COMBINING THE SILHOUETTE AND SKELETON DATA FOR GAIT RECOGNITION

Likai Wang   *Ruize Han** *   *Wei Feng*

College of Intelligence and Computing, Tianjin University, Tianjin, China

ABSTRACT

Gait recognition, a long-distance biometric technology, has aroused intense interest recently. Currently, the two dominant gait recognition works are appearance-based and model-based, which extract features from silhouettes and skeletons, respectively. However, appearance-based methods are greatly affected by clothes-changing and carrying conditions, while model-based methods are limited by the accuracy of pose estimation. To tackle this challenge, a simple yet effective two-branch network is proposed in this paper, which contains a CNN-based branch taking silhouettes as input and a GCN-based branch taking skeletons as input. In addition, for better gait representation in the GCN-based branch, we present a fully connected graph convolution operator to integrate multi-scale graph convolutions and alleviate the dependence on natural joint connections. Also, we deploy a multi-dimension attention module named STC-Att to learn spatial, temporal and channel-wise attention simultaneously. The experimental results on CASIA-B and OUMVLP show that our method achieves state-of-the-art performance in various conditions.

Index Terms— Gait recognition, two-branch neural network, graph convolution, convolutional neural network

1. INTRODUCTION

Gait recognition is a research topic for systematic study on human motion, which recently attracts increasing interest in computer vision. Due to the fact that gait can be recognized at a long-distance without the cooperation of subjects while other biometric characteristics cannot, gait recognition has considerable prospect for many practical applications, e.g., video surveillance, crime investigation and social security.

Currently, the two dominant gait recognition works are appearance-based and model-based. The first extracts features from silhouettes using CNNs [1] or transformer network [2, 3], which either compresses all silhouettes into one gait energy image (GEI) [4, 5] or regards gait as silhouette sequences [1, 6, 7, 8, 9]. However, such methods that rely on the human body shape are extremely sensitive to carrying, clothing and view-changing since these variations alter the human appearance drastically. In contrast to the above,

the second category takes raw skeleton data obtained by pose estimation algorithms [10] as input and performs feature extraction through CNNs [11, 12] or GCNs [13, 14]. Although such methods can overcome the interference caused by occlusion and view-changing to a certain extent, the accuracy of pose estimation methods and the lack of information contained in skeleton data bring limitations to the improvement of gait recognition accuracy. Recently, the work in [15] has attempted to combine these two methods, however, complex models are used with weak performance improvement.

Motivated by the above analysis, we propose a simple yet effective two-branch network for gait recognition task to realize the reciprocity of appearance- and model-based methods, which can significantly promote the enhancement of recognition accuracy. Specifically, the appearance-based branch leverages stacked CNNs to extract discriminative representations from a sequence of gait silhouettes. The model-based branch deploys GCNs to learn multi-order robust and compact representations from human skeleton data.

In addition, considering that existing GCN-based methods have not been well studied and have relatively low accuracy compared to appearance-based ones, we further design two modules for representing better skeleton feature in the GCN-based branch. Specifically, previous GCN-based methods [16, 13] learn the movement patterns of human joint from a spatial-temporal graph with joints as nodes and bones as edges. However, it is indisputable that the feature extractor should not only aggregate the information of directly connected joints, but also extract multi-scale structural features and long-range dependencies on account of strong correlations between joints that are physically apart. To achieve this, existing methods [16] apply *multiple adjacency matrices with different orders* for various joint connections, which suffers from the balance problem between training difficulty and the effective fusion of important information from distant joints, i.e., the *distant-joint connection problem*. To address this problem, we present a fully-connected graph structure without natural joint connections, to adaptively integrate multi-scale graph convolutions instead of the redundant adjacency matrixes for distant joints. Moreover, inspired by the effectiveness of the attention mechanism demonstrated in recent appearance-based works [17, 18], we propose an attention module named STC-Att in the model-based branch. Extensive experiments indicate that our model achieves state-of-

the-art performance in various conditions.

The main contributions of the proposed method are summarized as follows: (1) We propose a new two-branch framework for gait recognition task, which combines appearance- and model-based features to obtain their complementary advantages. (2) In the model-based branch, we propose an effective fully connected graph convolution operator, which can adaptively capture the dependencies between all nodes without multi-scale convolutions. Also, for more discriminative representations from raw skeleton data, we propose a multi-dimension attention module named STC-Att to learn the spatial, temporal and channel-wise attention simultaneously. (3) Experimental results demonstrate that our method achieves superior performance, especially on the cloth-changing condition, than the prior state-of-the-art methods.

2. PROPOSED METHOD

2.1. A Two-branch Framework

In this paper, we design a two-branch network to combine the appearance- and model-based methods, making the silhouettes and skeletons to be complementary with each other.

The architecture of the model-based branch is shown in Figure 1. As shown, we use the spatial-temporal graph convolutional (STGC) blocks [19] as the basic component of the model. In the STGC block, we apply the proposed fully-connected graph convolution (which will be discussed in Section 2.2) to capture the arbitrary-hop joint dependencies. The proposed STC-Att module (which will be discussed in Section 2.3) is employed after each STGC block to explicitly model the channel-, spatial- and temporal-wise correlations between feature maps. Finally, a global average pooling layer and a softmax classifier are utilized at the end of the model. In the appearance-based branch, we directly choose a state-of-the-art part-based model GaitPart [20] as the feature extractor.
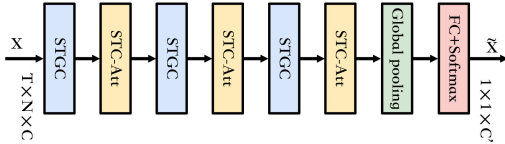

Fig. 1. Architecture of the model-based branch.

Suppose that the output feature vector of the model-based branch is $\mathbf{f}_{\mathrm{m}}$ and of the appearance-based branch is $\mathbf{f}_{\mathrm{a}}$, a straightforward concatenation operation is adopted to merge the two feature vectors together, which brings out the final gait descriptors for recognition. Note that because the two vectors, that are obtained from completely different raw data using diverse networks, have unequal vector size and value range, a ratio λ is needed to adjust the rate of them. Hence, we can get the overall feature vector through $\mathbf{f} = \mathbf{f}_{\mathrm{m}} \oplus \lambda \cdot \mathbf{f}_{\mathrm{a}}$, where $\oplus$ means the concatenation operation.
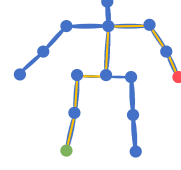

Fig. 2. An illustration of the human joints. The red vertex denotes right wrist, the green vertex denotes left ankle, and the yellow lines denote the smallest hop between them.

2.2. Fully Connected Graph Convolution

In current GCN-based works [19], to learn long-range dependencies, multi-scale graph convolution with multiple adjacency matrices are used to aggregate information of distant neighbors. The graph convolution can be formulated as

$$\mathbf{f}_{\mathrm{out}} = \sigma \left(\sum_k \boldsymbol{\Lambda}_k^{-\frac{1}{2}} \mathbf{A}_k \boldsymbol{\Lambda}_k^{-\frac{1}{2}} \mathbf{f}_{\mathrm{in}} \mathbf{W}_k \right), \qquad (1)$$

where $\mathbf{f}_{\mathrm{in}} \in \mathbb{R}^{N \times C_1}$ and $\mathbf{f}_{\mathrm{out}} \in \mathbb{R}^{N \times C_2}$ denote the input and output feature maps on each frame, respectively. Here C_1/C_2 and N denote the number of input/output feature channels and joints, respectively. Besides, $\mathbf{A}_k \in \{0,1\}^{N \times N}$ denotes the k-order adjacency matrix representing intra-frame connections, which is defined as

$$\mathbf{A}_k^{i,j} = \begin{cases} 1, & \text{if} \quad i = j \\ 1, & \text{if} \quad d(v_i, v_j) = k \\ 0, & \text{otherwise} \end{cases}, \qquad (2)$$

where $d(v_i, v_j)$ denotes the smallest hop between nodes v_i and v_j. $\boldsymbol{\Lambda}_k = \mathrm{diag}(\sum_j \mathbf{A}_k^{ij}) \in \mathbb{R}^{N \times N}$ is the degree matrix with the diagonal elements $\boldsymbol{\Lambda}_k^{ii} = \sum_j \mathbf{A}_k^{ij}$ and others as zeros, $\mathbf{W}_k \in \mathbb{R}^{C_1 \times C_2}$ is the learnable weight matrix and $\sigma(\cdot)$ is an activation function.

However, such method is suboptimal. First, it need multiple matrices for modeling the joint dependence relation. Take the right wrist (the red vertex in Figure 2) for example, if all the other joints are taken into account, eight adjacency matrices are needed since the smallest hop between the right wrist and the farthest node, i.e., the left ankle (see the green vertex in Figure 2), is seven. Second, we can see that all the adjacency matrixes include the self-connection (i.e., $i = j$), which makes the information of the right wrist itself will be aggregated eight times throughout the convolutional process, while other joints for only once. This may lead to a bias towards the local joint and the attenuation of other joint dependencies. We argue that such factorized modeling is dispensable and the weight of all joints can be learned simultaneously.

To address the problems, we propose a fully connected graph convolution operator to integrate multi-scale convolutions and balance the contributions from the joint itself and other joints. The operator no longer relies on the physical connections. Specifically, we define the adjacency matrix $\mathbf{A}$
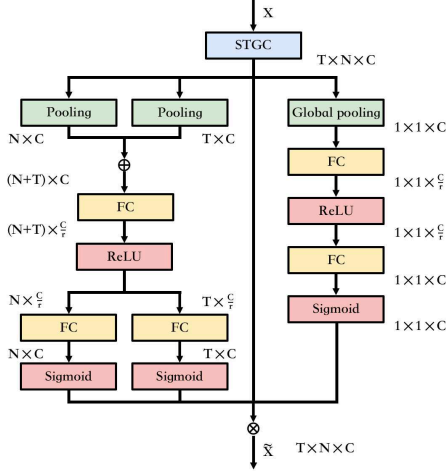
Fig. 3. Schema of the proposed STC-Att module.

as $\{\mathbf{A}^{ij} = 1 \mid i, j = 1, 2, \ldots, N\}$, that is, each pair of nodes of the skeleton graph is interconnected. We substitute $\mathbf{A}_k$ with $\mathbf{A}$ in Equation 1 and obtain

$$\mathbf{f}_{\text{out}} = \sigma\left(\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{A}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{f}_{\text{in}}\mathbf{W}\right), \quad \mathbf{\Lambda} = \text{diag}(\sum_j \mathbf{A}^{ij}). \quad (3)$$

Compared with multi-scale graph convolution where neighbors are divided into multiple scales based on the number of hops, the proposed fully connected convolution operator treats all nodes equally and automatically capture relationships between them.

2.3. Spatial-Temporal-Channel Attention (STC-Att)

Previous works [16, 13] treat the information equally without distinguishment in spatial and temporal domains. Actually, not all frames and joints make the same contribution to gait recognition, *i.e.*, the most discriminative information is commonly contained in some key joints from part of key frames. Inspired by [21], we propose a Spatial-Temporal-Channel Attention (STC-Att) module to dynamically adjust the weight of convolution kernel in spatial, temporal and channel domains.

The concrete structure of the STC-Att module is illustrated in Figure 3. Channel-wise attention (the right branch) is performed according to the correlation between channels. Given a feature map $\mathbf{f} \in \mathbb{R}^{T \times N \times C}$, the weight of each channel is learned in three steps. 1) The squeeze step, where global average pooling is performed to compact the feature map along the spatial and temporal dimension to a single value, that is, the feature map is converted into $\mathbf{f}_c \in \mathbb{R}^{1 \times 1 \times C}$. 2) The excitation step uses two FC layers with respective activation function to learn the nonlinear interaction among all channels, which finally generates the channel weight vector $\mathbf{w}_c \in \mathbb{R}^{1 \times 1 \times C}$ for the gait feature maps. 3) A scale layer is adopted to complete the recalibration of gait features through channel-wise product using the channel weight vector $\mathbf{w}_c$.

Spatial- and temporal-wise attentions (as shown in the left branch of Figure 3) are performed together to adaptively distinguish key joints and key frames from the whole skeleton sequence. Given a feature map $\mathbf{f} \in \mathbb{R}^{T \times N \times C}$, we first average the spatial information along the spatial dimension into the feature map $\mathbf{f}_t \in \mathbb{R}^{T \times C}$, and the temporal information along the temporal dimension into the feature map $\mathbf{f}_s \in \mathbb{R}^{N \times C}$. Then, the pooled feature maps are concatenated together into feature map $\mathbf{f}_{st} \in \mathbb{R}^{(V+T) \times C}$, followed by a FC layer and a ReLU activation function to reduce the feature dimension. After that, two independent FC layers followed by the activation functions are applied to learn the attention score vectors for temporal and spatial domains, respectively, which are denoted as $\mathbf{w}_t \in \mathbb{R}^{T \times C}$, $\mathbf{w}_s \in \mathbb{R}^{N \times C}$. Finally, the spatial/temporal attention vectors are multiplied on the input feature map, for spatial-temporal-adaptive feature refinement.

3. EXPERIMENTS

To evaluate our method, we conduct experiments on the public gait recognition dataset CASIA-B [22] and OUMVLP [23].

3.1. Experimental Setting

In model-based branch, the network is composed of 3 STGC blocks. The dimensions of output feature maps of each block are 96, 192, 384 in order. The size of the raw skeleton data is $2 \times 120 \times 15$, which means that there are 120 frames for each sequence, 15 joints for each frame, and 2 channels (horizontal and vertical coordinate) for each joint. The batch size is 128 and the training epoch is 65. The initial learning rate is 0.1 and decays with a factor of 0.1 after the 45-th and 55-th epoch. In appearance-based branch, we follow the settings in [20].

3.2. Comparisons with State-of-the-art Methods

We compare the proposed two-branch network to state-of-the-art gait recognition methods with the same experimental settings comprehensively, including GaitGraph2 [13], FR-GCN [14], GaitGL [8] and Lagrange [9]. Among that, the first two belong to model-based approaches and the last two belong to appearance-based approaches. Note that, the accuracies of the compared approaches are directly cited from their original papers. The detailed results are listed in Table 1 and 2. We can see that our model presents the superior performance under various conditions. Especially in the CL (Clothing) condition on CASIA-B, the accuracy of ours increases by 9.1% compared with GaitGL [8], which can indicate that the proposed network have great effectiveness in handling with occlusion and abstracting discriminative gait descriptors.

3.3. Ablation Study

Effectiveness of components in GCN. We validate the effectiveness of each proposed component in GCN-based branch in Table 3. From the comparison between the first and second

Table 1. Comparison of gait recognition performance on CASIA-B.

	Gallery angle	$0°-180°$											
	Probe angle	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
	GaitGraph2 [13]	78.5	82.9	85.8	85.6	83.1	81.5	84.3	83.2	84.2	81.6	71.8	82.0
	FR-GCN [14]	90.6	91.6	93.5	92.3	91.9	93.0	92.1	91.6	93.3	89.3	86.0	91.4
NM	GaitGL [8]	96.0	**98.3**	**99.0**	97.9	96.9	95.4	97.0	**98.9**	**99.3**	**98.8**	94.0	97.4
	Lagrange [9]	95.2	97.8	**99.0**	98.0	96.9	94.6	96.9	98.8	98.9	98.0	91.5	96.9
	Ours	**97.0**	97.9	98.4	**98.3**	**97.2**	**97.3**	**98.2**	98.4	98.3	98.1	**96.0**	**97.7**
	GaitGraph2 [13]	69.9	75.9	78.1	79.3	71.4	71.7	74.3	76.2	73.2	73.4	61.7	73.2
	FR-GCN [14]	77.9	85.2	84.0	81.2	82.5	78.9	81.3	79.5	80.2	77.8	71.0	80.0
BG	GaitGL [8]	**92.6**	**96.6**	**96.8**	95.5	93.5	89.3	92.2	**96.5**	**98.2**	96.9	**91.5**	**94.5**
	Lagrange [9]	89.9	94.5	95.9	94.6	93.9	88.0	91.1	96.3	98.1	**97.3**	88.9	93.5
	Ours	91.9	94.6	96.4	94.3	**94.4**	**91.6**	**94.1**	95.4	95.5	93.9	89.5	93.8
	GaitGraph2 [13]	57.1	61.1	68.9	66.0	67.8	65.4	68.1	67.2	63.7	63.6	50.4	63.6
	FR-GCN [14]	74.0	74.3	76.0	78.8	80.4	79.3	78.0	79.5	74.8	70.5	67.0	75.7
CL	GaitGL [8]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	Lagrange [9]	81.6	91.0	94.8	92.2	85.5	82.1	86.0	89.8	90.6	86.0	73.5	86.6
	Ours	**87.4**	**96.0**	**97.0**	**94.6**	**94.0**	**90.1**	**91.5**	**94.1**	**93.8**	**92.6**	**88.5**	**92.7**

Table 2. Comparison of gait recognition performance on OUMVLP.

Probe angle	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
GaitGraph2 [13]	54.3	68.4	76.1	76.8	71.5	75.0	70.1	52.2	60.6	57.8	73.2	67.8	70.8	65.3	67.1
FR-GCN [14]	48.3	53.5	56.8	58.9	58.3	55.2	50.6	36.6	49.0	45.5	60.6	60.4	57.4	53.6	53.2
GaitGL [8]	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
Lagrange [9]	84.5	89.8	91.0	91.2	90.7	90.5	90.2	88.5	87.9	89.9	90.0	89.2	89.2	88.7	89.4
Ours	**91.3**	**92.4**	91.2	89.9	**92.1**	90.9	90.2	**90.0**	**92.1**	90.3	89.5	**92.5**	90.7	**90.5**	**91.0**

Table 3. Ablation studies conducted on CASIA-B.

	Accuracy(%)		
	NM	BG	CL
GCN-Baseline	89.8	78.5	78.0
+ Fully-connected	91.3	79.4	78.5
+ STC-Att (Ours)	**91.8**	**79.8**	**79.4**

Table 4. Averaged accuracies with different values of λ.

	Accuracy(%)			
	NM	BG	CL	Mean
model-based branch	91.8	79.8	79.4	83.67
appearance-based branch	91.5	81.7	68.6	80.60
$\lambda = 300$	97.0	92.8	**93.4**	94.38
$\lambda = 350$	97.5	93.4	93.1	94.65
$\lambda = 400$ (Ours)	97.7	93.8	92.7	**94.74**
$\lambda = 450$	**97.7**	**93.9**	91.9	94.51
$\lambda = 500$	97.6	93.5	90.8	93.98

rows, we can clearly seen that the model with fully connected graph convolution gains higher accuracies for all conditions, which verifies the availability of this operator. In addition, it can be seen that with the integration of our attention mechanism (from the second row to the third row), the recognition accuracies of the network have been improved in all conditions, which demonstrates the superiority of STC-Att for capturing the discriminative gait representations. Note that all the results in Table 3 are generated using only skeleton features.

Discussion of two-branch fusion. The accuracies using different networks are shown in Table 4. The first two lines show the averaged accuracies of model- and appearance-based branch, respectively. As mentioned in Section 2.1, to combine the model- and appearance-based features, a ratio λ is needed to balance the disparity between them. The following five lines show the performance of the proposed two-branch network using different values of λ. We can see that the proposed two-branch network consistently gets a better performance with a large margin than using only one of them.

In order to determine the most suitable ratio of the two-branch network, we compare different values of λ from 300 to 500 with 50 as interval. We observe that as the value of λ increases, the accuracy rate under all conditions shows a trend of first increasing and then decreasing. NM and BG condition get the best results with setting $\lambda = 450$ and CL condition gets the best results with setting $\lambda = 300$. In addition, we calculate the mean value of accuracies under the three conditions (shown in the last column). Considering that the model should perform well in various states, we choose $\lambda = 400$ with the highest mean value as the final ratio.

4. CONCLUSION

In this paper, we have proposed a simple yet effective two-branch network, containing a CNN-based branch extracting gait features from silhouettes and a GCN-based branch extracting gait features from skeletons. We also provided a fully connected graph convolution operator and an attention module for efficient and effective gait feature representation. Extensive experiments show that our method outperforms the state-of-the-art methods under all conditions, which show its great potential in dealing with the challenges caused by the viewpoint, carrying and clothing condition variations.

5. REFERENCES

[1] 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, pp. 12909–12918.

[2] Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, and Jianfeng Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE TCSVT*, vol. 31, no. 1, pp. 260–274, 2020.

[3] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren, "Joint intensity transformer network for gait recognition robust against clothing and carrying status," *IEEE TIFS*, vol. 14, no. 12, pp. 3102–3115, 2019.

[4] 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*, vol. 39, no. 2, pp. 209–226, 2016.

[5] Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, and Hongdong Li, "Learning joint gait representation via quintuplet loss minimization," in *CVPR*, 2019, pp. 4700–4709.

[6] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *AAAI*, 2019, pp. 8126–8133.

[7] Jingran Su, Yang Zhao, and Xuelong Li, "Deep metric learning based on center-ranked loss for gait recognition," in *ICASSP*, 2020, pp. 4077–4081.

[8] Beibei Lin, Shunli Zhang, and Xin Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *ICCV*, 2021, pp. 14648–14656.

[9] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang, "Lagrange motion analysis and view embeddings for improved gait recognition," in *CVPR*, 2022, pp. 20249–20258.

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 7291–7299.

[11] Weizhi An, Rijun Liao, Shiqi Yu, Yongzhen Huang, and Pong C Yuen, "Improving gait recognition with 3d pose estimation," in *Chinese Conference on Biometric Recognition*, 2018, pp. 137–147.

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

[13] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll, "Towards a deeper understanding of skeleton-based gait recognition," in *CVPRW*, 2022, pp. 1569–1577.

[14] Likai Wang, Jinyan Chen, and Yuxin Liu, "Frame-level refinement networks for skeleton-based gait recognition," *CVIU*, vol. 222, pp. 103500, 2022.

[15] Ning Cai, Shiling Feng, Qing Gui, Lei Zhao, Huadong Pan, Jun Yin, and Bin Lin, "Hybrid silhouette-skeleton body representation for gait recognition," in *IHMSC*, 2021, pp. 216–220.

[16] Likai Wang, Jinyan Chen, Zhenghang Chen, Yuxin Liu, and Haolin Yang, "Multi-stream part-fused graph convolutional networks for skeleton-based gait recognition," *Connection Science*, vol. 34, no. 1, pp. 652–669, 2022.

[17] Shuangqun Li, Wu Liu, and Huadong Ma, "Attentive spatial–temporal summary networks for feature learning in irregular gait recognition," *IEEE TMM*, vol. 21, no. 9, pp. 2361–2375, 2019.

[18] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, and Xi Li, "Metagait: Learning to learn an omni sample adaptive representation for gait recognition," in *ECCV*, 2022, pp. 357–374.

[19] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *CVPR*, 2020, pp. 143–152.

[20] 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, pp. 14225–14233.

[21] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE TPAMI*, vol. 45, no. 2, pp. 1474–1488, 2022.

[22] 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, pp. 441–444.

[23] 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.