testing, and the remaining 6 sequences are divided into three subsets with different walking conditions.

OUMVLP: The OUMVLP [15] is a large-scale dataset in extant gait research as shown in Fig. 7. It contains 10307 subjects and 259,013 gait sequences with a balanced gender distribution and age range of 2-87 years. According to the dataset protocol, the lists of subjects 5153 and 5154 are predefined, and corresponding training and testing sets are provided. Each subject contains 14 views with different angles $(0, 15, \ldots, 90;\ 180, 195, \ldots, 270)$, where the angle change is 15 for each step. The sequence is classified into gallery and probe sets in the test phase according to the subscript index.

Training Details: For data preprocessing, we follow the protocol in [15] to align and crop the input silhouettes to 64×44. We follow the optimization strategy in the extant work [4], using the Adam optimizer with a learning rate of **0.0001** and a momentum of 0.9, while setting the margin of triplet loss to 0.2. We report the detailed parameter settings

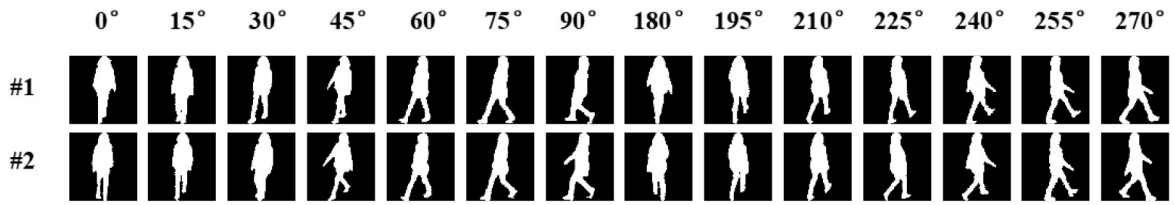


Fig. 7. Sampling on OUMVLP dataset. The two rows represent the cropped gait silhouettes of a subject captured from 14 different viewpoints in two sessions, respectively.

TABLE II
AVERAGED RANK-1 ACCURACIES ON CASIA-B UNDER NORMAL (NM) WALKING, EXCLUDING IDENTICAL-VIEW CASES

Gallery NM#1-4				0°-180°											Mean
Probe	Model	Year	Venue	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM#5-6	GaitNet [56]	2022	IEEE T-PAMI	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 [4]	2021	IEEE T-PAMI	91.1	99.0	99.9	97.8	95.1	94.5	96.1	98.3	99.2	98.1	88.0	96.1
	GaitPart [12]	2020	CVPR	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GLN [57]	2020	**ECCV**	93.2	99.3	99.5	98.7	96.1	95.6	97.2	98.1	99.3	98.6	90.1	96.8
	MT3D [25]	2020	ACM-MM	95.7	98.2	99.0	97.5	95.1	93.9	96.1	98.6	99.2	98.2	92.0	96.7
	GaitGL [58]	2021	ICCV	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	GPAN(ours)	-	-	97.2	99.2	99.3	99.1	96.9	94.8	97.5	98.5	99.7	99.0	93.2	97.7
	GPAN-L(ours)	-	-	97.5	99.3	99.4	98.4	96.7	94.5	97.4	98.7	99.7	99.4	94.5	**97.8**
	GPAN-P(ours)	-	-	96.5	99.2	99.5	98.6	96.9	94.5	97.0	97.9	99.7	98.8	93.0	97.4

TABLE III
AVERAGED RANK-1 ACCURACIES ON CASIA-B WITH CARRIED BAGS (BG), EXCLUDING IDENTICAL-VIEW CASES

Gallery NM#1-4				0°-180°											Mean
Probe	Model	Year	Venue	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
BG#1-2	GaitNet [56]	2022	IEEE T-PAMI	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 [4]	2021	IEEE T-PAMI	86.7	94.2	95.7	93.4	88.9	85.5	89.0	91.7	94.5	95.9	83.3	90.8
	GaitPart [12]	2020	CVPR	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	GLN [57]	2020	**ECCV**	91.1	97.7	95.2	97.8	92.5	91.2	92.4	96.0	97.5	95.0	88.1	94.0
	MT3D [25]	2020	ACM-MM	91.0	95.4	97.5	94.2	92.3	86.9	91.2	95.6	97.3	96.4	86.6	93.0
	GaitGL [58]	2021	ICCV	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	GPAN(ours)	-	-	93.6	97.4	97.8	95.8	92.3	88.3	93.3	95.8	97.7	96.3	88.5	**94.2**
	GPAN-L(ours)	-	-	93.5	97.7	97.6	95.1	90.9	86.7	91.1	95.8	97.8	96.5	89.4	93.8
	GPAN-P(ours)	-	-	91.6	96.8	96.7	94.9	91.6	87.3	91.8	95.4	97.1	96.4	88.5	93.5

TABLE IV
AVERAGED RANK-1 ACCURACIES ON CASIA-B WITH DIFFERENT CLOTHING (CL), EXCLUDING IDENTICAL-VIEW CASES

Gallery NM#1-4				0°-180°											Mean
Probe	Model	Year	Venue	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
CL#1-2	GaitNet [56]	2022	IEEE T-PAMI	50.1	60.7	72.4	72.7	74.6	78.4	70.3	68.2	53.5	44.1	40.8	62.3
	GaitSet [4]	2021	IEEE T-PAMI	59.5	75.0	78.3	74.6	71.4	71.3	70.8	74.1	74.6	69.4	54.1	70.3
	GaitPart [12]	2020	CVPR	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GLN [57]	2020	**ECCV**	70.6	82.4	85.2	82.7	79.2	76.4	76.2	78.9	77.9	78.7	64.3	77.5
	MT3D [25]	2020	ACM-MM	76.0	87.6	89.8	85.0	81.2	75.7	81.0	84.5	85.4	82.2	68.1	81.5
	GaitGL [58]	2021	ICCV	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GPAN(ours)	-	-	74.5	88.2	89.0	86.2	81.8	76.4	80.7	82.9	85.3	83.1	71.6	**81.8**
	GPAN-L(ours)	-	-	74.9	86.6	88.9	85.7	80.8	76.9	80.6	84.1	85.5	83.2	71.4	81.6
	GPAN-P(ours)	-	-	75.0	86.4	87.5	83.7	80.7	73.7	78.2	82.4	83.8	83.4	69.6	80.4

in Table VI for the training process. On CASIA-B [14], the batch size during training is (8, 8). The whole training process performs 130k iterations. The OUMVLP contains 20 times more sequences than CASIA-B. Considering the limitation of GPU memory, we set the batch size to (8, 16) during the training process, and the entire process is performed for 250k iterations. It is worth noting that the learning rate decreases to **0.00001** at 150k iterations, and other parameter values are kept constant. The whole process runs on a server equipped with 3 NVIDIA GeForce RTX 3090 GPUs.

B. Comparison With State-of-Art Methods

Evaluation on CASIA-B: In order to ensure an intuitive comparison of our proposed GPAN with other gait methods, the results in multiple angles and walking conditions are presented in Table II, Table III, Table IV. Specifically, compared to GaitNet [56], both methods have the same purpose of dealing with the carry problem, but the manner employed differs. GPAN outperforms GaitNet dramatically in all walking conditions, indicating that GPAN is more suitable for handling gait-complex environment problems. Compared to GaitSet [4], which treats the input sequence as an unordered set for gait recognition. However, such an approach does not consider the disparity of the body during walking. GPAN focuses on the global representation and the partially disparate representation of the human body in walking based on this baseline, resulting in improved performance, especially in the CL condition, with a 11.5% improvement in recognition accuracy. Compared with GaitPart [12], GPAN can not only learn the body discrepancy

Fig. 8. Comparison with multiple cross-view methods in 11 different probe angles (0°,..., 90°,..., 180°). Experiment samples are selected from walking with bags in CASIA-B.

but also excels in capturing the subtle local features, hence performing better in the BG and CL occlusion conditions. In contrast to GLN [57], although both methods use pyramid structures, they are fundamentally different. Moreover, GPAN adopts two attention mechanisms that focus more accurately on fine features targeted to silhouettes. Experimental results show that except for the BG condition, the other two walking states exhibit better performance. We speculate that the Compact block with dropout proposed by GLN [57] can provide a better fitting capability. Nevertheless, the recognition accuracy of GLN in the CL condition is much lower than GPAN, which indicates that the method is still highly constrained in the condition of noticeable appearance change. Compared with MT3D [25], which uses a 3D CNN to capture the temporal information in video sequences, achieving a significantly outperformed result than other gait methods proposed at the same stage. In contrast, our approach focuses more on fine-grained information on static frames. Except for the GPAN-P structure, the recognition performance of the other two variants is still better than MT3D, where the average recognition accuracy (three conditions) of the GPAN structure is superior to 0.83%. Compared to the latest GaitGL [58], which also extracts feature representations from global and local perspectives, our GPAN outperforms it by 0.4% in the NM condition. But in the complex situations BG and CL, our GPAN is slightly inferior to GaitGL. The reason for this phenomenon may be that GaitGL uses 3D convolution and local temporal aggregation to model the temporal features in gait sequences and thus it can extract more discriminative motion representations in the presence of occlusion in complex environments. To compare the performance of various methods in a more intuitive way, we take the subject walking in bags as an example, and Fig. 8 shows the comparison of recognition rates of multiple gait methods under 11 different views. Among them, GPAN including variants shows better recognition ability under 0°, 36°, 108°, 144°, 162°, and 180° angles. In 18° and 126°, the recognition accuracy remains essentially equivalent to the existing model. Where the GLN model shows higher performance than other contemporaneous gait methods in 54°, 72°, 90°, 108°, 126°, 144°, and 180° angles, we speculate that the Compact block proposed in GLN performs better than the baseline in fitting the temporal information, which also points out the direction for our further research. We plan to develop the modeling capability of GPAN for temporal information in the follow-up research. Anyway, the average recognition rate of GPAN is still superior to these methods.

Evaluation on OUMVLP: We have further evaluated the performance of GPAN on the OUMVLP dataset. For a fair comparison, we used the same training and testing protocols to compare the representative GaitSet, GaitPart methods for gait recognition, and the specific training parameters are shown in Table VI. We divide 10307 subjects into two parts, 5153 as the training set and 5154 as the testing set. Here, we select the GPAN structure with better data fitting ability for comparison. Table V shows the comparison results between GPAN and several representative methods, including GEINet,

TABLE V
AVERAGED RANK-1 ACCURACIES ON OUMVLP, EXCLUDING IDENTICAL-VIEW CASES

Model	Year	Venue	Probe View														Mean
			0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GEINet [59]	2016	ICB	11.4	29.1	41.5	45.5	39.5	41.8	38.9	14.9	33.1	43.2	45.6	39.4	40.5	36.3	35.8
GaitSet(our impl) [4]	2022	IEEE T-PAMI	63.3	77.5	83.9	84.3	78.5	82.1	77.9	67.0	75.4	81.7	82.4	76.7	80.0	75.3	77.6
GaitPart(our impl) [12]	2020	CVPR	59.7	74.6	83.6	84.5	76.0	80.8	77.1	62.5	73.1	82.2	82.8	74.3	79.1	74.6	76.1
GPAN(ours)	-	-	69.9	81.2	87.1	87.4	81.6	85.2	82.7	73.0	79.4	85.9	85.8	80.0	83.6	80.6	**81.7**

TABLE VI
SELECTED PARAMETER VALUES FOR THE NETWORK DURING TRAINING

Subnetwork	Parameter	Setting
GPAN	Loss Function	Triplet & Cross-Entropy
	Learning Rate	0.0001
	Hidden_dim	256
	Optimizer	Adam
	Margin Value	0.2
	Batch Size (CASIA-B)	(8, 8)
	Batch Size (OUMVLP)	(8, 16)
	Number of Epochs (CASIA-B)	130000
	Number of Epochs (OUMVLP)	250000

GaitSet, and GaitPart. It can be clearly seen that GPAN achieves better recognition performance than these methods in cross-view walking conditions. It is worth noting that we believe that GPAN will achieve higher recognition accuracy if the same batchsize of the state-of-the-art method is used without considering the GPU memory limitation.

C. Ablation Study

In this section, we perform ablation experiments for the proposed method on CASIA-B dataset, and the details of the parameters during the experiments refer to the previous experimental settings. Here, we present a detailed analysis of the roles and contributions for each component in the framework, where W/ denotes adopted and W/O denotes not.

Effectiveness of PyConvs: The proposed PyConvs, shown in Fig. 2, contains a pyramidal structure with three layers of different types of kernels. With the proposed PyConvs, the aim is to process the input using kernels with different perceptual scopes without increasing the computational cost or model complexity, thus providing rich internal features for the deeper layers of the network. In each layer of PyConvs, the kernels contain different sizes, and the kernel size decreases from the bottom to the top of the pyramid. We have conducted ablation experiments on the GPAN structure to investigate the effect of regular convolution (3×3) and pyramidal convolutions combination on network recognition ability. As shown in Fig. 9, the recognition accuracies of all walking conditions improved by PyConvs. The performance improvement is ∼0.5% in NM, ∼1.1% in BG, and ∼3.1% in CL. Experimental results fully validate our hypothesis that pyramidal convolutions combination can provide richer interior features.

Effectiveness of HS block: Some researches [4] on gait recognition have shown that partial regions of the human body contribute differently in the final recognition. Therefore, learning the relationship between these partial features can increase the robustness of the model against viewpoint changes, thus improving the whole performance of the gait recognition. The division of silhouettes in gait methods includes patches,

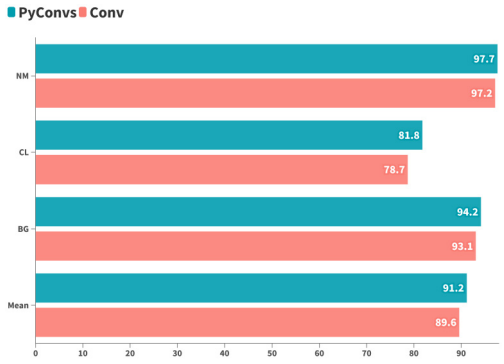


Fig. 9. Comparison of PyConvs and ordinary Conv in GPAN on CASIA-B, the average recognition rates of the probe data are NM, BG, and CL, respectively.

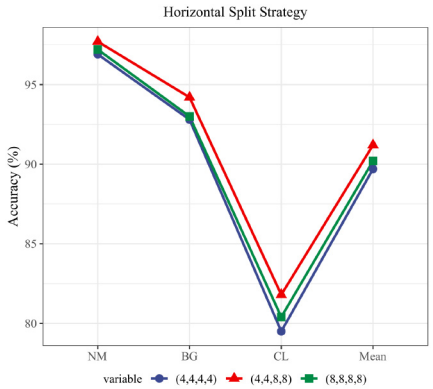
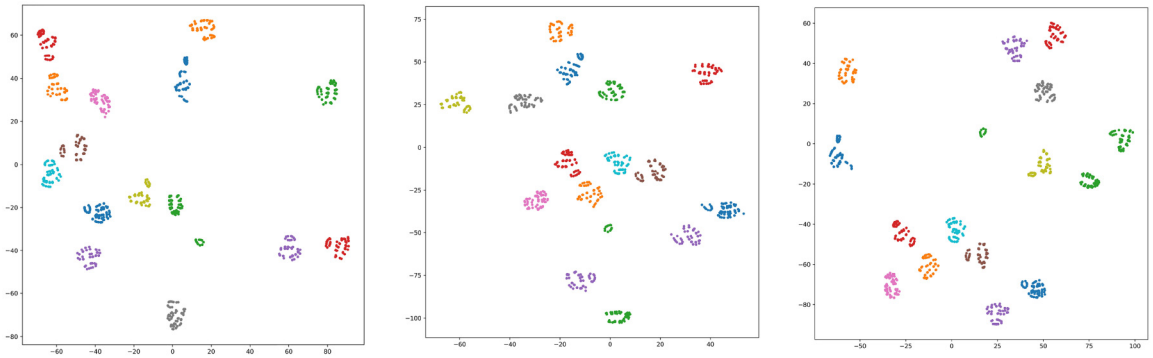


Fig. 10. Comparison results of HS blocks using different split strategies.

body components, and vertical/horizontal strips. We propose to use HS blocks to enhance the fine-grained learning of partial spatial features. Specifically, as shown in Fig. 1, we equally separate the four HS blocks into two groups. Fig. 10 shows the three different split strategies about HS. When HS block employs the split pattern of (4, 4, 4, 4), GPAN can achieve recognition rates of NM-96.9%, BG-92.8%, and CL-79.5%. In contrast, the performance of NM-97.7%, BG-94.2%, and CL-81.8% is obtained when (4, 4, 8, 8) is employed. Meanwhile, (8, 8, 8, 8) splits resulted in NM-97.2%, BG-93%, and CL-80.4%. Moreover, we report the average recognition rates of the three splits for various walking conditions, which intuitively shows that when (4, 4, 8, 8) has been employed, the average recognition performance is significantly higher than the other two modalities by ∼1% or more (89.7% vs. 91.2% vs. 90.2%), thus verifying that this strategy performs a better data fitting ability to the model.

Effectiveness of LPA Block: Local pyramidal attention integrates multi-scale spatial information and cross-channel



A. GPAN B. GPAN-L C. GPAN-P

Fig. 11. The t-SNE visualization of gait pyramid features f_{GPAN}, f_{GPAN-L}, f_{GPAN-P} on the test set of CASIA-B. We selected 15 subjects as the sample for the visualization experiment, where each subject is related to NM, BG and CL conditions. Each point represents a separate frame whose color is the subject ID.

TABLE VII
EFFECT OF USING LOCAL PYRAMIDAL ATTENTION MECHANISM FOR NETWORK PERFORMANCE ON CASIA-B

Walking Condition	LPA		
	W/	W/O	GAP
NM	97.4	97.0	0.4
BG	93.5	92.8	0.7
CL	80.4	79.4	1.0
Accuracy(mean)	**90.5**	89.7	0.8

TABLE VIII
EFFECT OF USING GLOBAL ATTENTION FUSION LEARNING MECHANISM FOR NETWORK PERFORMANCE ON CASIA-B

Walking Condition	GAFL		
	W/	W/O	GAP
NM	97.7	97.0	0.7
BG	93.8	92.8	1.0
CL	81.6	79.4	2.2
Accuracy(mean)	**91.0**	89.7	1.3

attention to the feature extraction phase of the model. Therefore, LPA block can extract multi-scale spatial information at a fine-grained level and favor long-range channel dependencies. This approach can obtain local information from human gait through a larger perceptual field and effectively address the detrimental effects of carrying conditions. In Table VII, we perform ablation experiments on the LPA block to explore the impact on the recognition performance. The experiments show that recognition performance improves for all three walking conditions when LPA is used, with CL improved by ∼1%. Evidently, LPA block can have a positive effect on network performance.

Effectiveness of GAFL Block: Global Attention Fusion learning treats the features from each layer as the responses to specific classes and correlates the responses to each other from different layers. By obtaining the dependencies between features of different depths, the network assigns corresponding attention weights to features of different depths, and improving the feature extraction ability. Thus, we propose to use the GAFL block to learn the relationships between features of different depths. Table VIII shows the ablation experiments of the GAFL block. The results show that the recognition ability of the model is qualitatively improved when using the GAFL block, i.e., ∼0.7% on NM, ∼1% on BG, and ∼2.2% on CL, respectively. It indicates that GAFL indeed promotes information interaction between shallow and deep layers by learning the relationships between features of different depths.

D. Feature Visualization

The results of the visualization of the features extracted by GPAN are further offered in Fig 11. The experiments are performed on CASIA-B training set consisting of 73 subjects. There are 110 sequences for each subject (11 views and 10 variants of the walk condition), and these sequences are used for visualization. Following the results of Fig 11, we found that all three structures have stronger feature aggregation ability and feature discriminative power in the feature extraction process. Secondly, to future examine the segmented blocks in PSF, we visualized the extracted features. As shown in Fig. 12, the middle part shows the input silhouette image, the left part refers to the global silhouette features learned from the initial layer in PSF, and the right part indicates the local silhouette features learned from the horizontal segmentation block in PSF. It can be roughly seen that the global silhouette features reflect a salient and complete appearance representation of the silhouette. In contrast, the local silhouette features focus on the representation of different body parts, thus ensuring that the network learns more fine-grained feature information.

Fig. 12. Visualization of Features Obtained From the Initial Convolution and Horizontal Segmentation Blocks in PSF.

TABLE IX
MORE PERFORMANCE COMPARISONS FOR THE BASELINE. THE ACCURACY OF REPORTED RESULTS IS RANK-1 (REP. MEANS REPRESENTATION). THE EXPERIMENTS ARE PERFORMED ON CASIA-B DATASET

Method	Venue	Year	Input	Feature Rep.	Temporal Rep.	NM	BG	CL
PoseGait [18]	PR	2020	Skeleton	Partial	Sequence Volume	68.7	44.5	36.0
ICD Net [60]	CVPR	2020	Silhouttes	Global	Tmp: GEI	94.5	-	-
ACL [27]	IEEE T-IP	2020	Silhouttes	Partial	Sequence Volume	96.0	-	-
MSPRT Net [61]	ICPR	2021	Silhouttes	Partial	Tmp: Set Pooling	95.7	90.7	72.4
PartialRNN [2]	IEEE T-Biom	2021	Silhouttes	Partial	Tmp: GCEM	95.2	89.7	74.7
GPAN(ours)	-	-	Silhouttes	Partial	Tmp: Hor. Pooling	**97.7**	**94.2**	**81.8**

E. Discussion

In Table IX, we report the results of performance comparisons on CASIA-B dataset with some other state-of-the-art baselines, including methods employing silhouettes, skeletons, etc. PoseGait uses skeletons for input, and its motion patterns and body structures are crafted using handcraft. However, handcrafted features may be missing some specific or fine-grained features. For example, it can learn short-range motion information by computing the difference between frame i_{th} and frame $(i+1)_{th}$ but fails to learn long-range motion features. In the silhouette-based approach, ICD Net proposes decoupled representation learning that combines identity and covariate features, but it still has limitations in decomposing identity and non-identity components to learn more differentiated gait representations. ACL extracts local gait features by using multiple individual 2D CNNs and reports slightly worse results, but the parallelization problem posed by LSTM cannot be ignored. MSPRT Net proposes to use RNN to learn local features and capsule network to extract the relationship between parts and the whole. Although this method also performs feature learning from local and global perspectives, it is fundamentally different from our proposed method and has lower performance since it fails to consider the fine-grained difference representation of human body parts. PartialRNN learns the relationship between partial features extracted from a sequence. However, its recognition performance is much inferior to our proposed, which benefits from the fact that GPAN considers the differences between human parts and focuses on valuable human regions from both global and local perspectives in a more targeted way. Fig. 13 visually compares the results of various baselines on CASIA-B dataset. In general, GPAN achieves better performance than other state-of-the-art baselines in the three walking cases. In addition, we also note that some excellent works [62], [63], [64], [65] in human activity recognition use the silhouette extraction technique. For example, Vishwakarma et al. perform human activity recognition by analyzing the effect of gradient space distribution on silhouette images.

F. Time Cost

In Table X, we show the time cost of the proposed three variants of the network. The entire validation process has been performed on the testing set of CASIA-B, which contains 50 subjects, each comprising 110 video sequences. In comparison, the GPAN complete structure achieves a real-time recognition speed of roughly 17 seqs/s, and the three structures are roughly consistent in terms of average efficiency.

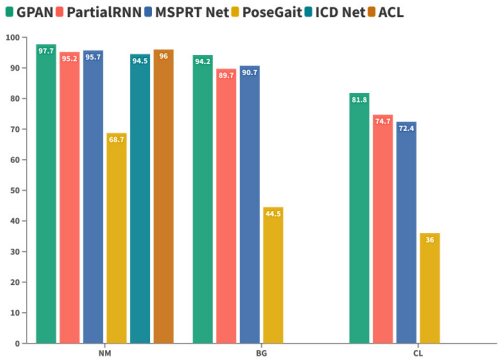


Fig. 13. Comparison with other state-of-the-art gait baselines. These baselines include multiple input forms, distinct temporal processing, and different feature representations.

TABLE X
THE TIME COST OF THE EXECUTION FOR THE PROPOSED NETWORK ON CASIA-B TESTING SET

Probe	Subnetworks		
	GPAN	GPAN-L	GPAN-P
Video Sequences	5500	5500	5500
Total Time	5min27s	4min58s	4min48s
Average Cost	17seqs/s	18seq/s	19seq/s

V. CONCLUSION

This paper proposes a novel Gait Pyramid Attention Network (GPAN) that learns discriminative and fine-grained representations from silhouettes for gait recognition. Mainly, we use a pyramid structure to learn multi-scale feature representations and then use horizontal split blocks to learn the differentiated representations of various body parts. Moreover, two different attention mechanisms are proposed to motivate the network to focus more purposefully on the regions of interest. By integrating different attentions, GPAN constructs three different variants. Significantly, GPAN enhances gait characterization by aggregating features between different layers and fine-grained local features to obtain the best recognition performance. Extensive experiments on CASIA-B and OUMVLP show that GPAN can bring consistent performance benefits in all walking conditions.

REFERENCES

[1] S. Hou, X. Liu, C. Cao, and Y. Huang, "Set residual network for silhouette-based gait recognition," *IEEE Trans. Biom., Behav., Ident. Sci.*, vol. 3, no. 3, pp. 384–393, Jul. 2021.

[2] A. Sepas-Moghaddam and A. Etemad, "View-invariant gait recognition with attentive recurrent learning of partial representations," *IEEE Trans. Biom., Behav., Ident. Sci.*, vol. 3, no. 1, pp. 124–137, Jan. 2021.

[3] Z. Zhang et al., "Gait recognition via disentangled representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4710–4719.

[4] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "GaitSet: Cross-view gait recognition through utilizing gait as a deep set," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3467–3478, Jul. 2022.

[5] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017.

[6] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, pp. 102–113, 2019.

[7] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," *Eng. Appl. Artif. Intell.*, vol. 77, pp. 21–45, Jan. 2019.

[8] C. Dhiman and D. K. Vishwakarma, "A robust framework for abnormal human action recognition using *R*-transform and Zernike moments in depth videos," *IEEE Sensors J.*, vol. 19, no. 13, pp. 5195–5203, Jul. 2019.

[9] D. K. Vishwakarma, R. Kapoor, R. Maheshwari, V. Kapoor, and S. Raman, "Recognition of abnormal human activity using the changes in orientation of silhouette in key frames," in *Proc. 2nd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, 2015, pp. 336–341.

[10] H. Aggarwal and D. K. Vishwakarma, "Covariate conscious approach for gait recognition based upon Zernike moment invariants," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 2, pp. 397–407, Jun. 2018.

[11] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations," in *Proc. Chin. Conf. Biom. Recognit.*, 2017, pp. 474–483.

[12] C. Fan et al., "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14225–14233.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[14] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, 2006, pp. 441–444.

[15] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–14, 2018.

[16] M. J. Nordin and A. Saadoon, "A survey of gait recognition based on skeleton model for human identification," *Res. J. Appl. Sci. Eng. Technol.*, vol. 12, no. 7, pp. 756–763, 2016.

[17] D. Zhang and M. Shah, "Human pose estimation in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2012–2020.

[18] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107069.

[19] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 3–20.

[20] W. An et al., "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Trans. Biom., Behav., Ident. Sci.*, vol. 2, no. 4, pp. 421–430, Oct. 2020.

[21] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[22] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, "Chrono-gait image: A novel temporal template for gait recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 257–270.

[23] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes," *Pattern Recognit. Lett.*, vol. 30, no. 11, pp. 977–984, 2009.

[24] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. Int. Conf. Imag. Crime Detection Prevent.*, 2009, pp. 1–6.

[25] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3D convolutional neural network," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3054–3062.

[26] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2016, pp. 4165–4169.

[27] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Trans. Image Process.*, vol. 29, pp. 1001–1015, 2020.

[28] N. Li, X. Zhao, and C. Ma, "JointsGait:A model-based gait recognition method based on gait graph convolutional networks and joints relationship pyramid mapping," 2020, *arXiv:2005.08625*.

[29] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5115–5124.

[30] C. Dhiman and D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics," *IEEE Trans. Image Process.*, vol. 29, pp. 3835–3844, 2020.

[31] T. Singh and D. K. Vishwakarma, "A deeply coupled ConvNet for human activity recognition using dynamic and RGB images," *Neural Comput. Appl.*, vol. 33, no. 1, pp. 469–485, 2021.

[32] D. K. Vishwakarma, "A two-fold transformation model for human action recognition using decisive pose," *Cogn. Syst. Res.*, vol. 61, pp. 1–13, Jun. 2020.

[33] Y. Fu et al., "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8295–8302.

[34] T. T. Verlekar, P. L. Correia, and L. D. Soares, "View-invariant gait recognition system using a gait energy image decomposition method," *IET Biom.*, vol. 6, no. 4, pp. 299–306, 2017.

[35] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10781–10790.

[36] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7036–7045.

[37] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 591–600.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[39] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[40] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3024–3033.

[41] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[42] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "a^2-nets: Double attention networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2018, pp. 352–361.

[43] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[44] H. Zhang et al., "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*.

[45] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 783–792.

[46] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1971–1980.

[47] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[48] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–37.

[49] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[50] Z. Zhong et al., "Squeeze-and-attention networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13065–13074.

[51] H. Sang, Q. Zhou, and Y. Zhao, "PCANet: Pyramid convolutional attention network for semantic segmentation," *Image Vis. Comput.*, vol. 103, Nov. 2020, Art. no. 103997.

[52] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep roots: Improving CNN efficiency with hierarchical filter groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1231–1240.

[53] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.
[54] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.
[55] H. B. Fredj, S. Bouguezzi, and C. Souani, "Face recognition in unconstrained environment with CNN," *Vis. Comput.*, vol. 37, no. 2, pp. 217–226, 2021.
[56] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On learning disentangled representations for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 345–360, Jan. 2022.
[57] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 382–398.
[58] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14648–14656.
[59] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. Int. Conf. Biometrics (ICB)*, 2016, pp. 1–8.
[60] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13309–13319.
[61] A. Sepas-Moghaddam, S. Ghorbani, N. F. Troje, and A. Etemad, "Gait recognition using multi-scale partial representation transformation with capsules," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 8045–8052.
[62] D. K. Vishwakarma and K. Singh, "Human activity recognition based on spatial distribution of gradients at sublevels of average energy silhouette images," *IEEE Trans. Cogn. Devel. Syst.*, vol. 9, no. 4, pp. 316–327, Dec. 2017.
[63] D. K. Vishwakarma, R. Kapoor, and A. Dhiman, "A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics," *Robot. Auton. Syst.*, vol. 77, pp. 25–38, Mar. 2016.
[64] D. K. Vishwakarma, R. Kapoor, and A. Dhiman, "Unified framework for human activity recognition: An approach using spatial edge distribution and *R*-transform," *AEU Int. J. Electron. Commun.*, vol. 70, no. 3, pp. 341–353, 2016.
[65] D. K. Vishwakarma and R. Kapoor, "Integrated approach for human action recognition using edge spatial distribution, direction pixel and *R*-transform," *Adv. Robot.*, vol. 29, no. 23, pp. 1553–1562, 2015.

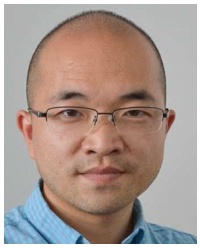
Jianyu Chen received the B.S. degree from the College of Information Engineering, Inner Mongolia University of Technology, China, in 2016, and the M.S. degree from the College of Computer Science and Information Technology, Northeast Normal University, in 2020. He is currently pursuing the Ph.D. degree with the National Engineering Research Center for Multimedia Software, Wuhan University. His research interests include gait recognition, video action recognition, and computer vision.

Zhongyuan Wang (Member, IEEE) received the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 2008, where he is currently a Professor with the School of Computer Science. He has been directing four projects funded by the National Natural Science Foundation of China. He has authored or coauthored over 80 refereed journal and conference papers and has been granted more than 30 invention patents. His research interests include biometrics and computer vision.

Peng Yi (Member, IEEE) received the B.S. degree from the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Computer, Wuhan University, under the supervision of Prof. Z. Wang. His research interests include image/video processing and computer vision.

Kangli Zeng (Student Member, IEEE) received the B.S. degree from Wuhan Donghu University, Wuhan, China, in 2016, and the M.S. degree from the School of Computer Science and Engineering, Wuhan Institute of Technology, in 2019. He is currently pursuing the Ph.D. degree with the National Engineering Research Center for Multimedia Software, Wuhan University. His research field is image/video processing and computer vision.

Zheng He received the Ph.D. degree in computer science from Hiroshima University, Hiroshima, Japan, in 2007. He is currently a Lecturer with the School of Computer, Wuhan University, Wuhan, China. His research interests include data mining, image processing, and multimedia communications.



Qin Zou received the B.E. degree in information engineering and the Ph.D. degree in computer vision from Wuhan University, Wuhan, China, in 2004 and 2012, respectively. From 2010 to 2011, he was a visiting Ph.D. student with the Computer Vision Laboratory, University of South Carolina, Columbia, SC, USA. He is currently an Associate Professor with the School of Computer Science, Wuhan University. His research interests include computer vision, pattern recognition, and machine learning. He was a co-recipient of the National Technology Invention Award of China in 2015.