GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition*

Hanqing Chao,[1†] **Yiwei He,**[1†] **Junping Zhang,**[1‡] **Jianfeng Feng**[2]
[1]Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science
[2]Institute of Science and Technology for Brain-inspired Intelligence
Fudan University, Shanghai 200433, China
{hqchao16, heyw15, jpzhang, jffeng}@fudan.edu.cn

Abstract

As a unique biometric feature that can be recognized at a distance, gait has broad applications in crime prevention, forensic identification and social security. To portray a gait, existing gait recognition methods utilize either a gait template, where temporal information is hard to preserve, or a gait sequence, which must keep unnecessary sequential constraints and thus loses the flexibility of gait recognition. In this paper we present a novel perspective, where a gait is regarded as a *set* consisting of independent frames. We propose a new network named *GaitSet* to learn identity information from the *set*. Based on the *set* perspective, our method is *immune to permutation* of frames, and can naturally *integrate frames from different videos* which have been filmed under different scenarios, such as diverse viewing angles, different clothes/carrying conditions. Experiments show that under normal walking conditions, our single-model method achieves an average rank-1 accuracy of 95.0% on the CASIA-B gait dataset and an 87.1% accuracy on the OU-MVLP gait dataset. These results represent new state-of-the-art recognition accuracy. On various complex scenarios, our model exhibits a significant level of robustness. It achieves accuracies of 87.2% and 70.4% on CASIA-B under bag-carrying and coat-wearing walking conditions, respectively. These outperform the existing best methods by a large margin. The method presented can also achieve a satisfactory accuracy with a small number of frames in a test sample, e.g., 82.5% on CASIA-B with only 7 frames. The source code has been released at https://github.com/AbnerHqC/GaitSet.

1 Introduction

Unlike other biometrics such as face, fingerprint and iris, gait is a unique biometric feature that can be recognized at a distance without the cooperation of subjects and intrusion to them. Therefore, it has broad applications in crime prevention, forensic identification and social security.

However, gait recognition suffers from exterior factors such as the subject's walking speed, dressing and carrying condition, and the camera's viewpoint and frame rate. There are two main ways to identify gait in literature, i.e., regarding gait as an image and regarding gait as a video sequence. The first category compresses all gait silhouettes into one image, or gait template for gait recognition (He et al. 2019; Takemura et al. 2018a; Wu et al. 2017; Hu et al. 2013). Simple and easy to implement, gait template easily loses temporal and fine-grained spatial information. Differently, the second category extracts features directly from the original gait silhouette sequences in recent years (Liao et al. 2017; Wolf, Babaee, and Rigoll 2016). However, these methods are vulnerable to exterior factors. Further, deep neural networks like 3D-CNN for extracting sequential information are harder to train than those using a single template like Gait Energy Image (GEI) (Han and Bhanu 2006).

Figure 1: From top-left to bottom-right are silhouettes of a completed period of a subject in **CASIA-B** gait dataset.

To solve these problems, we present a novel perspective which regards gait as a set of gait silhouettes. As a periodic motion, gait can be represented by a single period. In a silhouette sequence containing one gait period, it was observed that the silhouette in each ***position*** has unique appearance, as shown in Fig. 1. Even if these silhouettes are shuffled, it is not difficult to rearrange them into correct order only by observing the appearance of them. Thus, we assume the appearance of a silhouette has contained its ***position*** information. With this assumption, order information of gait sequence is not necessary and we can directly regard gait as a set to extract temporal information. We propose an end-to-end deep learning model called GaitSet whose scheme is shown in Fig. 2. The input of our model is a set of gait silhouettes. First, a CNN is used to extract frame-level features from each silhouette independently. Second, an operation called Set Pooling is used to aggregate frame-level features into a single *set-level* feature. Since this operation is applied on high-level feature maps instead of the original silhouettes, it can preserve spatial and temporal information better than gait

*This work is supported in part by National Natural Science Foundation of China (NSFC) (Grant No. 61673118) and in part by Shanghai Pujiang Program (Grant No. 16PJD009).

†H.C. and Y.H. are co-first authors.

‡Corresponding author.

template. This will be justified by the experiment in Sec. 4.3. Third, a structure called Horizontal Pyramid Mapping is used to map the set-level feature into a more discriminative space to obtain the final representation. The superiorities of the proposed method are summarized as follows:

- **Flexible** Our model is pretty flexible since there are no any constraints on the input of our model except the size of the silhouette. It means that the input set can contain any number of non-consecutive silhouettes filmed under different viewpoints with different walking conditions. Related experiments are shown in Sec. 4.4
- **Fast** Our model directly learns the representation of gait instead of measuring the similarity between a pair of gait templates or sequences. Thus, the representation of each sample needs to be calculated only once, then the recognition can be completed by calculating the Euclidean distance between representations of different samples.
- **Effective** Our model greatly improves the performance on the CASIA-B (Yu, Tan, and Tan 2006) and the OU-MVLP (Takemura et al. 2018b) datasets, showing its strong robustness to view and walking condition variations and high generalization ability to large datasets.

2 Related Work

In this section, we will give a brief survey on gait recognition and set-based deep learning methods.

2.1 Gait Recognition

Gait recognition can be grouped into template-based and sequence-based categories. Approaches in the former category first obtain human silhouettes of each frame by background subtraction. Second, they generate a gait template by rendering pixel level operators on the aligned silhouettes (Han and Bhanu 2006; Wang et al. 2012). Third, they extract the representation of the gait by machine learning approaches such as Canonical Correlation Analysis (CCA) (Xing et al. 2016), Linear Discriminant Analysis (LDA) (Bashir, Xiang, and Gong 2010) and deep learning (Shiraga et al. 2016). Fourth, they measure the similarity between pairs of representations by Euclidean distance or some metric learning approaches (Wu et al. 2017; Takemura et al. 2018a). Finally, they assign a label to the template by some classifier, e.g., nearest neighbor classifier.

Previous works generally divides this pipeline into two parts, template generation and matching. The goal of generation is to compress gait information into a single image, e.g., Gait Energy Image (GEI) (Han and Bhanu 2006) and Chrono-Gait Image (CGI) (Wang et al. 2012). In template matching approaches, View Transformation Model (VTM) learns a projection between different views (Makihara et al. 2006). (Hu et al. 2013) proposed View-invariant Discriminative Projection (ViDP) to project the templates into a latent space to learn a view-invariance representation. Recently, as deep learning performs well on various generation tasks, it has been employed on gait recognition task (Yu et al. 2017a; He et al. 2019; Takemura et al. 2018a; Shiraga et al. 2016; Yu et al. 2017b; Wu et al. 2017).

As the second category, video-based approaches directly take a sequence of silhouettes as input. Based on the way of extracting temporal information, they can be classified into LSTM-based approaches (Liao et al. 2017) and 3D CNN-based approaches (Wolf, Babaee, and Rigoll 2016; Wu et al. 2017). The advantages of these approaches are that **1)** focusing on each silhouette, they can obtain more comprehensive spatial information. **2)** They can gather more temporal information because specialized structures are utilized to extract sequential information. However, The price to pay for these advantages is high computational cost.

2.2 Deep learning on Unordered set

Most works in deep learning focus on regular input representations like sequence and images. The concept of unordered set is first introduced into computer vision by (Charles et al. 2017) (PointNet) to tackle point cloud tasks. Using unordered set, PointNet can avoid the noise and the extension of data caused by quantization, and obtain a high performance. Since then, set-based methods have been wildly used in point cloud field (Wang et al. 2018c; Zhou and Tuzel 2018; Qi et al. 2017). Recently, such methods are introduced into computer vision domains like content recommendation (Hamilton, Ying, and Leskovec 2017) and image captioning (Krause et al. 2017) to aggregate features in a form of a set. (Zaheer et al. 2017) further formalized the deep learning tasks defined on sets and characterizes the permutation invariant functions. To the best of our knowledge, it has not been employed in gait recognition domain up to now.

3 GaitSet

In this section, we describe our method for learning discriminative information from a set of gait silhouettes. The overall pipeline is illustrated in Fig. 2.

3.1 Problem Formulation

We begin with formulating our concept of regarding gait as a set. Given a dataset of N people with identities $y_i, i \in 1, 2, ..., N$, we assume the gait silhouettes of a certain person subject to a distribution $\mathcal{P}_i$ which is only related to its identity. Therefore, all silhouettes in one or more sequences of a person can be regarded as a set of n silhouettes $\mathcal{X}_i = \{x_i^j | j = 1, 2, ..., n\}$, where $x_i^j \sim \mathcal{P}_i$.

Under this assumption, we tackle the gait recognition task through 3 steps, formulated as

$$f_i = H(G(F(\mathcal{X}_i))) \tag{1}$$

where F is a convolutional network aims to extract frame-level features from each gait silhouette. The function G is a permutation invariant function used to map a set of frame-level feature to a set-level feature (Zaheer et al. 2017). It is implemented by an operation called Set Pooling (SP) which will be introduced in Sec. 3.2. The function H is used to learn the discriminative representation of $\mathcal{P}_i$ from the set-level feature. This function is implemented by a structure called Horizontal Pyramid Mapping (HMP) which will be discussed in Sec. 3.3. The input $\mathcal{X}_i$ is a tensor with four dimensions, i.e. set dimension, image channel dimension, image hight dimension, and image width dimension.

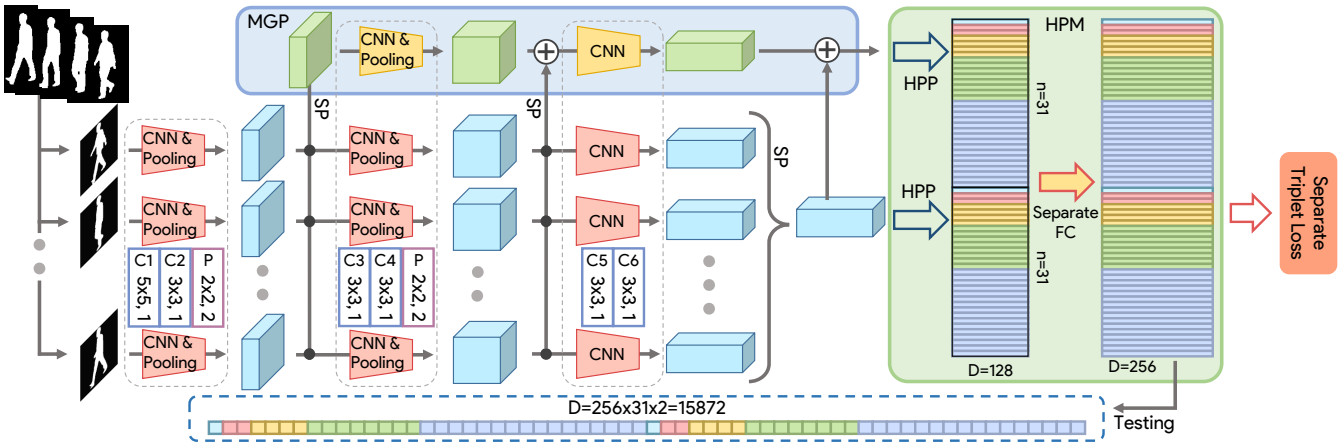

Figure 2: The framework of GaitSet. 'SP' represents Set Pooling. Trapezoids represent convolution and pooling blocks and those in the same column have the same configurations which are shown by rectangles with capital letters. Note that although blocks in MGP have same configurations with those in the main pipeline, parameters are only shared across blocks in the main pipeline but not with those in MGP. HPP represents horizontal pyramid pooling (Fu et al. 2018).

3.2 Set Pooling

The goal of Set Pooling (SP) is to aggregate gait information of elements in a set, formulated as $z = G(V)$, where z denotes the set-level feature and $V = \{v^j | j = 1, 2, ..., n\}$ denotes the frame-level features. There are two constraints in this operation. First, to take set as an input, it should be a permutation invariant function which is formulated as:

$$G(\{v^j|j = 1, 2, ..., n\}) = G(\{v^{\pi(j)}|j = 1, 2, ..., n\}) \quad (2)$$

where π is any permutation (Zaheer et al. 2017). Second, since in real-life scenario the number of a person's gait silhouettes can be arbitrary, the function G should be able to take a *set* with arbitrary cardinality. Next, we describe several instantiations of G. It will be shown in the experiments that although different instantiations of SP do have sort of influence on the performances, they do not differ greatly and all of them exceed GEI-based methods by a large margin.

Statistical Functions To meet the requirement of invariant constraint in Equ. 2, a natural choice of SP is to apply statistical functions on the set dimension. Considering the representativeness and the computational cost, we studied three statistical functions: $\max(\cdot)$, $\text{mean}(\cdot)$ and $\text{median}(\cdot)$. The comparison will be shown in Sec. 4.3.

Joint Functions We also studied two ways to join 3 statistical functions mentioned above:

$$G(\cdot) = \max(\cdot) + \text{mean}(\cdot) + \text{median}(\cdot) \quad (3)$$

$$G(\cdot) = 1\_1\text{C}(\text{cat}(\max(\cdot), \text{mean}(\cdot), \text{median}(\cdot))) \quad (4)$$

where cat means concatenate on the channel dimension, 1_1C means 1×1 convolutional layer, and $\max$, mean and median are applied on set dimension. Equ. 4 is an enhanced version of Equ. 3 where the 1×1 convolutional layer can learn a proper weight to combine information extracted by different statistical functions.

Attention Since visual attention was successfully applied in lots of tasks (Wang et al. 2018b; Xu et al. 2015; Li, Zhu, and Gong 2018), we use it to improve the performance of SP. Its structure is shown in Fig. 3. The main idea is to utilize the global information to learn an element-wise attention map for each frame-level feature map to refine it. Global

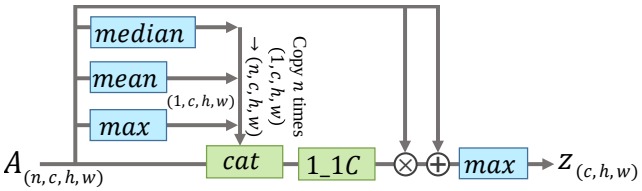

Figure 3: The structure of Set Pooling (SP) using attention. 1_1C and cat represents 1×1 convolutional layer and concatenate respectively. The multiplication and the addition are both pointwise.

information is first collected by the statistical functions in the left. Then it is fed into a 1×1 convolutional layer along with the original feature map to calculate an attention for the refinement. The final set-level feature z will be extracted by employing *MAX* on the set of the refined frame-level feature maps. The residual structure can accelerate and stabilize the convergence.

3.3 Horizontal Pyramid Mapping

In literature, splitting feature map into strips is commonly used in person re-identification task (Wang et al. 2018a; Fu et al. 2018). The images are cropped and resized into uniform size according to pedestrian size whereas the discriminative parts vary from image to image. (Fu et al. 2018) proposed Horizontal Pyramid Pooling (HPP) to deal with it. HPP has 4 scales and thus can help the deep network focus on features with different sizes to gather both local and global information. We improve HPP to make it adapt better for gait recognition task. Instead of applying a 1×1 convolutional

layer after the pooling, we use independent fully connect layers (FC) for each pooled feature to map it into the discriminative space, as shown in Fig. 4. We call it Horizontal Pyramid Mapping (HPM).

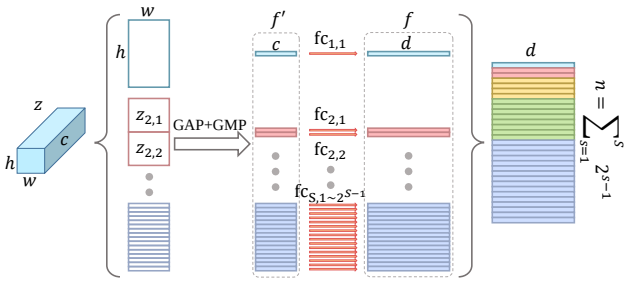

Figure 4: The structure of Horizontal Pyramid Mapping.

Specifically, HPM has S scales. On scale $s \in 1, 2, ..., S$, the feature map extracted by SP is split into 2^{s-1} strips on height dimension, i.e. $\sum_{s=1}^{S} 2^{s-1}$ strips in total. Then a Global Pooling is applied to the 3-D strips to get 1-D features. For a strip $z_{s,t}$ where $t \in 1, 2, ..., 2^{s-1}$ stands index of the strip in the scale, the Global Pooling is formulated as $f'_{s,t} = maxpool(z_{s,t}) + avgpool(z_{s,t})$, where $maxpool$ and $avgpool$ denote Global Max Pooling and Global Average Pooling respectively. Note that the functions $maxpool$ and $avgpool$ are used at the same time because it outperforms applying anyone of them alone. The final step is to employ FCs to map the features f' into a discriminative space. Since strips in different scales depict features of different receptive fields, and different strips in each scales depict features of different spatial positions, it comes naturally to use independent FCs, as shown in Fig. 4.

3.4 Multilayer Global Pipeline

Different layers of a convolutional network have different receptive fields. The deeper the layer is, the larger the receptive field will be. Thus, pixels in feature maps of a shallow layer focus on local and fine-grained information while those in a deeper layer focus on more global and coarse-grained information. The set-level features extracted by applying SP on different layers have analogical property. As shown in the main pipeline of Fig. 2, there is only one SP on the last layer of the convolutional network. To collect various-level *set* information, Multilayer Global Pipeline (MGP) is proposed. It has a similar structure with the convolutional network in the main pipeline and the set-level features extracted in different layers are added to MGP. The final feature map generated by MGP will also be mapped into $\sum_{s=1}^{S} 2^{s-1}$ features by HPM. Note that the HPM after MGP does not share parameters with the HPM after the main pipeline.

3.5 Training And Testing

Training Loss As aforementioned, the output of the network is $2 \times \sum_{s=1}^{S} 2^{s-1}$ features with dimension d. The corresponding features among different samples will be used to compute the loss. In this paper, Batch All (BA_+) triplet loss is employed to train the network (Hermans, Beyer, and Leibe 2017). A batch with size of $p \times k$ is sampled from the training set where p denotes the number of persons and k denotes the number of training samples each person has in the batch. Note that although the experiment shows that our model performs well when it is fed with the set composed by silhouettes gathered from arbitrary sequences, a sample used for training is actually composed by silhouettes sampled in one sequence.

Testing Given a query $\mathcal{Q}$, the goal is to retrieve all the *sets* with the same identity in gallery set $\mathbb{G}$. Denote the sample in $\mathbb{G}$ as $\mathcal{G}$. The $\mathcal{Q}$ is first put into GaitSet net to generate multi-scale features, followed by concatenating all these features into a final representations $\mathcal{F}_\mathcal{Q}$ as shown in Fig. 2. The same process is applied on each $\mathcal{G}$ to get $\mathcal{F}_\mathcal{G}$. Finally, $\mathcal{F}_\mathcal{Q}$ is compared with every $\mathcal{F}_\mathcal{G}$ using Euclidean distance to calculate Rank 1 recognition accuracy.

4 Experiments

Our empirical experiments mainly contain three parts. The first part compares GaitSet with other state-of-the-art methods on two public gait datasets: CASIA-B (Yu, Tan, and Tan 2006) and OU-MVLP (Takemura et al. 2018b). The Second part is ablation experiments conducted on CASIA-B. In the third part, we investigated the practicality of GaitSet in three aspects: the performance on limited silhouettes, multiple views and multiple walking conditions.

4.1 Datasets and Training Details

CASIA-B dataset (Yu, Tan, and Tan 2006) is a popular gait dataset. It contains 124 subjects (labeled in 001-124), 3 walking conditions and 11 views ($0°, 18°, ..., 180°$). The walking condition contains normal (NM) (6 sequences per subject), walking with bag (BG) (2 sequences per subject) and wearing coat or jacket (CL) (2 sequences per subject). Namely, each subject has $11 \times (6 + 2 + 2) = 110$ sequences. As there is no official partition of training and test sets of this dataset, we conduct experiments on three settings which are popular in current literatures. We name these three settings as small-sample training (ST), medium-sample training (MT) and large-sample training (LT). In ST, the first 24 subjects (labeled in 001-024) are used for training and the rest 100 subjects are leaved for test. In MT, the first 62 subjects are used for training and the rest 62 subjects are leaved for test. In LT, the first 74 subjects are used for training and the rest 50 subjects are leaved for test. In the test sets of all three settings, the first 4 sequences of the NM condition (NM #1-4) are kept in gallery, and the rest 6 sequences are divided into 3 probe subsets, i.e. NM subsets containing NM #5-6, BG subsets containing BG #1-2 and CL subsets containing CL #1-2.

OU-MVLP dataset (Takemura et al. 2018b) is so far the world's largest public gait dataset. It contains 10,307 subjects, 14 views ($0°, 15°, ..., 90°; 180°, 195°, ..., 270°$) per subject and 2 sequences (#00-01) per view. The sequences are divided into training and test set by subjects (5153 subjects for training and 5154 subjects for test). In the test set, sequences with index #01 are kept in gallery and those with index #00 are used as probes.

Table 1: Averaged rank-1 accuracies on **CASIA-B** under three different experimental settings, excluding identical-view cases.

Gallery NM#1-4			0°-180°											mean
Probe			0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
ST (24)	NM#5-6	ViDP (Hu et al. 2013)	–	–	–	59.1	–	50.2	–	57.5	–	–	–	–
		CMCC (Kusakunniran et al. 2014)	46.3	–	–	52.4	–	48.3	–	56.9	–	–	–	–
		CNN-LB (Wu et al. 2017)	54.8	–	–	77.8	–	64.9	–	76.1	–	–	–	–
		GaitSet(ours)	**64.6**	**83.3**	**90.4**	**86.5**	**80.2**	**75.5**	**80.3**	**86.0**	**87.1**	**81.4**	**59.6**	**79.5**
	BG#1-2	GaitSet(ours)	55.8	70.5	76.9	75.5	69.7	63.4	68.0	75.8	76.2	70.7	52.5	68.6
	CL#1-2	GaitSet(ours)	29.4	43.1	49.5	48.7	42.3	40.3	44.9	47.4	43.0	35.7	25.6	40.9
MT (62)	NM#5-6	AE (Yu et al. 2017b)	49.3	61.5	64.4	63.6	63.7	58.1	59.9	66.5	64.8	56.9	44.0	59.3
		MGAN (He et al. 2019)	54.9	65.9	72.1	74.8	71.1	65.7	70.0	75.6	76.2	68.6	53.8	68.1
		GaitSet(ours)	**86.8**	**95.2**	**98.0**	**94.5**	**91.5**	**89.1**	**91.1**	**95.0**	**97.4**	**93.7**	**80.2**	**92.0**
	BG#1-2	AE (Yu et al. 2017b)	29.8	37.7	39.2	40.5	43.8	37.5	43.0	42.7	36.3	30.6	28.5	37.2
		MGAN (He et al. 2019)	48.5	58.5	59.7	58.0	53.7	49.8	54.0	61.3	59.5	55.9	43.1	54.7
		GaitSet(ours)	**79.9**	**89.8**	**91.2**	**86.7**	**81.6**	**76.7**	**81.0**	**88.2**	**90.3**	**88.5**	**73.0**	**84.3**
	CL#1-2	AE (Yu et al. 2017b)	18.7	21.0	25.0	25.1	25.0	26.3	28.7	30.0	23.6	23.4	19.0	24.2
		MGAN (He et al. 2019)	23.1	34.5	36.3	33.3	32.9	32.7	34.2	37.6	33.7	26.7	21.0	31.5
		GaitSet(ours)	**52.0**	**66.0**	**72.8**	**69.3**	**63.1**	**61.2**	**63.5**	**66.5**	**67.5**	**60.0**	**45.9**	**62.5**
LT (74)	NM#5-6	CNN-3D (Wu et al. 2017)	87.1	93.2	97.0	94.6	90.2	88.3	91.1	93.8	96.5	96.0	85.7	92.1
		CNN-Ensemble (Wu et al. 2017)	88.7	95.1	98.2	96.4	**94.1**	91.5	93.9	97.5	98.4	95.8	85.6	94.1
		GaitSet(ours)	**90.8**	**97.9**	**99.4**	**96.9**	93.6	**91.7**	**95.0**	**97.8**	**98.9**	**96.8**	**85.8**	**95.0**
	BG#1-2	CNN-LB (Wu et al. 2017)	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(ours)	**83.8**	**91.2**	**91.8**	**88.8**	**83.3**	**81.0**	**84.1**	**90.0**	**92.2**	**94.4**	**79.0**	**87.2**
	CL#1-2	CNN-LB (Wu et al. 2017)	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
		GaitSet(ours)	**61.4**	**75.4**	**80.7**	**77.3**	**72.1**	**70.1**	**71.5**	**73.5**	**73.5**	**68.4**	**50.0**	**70.4**

Training Details In all the experiments, the input is a set of aligned silhouettes in size of 64×44. The silhouettes are directly provided by the datasets and are aligned based on methods in (Takemura et al. 2018b). The set cardinality in the training is set to be 30. Adam is chosen as an optimizer (Kingma and Ba 2015). The number of scales S in HPM is set as 5. The margin in BA_+ triplet loss is set as 0.2. The models are trained with 8 NVIDIA 1080TI GPUs. **1)** In CASIA-B, the mini-batch is composed by the manner introduced in Sec. 3.5 with $p = 8$ and $k = 16$. We set the number of channels in $C1$ and $C2$ as 32, in $C3$ and $C4$ as 64 and in $C5$ and $C6$ as 128. Under this setting, the average computational complexity of our model is 8.6GFLOPs. The learning rate is set to be $1e-4$. For ST, we train our model for 50K iterations. For MT, we train it for 60K iterations. For LT, we train it for 80K iterations. **2)** In OU-MVLP, since it contains 20 times more sequences than CASIA-B, we use convolutional layers with more channels ($C1 = C2 = 64, C3 = C4 = 128, C5 = C6 = 256$) and train it with larger batch size ($p = 32, k = 16$). The learning rate is $1e-4$ in the first 150K iterations, and then is changed into $1e-5$ for the rest of 100K iterations.

4.2 Main Results

CASIA-B Tab. 1 shows the comparison between the state-of-the-art methods [1] and our GaitSet. Except of ours, other results are directly taken from their original papers. All the results are averaged on the 11 gallery views and the identical views are excluded. For example, the accuracy of probe view $36°$ is averaged on 10 gallery views, excluding gallery view $36°$. An interesting pattern between views and accuracies can be observed in Tab. 1. Besides $0°$ and $180°$, the accuracy of $90°$ is a local minimum value. It is always worse than that of $72°$ or $108°$. The possible reason is that gait information contains not only those parallel to the walking direction like stride which can be observed most clearly at $90°$, but also those vertical to the walking direction like a left-right swinging of body or arms which can be observed most clearly at $0°$ or $180°$. So, both parallel and vertical perspectives lose some part of gait information while views like $36°$ or $144°$ can obtain most of it.

Small-Sample Training (ST) Our method achieves a high performance even with only 24 subjects in the training set and exceed the best performance reported so far (Wu et al. 2017) over 10 percent on the views they reported. There are mainly two reasons. **1)** As our model regards the input as a set, images used to train the convolution network in the main pipeline are dozens of times more than those models based on gait templates. Taking a mini-batch for an example, our model is fed with $30 \times 128 = 3840$ silhouettes while under the same batch size models using gait templates can only get 128 templates. **2)** Since the sample *sets* used in training phase are composed by frames selected randomly from the sequence, each sequence in the training set can generate multiple different *sets*. Thus any units related to set feature learning like MGP and HPM can also be trained well.

Medium-Sample Training (MT) & Large-Sample Training (LT) Tab. 1 shows that our model obtains very nice results on the NM subset, especially on LT where results of all views except $180°$ are over 90%. On the BG and CL subsets, although the accuracies of some views like $0°$ and $180°$ are still not high, the mean accuracies of our model exceed those of other models for at least 18.8%.

OU-MVLP Tab. 2 shows our results. As some of the previous works did not conduct experiments on all 14 views, we list our results on two kinds of gallery sets, i.e. all 14 views and 4 typical views ($0°, 30°, 60°, 90°$). All the results are averaged on the gallery views and the identical views are excluded. The results show that our methods can generalize well on the dataset with such a large scale and wide

[1] Since (Wu et al. 2017) proposed more than one model, the most competitive results under different experimental settings are cited.

Table 2: Averaged rank-1 accuracies on **OU-MVLP**, excluding identical-view cases. GEINet: (Shiraga et al. 2016). 3in+2diff: (Takemura et al. 2018a)

Probe	Gallery All 14 Views		Gallery 0°, 30°, 60°, 90°		
	GEINet	Ours	GEINet	3in+2diff	Ours
0°	11.4	**79.5**	8.2	25.5	**77.7**
15°	29.1	**87.9**	-	-	**86.3**
30°	41.5	**89.9**	32.3	50.0	**86.9**
45°	45.5	**90.2**	-	-	**89.1**
60°	39.5	**88.1**	33.6	45.3	**85.3**
75°	41.8	**88.7**	-	-	**87.6**
90°	38.9	**87.8**	28.5	40.6	**83.5**
180°	14.9	**81.7**	-	-	**80.5**
195°	33.1	**86.7**	-	-	**82.8**
210°	43.2	**89.0**	-	-	**87.2**
225°	45.6	**89.3**	-	-	**86.8**
240°	39.4	**87.2**	-	-	**85.4**
255°	40.5	**87.8**	-	-	**85.7**
270°	36.3	**86.2**	-	-	**85.0**
mean	35.8	**87.1**	-	-	**85.0**

view variation. Further, since representation for each sample only needs to be calculated once, our model can complete the test (containing 133780 sequences) in only 7 minutes with 8 NVIDIA 1080TI GPUs. It is note worthy that since some subjects miss several gait sequences and we did not remove them from the probe, the maximum of rank-1 accuracy cannot reach 100%. If we ignore the cases which have no corresponding samples in the gallery, the average rank-1 accuracy of all probe views is 93.3% rather than 87.1%.

4.3 Ablation Experiments

Tab. 3 shows the thorough results of ablation experiments. The effectiveness of every innovation in Sec. 3 is studied.

Set VS. GEI The first two lines of Tab. 3 show the effectiveness of regarding gait as a set. With fully identical networks, the result of using set exceeds that of using GEI by more than 10% on NM subset and more than 25% on CL subset. The only difference is that in GEI experiment, gait silhouettes are averaged into a single GEI before being fed into the network. There are mainly two reasons for this phenomenal improvement. **1)** Our SP extracts the set-level feature based on high-level feature map where temporal information can be well preserved and spatial information has been sufficiently processed. **2)** As mentioned in Sec. 4.2, regarding gait as a set enlarges the volume of training data.

Impact of SP In Tab. 3, the results from the third line to the eighth line show the impact of different SP strategies. SP with attention, 1×1 convolution (1_1C) joint function and $\max(\cdot)$ obtain the highest accuracy on the NM, BG, and CL subsets respectively. Considering SP with $\max(\cdot)$ also achieved the second best performance on the NM and BG subset and has the most concise structure, we choose it as SP in the final version of GaitSet.

Impact of HPM and MGP The second and the third lines of Tab. 3 compare the impact of independent weight in HPM. It can be seen that using independent weight improves the accuracy by about 2% on each subset. In the experiments, we also find out that the introduction of independent weight helps the network converge faster. The last two lines of Tab. 3 show that MGP can bring improvement on all three test subsets. This result is consistent the theory mentioned in Sec. 3.4 that set-level features extracted from different layers of the main pipeline contain different valuable information.

4.4 Practicality

Due to the flexibility of set, GaitSet has great potential in more complicated practical conditions. In this section, we investigate the practicality of GaitSet through three novel scenarios. **1)** How will it perform when the input set only contains a few silhouettes? **2)** Can silhouettes with different views enhance the identification accuracy? **3)** Whether can the model effectively extract discriminative representation from a set containing silhouettes shot under different walking conditions. It is worth noting that we did not retrain our model in these experiments. It is fully identical to that in Sec. 4.2 with setting LT. Note that, all the experiments containing random selection in this section are ran for 10 times and the average accuracies are reported.

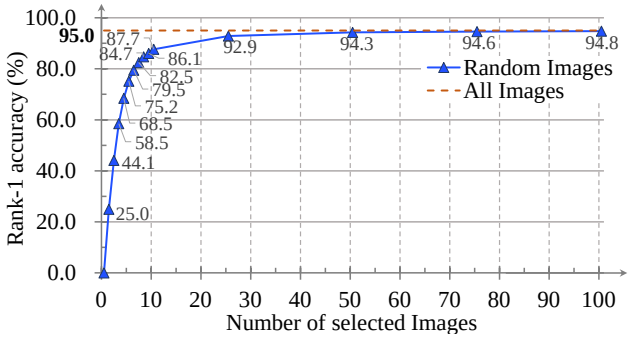

Figure 5: Average rank-1 accuracies with constraints of silhouette volume on **CASIA-B** using setting LT. Accuracies are averaged on all 11 views excluding identical-view cases, and the final reported results are averaged across 10 times experiments.

Limited Silhouettes In real forensic identification scenarios, there are cases that we do not have a continuous sequence of a subject's gait but only some fitful and sporadic silhouettes. We simulate such a circumstance by randomly selecting a certain number of frames from sequences to compose each sample in both gallery and probe. Fig. 5 shows the relationship between the number of silhouettes in each input set and the rank-1 accuracy averaged on all 11 probe views. Our method attains an 82% accuracy with only 7 silhouettes. The result also indicates that our model makes full use of the temporal information of gait. Since **1)** the accuracy rises monotonically with the increase of the number of silhouettes. **2)** The accuracy is close to the best performance when the samples contain more than 25 silhouettes. This number is consistent with the number of frames that one gait period contains.

Multiple Views There are conditions that different views of one person's gait can be gathered. We simulate these scenarios by constructing each sample with silhouettes selected

Table 3: Ablation experiments conducted on **CASIA-B** using setting LT. Results are rank-1 accuracies averaged on all 11 views, excluding identical-view cases. The numbers in brackets indicate the second highest results in each column.

GEI	Set	Set Pooling						HPM weight		MGP	NM	BG	CL
		Max	Mean	Median	Joint sum 3	Joint 1_1C 4	Attention	Shared	Independent				
√								√			80.4	68.1	40.8
	√	√						√			91.3	82.3	67.1
	√	√							√		93.2	84.7	(70.2)
	√		√						√		90.0	79.5	57.1
	√			√					√		89.5	78.1	53.5
	√				√				√		92.4	82.8	63.4
	√					√			√		93.3	(85.7)	66.3
	√						√		√		(93.7)	84.2	69.4
	√	√							√	√	**95.0**	**87.2**	**70.4**

from two sequences with the same walking condition but different views. To eliminate the effects of silhouette number, we also conduct an experiment in which the silhouette number is limited to 10. Specifically, in the contrast experiments of single view, an input set is composed by 10 silhouettes from one sequence. In the two-view experiment, an input set is composed by 5 silhouettes from each of two sequences. Note that in this experiment, only probe samples are composed by the way discussed above, whereas sample in the gallery is composed by all silhouettes from one sequence.

Table 4: Multi-view experiments conducted on **CASIA-B** using setting LT. Cases where the probe contains the view of the gallery are excluded.

View difference	18°/ 162°	36°/ 144°	54°/ 126°	72°/ 108°	90°	Single view
All silhouettes	97.0	97.9	98.7	99.1	99.0	95.0
10 silhouettes	87.9	90.6	92.7	93.7	93.7	87.7

Tab. 4 shows the results. As there are too many view pairs to be shown, we summarize the results by averaging accuracies of each possible view difference. For example, the result of 90° difference is averaged by accuracies of 6 view pairs (0°&$90^\circ, 18^\circ$&$108^\circ, ..., 90^\circ$&180°). Further, the 9 view differences are folded at 90° and those larger than 90° are averaged with the corresponding view differences less than 90°. For example, the results of 18° view difference are averaged with those of 162° view difference. It can be seen that our model can aggregate information from different views and boost the performance. This can be explained by the pattern between views and accuracies that we have discussed in Sec. 4.2. Containing multiple views in the input set can let the model gather both parallel and vertical information, resulting in performance improvement.

Multiple Walking Conditions In real life, it is highly possible that gait sequences of the same person are under different walking conditions. We simulate such a condition by forming input set with silhouettes from two sequences with same view but different walking conditions. We conduct experiments with different silhouette number constraints. Note that in this experiment, only probe samples are composed by the way discussed above. Any sample in the gallery is constituted by all silhouettes from one sequence. What's more, the probe-gallery division of this experiment is different. For each subject, sequences NM #02, BG #02 and CL #02 are kept in the gallery and sequences NM #01, BG #01 and CL #01 are used as probe.

Table 5: Multiple walking condition experiments conducted on **CASIA-B** using setting LT. Results are rank-1 accuracies averaged on all 11 views, excluding identical-view cases. The numbers in brackets indicate the constraints of silhouette number in each input set.

NM(10)	81.5	NM(10)+BG(10)	87.9	NM(20)	89.8
BG(10)	77.1	NM(10)+CL(10)	85.8	BG(20)	84.1
CL(10)	74.4	BG(10)+CL(10)	84.6	CL(20)	82.6

Tab. 5 shows the results. First, the accuracies will still be boosted with the increase of silhouette number. Second, when the number of silhouettes are fixed, the results reveal relationships between different walking conditions. Silhouettes of BG and CL contain massive but different noises, which makes them complementary with each other. Thus, their combination can improve the accuracy. However, silhouettes of NM contain few noises, so substituting some of them with silhouettes of other two conditions cannot bring extra information but only noises and can decrease the accuracies.

5 Conclusion

In this paper, we presented a novel perspective that regards gait as a set and thus proposed a GaitSet approach. The GaitSet can extract both spatial and temporal information more effectively and efficiently than those existing methods regarding gait as a template or sequence. It also provide a novel way to aggregate valuable information from different sequences to enhance the recognition accuracy. Experiments on two benchmark gait datasets has indicated that compared with other state-of-the-art algorithms, GaitSet achieves the highest recognition accuracy, and reveals a wide range of flexibility on various complex environments, showing a great potential in practical applications. In the future, we will investigate a more effective instantiation for Set Pooling (SP) and further improve the performance in complex scenarios.

References

Bashir, K.; Xiang, T.; and Gong, S. 2010. Gait recognition without subject cooperation. *Pattern Recognition Letters* 31(13):2052–2060.

Charles, R. Q.; Su, H.; Kaichun, M.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 77–85.

Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; and Huang, T. 2018. Horizontal pyramid matching for person re-identification. *ArXiv:1804.05275*.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *NIPS*, 1024–1034.

Han, J., and Bhanu, B. 2006. Individual recognition using gait energy image. *IEEE TPAMI* 28(2):316–322.

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.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *ArXiv:1703.07737*.

Hu, M.; Wang, Y.; Zhang, Z.; Little, J. J.; and Huang, D. 2013. View-invariant discriminative projection for multi-view gait-based human identification. *IEEE TIFS* 8(12):2034–2045.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. *ICLR*.

Krause, J.; Johnson, J.; Krishna, R.; and Fei-Fei, L. 2017. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, 3337–3345.

Kusakunniran, W.; Wu, Q.; Zhang, J.; Li, H.; and Wang, L. 2014. Recognizing gaits across views through correlated motion co-clustering. *IEEE TIP* 23(2):696–709.

Li, W.; Zhu, X.; and Gong, S. 2018. Harmonious attention network for person re-identification. In *CVPR*.

Liao, R.; Cao, C.; Garcia, E. B.; Yu, S.; and Huang, Y. 2017. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Chinese Conference on Biometric Recognition*, 474–483. Springer.

Makihara, Y.; Sagawa, R.; Mukaigawa, Y.; Echigo, T.; and Yagi, Y. 2006. Gait recognition using a view transformation model in the frequency domain. In *ECCV*, 151–163. Springer.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 5099–5108.

Shiraga, K.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2016. GEINet: View-invariant gait recognition using a convolutional neural network. In *ICB*, 1–8.

Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2018a. On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE TCSVT* 28(1):1–13.

Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2018b. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ TCVA* 10(4):1–14.

Wang, C.; Zhang, J.; Wang, L.; Pu, J.; and Yuan, X. 2012. Human identification using temporal information preserving gait template. *IEEE TPAMI* 34(11):2164–2176.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018a. Learning discriminative features with multiple granularities for person re-identification. *ArXiv:1804.01438*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018b. Non-local neural networks. In *CVPR*.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2018c. Dynamic graph CNN for learning on point clouds. *ArXiv:1801.07829*.

Wolf, T.; Babaee, M.; and Rigoll, G. 2016. Multi-view gait recognition using 3D convolutional neural networks. In *ICIP*, 4165–4169.

Wu, Z.; Huang, Y.; Wang, L.; Wang, X.; and Tan, T. 2017. A comprehensive study on cross-view gait based human identification with deep CNNs. *IEEE TPAMI* 39(2):209–226.

Xing, X.; Wang, K.; Yan, T.; and Lv, Z. 2016. Complete canonical correlation analysis with application to multi-view gait recognition. *Pattern Recognition* 50:107–117.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

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

Yu, S.; Chen, H.; Wang, Q.; Shen, L.; and Huang, Y. 2017b. Invariant feature extraction for gait recognition using only one uniform model. *Neurocomputing* 239:81–93.

Yu, S.; Tan, D.; and Tan, T. 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, volume 4, 441–444.

Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; and Smola, A. J. 2017. Deep sets. In *NIPS*, 3391–3401.

Zhou, Y., and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3D object detection. *CVPR*.