GaitPVD: Part-based View Distillation Network for Cross-View Gait Recognition

Lidan Shang
Key Laboratory of Electromagnetic Space Information of CAS, University of Science and Technology of China, Hefei, Anhui China
shangld@mail.ustc.edu.cn

Dong Yin*
Bin Hu†

ABSTRACT

Gait is a biometric feature used for long-distance identification which has important applications in video surveillance. However, the performance of gait recognition is limited by view angle variation. To solve the problem, we propose a Part-based View Distillation network (GaitPVD) which is a teacher-student framework. Firstly, we design a part-based network and use the part-based teacher network in GaitPVD to extract the gait feature from the gait sequences under the normative view. Secondly, we adopt knowledge distillation to solve cross-view gait recognition and design a loss function(PVDLoss) that is used to train the student network, which can achieve the propagation of view knowledge in GaitPVD. Finally, we extract the gait features from the gait sequences by the student network and calculate the Euclidean distance among the features to get the identity information. It is demonstrated by comprehensive experiments that our method(GaitPVD) can achieve state-of-the-art recognition accuracy on the two most popular cross-view gait recognition datasets CASIA-B and OU-MVLP.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**;

KEYWORDS

Deep learning, Gait recognition, Knowledge distillation

ACM Reference Format:
Lidan Shang, Dong Yin, and Bin Hu. 2021. GaitPVD: Part-based View Distillation Network for Cross-View Gait Recognition. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487100

*Key Laboratory of Electromagnetic Space Information of CAS, University of Science and Technology of China, Hefei, Anhui China, yindong@ustc.edu.cn
†Key Laboratory of Electromagnetic Space Information of CAS, University of Science and Technology of China, Hefei, Anhui China, hbhubin@mail.ustc.edu.cn

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487100

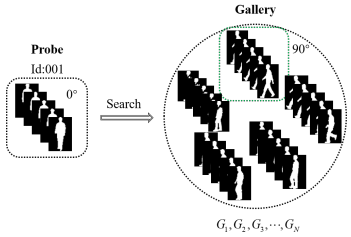

Figure 1: Schematic Diagram of Cross-View Gait Recognition.

1 INTRODUCTION

Gait is a biometric feature, similar to human faces, fingerprints, iris, etc., which can reflect the characteristics of the individual. However, compared with human faces, iris, and fingerprints, it can be obtained at a long distance and without interaction. Therefore, gait recognition is an effective way of identity recognition. There are usually several ways to represent gait: 1) the gait energy images (GEIs)[1]; 2) the gait silhouette images; 3) the RGB images. The methods[2][3] based on GEIs can greatly reduce the amount of calculation. However, GEIs neglect the difference in the importance among different silhouette images. The methods[4][5] based on RGB images usually use the human body structure information for gait recognition, such as the pose estimation. Although they have good performance, the amount of calculation is increased. Therefore, lots of methods based on the gait silhouette images emerge in the last two years. Chen et al.[6] propose an end-to-end network similar to Siamese network to extract and aggregate the gait features from the silhouette images. Chao et al.[7] assume that every gait silhouette image contains position information in the sequences, so they regard a gait silhouette sequence as a set and construct a network named GaitSet.

For cross-view gait recognition, the difficulty is to extract similar features from images with a large difference in appearance. As shown in Figure 1, we need to extract similar features from the images in the black box and in the green box. The Generative Adversarial Networks(GAN) is used in many methods [8] [9] [10] to solve the cross-view problem. Yu et al.[9] propose GaitGAN that can generate invariant gait images. He et al.[8] propose the multi-task GAN to learn view feature representations. These methods[8][9] can solve the cross-view problem to a certain extent. However, many GAN-based methods [8] [9] directly use the gait image from one view as a template to make images from other views closer to the template image. This means that they are transformed on

low-level features. But the recognition result is usually given by high-level features. The effect is usually not good enough.

For solving the problem, we propose a part-based view distillation network (GaitPVD) which is a teacher-student framework. And, our main contributions are summarized as follows: 1) We propose a new part-based network for gait recognition. 2)We apply knowledge distillation [11] [12] to the field of gait recognition to solve the cross-view problem. And, we propose a new loss function(PVDLoss) to optimize the propagation of view knowledge. 3) We perform lots of experiments to verify the effectiveness of our method, and achieve state-of-the-art on CASIA-B [13] and OU-MVLP [14].

2 RELATED WORKS

2.1 Cross-View Gait Recognition

The mainstream gait recognition methods can be roughly classified into three categories. In the first category [1], a gait sequence is directly compressed into a frame (such as GEI), and the features are extracted through convolution or traditional algorithms [15]. The methods in the second category [16] [17] usually obtain posture information or construct human body structure based on RGB images for gait recognition. Although they make full use of auxiliary information, the current mainstream pose estimation methods are based on RGB images. Unfortunately, many of the current mainstream gait recognition datasets only provide the gait silhouette images. In the third category [7], they usually regard a gait silhouette sequence as a whole and extract the spatio-temporal features of the gait based on the sequences.

For cross-view issues, the most commonly used method is generative adversarial networks. Yu et al.[9] propose GaitGAN to generate invariant gait images for solving cross-view gait recognition. In addition, some scholars have proposed a feature fusion network that can perform view conversion, so as to achieve the purpose of solving the cross-view problem.

2.2 Part-based Model

In many visual tasks, part-based models have achieved good performance. Especially in the field of person re-recognition, there are many part-based models [18] [19]. They usually divide the feature map into blocks along the horizontal direction, and optimize each block separately. The restricts of the part-based model is images are not aligned. However, the gait silhouette images are aligned during the construction process, so it is very suitable to adopt part-based models.

2.3 Knowledge Distillation

Hinton et al.[20] propose the concept of knowledge distillation, which is used to transfer knowledge from a large teacher network to a small student network in model compression. The main purpose is to make the small network reach or approach the effect of the large network under the guidance of the teacher network. With its rapid development, it is widely used in other visual tasks [21]. For example, in image-to-video person re-identification, some scholars use knowledge distillation to spread temporal information between the teacher network and the student network. Inspired by this, we propose the view distillation to solve the cross-view problem.

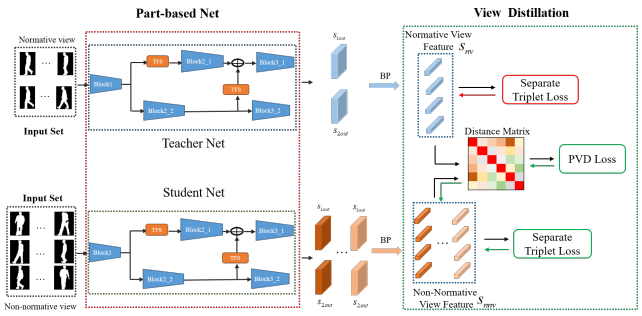

Figure 2: The Framework of GaitPVD. The Block1, 2 and 3 Are Composed of Convolution and Pooling Layer. The TFS Represents Temporal Feature Selection. The BP Represents the Block Pooling.

We train the teacher network to extract the gait feature under the normative view, and take it as a soft target to train the student network for reducing the difference between the features under the normative view and the non-normative view of the same person.

3 PROPOSED METHOD

3.1 Teacher Network (Part-based Net)

In this section, we first introduce the teacher network in GaitPVD which is a part-based network proposed by the paper, and then introduce the new method (view knowledge distillation) to solve cross-view gait recognition.

Block. As shown in Figure 2, a sequence of gait silhouette images denoted as$S = \{s_i | i = 1, ..., t\}$. is fed into the teacher network frame by frame, where$s_i$is the i-th frame in the sequence. The Block1 is composed of two convolution layers and a maxpooling layer, as shown in Table 1, which is used to extract the shallow features. The formula as follow:

$$F_{t\times c\times h\times w} = Block1(S) \tag{1}$$

Where $F_{t\times c\times h\times w}$ is a four-dimensional tensor, i.e. , the temporal, channel, height and width. It is obtained after a gait silhouette sequence passing through Block1. Similarly, other Blocks are also composed of two convolution layers and a pooling layer. In Table 1, In_S represents the shape of the input tensor of each layer, h and w represent the height and width of the input, respectively. t represents t the frame. There is no t variable in some In_S(such as $(n, 32, h/2, w/2)$) means it has completed the fusion among frames through TFS. n refers to the number of people included in one batch. In_C, Out_C, Kernel and Pad represent the input channels, output channels, kernel size and padding size.

Temporal Feature Selection. The TFS, which is aiming at extracting the motion features, can be formulated as:

$$m = TFS(F_{t\times c\times h\times w}) \tag{2}$$

where $F_{t\times c\times h\times w} = \{f_i | i = 1, ..., t\}$ is the set of sequence features, $f_i$is the feature of the i-th image in a gait sequence. $m$is the extracted motion feature. Compared with the input $F_{t\times c\times h\times w}$of TFS, its shape is $(c \times h \times w)$. The motion features contain more details than the pure image features, so mining motion features can

Table 1: The Structure of the Block

Block	In_S	Layer	In_C	Out_C	Kernel	Pad
Block1	(n,t,1,h,w)	Conv1	1	32	5	2
	(n,t,32,h,w)	Conv2	32	32	3	1
	(n,t,32,h,w)		Maxpool, kernel size = 2x2, stride=2			
Block2_1	(n,32,h/2,w/2)	Conv3_1	32	64	3	1
	(n,64,h/2,w/2)	Conv4_1	64	64	3	1
	(n,64,h/2,w/2)		Maxpool, kernel size = 2x2, stride=2			
Block2_2	(n,t,32,h/2,w/2)	Conv3_2	32	64	3	1
	(n,t,64,h/2,w/2)	Conv4_2	64	64	3	1
	(n,t,64,h/2,w/2)		Maxpool, kernel size = 2x2, stride=2			
Block3_1	(n,64,h/4,w/4)	Conv5_1	64	128	3	1
	(n,128,h/4,w/4)	Conv6_1	128	128	3	1
Block3_2	(n,t,64,h/4,w/4)	Conv5_2	64	128	3	1
	(n,t,128,h/4,w/4)	Conv6_2	128	128	3	1

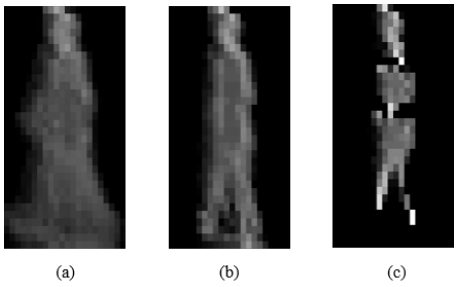

Figure 3: (a), (b) and (c) Are the Visualized Results of the Feature Map Extracted when the Statistical Function Is Set tomax(·)**,** *mean*(·)**, and** min(·) **in the TFS Respectively.**

bring better performance to gait recognition. For extracting motion features, some person choose to use statistical functions in the temporal dimension. They usually usemax(·), *mean*(·), etc., for joint extraction[7]. But we believe that using max(·)alone can achieve the best effect, and our experiments show that the performance of joint processing of multiple statistical functions is usually lower than using max(·) alone.

Thinking about this statistical process carefully, you can find that the process of taking the max value is the process of taking the silhouette information of all the images in a sequences. This is the reason why max(·) is chosen. As shown in Figure 3, compared withmax(·),*mean*(·)and min(·)lose a lot of response information.

Block Pooling. In order to obtain part-based features, we divide the feature map(s_{1out}, s_{2out}) into blocks along the horizontal direction and use the pooling layer in each block separately to obtain the final output feature, which is formulated as:

$$\{s_{1out1}, ..., s_{1outn_1}, s_{2out1}, ..., s_{2outn_2}\} = BP(s_{1out1}, s_{2out}) \quad (3)$$

where s_{1out} and s_{2out} are the output of the two branches in the pipeline respectively. The process of BP is as follows:1) We evenly divides$_{1out}$ands_{2out} into n_1 and n_2 parts along the horizontal direction respectively; 2) We use generalized-mean (GeM) pooling [22] to complete the final features output. This pooling layer allows us to determine the final fineness by adjusting p. After BP, we can get $n_1 + n_2$ parts.

Teacher Optimization. When training the teacher network, we only take a separate triplet loss[23] as the loss function. It can reduce the intra-class distance and increase the inter-class distance and be formulated as:

$$L_{tr} = \sum_{i}^{N} \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2 + \alpha \right]_+ \quad (4)$$

where $f(x_i^a)$ and $f(x_i^p)$ form a positive sample pair, $f(x_i^a)$ and $f(x_i^n)$ form a negative sample pair within a batch, N is the number of persons in one batch. $f(x_i^a)$ is the feature from an image x_i^a of a specific person, as an anchor. $f(x_i^p)$ is the feature from other image x_i^p (positive) of the same person. $f(x_i^n)$is the feature from any image x_i^n (negative) of any other person. α is the margin.

3.2 View Distillation

In order to solve the cross-view gait recognition, we propose to use the view distillation to obtain similar features under different views of the same person. We hope view knowledge can be spread among different views. Many experiments show that the accuracy of certain view is higher than other views. This indicates that the gait information under this view is more abundant and the extracted gait feature is easier to distinguish. Therefore, we regard the gait feature from the view as a person's normative feature. With the help of view distillation, we can extract the gait features which are closer to the normative feature, even in other views.

For the convenience of description, we define two concepts, namely normative view and non-normative views. The normative view refers to the high accuracy view, which corresponds to 36° and 45° in CASIA-B and OU-MVLP respectively. The non-normative views refer to the angles other than 36° or 45°. In order to achieve view distillation, we input the gait sequences of a person under the normative view into the teacher network, and input the gait sequences of the person under the non-normative views into the student network. After that, we use the output of the teacher network as a soft target to assist in the training of the student network.

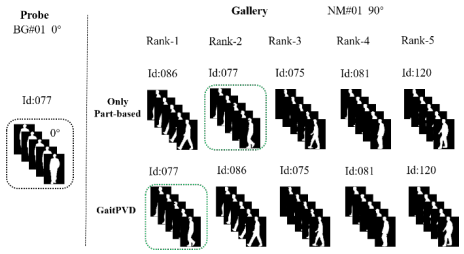

Figure 4: Recognition Rank List. Probe Is the Gait Sequence Belonging to Id077 under the 0° View in the Type Sequence Bg#1. Gallery Includes all Sequences under the 90° View in the Type Sequence NM#1.

Student optimization. The structure of the student network is the same as that of the teacher network. The difference lies in the input during training and the final objective function for optimization. In addition, we propose a new loss function (PVDLoss) to optimize the propagation of view knowledge. The loss function can make the non-normative features extracted under the non-normative views closer to the normative feature. It can be formulated as:

$$L_{pvd} = \frac{1}{M}\sum_{i}^{M} tr(\left\|S^i_{nnv} - S_{nv}\right\|_2) \tag{5}$$

where$M$represents the number of all non-normative views of the same person. S_{nv} is the feature under the normative view and its shape is $n \times d$, where n represents the number of blocks and d represents the number of channels. S^i_{nnv} is the feature of the i-th non-normative view and its shape is also $n \times d$. $\|S^i_{nnv} - S_{nv}\|$ is a $n \times n$ distance matrix, as shown in Figure 2. $tr(\|S^i_{nnv} - S_{nv}\|)$is the trace of the matrix, which is equal to the sum of the diagonal elements of the matrix. For the element from the i-th row and i-th column of the matrix, its physical meaning is the similarity between the i-th part of the non-normative feature and the i-th part of the normative feature. The non-diagonal elements of the matrix are meaningless.

The total objective function of the student network in GaitPVD includes the L_{pvd} used to distill the view knowledge and the L_{tr} used to reduce the intra-class distance and increase the inter-class distance. It can be formulated as:

$$L = L_{tr} + \lambda L_{pvd} \tag{6}$$

After experimental verification, we set λ to 0.3. As shown in Figure 4, the view distillation can improve correct matching from rank-2 to rank-1 in the final result.

4 EXPERIMENTS

4.1 Datasets

CASIA-B. CASIA-B is a classic dataset. It includes 124 subjects and each subject walks under 3 walking conditions to form 10 type sequences (NM#1-6, BG#1-2, CL#1-2). NM, BG, and CL indicate that the walking conditions are “normal case”(NM), “subjects carrying bags”(BG) and “subjects wearing coats or jackets”(CL), respectively. Each type sequence contains 11 views. We use the first 74 subjects as the training set and the remaining 50 subjects as the test set. In the test, the first four type sequences of NM are regarded as the gallery (NM#1-4), and the remaining 6 type sequences (as probe) are divided into three categories according to different walking conditions (NM#5-6, BG#1-2, CL#1-2).

OU-MVLP. OU-MVLP is currently the largest cross-view gait recognition dataset. Each subject includes 14 views (0, 15, 30...245, 270) and each view includes two sequences (#00, #01) [14]. And, it includes 10307 subjects. Among them, 5153 subjects are used for training and 5154 subjects are used for testing. During testing, we use the #00 sequence as the probe and #01 sequence as the gallery.

4.2 Comparison with State-of-art Methods

CASIA-B. As shown in Table 2, we compare GaitPVD with other state-of-the-art methods. Observing the results horizontally, we find that the accuracy can usually reach the maximum when the probe view is 36°. The reason may be that the gait silhouette images of this angle contain dual information of side view and front view. So we choose it as the normative view of CASIA-B. From Longitudinal comparison of results, we find that many methods (video-based) are better than CNN-LB (GEI), which reflects video-based has stronger representation ability than GEI.

OU-MVLP. We perform experiments on OU-MVLP which is the largest gait dataset at present, and the results are shown in Table 3. In Table 3, GaitSet(all) means using all training data for training, GaitSet(3/5) means using about 3/5 of training data for training, GaitPVD(ours)(3/5) uses the same data as GaitSet(3/5). Because the OU-MVLP is too large and there are too many small files, our experiment is limited by the capacity of the Solid State Drive and the IO bottleneck. So our model uses fewer data than other models. In this case, GaitPVD still exceeds other state-of-the-art methods. It further validates the effectiveness of GaitPVD.

4.3 Ablation Study

In order to verify the effectiveness of our method, we perform ablation study based on CASIA-B. We elaborate it from the following aspects: 1) the selection of the statistical function in TFS, 2) the selection of the pooling layer in BP, 3) the effect of view distillation.

The first six rows of Table 4 are the results of using different pooling layers in BP. It can be seen that the best effect is achieved when GeMpooling $p = 7$ is selected. In addition, the second best effect is to use Maxpooling and Avgpooling at the same time. This also shows that the most suitable fineness for recognition is between Maxpooling and Avgpooling. So choosing GeMpooling is a more appropriate way.

For the selection of statistical functions in TFS, we also perform ablation experiments. It can be seen from the 6-th to 8-th rows of Table 4 that the best effect can be achieved when max(·) is selected, while the effect of min(·)is very poor. As shown in Figure 3, max(·)can give the most comprehensive response information and min(·) has a lot of information missing.

It can be seen from the 6th and 9th rows of Table 4, in the NM and BG types, view distillation plays an important role in improving the final performance. For CL, there is a slight drop. The reason may be that wearing coats has a different influence on the silhouette images of different views. However, it can be seen from the results that it only has a slight impact on CL performance. And, it can

Table 2: Averaged Rank-1 Accuracies on CASIA-B, Excluding Identical-View Cases. The Results in the Table Omit%, the Red Means the Best Result, and the Green Means the Second Best Result

Gallery NM#1-4		0°-180°											Mean
	Probe	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	CNN-LB[2]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
#1-2	GaitSet[7]	90.8	**97.9**	**99.4**	**96.9**	**93.6**	91.7	**95.0**	**97.8**	**98.9**	**96.8**	85.8	**95.0**
	GaitNet[16]	**91.2**	92.0	90.5	95.6	86.9	**92.6**	93.5	96.0	90.9	88.8	**89.0**	91.6
	GaitPVD(ours)	**95.7**	**98.4**	**99.7**	**99.1**	**95.0**	**94.2**	**96.3**	**98.7**	**99.0**	**97.9**	**91.7**	**96.9**
BG #1-2	CNN-LB[2]	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[7]	**83.8**	**91.2**	**91.8**	88.8	**83.3**	**81.0**	84.1	90.0	**92.2**	**94.4**	79.0	**87.2**
	GaitNet[16]	83.0	87.8	88.3	**93.3**	82.6	74.8	**89.5**	**91.0**	86.1	81.2	**85.6**	85.7
	GaitPVD(ours)	**92.0**	**94.1**	**95.3**	**93.6**	**89.1**	**86.4**	**88.3**	**93.8**	**96.2**	**94.9**	**85.5**	**91.7**
CL	CNN-LB[2]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
#1-2	GaitSet[7]	**61.4**	**75.4**	**80.7**	**77.3**	**72.1**	70.1	**71.5**	**73.5**	**73.5**	**68.4**	**50.0**	**70.4**
	GaitNet[16]	42.1	58.2	65.1	70.7	68.0	**70.6**	65.3	69.4	51.5	50.1	36.6	58.9
	GaitPVD(ours)	**74.0**	**84.1**	**88.2**	**82.3**	**78.7**	**77.5**	**80.1**	**83.5**	**84.4**	**79.2**	**64.6**	**79.7**

Table 3: Averaged Rank-1 Accuracies on OU-MVLP, Excluding Identical-View Cases. The Results in the Table Omit%, the Red Means the Best Result, and the Green Means the Second Best Result

Probe	Gallery All 14 views			
	GEINet[2]	GaitSet[7](all)	GaitSet(3/5)	GaitPVD(ours)(3/5)
0°	11.4	79.5	77.8	79.4
15°	29.1	87.9	87.4	88.1
30°	41.5	89.9	89.6	90.0
45°	45.5	90.2	89.8	90.3
60°	39.5	88.1	87.6	88.5
75°	41.8	88.7	88.3	88.9
90°	38.9	87.8	87.2	87.8
180°	14.9	81.7	80.5	82.0
195°	33.1	86.7	86.3	87.0
210°	43.2	89.0	88.6	89.0
225°	45.6	89.3	88.8	89.3
240°	39.4	87.2	86.8	87.7
255°	40.5	87.8	87.2	87.9
270°	36.3	86.2	85.5	86.2
Mean	35.8	87.1	86.5	87.3

significantly improve the performance on NM and BG. So, our view distillation method deserves to be affirmed.

In order to determine the specific values of the α and λ in the loss function, we also conducted an ablation experiment. From Table 5, when α sets 0.2 and λ sets 0.3, the effect is best.

5 CONCLUSION

In this paper, we propose a part-based network, which consists of three Blocks, the Temporal Feature Selection and the Block Pooling. A new idea is given to solve cross-view gait recognition that is view knowledge distillation. And, we design the PVDLoss that can optimize the propagation of view knowledge.

Two Part-based networks and view knowledge distillation constitute GaitPVD, which is a teacher-student framework. By adjusting the input of the framework and using the separate triplet loss and the PVDLoss, GaitPVD can make the gait features under the non-normative views closer to the gait feature under the normative view to improve the recognition performance. In addition, GaitPVD can achieve state-of-the-art, *i.e.*,the average rank-1 accuracy of 96.9% on the CASIA-B and the 87.3% accuracy on the OU-MVLP.

ACKNOWLEDGMENTS

This work was supported by Anhui Province 2020 Major Science and Technology Special Project (202003a05020009).

Table 4: Ablation Experiments Performed in CASIA-B. Results Are Rank-1 Accuracies Averaged on All Views, Excluding Identical-View Cases. VD Represents View Distillation. ✓Means Adding this Operation into the Network. The Results in the Table Omit%, the Red Means the Best Result, and the Green Means the Second Best Result

Serial number	TFS			BP						VD	NM	BG	CL
	Max	Median	Min	Maxpool	Avgpool	Maxpool+Avgpool	GeM p=3	GeM p=5	GeM p=7				
1	✓			✓							95.7	89.5	79.8
2	✓				✓						95.9	89.5	78.4
3	✓					✓					95.7	90.0	79.9
4	✓						✓				95.3	89.0	79.2
5	✓							✓			95.4	90.1	79.2
6	✓								✓		**96.1**	**90.2**	**79.9**
7			✓						✓		81.2	75.0	58.7
8		✓							✓		92.8	82.2	69.0
9	✓								✓	✓	**96.9**	**91.7**	**79.7**

Table 5: Ablation Experiments about $\alpha$and λ Performed in CASIA-B. Results Are Rank-1 Accuracies Averaged on all Views, Excluding Identical-View Cases. The Results in the Table omit%

α	λ	NM	BG	CL
0.10	0.30	95.9	90.2	77.9
0.15	0.30	96.5	90.8	79.4
0.20	**0.30**	**96.9**	**91.7**	**79.7**
0.25	0.30	96.5	91.3	79.6
0.30	0.30	95.9	91.0	78.7
0.20	0.10	95.2	89.4	76.3
0.20	0.20	95.8	90.5	78.9
0.20	0.40	96.7	91.1	79.2
0.20	0.50	96.3	90.8	77.5

REFERENCES

[1] Shiraga, Kohei, *et al.* (2016). Geinet: View-invariant gait recognition using a convolutional neural network. *2016 international conference on biometrics (ICB).* IEEE.

[2] Wu, Zifeng, *et al.* (2016) "A comprehensive study on cross-view gait based human identification with deep cnns."*IEEE transactions on pattern analysis and machine intelligence* 39(2): 209-226.

[3] Bashir, Khalid, T. Xiang , and S. Gong . (2010) "Gait recognition using Gait Entropy Image." International Conference on Crime Detection & Prevention IET.

[4] Huynh-The, Thien, *et al.* (2020). Learning 3D spatiotemporal gait feature by convolutional network for person identification. *Neurocomputing* 397: 192-202.

[5] Sokolova, Anna, and Anton Konushin. (2019) "Pose-based deep gait recognition." *IET Biometrics* 8(2): 134-143.

[6] Chen, Qiang, *et al.* (2017). Feature map pooling for cross-view gait recognition based on silhouette sequence images. *2017 IEEE International Joint Conference on Biometrics (IJCB).* IEEE.

[7] Chao, Hanqing , *et al.*(2018) "GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition."

[8] He, Y.; Zhang, J.; Shan, H.; and Wang, L. (2019) Multi-task GANs for view-specific feature learning in gait recognition. IEEE TIFS 14(1):102–113

[9] Yu, S.; Chen, H.; Reyes, E. B. G.; and Poh, N. (2017). GaitGAN: Invariant gait feature extraction using generative adversarial networks. In CVPR Workshops, 532– 539.

[10] Zhang, Rui, *et al.* (2019) "Improving Cross-View Gait Recognition With Generative Adversarial Networks." Electrical Engineering and Computer Science (EECS) 3: 43-47

[11] Gu, X., Ma, B., Chang, H., Shan, S., Chen, X. (2019): Temporal knowledge propagation for image-to-video person re-identification. In: IEEE International Conference on Computer Vision.

[12] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. (2015) "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531.*

[13] Yu, Shiqi , D. Tan , and T. Tan . "A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition." 18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China IEEE.

[14] Hu, Maodi, *et al.* (2013) "View-invariant discriminative projection for multi-view gait-based human identification." IEEE Transactions on Information Forensics and Security 8(12): 2034-2045.

[15] Yang, Yazhou, *et al.* (2014) "Gait recognition using flow histogram energy image." *2014 22nd International Conference on Pattern Recognition.* IEEE.

[16] Zhang, Ziyuan, *et al.* (2019) "Gait recognition via disentangled representation learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[17] Chattopadhyay, Pratik, *et al.* (2014) "Pose Depth Volume extraction from RGB-D streams for frontal gait recognition." *Journal of Visual Communication and Image Representation* 25(1): 53-63.

[18] Liao, Rijun, *et al.* (2017) "Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations." Chinese Conference on Biometric Recognition. Springer, Cham.

[19] Sun, Yifan, *et al.* (2018) "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)." Proceedings of the European Conference on Computer Vision (ECCV).

[20] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. (2015) "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531.

[21] Gu, Xinqian, *et al.* (2019) "Temporal knowledge propagation for image-to-video person reidentification." Proceedings of the IEEE International Conference on Computer Vision.

[22] Radenovi´c, Filip, Giorgos Tolias, and Ondrej Chum. (2018) "Fine-tuning CNN image retrieval with no human annotation." IEEE transactions on pattern analysis and machine intelligence 41(7): 1655-1668.

[23] Hermans, Alexander, Lucas Beyer, and Bastian Leibe. (2017) "In defense of the triplet loss for person re-identification." arXiv preprint arXiv:1703.07737.