2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10221893

GAITMM: MULTI-GRANULARITY MOTION SEQUENCE LEARNING FOR GAIT RECOGNITION

Lei Wang *Bo Liu** *Bincheng Wang* *Fuqiang Yu*

School of Information Science and Technology, Hebei Agricultural University, China

ABSTRACT

Gait recognition aims to identify individual-specific walking patterns by observing the different periodic movements of each body part. However, most existing methods treat each part equally and fail to account for the data redundancy caused by the different step frequencies and sampling rates of gait sequences. In this study, we propose a multi-granularity motion representation network (GaitMM) for gait sequence learning. In GaitMM, we design a combined full-body and fine-grained sequence learning module (FFSL) to explore part-independent spatio-temporal representations. Moreover, we utilize a frame-wise compression strategy, referred to as multi-scale motion aggregation (MSMA), to capture discriminative information in the gait sequence. Experiments on two public datasets, CASIA-B and OUMVLP, show that our approach reaches state-of-the-art performances.

Index Terms— Gait Recognition, Multi-Granularity Motion Representation, Multi-Scale Motion Aggregation

1. INTRODUCTION

Gait recognition has emerged as a promising biometric technology that leverages human gait information for long-distance identification without the cooperation of subjects. This technology has shown great potential in many fields, including video surveillance, rail transit, and sports simulation. However, gait recognition performance is often affected by various factors in real-world scenarios, such as changing viewpoints [1], occlusion [2], and different wearing conditions [3, 4]. Therefore, learning gait representations that are invariant to these factors is a major challenge for gait recognition.

Most gait recognition methods utilize (convolutional neural networks) CNNs to extract spatio-temporal information from gait sequences. They can be categorized as set-based or sequence-based, depending on whether they consider the temporal order of frames. Set-based methods treat a gait sequence as an unordered set, which can either be compressed into a single gait template [5] or learn order-independent gait representations from silhouette sets [6, 7]. Although the ordering of inputs is not essential for gait assessment in these methods, they may ignore the temporal nature of the gait sequence, resulting in the loss of discriminative local motion information.

Sequence-based methods tend to explore individual gait patterns from multiple spatial and temporal scales [8, 9, 10, 11]. As the input sequences are usually aligned, a uniform horizontal division of intermediate layer features can improve recognition performance [9, 10]. Another approach introduces a body part-level localization module to achieve a more adaptive local representation [11]. However, localization errors caused by changes in wear conditions or movement amplitude may degrade recognition accuracy. Additionally, the redundancy of adjacent frames limits the recognition of spatio-temporal variation patterns. While some methods [10, 12] have been proposed to aggregate local clips, they may lack adaptability to motion aggregation.

To address the issues mentioned above, we propose a multi-scale motion learning framework named GaitMM for cross-view gait recognition. GaitMM comprises two main components: a combined full-body and fine-grained sequence learning module (FFSL) and a multi-scale motion aggregation (MSMA) operation. Rather than using shared convolutional kernels to extract part-specific features, in FFSL, the fine-grained motion patterns are independently obtained from body-part sequences. To reduce the redundancy of adjacent frames, MSMA compresses a sequence by aggregating information in each local clip. The contributions of our work are summarised as follows:

1) We propose a gait recognition framework named GaitMM, which combines global and fine-grained motion information for gait sequence learning.

2) We propose an adaptive MSMA module that reduces redundancy in the gait sequence.

3) Experimental results on two public datasets, CASIA-B and OUMVLP, demonstrate that our method achieves state-of-the-art performance.

2. RELATED WORK

According to order sensitivity, there are two main categories of gait recognition techniques: set-based and sequence-based. In set-based approaches, gait silhouettes are typically considered as an unordered set, from which a set-level representa-

* Bo Liu is the corresponding author.

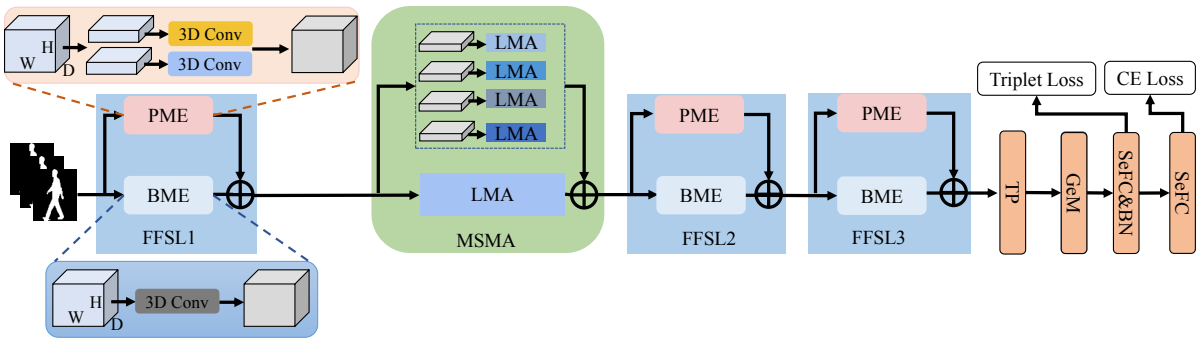

Fig. 1. Overview of GaitMM. The spatio-temporal dimensions of the feature map, i.e., D, H and W, are indicated in the figure, and we omit the channel dimension C for simplicity. The $\oplus$ represents the element-wise summation operation, and the TP represents the temporal pooling operation. The SeFC represents the separable fully connected layer.

tion is obtained by characterizing the complementarity of the silhouettes in the set [5, 6, 7, 13, 14]. A straightforward way to handle a set of silhouettes is to compress them into a single template, i.e., gait energy image (GEI), allowing the feature extraction and matching processes to be performed at the image level [5]. However, these template-based methods largely ignore the spatial and temporal properties during preprocessing. In order to maximally preserve the set information, some methods take the raw silhouettes as inputs [6, 7, 14]. Chao et al. [6] first propose a set-based gait recognition framework named GaitSet, which employs a max-pooling function to learn a permutation-invariant representation of a set. Hou et al. [7] further propose a lateral connection to fuse silhouette-level and set-level features. While the methods above provide flexibility by dropping the sequential constraints, the temporal cues are also essential for revealing subtle gait changes.

Sequence-based approaches emphasize continuous pose variations, aggregating multi-scale motion features associated with body models or silhouettes. The model-based methods extract geometric and dynamic gait features from human motion models[15, 16]. However, these approaches suffer from performance degradation caused by the inaccurate pose estimation results from low-resolution conditions. The silhouette-based methods usually extract spatio-temporal gait information from the sequence [8, 17, 9, 10, 11]. To capture the various temporal cues in the sequence, some researchers considered extracting gait information from multiple temporal scales [8, 17]. For example, Lin et al. [8] develop large and small temporal scales feature extractors for gait sequences using the designed 3D basic network blocks. Huang et al. [17] explore the temporal features at three scales: frame-level, short-term and long-term. However, these methods insufficiently consider the motion differences among body parts. Therefore, some studies horizontally divide the silhouette into several parts and extract part-specific features [9, 10, 12]. Moreover, Huang et al. [11] propose 3D local operations to extract 3D volumes of body parts. Nevertheless, some irregular gait patterns (such as wearing a coat) may affect the localization accuracy and reduce the recognition accuracy.

3. METHOD

This section outlines the framework of our proposed method and describes several components, including the full-body and fine-grained sequence learning module (FFSL) and the multi-scale motion aggregation (MSMA) operation.

3.1. Our Framework

In GaitMM, multiple FFSL modules are stacked to learn gait motion features, and an MSMA module is available for frame-level downsampling. The whole pipeline is illustrated in Fig. 1. Given a gait sequence $\mathcal{S} \in \mathbb{R}^{C_{in} \times D \times H \times W}$, where D means the number of frames, (H, W) is the image size of each frame, C denotes the number of input channels. First, we feed $\mathcal{S}$ into GaitMM. Next, after frame compression by the MSMA module, the output feature map $\mathcal{F}_{FFSL3} \in \mathbb{R}^{C_{out} \times \frac{D}{3} \times H \times W}$ of the third FFSL modoule (FFSL3) is mapped to the discriminative space via temporal pooling (TP) [8, 10] and generalized mean pooling (GeM) [10, 12] operations. Finally, we train the model using a combination of triplet and cross-entropy losses, which are commonly used for gait recognition [10, 7, 11, 18].

3.2. Full-body and Fine-grained Sequence Learning

The proposed FFSL module consists of a body-level motion feature extractor (BME) and a part-level motion feature extractor (PME). Specifically, the BME is implemented through a 3D convolution. Meanwhile, the PME learns part-independent spatio-temporal representations using non-shared 3D convolution filters that can account for diverse movement patterns of different body parts. For a input gait sequence $\mathcal{S}$, the process of BME can be formulated as:

$$\mathcal{F}_{BME} = \mathrm{3DConv}_{3\times3\times3}(\mathcal{S}), \quad (1)$$

Table 1. Rank-1 accuracy (%) on CASIA-B under all views and different conditions, excluding identical-view cases.

Gallery NM #1-4			$0^\circ - 180^\circ$											Mean
Probe			0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
LT	NM #5-6	GaitSet[6]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
		GaitPart[9]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
		GaitGL[10]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	**99.3**	98.8	94.0	97.4
		3D Local[11]	96.0	**99.0**	**99.5**	**98.9**	**97.1**	94.2	96.3	99.0	98.8	98.5	95.2	97.5
		LagrangeGait[18]	95.7	98.1	99.1	98.3	96.4	95.2	97.5	99.0	**99.3**	98.9	94.9	97.5
		Ours	**97.2**	98.6	99.2	98.1	97.0	**95.7**	**97.8**	**99.1**	**99.3**	**99.3**	**96.6**	**98.0**
	BG #1-2	GaitSet[6]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
		GaitPart[9]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
		GaitGL[10]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
		3D Local[11]	92.9	95.9	**97.8**	**96.2**	93.0	87.8	92.7	96.3	97.9	**98.0**	88.5	94.3
		LagrangeGait[18]	94.2	96.2	96.8	95.8	94.3	89.5	91.7	96.8	98.0	97.0	90.9	94.6
		Ours	**94.9**	**97.1**	97.6	96.1	**94.6**	**91.2**	**93.6**	**97.4**	**98.3**	97.0	**93.3**	**95.6**
	CL #1-2	GaitSet[6]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
		GaitPart[9]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
		GaitGL[10]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
		3D Local[11]	78.2	90.2	92.0	87.1	83.0	76.8	83.1	86.6	86.8	84.1	70.9	83.7
		LagrangeGait[18]	77.4	90.6	93.2	90.2	84.7	80.3	85.2	87.7	89.3	86.6	71.0	85.1
		Ours	**81.3**	**91.4**	**93.7**	**90.7**	**86.8**	**83.3**	**86.2**	**89.0**	**91.8**	**87.8**	**76.9**	**87.2**

where $\mathrm{3DConv}_{3\times3\times3}(\cdot)$ denotes a 3D convolution with a convolution kernel size of $3\times3\times3$, $\mathcal{F}_{BME} \in \mathbb{R}^{C_{out}\times D\times H\times W}$ is the output of the BME. For PME, the input $\mathcal{S}$ is evenly divided into k parts along the horizontal axis, which are denoted as $\mathcal{S}^j, j \in 1, 2, 3, \cdots, k$, where $\mathcal{S}^j \in \mathbb{R}^{C_{in}\times D\times \frac{H}{k}\times W}$. The PME process for j-th part sequence is written as:

$$\mathcal{F}^j_{PME} = \mathrm{3DConv}_{3\times3\times3}\left(\mathcal{S}^j\right), \tag{2}$$

where $\mathcal{F}^j_{PME} \in \mathbb{R}^{C_{out}\times D\times \frac{H}{k}\times W}$ is the output feature map. Note that each part sequence undergoes a separate 3D convolution, ensuring independence and diversity of the learned spatio-temporal representations. Next, these part-level feature maps are concatenated along the horizontal axis, which can be formulated as:

$$\mathcal{F}_{PME} = \mathcal{F}^1_{PME} © \mathcal{F}^2_{PME} \cdots © \mathcal{F}^k_{PME}, \tag{3}$$

where © represents the concatenation operation , $\mathcal{F}_{PME} \in \mathbb{R}^{C_{out}\times D\times H\times W}$ is the output of PME. The output of FFSL is obtained by fusing $\mathcal{F}_{BME}$ and $\mathcal{F}_{PME}$ with an element-wise summation, which can be expressed as:

$$\mathcal{F}_{FFSL} = \mathcal{F}_{BME} + \mathcal{F}_{PME}. \tag{4}$$

3.3. Multi-scale Motion Aggregation

MSMA is employed to reduce data redundancy and enhance the discriminability of motions. It consists of two parallel branches, as shown in Fig. 1. Each branch is based on a local motion aggregation (LMA) operation, which is designed to perform temporal-downsampling for each gait sequence. The part branch uses l separate LMAs for body parts to preserve distinctive movement patterns, while the global branch employs a body-level LMA to compress temporal information. The LMA can be formulated as:

$$\mathcal{F}_{LMA} = p_1 \mathrm{Max}_{3\times1\times1}(\mathcal{F}) + p_2 \mathrm{Avg}_{3\times1\times1}(\mathcal{F}), \tag{5}$$

where $\mathrm{Max}_{3\times1\times1}(\cdot)$ denotes max pooling operation with kernel size $(3\times1\times1)$, $\mathrm{Mean}_{t\times1\times1}(\cdot)$ denotes average pooling operation with kernel size $(3\times1\times1)$. $\mathcal{F} \in \mathbb{R}^{C_{in}\times D\times H\times W}$ and $\mathcal{F}_{LMA} \in \mathbb{R}^{C_2\times\frac{D}{3}\times H\times W}$ are the input and output of LMA, respectively. The p_1 and p_2 are two learnable parameters.

4. EXPERIMENTS

4.1. Datasets and Implementation Details

CASIA-B. The widely-used CASIA-B dataset [1] includes gait data for 124 subjects, captured from 11 camera views at regular intervals. Each view includes six normal walking (NM) sequences, as well as two sequences each of walking with a bag (BG) and walking with a coat (CL), resulting in a total of ten sequences per subject. Experiments in this study follow the large-sample training (LT) protocol [6], in which the first 74 subjects are used for training and the remaining 50 for testing. During testing, NM#01-04 sequences are used as the gallery, and NM#05-06, BG#01-02, and CL#01-02 sequences are used as the probe for evaluation.

OUMVLP. The OUMVLP dataset [19] is a large gait dataset, consisting of 10307 subjects. Each subject is captured at 14 camera views with a sampling interval of 15°, and each view

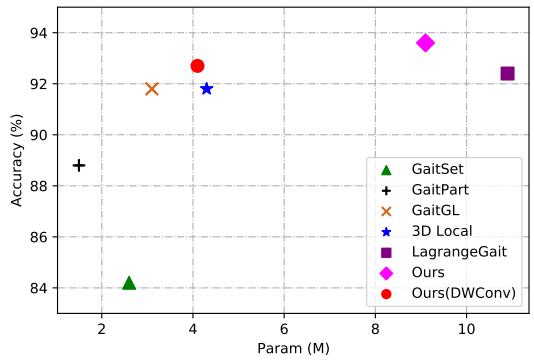

Fig. 2. The trade-off between accuracy and parameters of our method and other comparison methods on CASIA-B.

includes two groups of sequences. Following the protocol in [6], 5153 subjects are used for training and the remaining 5154 subjects for testing. During the testing phase, the sequences (Seq#01) are regarded as the gallery, while the sequences (Seq#00) are treated as the probe for evaluation.

Implementation Details. The k in Equ. 3 and the l in MSMA are both set to 8. The p_1 and p_2 in Equ. 5 are initialized to 0.5. The δ in GeM is initialized to 6.5, and the margin β of the triplet loss is set to 0.2. The number of FFSL is set to 3 for CASIA-B and double for OUMVLP. The gait silhouettes are aligned as [19] and the silhouette images uniformly crop to a size of 64×44. The batch size $(P \times K)$ is set to $(8, 8)$ on CASIA-B and $(32, 8)$ on OUMVLP. During training, the number of frames D is set to 30, and the model uses a Adam optimizer with the initial learning rate of $1e$-4. For CASIA-B, the number of iterations is 80K, and the learning rate reset to 1e-5 after 70K. For OUMVLP, the number of iterations is 160K, the learning rate reset to $1e$-5 after 150K iterations.

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

CASIA-B. Tab. 1 shows the performance comparison of our proposed GaitMM with the state-of-the-art (SOTA) methods on CASIA-B. Our approach achieves mean view recognition accuracies of 98.0%, 95.6% and 87.2% for the NM, BG and CL walking conditions, respectively, which are 0.5%, 1.0% and 2.1% higher than LagrangeGait [18], demonstrating the superiority of GaitMM in cross-view recognition. However, the operation of independent feature extraction in FFSL increases the number of parameters. To address this issue, we replace the 3D convolution in PME with the 3D depthwise separable convolution (DWConv) [20]. As shown in Fig. 2, we can observe that the proposed methods, especially the DWConv version, achieve a better trade-off between model size and accuracy.

OUMVLP. Tab. 2 presents the rank-1 accuracy of GaitMM evaluated on OUMVLP compared to several SOTA methods. The results demonstrate that our proposed method outperforms the current methods in all views, highlighting the generalization capability of GaitMM.

4.3. Ablation Study

The effects of FFSL and MSMA are shown in Tab. 3. Removing both PME and MSMA leads to a decrease in performance. FFSL is necessary for accurately modeling spatial scale information and capturing motion relationships between body parts, while MSMA is important for extracting discriminative temporal clues while compressing gait sequences.

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

	Gallery All 14 views				
Probe	GaitSet[6]	GaitPart[9]	GaitGL[10]	3D Local[11]	**Ours**
0°	84.5	88.0	90.5	-	**92.9**
15°	93.3	94.7	96.1	-	**97.1**
30°	96.7	97.7	98.0	-	**98.4**
45°	96.6	97.7	98.1	-	**98.4**
60°	93.5	95.5	97.0	-	**97.5**
75°	95.3	96.6	97.6	-	**98.0**
90°	94.2	96.2	97.1	-	**97.7**
180°	87.0	90.6	94.2	-	**95.8**
195°	92.5	94.2	94.9	-	**96.3**
210°	96.0	97.2	97.4	-	**97.8**
225°	96.0	97.1	97.4	-	**97.8**
240°	93.0	95.1	95.7	-	**96.4**
255°	94.3	96.0	96.5	-	**97.1**
270°	92.7	95.0	95.7	-	**96.6**
Mean	93.3	95.1	96.2	96.5	**97.0**

Table 3. Ablation study on FFSL and MSMA.

FFSL		MSMA	Rank-1 Accuracy			
BME	PME		NM	BG	CL	Mean
✓			97.1	94.4	84.1	91.9
✓	✓		97.8	95.2	85.2	92.7
✓		✓	97.1	94.0	85.1	92.1
✓	✓	✓	**98.0**	**95.6**	**87.2**	**93.6**

5. CONCLUSION

This paper proposes GaitMM, a novel gait recognition framework that integrates fine-grained and global motion properties. The FFSL module is designed to learn the part-based sequence and body representations, while the MSMA operation aggregates sequence information by compressing redundant frames. We conduct extensive experiments on two public datasets to demonstrate the effectiveness of GaitMM.

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61972132, 62106065) and the Research Project for Self-cultivating Talents at Hebei Agricultural University (Grant No. PY201810).

7. REFERENCES

[1] Shiqi Yu, Daoliang Tan, and Tieniu Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th International Conference on Pattern Recognition (ICPR)*, 2006, vol. 4, pp. 441–444.

[2] Imad Rida, Noor Almaadeed, and Somaya Almaadeed, "Robust gait recognition: a comprehensive survey," *IET Biometrics*, vol. 8, no. 1, pp. 14–28, 2019.

[3] TzeWei Yeoh, Hernán E Aguirre, and Kiyoshi Tanaka, "Clothing-invariant gait recognition using convolutional neural network," in *2016 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2016, pp. 1–5.

[4] Lingxiang Yao, Worapan Kusakunniran, Qiang Wu, Jingsong Xu, and Jian Zhang, "Collaborative feature learning for gait recognition under cloth changes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 3615–3629, 2021.

[5] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in *ICB*, 2016, pp. 1–8.

[6] Hanqing Chao, Kun Wang, Yiwei He, Junping Zhang, and Jianfeng Feng, "Gaitset: Cross-view gait recognition through utilizing gait as a deep set," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 3467–3478, 2021.

[7] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 382–398.

[8] Beibei Lin, Shunli Zhang, and Feng Bao, "Gait recognition with multiple-temporal-scale 3d convolutional neural network," in *Proceedings of the 28th ACM International conference on Multimedia*, 2020, pp. 3054–3062.

[9] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He, "Gaitpart: Temporal part-based model for gait recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020, pp. 14225–14233.

[10] Beibei Lin, Shunli Zhang, and Xin Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14648–14656.

[11] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14920–14929.

[12] Beibei Lin, Yu Liu, and Shunli Zhang, "Gaitmask: Mask-based model for gait recognition," in *32nd British Machine Vision Conference (BMVC)*, 2021, pp. 1–12.

[13] 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*, vol. 3, no. 3, pp. 384–393, 2021.

[14] Saihui Hou, Xu Liu, Chunshui Cao, and Yongzhen Huang, "Gait quality aware network: Toward the interpretability of silhouette-based gait recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[15] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, pp. 107069, 2020.

[16] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren, "End-to-end model-based gait recognition," in *Proceedings of the Asian conference on computer vision (ACCV)*, 2020, pp. 3–20.

[17] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12909–12918.

[18] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang, "Lagrange motion analysis and view embeddings for improved gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20249–20258.

[19] 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*, vol. 10, pp. 1–14, 2018.

[20] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.