HID 2021: Competition on Human Identification at a Distance 2021

Shiqi Yu[1], Yongzhen Huang[2,3], Liang Wang[4], Yasushi Makihara[5], Edel B. García Reyes[6], Feng Zheng[1], Md Atiqur Rahman Ahad[7,5], Beibei Lin[8], Yuchao Yang[9], Haijun Xiong[10], Binyuan Huang[11], Yuxuan Zhang[12]

[1]Southern University of Science and Technology, China. [2]Beijing Normal University, China.

[3]Watrix Technology Limited Co. Ltd. [4]Institute of Automation, Chinese Academy of Sciences, China.

[5]Osaka University, Japan. [6]Shenzhen Institute of Artificial Intelligence and Robotics for Society, China.

[7]University of Dhaka, Bangladesh. [8]Beijing Jiaotong University, China. [9]Sichuan University, China.

[10]Huazhong University of Science and Technology, China. [11]South China Normal University, China.

[12]South China University of Technology, China.

http://hid2021.iapr-tc4.org/

Abstract

The Competition on Human Identification at a Distance 2021 (HID 2021) is to promote the research in human identification at a distance and to provide a benchmark to evaluate different methods. HID 2021 is the second follow-up from the first one, HID 2020. The dataset size and the evaluation protocal are the same with the previous competition, but the data in the test set has been changed. The paper firstly introduces the dataset and the evaluation protocol, then describes the methods from the top teams and their results. The methods show how to achieve state-of-the-art performance on gait recognition. The results in HID 2021 are better than those in HID 2020. From the comparisons and analysis, some useful conclusions can be drawn. We hope more improvements can be achieved by better follow-up competitions.

1. Introduction

Human identification at a distance (HID) has great demands in applications since commonly-used biometric patterns, such as iris and fingerprint require to be acquired at a very close distance. Faces can be acquired at a relatively far distance, but faces are tend to be occluded by hats, sunglasses, etc. Especially in the pandemic caused by COVID-19, faces are heavily occluded by masks. Therefore, human identification at a distance is becoming a much more crucial for various public safety issues. A typical scenery of human identification is shown in Figure 1.

In a scene as Figure 1 gait should be the only biometric feature for identification since other features cannot be acquired. Besides, the walking direction may be various, and the clothing of a subject may be changed among different days. Some carried objects will also change the shape of the moving body and the style of walking. The resolution of subject is also a problem, and bad illuminations also can reduce the quality of images. If the subjects in these kind of scenes can be identified, the perception capability of the system will be improved greatly. But it is a very challenging task.

Figure 1. A typical scene of human identification at a distance [1].

HID research has gained obvious improvement in the past two decades. The improvement by deep learning is even more obvious in recent years. There is still a gap to meet the requirements in real applications. To promote the research and to exchange ideas in this area, we have organized competitions on HID. The first competition on HID was held with Asian Conference on Computer Vision (ACCV), 2020. This challenge is the follow-up from the 1st event. Compared with the previous competition HID 2020,

978-1-6654-3780-6/21/$31.00 ©2021 IEEE

the recognition accuracy is greatly improved.

In the rest part of the paper, the competition HID 2021 is introduced. The experimental protocol and evaluation method are presented in Section 2. The results from top teams are summarized in Section 3. The teams and their methods are detailed in Section 4. Finally, We give our conclusions and analysis in Section 5.

2. Experimental Protocol and Evaluation

Similar with the previous competition HID 2020, the HID 2021 had two phases with different test sets. The first phase was from March 1st, 2021 to the April 10th, 2021. The second stage was much shorter and from April 10th to April 20th, 2021. In both phases, the submitted results were evaluated at CodaLab online automatically. In the following part of this section, the dataset and the evaluation protocol are introduced.

2.1. Dataset

CASIA-E Dataset was employed for the competition. CASIA-E is a novel gait dataset created by the Institute of Automation, Chinese Academy of Sciences and Watrix company. The dataset contains 1008 subjects. There are about 600 video sequences for each subject. Those videos were collected from 28 views, which range from 0°to 180°. The data was collected in 3 scenes. The backgrounds and floors may be different. The walking conditions of each subject may be normal walking, walking in a coat or walking with a bag.

To reduce the burden of participants on data preprocessing, we provided human body silhouettes. The silhouettes were obtained from the original videos by a human body detection deep model and a segmentation deep model provided by Watrix company. All silhouette images were resized to a fixed size 128×128 as show in Figure 2. We did not remove bad quality silhouettes manually. All silhouettes are from automatic detection and segmentation algorithms. As show in the figure the silhouettes are not in perfect quality. Some noises in real applications are involved. The noises make the competition more challenging. It also makes the competition is a good simulation to real applications.

The dataset was separated into the training set and the test set. For each subject, 10 sequences were randomly selected from the dataset for the competition. The first 500 subjects are in the training sets, and the rest 508 ones are in the test set. Surely the labels of all sequences in the training set are given. But only 1 sequence of each subject in the gallery set is given its label. The other 9 sequences are in the probe set and need to be predicted their labels. The data for the competition is also described in Table 1. Since the 10 sequences of a subject was randomly selected, they should be in different views, different walking conditions and different clothing. Considering only 1 sequence is in the gallery set for each subject, to distinguish 508 subjects is a challenging task.

Web site of HID 2021, http://hid2021.iapr-tc4.org/
CodaLab site, https://competitions.codalab.org/competitions/29552

Figure 2. Example silhouette images from dataset CASIA-E.

	Training Set Subject #1-#500	**Test Set** Subject #501-#1008	
		Gallery	Probe
Num. of Seq.	10	1	9

Table 1. The numbers of sequences for the training set and the test set (including the gallery set and the probe set). The sequences of a subject were randomly selected from hundreds of ones of that subject.

2.2. Evaluation protocol

The evaluation should be user-friendly and convenient for participants. It should also be fare and safe to be hacked. We designed detailed rules as follows:

1. To avoid the ID labels of the probe set to be found by numerous submissions, we limited the number of submissions each day to 2. Only one CodaLab ID is allowed for a team.

2. The accuracy was evaluated automatically at CodaLab. The ranking will be updated in the scoreboard accordingly.

3. There were 40 days in the first phase. But only 25% of the probe samples were taken for the evaluation in the first phase.

4. There were 10 days in the second phase. The remaining 75% of the probe sample were for the evaluation. The data was different from that in the first phase.

5. The top six teams in the final scoreboard need to send their programs to the organizers. The programs were for being ran to reproduce their results. The reproduced results should be consistent with the results shown in the CodaLab scoreboard.

2.3. Performance metric

Rank 1 accuracy is for evaluating the methods from different teams. It is straightforward and easy to implement.

$$accuracy = \frac{TP}{N} \tag{1}$$

where, TP denotes the number of true positives, and N is for the number of the probe samples.

2.4. Awards

To attract more participants to the competition, 4 awards are offered for the teams in the second phase. We are grateful to Watrix Technology to sponsor the competition. The awards are listed as follows:

- First Prize: CNY15,000
- Second Prize: CNY5,000
- Third Prize: CNY2,500
- Fourth Prize: CNY1,000

where, CNY stands for Chinese Yuan.

3. Summary of Submissions

In the first phase, 128 participants registered and formed about 100 teams. Thirty-one of them moved to the second phase and submitted their results. The detailed information can be found at result page of the competition website.

3.1. Top 10 in the first phase

We selected the top 10 results in the first phase and listed them in Table 2. Considering that the dataset is a very challenging one, only one silhouette sequence for each subject in the gallery, and no color and texture, the silhouettes are noisy and the view directions are different, the achieved results are encouraging. The recognition rates are all greater than 60.0%, even reaches 81.4%.

3.2. Top 10 in the second phase

In the second phase, the accuracies were improved again. The average accuracy of the top 10 teams was improved from 64.8% to 67.5%. The results in the previous competition HID 2020 are listed in Table 4 for comparisons. The metric is the same and the protocols are similar in HID 2020 and HID 2021. Therefore, we can compare the results from the two competitions. Obviously, the accuracy has been improved greatly for the HID 2021. The best result in HID 2020 is 63.0%, which can only be ranked as the 8th in HID 2021. The average improvement is 18.0% (from 49.5% to 67.5%).

http://hid2021.iapr-tc4.org/results/

Rank	Team Name	CodaLab ID	Accuracy (in %)
1	GRGroup	BeibeiLin	81.4
2	HUST-MCLAB	Haijunxiong	68.2
3	-	hxhhust	66.8
4	SoGait	AlexFor	65.3
5	BRL	dtdtdt	62.4
6	BIPLAB	ZhangYuxuan	61.1
7	-	fuhui	61.1
8	-	robots	60.9
9	Alibaba	yunfeng	60.3
10	-	XinWang	60.2
Avg.	-	-	64.8

Table 2. The leaderboard of the top 10 teams in the first phase in HID 2021.

Rank	Team Name	CodaLab ID	Accuracy (in %)
1	GRGroup	BeibeiLin	83.9
2	BRL	dtdtdt	79.9
3	HUST-MCLAB	Haijunxiong	71.3
4	SoGait	AlexFor	66.8
5	SDU_gait	YaoJun	66.6
6	BIPLAB	ZhangYuxuan	63.7
7	–	robots	63.5
8	Alibaba	yunfeng	60.4
9	RUSH B	liguodong	59.6
10	Hrbeu-Xing	panfengzhang	59.2
Avg.	-	-	67.5

Table 3. The leaderboard of the top 10 teams in the second phase in HID 2021.

Rank	CodaLab ID	Accuracy (in %)
1	BeibeiLin	63.0
2	BRL	54.1
3	panfengzhang	53.4
4	ctsu-ca	51.5
5	ywang26	50.3
6	Wbz	49.3
7	HeHaodi	49.2
8	kouen93.	47.9
9	recognizer	43.6
10	color	33.1
Avg.	-	49.5

Table 4. The leaderboard of the top 10 teams in the second phase in the previous competition: HID 2020.

4. Top Teams and Their Methods

The top 6 teams and the team information are listed in Table 5. Each team can have several members and may have

supervisors. Five teams among the top 6 teams provided the descriptions of their methods. The 5th team did not provide. Those methods are presented in the following part of this section.

4.1. Team GRgroup: 1st position

The entire pipeline of their approach contains three steps: The first step is data preprocessing, the second step is feature extraction by GaitMask network, which was proposed by the team and GaitGL [6] network. The feature vectors from GaitMask and GaitGL are concatenated as the final feature. Finally, the query expansion and re-ranking [13] are employed to improve the recognition accuracy.

GaitMask network contains the following components. Firstly, Local Temporal Aggregation (LTA) is employed to aggregate the local temporal information [6]. Then Global and Mask Feature Extractor (GMFE) is proposed to learn more comprehensive gait features. In this component, the Global and Mask Convolution Layer (GMCL) is implemented to extract the gait features. GMCL includes two branches: Global Feature Extraction and Mask Feature Extraction. Global Feature Extraction is used to extract global feature representations, while Mask Feature Extraction is used to generate more comprehensive local feature representations. Finally, they employed Generalized-Mean (GeM) pooling layer [7] and temporal pooling to generate feature representations. In addition, the whole gait recognition method is built by 3D convolution [5]. In the training stage, Separate Triplet Loss [2] is employed to train GaitMask network.

Query Expansion is employed to improve the recognition accuracy. Specifically, it firstly concatenates all feature representations from the gallery and probe sets. Then, the clustering method based on the euclidean distance is adopted to find the most similar samples. Each feature representation from the gallery and probe sets is updated to the mean feature representation of the other representations in the same cluster.

A preprocessing method in [2] is selected to normalize the input images. The size of the normalized gait images is 64×64. The training details are shown in Table 6. The model is firstly trained using OU-MVLP [9] and then fine-tune it using the competition dataset and CASIA-B [11]. All training tasks take Adam as the optimizer, and the initial learning rate was $1e-4$. For the OU-MVLP dataset, the learning rate reset to $1e-5$ after 150K. For the CASIA-B and competition datasets, the learning rate reset to $1e-5$ after 10K.

4.2. Team BRL: 2nd position

The method also contains three steps: (1) preprocessing the competition dataset, (2) feature extraction by GaitSet network [2] and (3) optimizing the recognition results.

Preprocessing dataset: In order to distinguish and remove low quality silhouette images, a MobileNetV2 [8] is trained. 1103 images (including 212 low-quality images) were firstly manually labelled their qualities first. Then data augmentation was employed to generate more samples. The model can be used to distinguish the low quality images and remove them from the dataset.

Feature Extraction: If some silhouette images are remove the from a sequence, the temporal information inside will be distorted. GaitSet can regard the silhouettes as a set and is robust to silhouette lossing. GaitSet was used to extract feature and initially trained by OU-MVLP dataset [9], and fine-tuned by the competition dataset [12].

Results Optimization: In order to improve the recognition accuracy, the k-reciprocal encoding [13], which is widely used in pedestrian re-identification, was employed for the competition. For the sequence in the probe, the closest one among all the galleries was taken as the predicted label. Due to the data enhancement of the probe, there were two predictions for each sequence, and the one with the smallest distance was taken as the prediction label.

4.3. Team HUST-MCLAB: 3rd position

The team also employed GaitSet [2] for feature extraction, but also introduced Multi-branch Diverse Region Feature Generator (MDFG) and Global and Micro Motion Capturing Module (GMCM) for discriminative feature learning and global-local temporal feature learning respectively. To avoid using low quality images, a simple image-filtering strategy by considering the ratio of the foreground was used to filter out low-quality images.

Preprocessing Dataset: The preprocessing stage of the dataset consisted of two stages, which were data cleaning and data augmentation. In the data cleaning stage, the ratio of the foreground pixels of each image in each sequence was calculated, then to sort the images by the ratio of the foreground of each image in each sequence. The images with their ratios out of the range [0.85, 1.15] of the median will be removed.

Feature Extraction: GaitSet, MDFG and GMCM modules are for gait features extraction. GaitSet is the backbone of the whole network. A CNN part based on the GaitSet backbone was introduced to get more detailed information by extract feature maps with 32, 64, 128 and 256 channels.

MDGF is employed in both Set-level and Frame-level to generate visual clues in diverse regions for fine-grained feature learning. MDFG is applied to each frame $f_i \in \mathbb{R}^{1\times C\times H\times W}$, then produced N branches output $b_{i,j}$ by N independent 1×1 2D convolutions, where i denotes each branch feature and $j = 1, 2, 3, \cdots, N$. Global average pooling and global max pooling are used in each branch feature respectively to produce channel-compressed feature

Rank	Team Name	Team member	Supervisor
1	GRGroup	BeibeiLin, Shengdi Qin, Chengwei Wan	Shunli Zhang, Jiande Sun
2	BRL	Yuchao Yang, Shuiwang Li, Tao Ding, Yiwen Zhang	Qijun Zhao
3	HUST-MCLAB	Haijun Xiong, Xiaohu Huang	Bin Feng
4	SoGait	Binyuan Huang, Yongdong Luo, Jiahui Xie, Zhiwen Li	Chengju Zhou, Jiahui Pan
5	SDU_gait	Jun Yao, TianHuan Huang, Chang Liu, Lei Chen	Xianye Ben
6	BIPLAB	Yuxuan Zhang, Xin Wang, Hui Fu, Peng Zhao, Shizhe Liang	Wenxiong Kang

Table 5. The top 6 teams in the second phase (the final phase).

Datasets	Method	Epoch	T Frames	Batch Size
OU-MVLP	GaitGL GaitMask	250K	30	32*8
CASIA-B and CASIA-E	GaitGL GaitMask	15K	64	12*8

Table 6. The training details of Team GRgroup on different datasets. T means the length of input gait sequences in the training stage.

$H_{i,j} \in \mathbb{R}^{1\times H\times W}$. For each branch, the k-th maximum value of $H_{i,j}$ is denoted as $\sigma_{i,j}$. Then, $H_{i,j}$ and $\sigma_{i,j}$ were combined to obtain a focal mask $A_{i,j}$ by a Sigmoid function:

$$A_{i,j}(x,y) = \frac{1}{1+exp\{-H_{i,j}(x,y)-\sigma_{i,j}\}} \qquad (2)$$

where, (x, y) denotes any spatial location in $H_{i,j}$ and $\sigma_{i,j}$ represents the threshold of Sigmoid function. Thus, locations with values larger than $\sigma_{i,j}$ are highlighted, and locations with values smaller than $\sigma_{i,j}$ are suppressed. To generate different activated regions in different branches, Overlapped Activation Penalty (OAP) loss [10] was applied on $A_{i,j}$ for supervising. OAP loss aims to punish overlapped activated regions. The definition is as follows:

$$L^i_{oap} = \frac{1}{N}\Sigma_{x,y}(A_{i,1}, \odot A_{i,2}, \odot, \cdots, \odot A_{i,N}) \qquad (3)$$

where, L^i_{oap} represents OAP loss for i-th frame, $\odot$ denotes element-wise multiplication, and the overall L_{oap} is the mean of the L^i_{oap}.

GMCM contains two parts: Micro-motion Template Builder (MTB) and Global-motion Template Builder (GTB) [3]. MTB aims to map the frame-level part-informed feature vectors into the micro-motion feature vectors, and GTB is designed to map the frame-level global-informed feature vectors into the feature vectors.

In the training stage, the image of each frame was normalized to the size 128×88. The length of input gait sequences of the CASIA-B and the competition datasets were all set to 30. In the test stage, if the sequence length was less than 300, the whole gait sequences were put into the proposed model to extract gait features. Otherwise, the input sequence length was set to 300, and in all experiments Adam was taken as the optimizer. For the CASIA-B dataset, the iteration number set to 100K, and the learning rate was $1e-4$. For the competition dataset, the iteration number was set to 160K, and the learning rate was first set to $1e-4$ and set to $1e-5$ after 100K.

4.4. Team SoGait: 4th position

Due to the fact that gait recognition suffers from various covariates, including view, clothing, and carrying status, three key components are used to fuse multimodal features to learn more discriminative representations. They are Lateral Connection Feature Aggregator (LCFA), Multi-Scale Feature Extractor (MSFE) and Global and Local Feature Module (GLFM).

Particularly, LCFA is inspired by Gait Lateral Network(GLN) [4] and employs a serial multi-scale feature fusion method to combine gait features of different depth layers and different receptive fields.

MSFE module is a parallel multi-scale feature fusion method to aggregate features of different scales in gait images. MSFE module extract features by three convolution kernels of different sizes so that the network can capture different scales of gait characteristics. The features on different scales are merged by concatenating the channel dimension so that the network can learn more discriminative gait features.

GLFM is inspired by the idea of global and local feature extraction in gait-based age estimation [14]. GLFM is divided into global and local feature extractors. The global feature extractor uses convolution kernels to extract global information, while the local feature extractor divides the gait feature into multiple parts and executed the different parts separately. Each convolution network in the local feature extractor was used to learn the feature of specific parts and then integrated them by concatenating horizontally. Finally, they fused the global and local features through element-wise addition.

4.5. Team BIPLAB: 6th position

Data preprocessing: Similar with the methods of other teams, data cleaning and data augmentation are in data preprocessing. In data cleaning, a classification network is trained to remove low-quality images from the competition dataset. It is an easy task for the neural network to deter-

mine the qualities of silhouettes. The classifier can reach an accuracy of about 98% after only 30 epochs in training. Finally, about 150K silhouette images can be removed from the training set. In data augmentation, a variety of image augmentation strategies are adopted, such as Gaussian blurring and random occlusions, to increase the number of samples and improve the robustness of the trained model.

GaitGL feature extraction: The model used is GaitGL [6] and is firstly trained with OU-MVLP dataset [9] and then fine-tuned with the competition data. All silhouettes are resized to 64×64. The batch size is set to $16 \times 8 = 128$. In the training stage, the number of frames of each sequence from OU-MVLP was set to 30. Adam was the optimizer, and the initial learning rate is 0.0001. The epoch number was set to 250K. The learning rate was reset to $1e-5$ after 150K epochs. In the fine-tune step with the competition data, the initial learning rate was first set to $1e-4$ and reset to $1e-5$ after 80k epochs.The whole training process ends at 100k epoch. The value 0.2 is for the triplet margin, and the batch size was $8 \times 8 = 64$.

4.6. Methods comparison of the top teams

Five teams reported their methods and implementation details. Because of the limitation of the space, we only summaried the methods in the previous part of this section. Detailed implementation details can be found at the competition result page. To easily compare those methods for a better understanding, we list the modules and models used in those methods in Table 7. From the Table, it can be noticed that all teams explored data cleaning, data augmentation and model pre-training. Especially, the top two teams used the re-ranking strategy, which is one of the key factors to have a good accuracy. In the train stage, all teams combined global features and local features by some methods such as GaitGL, GLFM or GMCM. The 1st and 6th teams employed 3D convolution to extract temporal and spatial information simultaneously. Other teams used two branches to extract temporal and spatial features. In addition, the 1st team used the Query Expansion (QE) strategy, which can increase two percentages.

5. Conclusions and Future Improvements

The methods from different teams can help to understand how a good performance can be achieved. From the methods we can find that the quality of the input data is still essential to the performance. Better human body silhouettes can help obviously, and it is the reason why most teams have a quality evaluation module before feature extraction. Secondly, since the input data is noisy most methods using some aggregate methods for robust feature extraction, and a typical one is GaitSet [2]. The temporal features especially some local temproal features are not widely employed even this kind of features should benefit the identification. We believe HID will be continuesly improved with the improvement on human body modeling and analysis since the human body related features will be in a higher quality.

The main goal of the competition is to promote research in this area. A competition can provide a fair benchmark for different methods. It is essential to compare different methods with the same benchmark. We can happily conclude that the competition achieved the goal. The accuracy has been obviously improved in HID 2021 compared with the previous competition HID 2020. We still hope that more improvements can be gained in the future competitions, and the accuracy can reach a level, which is similar with face recognition. Despite of the success of this competition, there is still something to ponder in the future to have a better competition:

Data modality: Due to the consideration on privacy the original RGB videos were not provided. Only binary silhouettes were provided for the competition. In the future we may provide some other kinds of data such as human skeletons, optical flow. A new task can be created for skeleton data or some other kinds of data.

Data size: The successes of different topics such as face recognition show that the size of data is essential. A deep model can be trained much better if with much more samples. If the size is 10 times larger than the current one, we believe the competition will be more challenging and more useful to promote the research in the area. To collect and share big data on HID is a relatively sensitive problem according to the laws and regulations of different countries and regions. We also would like to hear feed backs from the academic community on this topic.

Metric: Currently only top 1 accuracy is for evaluation. Some more metrics can be employed for a comprehensive evaluation. Some possible metrics can be the speed of the algorithm, the memory comsumed, another kind of accuracy metric such as equal error rate (EER).

Acknowledgement

We would like to thank Watrix Technology Limited Co. Ltd. for sponsoring the competition, and the Institute of Automation, Chinese Academy of Sciences for providing the dataset CASIA-E. We would also like to acknowledge CodaLab for the result evaluation platform. Our special gratitude goes to our advisors Prof. Tieniu Tan, Prof. Yasushi Yagi and Prof. Mark Nixon. Their supports and suggestions made the competition successful. We also acknowledge Mr. Jingzhe Ma, Mr. Zihao Mu and Mr. Chuanfu Shen for their technical supports.

Rank	Team Name	CodaLab ID	Algorithm	Accuracy (%)
1	GRGroup	BeibeiLin	GaitGL	64.1
			GaitGL+Re-ranking	78.4
			GaitGL+Re-ranking+QE	80.7
			GaitGL+GaitMask+Re-ranking+QE	**83.9**
2	BRL	dtdtdt	GaitSet	54.6
			GaitSet+train data augmentation(TDA)	57.4
			GaitSet+TDA+pre_trained	66.4
			GaitSet+TDA+pre_trained+Re-ranking	76.8
			GaitSet+TDA+pre-trained+Re-ranking+probe data augmentation	**79.9**
3	HUST-MCLAB	Haijunxiong	**GaitSet+MDFG+GMCM**	**71.3**
4	SoGait	AlexFor	GaitSet+GLFM+MSFE	64.7
			GaitSet+GLFM+MSFE+LCFA	**66.8**
6	BIPLAB	ZhangYuxuan	GaitSet+cleaned data	58.6
			GaitGL+original data	57.9
			GaitGL+cleaned data	60.0
			GaitGL+cleaned data+data augmentation	**63.7**

Table 7. Summary of different methods and their accuracies in the second phase.

References

[1] Visual tracker benchmark, http://cvlab.hanyang.ac.kr/tracker_benchmark/index.html. 1

[2] H. Chao, Y. He, J. Zhang, and J. Feng. GaitSet: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8126–8133, 2019. 4, 6

[3] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14233, 2020. 5

[4] S. Hou, C. Cao, X. Liu, and Y. Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European Conference on Computer Vision*, pages 382–398. Springer, 2020. 5

[5] B. Lin, S. Zhang, and F. Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3054–3062, 2020. 4

[6] B. Lin, S. Zhang, X. Yu, Z. Chu, and H. Zhang. Learning effective representations from global and local features for cross-view gait recognition. *arXiv preprint arXiv:2011.01461*, 2020. 4, 6

[7] F. Radenović, G. Tolias, and O. Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 4

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4

[9] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):1–14, 2018. 4, 6

[10] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1398, 2019. 5

[11] S. Yu, D. Tan, and T. 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'06)*, volume 4, pages 441–444. IEEE, 2006. 4

[12] Y. Zhang, Y. Huang, L. Wang, and S. Yu. A comprehensive study on gait biometrics using a joint cnn-based method. *Pattern Recognition*, 93:228–236, 2019. 4

[13] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 4

[14] H. Zhu, Y. Zhang, G. Li, J. Zhang, and H. Shan. Ordinal distribution regression for gait-based age estimation. *Science China Information Sciences*, 63(2):1–14, 2020. 5