Gait Lateral Network: Learning Discriminative and Compact Representations for Gait Recognition

Saihui Hou[1,3], Chunshui Cao[3], Xu Liu[2,3], and Yongzhen Huang[1,3](✉)

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
yzhuang@nlpr.ia.ac.cn
[2] Beijing University of Technology, Beijing, China
[3] WATRIX AI, Beijing, China

Abstract. Gait recognition aims at identifying different people by the walking patterns, which can be conducted at a long distance without the cooperation of subjects. A key challenge for gait recognition is to learn representations from the silhouettes that are invariant to the factors such as clothing, carrying conditions and camera viewpoints. Besides being discriminative for identification, the gait representations should also be compact for storage to keep millions of subjects registered in the gallery. In this work, we propose a novel network named Gait Lateral Network (GLN) which can learn both *discriminative* and *compact* representations from the silhouettes for gait recognition. Specifically, GLN leverages the inherent feature pyramid in deep convolutional neural networks to enhance the gait representations. The silhouette-level and set-level features extracted by different stages are merged with the lateral connections in a top-down manner. Besides, GLN is equipped with a *Compact Block* which can significantly reduce the dimension of the gait representations without hindering the accuracy. Extensive experiments on CASIA-B and OUMVLP show that GLN can achieve state-of-the-art performance using the 256-dimensional representations. Under the most challenging condition of walking in different clothes on CASIA-B, our method improves the rank-1 accuracy by 6.45%.

Keywords: Gait recognition · Lateral connections · Discriminative representations · Compact representations

1 Introduction

Gait recognition aims at identifying different people using videos recording the walking patterns [38]. Compared to other biometrics such as face [33], fingerprint [27] and iris [39], human gait can be obtained at a long distance without the cooperation of subjects, which contributes to its broad applications in crime prevention, forensic identification and social security [4,18]. However, gait recognition suffers from a lot of variations such as clothing, carrying conditions and

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12354, pp. 382–398, 2020.
https://doi.org/10.1007/978-3-030-58545-7_22

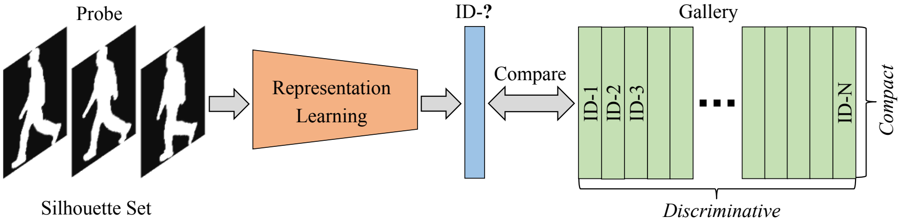

Fig. 1. Illustration of the silhouette-based gait recognition. The learned representations should be *discriminative* to identify different people, and should also be *compact* for the convenience of storage

camera viewpoints [36,42]. A key challenge is to learn representations from the silhouettes of gait sequences that are invariant to the factors mentioned above.

To address the issue, various methods have been proposed which can be roughly divided into three categories. The first category [8,10,35,41] aggregates the silhouettes of a complete gait sequence into an image (or template) for recognition, *e.g.* Gait Energy Image [8]. Despite the simplicity, the temporal and fine-grained spatial information is inevitably lost in the pre-processing. The second [20,40] regards the silhouettes of a gait sequence as a video. For example, in [40], a 3D-CNN [14] is adopted to extract the spatial and temporal information while the model is relatively hard to train. The third [6] is recently proposed and treats the silhouettes of a gait sequence as an unordered set, which is robust to the number of the silhouettes and achieves significant improvements. However, the dimension of the representations learned by [6] reaches up to 15872 that is much higher than those for face recognition (*e.g.* 180 [33]) or person re-identification (*e.g.* 2048 [25]).

In this work, we deal with gait recognition with the aims of learning both *discriminative* and *compact* representations from the silhouettes for gait recognition. We propose a novel network named Gait Lateral Network (denoted as GLN) where the silhouettes of each gait sequence are regarded as an unordered set. As illustrated in Fig. 1, besides being discriminative to identify different people, the learned representation for each silhouette set should also be as compact as possible, which would otherwise incur a heavy storage burden to keep millions of subjects registered in the gallery. It is noteworthy that, the dimension of the representations learned by GLN is fixed to 256 which is reduced by nearly two orders of magnitude compared to [6] and the performance for all walking conditions are improved simultaneously.

Specifically, we propose to leverage the inherent feature pyramid in deep CNNs to learn *discriminative* gait representations. The features extracted by different layers capture various visual details of the input [43]. We notice that the silhouettes for different subjects only have subtle differences in many cases, which makes it vital to explore the shallow features encoding the local spatial structural information for gait recognition. Particularly, we modify the network of [6] as the backbone and explicitly divide the layers into three stages. The silhouette-level and set-level features extracted by different stages are merged with the lateral connections in a top-down manner, which tries to aggregate the visual details extracted by different layers for accurate recognition. The features after refinement of different stages are then split horizontally to learn part representations and the triplet loss is added at all stages as the intermediate supervision [19]. Besides, we propose a novel *Compact Block* to learn *compact* gait representations. The preliminary study reveals that there exists a lot of redundancy in the high-dimensional representations learned by HPM [6,7] which is widely adopted for part representation learning. The proposed *Compact Block* can distill the knowledge of high-dimensional gait representations into compact ones without hindering the accuracy. Its architecture is simple but non-trivial which can be seamlessly integrated with the backbone and trained in an end-to-end manner. We regard the high-dimensional representations as an ensemble of low-dimensional ones and utilize *Dropout* to select a small subset, which is then mapped into a compact space by *Fully Connected Layer*.

In summary, our contributions of this work lie in three folds: (1) We propose to leverage the inherent feature pyramid in deep CNNs to enhance the gait representations for accurate recognition. The silhouette-level and set-level features extracted by different stages are merged with the lateral connections in a top-down manner. (2) We propose a *Compact Block* which can significantly reduce the dimension of the gait representations without hindering the accuracy. (3) The resulting GLN can learn both *discriminative* and *compact* representations from the silhouettes for gait recognition. The experiments on CASIA-B [42] and OUMVLP [36] show that GLN can achieve state-of-the-art performance for all walking conditions using the 256-dimensional representations. In particular, under the most challenging condition of walking in different clothes on CASIA-B, the rank-1 accuracy achieved by GLN exceeds GaitSet [6] with the 15872-dimensional representations by 6.45%.

2 Related Work

Motion-Based Gait Recognition. These methods including [1,3,16] attempt to model the human body structures and then extract motion features for gait recognition, which have the advantage of being robust to clothing and carrying conditions. Nevertheless, they usually fail on low-resolution videos where it is difficult to estimate the body parameters accurately.

Appearance-Based Gait Recognition. These methods including [8,17,26,37] directly learn features from the gait sequences without explicitly modeling the body structures, which suit for the low-resolution conditions and thus attract increasing attention [44,46]. The silhouettes are usually taken as the input and a key challenge is to learn representations from the silhouettes that are robust to the factors such as clothing, carrying conditions and camera viewpoints [36,42]. The silhouette-based gait recognition can be roughly divided into three categories where the silhouettes of a complete gait sequence are respectively regarded as an image [8,10,35,41], a video [20,40] or an unordered image set [6].

Deep learning that innovates the field of computer vision is also widely used for gait recognition. Specifically, a comprehensive study on deep convolutional neural networks for gait recognition is conducted in [41]. An auto-encoder framework is proposed by [46] to explicitly disentangle the appearance and pose features in the representation learning. JUCNet [44] integrates the cross-gait and unique-gait supervision with a tailored quintuplet loss. DiGGAN [12] takes advantage of a Conditional GAN [28] to learn the view-invariant gait features. GaitSet [6] treats the silhouettes of each gait sequence as an unordered set and splits the features horizontally to learn part representations for gait recognition, which achieves significant improvements and holds the best performance across different datasets. However, the dimension of the final representations learned by [6] is too high, *i.e.* 15872.

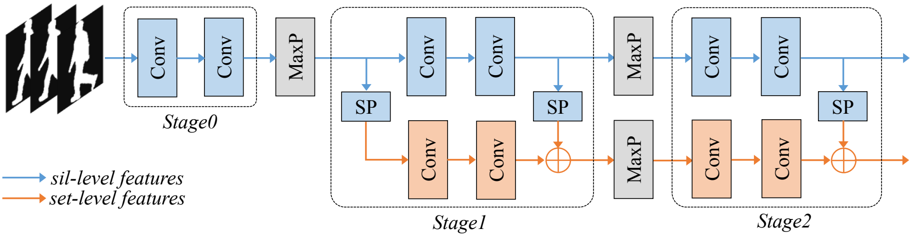

Fig. 2. Illustration of the division for the backbone, *Sil-level* for *Silhouette-level*, *MaxP* for *Max Pooling*, *SP* for *Set Pooling*. The *silhouette-level* features are extracted from each silhouette separately while the *set-level* features are extracted from all silhouettes. *Set Pooling* is a function to aggregate the features in a silhouette set

Inherent Feature Pyramid. The inherent feature pyramid in deep convolutional neural networks has been exploited in many visual tasks. For example, FCN [24] utilizes the features of different layers to progressively refine the predictions for semantic segmentation. Hypercolumns [9] proposes an efficient computation strategy to aggregate the features of different layers for object segmentation and localization. SSD [23] detects the objects using the features of different layers separately without fusing features or scores.

The top-down manner to merge the features of different stages in GLN is inspired by FPN [21] for object detection. However, our approach differs from FPN in three aspects. First, there are two branches in the last two stages of GLN as shown in Fig. 2 and the lateral connections in GLN are utilized to merge the silhouette-level and set-level features simultaneously. Second, the training labels for different stages in FPN are assigned according to the receptive fields, while the supervision signals for different stages in GLN are the same. Third, FPN shares the parameters in the heads following different stages, while the subsequent layers for different stages in GLN have independent parameters.

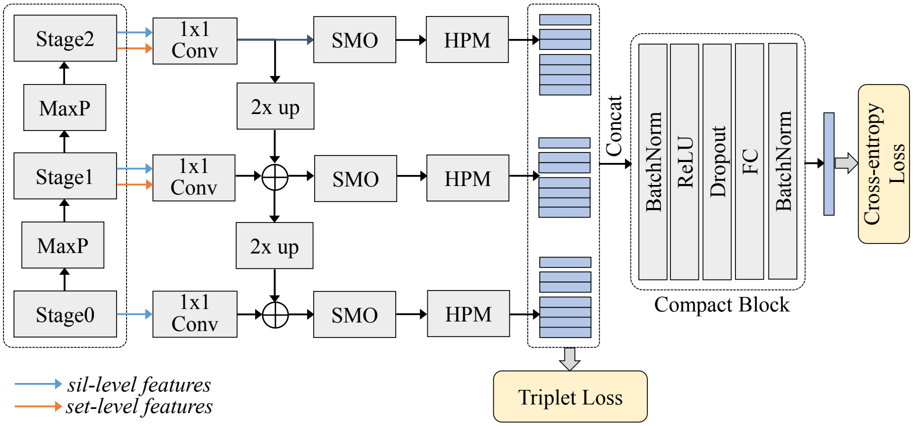

Fig. 3. Illustration of Gait Lateral Network, *Sil-level* for *Silhouette-level*, *MaxP* for *Max Pooling*, *SMO* for *Smooth Layer*, *HPM* for *Horizontal Pyramid Mapping*. For simplicity, *Set Pooling* for the silhouette-level features before each 1×1 convolutional layer is omitted and we use the scales $S = \{1, 2, 4\}$ to split the features horizontally in HPM. The output of *Compact Block* is taken as the final representation

3 Our Approach

In this work, we propose a novel network named Gait Lateral Network (GLN) which can learn both *discriminative* and *compact* representations from the silhouettes for gait recognition. The silhouettes of a complete gait sequence are regarded as an unordered set. The network structure is illustrated in Fig. 3. The silhouette-level and set-level features extracted by different stages in the backbone are merged with the lateral connections in a top-down manner, which aims to enhance the gait representations for accurate recognition. And we propose a *Compact Block* which can significantly reduce the dimension of the gait representations without hindering the accuracy. In what follows, we will first elaborate the lateral connections in GLN. Then we will introduce the composition of *Compact Block*. Finally, we will describe the corresponding training strategy for GLN.

3.1 Lateral Connections

In GLN, we propose to leverage the inherent feature pyramid in deep convolutional neural networks to learn *discriminative* gait representations. The features extracted by different layers in the backbone are aggregated to enhance the gait representations.

Specifically, we modify the network of [6] as the backbone where the order of the second *Set Pooling* and *Max Pooling* is switched. As shown in Fig. 2, we explicitly divide the layers in the backbone into three stages. The first stage is comprised of two convolutional layers which transform the silhouettes into the internal features. The second and third stages consist of two branches that learn the silhouette-level and set-level features respectively. *Set Pooling* is a function to aggregate the features in a silhouette set, which should be permutation invariant to the order of the silhouettes and is implemented by *Max Pooling* for simplicity. Note that, different from *Max Pooling* between different stages operating along the spatial dimensions (height and width), *Max Pooling* for *Set Pooling* operates along the set dimension[1]. The backbone extracts the silhouette-level features as well as the set-level features in a *bottom-up* way. The features extracted by the three stages are respectively denoted as $\{C_0, C_1, C_2\}$, which have the strides of $\{1, 2, 4\}$ with respect to the input silhouettes.

The features of different stages in the backbone capture various visual details of the silhouettes [43], and we propose to merge the features extracted by different stages with the lateral connections in a *top-down* manner. The strategy is illustrated in Fig. 3. Specifically, first, at the last two stages, we adopt *Set Pooling* to deal with the silhouette-level features and concatenate the output with the set-level features along the channel dimension. And at the first stage, only the silhouette-level features are available which are also processed by *Set Pooling*. Then at each stage, a 1×1 convolutional layer is taken to rearrange the features and adjust the channel dimension. Next, starting from the features generated at the last stage, we upsample the spatial dimensions (height and width) by a factor of 2 and add to the features generated at the previous stage (which have the same channel dimensions after the 1×1 convolutional layers) by element-wise addition. This process is iterated until the features generated at all stages are merged. Finally, a smooth layer is appended after each stage to alleviate the aliasing effect caused by upsampling and semantic gaps between different stages. The output of the smooth layers are denoted as $\{F_0, F_1, F_2\}$ for the three stages, corresponding to $\{C_0, C_1, C_2\}$, which respectively have the same spatial dimensions.

It is worth noting that, the output of the 1×1 convolutional layers as well as the smooth layers have the same channel dimensions, which are fixed to 256 through our experiments. Every smooth layer is implemented by a 3×3 convolutional layer. Besides, there are no non-linear activation functions involved in the lateral connections and we use the nearest neighbor upsampling between different stages.

3.2 Compact Block

In this section, we will elaborate the composition of *Compact Block* which is proposed to learn *compact* gait representations. Before that, we first review *Horizontal Pyramid Mapping* (HPM) [6] as the background which results in the high representation dimension.

HPM is equivalent to *Horizontal Pyramid Pooling* (HPP) [7] for person Re-ID, which is adopted by GLN to learn part representations for gait recognition.

[1] The silhouette-level features have the shape of [*batch, set, channel, height, width*] where the *set* dimension denotes the number of the silhouettes in an unordered set. And the set-level features have the shape of [*batch, channel, height, width*].

Despite the effectiveness, the representations obtained by HPM hold a very high dimension, *e.g.* 15872 [6]. Specifically, HPM first splits the features horizontally using multiple scales S, *e.g.* $S = \{1, 2, 4, 8, 16\}$. For each scale $s \in S$, the features F are sliced into s bins horizontally and equally. Then, the global max and average pooling are taken to generate the features $G_{s,t}$ for each bin:

$$G_{s,t} = \mathrm{MaxPool}\left(F_{s,t}\right) + \mathrm{AvgPool}\left(F_{s,t}\right) \tag{1}$$

where $s \in S$ and $t \in \{1, \cdots, s\}$. Finally, a fully connected layer is applied to $G_{s,t}$ and the output is denoted as $\widehat{G}_{s,t}$. In the training phase, the loss is added to the features of each part. And in the test phase, the features of all parts are concatenated as the final representations. As a result, the dimension of the final representations is proportional to the sum of scales ($e.g.\ sum(S) = 31$ with $S = \{1, 2, 4, 8, 16\}$) and the feature dimension of each part (*e.g.* 256), which is infeasible for the real-world applications. By delving into the formulation of HPM, we observe that the representations across different scales encode some duplicate information. For example, the part representations with the indexes $(s, t) = (4, 1)$ and $(s, t) = \{(8, 1), (8, 2)\}$ correspond to the same regions in the input silhouettes. Thus we conjecture that there exists a lot of redundancy in the high-dimensional representations obtained by HPM.

To tackle the issue, we propose a *Compact Block* with the aims of distilling the knowledge of high-dimensional representations into compact ones without hindering the accuracy. As shown in Fig. 3, *Compact Block* has a plain structure which is composed of *Batch Normalization (BN-I)* [13], *ReLU* [15], *Dropout* [32], *Fully Connected Layer (FC)* and another *Batch Normalization (BN-II)*. The block is simple yet effective, and here we provide the design principles for each layer:

(1) *BN-I* is adopted to normalize the concatenated features obtained by HPM, which helps stabilize the training processing.
(2) *ReLU* is introduced as the activation function to increase the non-linearity for *Compact Block*.
(3) *Dropout* is the key of *Compact Block*. As mentioned above, the representations obtained by HPM can be regarded as an ensemble of low-dimensional ones. Here we take advantage of *Dropout* to select a small subset from each high-dimensional representation.
(4) *FC* is used to map the small subset from *Dropout* into a more discriminative space. The output of *FC* determines the dimension of the final representations which is set to 256 through our experiments.
(5) *BN-II* is introduced for the convenience of optimizing the cross-entropy loss inspired by [22,25], where each subject in the training set is treated as a separate class.

In summary, *Compact Block* significantly reduces the gait representations to a fixed dimension (*e.g.* 256 in our experiments), which is seamlessly integrated with the backbone and trained in an end-to-end manner. It is worth noting that, we adopt the implementation of *Dropout* that is available in PyTorch [29]. It only works in the training phase, and at inference time the final representations can be treated as an ensemble of multiple reductions.

3.3 Training Strategy

The training strategy for GLN consists of two steps: *Lateral Pretraining* and *Global Training*. As shown in Fig. 3, there are two types of losses involved in the training, *i.e.* triplet loss and cross-entropy loss. The triplet loss is deployed after HPM as the intermediate supervision [19], while the cross-entropy loss is added at the end of GLN to learn the global representations.

First, in order to obtain a reasonable initialization for the lateral connections, we propose *Lateral Pretraining* supervised by the triplet loss only. Specifically, the *batch all* version of triplet loss [11] is added to the features of each part obtained by HPM at all stages. Formally:

$$L_{tp} = \frac{1}{N_{tp_+}} \overbrace{\sum_{s\in S}\sum_{t=1}^{s}}^{bins} \overbrace{\sum_{i=1}^{P}\sum_{j=1}^{K}}^{anchors} \overbrace{\sum_{\substack{a=1\\a\neq j}}^{K}}^{pos.} \overbrace{\sum_{\substack{b=1\\b\neq i}}^{P}\sum_{c=1}^{K}}^{negative} \left[m + d_{s,t,i,j,i,a}^{s,t,i,j,b,c}\right]_+ \tag{2}$$

$$d_{s,t,i,j,i,a}^{s,t,i,j,b,c} = dist(f(sil_{i,j}^{s,t}), f(sil_{i,a}^{s,t})) - dist(f(sil_{i,j}^{s,t}), f(sil_{b,c}^{s,t}))$$

where N_{tp_+} is the number of triplets resulting in the non-zero loss terms over a mini-batch, S is the multiple scales for HPM, (P, K) are the number of subjects and the number of sequences for each subject in a mini-batch, m is the margin threshold, f denotes the feature extraction, sil denotes the silhouette set, $dist$ measures the similarity between two features, *e.g.* euclidean distance. Note that, in *Lateral Pretraining*, we do not decrease the learning rate to prevent overfitting [47].

Then, *Global Training* is conducted to train the whole network with the sum of triplet loss and cross-entropy loss. For cross-entropy loss, each subject in the training set is treated as a separate class and the label smooth technique [34] is adopted. Formally:

$$L_{ce} = -\frac{1}{P \times K} \sum_{i=1}^{P}\sum_{j=1}^{K}\sum_{n=1}^{N} q_n^{ij} \log p_n^{ij} \tag{3}$$

where N is the number of all subjects in the training set, p is the probabilities belonging to each subject, q encodes the identity information which is computed as follows (taking the y-th subject as an example):

$$q_n^{ij} = \begin{cases} 1 - \frac{N-1}{N}\epsilon & \text{if } n = y \\ \frac{\epsilon}{N} & \text{otherwise} \end{cases} \tag{4}$$

where ϵ is a small constant to encourage the model to be less confident on the training set. In our experiments, ϵ is set to 0.1. The total loss for *Global Training* is computed as:

$$L = L_{tp} + L_{ce} \tag{5}$$

It is worth noting that, in the training phase, another *Fully Connected Layer* is introduced after *Compact Block* to compute the probabilities for each subject, which, however, is deprecated at inference time. The output of *Compact Block* is taken as the final representation for each silhouette set to match the probe and gallery.

Table 1. The dataset statistics. *NM* for *normal walking*, *BG* for *walking with bags*, *CL* for *walking in different clothes*

Dataset	Subjects		Walking conditions			Views
	Train	Test	NM	BG	CL	
CASIA-B	74	50	6	2	2	11
OUMVLP	5153	5154	2	–	–	14

4 Experiment

4.1 Settings

Datasets. The experiments are conducted on two popular gait datasets: CASIA-B [42] and OUMVLP [36]. The dataset statistics are shown in Table 1.

CASIA-B. It is a typical gait dataset that consists of 124 subjects. The walking conditions contain normal walking (NM, 6 variants per subject), walking with bags (BG, 2 variants per subject) and walking in different clothes (CL, 2 variants per subject). The 11 views for each walking condition are uniformly distributed in $[0°, 180°]$ at an interval of 18°. In total, there are $(6 + 2 + 2) \times 11 = 110$ sequences for each subject. There is no partition for training and test provided in this dataset. In our experiments, we take the first 74 subjects as the training set and the rest 50 as the test set. For evaluation, we regard the first 4 variants of normal walking (NM) for each subject as the gallery with the rest as the probe. The probe can be further divided into three subsets according to the walking conditions, *i.e.* NM, BG, CL.

OUMVLP. It is the largest gait dataset in public which consists of 10307 subjects. However, only the sequences of normal walking (NM, 2 variants per subject) are available for each subject. The 14 views are uniformly distributed between $[0°, 90°]$ and $[180°, 270°]$ at an interval of 15°. In total, there are $2 \times 14 = 28$ sequences for each subject. According to the provided partition, we take 5153 subjects as the training set with the rest 5154 as the test set. For evaluation, the first variant of normal walking (NM) for each subject is treated as the gallery with the rest as the probe.

Implementation Details. All models are implemented with PyTorch [29]. The silhouettes in both datasets are pre-processed using the methods in [35]. The number of subjects and the sequences for each subject in a mini-batch as well as the input size of each silhouette, are set to $(8, 16, 128 \times 88)$ for CASIA-B and $(32, 16, 64 \times 44)$ for OUMVLP. In the training phase, we randomly select 30 silhouettes for each gait sequence. For evaluation, all silhouettes of a gait sequence are taken to obtain the final representation.

The convolutional channels in the three stages shown in Fig. 2 are set to $(32, 64, 128)$ for CASIA-B and $(64, 128, 256)$ for OUMVLP. In the lateral connections shown in Fig. 3, the output dimensions of the 1×1 convolutional layers and the smooth layers are all set to 256. We use the multiple scales $S = \{1, 2, 4, 8, 16\}$ to split the features horizontally at all stages and the feature dimension of each part obtained by HPM is set to 256. For *Compact Block*, an aggressive dropping ratio 0.9 is adopted for *Dropout* and the output dimension is set to 256.

We adopt SGD with momentum [30] as the optimizer. The initial learning rate is set to 0.1 which is not decreased in *Lateral Pretraining*. While in *Global Training*, the learning rate is scaled to its 1/10 three times until convergence. The step size is set to 10000 iterations for CAISA-B and 50000 iterations for OUMVLP. We use the momentum 0.9 and the weight decay 5e$-$4 for the optimization. The margin threshold m for L_{tp} in Eq. 2 is set to 0.2. Besides, the warmup strategy [25] is adopted at the start of training.

Baselines. GaitSet [6] holds the best performance for the silhouette-based gait recognition and is taken as an important baseline in our experiments. It proposes to treat the silhouettes of a gait sequence as an unordered set and splits the features horizontally to learn part representations for gait recognition, which outperforms the previous works [31,41] by a large margin. It is worth mentioning that, we reproduce the results for GaitSet by ourselves which are a little higher than those reported in [6]. Besides, for a comprehensive study, we also re-implement GEINet [31] which is a representative method taking Gait Energy Image [8] as the input. It customizes a network for gait recognition and treats each subject as a separate class in the training. The features before the softmax layer are taken to match the probe and gallery for evaluation. Finally, to enable a more fair comparison on CASIA-B, we implement an improved version of GaitSet (denoted as GaitSet-L) where the input size of each silhouette is enlarged from 64×44 to 128×88.

4.2 Performance Comparison

CASIA-B. Table 2 shows the performance comparison on CASIA-B. The dimensions of the final representations learned by different methods are also compared. The probe sequences are divided into three subsets, *i.e.* NM, BG, CL, which are respectively evaluated. The accuracy for each probe view is averaged on all gallery views excluding the identical-view cases.

From the results in Table 2, we observe that GaitSet and GaitSet-L outperform GEINet by a large margin, which, however, generate the gait representations with a very high dimension (*i.e.* 15872). The comparisons between GaitSet-L and GaitSet indicate that enlarging the input size is beneficial to gait recognition especially for walking with bags (BG) and walking in different clothes (CL), although the consumption of GPU memory is simultaneously increased. Particularly, compared to GaitSet and GaitSet-L, GLN reduces the representation dimension by nearly two orders of magnitude ($15872 \rightarrow 256$) and achieves state-of-the-art performance under all walking conditions (NM-96.88%, BG-94.04%, CL-77.50%). Under the most challenging condition of walking in different clothes (CL), GLN exceeds GaitSet by 6.45% with the representation dimension significantly reduced to 256. The improvements under the other two walking conditions compared to GaitSet are also impressive, *i.e.* +1.67% for normal walking (NM) and +5.96% for walking with bags (BG). Besides, we notice that, though the average performance is inferior to GLN, GaitSet-L achieves the best performance in some probe views (*e.g.* 126°) for walking in different clothes (CL). This phenomenon needs further exploration.

Table 2. The rank-1 accuracy (%) on CAISA-B across different views excluding the identical-view cases, *DIM* for *Dimension*. For evaluation, the first 4 variants of normal walking (NM) for each subject are taken as the gallery. The probe sequences are divided into three subsets according to the walking conditions, *i.e.* NM, BG and CL

Probe	Method	DIM	Probe view											Average
			0°	18°	36°	54°	72°	90°	108°	126°	134°	162°	180°	
NM	GEINet [31]	1024	40.20	38.90	42.90	45.60	51.20	42.00	53.50	57.60	57.80	51.80	47.70	48.11
	GaitSet [6]	15872	**93.40**	98.10	98.50	97.80	92.60	90.90	94.20	97.30	98.40	97.00	89.10	95.21
	GaitSet-L	15872	91.40	98.50	98.80	97.20	94.80	92.90	95.40	97.90	98.80	96.50	89.10	95.57
	GLN **(ours)**	**256**	93.20	**99.30**	**99.50**	**98.70**	**96.10**	**95.60**	**97.20**	**98.10**	**99.30**	**98.60**	**90.10**	**96.88**
BG	GEINet [31]	1024	34.20	29.29	31.21	35.20	35.20	27.60	35.90	43.50	45.00	38.99	36.80	35.72
	GaitSet [6]	15872	85.90	92.12	93.94	90.41	86.40	78.70	85.00	91.60	93.10	91.01	80.70	88.08
	GaitSet-L	15872	89.00	95.25	95.56	93.98	89.70	86.70	89.70	94.30	95.40	92.73	84.40	91.52
	GLN **(ours)**	**256**	**91.10**	**97.68**	**97.78**	**95.20**	**92.50**	**91.20**	**92.40**	**96.00**	**97.50**	**94.95**	**88.10**	**94.04**
CL	GEINet [31]	1024	19.90	20.30	22.50	23.50	26.70	21.30	27.40	28.20	24.20	22.50	21.60	23.46
	GaitSet [6]	15872	63.70	75.60	80.70	77.50	69.10	67.80	69.70	74.60	76.10	71.10	55.70	71.05
	GaitSet-L	15872	66.30	79.40	84.50	80.70	74.60	73.20	74.10	**80.30**	**79.70**	72.30	62.90	75.27
	GLN **(ours)**	**256**	**70.60**	**82.40**	**85.20**	**82.70**	**79.20**	**76.40**	**76.20**	78.90	77.90	**78.70**	**64.30**	**77.50**

OUMVLP. Table 3 displays the performance comparison on OUMVLP where GLN also achieves state-of-the-art performance with the 256-dimensional representations. The input size of each silhouette on this dataset is set to 64×44 due to the limits of GPU memory and thus the performance of GaitSet-L is not available. It is worth noting that, the dimension of the representations learned by GEINet is doubled to 2048 for this large-scale dataset, and the reproduced results for GEINet is much higher than those reported in [6]. In spite of the high representation dimension, GaitS et al. so holds the best performance on this large-scale dataset before this work and outperforms GEINet by a large margin. According to the results shown in Table 3, we observe that GLN improves the rank-1 accuracy by 2.13% compared to GaitSet and the representation dimension is significantly reduced to 256.

Besides, we notice that the gait data for some subjects in OUMVLP is incomplete. As a result, for some probe sequences, there are not the corresponding sequences in the gallery. Thus we further conduct the evaluation ignoring the probe sequences which have no corresponding ones in the gallery. As shown in the last three rows of Table 3, GLN finally achieves the rank-1 accuracy of 95.57% with the 256-dimensional representations, which exceeds GaitSet with the 15872-dimensional representations by 2.32% on this large-scale dataset.

4.3 Ablation Study

In this section we provide the ablation study to further analyze GLN. The experiments are conducted on CASIA-B using the settings described in Sect. 4.1.

Table 3. The rank-1 accuracy (%) on OUMVLP across different views excluding the identical-view cases, *DIM* for *Dimension*. For evaluation, the first variant of normal walking (NM) for each subject is taken as the gallery with the rest as the probe. The last three rows show the results ignoring the probe sequences which have no corresponding ones in the gallery

Method	DIM	Probe view														Average
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GEINet [31]	2048	23.20	38.09	47.95	51.81	47.53	48.09	43.75	27.25	37.89	46.78	49.85	45.94	45.65	40.96	42.48
GaitSet [6]	15872	79.33	87.59	89.96	90.09	87.96	88.74	87.69	81.82	86.46	88.95	89.17	87.16	87.60	86.15	87.05
GLN(**ours**)	**256**	**83.81**	**90.00**	**91.02**	**91.21**	**90.25**	**89.99**	**89.43**	**85.28**	**89.09**	**90.47**	**90.59**	**89.60**	**89.31**	**88.47**	**89.18**
GEINet [31]	2048	24.91	40.65	51.55	55.13	49.81	51.05	46.37	29.17	40.67	50.53	53.27	48.39	48.64	43.49	45.26
GaitSet [6]	15872	84.50	93.27	96.72	96.58	93.48	95.28	94.15	87.04	92.50	96.00	95.96	92.99	94.34	92.69	93.25
GLN(**ours**)	**256**	**89.28**	**95.84**	**97.87**	**97.82**	**96.01**	**96.68**	**96.07**	**90.71**	**95.34**	**97.66**	**97.54**	**95.69**	**96.24**	**95.27**	**95.57**

Lateral Connections. In GLN, the silhouette-level and set-level features extracted by different stages are merged with the lateral connections in a top-down manner. Here we separately evaluate the effect of lateral connections. Specifically, the network shown in Fig. 3 is trained without *Compact Block* until convergence. HPM is applied to the features generated by the lateral connections at the three stages. And the features of all parts are concatenated as the final representations where the dimension reaches up to 23808. According to the results shown in Table 4, the lateral connections can improve the performance under all walking conditions especially for walking in different clothes (CL).

Label Smooth. As stated in Sect. 3.3, the label smooth is adopted to prevent overfitting on the subjects in the training set. Here we conduct the experiment ignoring the label smooth and the standard cross-entropy loss [15] is computed for GLN. As shown in the last two rows of Table 4, the label smooth is beneficial to gait recognition under all walking conditions. Besides, the experimental results in Table 4 indicate that *Compact Block* can simultaneously reduce the dimension of the representations and improve the performance especially for normal walking (NM) and walking with bags (BG). The performance for walking in different clothes (CL) is comparable before and after reduction.

Training Strategy. The training strategy for GLN consists of two steps: *Lateral Pretraining* and *Global Training*. We have also tried to train the whole network globally from scratch, however, the performance is inferior under all walking conditions on CASIA-B (NM-96.48%, BG-93.07%, CL-77.07%). The comparison indicates that it is necessary to pretrain the lateral connections.

Table 4. The ablation study for lateral connections and label smooth, *DIM* for *Dimension*, *CBlock* for *Compact Block*. The results are reported on CASIA-B

Method	DIM	Label Smooth	NM	BG	CL
GaitSet [6]	15872	–	95.21	88.08	71.05
GaitSet-L	15872	–	95.57	91.52	75.27
GLN (without *CBlock*)	23808	–	95.58	91.98	77.22
GLN (with *CBlock*)	256	×	96.48	94.03	77.03
GLN (with *CBlock*)	256	√	**96.88**	**94.04**	**77.50**

Output Dimensions. The dimension of the final representations learned by GLN is empirically set to 256 through our experiments. Here we provide the experimental results with different output dimensions including 128 and 512. As shown in Table 5, the performance for normal walking (NM) and walking with bags (BG) are comparable with the three dimensions. The dimension 256 achieves the best performance for walking in different clothes (CL) which is the most challenging and occurs frequently in the real-world applications.

Variants of Compact Block. As shown in Table 5, we conduct the experiments in comparison to some variants of *Compact Block*. We use the same settings as described in Sect. 4.1 except that the structure of *Compact Block* is replaced by the variants shown in Table 5. And we have also tried some classical methods for dimension reduction such as Principal Components Analysis (PCA, NM-95.47%, BG-91.90%, CL-76.99%) and Linear Discriminant Analysis (LDA, NM-87.97%, BG-81.85%, CL-63.19%). The performance comparisons indicate that *Compact Block* can be treated as a reasonable choice to reduce the dimension of the gait representations.

Time Statistics. Here we provide the running time comparison on CASIA-B between GaitSet-L and GLN in the training (GaitSet-L: 0.96s per iteration v.s. GLN: 1.01s per iteration) and test (GaitSet-L: 0.021s per sequence v.s. GLN: 0.022s per sequence). Though the running time of GLN is marginally increased compared to GaitSet-L, our method can reduce the representation dimension by nearly two orders of magnitude (15872 → 256) and the performance for all walking conditions are improved simultaneously.

Table 5. The ablation study for output dimensions and variants of *Compact Block*, *DIM* for *Dimension*. The results are reported on CASIA-B

DIM	Variants of *compact block*	NM	BG	CL
512	*BN+ReLU+Dropout+FC+BN*	96.98	94.09	77.10
256	*BN+ReLU+Dropout+FC+BN*	96.88	94.04	**77.50**
128	*BN+ReLU+Dropout+FC+BN*	**97.07**	**94.10**	76.84
256	*BN+FC+BN*	94.58	90.81	70.05
256	*BN+ReLU+FC+BN*	95.29	90.74	71.48
256	*BN+Dropout+FC+BN*	96.37	93.83	75.23
256	*BN+ReLU+Dropout+FC+BN*	**96.88**	**94.04**	**77.50**

Comparison to More Baselines. As stated in Sect. 4.1, GaitSet [6] holds state-of-the-art performance for the silhouettes-based gait recognition before this work and GEINet [31] is a representative method taking Gait Energy Image [8] as input, which are more related to our work and compared thoroughly in our experiments. Here we provide more silhouette-based methods for comparison such as CNN-LB [41] (NM-89.9%, BG-72.4%, CL-54.0%) and J-CNN [45] (NM-91.2%, BG-75.0%, 54.0%). Besides, we notice that there are some methods taking other types of input for gait recognition such as GaitNet [46] (RGB frames, NM-92.3%, BG-88.9%, CL-62.3%), GaitMotion [2] (optical flow, NM-97.5%, BG-83.6%, CL-48.8%), SM-Prod [5] (gray images and optical flow, NM-99.8%, BG-96.1%, CL-67.0%). Though some methods [2,5] report a little higher performance for NM, the optical flow needs a lot of computation cost and the performance for the challenging CL is much inferior to our method (CL-77.50%). The results here are all reported on CASIA-B.

5 Conclusion

In this work, we propose a novel network named Gait Lateral Network (GLN) which can learn both *discriminative* and *compact* representations from the silhouettes for gait recognition. Specifically, the inherent feature pyramid in deep convolutional networks is leveraged to learn *discriminative* gait representations. The silhouette-level and set-level features extracted by different stages in the backbone are merged with the lateral connections in a top-down manner, which enhances the gait representations by aggregating more visual details. And we propose a *Compact Block* to learn *compact* gait representations, which can significantly reduce the dimension of the gait representations without hindering the accuracy. Extensive experiments on CASIA-B and OUMVLP demonstrate that GLN achieves state-of-the-art performance under all walking conditions using the 256-dimensional representations.

Acknowledgments. We are grateful to Prof. Dongbin Zhao for his support to this work.

References

1. Ariyanto, G., Nixon, M.S.: Model-based 3D gait biometrics. In: International Joint Conference on Biometrics, pp. 1–7 (2011)
2. Bashir, K., Xiang, T., Gong, S., Mary, Q.: Gait representation using flow fields. In: BMVC, pp. 1–11 (2009)
3. Bodor, R., Drenner, A., Fehr, D., Masoud, O., Papanikolopoulos, N.: View-independent human motion classification using image-based reconstruction. Image Vis. Comput. **27**(8), 1194–1206 (2009)
4. Bouchrika, I., Goffredo, M., Carter, J., Nixon, M.: On using gait in forensic biometrics. J. Forensic Sci. **56**(4), 882–889 (2011)
5. Castro, F.M., Marín-Jiménez, M.J., Guil, N., de la Blanca, N.P.: Multimodal feature fusion for CNN-based gait recognition: an empirical comparison. Neural Comput. Appl. **32**, 14173–14193 (2020). https://doi.org/10.1007/s00521-020-04811-z
6. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: regarding gait as a set for cross-view gait recognition. In: AAAI, vol. 33, pp. 8126–8133 (2019)
7. Fu, Y., et al.: Horizontal pyramid matching for person re-identification. In: AAAI, vol. 33, pp. 8295–8302 (2019)
8. Han, J., Bhanu, B.: Individual recognition using gait energy image. TPAMI **28**(2), 316–322 (2005)
9. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR, pp. 447–456 (2015)
10. He, Y., Zhang, J., Shan, H., Wang, L.: Multi-task GANs for view-specific feature learning in gait recognition. IEEE Trans. Inf. Forensics Secur. **14**(1), 102–113 (2018)
11. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
12. Hu, B., Gao, Y., Guan, Y., Long, Y., Lane, N., Ploetz, T.: Robust cross-view gait identification with evidence: a discriminant gait GAN (DIGGAN) approach on 10000 people. arXiv preprint arXiv:1811.10493 (2018)
13. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. ICML **37**, 448–456 (2015)
14. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. TPAMI **35**(1), 221–231 (2012)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NeurIPS, pp. 1097–1105 (2012)

16. Kusakunniran, W., Wu, Q., Li, H., Zhang, J.: Multiple views gait recognition using view transformation model based on optimized gait energy image. In: ICCV Workshops, pp. 1058–1064 (2009)
17. Kusakunniran, W., Wu, Q., Zhang, J., Ma, Y., Li, H.: A new view-invariant feature for cross-view gait recognition. IEEE Trans. Inf. Forensics Secur. **8**(10), 1642–1653 (2013)
18. Larsen, P.K., Simonsen, E.B., Lynnerup, N.: Gait analysis in forensic medicine. J. Forensic Sci. **53**(5), 1149–1153 (2008)
19. Lee, C.Y., Xie, S., Gallagher, P.W., Zhang, Z., Tu, Z.: Deeply-supervised nets. ArXiv abs/1409.5185 (2014)
20. Liao, R., Cao, C., Garcia, E.B., Yu, S., Huang, Y.: Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In: Zhou, J., et al. (eds.) CCBR 2017. LNCS, vol. 10568, pp. 474–483. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69923-3_51
21. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR, pp. 2117–2125 (2017)
22. Liu, C.T., Wu, C.W., Wang, Y.C.F., Chien, S.Y.: Spatially and temporally efficient non-local attention network for video-based person re-identification. arXiv preprint arXiv:1908.01683 (2019)
23. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
25. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPR Workshops (2019)
26. Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., Yagi, Y.: Gait recognition using a view transformation model in the frequency domain. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 151–163. Springer, Heidelberg (2006). https://doi.org/10.1007/11744078_12
27. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer, London (2009). https://doi.org/10.1007/978-1-84882-254-2
28. Mirza, M., Osindero, S.: Conditional generative adversarial nets. ArXiv abs/1411.1784 (2014)
29. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: NeurIPS, pp. 8024–8035 (2019)
30. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
31. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: GeiNet: view-invariant gait recognition using a convolutional neural network. In: International Conference on Biometrics, pp. 1–8 (2016)
32. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
33. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NeurIPS,pp. 1988–1996 (2014)
34. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR, pp. 2818–2826 (2016)
35. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: On input/output architectures for convolutional neural network-based cross-view gait recognition. IEEE Trans. Circuits Syst. Video Technol. (2017)

36. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Trans. Comput. Vis. Appl. **10**(1), 1–14 (2018). https://doi.org/10.1186/s41074-018-0039-6
37. Wang, C., Zhang, J., Wang, L., Pu, J., Yuan, X.: Human identification using temporal information preserving gait template. TPAMI **34**(11), 2164–2176 (2011)
38. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. TPAMI **25**(12), 1505–1518 (2003)
39. Wildes, R.P.: Iris recognition: an emerging biometric technology. Proc. IEEE **85**(9), 1348–1363 (1997)
40. Wolf, T., Babaee, M., Rigoll, G.: Multi-view gait recognition using 3D convolutional neural networks. In: ICIP, pp. 4165–4169 (2016)
41. Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T.: A comprehensive study on cross-view gait based human identification with deep CNNs. TPAMI **39**(2), 209–226 (2016)
42. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: International Conference on Pattern Recognition, vol. 4, pp. 441–444 (2006)
43. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
44. Zhang, K., Luo, W., Ma, L., Liu, W., Li, H.: Learning joint gait representation via quintuplet loss minimization. In: CVPR, pp. 4700–4709 (2019)
45. Zhang, Y., Huang, Y., Wang, L., Yu, S.: A comprehensive study on gait biometrics using a joint CNN-based method. Pattern Recogn. **93**, 228–236 (2019)
46. Zhang, Z., et al.: Gait recognition via disentangled representation learning. In: CVPR, pp. 4710–4719 (2019)
47. Zhu, W., Hu, J., Sun, G., Cao, X., Qiao, Y.: A key volume mining deep framework for action recognition. In: CVPR, pp. 1991–1999 (2016)