# Enhanced Spatial-Temporal Salience for Cross-view Gait Recognition

Tianhuan Huang, Xianye Ben, *Senior Member, IEEE,* Chen Gong, *Member, IEEE,*

Baochang Zhang, *Senior Member, IEEE,* Rui Yan, and Qiang Wu, *Senior Member, IEEE*

***Abstract*—Gait recognition can be used in person identification and re-identification by itself or in conjunction with other biometrics. Although gait has both spatial and temporal attributes, and it has been observed that decoupling spatial feature and temporal feature can better exploit the gait feature on the fine-grained level. However, the spatial-temporal correlations of gait video signals are also lost in the decoupling process. Direct 3D convolution approaches can retain such correlations, but they also introduce unnecessary interferences. Instead of common 3D convolution solutions, this paper proposes an integration of decoupling process into a 3D convolution framework for cross-view gait recognition. In particular, a novel block consisting of a Parallel-insight Convolution layer integrated with a Spatial-Temporal Dual-Attention (STDA) unit is proposed as the basic block for global spatial-temporal information extraction. Under the guidance of the STDA unit, this block can well integrate spatial-temporal information extracted by two decoupled models and at the same time retain the spatial-temporal correlations. In addition, a Multi-Scale Salient Feature Extractor is proposed to further exploit the fine-grained features through context awareness extension of part-based features and adaptively aggregating the spatial features. Extensive experiments on three popular gait datasets, namely CASIA-B, OULP and OUMVLP, demonstrate that the proposed method outperforms state-of-the-art methods.**

***Index Terms*—Gait Recognition, Cross View, Spatial-Temporal Enhance, Multi-Scale Salient Feature Extraction.**

## I. Introduction

COMPARED with traditional biometrics such as fingerprint, iris and face, gait is hard to disguise and has the advantages of being identified under distant, non-cooperative or low-resolution scenarios. These characters make it widely used in many important applications, e.g., crime investigation, forensic identification, and security systems.

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010186, the Natural Science Foundation of China under Grant 61971468, and in part by the Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) under Grant 2019JZZY010119. (Corresponding author: Xianye Ben.)

T. Huang, X. Ben are with the School of Information Science and Engineering, Shandong University, Qingdao 266237, China (e-mail: huangtianhuan@mail.sdu.edu.cn; benxianye@gmail.com).

C. Gong is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information, School of Computer Science and Engineering, Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: chen.gong@njust.edu.cn).

B. Zhang is with Institute of Artificial Intellgence, Beihang University, Beijing 100191, China (e-mail: bczhang@buaa.edu.cn).

R. Yan is with Microsoft Corp, Bing, Bellevue, WA 98004, USA (e-mail: raymondino.yan@gmail.com).

Q. Wu is with the School of Electrical and Data Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: qiang.wu@uts.edu.au).

However, gait recognition, especially cross-view gait recognition that seeks to identify gaits in unknown views, is very challenging due to external interferences such as clothing, carrying conditions and camera view changes [1, 2].

There are several potential ways to tackle this problem and many advanced methods have been proposed for cross-view gait recognition. In general, existing works fall into three categories: i) model-based approaches, ii) template-based approaches, and iii) sequence-based approaches. The methods in the first category focus on reconstructing the 3D model of a pedestrian [3–7] so that the motion information under any view can be obtained in theory and an acceptable performance can be achieved against view variations. However, these solutions are vulnerable to the accuracy of pose estimation and the quality of silhouette sequences. The methods in the second category often obtain silhouettes using a gait cycle detection method [8], and aggregate them into a template, e.g., a gait energy image (GEI) [9–12]. These methods then optimize the intra-individual distance by learning common subspace projections across different views [13–17], building deep view transformation models [18] or using metric learning loss functions [1, 19–21] without modeling the gait cycle. Compared with model-based methods, template-based methods are simpler and better in feature representation. However, they ignore the temporal information contained in gait sequences, which limits the models' performance, especially when the view of input query gait is quite different from that of the gallery gaits. The methods in the third category take the original gait silhouette sequences as input and directly extract features from the sequences to retain necessary temporal information. Two widely used approaches are 1) decoupled modeling [2, 22–27], and 2) 3D convolution [1, 28, 29], which pick up features along the spatial and temporal domains simultaneously. Although the cross-view problem is not addressed explicitly by this category of approaches, effective spatial-temporal feature extraction shows encouraging accuracy on cross-view recognition tasks. Nonetheless, both 1) and 2) have limitations. Decoupled modeling usually uses different processes to extract spatial and temporal features, which overlooks the explicit correlation information across spatial and temporal domains. Even though direct 3D convolution can solve the correlation problem between the two domains, it will inevitably introduce external interference from spatial-temporal changes. Such changes could be caused by variations in clothing or carrying bags when pedestrians are walking.

In this paper, we propose a novel 3D CNN-based framework, called Enhanced Spatial-Temporal Salience Extraction

Network (ESNet). This model can fully exploit the integration effectiveness of individual decoupled spatial feature, temporal feature, and the common joint spatial-temporal information through 3D convolution pipelines, so that better spatial-temporal gait representations can be achieved and the model's performance under various walking conditions can be improved. As shown in Fig. 1, ESNet consists of initial layers and two well-designed components, i.e., the Dual-Attention Guided Feature Extractor (DAGFE) and Multi-Scale Salient Feature Extractor (MSSFE). Specifically, DAGFE is a special stacked CNN which has multiple specially designed blocks and takes shallow spatial-temporal features as input. Each block in DAGFE can extract information for both joint spatial-temporal and individual spatial/temporal domain in parallel, and then all the information is combined in an intuitive but effective way. Thus, we can extract higher-quality global spatial-temporal features without losing the correlations between two domains. Subsequently, MSSFE is proposed for fine-grained feature mining from the global spatial-temporal features, which can further improve the feature representation capability of the ESNet. More specifically, we make the following three major contributions:

- In DAGFE, we propose a novel spatial-temporal extraction layer, called the Parallel-insight Convolution (Pi-Conv) layer, to realize the synergy of direct 3D convolution and decoupled modeling. The core idea is to enable parallel convolutions to perceive and extract different domains' information.
- In DAGFE, we design a simple yet effective attention unit, namely the STDA unit. It is the first specifically designed attention method for silhouette sequence-based gait recognition. Using an attention approach, the STDA unit can adaptively adjust the output of the Pi-Conv layer and achieve a better integration of direct 3D convolution and decoupled modeling.
- We propose MSSFE to acquire salient and compact fine-grained features further. Instead of performing horizontal slicing on the global spatial-temporal feature directly to get part-based features, MSSFE expands the context-aware scope of each part to capture the relationship between adjacent parts and aggregate the spatial feature adaptively, which more efficiently facilitates the robust part-based feature learning.

The rest of this paper is organized as follows. Section II briefly introduces the related work. Section III explains the proposed ESNet in detail. In Section IV, the training and testing phases with the ESNet are introduced. Meanwhile, the experimental validation and performance evaluation of the ESNet are presented. Section V concludes the entire paper.

## II. Related Work

In this section, we discuss related works on (1) sequence-based gait recognition by taking decoupled modeling of temporal and spatial domains and simultaneous extraction of spatial-temporal information using 3D convolution into account, (2) the 3D CNN-based framework, and also (3) the attention mechanism, wherein the latter two inspired us to propose the ESNet.

### A. Sequence-based Gait Recognition

Sequence-based gait recognition methods [1, 2, 22–29] take the original gait sequences as input, and this paper also belongs to it. Recently, sequence-based spatial-temporal feature extraction is popular in gait recognition [30–32]. Sequence-based gait recognition mainly refers to the decoupled modeling and simultaneous modeling with 3D convolution.

The decoupled modeling of spatial and temporal information often first extracts frame-level features and then applies temporal models to encode the information along the time axis. For instance, after extracting the spatial feature of each frame, Zhang et al. [2] used an LSTM-based attention network to fuse the features in the time dimension, Fan et al. [23] proposed a Micro-motion Capture Module for local short-range time modeling, and Li et al. [27] designed the Residual Frame Attention Mechanism to acquire and highlight critical frames of sequences and then aggregated temporal features using max aggregation. Similarly, Chao et al. [22], Han et al. [26] and Qin et al. [24] used a statistical function for temporal modeling after frame-level feature extraction. And Zhang et al. [33] extracted temporal and spatial features by two separate processes through disentangled representation learning. Although these methods have exhibited encouraging performance, the decoupled modeling strategy ignores the synergy of spatial-temporal information and loses the correlations between the spatial and temporal domains.

Using 3D convolution is the other spatial-temporal feature extraction method. Wolf et al. [28] proposed a 3D convolution network for gait recognition. Lin et al. [29] proposed a Multiple-Temporal-Scale 3D (MT3D) network to extract spatial-temporal information on multiple temporal scales. Since the 3D convolution can extract both the spatial and temporal information simultaneously and preserve the correlations between the two domains, [28] and MT3D achieved performance improvements. However, common 3D convolution will inevitably introduce external spatial-temporal interference information, e.g., spatial-temporal changes caused by variations in clothing or carrying.

Different from previous works which decouple spatial-temporal feature extraction into two processes [2, 22–27], or use direct 3D convolution [1, 28, 29], the proposed method aims to integrate individual decoupled spatial feature, temporal feature, and the common joint spatial-temporal information through 3D convolution pipelines.

### B. 3D CNN-based Framework

The 3D CNN-based framework has the spatial-temporal modeling capability, which can effectively enhance the model performance for video-based tasks [34, 35]. The proposal of C3D [36] and its success in action recognition attracted the attention of researchers and promoted the application of 3D CNN-based frameworks in various computer vision tasks [35, 37, 38]. However, a 3D CNN-based framework built with standard single 3D convolution may not achieve superior performance. In response, Carreira et al. [39] proposed an I3D network with a two-stream architecture, which achieves

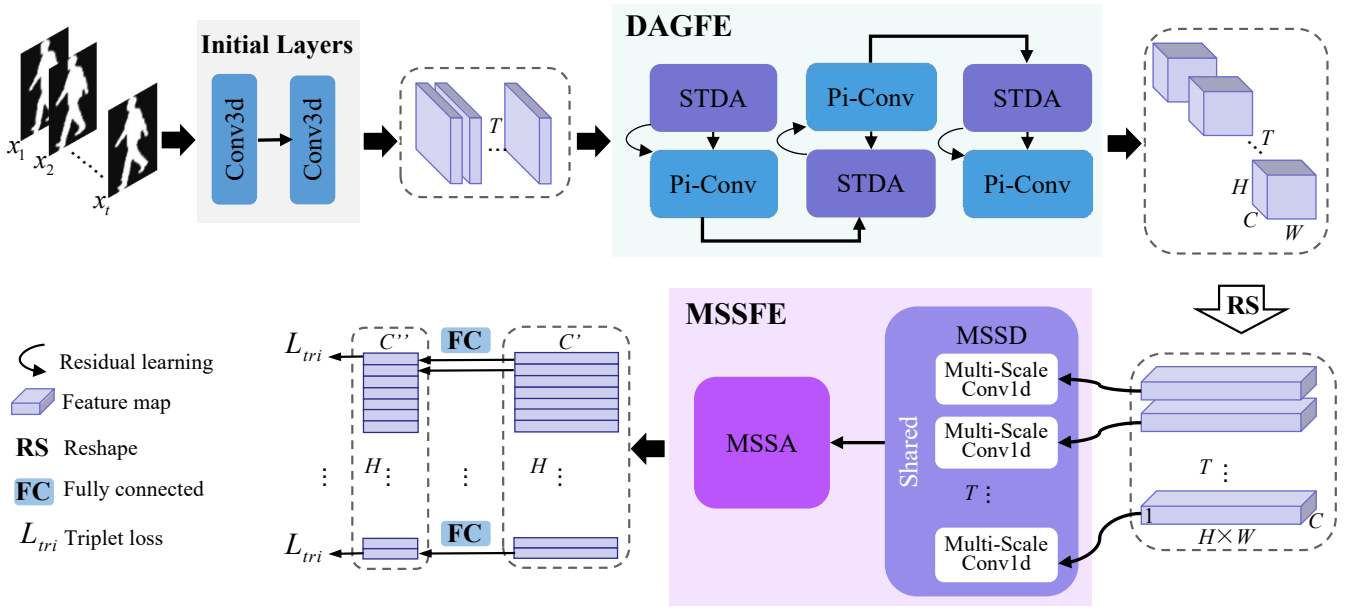



Fig. 1. The framework of the Enhanced Spatial-Temporal Salience Extraction Network. Our framework consists of initial layers and two well-designed components, i.e., the Dual-Attention Guided Feature Extractor (DAGFE) and Multi-Scale Salient Feature Extractor (MSSFE). 1) Initial layers are introduced to obtain shallow spatial-temporal features. 2) DAGFE is designed for global high-quality spatial-temporal extraction, which contains three tandem residual blocks, and each block is composed of an STDA unit and a Pi-Conv layer. In particular, there is a pooling layer after the first block, which is omitted in this figure. 3) MSSFE is utilized to acquire salient fine-grained gait features. Multi-Scale Salience Descriptor (MSSD) and Multi-Scale Salience Aggregator (MSSA) are two components in the MSSFE, where MSSD contains multiple parameter-shared Multi-Scale Conv1ds. The final gait representation is the output after separate fully connected mapping of the fine-grained gait features.

more significant performance improvement than C3D by adjusting the network structure and increasing the network width. Unfortunately, although I3D has exhibited exciting results on several benchmark datasets, it contains massive parameters. In this case, problems such as huge resource consumption and convergence difficulties in the training phase follow.

Many attempts have been made to introduce sparsity into the network by factorizing the 3D convolution, thereby reducing parameter redundancy. For example, through decomposing the 3D parameter matrix into 1D and 2D parameter matrix, Tran et al. [40] proposed a new 3D convolution block named R(2+1)D to improve model's performance. Qiu et al. [41] proposed a Pseudo-3D (P3D) network structure to reduce the parameters of the model and extract more robust features.

Such network construction methods of decomposing 3D parameter matrix into low-rank matrices also give us inspiration. Inspired by these two works, we first propose the Pi-Conv layer, which implements the integration of the direct 3D convolution and the decoupled modeling by controlling the scope of perceptual domains (temporal, spatial, or spatial-temporal) of 3D convolution kernels. The details of the Pi-Conv layer will be introduced in III-B1.

### C. Attention Mechanism and Beyond

Attention mechanism aims to focus on key features and suppressing unnecessary ones. For CNNs, spatial attention and channel attention are two main types of attention operation. Hu et al. [42] introduced a SENet architecture and proposed a squeeze-and-excitation block. SENet is a typical representative of channel attention, which can adaptively calibrate the channel-wise features by explicitly modeling the channel interdependencies. In addition to considering the importance of different channel pixels like SENet, Woo et al. [43] proposed a Convolutional Block Attention Module (CBAM) combining channel attention and spatial attention. Although SENet and CBAM have been proven useful for various computer vision tasks [44, 45], limited by the practical receptive field, neithor of them can effectively capture the large scope information.

Therefore, non-local/global attention exploration began. In [46], a non-local block is inserted before the encoder-decoder style attention module to enable attention learning based on globally refined features. In [47], researchers adopted the non-local mean idea which computes a weighted summation of the non-local pixels/features as the refined representation of the target pixels/features. The weight value connecting every two positions represents their relationship and is calculated from the similarity/correlation of the pair.

However, neither the non-local attention mechanism nor the local attention-based SENet and CBAM are suitable for silhouette sequence-based gait recognition, since its input are simple binary image sequences which are lack of color and texture information. Therefore, the similarity in each pair of pixel positions does not necessarily show the relevance but may introduce noise, especially for shallow gait feature maps. Similarly, the introduction of channel attention will interfere with the original feature extraction and lead to performance degradation. Based on the above facts, we propose a novel local attention structure, i.e., STDA unit. Unlike SENet and CBAM processing each image without considering critical information such as temporal properties for videos, the STDA

unit introduces temporal attention and replaces the original channel attention, in which case, the critical spatial and temporal information in input can be activated. The details of the STDA unit will be introduced in Sec. III-B2.

## III. The Proposed Method

In this section, we introduce the details of the proposed ESNet which aims to extract robust and discriminative fine-grained features from gait silhouettes. We first present the pipeline of the ESNet, followed by the Dual-Attention Guided Feature Extractor (DAGFE) and Multi-Scale Salient Feature Extractor (MSSFE), and also the related loss functions. The overall framework is illustrated in Fig. 1.

### A. Pipeline

As shown in Fig. 1, a gait silhouette sequence containing $t$ frames $\mathbf{X}_S = \{x_i \,|i = 1, 2, ..., t\}$ is fed into the ESNet, and shallow spatial-temporal features are first obtained through two initial layers. Second, the Dual-Attention Guided Feature Extractor (DAGFE), a specially designed 3D convolution network, is utilized to further extract high-quality spatial-temporal features $\mathbf{X}_F$:

$$\mathbf{X}_F = \mathrm{DAGFE}(\mathbf{X}_f), \tag{1}$$

where $\mathbf{X}_f$ denotes the shallow spatial-temporal features, and both $\mathbf{X}_f$ and $\mathbf{X}_F$ are five-dimensional tensors, i.e., the batch size, channels, frames, height and width. DAGFE consists of several well-designed blocks, and the details of it will be introduced in Sec. III-B. In this way, the global spatial-temporal representation of a gait silhouette sequence can be obtained.

In order to meet the input requirement of the subsequent module, the Reshape (RS) operation is used to flat the spatial feature of $\mathbf{X}_F$ and obtain the output $\mathbf{X}_F^{'}$. Third, the Multi-Scale Salient Feature Extractor (MSSFE), aiming at extracting salient local information and improve the discrimination of parted-based features, is executed over $\mathbf{X}_F^{'}$:

$$\mathbf{Sal}_F = \mathrm{MSSFE}(\mathbf{X}_F^{'}), \tag{2}$$

where $\mathbf{Sal}_F$ denotes the salient output of MSSFE. Actually, MSSFE contains two sequential manipulations for input $\mathbf{X}_F^{'}$:

$$\mathbf{Z}_F = \mathrm{MSSD}(\mathbf{X}_F^{'}), \tag{3}$$

$$\mathbf{Sal}_F = \mathrm{MSSA}(\mathbf{Z}_F), \tag{4}$$

where MSSD represents the Multi-Scale Salience Descriptor, which can improve the spatial context awareness of each part and efficiently extract salient part-based features. MSSA represents the Multi-Scale Salience Aggregator, which is used to adaptively aggregate the features acquired by MSSD and obtain compact feature representations. Finally, several separate FC layers are employed to map the feature vectors to metric space for the final individual identification.

### B. Dual-Attention Guided Feature Extractor

The Dual-Attention Guided Feature Extractor (DAGFE), aiming at extracting high-quality global spatial-temporal features, is composed of three tandem blocks. As shown in Fig. 2(a), each block consists of two main components, a Parallel-insight Convolution (Pi-Conv) layer and a Spatial-Temporal Dual-Attention (STDA) unit. In this part, the Pi-Conv layer and the STDA unit will be described in detail first and followed by the overall illustration of each block.

*1) **Pi-Conv layer**:*

**Definition.** The Pi-Conv layer is a novel spatial-temporal feature extraction layer based on 3D convolution, which contains three parallel 3D convolutions with different kernels. These three parallel 3D convolutions are executed on the input feature map, separately, and their outputs are added in an element-wise manner.

**Motivation.** In order to assemble the decoupled learning of spatial and temporal features and relation retainment through 3D convolution, the Pi-Conv layer is developed. As shown in Fig. 2(b), the insight domain of a neuron can be determined by setting the kernel size in the Pi-Conv layer. The 3D convolution with the kernel size of $k_1 \times k_1 \times k_1$ (left in Fig. 2(b)) is a regular 3D convolution operation, which can extract spatial-temporal information simultaneously. The 3D convolutions with the kernel sizes of $k_2 \times 1 \times 1$ (middle in Fig. 2(b)) and $1 \times k_3 \times k_3$ (right in Fig. 2(b)) can realize the separate feature extraction of the temporal and spatial domains, respectively. This parallel-insight convolution design ensures the synergy of spatial-temporal information while extracting high-quality spatial and temporal features, which makes it possible to give full play to the advantages of two sequence-based modeling methods.

**Operation.** For convenience, the input feature map of the Pi-Conv layer is expressed as $\mathbf{X} \in \mathbf{R}^{N\times C\times T\times H\times W}$. $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ denote the three kernels with the sizes of $k_1 \times k_1 \times k_1$, $k_2 \times 1 \times 1$ and $1 \times k_3 \times k_3$, respectively. The outputs of parallel-insight convolutions are added element-wisely as the final output of the Pi-Conv layer. As shown in Fig. 2(b), the output $\mathbf{Y}_{Pi-Conv} \in \mathbf{R}^{N\times C\times T\times H\times W}$ of the Pi-Conv layer can be presented as follows:

$$\mathbf{Y}_{Pi-Conv} = \mathbf{W}_1 * \mathbf{X} + \mathbf{W}_2 * \mathbf{X} + \mathbf{W}_3 * \mathbf{X}, \tag{5}$$

where $*$ denotes the convolution operation.

*2) **STDA unit**:*

**Definition.** The STDA unit is a simple yet effective attention module that is specially designed for silhouette sequence-based gait recognition. As shown in Fig. 2(c), the STDA unit consists of two simple attention branches and a few element-wise arithmetic operations.

**Motivation.** Effective integration of the decoupled (spatial and temporal) modeling and the direct 3D convolution can obtain more robust spatial-temporal features. To this end, the STDA unit is devised and injected into the head of each block ( Fig. 2(c) is injected into Fig. 2(a)). Considering the binary silhouette input of gait recognition, we discard the non-local attention mechanism that computes similarity in each pair of pixel positions or modeling the channel interdependencies

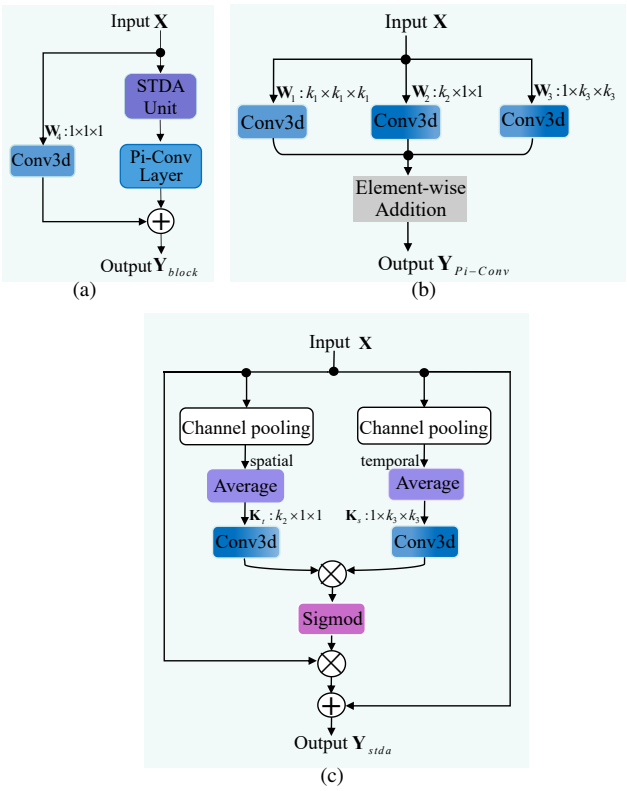



Fig. 2. (a): The overall illustration of the proposed block. (b): The detailed architecture of the Pi-Conv layer. (c): The detailed architecture of the STDA unit.

explicitly (e.g., CBAM [43] or SENet [42]) since the input is lack of color and texture. Instead, the STDA unit consists of two simple branches. The temporal attention branch (left branch in Fig. 2(c)) explores the correlations between temporal features, and the spatial attention branch (right branch in Fig. 2(c)) explores semantically robust features in the spatial domain. Then the critical spatial and temporal information in input can be activated by this parallel dual-attention branch design. By embedding the STDA unit in the head of the Pi-Conv layer, the STDA unit is expected to adaptively adjust the spatial-temporal information extracted by the Pi-Conv layer in such an attention manner, and achieve a better integration of decoupled spatial and temporal feature extraction and the direct 3D convolution.

**Operation.** The detailed architecture of the STDA unit is shown Fig. 2(c). To obtain the temporal attention weight, the input feature map $\mathbf{X} \in \mathbf{R}^{N\times C\times T\times H\times W}$ is first averaged across channels to get a global spatial-temporal tensor $\mathbf{F} \in \mathbf{R}^{N\times1\times T\times H\times W}$. Subsequently, the spatial-wise information of $\mathbf{F}$ is aggregated by average pooling operation to obtain $\mathbf{F}_t \in \mathbf{R}^{N\times1\times T\times1\times1}$, which is then fed to a 3D convolution layer $\mathbf{K}_t$ with kernel size $k_2\times1\times1$ corresponding to $\mathbf{W}_2$ in the Pi-Conv layer. Finally, the temporal attention score $\mathbf{S}_t \in \mathbf{R}^{N\times1\times T\times1\times1}$ can be formulated as

$$\mathbf{S}_t = \mathbf{K}_t * \mathbf{F}_t, \tag{6}$$

Similarly, to obtain the spatial attention weight, the input tensor is averaged on the channel axis as above to obtain $\mathbf{F}$ and the temporal dimension of $\mathbf{F}$ is squeezed by average pooling to obtain $\mathbf{F}_s \in \mathbf{R}^{N\times1\times1\times H\times W}$. Then a 3D convolution layer $\mathbf{K}_s$ with kernel size $1\times k_3\times k_3$ corresponding to $\mathbf{W}_3$ in the Pi-Conv layer is employed. The spatial attention score $\mathbf{S}_s \in \mathbf{R}^{N\times1\times1\times H\times W}$ can be formulated as follows:

$$\mathbf{S}_s = \mathbf{K}_s * \mathbf{F}_s, \tag{7}$$

In order to take full advantage of spatial-temporal attention, the attention scores from these two attention branches are multiplied in an element-wise manner, and is fed into a sigmoid activation function $\sigma$ to get the final spatial-temporal mask $\mathbf{M} \in \mathbf{R}^{N\times1\times T\times H\times W}$, which can be represented as:

$$\mathbf{M} = \sigma\left(\mathbf{S}_t \odot \mathbf{S}_s\right), \tag{8}$$

where $\odot$ denotes the element-wise multiplication.

Then, the final output $\mathbf{Y}_{stda} \in \mathbf{R}^{N\times C\times T\times H\times W}$ of the STDA unit can be interpreted as:

$$\mathbf{Y}_{stda} = \mathbf{X} + \mathbf{X} \odot \mathbf{M}, \tag{9}$$

*3) **Overall illustration of the proposed block**:*

After injecting the STDA unit, we formulate the proposed block so as to better illustrate its specific structure. The overall illustration of the proposed block is shown in Fig. 2(a), the input feature map is first sent to the STDA unit and then a Pi-Conv layer is performed over it, finally the residual learning mechanism is adopted.

Formally, let $\mathbf{X} \in \mathbf{R}^{N\times C\times T\times H\times W}$ denote the input of the proposed block, $\mathbf{Y}_{stda} \in \mathbf{R}^{N\times C\times T\times H\times W}$ and $\mathbf{Y}_{Pi-Conv} \in \mathbf{R}^{N\times C\times T\times H\times W}$ denote the output of the STDA unit and the Pi-Conv layer, respectively. To obtain $\mathbf{Y}_{block}$, we first feed $\mathbf{X}$ into the STDA unit to get $\mathbf{Y}_{stda}$. Subsequently, $\mathbf{Y}_{stda}$ is sent to three parallel-insight convolutions, respectively, and the obtained results are added by element-wise, which can be represented as:

$$\mathbf{Y}_{Pi-Conv} = \mathbf{W}_1 * \mathbf{Y}_{stda} + \mathbf{W}_2 * \mathbf{Y}_{stda} + \mathbf{W}_3 * \mathbf{Y}_{stda}, \tag{10}$$

Finally, residual learning is performed with the input:

$$\mathbf{Y}_{block} = \delta\left(\mathbf{Y}_{Pi-Conv} + \mathbf{W}_4 * \mathbf{X}\right), \tag{11}$$

where $\mathbf{W}_4$ is a 3D convolution layer with the kernel size of $1\times1\times1$, which is used to match the number of channels. The notation $\delta$ in Eq. (11) denotes the Leaky ReLU activation.

The above is the formulaic description of the proposed block. The exact structure of DAGFE is listed in Tab. I and the ablation study of the Pi-Conv layer and STDA unit will be discussed in Sec. IV-F.

### C. Multi-Scale Salient Feature Extractor

The Multi-Scale Salient Feature Extractor (MSSFE) is designed for further salient and discriminative part-based feature extraction. The detailed structure is shown in Fig. 3. For the output $\mathbf{X}_F \in \mathbf{R}^{N\times C\times T\times H\times W}$ of DAGFE, we hope to mine diverse and robust fine-grained features through different horizontal parts along the $H$ axis. As mentioned in Sec. III-A, the Reshape (RS) operation is firstly conducted to flat the spatial feature of $\mathbf{X}_F$ so as to meet the input requirement of MSSFE. The MSSFE can be decomposed into a Multi-Scale Salience Descriptor (MSSD) and a Multi-Scale Salience Aggregator (MSSA). Next, the MSSD and MSSA will be described in detail, respectively.

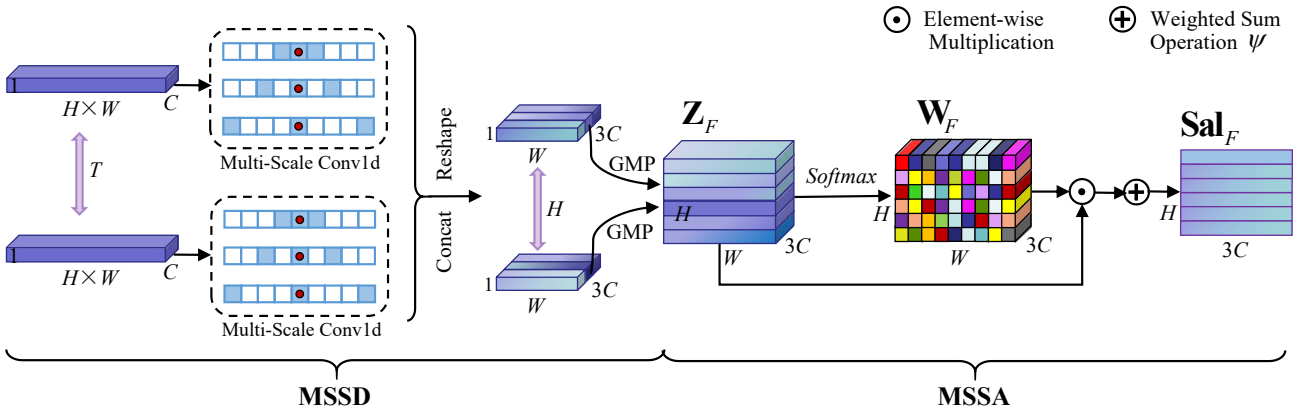



Fig. 3. The detailed architecture of the Multi-Scale Salient Feature Extractor (MSSFE). Taking $n = 3$ as an example and the dilation rates of these three parallel 1D convolutions are set to 1, 2, 4, respectively. For the convenience of description, the dimension $T$ is omitted before the Global Max Pooling (GMP) operation.

*1) **MSSD**:*

**Definition.** The MSSD is a new application of the dilation convolution which contains multiple multi-scale 1D convolutions, namely Multi-Scale Conv1d (as shown in Fig. 3), and each Multi-Scale Conv1d consists of multiple parallel 1D convolutions with different dilation rates. Let $n$ be the number of pre-defined parallel 1D convolutions and $d_1, d_2, ..., d_n$ denote the dilation rates for these parallel 1D convolutions, respectively. In particular, the Multi-Scale Conv1d is equivalent to an ordinary 1D convolution layer when $n = 1$.

**Motivation.** Each part of the human body has a dependent relationship with each other, especially between adjacent parts. In order to enhance the fine-grained learning of part-based spatial features and avoid losing the relationship between adjacent parts, MSSD is designed. As shown in Fig. 4, compared to the direct horizontal splicing of $\mathbf{X}_F$, the context-aware scope of each part is expanded with the dilation rate progressively increaseing, which makes it possible to capture the relationship between adjacent parts. Furthermore, parallel 1D convolutions with multiple different dilation rates enable each part to be aware of multi-scale contexts. By this means, more diverse and robust part-based feature representations can be obtained.

**Operation.** As shown in Fig. 3, the input feature map $\mathbf{X}_F^{'} \in R^{N\times C\times T\times(H\times W)}$ after reshaping is first sliced along the $T$ dimension and each slice is sent to a Multi-Scale Conv1d, separately. Then regular convolution operations are performed over each slice. Note that these Multi-Scale Conv1ds for each slice are parameter shared. After that, the outputs of each slice are concatenated along the channel. Then all these slices' outputs are combined as the whole output feature maps and reshaped reversely to recover the shape like $\mathbf{X}_F$. Thus, the multi-scale feature descriptor with the shape of $N\times nC\times T\times H\times W$ is generated. Subsequently, a Global Max Pooling (GMP) operation is applied on the feature descriptor to get the final multi-scale feature $\mathbf{Z}_F \in R^{N\times nC\times H\times W}$.

*2) **MSSA**:*

**Definition.** MSSA behaves like a multi-scale salience feature receptor that can perceive which part-based multi-scale features are discriminative and need to be retained. It performs salient feature selection as well as adaptive spatial feature aggregation for each part.

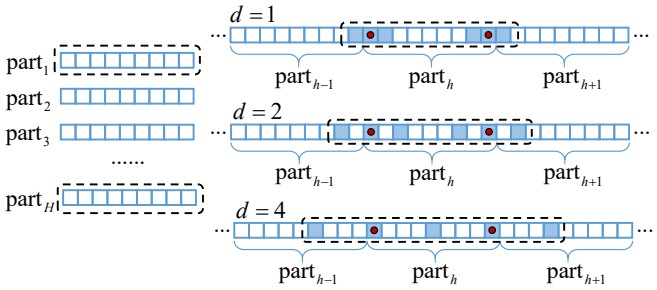



Fig. 4. An illustration of expanding the context-aware scope for each part by 1D convolution with different dilation rates.

**Operation.** MSSA is composed of a softmax activation, an element-wise multiplication $\odot$ and a weighted sum operation $\psi$, which is analogous to an attention mechanism. As shown in Fig. 3, the salience-sensitive weight tensor $\mathbf{W}_F \in R^{N\times nC\times H\times W}$ can be obtained after the softmax activation operated on the $W$ dimension of $\mathbf{Z}_F$. And then, the salience part-based feature $\mathbf{Sal}_F \in \mathbf{R}^{N\times nC\times H}$ is achieved and represented as follows:

$$Sal(\mathbf{X}) = \psi(\mathbf{W}_F \odot \mathbf{Z}_F), \tag{12}$$

$$\mathbf{W}_F = Softmax(\mathbf{Z}_F), \tag{13}$$

Compared with common statistical functions, e.g., max and mean, MSSA can integrate the spatial features of each part adaptively while preserving the salience of features.

## D. Loss Function

Before training the proposed gait recognition model, for each horizontal slice of $\mathbf{Sal}_F$ shown in Fig. 3, a fully connected operation is performed and the final feature $\mathbf{Y} \in R^{N\times C^{''}\times H}$ is obtained. Then inspired by the success of triplet loss in person re-identification task [48], we employ the batch all version of triplet loss and embed it into the proposed ESNet. In the training stage, a sample triplet consists of an anchor, a positive example (pos.), and a negative example (neg.). Specifically, the anchor and positive examples have the same identity label but are different from the negative example. To fully exploit fine-grained features, the triplet constraint is

separately imposed on each horizontal slice. The complete triplet loss is defined as:

$$L_{tri\_all} = \frac{1}{N_{tri}} \sum_{h=1}^{H} \overbrace{\sum_{u=1}^{U} \sum_{v=1}^{V}}^{anchor} \overbrace{\sum_{\substack{a=1 \\ a \neq v}}^{V}}^{\text{pos.}} \overbrace{\sum_{\substack{b=1 \\ b \neq u}}^{U} \sum_{c=1}^{V}}^{neg.} \max\left\{dist + m, 0\right\}, \tag{14}$$

where $N_{tri}$ is the number of triplets resulting in the non-zero loss terms; $(U, V)$ are the number of subjects and the number of sequences for each subject in a mini-batch, i.e., a total of $U \times V$ sequences are input into the model and constitute $UV\left(V-1\right)\left(UV-V\right)$ sample triplets; $H$ is the scale to slice the features horizontally, which is also the height of $\mathbf{X}_F$; and $m$ is the margin. In a sample triplet, each example has $H$ part-based features, we calculate the triplet loss for each corresponding feature triplet, i.e., $H$ triplet losses are calculated. Thus, the complexity of the triplet loss in Eq. (14) is $O\left(HUV\left(V-1\right)\left(UV-V\right)\right)$, which is $H$ times that of original batch all version of triplet loss [48]. Although the complexity is increased, the computational load is reduced to $1/H$ of the original. Furthermore, this sum of separate triplet loss functions for the horizontal slicing can facilitate the fully exploitation of fine-grained features effectively. The $dist$ in Eq. (14) can be formulated as follows:

$$dist = d_+\left(y_{u,v}^h, y_{u,a}^h\right) - d_-\left(y_{u,v}^h, y_{b,c}^h\right), \tag{15}$$

where $y_{u,v}^h$ denotes the $h$-th part feature in the $v$-th gait sequence of the $u$-th subject ( $y_{u,a}^h$ and $y_{b,c}^h$ are similar to $y_{u,v}^h$), and $d_+$ and $d_-$ measure the similarity between positive sample pairs and negative sample pairs, respectively, e.g., Euclidean distance.

## IV. Experiments

Three public datasets have been applied to evaluate the proposed method, namely CASIA-B [49], OULP [50] and OVMVLP [51]. In this section, datasets and implementation details will be described firstly. Then, the performance of the proposed method will be compared with that of other state-of-the-art methods. Finally, the detailed ablation studies will be conducted strictly on CASIA-B to verify the effectiveness of each component in the proposed method.

### A. Datasets

**CASIA-B** is a widely applied gait dataset with 124 subjects. Each subject has 110 sequences, and the sequences are collected under three conditions, i.e., normal (NM) (6 video groups per subject indexed as NM#01-06), walking with a bag (BG) (2 video groups per subjects indexed as BG#01-02) and wearing a coat or jacket (CL) (2 video groups per subject indexed as CL#01-02). Each group is simultaneously taken under 11 different views (0°-180°with interval 18°). Therefore, the gait dataset contains 124×(6+2+2)×11=13640 sequences in total. Under 90°, a subject's gait sequences under the NM, BG, and CL conditions are shown in Fig. 5. To evaluate the performance of the proposed method fairly, we strictly follow the popular protocol as [19] and [22]. In this paper, the first 74 subjects are used for training and the remaining 50 subjects are reserved for testing with no overlap. During the testing phase, the first 4 sequences of the NM condition (i.e., NM#01-04) are grouped into the gallery, and the rest sequences NM#05-06, BG#01-02, and CL#01-02 are used as the probe, respectively.

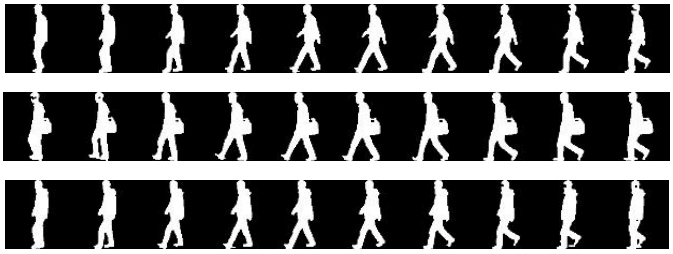


Fig. 5. Three gait sequences under 90° from the CASIA-B dataset. From top to bottom, these gait sequences are collected under the NM, BG and CL walking conditions, respectively.

**OULP** is a gait dataset with large population. It is made up of 4007 subjects including 2135 males and 1872 females ranging in age from 1 to 94 years and has four view angles (55°, 65°, 75°, 85°). Two sequences taken under normal walking condition are available for each subject, one for gallery and the other for the probe. Our experimental settings are consistent with [19]. Thus, a total of 3836 subjects are used in the subsequent experiments and the five-fold cross-validation is adopted. At the testing phase, the sequences with index #01 are used as the gallery while the rest sequences with index #02 are used as the probe.

**OUMVLP** is currently the largest public gait dataset available. It contains 10307 subjects with 14 views per subject (0°, 15°, ... , 90°; 180°, 195°, ... , 270°) and two sequences per view (indexed as #00-01). Consistent with its protocols, we take the sequences from 5,153 subjects for training and the sequences from the remaining 5,154 subjects for testing. During the testing phase, the sequences with index #01 for each subject are kept in the gallery and the rest sequences with index #00 are taken as the probe.

### B. Implementation Details

In all the experiments, the input gait silhouettes are first pre-processed using the method in [52], and then resized to 64×44. The Adam optimizer [53] with the momentum of 0.9 and the learning rate of 1e-4 is utilized for training the proposed ESNet. The margin in Eq. (14) is set to 0.2. **1)** On the **CASIA-B** and **OULP** datasets, the convolution channels of initial layers and the three stacked blocks in DAGFE are set to (32, 32, 64, 128, 128), respectively, and the dilation rates in MSSFE are set to (1, 2, 4). The exact structure of ESNet is listed in Tab. I. The number of subjects and the sequences for each subject in a mini-batch are set to (8, 8) for CASIA-B and (32, 4) for OULP. We train the model for 100K and 60K iterations on CASIA-B and OULP, respectively. **2)** On the **OUMVLP** dataset, since it contains almost 20 times more sequences than CASIA-B, two additional blocks with output channels set to 256 are stacked into the DAGFE. The batch size on OUMVLP is set to (32, 8), the iteration number is set to 250K, and the learning rate is decreased to 1e-5 after

TABLE I
THE STRUCTURE OF ESNET. (IN MSSFE, TAKING $n = 3$ AND THE DILATION RATES ARE SET TO 1, 2, 4, RESPECTIVELY.)

| Stage | | | Details | Output Size |
|---|---|---|---|---|
| Input | | | T×C×H×W | 30×1×64×44 |
| Initial Layers | Layer1 | | Conv3D [1, 3×3×3, 32], stride 1, 1, 1 | 30×32×64×44 |
| | Layer2 | | Conv3D [32, 3×1×1, 32], stride 3, 1, 1 | 10×32×64×44 |
| DAGFE | Block1 | STDA | Conv3D $\begin{bmatrix} 1,3\times1\times1,1 \\ 1,1\times3\times3,1 \end{bmatrix}$, stride 1, 1, 1 | 10×64×64×44 |
| | | Pi-Conv | Conv3D $\begin{bmatrix} 32,3\times3\times3,64 \\ 32,3\times1\times1,64 \\ 32,1\times3\times3,64 \end{bmatrix}$, stride 1, 1, 1 | |
| | MaxPool3D [1×2×2], stride 1, 2, 2 | | | 10×64×32×22 |
| | Block2 | STDA | Conv3D $\begin{bmatrix} 1,3\times1\times1,1 \\ 1,1\times3\times3,1 \end{bmatrix}$, stride 1, 1, 1 | 10×128×32×22 |
| | | Pi-Conv | Conv3D $\begin{bmatrix} 64,3\times3\times3,128 \\ 64,3\times1\times1,128 \\ 64,1\times3\times3,128 \end{bmatrix}$, stride 1, 1, 1 | |
| | Block3 | STDA | Conv3D $\begin{bmatrix} 1,3\times1\times1,1 \\ 1,1\times3\times3,1 \end{bmatrix}$, stride 1, 1, 1 | 10×128×32×22 |
| | | Pi-Conv | Conv3D $\begin{bmatrix} 128,3\times3\times3,128 \\ 128,3\times1\times1,128 \\ 128,1\times3\times3,128 \end{bmatrix}$, stride 1, 1, 1 | |
| MSSFE | Scale Branch1 | | Conv1D [128, 3, 128], stride 1, dilation=1 | 384×32 |
| | Scale Branch2 | | Conv1D [128, 3, 128], stride 1, dilation=2 | |
| | Scale Branch3 | | Conv1D [128, 3, 128], stride 1, dilation=4 | |
| Separate FC | | | For each part, FC [384, 128] | 128×32 |

150K iterations. In the training phase, we randomly select 30 consecutive frames from each gait sequence as input. While in the testing phase, all silhouette images of each gait sequence are used to obtain the final representation. Furthermore, rank-1 identification accuracy is adopted to measure the identification performance in the subsequent experiments.

### C. Performance Comparison on CASIA-B

In this section, we evaluate the performance of the proposed method on the CASIA-B dataset, and several state-of-the-art methods are chosen for comparison, including GEINet [52], CNN-LB [1], GaitNet [33], ACL [2], GaitPart [23], GaitSet [22], MT3D [29], GaitSlice [27] and MvGAN [54]. For a systematical and comprehensive comparison, the experiments under all cross-view and cross-walking-condition cases are conducted. To alleviate the influence of randomness, all experiments in this subsection are conducted five times with different random seeds, and the mean and standard deviation of the experimental results are reported. Tab. II lists the average rank-1 accuracy for each probe view on all gallery views excluding the identical-view case, and the best record under each probe view is marked in bold.

As listed in Tab. II, the proposed method achieves the best performance with the mean recognition rates of 97.4%, 94.0% and 84.0% under the condition of NM, BG and CL, respectively, which demonstrates the superiority of ESNet. Furthermore, some interesting experimental phenomena can be also analyzed from Tab. II:

• Effective extraction of temporal information can improve recognition rates. Compared with GEINet and CNN-LB, the methods that take temporal information into account, such as GaitNet, ACL, GaitSet, GaitPart, GaitSlice, MT3D and ESNet, have clear performance advantages. This indicates that fully exploring the spatial-temporal information from original gait sequences is the key to improving the recognition performance.

• The temporal-spatial correlation also contributes to superior performances. This phenomenon is clearly revealed in Tab. II that MT3D and ESNet surpass GaitNet, ACL, GaitSet, GaitPart and GaitSlice. Moreover, ESNet is superior to MT3D which also uses 3D convolution. The reason is that the proposed ESNet can obtain more effective spatial-temporal gait representations by integrating spatial and temporal decoupled modeling and the direct 3D convolution.

• The proposed ESNet is more robust, i.e., the recognition accuracy of ESNet drops less under more difficult testing conditions. For example, the mean accuracy of GaitSet drops by almost 27% when the walking condition changes from NM (96.1%) to CL (70.3%). Corresponding to that, the performance degradation of ESNet is only 13.7% (from 97.4% to 84.0%). In the NM scenario, both temporal and spatial information contribute to gait recognition performance. However, in the CL scenario, large appearance changes make temporal characteristics more dominant. Therefore, compared with direct 3D convolution which may introduce extra interference as extracting spatial-temporal information simultaneously, the proposed block is much more adept at learning high-quality spatial-temporal gait features. In addition, MSSFE enables ESNet to capture more discriminative fine-grained gait features. All of the above enhance the robustness of ESNet to various walking conditions.

TABLE II
CROSS-VIEW AVERAGE RANK-1 ACCURACIES (%) ON CASIA-B FOR DIFFERENT PROBE VIEWS EXCLUDING THE IDENTICAL-VIEW CASES.

| Gallery NM#1-4 | | 0°-180° | | | | | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Probe | | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | |
| NM#5-6 | GEINet [52] | 40.2 | 38.9 | 42.9 | 45.6 | 51.2 | 42.0 | 53.5 | 57.6 | 57.8 | 51.8 | 47.7 | 48.1 |
| | CNN-LB [1] | 82.6 | 90.3 | 96.1 | 94.3 | 90.1 | 87.4 | 89.9 | 94.0 | 94.7 | 91.3 | 78.5 | 89.9 |
| | GaitNet [33] | 93.1 | 92.6 | 90.8 | 92.4 | 87.6 | **95.1** | 94.2 | 95.8 | 92.6 | 90.4 | 90.2 | 92.3 |
| | ACL [2] | 92.0 | 98.5 | **100.0** | 98.9 | 95.7 | 91.5 | 94.5 | 97.7 | 98.4 | 96.7 | 91.9 | 96.0 |
| | GaitSet [22] | 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 [23] | 94.1 | 98.6 | 99.3 | 98.5 | 94.0 | 92.3 | 95.9 | 98.4 | 99.2 | 97.8 | 90.4 | 96.2 |
| | MvGAN [54] | 94.8 | 99.0 | 99.7 | **99.2** | 96.6 | 93.7 | 96.3 | 98.6 | 99.2 | 98.2 | 92.3 | 97.1 |
| | GaitSlice [27] | 95.5 | **99.2** | 99.6 | 99.0 | 94.4 | 92.5 | 95.0 | 98.1 | **99.7** | 98.3 | 92.9 | 96.7 |
| | MT3D [29] | **95.7** | 98.2 | 99.0 | 97.5 | 95.1 | 93.9 | 96.1 | 98.6 | 99.2 | 98.2 | 92.0 | 96.7 |
| | Ours | 95.6 ± 0.218 | 98.6 ± 0.050 | 99.1 ± 0.076 | 97.9 ± 0.041 | **96.7** ± 0.119 | 94.4 ± 0.326 | **96.9** ± 0.233 | **98.7** ± 0.218 | 99.3 ± 0.041 | **98.6** ± 0.076 | **95.1** ± 0.292 | **97.4** ± 0.043 |
| BG #1-2 | GEINet [52] | 34.2 | 29.3 | 31.2 | 35.2 | 35.2 | 27.6 | 35.9 | 43.5 | 45.0 | 39.0 | 36.8 | 35.7 |
| | CNN-LB [1] | 64.2 | 80.6 | 82.7 | 76.9 | 64.8 | 63.1 | 68.0 | 76.9 | 82.2 | 75.4 | 61.3 | 72.4 |
| | GaitNet [33] | 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 [22] | 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 [23] | 89.1 | 94.8 | 96.7 | **95.1** | 88.3 | 84.9 | 89.0 | 93.5 | 96.1 | 93.8 | 85.8 | 91.5 |
| | MvGAN [54] | 92.4 | 94.7 | 97.2 | 94.6 | 88.7 | 83.6 | 87.8 | 93.8 | 96.3 | 95.2 | 86.8 | 91.9 |
| | GaitSlice [27] | 90.2 | **96.4** | 96.1 | 94.9 | 89.3 | 85.0 | 90.9 | 94.5 | 96.3 | 95.0 | 88.1 | 92.4 |
| | MT3D [29] | 91.0 | 95.4 | **97.5** | 94.2 | 92.3 | 86.9 | **91.2** | 95.6 | **97.3** | 96.4 | 86.6 | 93.0 |
| | Ours | **92.7** ± 0.139 | 95.9 ± 0.050 | 96.3 ± 0.050 | 94.9 ± 0.082 | **93.2** ± 0.100 | 87.7 ± 0.100 | 90.9 ± 0.119 | **96.2** ± 0.091 | **97.3** ± 0.076 | **96.9** ± 0.146 | **91.7** ± 0.122 | **94.0** ± 0.026 |
| CL #1-2 | GEINet [52] | 19.9 | 20.3 | 22.5 | 23.5 | 26.7 | 21.3 | 27.4 | 28.2 | 24.2 | 22.5 | 21.6 | 23.5 |
| | CNN-LB [1] | 37.7 | 57.2 | 66.6 | 61.1 | 55.2 | 54.6 | 55.2 | 59.1 | 58.9 | 48.8 | 39.4 | 54.0 |
| | GaitNet [33] | 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 [22] | 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 [23] | 70.7 | 85.5 | 86.9 | 83.3 | 77.1 | 72.5 | 76.9 | 82.2 | 83.8 | 80.2 | 66.5 | 78.7 |
| | MvGAN [54] | 70.5 | 77.9 | 82.5 | 82.7 | 77.4 | 73.6 | 73.8 | 77.8 | 77.6 | 72.5 | 64.8 | 75.6 |
| | GaitSlice [27] | 75.6 | 87.0 | 88.9 | 86.5 | 80.5 | 77.5 | 79.1 | 84.0 | 84.8 | 83.6 | 70.1 | 81.6 |
| | MT3D [29] | **76.0** | 87.6 | 89.8 | 85.0 | 81.2 | 75.7 | 81.0 | 84.5 | 85.4 | 82.2 | 68.1 | 81.5 |
| | Ours | 75.6 ± 0.887 | **89.2** ± 0.345 | **92.4** ± 0.227 | **90.3** ± 0.802 | **84.3** ± 0.178 | **80.2** ± 1.252 | **83.0** ± 0.761 | **86.3** ± 0.473 | **89.0** ± 0.305 | **83.9** ± 0.536 | **69.8** ± 0.657 | **84.0** ± 0.024 |

### *D. Performance Comparison on OULP*

To verify the generalization of the proposed method, we perform the evaluation of ESNet on the OULP dataset and all experiments are conducted five times using different random seeds as in Sec. IV-C. The detailed experimental results of ESNet and several state-of-the-art methods including NN [50], MGAN [19], CNNS [1] and MT3D [29] for each view are reported in Tab. III. It can be found that ESNet achieves the highest accuracy in all cross-view cases with a clear performance advantage. In the identical view case, although the proposed method achieves sub-optimal results, the gap between the results of the proposed method and the optimal method is no more than 0.2%, which is negligible compared with the performance improvement in cross-view cases. Moreover, as can be seen in Tab. III, the accuracy of CNNS, MGAN and MT3D may drop heavily when the angle between the probe and the gallery becomes larger. For example, the recognition rate of the probe and gallery with angles of (55°, 85°) is significantly lower than that of the probe and gallery with (55°, 65°) and (55°, 75°). The same problem also exists when the probe is another view. Nevertheless, the proposed ESNet can still obtain excellent recognition performance when the angles of the probe and gallery are quite different, indicating that ESNet is more robust to the variable of view and has better generalization ability.

### *E. Performance Comparison on OUMVLP*

To further evaluate the performance of the proposed method, the evaluation of ESNet is completed on the largest public gait dataset, i.e., OUMVLP. As in Sec. IV-C, experiments in this subsection are also performed five times with different random seeds. Tab. IV lists the comparison results between ESNet and other five famous methods, including DULE [55], GaitSet [22], GaitPart [23], GLN [25] and GaitSlice [27]. As listed in Tab. IV, ESNet achieves the best recognition performance in most cases, which demonstrates the generalization capability of ESNet on a large-scale dataset. For some probe sequences, the corresponding sequences are not available in the gallery due to the incomplete sample collection for some subjects. If we exclude the case where there is no corresponding sample in the gallery, the average rank-1 accuracy of all probe views can rise to 95.8%.

### *F. Ablation Study*

*1) **Incremental evaluation of each component**:*

To validate the effectiveness of the Pi-Conv layer, STDA unit and MSSFE, we incrementally evaluate each component on the CASIA-B dataset. It is worth noting that unlike most gait recognition methods that decouple spatial and temporal feature extraction into two processes or use direct 3D convolution to extract spatial-temporal information, the Pi-Conv layer achieves the integration of both above to extract

TABLE III
CROSS-VIEW AVERAGE RANK-1 ACCURACIES (%) ON OULP FOR FOUR VIEWS EXCLUDING THE IDENTICAL-VIEW CASES.

| Probe angle | Method | Gallery angle | | | | | Identical angle |
|---|---|---|---|---|---|---|---|
| | | 55° | 65° | 75° | 85° | Mean | |
| 55° | NN | | - | - | - | - | 84.7 |
| | CNNS | | 98.3 | 96.0 | 80.5 | 91.6 | 98.8 |
| | MGAN | | 99.4 | 96.1 | 77.9 | - | 98.8 |
| | MT3D | | 99.6 | 98.1 | 84.7 | 94.2 | 100 |
| | Ours | | **99.7** ± 0.052 | **99.2** ± 0.052 | **98.7** ± 0.000 | **99.2** ± 0.027 | 99.9 ± 0.000 |
| 65° | NN | - | | - | - | - | 86.6 |
| | CNNS | 96.3 | | 97.3 | 83.3 | 92.3 | 98.9 |
| | MGAN | 97.7 | | 98.5 | 84.4 | - | - |
| | MT3D | 97.8 | | 98.5 | 84.9 | 93.7 | 99.9 |
| | Ours | **99.5** ± 0.000 | | **99.5** ± 0.000 | **98.8** ± 0.000 | **99.3** ± 0.000 | 99.8 ± 0.052 |
| 75° | NN | - | - | | - | - | 86.9 |
| | CNNS | 94.2 | 97.8 | | 85.1 | 92.4 | 98.9 |
| | MGAN | 94.8 | 98.9 | | 86.4 | - | - |
| | MT3D | 96.8 | 99.0 | | 86.1 | 94.0 | 99.9 |
| | Ours | **99.0** ± 0.064 | **99.6** ± 0.064 | | **99.1** ± 0.053 | **99.2** ± 0.032 | 99.7 ± 0.052 |
| 85° | NN | - | - | - | | - | 85.7 |
| | CNNS | 90.0 | 96.0 | 98.4 | | 94.8 | 98.9 |
| | MGAN | 86.9 | 97.4 | 99.5 | | - | - |
| | MT3D | 96.4 | 98.4 | 99.5 | | 98.1 | 99.8 |
| | Ours | **99.2** ± 0.097 | **99.5** ± 0.056 | **99.8** ± 0.052 | | **99.5** ± 0.058 | 99.7 ± 0.064 |

TABLE IV
CROSS-VIEW AVERAGE RANK-1 ACCURACIES (%) ON OUMVLP EXCLUDING IDENTICAL-VIEW CASES.

| Probe | Gallery all 14 views | | | | | |
|---|---|---|---|---|---|---|
| | DULE | GaitSet | GaitPart | GLN | GaitSlice | Ours |
| 0° | 56.2 | 81.3 | 82.6 | 83.8 | 84.1 | **84.8**±0.065 |
| 15° | 73.7 | 88.6 | 88.9 | **90.0** | 89.0 | 89.6±0.013 |
| 30° | 81.4 | 90.2 | 90.8 | 91.0 | **91.2** | 91.0±0.110 |
| 45° | 82.0 | 90.7 | 91.0 | 91.2 | **91.6** | 91.3±0.004 |
| 60° | 78.4 | 88.6 | 89.7 | 90.3 | 90.6 | **90.7**±0.007 |
| 75° | 78.0 | 89.1 | 89.9 | 90.0 | 89.9 | **90.4**±0.003 |
| 90° | 76.5 | 88.3 | 89.5 | 89.4 | 89.8 | **89.9**±0.027 |
| 180° | 60.2 | 83.1 | 85.2 | 85.3 | 85.7 | **88.5**±0.040 |
| 195° | 72.0 | 87.7 | 88.1 | 89.1 | **89.3** | 87.5±0.042 |
| 210° | 79.8 | 89.4 | 90.0 | 90.5 | **90.6** | 90.1±0.009 |
| 225° | 80.2 | 89.7 | 90.1 | 90.6 | **90.7** | 90.2±0.015 |
| 240° | 76.7 | 87.8 | 89.0 | 89.6 | **89.8** | 89.4±0.014 |
| 255° | 76.3 | 88.3 | 89.1 | 89.3 | **89.6** | 89.3±0.015 |
| 270° | 73.9 | 86.9 | 88.2 | 88.5 | 88.5 | **88.5**±0.042 |
| Mean | 74.7 | 87.9 | 88.7 | 89.2 | 89.3 | **89.4**±0.022 |

more comprehensive gait information. The Pi-Conv layer is composed of three parallel 3D convolutions, i.e., the right two in Fig. 2(b) implement the decoupled spatial and temporal feature extraction, and the rest in Fig. 2(b) is the direct 3D convolution. The detailed experimental results are presented in Tab. V, where CBAM and SEM denote the Convolutional Block Attention Module [43] and Squeeze-and-Excitation module [42], respectively. From Tab. V, several observations can be drawn:

• Under three walking conditions, the average accuracy when using either decoupled extraction or direct 3D convolution is 88.9% and 89.5%, while the average accuracy when the Pi-Conv layer uses both (group g) is 90.1%. It increases by 1.2% and 0.6%, respectively. With the assistance of the STDA unit, the improvements are clearer, especially under difficult testing conditions. For example, under the CL condition, compared with the cases using either decoupled extraction or direct 3D convolution, the combination of the Pi-Conv layer and the STDA unit (group h) obtains significant improvements of 3.3% and 1.7%, respectively, which powerfully demonstrates the necessity and effectiveness of the Pi-Conv layer and the STDA unit.

• As shown in groups (h-j), only the STDA unit that combines the spatial and temporal attention mechanism can improve the performance. Especially, both SEM and CBAM severely reduce the accuracy under all three walking conditions, which means that channel attention is not suitable for silhouette sequence-based gait recognition. Moreover, an inappropriate combination of attention may have an additional negative impact, as the accuracy with CBMA is reduced again compared to SEM. It fully validates the rationality of the STDA unit's design and its applicability for silhouette sequence-based gait recognition.

• Combined with MSSFE, the recognition accuracy can be boosted once again. As listed in Tab. V, based on the joint use of the Pi-Conv layer and the STDA unit (group h), MEEFS improves the performance under three walking conditions by 0.9%, 1.2%, and 1.3%, respectively, and makes the proposed model achieve the best accuracy. This can be attributed to MSSFE, by which more robust and salient part-based representations for gait recognition can be obtained.

*2) **Impact of each branch in the Pi-Conv layer**:*

As shown in Fig. 2(b), the Pi-Conv layer is composed of temporal, spatial and direct 3D convolution branches. To explore the role of each branch, we conduct comparison experiments of ESNet and its three degradation models on the CASIA-B dataset, where each of these degradation models is implemented by using only one branch in the Pi-Conv layer. The experimental results are shown in Fig. 6, from which we can see that the direct 3D convolution branch shows the biggest contribution among the three branches. Although the temporal and spatial branches contribute relatively weakly compared to the direct 3D convolution branch, they are both useful because when they are combined with the direct 3D convolution branch, i.e., the proposed Pi-Conv layer, our ESNet achieves the best results under all three walking conditions.

*3) **Impact of each branch in the STDA unit**:*

The STDA unit consists of two parallel attention branches, i.e., the temporal attention branch and the spatial attention branch as shown in Fig. 2(c). Similarly, to explore the role of each branch in the STDA unit, we compare ESNet with its degradation models that implemented with only one attention branch of the STDA unit. The experiments are conducted on the CASIA-B dataset and the experimental results are shown in Fig. 7. It can be observed that the contribution of the spatial attention branch is slightly better than that of the temporal attention branch. The combination of these two branches, i.e., the proposed STDA unit, can further improve performance, which also demonstrates the role of the STDA unit.

TABLE V
AVERAGED RANK-1 ACCURACIES (%) OF ESNET FOR ABLATION STUDIES ON CASIA-B.

| Groups | Pi-Conv | | STDA | CBAM | SEM | MSSFE | Accuracy | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Decoupled | Direct 3D | | | | | NM | BG | CL | Mean |
| a | ✓ | | | | | | 95.8 | 91.5 | 79.4 | 88.9 |
| b | ✓ | | ✓ | | | | 95.8 | 92.0 | 79.4 | 89.0 |
| c | ✓ | | ✓ | | | ✓ | 96.7 | 92.6 | 82.1 | 90.5 |
| d | | ✓ | | | | | 95.8 | 91.8 | 81.0 | 89.5 |
| e | | ✓ | ✓ | | | | 96.1 | 92.3 | 81.8 | 90.1 |
| f | | ✓ | ✓ | | | ✓ | 96.9 | 93.2 | 83.4 | 91.2 |
| g | ✓ | ✓ | | | | | 96.3 | 92.4 | 81.6 | 90.1 |
| h | ✓ | ✓ | ✓ | | | | 96.5 | 92.8 | 82.7 | 90.7 |
| i | ✓ | ✓ | | ✓ | | | 91.1 | 83.9 | 68.1 | 81.0 |
| j | ✓ | ✓ | | | ✓ | | 93.3 | 87.2 | 72.6 | 84.4 |
| k | ✓ | ✓ | ✓ | | | ✓ | **97.4** | **94.0** | **84.0** | **91.8** |

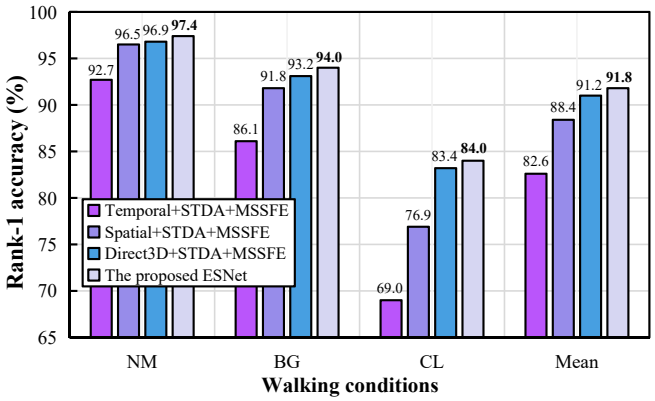



Fig. 6. The impact of each branch within the Pi-Conv layer on the performance of the proposed ESNet on CASIA-B.

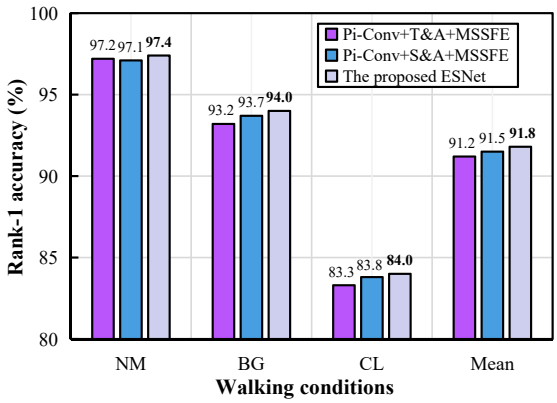



Fig. 7. The impact of each branch within the STDA unit on the performance of the proposed ESNet on CASIA-B. (T&A and S&A represent the temporal attention branch and the spatial attention branch, respectively.)

To further analyze the role of the STDA unit in ESNet, on the CASIA-B dataset, we visualize the attention scores of temporal and spatial branches of the STDA unit in Block3. As shown in Fig. 8, by normalizing the temporal and spatial attention scores, the heat maps of one subject's gait sequences under three different conditions (NM, BG, and CL) are drawn, where each condition contains 11 different views. It can be observed that spatial attention can focus on crucial body regions. By comparing the second, fourth and sixth rows in Fig. 8, we can find that the temporal attention scores vary under different walking conditions even for the same view of the same subject. As the walking condition becomes more complex, the temporal attention scores become larger. Obviously, the temporal attention scores of all 11 views in the CL case are larger than those in the NM case. It strongly indicates that the STDA unit can indeed adjust the output of the Pi-Conv layer adaptively under different walking conditions, and verifies the effectiveness and necessity of the STDA unit again.

*4) **Impact of multi-scale in MSSFE**:*

We achieve the robust fine-grained feature learning through multi-scale feature extraction by using parallel multiple 1D convolutions with different dilation rates in MSSFE. For each part, different settings of dilation rates enable them to have different contextual perception capabilities, and different multi-scale combinations may have different effects on the model. Therefore, we conduct experiments of different multi-scale combinations on the CASIA-B dataset and report the experimental results in Tab. VI.

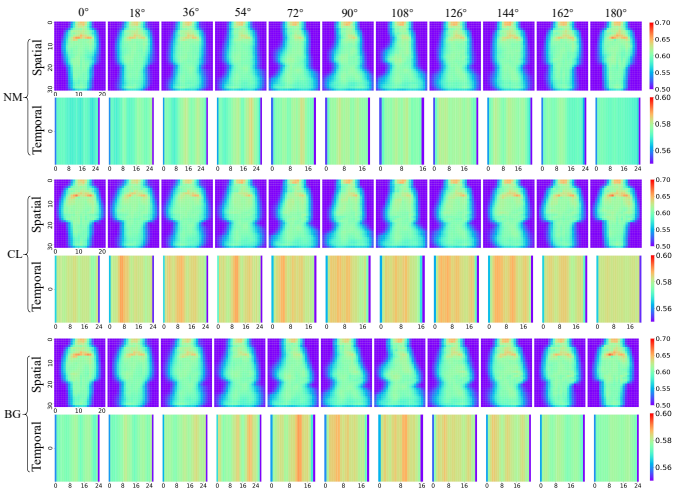



Fig. 8. Visualization of attention scores for spatial and temporal branches in the STDA unit on CASIA-B. The first two rows are heat maps of the spatial and temporal attention scores for 11 views under the NM walking condition, the middle two rows are the attention scores under the CL walking condition, and the last two rows are under the BG walking condition.

From Tab. VI, we can observe that under all three walking conditions, the accuracy first increases and then decreases as the number of different scales increases from 1 to 4. When the combination includes more scales from (1, 2, 4) to (1, 2, 4, 8), a significant decrease in recognition performance occurs, which indicates that too many combinations of scales would instead lead to parameter redundancy. Therefore, the number

TABLE VI
AVERAGED RANK-1 ACCURACIES (%) OF DIFFERENT SCALES SET IN MSSFE ON CASIA-B.

| Groups | Different scales in MSSFE | Accuracy | | |
|---|---|---|---|---|
| | | NM | BG | CL |
| a | (1) | 96.7 | 92.8 | 82.4 |
| b | (1, 2) | 96.8 | 93.2 | 83.4 |
| c | (1, 4) | 97.1 | 93.6 | 83.5 |
| d | (2, 4) | 97.1 | 93.0 | 82.2 |
| e | (1, 2, 4) | **97.4** | **94.0** | **84.0** |
| f | (1, 2, 8) | 96.6 | 92.7 | 82.5 |
| g | (1, 4, 8) | 96.9 | 93.0 | 82.8 |
| h | (2, 4, 8) | 96.9 | 92.6 | 82.0 |
| i | (1, 2, 4, 8) | 96.8 | 93.0 | 82.9 |

TABLE VII
AVERAGED RANK-1 ACCURACIES (%) OF DIFFERENT FEATURE MAP SIZES INPUT TO MSSFE ON CASIA-B.

| Groups | Different feature map sizes input to MSSFE (H×W) | Accuracy | | | |
|---|---|---|---|---|---|
| | | NM | BG | CL | Mean |
| a | (16×11) | 96.0 | 90.7 | 78.1 | 88.3 |
| b | (32×22) | **97.4** | 94.0 | 84.0 | 91.8 |
| c | (64×44) | 97.1 | **94.2** | **85.1** | **92.1** |

of different scales is set to 3. Moreover, when the number of scales is the same, an appropriate combination of scales can bring performance gain, while expanding the context-aware scope of each part excessively may be contrary to the purpose of the STDA unit, resulting in performance degradation. For example, under the conditions of NM, BG and CL, a multi-scale combination of (1, 4) has higher accuracy than (1, 2), while (1, 2, 8) has lower accuracy than (1, 2, 4). In all multi-scale combinations, when 3 different scales are combined and set to (1, 2, 4), the proposed model achieves the highest recognition rates. Thereby, the multi-scale combination of (1, 2, 4) is selected to implement the proposed model.

*5) **Impact of input feature map size in MSSFE**:*

To further analyze the influence of the input feature map size on MSSFE, we fix the multi-scale combination in MSSFE as (1, 2, 4) and conduct experiments with different input feature map sizes for MSSFE on the CASIA-B dataset. The experimental results are listed in Tab. VII, where input feature map sizes set to (64×44) and (16×11) are implemented by removing the MaxPool3D layer (listed in Tab. I) after Block1 or adding another MaxPool3D layer after Block2, respectively. It can be observed that when the scale of the feature maps input to MSSFE is reduced from (32×22) to (16×11), the mean accuracy under three walking conditions drops from 91.8% to 88.3%, which shows a significant performance degradation. When the feature map scale is extended to (64×44), the average recognition rate under three walking conditions is 92.1%, with a weak improvement but a significant increase in model training time and GPU occupation. Therefore, balancing the accuracy and calculation load, we set the scale of the feature maps input to MSSFE as (32×22).

*6) **Impact of spatial aggregator in MSSFE**:*

To verify the effectiveness of the MSSA, we design the comparison experiment by implementing the proposed framework with different spatial aggregation methods on CASIA-B, including aggregating with a single statistical function, e.g., Max(), Mean(), and the sum of them, as well as the deformation of them, e.g., Global contrastive pooling (GCP) [56]. The comparison results are listed in Tab. VIII. It can be observed that compared with Mean(), Max() and the sum of Max() and Mean() have a significant improvement, especially under the condition of CL. The performance of GCP in this experiment is normal. What's exciting is that the MSSA proposed in this paper achieves the best performance under all three walking conditions, and has obvious advantages compared with other four spatial feature aggregation methods.

TABLE VIII
AVERAGED RANK-1 ACCURACIES(%) OF DIFFERENT SPATIAL AGGREGATORS IN MSSFE ON CASIA-B.

| Groups | Different spatial aggregators in MSSFE | Accuracy | | |
|---|---|---|---|---|
| | | NM | BG | CL |
| a | Max() | 96.9 | 93.3 | 82.0 |
| b | Mean() | 96.4 | 92.6 | 79.9 |
| c | Max()+Mean() | 96.7 | 92.8 | 82.7 |
| d | GCP | 96.7 | 92.8 | 82.1 |
| e | MSSA | **97.4** | **94.0** | **84.0** |

## V. CONCLUSION AND FUTURE WORK

In this work, we present a novel insight that integrating the extraction of temporal and spatial information separately in a decoupled manner and the simultaneous extraction of spatial-temporal information using 3D convolution can yield better spatial-temporal feature representations of gait. The proposed ESNet for cross-view gait recognition consists of the Dual-Attention Guided Feature Extractor (DAGFE) with stacked well-designed blocks and the Multi-Scale Salient Feature Extractor (MSSFE). Specially, the proposed block is a novel residual learning block with a Pi-Conv layer and a STDA unit, the core of these two components is to enhance high-quality spatial-temporal feature learning, and MSSFE is designed for further part-based salient feature extraction. Thus, discriminative and robust fine-grained feature representations can be obtained by the ESNet. The experiments on CASIA-B, OULP and OUMVLP demonstrate that the proposed ESNet can bring improvement for cross-view gait recognition.

In the future work, we will thoroughly investigate the influence of view changes on the gait spatial-temporal feature extraction. To reduce the performance degradation due to view changes, the explicit modeling of viewpoints will be considered. In addition, although MSSFE is able to extract part-based features for fine-grained mining, the contribution of different part-based features is not considered. Therefore, fully considering the synergy between different part-based features and finding a better integration method for them is still the future work. Moreover, we will explore a unified framework that integrates gait segmentation [57–60] and recognition for online applications. For gait segmentation, the combination of local and global attention algorithms with segmentation technologies will also be considered, since attention algorithms can highlight important parts of silhouettes and suppress unnecessary parts.

## REFERENCES


[1] 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, 2016.

[2] 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, 2019.

[3] G. Zhao, G. Liu, H. Li, and M. Pietikainen, "3D gait recognition using multiple cameras," in *Proc. Int. Conf. Automat. Face Gesture Recog. (FG)*, 2006, pp. 529–534.

[4] G. Ariyanto and M. S. Nixon, "Model-based 3D gait biometrics," in *IEEE Int. Joint Conf. on Bio-metrics (IJCB)*, 2011, pp. 1–7.

[5] 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, p. 107069, 2020.

[6] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. Yagi, "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Trans. Biom.*, vol. 2, no. 4, pp. 421–430, 2020.

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

[8] H. Aggarwal and D. K. Vishwakarma, "Covariate conscious approach for gait recognition based upon Zernike moment invariants," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 2, pp. 397–407, 2017.

[9] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, 2005.

[10] 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, 2019.

[11] D. 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 *IEEE Int. Conf. Comput. Sustain. Glob. Develop.* IEEE, 2015, pp. 336–341.

[12] C. Dhiman and D. K. Vishwakarma, "A Robust Framework for Abnormal Human Action Recognition Using $\mathcal{R}$-Transform and Zernike Moments in Depth Videos," *IEEE Sensors J.*, vol. 19, no. 13, pp. 5195–5203, 2019.

[13] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2006, pp. 151–163.

[14] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Trans. Inf. Forensic Secur.*, vol. 8, no. 12, pp. 2034–2045, 2013.

[15] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3142–3157, 2019.

[16] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 734–747, 2020.

[17] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng, "A general tensor representation framework for cross-view gait recognition," *Pattern Recognit.*, vol. 90, pp. 87–98, 2019.

[18] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 260–274, 2021.

[19] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensic Secur.*, vol. 14, no. 1, pp. 102–113, 2018.

[20] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2016, pp. 2832–2836.

[21] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, 2019.

[22] 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.*, 2021, [Online]. Available: https://doi.org/10.1109/TPAMI.2021.3057879.

[23] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 14 225–14 233.

[24] H. Qin, Z. Chen, Q. Guo, Q. J. Wu, and M. Lu, "RPNet: Gait recognition with relationships between each body-parts," *IEEE Trans. Circuits Syst. Video Technol.*, 2021, [Online]. Available: https://doi.org/10.1109/TCSVT.2021.3095290.

[25] 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. (ECCV)*, 2020, pp. 382–398.

[26] F. Han, X. Li, J. Zhao, and F. Shen, "A unified perspective of classification-based loss and distance-based loss for cross-view gait recognition," *Pattern Recognit.*, 2022, [Online]. Available: https://doi.org/10.1016/j.patcog.2021.108519.

[27] H. Li, Y. Qiu, H. Zhao, J. Zhan, R. Chen, T. Wei, and Z. Huang, "GaitSlice: A gait recognition model based on spatio-temporal slice features," *Pattern Recognit.*, vol. 124, 2022, [Online]. Available: https://doi.org/10.1016/j.patcog.2021.108453.

[28] 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.

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

[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," *Cognit. Syst. Res.*, vol. 61, pp. 1–13, 2020.

[33] 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, 2022.

[34] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 5794–5803.

[35] X. Liao, L. He, Z. Yang, and C. Zhang, "Video-based person re-identification via 3D convolutional networks and non-local attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2018, pp. 620–634.

[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4489–4497.

[37] K. Yang, P. Qiao, D. Li, S. Lv, and Y. Dou, "Exploring temporal preservation networks for precise temporal action localization," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 32, no. 1, pp. 7477–7484.

[38] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen, "Appearance-preserving 3D convolution for video-based person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 228–243.

[39] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 4724–4733.

[40] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 6450–6459.

[41] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 5533–5541.

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

[43] 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.

[44] X. Cun and C.-M. Pun, "Improving the harmony of the composite image by spatial-separated attention module," *IEEE Trans. Image Process.*, vol. 29, pp. 4759–4771, 2020.

[45] N. Martinel, G. L. Foresti, and C. Micheloni, "Deep pyramidal pooling with attention for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 7306–7316, 2020.

[46] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," *arXiv:1903.10082*, 2019.

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

[48] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv:1703.07737*, 2017.

[49] 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. Int. Conf. Pattern Recog. (ICPR)*, vol. 4, 2006, pp. 441–444.

[50] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensic Secur.*, vol. 7, no. 5, pp. 1511–1521, 2012.

[51] 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. Vision Appl*, vol. 10, no. 1, pp. 1–14, 2018.

[52] 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.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[54] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian, "Multi-view gait image generation for cross-view gait recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 3041–3055, 2021.

[55] S. Zhang, Y. Wang, and A. Li, "Cross-view gait recognition with deep universal linear embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 9095–9104.

[56] H. Park and B. Ham, "Relation network for person re-identification," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, vol. 34, no. 7, 2020, pp. 11 839–11 847.

[57] 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. Develop. Syst.*, vol. 9, no. 4, pp. 316–327, 2016.

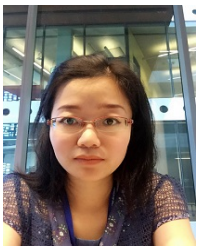
[58] D. K. Vishwakarma, R. Kapoor, and A. Dhiman, "A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics," *Rob. Auton. Syst.*, vol. 77, pp. 25–38, 2016.

[59] D. Vishwakarma, R. Kapoor, and A. Dhiman, "Unified framework for human activity recognition: an approach using spatial edge distribution and $\mathcal{R}$-transform," *AEU-*

*Int. J. Electron. Commun.*, vol. 70, no. 3, pp. 341–353, 2016.

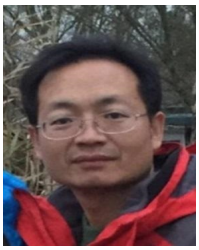
[60] D. K. Vishwakarma and R. Kapoor, "Integrated approach for human action recognition using edge spatial distribution, direction pixel and $\mathcal{R}$-transform," *Adv. Robot.*, vol. 29, no. 23, pp. 1553–1562, 2015.

**Tianhuan Huang** received the B.E. degree in Electronic Information Engineering from School of Physical Science and Technology, Nanjing normal university, Nanjing, China, in 2018. She is currently a Ph.D candidate with the School of Information Science and Engineering, Shandong University, Qingdao, China. Her current research interests include gait recognition, computer vision and machine learning.



**Xianye Ben** received the Ph.D. degree in pattern recognition and intelligent system from the College of Automation, Harbin Engineering University, Harbin, China, in 2010. She is currently working as a Full Professor with the School of Information Science and Engineering, Shandong University, Qingdao, China. She has authored or coauthored more than 100 papers in major journals and conferences, such as IEEE T-PAMI, IEEE T-IP, IEEE T-CSVT, IEEE T-MM, PR, CVPR, etc. Her current research interests include pattern recognition and image processing. She received the Excellent Doctorial Dissertation awarded by Harbin Engineering University. She was also enrolled by the Qilu Young Scholars Program of Shandong University.

**Chen Gong** received his B.E. degree from East China University of Science and Technology (ECUST) in 2010, and dual doctoral degree from Shanghai Jiao Tong University (SJTU) and University of Technology Sydney (UTS) in 2016 and 2017, under the supervision of Prof. Jie Yang and Prof. Dacheng Tao, respectively. Currently, he is a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests mainly include machine learning, data mining, and learning-based vision problems. He has published more than 50 technical papers at prominent journals and conferences such as IEEE T-NNLS, IEEE T-IP, IEEE T-CYB, IEEE T-CSVT, IEEE T-MM, IEEE T-ITS, CVPR, AAAI, IJCAI, ICDM, etc. He received the Excellent Doctorial Dissertation awarded by Shanghai Jiao Tong University (SJTU) and Chinese Association for Artificial Intelligence (CAAI). He was also enrolled by the Summit of the Six Top Talents Program of Jiangsu Province, China, and the Lift Program for Young Talents of China Association for Science and Technology.



**Baochang Zhang** received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of the Technology, Harbin, China, in 1999, 2001, and 2006, respectively. He is currently a full Professor with Institute of Artificial Intelligence, Beihang University, Beijing, China. His current research interests include explainable deep learning, computer vision and patter recognition.

**Rui Yan** graduated from Rensselaer Polytechnic Inistitute majoring in Computer Science with a Ph.D. degree. He is now a data scientist in Microsoft AI & R. His research interests include knowledge graph, machine learning and pattern recognition.

**Qiang Wu** received the B.Eng. and M.Eng. degrees in electronic engineering from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree in computing science from the University of Technology Sydney, Sydney, Australia, in 2004. He is currently an Associate Professor and a Core Member of the Global Big Data Technologies Centre, University of Technology Sydney. His major research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. He has published more than 70 refereed papers, including those published in prestigious journals and top international conferences. Dr. Wu has been a Guest Editor of several international journals, such as the Pattern Recognition Letters (PRL) and the International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI). He has served as a Chair and/or a Program Committee Member for a number of international conferences.