Gait Quality Aware Network: Toward the Interpretability of Silhouette-Based Gait Recognition

Saihui Hou, Xu Liu, Chunshui Cao, and Yongzhen Huang

***Abstract*— Gait recognition receives increasing attention since it can be conducted at a long distance in a nonintrusive way and applied to the condition of changing clothes. Most existing methods take the silhouettes of gait sequences as the input and learn a unified representation from multiple silhouettes to match probe and gallery. However, these models are all faced with the lack of interpretability, e.g., it is not clear which silhouette in a gait sequence and which part in the human body are relatively more important for recognition. In this work, we propose a gait quality aware network (GQAN) for gait recognition which explicitly assesses the quality of each silhouette and each part via two blocks: frame quality block (FQBlock) and part quality block (PQBlock). Specifically, FQBlock works in a squeeze-and-excitation style to recalibrate the features for each silhouette, and the scores of all the channels are added as frame quality indicator. PQBlock predicts a score for each part which is used to compute the weighted distance between the probe and gallery. Particularly, we propose a part quality loss (PQLoss) which enables GQAN to be trained in an end-to-end manner with only sequence-level identity annotations. This work is meaningful by moving toward the interpretability of silhouette-based gait recognition, and our method also achieves very competitive performance on CASIA-B and OUMVLP.**

***Index Terms*— Frame quality, gait quality aware network (GQAN), part quality, silhouette-based gait recognition.**

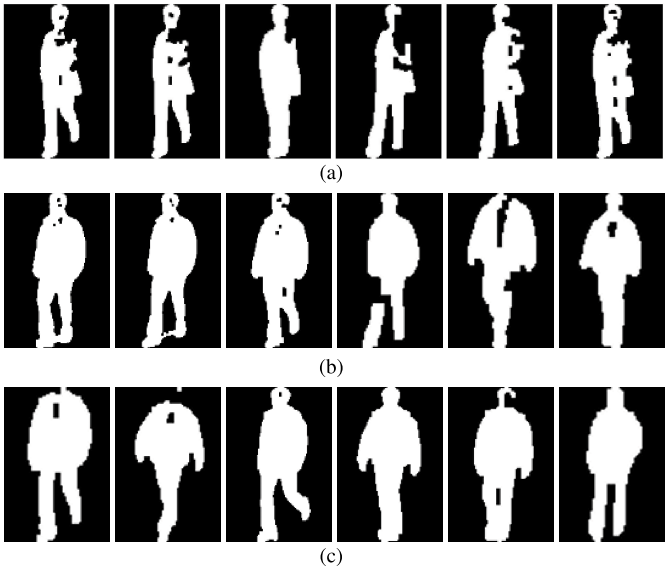


Fig. 1. Illustration of hard cases for silhouette-based gait recognition. (a) and (b) Same subject with a low similarity score. (b) and (c) Different subjects with a high similarity score. Due to the lack of interpretability, it is difficult to analyze the reasons that lead to these cases. (a) ID = 112. (b) ID = 112. (c) ID = 123.

I. Introduction

HUMAN gait, as a unique biometrics, receives increasing attention since it can be obtained at a long distance in a nonintrusive way [1]–[4]. Gait recognition aims to recognize the identity of people according to the walking patterns, and the existing methods for gait recognition can be roughly divided into two categories, i.e., *model-based* and *appearance-based*. The model-based methods [5]–[7] try to extract the human body structures from videos which have the advantage of being robust to clothing and carrying conditions. However, in low-resolution conditions, it is difficult to estimate the body parameters accurately which has a large adverse effect on the model-based methods. In contrast, the appearance-based methods [8]–[10] try to learn gait features from videos *without explicitly modeling the human body structures*. The silhouettes, which are simple yet effective, are usually taken as the input for the appearance-based methods.

In previous literature, the silhouette-based methods [11]–[13] hold state-of-the-art performance for gait recognition; however, all these methods are faced with a key challenge, i.e., the lack of interpretability. For example, it is not clear which silhouette in a gait sequence and which part in the human body are relatively more important for final recognition. In real-world applications, the cases shown in Fig. 1 occur frequently: 1) two sequences belonging to different subjects are assigned a high similarity score and 2) two sequences belonging to the same subject are assigned a low similarity score. Due to the lack of interpretability, it is hard to analyze the reasons that lead to these cases.

In this work, we focus on the interpretability of silhouette-based gait recognition. Interpretability is a broad

Manuscript received May 26, 2021; revised October 19, 2021 and February 11, 2022; accepted February 14, 2022. This work was supported in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Yongzhen Huang.)*

Saihui Hou and Yongzhen Huang are with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China (e-mail: housaihui@bnu.edu.cn; huangyongzhen@bnu.edu.cn).

Xu Liu and Chunshui Cao are with Watrix Technology Company Limited, Beijing 100088, China (e-mail: xu.liu@watrix.ai; chunshui.cao@watrix.ai).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3154723.

Digital Object Identifier 10.1109/TNNLS.2022.3154723

topic which can be addressed from many different aspects. To the best of our knowledge, the previous works that deal with the interpretability in the context of deep learning mostly try to analyze the role of each neuron [14], [15] and discover the discriminative area in the input [16]. Differently, in this work, we try to *find out the relative importance of each silhouette and each part for silhouette-based gait recognition*. To this end, we propose a gait quality aware network (denoted as GQAN) which explicitly assesses the quality of each silhouette and each part via two blocks, i.e., frame quality block (FQBlock) and part quality block (PQBlock).

Specifically, the silhouettes in a gait sequence are complementary to each other and contain the walking pattern of a subject. Due to occlusion, geometry distortion, segmentation errors, and so on, the quality of each silhouette cannot be guaranteed [17], [18], which is likely to hurt feature learning from silhouettes. In GQAN, the silhouettes of a gait sequence are regarded as an unordered set. FQBlock is proposed to incorporate the frame quality of each silhouette for set-based feature learning which can learn more discriminative features and *enhance the interpretability*. Specifically, FQBlock works in a squeeze-and-excitation style [19] to recalibrate the features for each silhouette, and the scores of all the channels are added as frame quality indicator for the corresponding silhouette.

Besides, learning part representations by horizontally slicing the features has been widely used for gait recognition [10]–[13]. However, all part representations are treated equally when computing the distance of two gait sequences, which is not optimal for gait recognition. For example, in the cases of changing coats or jackets, the features of head and legs are usually more important than those of upper body to match the probe and gallery. PQBlock addresses the issue by learning an adaptive weight for each part. It operates on set-level part representations and predicts a score for each part which is taken to compute the weighted distance between the probe and gallery. Particularly, we propose a loss function named part quality loss (PQLoss) to train PQBlock with only sequence-level identity annotations.

In summary, the main contributions of this work lie in three folds.

1) We propose a GQAN toward the interpretability of silhouette-based gait recognition by explicitly assessing the quality of each silhouette and each part. GQAN can *automatically* sort the silhouettes and parts according to the relative importance.
2) We propose a PQLoss which enables GQAN to be trained in an end-to-end manner with only sequence-level identity annotations.
3) GQAN achieves very competitive performance on CASIA-B and OUMVLP under all walking conditions.

II. Related Work

A. Gait Recognition

The methods for gait recognition can be roughly divided into two categories: *model-based* and *appearance-based*, which will be briefly reviewed in this section.

The *model-based* methods try to explicitly extract the human body structures from the videos for gait recognition. For example, PoseGait [6] takes the 3-D pose estimated from RGB frames as the input for gait recognition. OUMVLP-Pose [5] constructs a large-scale pose-based gait dataset and evaluates different pose estimation methods for gait recognition. End2EndGait [7] first extracts pose and shape features by fitting SMPL [20] and subsequently feeds into a recognition network. These methods are theoretically robust to clothing and carrying conditions; however, they are difficult to adapt to low-resolution conditions where it is hard to estimate the human body parameters accurately.

The *appearance-based* methods try to learn gait features from the videos *without explicitly modeling the human body structures*. In most cases, the silhouettes are taken as the input which can adapt to conditions of low resolution and changing clothes. The silhouette-based methods can be further divided into three subcategories: *template-based*, *video-based*, and *set-based*. The *template-based* methods [3], [8], [21]–[23] aggregate the silhouettes of a gait sequence into a template, e.g., gait energy image (GEI) [24], which is simple but ignores the temporal information. The *video-based* methods [9], [25], [26] treat the silhouettes of a gait sequence as a video to extract the spatial and temporal information, while the models (e.g., 3-D-convolutional neural network (CNN) [25] and multiple-temporal-scale 3-D (MT3D) [9]) are relatively hard to train. The *set-based* methods [10]–[13] regard the silhouettes of each gait sequence as an unordered set which is on the basis that the appearance of the silhouettes also encodes some temporal information [10]. Besides, there are also some works taking other types of input for appearance-based gait recognition, such as GaitNet [27] (RGB frames), GaitMotion [28] (optical flow), and SM-Prod [29] (gray images and optical flow).

Our work belongs to the *appearance-based* methods where the *silhouettes* of each gait sequence are taken as the input and regarded as an *unordered set*. We move toward the interpretability of silhouette-based gait recognition by explicitly assessing the relative importance of each silhouette and each part for recognition.

B. Unordered Set

The unordered set is first introduced to the visual community by PointNet [30] for 3-D classification and segmentation, which is then adopted for many other visual tasks [10], [31]–[33]. GaitSet [10] first proposes to treat the silhouettes of a gait sequence as an unordered set which is now widely used for silhouette-based gait recognition. In GQAN, the silhouettes of a gait sequence are also regarded as an unordered set for quality assessment.

More related to our work lie in [31] and [32], which learn adaptive weights according to the image quality for set-based person reidentification [31] and face recognition [32]. In GQAN, FQBlock is proposed to assess the frame quality of each silhouette for gait recognition which differs from [31] and [32] in three aspects. First, we adopt a different way to obtain set-level representations according to the frame quality, and the details will be described in Section III-A.

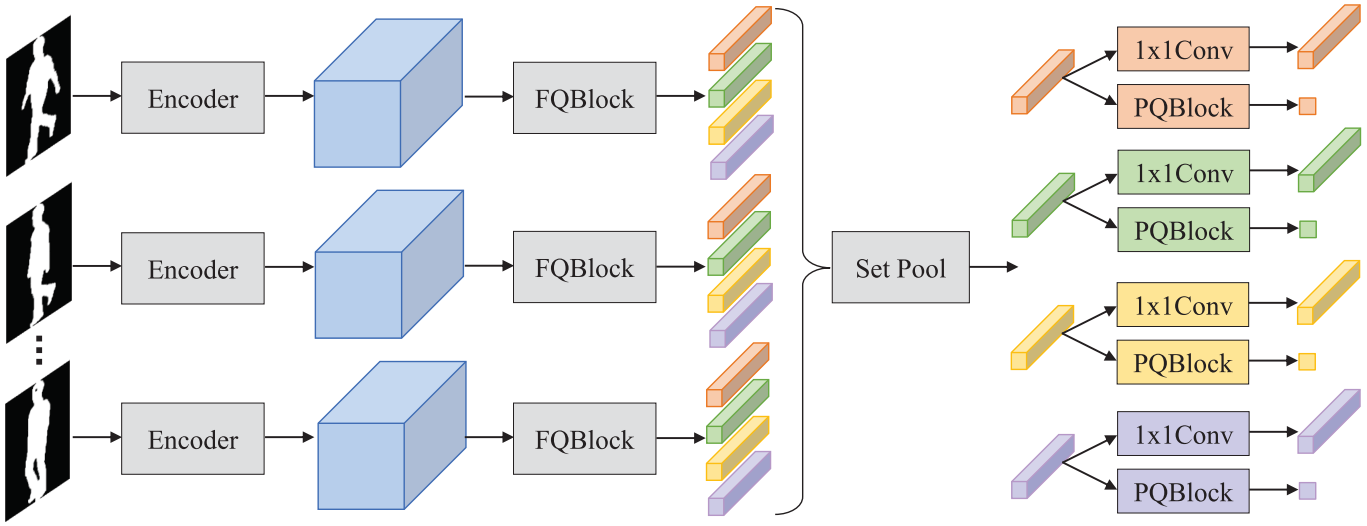


Fig. 2. Illustration of GQAN. The encoder mainly consists of convolutional layers to extract the features from each silhouette separately. FQBlock is taken to assess the quality of each silhouette where its weights are shared across the silhouettes but independent for different bins (annotated by different colors). PQBlock operates on set-level part representations and predicts a score to assess each part separately.

Second, FQBlock holds independent weights for features of different bins which are obtained by horizontally slicing the features of each silhouette. Third, we further propose a PQBlock to assess the quality of each part for gait recognition.

III. Our Approach

In this work, we propose a GQAN toward the interpretability of silhouette-based gait recognition. The network structure is illustrated in Fig. 2. It mainly consists of two blocks, i.e., FQBlock and PQBlock, to explicitly assess the quality of each silhouette and each part for recognition. Specifically, FQBlock works in a squeeze-and-excitation style to recalibrate the features for each silhouette, and the scores of all the channels are added as frame quality indicator. PQBlock predicts a score for each part to generate adaptive weights for distance computation between the probe and gallery. In what follows, we will first introduce the composition and working mechanism of FQBlock. Then we will describe the structure of PQBlock and how its output is taken to compute the distance between the probe and gallery. Finally, a PQLoss will be presented which enables GQAN to be trained with only sequence-level identity annotations.

A. Frame Quality

The silhouettes of a gait sequence contain the walking pattern of a subject and are complementary to each other. However, due to various factors such as occlusion, geometry distortion, and segmentation errors, the quality of each silhouette cannot be guaranteed, which has an adverse effect on gait feature learning. In GQAN, FQBlock is proposed to confront this problem where the quality of each silhouette can be automatically learned although such supervision is not explicitly provided in training.

Specifically, FQBlock works in a squeeze-and-excitation style [19] which mainly consists of two fully connected layers followed by rectified linear unit (ReLU) and sigmoid, respectively. Global average pooling (GAP) and global max pooling (GMP) are first adopted to squeeze the information in each channel. Then two fully connected layers are adopted for excitation to adaptively recalibrate the features for each silhouette. Particularly, considering large shape variance in a silhouette from head to feet, FQBlock horizontally and equally slices the features of each silhouette into multiple bins and holds independent weights for different bins. The structure of FQBlock is illustrated in Fig. 3 where silhouette-level features are horizontally split into four bins for the convenience of illustration.

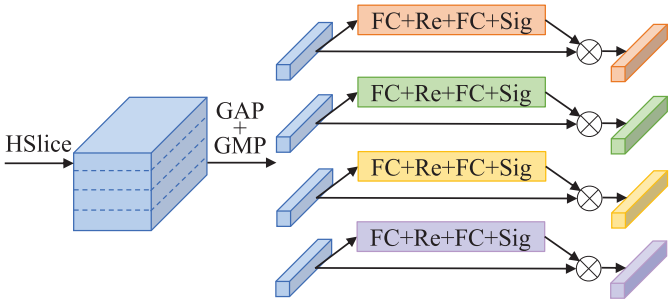


Fig. 3. Illustration of FQBlock. *HSlice* for *horizontal slice*, *GAP* for *global average pooling*, *GMP* for *global max pooling*, *FC* for *fully connected layer*, *Re* for *ReLU*, and *Sig* for *sigmoid*.

Formally, G denotes a gait sequence consisting of N silhouettes, $F = \{F_1, F_2, \ldots, F_N\}$ denotes the features extracted from each silhouette. FQBlock first horizontally and equally slices the features of each silhouette into S bins (e.g., $S = 16$) which are denoted as F_{ij} ($i \in [1, 2, \ldots, N]$, $j \in [1, 2, \ldots, S]$). Then, F_{ij}, i.e., the features of the ith silhouette and jth bin, is processed as follows:

$$X_{ij} = \mathrm{GAP}(F_{ij}) + \mathrm{GMP}(F_{ij}) \tag{1}$$

$$Y_{ij} = \sigma\left(W_j^2 \delta\left(W_j^1 X_{ij}\right)\right) \tag{2}$$

$$Z_{ij} = X_{ij} \otimes Y_{ij} \tag{3}$$

where GAP and GMP are the squeeze operations along the spatial dimension, W_j^1 and W_j^2 denote the weights of two fully connected layers for excitation, δ denotes the ReLU function, σ denotes the sigmoid function, and $\otimes$ denotes the elementwise multiplication. Finally, set-level representation of the jth bin is obtained as follows:

$$P_j = \text{SetPool}\big(Z_{1j}, Z_{2j}, \ldots, Z_{Nj}\big) \tag{4}$$

where SetPool denotes the set pooling to aggregate the features in an unordered set [30] and is implemented by max pooling along the set dimension. Particularly, we use Y_{ij} to recalibrate the features and add Y_{ij} along the channel dimension as frame quality indicator for the ith silhouette and jth region.

It is worth noting that the squeeze-and-excitation style is first proposed in squeeze-and-excitation network (SENet) [19] and FQBlock differs from it in three aspects. First, SENet is proposed for single image classification, while FQBlock is proposed to assess the frame quality in an unordered set, and the intermediate output of FQBlock is taken as frame quality indicator for each silhouette. Second, FQBlock takes GAP and GMP to squeeze the information in each channel. Third, FQBlock holds independent weights for features of different bins to deal with large shape variances in the silhouettes. Moreover, the attention mechanism is widely used in the field of action recognition [34]–[36]. Specifically, the methods in [34] and [35] use a *recurrent model* to discover the representative area in the *consecutive frames*, while FQBlock mainly consists of two *fully connected layers* and deals with the *unordered sets*. Self-attention network (SAN) [36] uses the *self-attention mechanism* [37] to capture the *correlation of position and motion* among different frames, while FQBlock deals with the quality of *each silhouette* and the following PQBlock deals with the quality of *each part* for a gait sequence *separately*. It is worth noting that the proposed quality modules are different from the popular self-attention mechanism [37]. For example, the vectors of query, key, and values, which are essential for the self-attention mechanism, are not involved in our method.

B. Part Quality

Horizontally and equally slicing the features to obtain the part representations has been widely used for gait recognition [10]–[12]. To keep the notations consistent, we use $\widehat{P}_j$ ($j \in [1, 2, \ldots, S]$) to denote the set-level jth part representation extracted from the gait sequence G, which is obtained by applying 1×1 convolution on P_j output by (4). Then, the distance of two gait sequences (denoted as G_1 and G_2) is computed as

$$D_{\text{eq}}(G_1, G_2) = \frac{1}{S}\sum_{j=1}^{S} D\Big(\widehat{P}_j^{G_1}, \widehat{P}_j^{G_2}\Big) \tag{5}$$

where S is the number of parts, and $D()$ measures the distance of two part representations, e.g., Euclidean distance.[1] All the parts are treated equally which is not optimal for gait recognition. For example, in the cases of changing coats or jackets, the head and legs should be assigned larger weights compared with the upper body. In GQAN, PQBlock is proposed to learn an adaptive weight for each part to match the gait sequences.

The structure of PQBlock is simple yet effective which consists of a fully connected layer followed by a Sigmoid function. It operates on set-level part representations and predicts a score to assess the relative importance of each part. The weights of the fully connected layer are independent for different parts. Formally, for the jth part, P_j in (4) is taken as the input, and the output score q_j is computed as

$$q_j = \sigma\left(M_j P_j\right) \tag{6}$$

where M_j denotes the weights of the fully connected layer, and σ denotes the sigmoid function. Then, the distance of G_1 and G_2 is computed as

$$D_{\text{ada}}(G_1, G_2) = \frac{1}{\sum_{j=1}^{S} q_j^{G_1} q_j^{G_2}} \sum_{j=1}^{S} q_j^{G_1} q_j^{G_2} D\Big(\widehat{P}_j^{G_1}, \widehat{P}_j^{G_2}\Big) \tag{7}$$

where $q_j^{G_1}$ and $q_j^{G_2}$ are the predicted scores of PQBlock for the jth part of G_1 and G_2, respectively. In this way, different parts are adaptively weighted for distance computation of two gait sequences.

C. Part Quality Loss

A key challenge for GQAN is how to train PQBlock. The current gait datasets [17], [18] are provided only with sequence-level identity annotations, and it is impossible to manually annotate weights for each part. To address the issue, we propose a PQLoss which enables PQBlock to be trained with only sequence-level identity annotations.

Before introducing PQLoss, we first review some nouns including *anchor*, *positive*, and *negative* which are widely used in the literature of metric learning [38], [39]. Traditionally, anchor-positive represents two samples belonging to the same class, while anchor-negative represents two samples belonging to different classes. In this work, we use *anchor-positive* to denote *two gait sequences belonging to the same subject* and *anchor-negative* to denote *two gait sequences belonging to different subjects*. The core idea of PQLoss is to *make anchor-positive closer and anchor-negative further in the feature space*.

Formally, PQLoss is computed as follows:

$$\begin{aligned} L_{pq} &= \frac{1}{N_{\text{ap}^+}} \sum_{l(G_1)=l(G_2)} \big[m_{\text{ap}} + D_{\text{ada}}(G_1, G_2) - D_{\text{eq}}(G_1, G_2)\big]_+ \\ &\quad + \frac{1}{N_{\text{an}^+}} \sum_{l(G_1)\neq l(G_3)} \big[m_{\text{an}} + D_{\text{eq}}(G_1, G_3) - D_{\text{ada}}(G_1, G_3)\big]_+ \end{aligned} \tag{8}$$

where the first term encourages the distance of *anchor-positive* with the adaptive weights to be smaller than that with equal weights, and the second term encourages the distance of *anchor-negative* with the adaptive weights to be larger than that with equal weights. $l()$ indicates the identity label of the gait sequence, N_{ap^+} and N_{an^+} are the number of *anchor-positive* and *anchor-negative* pairs resulting in nonzero loss

[1] In some works such as [10], the representations of all the parts are concatenated to compute the distance of two gait sequences while the performance is no better than that using the distance computed in (5).

terms, respectively, and m_{ap} and m_{an} are the margin thresholds to avoid correcting "already correct" pairs. It can be observed that the computation of PQLoss only requires sequence-level annotations, and the relative importance of each part can be automatically learned.

D. Training and Evaluation

GQAN is proposed to enhance the interpretability of silhouette-based gait recognition by explicitly assessing the quality of each silhouette and each part via FQBlock and PQBlock. A PQLoss is proposed to make it feasible to train GQAN with only sequence-level identity annotations. Besides, we design an efficient and effective backbone for GQAN. The backbone is similar to GaitSet [10], and the modification details will be provided in Section IV-A. In this section, we will respectively introduce the training and evaluation for GQAN-Backbone and GQAN which constitute a fair comparison.

1) Training and Evaluation for GQAN-Backbone: For the training of GQAN-Backbone, the loss denoted as L_1 consists of a triplet loss L_{tp} and a cross-entropy loss L_{ce}.

First, the triplet loss L_{tp} is computed on the features of each part. Formally, for the jth part, L_{tp}^j is computed as

$$L_{\text{tp}}^j = \frac{1}{N_{\text{tp}^+}^j} \sum_{\substack{l(G_1)=l(G_2) \\ l(G_1)\neq l(G_3)}} \left[m + D\left(\widehat{P}_j^{G_1}, \widehat{P}_j^{G_2}\right) - D\left(\widehat{P}_j^{G_1}, \widehat{P}_j^{G_3}\right) \right]_+ \tag{9}$$

where $N_{\text{tp}^+}^j$ is the number of triplets resulting in nonzero terms for the jth part, m is the margin threshold, and L_{tp} is obtained by averaging L_{tp}^j for all the parts. It is worth noting that different from L_{tp}^j computed on the *part-level* distance, L_{pq} in (8) is computed on the *sequence-level* distance, and L_{pq} consists of two terms for the distance of *anchor-positive anchor-negative pairs, respectively* (instead of triplets).

Second, the cross-entropy loss L_{ce} is computed in a similar way to that for image classification [40], [41]. In our experiments, it takes the concatenated features of all the parts as the input and treats each subject in the training set as a separate class. For simplicity, we omit the computation of L_{ce} here; refer to [12] for details.

Finally, the loss L_1 to train GQAN-Backbone is computed as

$$L_1 = L_{\text{tp}} + \alpha L_{\text{ce}} \tag{10}$$

where α is a loss weight to balance the two terms.

For the evaluation of GQAN-Backbone, we compute the Euclidean distance averaged on all the parts to match the probe and gallery as shown in (5).

2) Training and Evaluation for GQAN: For the training of GQAN including FQBlock and PQBlock, the loss denoted as L_2 consists of a cross-entropy loss L_{ce} and a triplet loss L_{tp} as well as a PQLoss L_{pq} shown in (8).

Formally, the loss L_2 to train GQAN is computed as

$$L_2 = L_{\text{tp}} + \alpha L_{\text{ce}} + \beta L_{pq} \tag{11}$$

where β is a loss weight for PQLoss.

For the evaluation of GQAN, we compute the Euclidean distance between the probe and gallery with the parts adaptively weighted as shown (7).

IV. Experiments

A. Experimental Settings

The experimental settings for GQAN are similar to the baseline methods [10], [12], [13] to ensure fair comparisons. The methods in [12] and [13] are our previous works which aim to learn more discriminative features from the silhouettes for gait recognition. Differently, the main goal of GQAN is to enhance the *interpretability* of silhouette-based gait recognition, which also achieves very competitive performance under all walking conditions.

1) GQAN-Backbone: In our experiments, we design an effective and efficient backbone for GQAN. Specifically, GQAN-Backbone is modified from GaitSet [10], and our modifications mainly lie as follows.

1) We use $S = 16$ instead of $S = \{1, 2, 4, 8, 16\}$ for simplicity to horizontally slice the features in horizontal pyramid matching.
2) We remove multilayer global pipeline to make it feasible to separately assess the quality of each silhouette in high layers, which can also accelerate training and reduce GPU memory consumption.
3) We add BNNeck [43] and compute the cross-entropy loss on the concatenated features of all the parts.
4) We use the warmup strategy [41] to adjust the learning rate at the start of training.
5) We adopt random erasing data augmentation [44] to alleviate the overfitting on CASIA-B [17].
6) We add two additional convolutional layers in the encoder for the experiments on OUMVLP [18] to adapt to large-scale dataset.

It is worth noting that the networks in GaitSet [10] and [11]–[13] cannot be directly adopted as the backbone for GQAN. Specifically, the networks in GaitSet [10], gait lateral network (GLN) [12], and set residual network (SRN) [13] consist of a global branch (e.g., multilayer global pipeline in GaitSet [10]) to aggregate silhouette-level features at the early layers which makes it infeasible to separately assess the quality of each silhouette in high layers. Besides, GaitPart [11] relies on an micro-motion capture module (MCM) to model the micromotion features in adjacent frames, and we provide the comparison between FQBlock and MCM in Section IV-D1.

2) Datasets: The experiments are mainly conducted on two typical gait datasets, i.e., CASIA-B [17] and OUMVLP [18].

CASIA-B consists of 124 subjects and collects the videos of normal walking (NM-1,2,3,4,5,6), walking with bags (BG-1,2), and walking in different coats/jackets (CL-1,2). There are 11 views for each walking condition. Since there is no split way provided in the dataset, we take the first 74 subjects as the training set with the remaining 50 subjects as the test set. For evaluation, the sequences of NM-1,2,3,4 for each subject are taken as the gallery, and the sequences of NM-5,6, BG-1,2, CL-1,2 are taken as the probe.

OUMVLP consists of 10 307 subjects which is the largest public gait dataset so far. However, it only provides the

TABLE I

RANK-1 ACCURACY (%) ON CASIA-B FOR DIFFERENT PROBE VIEWS EXCLUDING THE IDENTICAL VIEW CASES. FOR EVALUATION, THE SEQUENCES OF NM-1,2,3,4 FOR EACH SUBJECT ARE TAKEN AS THE GALLERY

	Method	Probe View											Average
		0°	18°	36°	54°	72°	90°	108°	126°	134°	162°	180°	
NM	GEINet [42]	40.20	38.90	42.90	45.60	51.20	42.00	53.50	57.60	57.80	51.80	47.70	48.11
	CNN-LB [3]	82.60	90.30	96.10	94.30	90.10	87.40	89.90	94.00	94.70	91.30	78.50	89.93
	GaitSet [10]	93.40	98.10	98.50	97.80	92.60	90.90	94.20	97.30	98.40	97.00	89.10	95.21
	GaitPart [11]	94.10	98.60	99.30	98.50	94.00	92.30	95.90	98.40	99.20	97.80	90.40	96.23
	GLN [12]	93.20	99.30	99.50	98.70	96.10	95.60	97.20	98.10	99.30	98.60	90.10	96.88
	SRN [13]	94.70	99.40	99.40	98.40	96.50	94.80	96.00	98.20	99.30	98.40	92.90	97.09
	GQAN-Backbone(**ours**)	95.80	99.60	99.70	99.10	**97.90**	96.30	**97.80**	98.70	99.50	98.30	94.10	97.89
	GQAN(**ours**)	**98.00**	**99.80**	**99.80**	**99.20**	97.70	**97.30**	**97.80**	**98.80**	**99.80**	**99.20**	**96.20**	**98.51**
BG	GEINet [42]	34.20	29.29	31.21	35.20	35.20	27.60	35.90	43.50	45.00	38.99	36.80	35.72
	CNN-LB [3]	64.20	80.60	82.70	76.90	64.80	63.10	68.00	76.90	82.20	75.40	61.30	72.37
	GaitSet [10]	85.90	92.12	93.94	90.41	86.40	78.70	85.00	91.60	93.10	91.01	80.70	88.08
	GaitPart [11]	89.10	94.80	96.70	95.10	88.30	84.90	89.00	93.50	96.10	93.80	85.80	91.55
	GLN [12]	91.10	97.68	97.78	95.20	92.50	91.20	92.40	**96.00**	97.50	94.95	88.10	94.04
	SRN [13]	92.00	97.37	97.58	95.82	91.80	90.40	93.20	95.30	97.60	95.35	87.80	94.02
	GQAN-Backbone(**ours**)	93.90	97.27	97.37	96.43	**94.00**	**92.60**	93.10	95.40	97.40	96.97	88.70	94.83
	GQAN(**ours**)	**96.00**	**98.69**	**98.38**	**96.94**	93.40	90.80	**93.70**	95.90	**97.70**	**97.07**	**90.50**	**95.37**
CL	GEINet [42]	19.90	20.30	22.50	23.50	26.70	21.30	27.40	28.20	24.20	22.50	21.60	23.46
	CNN-LB [3]	37.70	57.20	66.60	61.10	55.20	54.60	55.20	59.10	58.90	48.80	39.40	53.98
	GaitSet [10]	63.70	75.60	80.70	77.50	69.10	67.80	69.70	74.60	76.10	71.10	55.70	71.05
	GaitPart [11]	70.70	85.50	86.90	83.30	77.10	72.50	76.90	82.20	83.80	80.20	66.50	78.69
	GLN [12]	70.60	82.40	85.20	82.70	79.20	76.40	76.20	78.90	77.90	78.70	64.30	77.50
	SRN [13]	75.10	88.20	89.90	86.30	81.20	78.80	80.00	84.00	86.30	80.70	68.80	81.75
	GQAN-Backbone(**ours**)	71.70	84.00	88.70	84.30	83.20	78.30	81.80	83.20	83.60	77.50	66.10	80.22
	GQAN(**ours**)	**80.20**	**90.30**	**90.20**	**87.40**	**85.50**	**81.50**	**83.70**	**85.30**	**86.90**	**83.30**	**75.30**	**84.51**

silhouettes of normal walking (NM-00,01) for each subject. There are 14 views available for normal walking. According to the split way provided in the dataset, we take 5153 subjects as the training set with the remaining 5154 subjects as the test set. For evaluation, the sequences of NM-01 for each subject are taken as the gallery, and the sequences of NM-00 are taken as the probe.

3) Implementation Details: All the models are implemented with PyTorch [45] and trained on TITAN-V GPUs.

The silhouettes in both the datasets are preprocessed using the method in [22]. The input size of each silhouette is set to 128×88 for CASIA-B and 64×44 for OUMVLP. In the training phase, we randomly select 30 silhouettes from each sequence as the input. The number of subjects and the number of sequences for each subject in a batch are set to $(8, 16)$ for CASIA-B and $(32, 16)$ for OUMVLP. For evaluation, all the silhouettes for each sequence are taken as the input to obtain the representations.

The convolutional channels are set to $\{32, 64, 128\}$ for CASIA-B and $\{64, 128, 256, 512\}$ for OUMVLP. The output dimension for each part representation is set to 256. For FQBlock, there is no dimension reduction, and the channels in the two fully connected layers are the same as the output channel of the last convolutional layer (128 for CASIA-B and 512 for OUMVLP). For PQBlock, the output dimension of the fully connected layer is set to 1.

SGD with momentum is taken as the optimizer. The learning rate is initialized with 0.1 and scaled to its $1/10$ three times for training. The stepsize is set to 10 000 iterations for CASIA-B and 50 000 iterations for OUMVLP. The momentum and the weight decay are set to 0.9 and 0.0005, respectively, for optimization. Particularly, GQAN is pretrained without PQLoss using the initial learning rate 0.1 for 10 000 iterations for CASIA-B and 50 000 iterations for OUMVLP.

Besides, the margin thresholds m_{ap} and m_{an} in (8) are both set to 0.01, the margin threshold m in (9) is set to 0.2, the loss weight α in (10) and (11) is set 0.1, and the loss weight β in (11) is set to 10.0.

B. Performance Comparison

1) CASIA-B: Table I shows the performance comparison on CASIA-B. The probe sequences are divided into three categories according to walking conditions, i.e., NM, BG, and CL, which are respectively evaluated. We report rank-1 accuracy for each probe view averaged on all the gallery views excluding the identical view cases [3].

In the methods listed in Table I, GEINet [42] and CNN-LB [3] are two representative methods taking GEIs as the input. GaitSet [10] first proposes to treat the silhouettes of a gait sequence as an unordered set and horizontally slices the features to learn part representation for gait recognition, which achieves significant improvement compared with the GEI-based methods. GaitPart [11] uses a focal convolutional layer and MCM to enhance part representations. GLN [12] takes the lateral connections to merge multilayer features and proposes a compact block to reduce representation dimension. SRN [13] proposes a set residual block to effectively coordinate the silhouette-level and set-level information in feature learning.

For the results in Table I, we can observe that the backbone we design for GQAN achieves competitive performance (NM-97.89%, BG-94.83%, and CL-80.22%) compared with previous works. FQBlock and PQBlock for GQAN, which

TABLE II

RANK-1 ACCURACY (%) ON OUMVLP FOR DIFFERENT PROBE VIEWS EXCLUDING THE IDENTICAL VIEW CASES. FOR EVALUATION, THE SEQUENCES OF NM-01 FOR EACH SUBJECT ARE TAKEN AS THE GALLERY. THE PROBE SEQUENCES WHICH HAVE NO CORRESPONDING ONES IN THE GALLERY ARE INCLUDED AND IGNORED, RESPECTIVELY

Method	Probe View														Average
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GEINet [42]	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 [10]	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
GaitPart [11]	82.57	88.93	90.84	91.00	89.75	89.91	89.50	85.19	88.09	90.02	90.15	89.03	89.10	88.24	88.74
GLN [12]	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
SRN [13]	83.76	89.70	90.94	91.19	89.88	90.25	89.61	85.76	88.79	90.11	90.41	89.03	89.36	88.47	89.09
GQAN-Backbone(**ours**)	84.38	90.06	91.15	91.30	90.41	90.41	89.86	86.85	89.13	90.34	90.51	89.75	89.50	88.77	89.46
GQAN(**ours**)	**84.99**	**90.34**	**91.26**	**91.40**	**90.63**	**90.57**	**90.14**	**87.09**	**89.37**	90.46	**90.64**	**90.02**	**89.81**	**89.10**	**89.70**
GEINet [42]	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 [10]	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
GaitPart [11]	87.95	94.70	97.69	97.59	95.46	96.60	96.15	90.61	94.25	97.17	97.06	95.07	96.02	95.02	95.10
GLN [12]	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
SRN [13]	89.22	95.52	97.79	97.81	95.62	96.97	96.28	91.22	95.01	97.26	97.35	95.07	96.31	95.28	95.48
GQAN-Backbone(**ours**)	89.88	95.92	98.03	97.94	96.20	97.17	96.57	92.37	95.37	97.51	97.46	95.88	96.48	95.62	95.89
GQAN(**ours**)	**90.53**	**96.20**	**98.14**	**98.04**	**96.44**	**97.34**	**96.88**	**92.63**	**95.63**	97.64	**97.61**	**96.18**	**96.82**	**95.98**	**96.15**

TABLE III

PERFORMANCE COMPARISON ON HID COMPETITION DATASET 2021. THE RESULTS ARE REPORTED IN RANK-1 ACCURACY

Method	SRN [13]	GQAN-Backbone	GQAN
Rank-1 Acc	64.31	58.19	**65.61**

are proposed to assess the quality of each silhouette and each part for gait recognition, can further boost the performance under all walking conditions to state-of-the-art (NM-98.51%, BG-95.37%, and CL-84.51%). Particularly, under the most challenging condition of walking in different coats/jackets, rank-1 accuracy achieved by GQAN exceeds GQAN-Backbone by a large margin (+4.29%). In Section IV-A1, we have explained why the networks in [10]–[13] cannot be directly adopted as the backbone for GQAN. While in Table I, GQAN-Backbone and GQAN constitute a fair comparison, and the performance gain brought by GQAN validates the effectiveness of FQBlock and PQBlock.

2) OUMVLP: Table II shows the performance comparison on OUMVLP. Although a large number of subjects are available, the lack of walking with bags (BG) and walking in different clothes (CL) makes it less challenging than CASIA-B. Due to the incomplete data for some subjects, we respectively conduct the evaluation *including* and *ignoring* the probe sequences which have no corresponding ones in the gallery.

Compared with the methods listed in Table I, CNN-LB [3] is too time-consuming for training and test which is thus not listed in Table II for large-scale dataset. Here, we mainly compare rank-1 accuracy obtained by *ignoring* the probe sequences which do not have the corresponding ones in the gallery. From the results in Table II, we can observe that the baselines including GaitPart [11], GLN [12], and SRN [13] all report rank-1 accuracy of more than 95% on this dataset. Particularly, GQAN-Backbone achieves the competitive accuracy of 95.89% which validates the effectiveness of the backbone we designed for GQAN. Besides, we note that the frame quality of each silhouette in OUMVLP is obviously higher than that in CASIA-B, and the silhouettes for each subject walking in different clothes (CL) are not available. FQBlock and PQBlock, which work by explicitly assessing the quality of each silhouette and each part, can still benefit from gait recognition and improve rank-1 accuracy to 96.15%.

3) HID Competition Dataset 2021: CASIA-B and OUMVLP are the two most popular benchmarks which are widely used in previous literature [10]–[13]. The gait dataset is lacking due to privacy issue and the requirement for cameras and sites. To further verify the effectiveness of GQAN, we conduct the experiments on HID Competition Dataset 2021 [46] using settings similar to CASIA-B. The competition provides the sequences for 500 subjects, and each subject consists of about ten sequences. We take the sequences of the first 300 subjects as the training set and the remaining 200 subjects as the test set. For evaluation, we gather the first sequence of each subject as the gallery and take the remaining sequences as the probe. The performance comparison is provided in Table III, and the results are reported in rank-1 accuracy. Along with GQAN-Backbone and GQAN, we also reimplement SRN on this dataset which holds the best performance for gait recognition before this work. The performance comparison shown in Table III, especially the performance gain from GQAN-Backbone to GQAN, further validates the effectiveness of the proposed method.

C. Ablation Study

In this section, we conduct more experiments to further analyze GQAN. For simplicity, we report rank-1 accuracy under different walking conditions averaged on all the probe and gallery views excluding the identical view cases.

1) Effect of Each Block: GQAN mainly consists of two blocks, i.e., FQBlock and PQBlock, to explicitly assess the quality of each silhouette and each part. It is proposed toward the interpretability of silhouette-based gait recognition which also achieves very competitive performance. Here, we conduct the experiments to separately evaluate the effect of each block on recognition accuracy. The experimental results are provided in Table IV. From the results on CASIA-B, we can observe

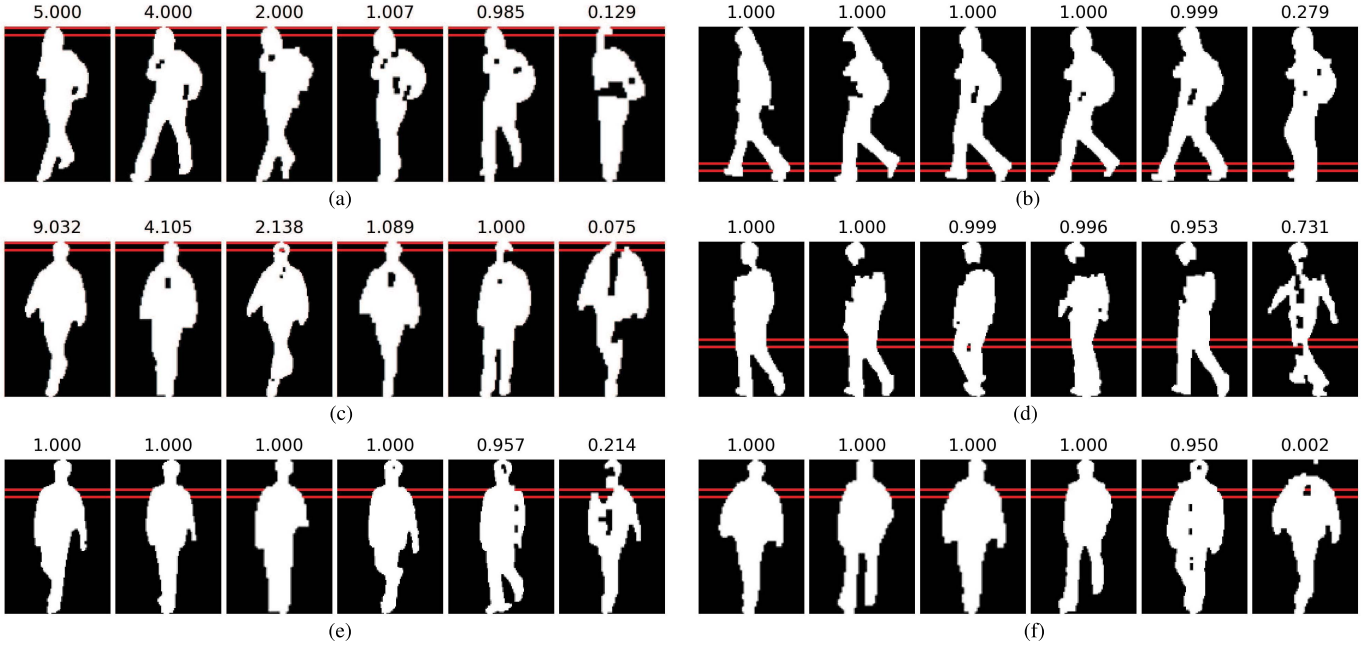


Fig. 4. Examples of frame quality visualization on CASIA-B. The silhouettes in each figure are randomly selected from each gait sequence and sorted in descending order according to the frame quality scores of a certain bin. The corresponding region in each silhouette is marked in red. The number above each silhouette is the frame quality score which is computed by adding Y_{ij} in (2) along the channel dimension. (a) ID = 110, Type = BG-01, View = 036, Bin = 1. (b) ID = 110, Type = BG-01, View = 090, Bin = 15. (c) ID = 112, Type = CL-02, View = 018, Bin = 1. (d) ID = 112, Type = NM-06, View = 144, Bin = 11. (e) ID = 123, Type = BG-02, View = 018, Bin = 4. (f) ID = 123, Type = CL-02, View = 018, Bin = 4.

TABLE IV

EFFECT OF EACH BLOCK FOR GQAN. THE RESULTS ARE REPORTED IN RANK-1 ACCURACY EXCLUDING THE IDENTICAL VIEW CASES

Dataset	CASIA-B			OUMVLP
Method	NM	BG	CL	NM
GQAN-Backbone	97.89	94.83	80.22	95.89
GQAN-Backbone+FQBlock	98.60	95.41	83.72	96.01
GQAN-Backbone+PQBlock	97.96	94.68	82.36	96.01
GQAN	98.51	95.37	84.51	96.15

TABLE V

PERFORMANCE COMPARISON UNDER LOW-RESOLUTION CONDITIONS. THE EXPERIMENTS ARE CONDUCTED ON THE TEST SET OF CASIA-B AND THE RESULTS ARE REPORTED IN RANK-1 ACCURACY EXCLUDING THE IDENTICAL VIEW CASES

Downsample Size	Method	NM	BG	CL
-	GQAN-Backbone	97.89	94.83	80.22
	GQAN	98.51	95.37	84.51
64×44	GQAN-Backbone	96.77	93.29	74.93
	GQAN	97.89	94.16	79.67
32×22	GQAN-Backbone	72.11	65.11	37.29
	GQAN	76.63	69.85	44.10

that FQBlock can improve the performance for all walking conditions (NN, BG, and CL). The probable reason is that the silhouettes in CASIA-B are obtained by subtracting the background [17] and contain a lot of noise. PQBlock is mainly beneficial for walking in different coats/jackets (CL) which causes a lot of shape variance for the upper body. For OUMVLP, as mentioned above, the overall improvement is not that significant due to the high quality of each silhouette and the lack of walking in different clothes (CL), while the performance comparison shown in Table IV indicates that FQBlock and PQBlock can still help improve the recognition accuracy respectively.

2) Performance Comparison Under Low Resolution: In this section, we provide the performance comparison between GQAN-Backbone and GQAN under low-resolution conditions for a more comprehensive study. Specifically, the experiments are conducted on the test set of CASIA-B covering different walking conditions. In the above experiments, the input size of each silhouette is set to 128×88 for training and test. While in this section, we first downsample the silhouettes in the test set and then restore to 128×88 to simulate the low-resolution conditions. The results shown in Table V show that GQAN can consistently outperform the backbone, which indicates that the proposed method is more robust to low-resolution conditions.

3) Frame Quality Visualization: As described in Section III-A, FQBlock works in a squeeze-and-excitation style to assess the frame quality of each silhouette for gait recognition. Specifically, it predicts the scores Y_{ij} as shown in (2) to recalibrate the features of the ith silhouette and j-bin, and we add Y_{ij} along the channel dimension as the frame quality indicator for the ith silhouette and jth region. In Fig. 4, we present some examples from CASIA-B, and the silhouettes are sorted in descending order according to the predicted scores of FQBlock. As aforementioned, the silhouettes in CASIA-B contain a lot of noise due to the errors in background extraction [17]. From the results in Fig. 4, we can observe that Y_{ij} can be treated as an indicator of frame quality for each silhouette in a gait sequence.

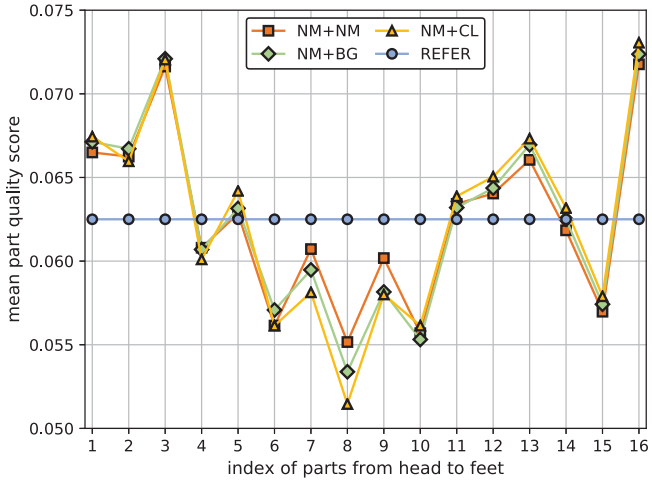


Fig. 5. Statistics of part quality scores on CASIA-B. We conduct the experiment on the test set and average the scores for each part when computing the distance of gait sequences belonging to different types. *REFER* for *the reference scores of treating all the parts equally (0.0625).*

Fig. 6. Example silhouettes from CASIA-B. The silhouettes in CASIA-B contain a lot of noise for the 15th region due to the shadow in the floor. The corresponding region in each silhouette is marked in red.

4) Part Quality Visualization: As described in Section III-B, PQBlock operates on set-level part representations and it predicts a score encoding the part quality to generate adaptive weights as shown in (7). For visualization of PQBlock, we obtain the statistics of part quality scores on the test set of CASIA-B when computing the distance of gait sequences belonging to different walking conditions, including NM-NM, NM-BG, and NM-CL. The results are displayed in Fig. 5. It can be observed that the part quality scores for the upper body (sixth to tenth parts) are relatively low especially for the cases of NM-CL. Besides, the mean score of the 15th part is relatively smaller than the reference score of treating all the parts equally, which is possibly caused by the segmentation errors due to the shadow in the floor [17]. In Fig. 6, we display some example silhouettes from CASIA-B to illustrate the segmentation errors caused by the shadow in the floor.

D. Discussion

1) Comparison With MCM: In this section, we provide the comparison between FQBlock and MCM [11] which take the part features extracted by the backbone as the input. First, the motivation behind the two blocks is different. MCM is designed to model the micromotion features in adjacent frames, while FQBlock is proposed to measure the quality of each frame. Second, the working mechanism of the two blocks is different. For simplicity, given a gait sequence, we denote the three dimensions of tensor to FQBlock and MCM as *D-S*, *D-P*, and *D-C*. Specifically, *D-S* denotes the number of silhouettes, *D-P* denotes the number of human parts, and *D-C* denotes the number of feature channels. MCM mainly works on *D-S* dimension to deal with the relation of the adjacent frames, while FQBlock works on *D-C* dimension to deal with the features of each frame separately. Third, the attention values in the two blocks have different characteristics. Specifically, the attention values of each frame in MCM rely on adjacent ones, which are susceptible to temporal kernel size, silhouette order, and missing frames. Moreover, the attention values for the frames in different sequences are not comparable due to the variation in context. As a result, the attention values in MCM are not suitable as frame quality indicator. In contrast, the attention values of each frame in FQBlock only rely on the features of itself and are permutation-invariant to silhouette order. Besides, FQBlock shares the weights across different silhouettes, which makes the attention values of the frames in different sequences comparable. The visualization results shown in Fig. 4 indicate that the attention values in FQBlock can be taken as frame quality indicator.

In addition, we conduct the experiment with MCM using the proposed strong baseline on CASIA-B. The performance of GQAN-Backbone + MCM (NM-97.96%, BG-94.92%, and CL-81.74%) is inferior to GQAN-Backbone + FQBlock (NM-98.60%, BG-95.41%, and CL-83.72%), which further demonstrates the effectiveness of FQBlock.

2) Comparison With More Methods: For a comprehensive study, we provide the performance comparison with more methods on CASIA-B and OUMVLP in Table VI. Most of these works are orthogonal to GQAN such as the model-based methods (e.g., PoseGait [6] and End2EndGait [7]) and the appearance methods taking other types of input for gait recognition (e.g., GaitNet [27] and GaitMotion [28]). Particularly, SM-Prod [29] reports a little higher accuracy for NM and BG on CASIA-B. However, the optical flow needs a lot of computation cost, and the performance for the most challenging CL is much inferior to GQAN. Besides, SRN + CBlock [13] is a variant of SRN which integrates SRN with compact block proposed in [12]. It reports a little higher accuracy on OUMVLP, while its performance on CASIA-B is inferior to GQAN especially for the challenging CL (77.7% versus 84.51%).

3) Discussion on Training Tricks: As described in Section IV-A1, we adopt some useful tricks to train GQAN-Backbone and achieve competitive performance. In this section, we try to add these tricks to SRN [13] which holds the best performance before this work. We conduct the experiments on CASIA-B, and the performance under different walking conditions is moderately improved (NM-98.01%, BG-95.09%, and CL-83.34%) which yet is still inferior to GQAN (NM-98.51%, BG-95.37%, and CL-84.51%). More importantly, GQAN can enhance the interpretability of silhouette-based gait recognition by trying to find out the relative importance of each silhouette and each part.

4) Discussion on Generalization Ability: In this section, we conduct the experiments to verify the generalization ability of the learned gait quality. Specifically, we perform frame quality visualization on HID Competition Dataset 2021 using

TABLE VI

PERFORMANCE COMPARISON WITH MORE BASELINES. THE RESULTS ARE REPORTED IN RANK-1 ACCURACY EXCLUDING THE IDENTICAL VIEW CASES. SIL FOR SILHOUETTES, RGB FOR RGB FRAMES, OF FOR OPTICAL FLOW

Dataset	Method	Input	NM	BG	CL
CASIA-B	J-CNN [47]	Sil	91.2	75.0	54.0
	GaitSet-L [12]	Sil	95.6	91.5	75.3
	GLN-Backbone [12]	Sil	95.5	92.0	77.2
	SRN+CBlock [13]	Sil	97.5	94.3	77.7
	MT3D [9]	Sil	96.7	93.0	81.5
	PoseGait [6]	RGB	68.7	44.5	36.0
	GaitNet [27]	RGB	92.3	88.9	62.3
	End2EndGait [7]	RGB	97.9	93.1	77.6
	GaitMotion [28]	OF	97.5	83.6	48.8
	SM-Prod [29]	Gray+OF	**99.8**	**96.1**	67.0
	GQAN(**ours**)	Sil	98.51	95.37	**84.51**
OUMVLP	GLN-Backbone [12]	Sil	94.2	-	-
	SRN+CBlock [13]	Sil	**96.4**	-	-
	End2EndGait [7]	RGB	95.8	-	-
	GQAN(**ours**)	Sil	96.15	-	-

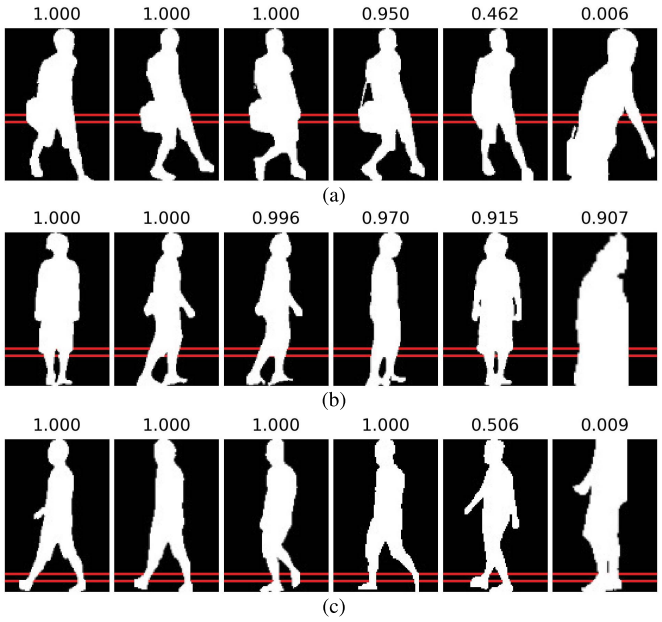


Fig. 7. Examples of frame quality visualization on HID Competition Dataset 2021. The settings are the same as those in Section IV-C3 except that the model is only trained on CASIA-B. (a) ID = 400, BIN = 10. (b) ID = 401, BIN = 13. (c) ID = 403, BIN = 15.

the same settings as those in Section IV-C3 except that the model is only trained on CASIA-B. We provide some visualization results randomly selected from HID Competition Dataset 2021 in Fig. 7, which indicates that the learned gait quality can be generalized to an unseen dataset.

5) Discussion on Gait Interpretability: The interpretability of gait recognition is greatly important for real-world applications. In this work, we move toward the interpretability of silhouette-based gait recognition by explicitly assessing the quality of each silhouette and each part. However, the problem needs further exploration. For example, part representations for gait recognition in most works [10]–[12] are obtained by horizontally and equally slicing the features which are inconsistent with the semantic parts of human body, e.g., hands or feet. The interpretability of model-based gait recognition also needs to be explored where the weights for different key points are not explicitly modeled in the previous works [6], [7]. Moreover, fusing the multimodal features proves to be useful for many other visual tasks [48], [49]. It is also promising to fuse the silhouettes and key points to obtain more rich features for gait recognition, and how to enhance the interpretability in this case remains a challenge.

V. CONCLUSION

In this work, we propose a GQAN toward the interpretability of silhouette-based gait recognition. It tries to assess the quality of each silhouette and each part for the recognition via two blocks: FQBlock and PQBlock. Specifically, FQBlock works in a squeeze-and-excitation style to recalibrate the features for each silhouette, and the scores for all the channels are added as frame quality indicator. PQBlock predicts a score for each part to compute the weighted distance between the probe and gallery. A PQLoss is proposed which enables GQAN to be trained with only sequence-level identity annotations. Besides, GQAN achieves very competitive performance under all walking conditions on CASIA-B and OUMVLP.

REFERENCES

[1] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon, "On using gait in forensic biometrics," *J. Forensic Sci.*, vol. 56, no. 4, pp. 882–889, 2011.
[2] X. Gu, Y. Guo, F. Deligianni, B. Lo, and G.-Z. Yang, "Cross-subject and cross-modal transfer for generalized abnormal gait pattern recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 546–560, Feb. 2020.
[3] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2016.
[4] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," 2021, *arXiv:2102.09546*.
[5] W. An *et al.*, "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 4, pp. 421–430, Oct. 2020.
[6] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, Oct. 2020, Art. no. 107069.
[7] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proc. ACCV*, 2020, pp. 1–17.
[8] K. Zhang, W. Luo, L. Ma, W. Liu, and H. Li, "Learning joint gait representation via quintuplet loss minimization," in *Proc. CVPR*, 2019, pp. 4700–4709.
[9] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3d convolutional neural network," in *Proc. ACM Multimedia*, 2020, pp. 3054–3062.
[10] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8126–8133.
[11] C. Fan *et al.*, "GaitPart: Temporal part-based model for gait recognition," in *Proc. CVPR*, 2020, pp. 14225–14233.
[12] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Proc. ECCV*, 2020, pp. 382–398.
[13] S. Hou, X. Liu, C. Cao, and Y. Huang, "Set residual network for silhouette-based gait recognition," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 384–393, Jul. 2021.
[14] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.
[15] B. A. L. A. Zhou Khosla and A. A. O. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016, pp. 2921–2929.
[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, 2017, pp. 618–626.

[17] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. IEEE Int. Conf. Pattern Recognit.*, vol. 4, Aug. 2006, pp. 441–444.

[18] 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. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–14, Dec. 2018.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Oct. 2015.

[21] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 1, pp. 102–113, Jan. 2019.

[22] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2019.

[23] B. Hu, Y. Guan, Y. Gao, Y. Long, N. Lane, and T. Ploetz, "Robust cross-view gait recognition with evidence: A discriminant gait GAN (DiGGAN) approach," 2018, *arXiv:1811.10493*.

[24] J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[25] W. Liu, C. Zhang, H. Ma, and S. Li, "Learning efficient spatial-temporal gait features with deep learning for human identification," *Neuroinformatics*, vol. 16, nos. 3–4, pp. 457–471, Oct. 2018.

[26] S. Tong, Y. Fu, X. Yue, and H. Ling, "Multi-view gait recognition based on a spatial-temporal deep neural network," *IEEE Access*, vol. 6, pp. 57583–57596, 2018.

[27] Z. Zhang *et al.*, "Gait recognition via disentangled representation learning," in *Proc. CVPR*, 2019, pp. 4710–4719.

[28] K. Bashir, T. Xiang, S. Gong, and Q. Mary, "Gait representation using flow fields," in *Proc. BMVC*, 2009, pp. 1–11.

[29] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. P. de la Blanca, "Multimodal feature fusion for CNN-based gait recognition: An empirical comparison," *Neural Comput. Appl.*, vol. 32, no. 17, pp. 14173–14193, Sep. 2020.

[30] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3d classification and segmentation," in *Proc. CVPR*, 2017, pp. 652–660.

[31] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. CVPR*, 2017, pp. 5790–5799.

[32] W. Xie and A. Zisserman, "Multicolumn networks for face recognition," in *Proc. BMVC*, 2018, pp. 1–12.

[33] L. Zhang, M. Xu, J. Yin, C. Zhang, and L. Shao, "Weakly supervised complets ranking for deep image quality modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5041–5054, Dec. 2020.

[34] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 1–11.

[35] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI*, 2017, vol. 31, no. 1, pp. 4263–4270.

[36] S. Cho, M. H. Maqbool, F. Liu, and H. Foroosh, "Self-attention network for skeleton-based human action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 635–644.

[37] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[38] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

[39] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[42] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.

[43] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.

[44] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13001–13008.

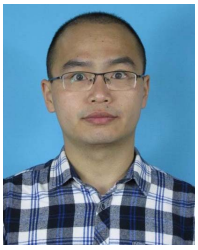
[45] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019, pp. 8024–8035.

[46] S. Yu *et al.*, "Hid 2021: Competition on human identification at a distance 2021," in *Proc. Int. Joint Conf. Biometrics*, 2021, pp. 1–7.

[47] Y. Zhang, Y. Huang, L. Wang, and S. Yu, "A comprehensive study on gait biometrics using a joint CNN-based method," *Pattern Recognit.*, vol. 93, pp. 228–236, Sep. 2019.

[48] Y.-D. Zhang *et al.*, "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation," *Inf. Fusion*, vol. 64, pp. 149–187, Dec. 2020.

[49] Y.-D. Zhang, S. C. Satapathy, D. S. Guttery, J. M. Górriz, and S.-H. Wang, "Improved breast cancer classification through combining graph convolutional network and convolutional neural network," *Inf. Process. Manage.*, vol. 58, no. 2, Mar. 2021, Art. no. 102439.



Saihui Hou received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2014 and 2019, respectively.

He is currently an Assistant Professor with the School of Artificial Intelligence, Beijing Normal University, Beijing, China. His research interests include computer vision and machine learning. He focuses on gait recognition which aims to identify different people according to walking patterns.

Xu Liu received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2013 and 2018, respectively.

He is currently a Research Scientist with Watrix Technology Company Limited, Beijing, China. His research interests include gait recognition, object detection, and image segmentation.

Chunshui Cao received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2013 and 2018, respectively.

During his Ph.D. study, he joined the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. From 2018 to 2020, he was a Post-Doctoral Fellow with the PBC School of Finance, Tsinghua University, Beijing. He is currently a Research Scientist with Watrix Technology Company Limited, Beijing. His research interests include pattern recognition, computer vision, and machine learning.

Yongzhen Huang received the B.E. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2006, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently an Associate Professor with the School of Artificial Intelligence, Beijing Normal University, Beijing. He has published one book and more than 80 articles at international journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *International Journal of Computer Vision* (IJCV), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—B: CYBERNETICS (TSMCB), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), Conference on Neural Information Processing Systems (NIPS), and AAAI Conference on Artificial Intelligence (AAAI). His research interests include pattern recognition, computer vision, and machine learning.