HID 2022: The 3rd International Competition on Human Identification at a Distance

Shiqi Yu[1], Yongzhen Huang[2,3], Liang Wang[4], Yasushi Makihara[5],
Shengjin Wang[6], Md Atiqur Rahman Ahad[7], and Mark Nixon[8]
[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]Tsinghua University, China. [7]University of East London, UK.
[8]University of Southampton, UK.
https://hid2022.iapr-tc4.org/

Abstract

The paper provides a summary of the Competition on Human Identification at a Distance 2022 (HID 2022), which is the third one in a series of competitions. HID 2022 is for promoting the research in human identification at a distance by providing a benchmark to evaluate different methods. The competition attracted 112 valid registered teams. 71 teams and 51 teams submitted their results in the first phase and the second phase, respectively. Very encouraging results have been achieved, and the accuracies of the top teams are much higher than those achieved in the previous two competitions. In this paper, we introduce the competition including the dataset, experimental settings, competition organization, results from the top teams and their analysis. The methods used by the top teams are also presented in the paper. The progress of this competition can give us an optimistic view on gait recognition.

1. Introduction

Human identification at a distance is a challenging task. Most biometric modalities such as face, fingerprint, etc. cannot be acquired at a distance easily. But human identification at a distance has great demands for improving the security of our society. Gait is the most popular biometric for human identification at a distance since it can be acquired when faces are blurred or too small and other biometric modalities cannot be perceived. Person re-identification (ReID) [4] is to associate images or videos of the same person from different cameras or from the same camera in different time. ReID is also in the scope of human identification at a distance. ReID widely uses colors and textures. But colors and textures are considered as variations in gait recognition, and also not considered in the competition.

Gait recognition has been improved greatly since it was firstly proposed in late 1990s. Especially in the recent years, the progress in deep learning greatly boosted the development of gait recognition. Encouraging results have been presented along with many innovative algorithms [1, 3, 8]. Similar to most research topics in artificial intelligence, the performance of an algorithm is affected by many factors. Different experimental settings of the same method will achieve different results. Many people – even the researchers on this topic – wonder whether HID methods can be deployed in real complex environments. So, a benchmark is needed to compare different methods fairly and to evaluate them in complex environments. The evaluation based on the benchmark can promote the research in this domain.

Following the HID 2020[1] and HID 2021[2], we organized *the third international competition on human identification at a distance* (HID 2022)[3]. The first competition on HID was held in conjunction with Asian Conference on Computer Vision (ACCV) in 2020, and the second one was held with International Joint Conference on Biometrics (IJCB) in 2021 [15]. As shown in Figure 2, the performance of this competition has been greatly improved compared with the previous two. The best one has reached to 95.9% on the challenging dataset of the competition. This paper introduces the competition, which includes dataset, evaluation protocol, competition organization details, competition results and the methods of the top teams.

The paper is organized as follows. Section 2 introduces the dataset and the evaluation metric. The organizing details for a fair competition are introduced in Section 3, with some statistics of the competition. The results of the top 10 teams

[1] https://hid2020.iapr-tc4.org/.
[2] https://hid2021.iapr-tc4.org/.
[3] https://hid2022.iapr-tc4.org/.

are presented in Section 4. The methods from the top 10 teams and the team information are in Section 5. We also discuss technologies by the top 10 teams in Section 6. The last section, Section 7, concludes the paper.

2. Experimental Settings

There are two main concerns for the competition. The first is the experiment should be challenging, and the second is the competition should be fair to all participants. For the first concern, we chose a large dataset, CASIA-E [13], to evaluate the algorithms from different participants. CASIA-E was also used in the previous two competitions: HID 2020 and HID 2021. Note that only a subset of CASIA-E was used for each competition. The training set is exactly the same with the previous two competitions. But the samples in the test set are randomly selected from the whole CASIA-E and different from the ones in HID 2020 and HID 2021. To avoid the class labels of the test samples being hacked by frequent submissions, the competition has two phases. In the first phase, March 10 to April 20, 2022, only 25% of the test samples were evaluated by the competition system. All the test samples were evaluated only in the second phase, which has only 10 days, April 21 to April 30. Whatever in the first phase or the second phase, 2 submissions were allowed for each team per day.

The detailed description of the dataset and the evaluation protocol can be found in the following part of this section.

2.1. Dataset

The CASIA-E [13] dataset was employed for the competition. CASIA-E is a novel gait dataset created by the Institute of Automation, Chinese Academy of Sciences and the company Watrix. The dataset for the competition contains 1,005 subjects[4]. 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 several 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 the company Watrix. It was the best segmentation algorithm we could found for HID 2020 in 2020. We used the same segmentation algorithm for HID 2021 and HID 2022 for fair comparisons among the three competitions. We may adapt it to an open-source segmentation algorithm in the successor competitions. All silhouette images were resized to a fixed size of 128 × 128, as shown in Figure 1. We did not remove bad quality silhouettes manually. All silhouettes are from automatic detection and segmentation algorithms. As shown in Figure 1, the silhouettes are not of perfect quality. Some noise exists as in real applications and silhouettes we presented as they are. This makes the competition more challenging. The challenges also make the competition a good simulation platform for real applications.

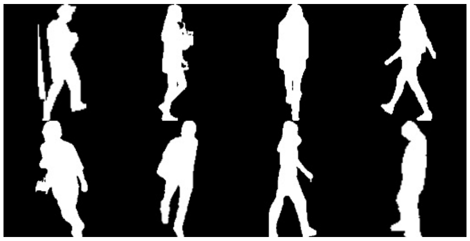


Figure 1. Example silhouette images from dataset CASIA-E.

The dataset was separated into the training set and the test set. For each subject, 10 sequences were randomly selected for the competition from the dataset. The first 500 subjects are in the training sets, and the last 505 ones are in the test set. The labels of all sequences in the training set were released to participants. But the label of only 1 sequence of each subject in the gallery set was released. The other 9 sequences were in the probe set and their labels were predicted. The data for the competition is summarized in Table 1. Since the 10 sequences of a subject were randomly selected, they should be in different views, different walking conditions and different clothing. Considering only 1 sequence is put into the gallery set for each subject, the competition is very challenging.

	Training Set Subject #1 $\sim$ #500	**Test Set** Subject #501 $\sim$ #1,005	
		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 sequences of that subject.

2.2. Performance metric

Rank 1 accuracy is used for evaluating the methods from different teams. It is straightforward and can be implemented as follows.

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

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

[4] Three subjects without data (empty folders) were deleted from the dataset. So the number of subjects is reduced to 1,005 this year from 1,008.

3. Competition Organization

The evaluation should be user-friendly and convenient for participants. It should also be safe and not hacked. To meet those requirements, we designed detailed rules as follows:

1. To avoid the ID labels of the probe set being detected by numerous submissions, we limited the number of submissions each day to 2. Only one CodaLab ID is allowed per team. Only institutional emails can be accepted to register for the competition.
2. The accuracy was evaluated automatically at CodaLab. The ranking will be updated in the scoreboard accordingly. The immediate feedback made the evaluation user-friendly.
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. The second phase was much shorter than the first phase, and there were only 10 days. The results by the whole probe set would be given in the second phase.
4. The top 10 teams in the final scoreboard need to send their programs to the organizers. The programs were ran to reproduce their results. The reproduced results should be consistent with the results shown on the CodaLab scoreboard.

We received altogether 162 registrations, and rejected all registrations with public emails such as Gmail. There were 112 valid registrations. 71 of them submitted their results to CodaLab in the first phase, and 51 teams submitted in the second phase. We also evaluated the programs from top teams to make sure that their results could be reproduced. After careful evaluations, the top 10 teams were selected, and their results are presented in the following section. The details of their methods are also briefly introduced in the paper.

4. The Results of the Top 10 Teams

The results of the top 10 teams are listed in Table 2. The best accuracy reaches 95.9%, and the average of the top 10 teams is 93.0%. For better understanding of the employed methods, we also list the most commonly used ones in Table 2. More details about the methods from those teams can be found in the next section. The analysis of the results is presented in Section 6.

The results achieved in HID 2022 obviously surpass those in HID 2021 and HID 2020 as shown in Figure 2. Note that the results in HID 2021 and HID 2020 were not correctly calculated. The number of the total probe samples is $505 \times 9 = 4,545$, not $505 \times 10 = 5,050$, which was used in HID 2020 and HID 2021. In this paper, all results in HID 2021 and HID 2020 have been calibrated to the correct ones by multiplying a factor of 1.111 (=5,050/4,545). From the results of the three competitions, the results were improved obviously year by year.

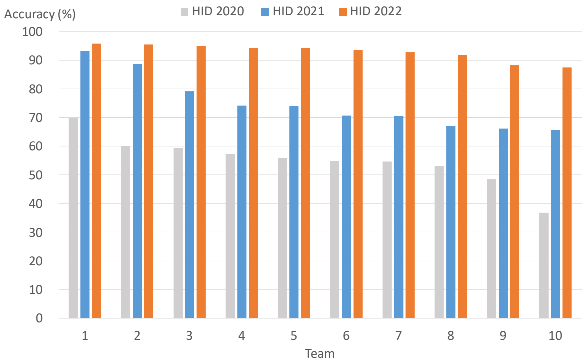


Figure 2. The top 10 result comparison among HID 2020, HID 2021 and HID 2022. The results in HID 2020 and HID 2021 have been calibrated according to the same standard with HID 2022.

5. Descriptions of the Top Methods

The method descriptions from 9 teams of the top 10 are presented in the following part of the section. One team, Team damowang, did not provide their description, but provided the modules used in the methods as shown in Table 2.

5.1. Team league

Team name: *league*
Members: Li Wang and Lichen Song (Dalian Everspry Sci&Tech Co., Ltd.) {challenge@everspry.com}
Method: The method is mainly divided into the following parts: data preprocessing, augmentation, network design, feature fusion, and ensemble learning. Data preprocessing is the same with the baseline model in OpenGait [10]. They used all the provided data and did not clean low quality data. Data augmentation includes horizontal flip, random horizontal and vertical translation. We trained two models with different augmentation methods as follows:

- Model A: Random horizontal flip and random translation for a whole sequence.
- Model B: Random horizontal flip for a whole sequence, and random translation for each image in the sequence.

A feature fusion method combines the features before and after horizontal flip with a mean operation. To achieve a better performance, ensemble learning is employed to combine the Euclidean distances from both Model A and Model

Team rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CodaLab ID	league	MingWang	ggyyll	huihuihui	damowang	YiYShuxiao	xiaohao1	RammusLeo	AIG	yzzhang
Data cleaning	×	×	✓	✓	×	✓	×	×	✓	✓
Data alignment	✓	×	✓	×	×	×	×	×	×	×
Data augmentation	✓	✓	✓	✓	✓	✓	✓	✓	✓	×
Query expansion	✓	✓	×	✓	×	×	✓	×	×	×
Re-ranking	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
Extra data	×	OUMVLP	×	×	×	×	OUMVLP	×	×	×
OpenGait	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GPU	RTX3090 × 2	V100 × 8	RTX2080ti × 4 RTX3090 × 4	RTX3090 × 4	N/A	V100S × 2	RTX2080ti × 4	RTX3090 × 1	V100 × 8	A100 × 4
Accuracy(%)	95.9	95.5	95.1	94.4	94.3	93.6	92.9	92.0	88.3	87.5

Table 2. The technologies used by the top 10 teams and their accuracies in HID 2022.

Layer Name	In_C, Out_C	Kernel	Stride
Conv-LeakyReLU	1, 64	(5, 5)	1
Conv-LeakyReLU	64, 64	(3, 3)	1
MaxPool	64, 64	(2, 2)	2
Conv-LeakyReLU	64, 128	(3, 3)	1
Conv-LeakyReLU	128, 128	(3, 3)	1
Conv-LeakyReLU	128, 128	(3, 3)	1
MaxPool	128, 128	(2, 2)	2
Conv-LeakyReLU	128, 256	(3, 3)	1
Conv-LeakyReLU	256, 256	(3, 3)	1
MaxPool	256, 256	(2, 2)	2
Conv-LeakyReLU	256, 512	(3, 3)	1
Conv-LeakyReLU	512, 512	(3, 3)	1
Set pooling	512, 512	-	-
HorizontalPoolingPyramid	512, 512	-	-
SeparateFCs	512, 256	-	-
SeparateBNNecks	256, 500	-	-

Table 3. The backbone model details from Team *league*.

B, and re-ranking [17] is also employed. The structure of the proposed network is described in Table 3.

Experimental settings: Hardware: Intel(R) Core(TM) i9-10940X CPU @ 3.30GHz CPU and GeForce RTX 3090 × 2 GPU. The hyper-parameters are almost the same with the baseline model in OpenGait, except the model structure settings.

5.2. Team GRgroup (MingWang@CodaLab)

Team name: *GRgroup*

Members: Ming Wang[1], Beibei Lin[2], Shengdi Qin[1], Yu Liu[1], Lincheng Li[2], Shunli Zhang[1], Xin Yu[3] ([1]School of Software Engineering, Beijing Jiaotong University. [2]Netease Fuxi AI Lab. [3]University of Technology Sydney.) {21121736@bjtu.edu.cn}

Method: The framework of their method is shown in Figure 3. The whole pipeline is from OpenGait [10].

In the training phase, they trained 3 backbones, including the baseline in OpenGait [10], GaitMask [7], and GaitGL [8]. The horizontal flip is used for data augmentation. In the test phase, they first used the three backbones to extract gait features. Then, they employed three strategies, query expansion [15], re-ranking [17], and vote mechanism, to improve the recognition accuracy.

The experimental results are shown in Table 4. The final recognition accuracy is 95.500%.

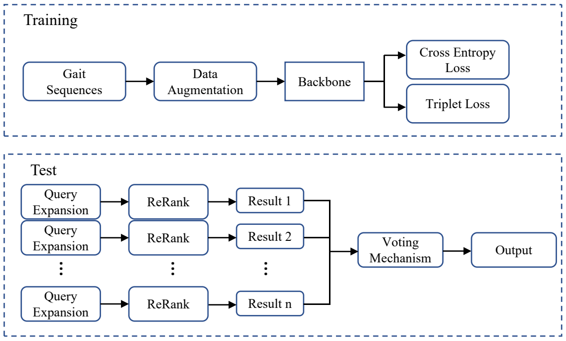


Figure 3. The framework of Team *GRgroup*.

Model	RK	QE	VM	Accuracy (%)
GaitGL	✓			94.368
GaitGL	✓	✓		94.746
Three models	✓	✓	✓	95.500

Table 4. Rank-1 accuracy (%) of different techniques. RK, QE and VM are for re-ranking, query expansion and vote mechanism respectively.

Experimental settings: The channels of different layers in all the 3 networks are set to 64, 128, 256 and 512. The images were normalized to 128 × 88. The batch size was set to 16 × 8 in the training phase. All experiments took SGD as the optimizer. The learning rate was 0.1, and the number of iterations was $60K$. Note that they firstly train their 3 models using OUMVLP dataset [14] and then fine-tuned them using the competition dataset.

5.3. Team ggyyll

Team name: *ggyyll*

Members: Xiaohui Xu[1], Huang Huang[1], Lian Zhang[1], Guohe Li[2], Guoqing Gao[2], Fei Suo[2] and Rui Xu[3] ([1]InChao Institute, Ltd. [2]State Grid Xinyuan Group Co., Ltd. [3]Xiangtan University.) {201921001086@smail.xtu.edu.cn}

Method: This method mainly focuses on data processing and parameter optimization. The data processing includes data cleaning, data alignment and data augment. Data cleaning can remove some noisy and distorted silhouettes. Data alignment can adjust the images and make the body parts aligned to the same location. The parameter optimization is mainly on finding out best checkpoint and re-ranking parameters. The entire pipeline shown in Figure 4 that contains 3 parts: (1) data cleaning and data alignment, (2) The OpenGait baseline model training with the help of data augmentation, and (3) recognition with TTAs (testing time augment) and re-rank.

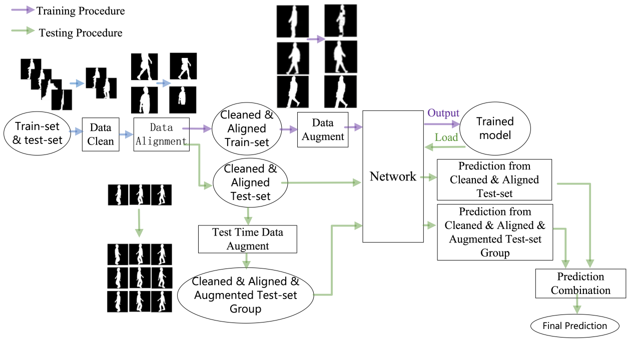

Training Procedure
Testing Procedure
Train-set & test-set
Data Clean
Data Alignment
Cleaned & Aligned Train-set
Data Augment
Network
Output
Load
Trained model
Cleaned & Aligned Test-set
Test Time Data Augment
Cleaned & Aligned & Augmented Test-set Group
Prediction from Cleaned & Aligned Test-set
Prediction from Cleaned & Aligned & Augmented Test-set Group
Prediction Combination
Final Prediction

Figure 4. The method pipeline of Team *ggyyll*

Experimental settings: They used the default configurations of the baseline in OpenGait. The configurations were image size=128 × 128, learning rate=0.1 and batch size=[16,8]. But they also changed some configurations. The learning rate reduction scheduler was [20000,40000,70000], total iteration was 80000 and store iteration period was 2500. About the parameter optimization, they chose the weights trained at step 72,500.

5.4. Team SCUT-BIPLAB (huihuihui@CodaLab)

Team name: *SCUT-BIPLAB*
Members: Xin Wang, Hui Fu, Yuxuan Zhang and Wenxiong Kang (South China University of Technology) {202121018458@mail.scut.edu.cn}
Method: They employed the baseline model in OpenGait [10], and the following tricks were employed to improve the performance.

- Data augmentation: The method used a data augmentation strategy of random horizontal flipping in the training phase to improve the robustness of the model in the walking direction.

- Data cleaning: It explored a simple strategy to filter out some of low quality silhouettes by their ratios of foreground pixels.

- A wider model: They made the baseline model wider from 64-128-256-512 channels to 128-256-512-1024 by using more kernels.

- Remove the last pooling layer: For more fine-grained features, it removed the last pooling layer to obtain the feature map with a higher spatial resolution of 32×22 instead of 16×11.

- Multi-scale feature supervision: Embeddings of different scale features can be extracted by multiple branches, and the losses for the embeddings at different scales can be calculated as shown in Figure 5, respectively.

- Generalized-Mean Pooling: HPP [4] is replaced with Generalized Mean Pooling (GeM) [11] to integrate the spatial information adaptively.

- Re-ranking [17].

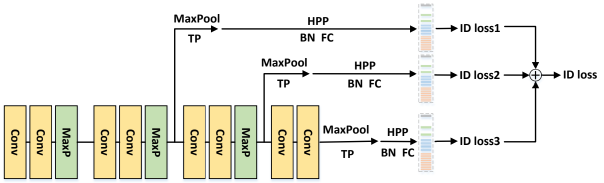


Figure 5. Multi-scale feature supervision of Team *SCUT-BIPLAB*

Experimental settings: All ablation experiments listed in Table 5 follow the same settings, e.g., the batch size is all 8 × 4.

Model	Training data	Test data	Acc-P1	Acc-P2
baseline	E	E	77.89	84.63
baseline	E-flip	E	79.40	-
baseline	E-flip	E-flip	83.28	-
baseline	E-clean	E-clean	-	86.36
baseline	E-clean-flip	E-clean-flip	-	91.11
wider	E	E	-	85.77
lastpool	E-flip	E+E-flip	-	91.43
multi-scale	E-clean-flip	E-clean-flip	-	92.46
GeM	E	E	78.69	85.63

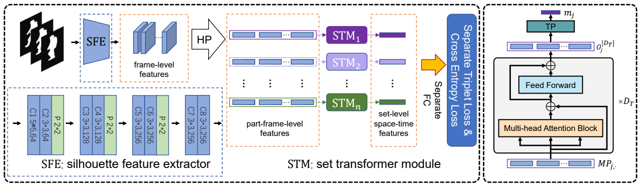
Table 5. Results of the ablation study. The E, E-flip, E-clean, and E-clean-flip represent the CASIA-E, CASIA-E with random horizontal flipping, CASIA-E with data cleaning, and CASIA-E with both data cleaning and random horizontal flipping, respectively. The Acc-P1 and Acc-P2 mean the test accuracy obtained by the first phase and the second phase of the competition, respectively.

5.5. Team SetTrans (YiYShuxiao@CodaLab)

Team name: *SetTrans*
Members: Xianchun Wang, Guodong Li, Lijun Guo and Rong Zhang (Ningbo University) {2011082328@nbu.edu.cn}
Method: In this competition, they designed a network called SetTrans, which derives from the set transformer

module (STM), proposed in [6]. STM performs a temporal aggregation operation for obtaining set-level spatio-temporal features from an image sequence. Also, the multi-head attention mechanism in STM helps to extract more abundant movement patterns on different time scales.

The structure of SetTrans is shown in Figure 6(a). The feature extraction modules (SFE) is a combination of CNN and maxpooling layers to get frame-level features. By using horizontal segmentation, it can get the part-frame-level features. For each part, it used a STM to extract movement patterns on different time scales of the gait sequence and obtain spatio-temporal fine-grained features through temporal aggregation. The structure of STM is given in Figure 6(b).



(a) The structure of SetTrans (b) The structure of STM

Figure 6. The proposed network SetTrans by Team *SetTrans*.

Experimental settings: (1) The input size of the silhouettes was 128×88. They randomly took 30 frames of silhouettes from the sequence during each training epoch. Adam optimizer was used with the learning rate of 1e-4. The margin of the triplet loss was set to 0.2. (2) In baseline of OpenGait [10]: The batch size was set as (8,8), the number of training epochs was 200K, and the learning rate would be reduced to 1e-5 at 80K iterations. (3) In SetTrans: The batch size was set to (7,8), the iterations were set to 100K, and the learning rate was reduced to 1e-5 at 60K iterations. The entire method contains three steps:

- Data clean: They manually selected 2,134 silhouettes (including 292 inferior quality images) to train a classification network MobileNetV2, which was used to remove those inferior quality silhouettes in CASIA-E.
- Data augmentation: It randomly flipped the images in each sequence horizontally to do data augmentation.
- Feature extraction: In this competition, they finally used the baseline and the SetTrans to extract features and concatenated the feature vectors from the two networks as the distinguishing features. Table 6 shows five experimental results, which prove the effectiveness of each trick they explored.

5.6. Team xiaohao1

Team name: *xiaohao1*
Members: Haomiao Li and Xianglei Xing (Harbin Engineering University) {lhm@hrbeu.edu.cn}
Method: The method includes the following techniques:

Method	Acc. (%)
SetTrans	87.533
SetTrans+Data Clean	87.797
SetTrans+Data Clean+Randomly Flip	91.874
SetTrans+Baseline+Data Clean+Randomly Flip	93.634

Table 6. Ablation study of Team *SetTrans*.

- Data augmentation: It flipped the silhouette images horizontally to enhance the data.
- Feature extraction: It used the input order independent model, GaitSet [1], for the experiments. In addition, the model was pre-trained on OUMVLP.
- Re-ranking: Re-ranking [17] can be based on feature similarity or adjacent sample similarity. In general, the re-ranking method based on the similarity of adjacent samples has better performance, and they chose it.
- Query expansion: In order to improve the accuracy of recognition, they used the query expand method.

Datasets	Epoch	T Frames	Batch Size
OUMVLP	25K	30	16*8
CASIA-E	100k	30	16*8

Table 7. Training details in the experiments of Team *xiaohao1*

Experimental settings: The training details are listed in Table 7 on the OUMVLP database. The network structure of GaitSet they used is shown in Table 8. All experiments take Adam as the optimizer and the learning rate is 1e-4.

Layer Name	In$_C$	Out$_C$	Kernel
setblock1	1	64	3
setblock2	64	128	3
setblock3	128	256	3
setblock4	256	512	3
glblock1	1	64	3
glblock2	64	128	3
glblock3	128	256	3
glblock4	256	512	3
FcS	512	256	-

Table 8. The network structure of GaitSet used by Team *xiaohao1*

5.7. Team BNU II (RammusLeo@ColaLab)

Team name: *BNU II*
Members: Zhenye Luo and Ao Li (Beijing Normal University) {202011081020@mail.bnu.edu.cn}
Method: The method is based on the baseline in OpenGait [10]. The following techniques were employed to improve the performance.

- Data augmentation: All the silhouettes in CASIA-E were flipped horizontally.

- New ResNet module: Some convolution layers in the baseline of OpenGait were changed to a new ResNet module, which is shown in Figure 7. The core idea of the design is to extract global information with basic convolution layers, and extract local feature information with focal convolution layers. The new ResNet module designed could also improve the training speed and prevent vanishing gradient.

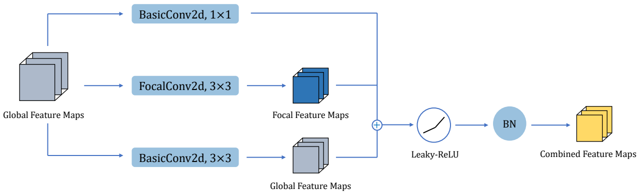


Figure 7. The proposed ResNet module by Team *BNU II*.

The proposed network structure is divided into four stages. The network structure is shown in Table 9. The setting of the halving parameter is inspired by GaitPart [3].

Block	Layer Name	In_C, Out_C	Halving
Block1	2*(Conv+BN)	1,64	-
	Maxpool	64,64	-
Block2	3*(Res+BN)	64,128	2
	Maxpool	128,128	-
Block3	3*(Res+BN)	128,256	3
	Maxpool	256,256	-
Block4	2*(Conv+BN)	256,512	-
-	SeparateFCs	512,256	-
-	SeparateBNNecks	256,500	-

Table 9. The proposed network structure by Team *BNU II*.

Experimental settings: The data for the experiments was from CASIA-E. Besides the following hyper-parameters, the rest of the hyper-parameters are the same with the default ones of of the baseline in OpenGait: The hyper-parameters changed: num_works = 2, evaluator_cfg/sampler/batch_size = 4, milestones={20000, 40000, 60000}, total_iter = 70000, trainer_cfg/sampler/batch_size = {-8, -4}.

5.8. Team AIG

Team name: *AIG*
Members: Changyuan Zhou (Onewo Space-Tech Service Co., Ltd.) {zhoucy26@vanke.com}
Method: This method is mainly implemented based on the baseline in OpenGait [10]. The highlights of the implementation are as follows:

- Data preprocessing: Two models, a MobileNetV2 and a YOLOv5, were trained to recognize whether images are high-quality or low-quality.

- Feature extraction: To enhance the model's representation capability, the numbers of all convolutional layers of the baseline in OpenGait were doubled.

- Re-ranking: Re-ranking technique [17] is exploited as post-processing in this model.

Experimental settings: They trained the model with the selected good quality silhouettes. The model was trained from the scratch, and no model fine-tuning was employed. During the training, the optimizer was SGD with a total of 70k iterations. The batch size was 8, and the number of sequence for each subject was 8. The initial learning rate was 0.1, and it decreased by the factor of 0.1 for every 20k iterations.

5.9. Team pami-gait (yzzhang@CodaLab)

Team name: *pami-gait*
Members: Yuzhen Zhang and Jingqi Li (Fudan University) {21210240434@m.fudan.edu.cn}
Method: Before the training, they preprocessed the data and only kept good quality data. The raw data was classified by a ResNet18 into good quality samples and bad quality samples. They employed a simple but effective model, which contains eight conv2d layers. Both triplet loss and cross-entropy loss were used in the experiments. Considering the size of CASIA-E (which is larger dataset, having a lot of variations in it), we set 8 conv2d layers and the number of channels were set to 64, 64, 128, 128, 256, 256, 512 and 512 respectively. The settings followed the baseline in OpenGait [10]. They also fine-tuned the pre-trained model for 20,000 iterations with the cleaned data for efficiency. This model was optimized by SGD, and the learning rate was set to a constant 1.0e-3 in the training process.

6. Analysis

In this section we analyze the different modules one by one by comparing the performance in Table 2.

- **Data cleaning**: It seems that data cleaning is not effective as expected. Only half of the 10 teams used it. The best two teams, Team 1 and Team 2, did not even use it in their methods. Since both the test data and the training data are noisy, models trained on noisy data can be robust. But we cannot say data cleaning is not useful. It should be investigated more deeply.

- **Data alignment**: Data alignment in the spatial domain or in temporal domain should be helpful. But only Team 1 and Team 3 used it in their methods. According to the different walking speeds of different subjects and the noisy silhouettes, data alignment is not so easy to implement.

- **Data augmentation**: Unsurprisingly, almost all teams (9 of 10) employed data augmentation. Data augmentation has become a standard prepossessing step for most deep learning tasks nowadays. It can enrich the samples and make the trained models robust to many variations.

- **Query expansion**: Query expansion can improve the accuracy by combining highly ranked samples from an original query into an expanded query that is then reissued [5]. Four teams, including the first two, chose this technology.
- **Re-ranking**: As stated in [15], Re-ranking [17] can obviously improve the accuracy . All 10 teams employed re-ranking in their experiments. Re-ranking can bring more computational cost. Since the competition did not evaluate the computational cost, the participants could use re-ranking even heavy models.
- **OpenGait**: All teams used OpenGait [10] to design their methods. OpenGait is based on PyTorch and is a flexible and extensible framework for gait recognition. Some popular gait recognition methods have been implemented in OpenGait. Researchers can focus their attentions on algorithm designing since OpenGait can provide fair and easy comparisons with start-of-the-art methods. OpenGait also provides a baseline deep model. We also noticed that many teams designed their own methods based on the baseline model in OpenGait, not on GaitSet [1] as in HID 2021 [15].
- **Extra Data**: Large datasets should be essential for model training, but only 2 teams trained their models with extra data (it was allowed in the competition). A useful extra dataset should be large and may contain more variations than CASIA-E dataset. There are not so many those kinds of public datasets. It may be the reason of fewer teams using extra data.
- **GPU**: Deep model training heavily depends on good hardware. All the 9 teams (one team did not provide) had very powerful GPUs for model training. Some even used 8 of the latest GPUs.

In the competition the participants would try their best to achieve good accuracies. The brief description in the previous section and the analysis in this section can only cover some main technologies but cannot cover all details. We have encouraged the participants to open-source their implementations and publish their methods.

We still believe good algorithms should be developed and evaluated with a large dataset. To collect a large dataset and label all samples are very challenging. One of the large dataset is OUMVLP [14], which was collected in a controlled indoor environment. There are two recent large datasets in the wild, GREW [18] and Gait3D [16]. GREW was collected outdoor and contains 26,345 subjects, 128,671 sequences, 14,185,478 human boxes and a distractor set contains 233,857 sequences. Gait3D contains 4,000 subjects and over 25,000 sequences extracted from 39 cameras in an unconstrained indoor scene. Gait3D provides dense 3D body shape, 2D silhouettes and body skeletons.

With the help of large datasets in the wild, deep learning based models can be wider and deeper without much concern about over-fitting. Another trend is feature extraction from the temporal domain. In GaitSet [1], temporal information is not considered but some recent works such as GaitPart [3] and GaitGL [8] have tried to extract temporal features and combined temporal features with spatial features. Considering the difficulties of collecting and labelling data, learning from unlabelled data with the help of contrastive learning [2] could also be a potential topic of interest. Detailed discussion on the future directions can be found in [9] and [12].

7. Conclusions and Plans

By their nature, competitions can be used to gauge progress and this competition demonstrates that gait can be used for identification in any increasingly challenging scenarios, and to good effect. Gait is one of the most convenient biometrics, since walking is part of daily life. Gait recognition is being improved greatly from the results of the three competitions, and the new competition has reached a new peak. These competitions show that it is more likely that gait can be deployed in a wider selection of environments since many entrants have shown that good results can be achieved, even within the short time frame imposed by this competition.

We find that the best accuracy in this competition is 95.9%. This is perhaps near the upper limit for the dataset used in the competition. To better evaluate the algorithms, a larger and more challenging dataset is required for the next competitions. Besides of the silhouette data, some other modalities, such as human skeletons, 3D mesh and 3D point cloud, can also be involved to encourage researchers to test different kinds of data for human identification at a distance.

Acknowledgements

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. Mark Nixon, Prof. Tieniu Tan and Prof. Yasushi Yagi. Their support and suggestions made the competition successful. We also acknowledge Mr. Jingzhe Ma and Mr. Rui Wang for their technical support.

References

[1] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng. Gaitset: Cross-view gait recognition through utilizing gait as a deep

set. *IEEE TPAMI*, 44(7):3467–3478, 2022. 1, 6, 8

[2] C. Fan, S. Hou, J. Wang, Y. Huang, and S. Yu. Learning gait representation from massive unlabelled walking videos: A benchmark, ARXIV.2206.13964, 2022. 8

[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 *CVPR*, pages 14213–14221, 2020. 1, 7, 8

[4] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019. 1, 5

[5] A. Gordo, F. Radenovic, and T. Berg. Attention-based query expansion learning. In *ECCV*, pages 14213–14221, 2020. 8

[6] G. Li, L. Guo, R. Zhang, J. Qian, and S. Gao. Transgait: Multimodal-based gait recognition with set transformer. *Applied Intelligence*, pages 1–13, 04 2022. 6

[7] B. Lin, X. Yu, and S. Zhang. Gaitmask: Mask-based model for gait recognition. In *BMVC*, pages 1–12, 2021. 4

[8] B. Lin, S. Zhang, and X. Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pages 14648–14656, 2021. 1, 4, 8

[9] Y. Makihara, M. S. Nixon, and Y. Yagi. *Gait Recognition: Databases, Representations, and Applications*, pages 1–13. Springer International Publishing, Cham, 2020. 8

[10] OpenGait. A flexible and extensible framework for gait recognition. https://github.com/ShiqiYu/OpenGait. 3, 4, 5, 6, 7, 8

[11] F. Radenovic, G. Tolias, and O. Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE TPAMI*, 41(7):1655–1668, 2019. 5

[12] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang. A comprehensive survey on deep gait recognition: Algorithms, datasets and challenges, ARXIV.2206.13732, 2022. 8

[13] C. Song, Y. Huang, W. Wang, and L. Wang. CASIA-E: a large comprehensive dataset for gait recognition. *IEEE TPAMI, accepted*, 2022. 2

[14] 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 Trans. on Computer Vision and Applications*, 10(4):1–14, 2018. 4, 8

[15] S. Yu, Y. Huang, L. Wang, Y. Makihara, E. B. Garcia Reyes, F. Zheng, M. A. R. Ahad, B. Lin, Y. Yang, H. Xiong, B. Huang, and Y. Zhang. HID 2021: Competition on human identification at a distance 2021. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7, 2021. 1, 4, 8

[16] J. Zheng, X. Liu, L. H. Wu Liu, C. Yan, and T. Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *CVPR*, 2022. 8

[17] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *CVPR*, pages 3652–3661, 2017. 4, 5, 6, 7, 8

[18] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu, and J. Zhou. Gait recognition in the wild: A benchmark. In *ICCV*, 2021. 8