Dynamic Aggregated Network for Gait Recognition

Kang Ma[1], Ying Fu[1]*, Dezhi Zheng[1]*, Chunshui Cao[2], Xuecai Hu[2], Yongzhen Huang[2,3]
[1]Beijing Institute of Technology, [2]WATRIX.AI, [3]Beijing Normal University
kangx.ma@gmail.com, {fuying, zhengdezhi}@bit.edu.cn, {chunshui.cao, xuecai.hu}@watrix.ai,
huangyongzhen@bnu.edu.cn

Abstract

Gait recognition is beneficial for a variety of applications, including video surveillance, crime scene investigation, and social security, to mention a few. However, gait recognition often suffers from multiple exterior factors in real scenes, such as carrying conditions, wearing overcoats, and diverse viewing angles. Recently, various deep learning-based gait recognition methods have achieved promising results, but they tend to extract one of the salient features using fixed-weighted convolutional networks, do not well consider the relationship within gait features in key regions, and ignore the aggregation of complete motion patterns. In this paper, we propose a new perspective that actual gait features include global motion patterns in multiple key regions, and each global motion pattern is composed of a series of local motion patterns. To this end, we propose a Dynamic Aggregation Network (DANet) to learn more discriminative gait features. Specifically, we create a dynamic attention mechanism between the features of neighboring pixels that not only adaptively focuses on key regions but also generates more expressive local motion patterns. In addition, we develop a self-attention mechanism to select representative local motion patterns and further learn robust global motion patterns. Extensive experiments on three popular public gait datasets, i.e., CASIA-B, OUMVLP, and Gait3D, demonstrate that the proposed method can provide substantial improvements over the current state-of-the-art methods.[1]

1. Introduction

Gait recognition aims to retrieve the same identity at a long distance, and has been widely used throughout social security [28], video surveillance [4, 15, 49], crime investigation [25], and so on. Compared with action recognition [17, 53, 54] and person re-identification [2, 55, 60, 61], the gait recognition task is one of the most challenging fine-grained label classification problems. On the one hand, silhouette data is a binary image of a person suffering from the limitations of the segmentation algorithm [26, 62, 63], with occasional holes and broken edges. On the other hand, gait recognition is also impacted by various exterior factors in real scenes, such as carrying conditions, wearing coats, and diverse viewing angles. Different angles and clothing conditions will greatly change the silhouette appearance of the same person, resulting in the intra-class variance being much greater than inter-class. We ask: *How to learn more robust features adaptively for each person under the influence of various external factors*? We attempt to answer this question from the following perspectives:

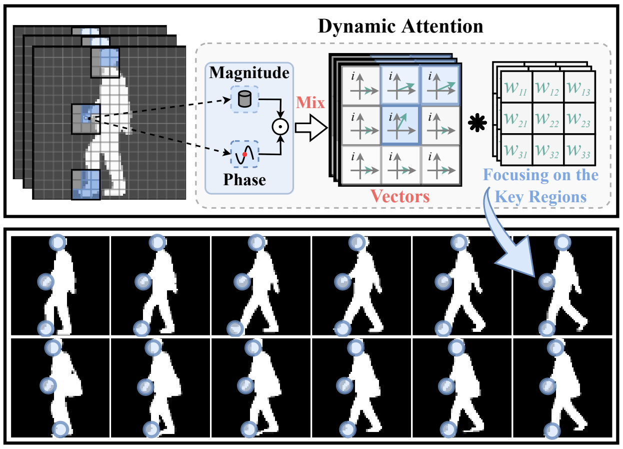

Dynamic Attention
Magnitude
Phase
Mix
Vectors
W_{11} W_{12} W_{13}
W_{21} W_{22} W_{23}
W_{31} W_{32} W_{33}
Focusing on the Key Regions

Figure 1. The features of each pixel are mapped as a vector with both magnitude and phase components. The magnitude represents contextual information, while the phase direction is used to construct dynamic attention models for the key regions. The convolution operation is denoted by "$*$", and the blue circles in the diagrams represent the key regions learned by the dynamic attention.

(i) Local Motion Patterns. Gait, or the act of walking, is essentially the coordinated movement of body parts. In a gait sequence, we observe that each part has a unique representative motion pattern, and each motion pattern is composed of a set of localized sub-movements. Therefore, it

*Corresponding Authors
[1] Code available at https://github.com/XKMar/FastGait

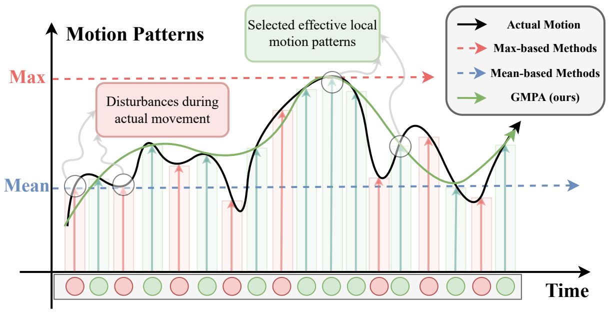


Figure 2. Comparison of the actual motion pattern with the Max-based method, Mean-based method, and Global Motion Pattern Aggregator (GMPA) module. The black curve represents a single periodic action that is affected by disturbances, whereas the green curve represents a synthesized periodic action consisting of distinct local motion patterns selected by the GMPA.

is critical to accurately locate the discriminative parts and obtain representative local motion patterns under the interference of various external factors. However, previous gait-based approaches [7, 8, 13, 14, 20, 24, 33] simply use convolutional networks with non-linear activation to model the dynamic movements. Once the network has been trained, the parameters and the non-linear function can only focus on the fixed patterns. To this end, we propose to encode the features of each pixel as a vector with magnitude and phase, as shown in Fig. 1, which allows learning the dynamic attention mapping functions among the neighboring pixel of focusing. By modeling the relationship, the network can further focus on local motion patterns in key regions.

(ii) Global Motion Patterns. Gait is a periodic movement. We assume that the actual motion pattern is a one-dimensional signal, as shown in Fig. 2, whereby the local motion patterns are the points on the signal. Therefore, it is essential to use a series of local motion patterns to further fit the actual motion patterns for obtaining discriminative gait features. However, recent gait-based methods [8,20,33] only use Max- or Mean-based methods to extract one of the significant local features. These methods are susceptible to disturbances and can not fit the actual motion patterns. According to the Nyquist-Shannon sampling theorem [37, 39] in signal processing theory, when a continuous signal is sampled at a frequency greater than twice the frequency of the signal, the information of the original signal is retained intact. In this regard, we propose to construct a global attention model and use it to dynamically select a preset number of distinguishable local motion patterns (green arrows), while excluding the effect of noise (red arrows). By selecting sufficient discriminative local motion patterns, the network can further obtain robust global motion patterns.

Driven by this analysis, we propose a novel and effective Dynamic Aggregated Network (DANet) for gait recognition. As shown in Fig. 3, DANet consists of two well-designed components, *i.e.*, Local Conv-Mixing Block (LCMB) and Global Motion Patterns Aggregator (GMPA). **Firstly**, we encode the features of each pixel into the complex domain including magnitude and phase, where the magnitude term represents the contextual information and the phase term is used to establish the relationship between each vector. The local motion pattern is generated by aggregating the magnitude and phase of the vectors in the neighboring regions of focus. **Secondly**, we use the self-attentive mechanism in the GMPA model to dynamically select sufficient discriminative local motion patterns and further learn to fit the actual gait patterns. **Finally**, with our proposed modules, we obtain the most representative stable gait features for each person and outperform the state-of-the-art (SOTA) methods, especially under the most challenging condition of cross-dressing.

Our main contributions can be summarized as follows:

- We propose a novel LCMB to extract the representative local motion patterns, which can dynamically model the relationships among the features of neighboring pixels and then accurately locate key regions.
- We design an effective GMPA to select the discriminative local motion patterns and then aggregate them to obtain a robust global representation. To the best of our knowledge, it is the first attempt to explore the potential of self-attention model in this task.
- Experimental results are illustrated to demonstrate the effectiveness of the proposed method, outperforming the SOTA method on CASIA-B [56], OUMVLP [41] and Gait3D [59] datasets. In addition, many rigorous ablation experiments on CASIA-B [56] further validated the effectiveness of each component in DANet.

2. Related Works

In this section, we provide a brief overview of relevant research in the fields of gait recognition, local action modeling, and global action modeling.

Gait Architectures. Gait recognition approaches mainly fall into two typical categories, *i.e.*, model-based approaches and appearance-based approaches. Model-based approaches [1, 3, 27, 29–31, 44] attempt to explicitly fit human pose structures [10] to images. However, the predefined points are empirically designed and limited by the inaccuracy of estimation results for low-quality images. Therefore, the model-based methods are generally inferior to appearance-based methods in performance. Appearance-based approaches [7, 8, 13, 18, 20, 21, 24, 32, 33, 43, 47, 51, 52, 58] are the mainstream frameworks for gait and have benefited from the quick growth of deep learning. It can be roughly divided into three classes, namely template-based

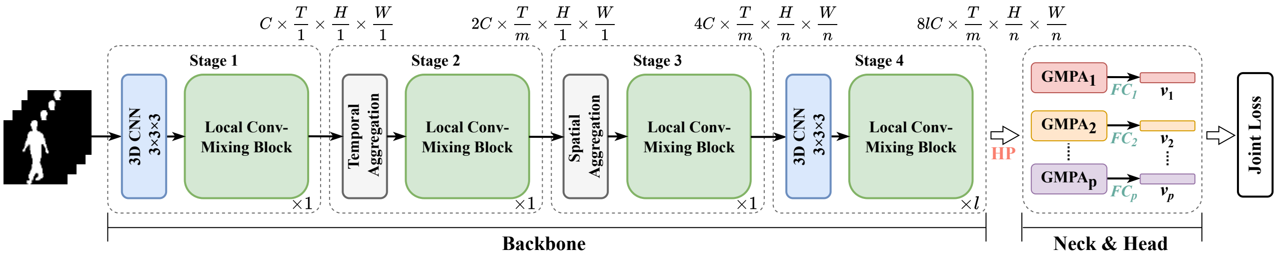


Figure 3. The overview of the proposed DANet. Each stage contains the Local Conv-Mixing Block (LCMB), which utilizes a dynamic attention model to establish relationships among neighboring pixels of interest. The HP denotes Horizontal Pooling, GMPA represents Global Motion Pattern Aggregator, and l indicates the number of last stages. In particular, the GMPA_j module is responsible for aggregating the local motion patterns of the j-th part and producing the final global motion patterns v_i for recognition.

methods, set-based methods, and sequence-based methods. The template-based approaches [18, 36, 43, 47, 52, 57] extracted gait spatio-temporal features by compressing a sequence of gait silhouettes, *e.g*., Gait Energy Image (GEI), which inevitably destroyed the representation of discriminative local motion patterns in gait sequences. The set-based approaches [8,20,21,23] assumed that the appearance of a silhouette contained its position information, which could not construct local motion patterns using continuous frames. Recently some advanced sequence-based methods [7, 24, 32, 33, 51] used 3D convolutional (C3D) neural networks to extract gait features from the gait sequence and achieve SOTA results. Our approach belongs to the sequence-based method, in contrast to other methods, we propose to use variable-length frames as the input.

Local Action Modeling. Local action modeling [14,24,32, 33] aims at building short-range spatio-temporal features, which have been shown to be beneficial for gait recognition in various literature. GaitPart [14] proposed a micro-motion capture module to model the short-range spatio-temporal features. MT3D [32] proposed multiple temporal-scale 3D convolutional layers to extract the small and large temporal-scale motion features. GaitGL [33] utilized a local temporal aggregation module to extract the local temporal information. 3DLocal [24] proposed a localization module to adaptively sample the local action features. In contrast to these strategies, we propose to map each pixel of the gait sequence to the complex-valued domain, using the phase term to encode the relationship between gait features. By fully exploiting the phase term, we construct a dynamic attention model among each pixel of the feature to extract local motion patterns in key regions.

Global Action Modeling. Global action modeling, aiming at capturing long-range dependencies, has been demonstrated to be advantageous to a wide range of recognition tasks such as action recognition [6, 45, 48, 54] and person re-identification [9,16,19,35,40,61]. Many attention-based approaches [12,34,46] built global relationships in the spatial dimension [5, 50] or channel dimension [22] with remarkable results. However, current state-of-the-art works [7, 8, 32, 33] in gait recognition still directly use Max- or Mean-based methods to extract global temporal features, which only focus on the most salient features. Different from these methods, we design an efficient global self-attention model to obtain a robust representation for each person, which can select discriminative local motion patterns, and further map them to global motion patterns.

3. Methodology

In this section, we first describe the overall architecture of our method in Sec. 3.1, and then introduce the proposed two novel well-designed modules, *i.e*., Local Conv-Mixing Block (LCMB) in Sec. 3.2 and Global Motion Pattern Aggregator (GMPA) in Sec. 3.3. Finally, we will discuss the joint loss functions in Sec. 3.4.

3.1. Formulation and Motivation

Gait recognition aims to identify the same person under the influence of various external factors. Let $\mathbf{X} \in \mathbb{R}^{T\times H\times W}$ denote silhouette data containing consecutive T frames, where T, H, and W represent the temporal, height, and width dimensions of input frames. In our implementation, we sample variable-length frames $T \in [20, 40]$ from a continuous sequence as the input. The extraction of gait features can be expressed as

$$f = \mathcal{G}(\mathcal{L}(\mathbf{X})), \tag{1}$$

where $f \in \mathbb{R}^{P\times C}$ is the output features, P is the number of horizontally sliced parts, C is the feature channels, $\mathcal{L}$ represents the local motion pattern extraction, and $\mathcal{G}$ denotes the global motion pattern aggregation.

To learn distinctive representation for each person, the previous C2D-based methods [8, 20] and C3D-based methods [32, 33] only use convolutional layers and non-linear functions to learn gait features. However, the trained network can only recognize certain movement patterns that are vulnerable to noise. In this work, we propose a novel

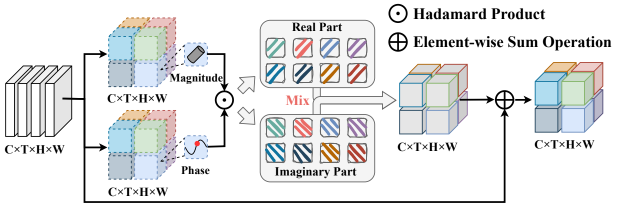


Figure 4. The architecture of LCMB, where "⊙" represents Hadamard product, "⊕" represents element-wise sum operation, and Mix denotes vector aggregation operation.

LCMB in the backbone of DANet, which enables the network to focus on the key regions and extract the local motion patterns by dynamically building the relationships among pixels. Furthermore, inspired by the Nyquist-Shannon sampling theorem [37, 39], we propose a new perspective that a complete gait pattern should contain many distinguishable local motion patterns. To this end, we developed an effective GMPA to select sufficient distinguishable local motion patterns while effectively excluding the interference of noise. The selected local motion patterns are then aggregated to generate a robust global motion pattern.

3.2. Local Conv-Mixing Block

In this section, we provide a detailed description of the vector representation and vector aggregation in the Local Conv-Mixing Block (LCMB) module.

Vector Representation. In the LCMB module, the input features are denoted as $\mathbf{V} = [v_1, v_2, ..., v_N] \in \mathbb{R}^{N \times C_i}$, where N is the number of pixels in the gait sequence, C_i is the dimension of input features. As shown in Fig. 4, we obtain the magnitude $|v_j|$ and phase θ_j of each vector by multiply with the learnable parameters $W^m \in \mathbb{R}^{C_i \times C_l}$ and $W^t \in \mathbb{R}^{C_i \times C_l}$, separately, *i.e.*,

$$|v_j| = W^m v_j, j = 1, 2, \cdots, N, \tag{2}$$

$$\theta_j = \max(0, W^t v_j), j = 1, 2, \cdots, N, \tag{3}$$

where the subscript j is the feature of the j-th pixel. The content of each vector is a real-value feature modeled by the magnitude term $|v_j|$, while the relationship of each vector is modulated by the phase term θ_j, using grouped convolutional layers with rectified linear activation ReLU. The complex vectors $\tilde{v}_j \in \mathbb{C}^{C_l}$ modulated by the magnitude and phase terms using Euler's formula, *i.e.*,

$$\tilde{v}_j = \overbrace{|v_j| \odot \cos\theta_j}^{\text{real part}} + i \overbrace{|v_j| \odot \sin\theta_j}^{\text{imaginary part}}, j = 1, 2, \cdots, N, \tag{4}$$

where i is the imaginary unit satisfying $i^2 = -1$, and $\odot$ is element-wise multiplication.

Vector Aggregation. After representing the features of each pixel as a vector, we further aggregate the local spatio-temporal domain of each vector as shown in Fig. 4. In particular, the complex-value representation of the output $\tilde{o}_j \in \mathbb{C}^{C_i}$ that aggregated by learnable convolution kernels $\mathcal{K} \in \mathbb{R}^{C_i \times C_l \times K_t \times K_s \times K_s}$, *i.e.*,

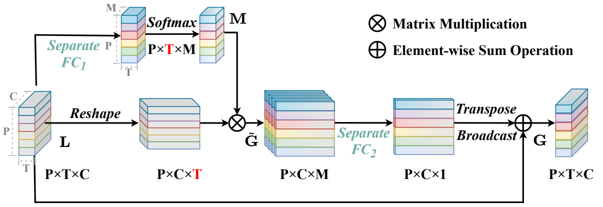


Figure 5. The architecture of GMPA and the feature maps are shown by their dimensions, where "⊗" is matrix multiplication.

$$\tilde{o}_j = \sum_{m \in \mathcal{N}(j)} \mathcal{K}[j-m] \tilde{v}_m + v_j, j = 1, 2, \cdots, N, \tag{5}$$

where $\mathcal{N}(j)$ denotes the neighboring pixels set of j, and $\tilde{v}_m$ represents the vector belong to the neighboring pixels of $\tilde{v}_j$. Following [42], we obtain the real-value output feature $o_j \in \mathbb{R}^{C_i}$ by summing the real and imaginary parts of $\tilde{v}_j$ for convenient computation, *i.e.*,

$$\begin{aligned} o_j = \sum_{m \in \mathcal{N}(j)} \big(&\mathcal{K}[j-m] \, |v_m| \odot \cos\theta_m + \\ &\mathcal{K}[j-m] \, |v_m| \odot \sin\theta_m\big) + v_j, j = 1, 2, \cdots, N, \end{aligned} \tag{6}$$

where $(\cos\theta_m + \sin\theta_m)$ denotes the dynamic attention among the neighboring pixels of j. To further comprehend the dynamic aggregation model, heatmaps showcasing the phase values are visualized in Fig. 6.

3.3. Global Motion Patterns Aggregator

In this part, we propose a new instantiation of the global action modeling framework for gait recognition named the global motion patterns aggregator (GMPA), which selects a preset number of distinguished local motion patterns and then utilizes an attention mechanism to aggregate query-specific global motion patterns of each query location. The GMPA adopts separate parameters for each part and models the global movement patterns of the corresponding part.

Lower-order Global Motion Patterns. Gait data is affected by a number of variable factors, such as segmentation holes or broken edges, further impairing the actual movement pattern. To this end, we first propose to squeeze variable local motion patterns into a preset number of channel descriptors. Then, we utilize softmax to construct an attention map in the temporal dimension, and multiply the squeezed attention maps with the reshaped local motion pattern features to obtain the global low-order motion patterns. Formally, the global low-order motion patterns $\tilde{\mathbf{G}} \in \mathbb{R}^{P \times C \times M}$ are generated by local motion patterns $\mathbf{L} \in \mathbb{R}^{P \times T \times C}$ and attention maps $\mathbf{M} \in \mathbb{R}^{P \times T \times M}$, and can be expressed as

$$\mathbf{M} = \frac{\exp(W_1 \mathbf{L}_i)}{\sum_{i=1}^{T} \exp(W_1 \mathbf{L}_i)}, \tag{7}$$

$$\tilde{\mathbf{G}} = \mathbf{M} \otimes \mathbf{L}, \tag{8}$$

where $W_1 \in \mathbb{R}^{P \times C \times M}$ is the weight of *Separate* FC_1, i is the index of frame, and $\otimes$ denotes matrix multiplication.

Higher-order Global Motion Patterns. To take advantage of the information aggregated in the low-order global motion patterns, we perform a further mapping aiming at fully capturing the high-order global motion patterns. In addition, we also introduce residual learning into GMPA to ease the training. Concretely, we further map the preset number of low-order global motion patterns $\tilde{\mathbf{G}}$ into a high-order global feature $\mathbf{G}$, *i.e.*,

$$\mathbf{G} = \delta(W_2 \tilde{\mathbf{G}}) \oplus \mathbf{L}, \tag{9}$$

where $W_2 \in \mathbb{R}^{P \times M \times 1}$ is the weight of *Separate* FC_2, δ represents the LeakyReLU activation function, and $\oplus$ denotes the broadcast element-wise addition.

3.4. Joint Loss

In this work, there are two types of loss functions involved, *i.e.*, triplet loss $\mathcal{L}_{tp}$ and cross-entropy loss $\mathcal{L}_{ce}$, which constrain the features of each part separately. Formally, triplet loss $\mathcal{L}_{tp}$ [11] can be expressed as:

$$\mathcal{L}_{tp} = \frac{1}{N_{tp}} \overbrace{\sum_{p=1}^{P}}^{\text{parts}} \overbrace{\sum_{i=1}^{S} \sum_{a=1}^{K}}^{\text{anchors}} \overbrace{\sum_{s=1}^{K}}^{\text{pos.}} \overbrace{\sum_{\substack{j=1 \\ j \neq i}}^{S} \sum_{n=1}^{K}}^{\text{neg.}} \max\big(0, m + D(\mathcal{F}(x^p_{a,i}), \mathcal{F}(x^p_{s,i})) - D(\mathcal{F}(x^p_{a,i}), \mathcal{F}(x^p_{n,j}))\big), \tag{10}$$

where P is the number of sliced parts horizontally on the gait features, N_{tp} is a positive integer obtained by multiplying the non-zero terms in triplets with the number of parts, (S, K) denotes the number of subjects with different identities and the number of samples per person, m is the margin value, $D(f_1, f_2)$ represents the euclidean distance between two features, $\mathcal{F}$ denotes the feature extraction model, and x represents the input sequence. We retain the case where the anchor and positive labels are the same, in other words, our triplet loss function requires the distance between each anchor sample and the negative sample is greater than the margin, which is an important trick for gait recognition.

Here, we propose to use the cross entropy loss $\mathcal{L}_{ce}$ to constrain each part with the label smoothing. Formally,

$$\mathcal{L}_{ce} = -\frac{1}{N_{ce}} \overbrace{\sum_{p=1}^{P}}^{\text{parts}} \overbrace{\sum_{i=1}^{S} \sum_{j=1}^{K}}^{\text{mini-batch}} \overbrace{\sum_{b=1}^{B}}^{\text{subjects}} q^{p,b}_{i,j} \log p^{p,b}_{i,j}, \tag{11}$$

where N_{ce} is a positive integer obtained by multiplying the mini-batch with the number of parts, p is the distribution of predicted probabilities, and q is the label of the mini-batch. In our experiments, the combined loss function $\mathcal{L}_c$ can be expressed as:

$$\mathcal{L}_c = \mathcal{L}_{tp} + \beta \mathcal{L}_{ce}, \tag{12}$$

where β is the hyper-parameter to balance the two terms and is set to 0.2 through the experiments.

4. Experiments

In this section, we evaluate our network on three typical gait datasets, *i.e.*, CASIA-B [56], OUMVLP [41], and Gait3D [59], and provide the implementation details. Then, we compare our approach with the current state-of-the-art methods. Finally, we conduct comprehensive ablation studies to verify the effectiveness of the proposed method.

4.1. Datasets and Implementation Details

CASIA-B [56] is a widely used dataset for gait recognition. It contains 124 subjects, 3 different walking conditions, and 11 different camera viewpoints uniformly distributed in $[0°, 180°]$. The different walking conditions include normal walking (NM), walking with bags (BG), and walking wearing a coat (CL). In summary, there are 110 sequences for each person, and each sequence has an uncertain length of frames. We take the first 74 subjects as the training set and the rest 50 subjects for the test. In the test phase, the first 4 normal walking conditions are taken as the gallery, and the rest are taken as the probe.

OUMVLP [41] has the largest number of sequences in the public gait dataset. It consists of 10,307 subjects, 2 different walking sequences ("00-01"), and 14 different camera viewpoints uniformly distributed in $[0°, 90°]$ and $[180°, 270°]$. In general, each subject contains up to 28 sequences. According to the official split way, we take 5153 subjects as the training set and the rest of 5154 as the test set. In the test phase, the sequences of "01" are taken as the gallery, and the rest of "00" is taken as the probe.

Gait3D [59] is a large-scale comprehensive dataset for gait recognition, containing silhouettes, 2D/3D human body pose, and 3D Mesh. Compared with the above datasets, Gait3D collected from more complex scenes in the wild is more challenging for the gait recognition task. It contains 4000 subjects, 25309 sequences, and 39 different camera viewpoints. Following the official splitting approach, we take 3000 subjects as the training set and the remaining 1000 subjects as the test set. In the testing phase, we calculate the similarity of one sequence of query set to all sequences in gallery set, and then report the average rank-1 and rank-5 recognition rates for all query sequences.

Table 1. The performance comparisons on CASIA-B are reported with rank-1 accuracy (%), excluding the identical-view cases. The defaults input silhouette size is 64×44 and "(*)" indicates that the input size is 128×88.

	Method	Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	134°	162°	180°	
NM	CNN-LB [52]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	GaitSet [8]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart [14]	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(*) [20]	93.2	**99.3**	**99.5**	**98.7**	96.1	**95.6**	97.2	98.1	99.3	98.6	90.1	96.9
	MT3D [32]	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 [33]	96.0	98.3	99.0	97.9	**96.9**	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	LagrangeGait [7]	95.7	98.1	99.1	98.3	96.4	95.2	97.5	99.0	99.3	98.9	94.9	97.5
	DANet(ours)	**96.4**	99.1	99.2	98.2	96.6	95.5	**97.6**	**99.4**	**99.5**	**99.3**	**96.9**	**98.0**
BG	CNN-LB [52]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	GaitSet [8]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart [14]	89.1	94.8	96.7	95.1	88.3	**94.9**	89.0	93.5	96.1	93.8	85.8	91.5
	GLN(*) [20]	91.1	**97.7**	97.8	95.2	92.5	91.2	92.4	96.0	97.5	95.0	88.1	94.0
	MT3D [32]	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 [33]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	**98.2**	96.9	91.5	94.5
	LagrangeGait [7]	94.2	96.2	96.8	95.8	94.3	89.5	91.7	96.8	98.0	97.0	90.9	94.6
	DANet(ours)	**95.0**	97.3	**98.3**	**97.4**	**94.7**	91.0	**93.9**	**97.4**	**98.2**	**97.6**	**94.2**	**95.9**
CL	CNN-LB [52]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	GaitSet [8]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart [14]	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(*) [20]	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 [32]	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 [33]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	LagrangeGait [7]	77.4	90.6	93.2	90.2	84.7	80.3	85.2	87.7	89.3	86.6	71.0	85.1
	DANet(ours)	**82.8**	**94.8**	**96.9**	**94.3**	**89.0**	**83.9**	**87.9**	**92.3**	**95.1**	**92.0**	**80.3**	**89.9**

Implementation Details. We implement our network in PyTorch [38] for all experiments. Following the preprocessing method mentioned in [8], we align and resize the input silhouettes to 64×44. During the training phase, the sampling module randomly selects [20, 40] sequences as the inputs. In the test phase, we utilize all silhouettes to obtain the gait feature. We train the model in an end-to-end manner with an optimizer of SGD and an initial learning rate of 0.1, which is reduced by a factor of 10 until convergence. The parameter l in Fig. 3 indicates the number of stages, where $l = 0$ for CASIA-B and $l = 1$ for OUMVLP and Gait3D. **(1)** In CASIA-B, the model is trained for a total of 40K iterations with the step size set every 10K iterations, using a mini-batch size of (8, 16). The convolutional channels are set to (64, 128, 256), and the stride of the temporal pooling and spatial pooling modules are set to m=3 and n=1, respectively. **(2)** In the case of OUMVLP and Gait3D, we consider that the number of sequences in OUMVLP is 20 times greater than that in CASIA-B, and the sequences in Gait3D are collected in the wild with more views. To account for these differences, we set the number of channels in four stages to (64, 128, 256, 512) and the training mini-batch size to (32, 16). The model is trained for a total of 200K iterations, with the step size set every 50K iterations. Additionally, the stride of spatial pooling is set to n=2.

4.2. Comparison with State-of-the-art Methods

To verify the effectiveness of our method, several latest gait recognition methods are introduced for comparison, including CNN-LB [52], GaitSet [8], GaitPart [14], GLN [20], MT3D [32], GaitGL [33], and LagrangeGait [7].

Evaluation on CASIA-B. The performance comparison on CASIA-B is provided in Tab. 1, where the probe sequence is divided into three subsets according to the walking conditions. **(1)** Comparing with the template-based approach, *i.e.*, CNN-LB [52], our method achieves significantly better results in all walking conditions and viewpoints. The possible reason is that the template-based approach directly compressing the gait sequence into a gait energy map would greatly compromise the temporal information in gait sequence. Once the temporal features are neglected, the motion pattern of the gait can not be adequately represented. **(2)** In contrast to the set-based methods, *i.e.*, GaitSet [8] and GLN [20], the proposed method achieves higher recog-

Table 2. Rank-1 accuracy (%) on OUMVLP under all view angles, excluding the identical-views cases.

Method	Probe View														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GaitSet [8]	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
GaitPart [14]	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
GLN [20]	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
GaitGL [33]	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7
LagrangeGait [7]	85.9	90.6	91.3	91.5	91.2	91.0	90.6	88.9	89.2	90.5	90.6	89.9	89.8	89.2	90.0
CSTL [23]	87.1	91.0	91.5	**91.8**	90.6	90.8	90.6	89.4	90.2	90.5	90.7	89.8	90.0	89.4	90.2
DANet(ours)	**87.7**	**91.3**	**91.6**	**91.8**	**91.7**	**91.4**	**91.1**	**90.4**	**90.3**	**90.7**	**90.9**	**90.5**	**90.3**	**89.9**	**90.7**

nition accuracy. This may be due to the fact that the set-based approach assumes that each silhouette contains positional information but ignores the motion patterns between sequences. We believe that motion patterns are the essential features of each person, and adequately expressing them is the key to identifying cross-views and cross-dressing conditions. **(3)** Compared to the sequence-based methods, *i.e.*, GaitPart [14], MT3D [32], GaitGL [33], and LagrangeGait [7] our method exceeds GaitPart by 11.2%, MT3D by 8.4%, GaitGL by 6.3%, and LagrangeGait by 4.8% under the most challenging cross-dressing condition. The experimental results prove the effectiveness of our method and also confirm that the combination of LCMB and GMPA modules can extract more discriminating gait features.

Evaluation on OUMVLP. To further demonstrate the effectiveness of our proposed method, as shown in Tab. 2, DANet is also evaluated on the largest gait datasets, *i.e.*, OUMVLP. **(1)** Compared with other methods, the proposed method achieves the state-of-the-art performance under all cross-view conditions. The comparison results also demonstrate that the proposed method can also effectively obtain representative and stable global motion patterns on the large-scale dataset. **(2)** It is worth noting that some subjects had no sequences corresponding to the probes in the gallery because of missing sequences, therefore the recognition accuracy in the Tab. 2 is lower than the actual performance.

Evaluation on Gait3D. The effectiveness of our method was further validated on a wild dataset, *i.e.*, Gait3D. **(1)** We observe a significant degradation in rank-1 performance of all methods, *i.e.*, GaitPart [14], GLN [20], and GaitGL [33], on the Gait3D dataset. The possible reason is that Gait3D contains more complex conditions, such as misaligned and occlusions in the silhouette data. **(2)** The experimental comparison results are illustrated in Tab. 3, which confirms that our proposed method is effective in extracting robust global motion patterns on the wild dataset.

4.3. Ablation Study

In this section, we provide an ablation study in DANet to gain a better understanding of the effect of different configurations on the LCMB and GMPA. All experiments in the ablation study are performed on CASIA-B, excluding the identical-views cases.

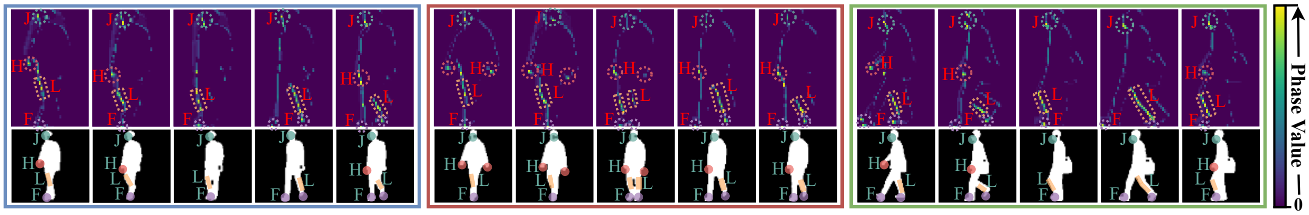
Table 3. The performance comparisons on Gait3D are reported with rank-1 accuracy(%) and rank-5 accuracy(%).

Methods	Publication	R-1(%)	R-5(%)
GaitSet [8]	AAAI 2019	36.7	58.3
GaitPart [14]	CVPR 2020	28.2	47.6
GLN [20]	ECCV 2020	31.4	52.9
GaitGL [33]	ICCV 2021	29.7	48.5
SMPLGait [59]	CVPR 2022	46.3	64.5
DANet	Ours	**48.0**	**69.7**

Table 4. The ablation study of the proposed modules on CASIA-B with rank-1 accuracy (%). Where the Conv. represents the convolutional layer, and the Aggr. denotes the aggregator.

Method	Conv.	Aggr.	NM	BG	CL
GaitSet [8]	C2D	Max	95.0	87.2	70.4
GaitPart [14]	C2D	MCM	96.2	91.5	78.7
GaitGL [33]	C3D	GeM	97.4	94.5	83.6
DANet(C2D)	C2D	GMPA	97.8	94.5	84.3
DANet(C3D)	C3D	GMPA	97.8	95.4	87.7
DANet(Max)	LCMB	Max	97.8	95.6	86.9
DANet(Mean)	LCMB	Mean	97.4	95.0	87.1
DANet(Gem)	LCMB	GeM	97.1	95.0	85.3
DANet(MCM)	LCMB	MCM	97.5	94.8	85.4
DANet	LCMB	GMPA	**97.9**	**96.2**	**89.9**

The Effectiveness of Local Conv-Mixing Block. As mentioned in Sec. 3.2, we design a novel Local Conv-Mixing Block (LCMB) module to effectively aggregate local motion patterns. **(1)** As shown in Tab. 4, compared with the benchmark approach, *i.e.*, GaitSet [8], GaitPart [14], and GaitGL [33], the LCMB-based method achieves higher accuracy under the same conditions. **(2)** With other conditions remaining the same, we directly replace our LCMB with C2D or C3D convolutional layers. As shown in Tab. 4, our LCMB is significantly higher than the C2D or C3D convo-



(a) (b) (c)

Figure 6. Visualization of attention maps (dash boxes) for phase values and corresponding positions (solid points) in the silhouettes of the same person, where "J" represents the face, "H" represents the hand, "L" represents the leg, and "F" represents the foot.

lutional methods, which demonstrates that the LCMB can effectively extract the local motion patterns of gait.

The Effectiveness of Global Motion Patterns Aggregator. To demonstrate the effectiveness of GMPA, as shown in Tab. 4, we fully compare it with other aggregators, *i.e.*, Max [8], MCM [14], GeM [33]. **(1)** The C2D and C3D convolutional layers combined with our GMPA module can significantly improve the performance. **(2)** In contrast to MCM, our GMPA focus on establishing a complete global motion pattern, which we believe is the essence of the actual motion of the gait. **(3)** Compared to Max- or Mean-based methods, the experimental results show that GMPA can effectively aggregate robust global motion patterns.

Analysis of the Number of Lower-order Global Motion Patterns. The number of low-order global motion patterns N_{patt} (M in GMPA module) selected in the GMPA module will directly affect the final feature representation. **(1)** As shown in Tab. 5, when N_{patt}=1 or 8, the network cannot fit actual movement patterns due to insufficiently sampled low-order motion patterns. **(2)** The experimental results show that when N_{patt} = 32, the network over-fits real motion pattern when the number of samples is excessive. Therefore, we adopt N_{patt}=16 as the default parameter of DANet.

Analysis of the Number of Horizontal Slice Parts. As shown in Tab. 5, we further analyze the effect of the number spatial parts N_{part} (P in the Horizontal Pooling module). **(1)** In DANet, the experimental results rise as the number of N_{part} increases. **(2)** The possible reason is that the LCMB and GMPA modules can effectively select the key regions and aggregate robust global motion patterns. As a result, the network expresses more sufficient global motion patterns when the number of N_{part} increases.

4.4. Visualization

To gain further insight into the local motion patterns of salient parts established by the phase term of the vector, as shown in Fig. 6, we visualize the heatmaps for the phase values and corresponding silhouettes of the same person. **(1)** In terms of the spatial dimension, the phase term of the vector can be effectively located in the key regions. **(2)** In terms of the temporal dimension, the phase terms of the vectors in the sequence move with the motion of the body parts, precisely localizing the disappearance and appearance of hand movements and the alternate walking of the legs. **(3)** The visualization results show that, in contrast to the fixed parameters in the CNN, the phase term of the vector can dynamically distinguish the key regions.

Table 5. Comparison of the number of spatial bins N_{part} in HP and low-order global motion patterns N_{patt} in GMPA.

N_{patt}	N_{part}	NM	BG	CL	Mean
16	16	97.5	94.4	85.7	92.5
16	32	97.7	95.6	89.3	94.2
1	64	97.8	95.6	86.9	93.4
8	64	97.4	95.0	87.1	93.1
32	64	97.5	94.7	88.5	93.6
16	64	**97.9**	**96.2**	**89.9**	**94.6**

5. Conclusion

In this paper, we propose a novel Dynamic Aggregated (DANet) for gait recognition, which consists of a serial of Local Conv-Mixing Block (LCMB) and Global Motion Pattern Aggregator (GMPA) to adaptively aggregate the robust discriminative global motion patterns. The proposed method can dynamically locate the key regions and extract the local motion patterns, and then adaptively select the distinguishing local motion patterns to further construct robust global motion patterns. The experimental results on three popular gait datasets, *i.e.*, CASIA-B, OUMVLP, and Gait3D, verify the effectiveness of the proposed method and show great potential for practical applications. In the future, we will further investigate adaptive learning of the local and global motion patterns in the complex-valued domain to aggregate more representative gait features.

6. Acknowledgement

This work was supported by National Natural Science Foundation of China (Grants No.62171038, No.61827901, No.62088101, No.62276031, and No.62276025).

References

[1] Gunawan Ariyanto and Mark S Nixon. Model-based 3d gait biometrics. In *international joint conference on biometrics*, 2011. 2

[2] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, and Xilin Chen. Salient-to-broad transition for video person re-identification. In *CVPR*, 2022. 1

[3] Robert Bodor, Andrew Drenner, Duc Fehr, Osama Masoud, and Nikolaos Papanikolopoulos. View-independent human motion classification using image-based reconstruction. *Image and Vision Computing*, 2009. 2

[4] Imed Bouchrika, Michaela Goffredo, John Carter, and Mark Nixon. On using gait in forensic biometrics. *Journal of forensic sciences*, 2011. 1

[5] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *ICCV*, 2019. 3

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[7] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang. Lagrange motion analysis and view embeddings for improved gait recognition. In *CVPR*, 2022. 2, 3, 6, 7

[8] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, 2019. 2, 3, 6, 7, 8

[9] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *CVPR*, 2019. 3

[10] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 2

[11] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *ECCV*, 2018. 5

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[13] Huanzhang Dou, Pengyi Zhang, Yuhan Zhao, Lin Dong, Zequn Qin, and Xi Li. Gaitmpl: Gait recognition with memory-augmented progressive learning. *IEEE TIP*, 2022. 2

[14] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, 2020. 2, 3, 6, 7, 8

[15] Yang Feng, Yuncheng Li, and Jiebo Luo. Learning effective gait features using lstm. In *ICPR*, 2016. 1

[16] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019. 3

[17] Hongji Guo, Hanjing Wang, and Qiang Ji. Uncertainty-guided probabilistic transformer for complex action recognition. In *CVPR*, 2022. 1

[18] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE TPAMI*, 2005. 2, 3

[19] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, 2021. 3

[20] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, 2020. 2, 3, 6, 7

[21] Saihui Hou, Xu Liu, Chunshui Cao, and Yongzhen Huang. Set residual network for silhouette-based gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021. 2, 3

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[23] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *ICCV*, 2021. 3, 7

[24] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3d local convolutional neural networks for gait recognition. In *CVPR*, 2021. 2, 3

[25] Haruyuki Iwama, Daigo Muramatsu, Yasushi Makihara, and Yasushi Yagi. Gait verification system for criminal investigation. *Information and Media Technologies*, 2013. 1

[26] Pakorn KaewTraKulPong and Richard Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-based surveillance systems*. Springer, 2002. 1

[27] Worapan Kusakunniran, Qiang Wu, Hongdong Li, and Jian Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *ICCV*, 2009. 2

[28] Peter K Larsen, Erik B Simonsen, and Niels Lynnerup. Gait analysis in forensic medicine. *Journal of forensic sciences*, 2008. 1

[29] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. End-to-end model-based gait recognition using synchronized multi-view pose constraint. In *ICCV*, 2021. 2

[30] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *ACCV*, 2020. 2

[31] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *PR*, 2020. 2

[32] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *ACM MM*, 2020. 2, 3, 6, 7

[33] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, 2021. 2, 3, 6, 7, 8

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3

[35] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR*, 2019. 3

[36] Yasushi Makihara, Atsuyuki Suzuki, Daigo Muramatsu, Xiang Li, and Yasushi Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *CVPR*, 2017. 3

[37] Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 1928. 2, 4

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 6

[39] Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 1949. 2, 4

[40] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 3

[41] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 2018. 2, 5

[42] Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Yanxi Li, Chao Xu, and Yunhe Wang. An image patch is a wave: Phase-aware vision mlp. In *CVPR*, 2022. 4

[43] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE TPAMI*, 29, 2007. 2, 3

[44] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: graph convolutional network for skeleton-based gait recognition. In *ICIP*, 2021. 2

[45] Thanh-Dat Truong, Quoc-Huy Bui, Chi Nhan Duong, Han-Seok Seo, Son Lam Phung, Xin Li, and Khoa Luu. Direcformer: A directed attention in transformer approach to robust action recognition. In *CVPR*, 2022. 3

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3

[47] Chen Wang, Junping Zhang, Liang Wang, Jian Pu, and Xiaoru Yuan. Human identification using temporal information preserving gait template. *IEEE TPAMI*, 2012. 2, 3

[48] Jue Wang and Lorenzo Torresani. Deformable video transformer. In *CVPR*, 2022. 3

[49] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE TPAMI*, 2003. 1

[50] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[51] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *ICIP*, 2016. 2, 3

[52] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE TPAMI*, 2016. 2, 3, 6

[53] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *CVPR*, 2022. 1

[54] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *CVPR*, 2022. 1, 3

[55] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE TIP*, 2019. 1

[56] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, 2006. 2, 5

[57] Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, and Hongdong Li. Learning joint gait representation via quintuplet loss minimization. In *CVPR*, 2019. 3

[58] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, 2019. 2

[59] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *CVPR*, 2022. 2, 5, 7

[60] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Reranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 1

[61] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *CVPR*, June 2022. 1, 3

[62] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *ICPR*, 2004. 1

[63] Zoran Zivkovic and Ferdinand Van Der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 2006. 1