

GaitEdge: Beyond Plain End-to-end Gait Recognition for Better Practicality

Junhao Liang^{1*} , Chao Fan^{1*} , Saihui Hou^{2,4} , Chuanfu Shen^{3,1} ,
Yongzhen Huang^{2,4} , and Shiqi Yu¹  

¹ Southern University of Science and Technology

² School of Artificial Intelligence, Beijing Normal University

³ The University of Hong Kong

⁴ WATRIX.AI

Abstract. Gait is one of the most promising biometrics to identify individuals at a long distance. Although most previous methods have focused on recognizing the silhouettes, several end-to-end methods that extract gait features directly from RGB images perform better. However, we demonstrate that these end-to-end methods may inevitably suffer from the gait-irrelevant noises, *i.e.* low-level texture and color information. Experimentally, we design the **cross-domain** evaluation to support this view. In this work, we propose a novel end-to-end framework named **GaitEdge** which can effectively block gait-irrelevant information and release end-to-end training potential. Specifically, GaitEdge synthesizes the output of the pedestrian segmentation network and then feeds it to the subsequent recognition network, where the synthetic silhouettes consist of trainable edges of bodies and fixed interiors to limit the information that the recognition network receives. Besides, **GaitAlign** for aligning silhouettes is embedded into the GaitEdge without losing differentiability. Experimental results on CASIA-B and our newly built TTG-200 indicate that GaitEdge significantly outperforms the previous methods and provides a more practical end-to-end paradigm. All the source code are available at <https://github.com/ShiqiYu/OpenGait>.

Keywords: Gait Recognition; End-to-end; Gait Alignment; Cross-domain

1 Introduction

In recent years, human identification by walking pattern, *i.e.* gait, has become an attractive research topic. Compared with other biometrics, *e.g.*, face, fingerprint and iris, human gait can be easily captured at a long distance without the cooperation of subjects, which means gait can be promising for crimes investigation and suspects tracing under real-world uncontrolled conditions. It is noticed that most of the studies treat gait recognition as a two-step approach, including extracting the intermediate modality, *e.g.*, silhouette mask or skeleton keypoints,

* Equal contributions.

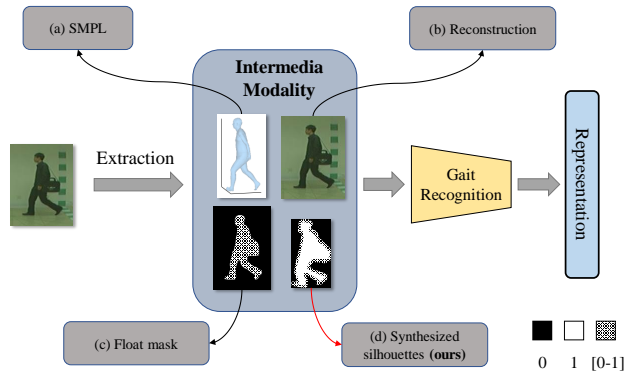


Fig. 1. Three typical end-to-end approaches: (a) model-based end-to-end [14,13], (b) Zhang’s GaitNet [30], (c) Song’s GaitNet [20] and (d) our **GaitEdge**. Shaded areas for the float-point numbers ranging from 0 to 1

from RGB images and putting them into the downstream gait recognition network. However, some researches [4,5,7,16] indicate that those multi-step pipelines usually give rise to the weakness in efficiency and effectiveness; increasing works tend to infer the final results directly in end-to-end [30,20,14].

To the best of our knowledge, there are three typical end-to-end gait recognition methods in the recent literature. As illustrated in Fig. 1 (a), Li *et al.* [14,13] utilize a fashion human mesh recovery model [11] to reconstruct the three-dimensional human body and train the recognition network by taking the parameters of the skinned multi-person linear (SMPL) [17] model as inputs. Another typical approach proposed by Zhang *et al.* [30] introduces an autoencoder framework to disentangle the motion-relevant gait patterns and motion-irrelevant appearance features explicitly from the sequential RGB images, as shown in Fig. 1 (b). In addition, Song *et al.* propose GaitNet [20], which integrates two tasks, *i.e.* pedestrian segmentation and gait recognition, as illustrated in Fig. 1 (c). It extracts gait features straightly from the intermediate float mask instead of the classical binary silhouettes.

To achieve higher performance than those two-step methods, those end-to-end methods cannot ensure that the learned characteristics for human identification only consist of walking patterns. Since the intermediate modalities of previous end-to-end frameworks, *e.g.*, the SMPL reconstruction in [14,13], posture features disentanglement in [30], and pedestrian segmentation supervision in [20], are float-coding, they may introduce some background and texture information. Furthermore, while the previous methods all attempt to exclude gait-irrelevant features, they lack convincing experiments to validate.

To alleviate these issues, we notice that gait features are generally more robust to the variations of camera viewpoints, carrying and clothing than other gait-irrelevant noises, *i.e.* texture and color. If those uncorrelated texture and color features dominate in the extracted gait representations, the recognition

performance will drop obviously when the model is directly exploited in unseen domains (new dataset) [8]. Hence, in this paper, we introduce the **cross-domain** evaluation to expose the side effects of texture and color. More importantly, we propose a concise yet compelling end-to-end framework named **GaitEdge** to deal with this challenging evaluation. As shown in Fig. 1 (d), the intermediate modality of GaitEdge is a novel synthetic silhouette, while its edge is composed of the trainable float mask, and other regions are classical binary silhouettes. The design is inspired by two intuitive phenomena. First, the RGB-informed noises are mainly distributed in the non-edge regions, *e.g.*, the human body and background. Therefore, treating these regions as binary silhouettes can effectively prevent the leakage of gait-irrelevant noises. Second, the edge region plays a vital role in describing the shape of the human body. Hence, making the only edge region trainable can liberate the potential of the end-to-end training strategy. In addition, we observe that the size-normalized alignment [9] is necessary for the silhouette pre-processing to keep the body in aspect ratio. Unfortunately, this operation used to be offline and thus non-differentiable, which means it can not be directly applied to align the synthetic silhouette. To solve this problem, inspired by the RoIAlign [6], we propose **GaitAlign** module to complete the framework of GaitEdge, which can be regarded as a differentiable version of the alignment method proposed by [9].

In summary, the three major contributions are as follows:

- We point out that gait-irrelevant noises may be mixed into gait representations in end-to-end solutions, and introduce a cross-domain testing to verify the leakage of RGB-informed noises. Besides, due to the lack of a gait dataset to provide RGB videos, we collect the *Ten Thousand Gaits* (TTG-200) dataset which has a similar size to the popular CASIA-B [26] dataset.
- We propose GaitEdge, a concise yet compelling end-to-end gait recognition framework. Experiments on both CASIA-B and TTG-200 demonstrate that GaitEdge outperforms state-of-the-arts. GaitEdge can also effectively prevent irrelevant RGB-informed noises.
- We propose a differentiable size-normalization module, GaitAlign, for silhouette-based end-to-end gait recognition. GaitAlign makes the total solution is end-to-end and trainable.

2 Related Work

2.1 Gait Recognition

As a kind of biometrics, gait is defined by early research [23] as the walking pattern that a given person will perform in a fairly repeatable and characteristic way. On the other hand, another similar task, *i.e.* person re-identification (ReID) [31], aims to find a person presented in one camera in another place by another camera. Despite the similarity, they are still fundamentally different: gait recognition focuses on walking patterns, while ReID uses clothing primarily for identification. Therefore, it is worth emphasizing that we can not let the

gait recognition network acquire information other than gait patterns, such as RGB-informed texture and color.

At present, the mainstream visual-based gait recognition methods can be roughly divided into model-based and appearance-based. The former model-based approaches [15,2,14,22] usually extract the underlying structure of the human body first, *e.g.*, 2D or 3D skeleton key-points, and then model the human walking patterns. In general, such methods can better mitigate the effects of clothing and more accurately describe the body’s posture. Nonetheless, all of them are difficult to model the human body structure under the practical surveillance scene due to the low quality of the video.

Appearance-based gait recognition [24,28,4,29,5,7,16] is more popular than model-based gait recognition in the literature. A recent method, GaitSet [4], takes a sequence of silhouettes as input and makes great progress. Subsequently, Fan *et al.* [5] propose a focal convolutional layer to learn the part-level feature and utilize micro-motion capture module to model short-range temporal patterns. Besides, Lin *et al.* [16] propose a 3D CNN-based global and local feature extractor module to extract discriminative global and local representations from frames, which outperforms most other methods remarkably.

2.2 End-to-end Learning

End-to-end learning refers to integrating several separate gradient-based deep learning modules in a differentiable manner. This training paradigm has a natural advantage in that the system optimizes components for overall performance rather than optimizing human-selected intermediates [3].

Recently, some research topics have been benefited from the end-to-end learning paradigm. Amodei *et al.* [1] replace entire pipelines of hand-engineered components with neural networks to overcome the diverse variety of speech by end-to-end learning. Another notable work [3] is Nvidia’s end-to-end training for autonomous driving systems. It only gives the system the human steering angle as the training signal. Still, the system can automatically learn the internal representation of the necessary processing steps, such as detecting lane lines.

As the end-to-end philosophy becomes increasingly popular, several studies [30,14,20] have applied it to gait recognition. Zhang *et al.* [30] propose an autoencoder to disentangle the appearance and gait information without the explicit appearance and gait label for supervision. Li *et al.* [14,13] use the newly developed 3D human mesh model [17] as an intermediate modality and make the silhouettes generated by the neural 3D mesh renderer [12] consistent with the silhouette segmented from RGB images. Because the 3D mesh model provides more helpful information than silhouettes, this approach achieves state-of-the-art results. However, using a 3D mesh model requires high resolution input RGB images which is not feasible in the real surveillance scenario. Different from the previous two, Song *et al.* [20] propose another type of end-to-end gait recognition framework. It is formed by directly connecting the pedestrian segmentation and gait recognition networks, which is supervised by a joint loss, *i.e.* segmentation loss and recognition loss. This approach looks more applicable, but it may make

the gait-irrelevant noises leak into the recognition network due to the absence of explicit restrictions. Differently, our GaitEdge mainly poses and addresses two pivotal problems: cross-domain evaluation and silhouette misalignment.

3 Cross Domain Problem

From the previous perspective, we argue that although the existing end-to-end approaches [30,20,14,13] greatly improve the accuracy, it is natural to suspect that the introduction of RGB information is the cause of the improvement. To verify our conjecture, we introduce two gait recognition paradigms and compare them experimentally.

Firstly, one of the best-performing two-step gait recognition methods, *i.e.* **GaitGL** [16], is adopted as a baseline. In addition, a simple and straightforward end-to-end model named **GaitGL-E2E** that provides a fair comparison is introduced. As shown in Fig. 2 (a) and (b), both methods use the same modules except that GaitGL-E2E replaces binary mask with float-coding silhouettes through a trainable segmentation network, *i.e.* U-Net [18]. Experimentally, we define the **single-domain** evaluation as training and testing on CASIA-B*⁵ [26]. Correspondingly, the **cross-domain** evaluation is defined as training on another dataset, *i.e.* TTG-200, but testing the trained model on CASIA-B*. More implementation details will be elaborated in Section 5.

As shown in the single-domain part of Fig. 2 (d), GaitGL-E2E easily outperforms GaitGL because it has more trainable parameters, and more information is contained in the float-point mask than the binary mask. However, it is inevitable to doubt that float-point numbers flowing into the recognition network bring in texture and color from RGB images, which makes the recognition network learn gait-irrelevant information and leads to the degradation of cross-domain performance. On the other hand, the cross-domain part of Fig. 2 (d) shows that GaitGL-E2E does not achieve the same advantages as it does in single-domain and is even much lower than GaitGL (GaitGL: 40.34%, GaitGL-E2E: 27.18%) in the most challenging case, *i.e.* CL (walking with cloth change). This phenomenon indicates it is easier for end-to-end models to learn easily identifiable coarse-grained RGB information rather than fine-grained imperceptible gait patterns.

The above two experiments demonstrate that GaitGL-E2E does absorb RGB noises so that it is no longer reliable for gait recognition with practical cross-domain requirements. Therefore, we propose a novel framework GaitEdge composed of our carefully designed Gait Synthesis module and differentiable GaitAlign module, as shown in Fig. 2 (c). The most significant difference between GaitEdge and GaitGL-E2E is that we control the transmission of RGB information through manual silhouettes synthesis.

⁵ We reprocess CASIA-B and denote the newly processed one as CASIA-B*.

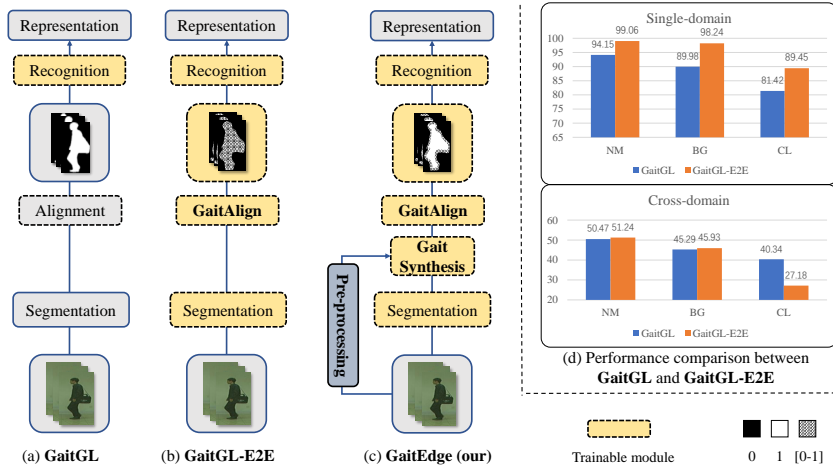


Fig. 2. (a), (b) and (c) are three different frameworks. (d) The rank-1 accuracy (%) on CASIA-B* excluding identical-view cases. *NM* for normal walking, *BG* for walking with bags, and *CL* for walking with cloth change

4 Our Framework

4.1 Gait Synthesis

We generally believe that the edge (the contour of the silhouette) contains the most discriminative information in silhouette images [25]. The interior of a silhouette can be regarded as low-frequency content with less information, whereas the information will be too compact to train the recognition network if we get rid of the interior. Therefore, the designed module, named **Gait Synthesis**, focuses on combining trainable edges with fixed interiors through mask operation. It only trains the edge part of the silhouette image, and the region other than edges are extracted from the frozen segmentation network. As shown in Fig. 3, to clarify how our framework works, we use yellow for the trainable module and illustrate the flow of gradient transfer, in which the dotted orange line represents the backward propagation, and the solid blue line represents the forward propagation. The masks of edge and interior are denoted as M_e and M_i . The output probability of the segmentation network is denoted as P . Then, the output of Gait Synthesis denoted M_s can be obtained by several element-wise operations:

$$M_s = M_e \times P + M_i \quad (1)$$

As shown in Equation 1, we explicitly multiply P_s by M_e and then add it to M_i , which blocks most information, including the gait-relevant and gait-irrelevant. However, we can still fine-tune the edges of the silhouettes, making it automatically optimized for recognition.

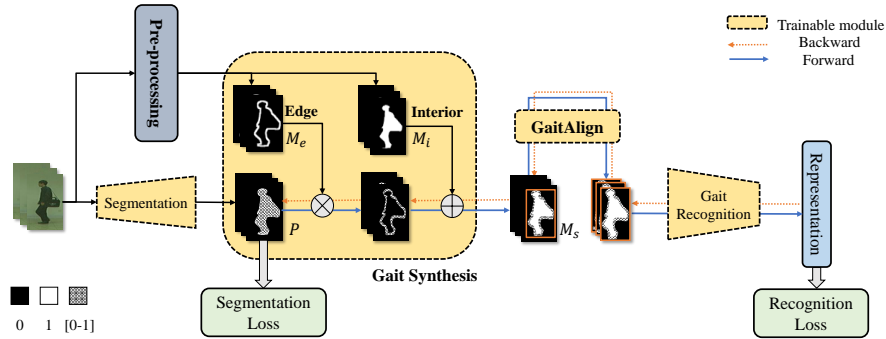


Fig. 3. Illustration of GaitEdge, \oplus for *add element-wise*, \otimes for *multiply element-wise*, and *shaded areas* for float-point numbers ranging from 0 to 1. More details about **Pre-processing** module can be found in Fig. 4.

Pre-processing. We design an untrainable pre-processing operation to get M_e and M_i , shown in Fig. 4. Specifically, we divide it to three steps. First, we segment the input RGB image with the trained segmentation model to obtain the silhouette M . Then, in the second step, we use the classic morphological algorithms to get the dilation and erosion silhouettes (M_i) with a 3×3 flat structuring element. Finally, we get M_e by element-wise exclusive or \vee . Formally:

$$\begin{aligned} M_i &= \text{erode}(M) \\ M_e &= M_i \vee \text{dilate}(M) \end{aligned} \quad (2)$$

Overall, *Gait Synthesis* takes the most intuitive approach by limiting the adjustable region to retain the most valuable silhouette features while eliminate most of low-level RGB-informed noises. It is worth mentioning that, *Gait Synthesis* can be detachably integrated into previous silhouette-based end-to-end methods due to the simplicity of the design.

4.2 Gait Alignment Module

Alignment is very crucial for all silhouette-based gait recognition methods. Since the size-normalization of the silhouette was used for the first time on the OU-ISIR Gait Database [9], almost all silhouette-based methods pre-process the

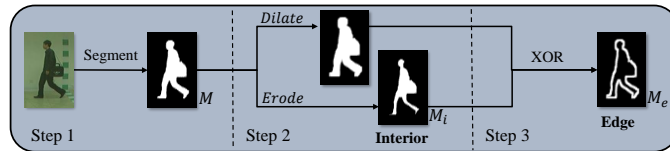


Fig. 4. The **Pre-processing** module in GaitEdge

Algorithm 1 Pseudocode of GaitAlign in a PyTorch-like style.

```

# s_in : silhouettes from segmentation output, (nx1xhxw)
# size : the target size, (H,W)
# r : aspect ratio of human body, (n)
# s_out : aligned silhouettes, (nx1xHxW)

# pad along the x axis so as not to exceed the boundary
s_in = ZeroPad2d((w // 2, w // 2), 0, 0) # (nx1xhx2w)
binary_mask = round(s_in) # binary silhouette

# compute the coordinates and restore the aspect ratio r
left, top, right, bottom = bbox(binary_mask, r, size)

# get the new silhouettes by differentiable roi_align
s_out = roi_align(s_in, (left, top, right, bottom), size)

```

`bbox`: Get the four regularly locations of bounding box keeping aspect ratio. We hide this tedious engineering trick, and the source code will be released.
`roi_align`: Crop and resize the interested region without the loss of spatial alignment.

silhouette input via size-normalization, which removes the noise and benefits the recognition. However, the previous end-to-end approach, *i.e.* GaitNet [20], feeds the segmented silhouette into the recognition network directly, which hardly handles the situation mentioned above. Therefore, we propose a differentiable Gait Alignment Module called *GaitAlign* to make the body be the center of the image and fill the entire image vertically.

We first review the size-normalization [9] procedure because GaitAlign can be regarded as a differentiable version. In size-normalization, by figuring out the top, bottom, and horizontal center of the body, we can scale the body to the target height in aspect ratio and then pad the x-axis with zeros to reach the target width. In our case, pseudo-code in Algorithm 1 depicts the procedure of GaitAlign. We first need to pad the left and right sides with half the width of zeros, which ensures that crop operation will not exceed the boundary. According to the aspect ratio and the target size, then we compute the exact values of four regularly sampled locations. Finally, RoIAlign [6] is applied to the locations given by the previous step. As a result, we get the standard-size, image-filled silhouettes, and its aspect ratio remains the same (refer to the output of GaitAlign in Fig. 3). Another noteworthy point is that the GaitAlign module is still **differentiable**, making our end-to-end training feasible.

5 Experiment

5.1 Settings

Datasets. There are a few available datasets for gait recognition, *e.g.*, CASIA-B [26], OUMVLP [21], Outdoor-Gait [20], FVG [30], GREW [32], and so on. However, not all of them are useful for the end-to-end based gait recognition methods. For example, the proposed work cannot apply the two worldwide most enormous gait datasets, OUMVLP [21] and GREW [32], because neither provides



Fig. 5. Examples of CASIA-B and TTG-200. The left (CASIA-B) consists of six views of one sequence. The right (TTG-200) consists of six subjects with different views

RGB videos. In short, our ideal gait dataset owns several vital attributes: RGB videos available, rich camera viewpoints, and multiple walking conditions.

CASIA-B [26] seems to be a good choice. Nevertheless, there still needs another similar dataset to fit our cross-domain settings. Consequently, we collect a private dataset named *Ten Thousand Gaits 200 (TTG-200)* and show its statistics in Table 1.

CASIA-B. There are 124 subjects walking indoor in CASIA-B. It is probably the most popular dataset that consists of 11 views ($[0^\circ-180^\circ]$) and three walking conditions, *i.e.* normal walking (NM#01-06), walking with bags (BG#01-02), and walking with cloth change (CL#01-02). We strictly follow previous studies, which group the first 74 subjects into the training set and others into the test set. Furthermore, for the test stage, the first 4 sequences (NM#01-04) are regarded as the gallery set, while the left 6 sequences are grouped into 3 probe subsets, *i.e.* NM#05-06, BG#01-02, CL#01-02. Besides, since the silhouettes of CASIA-B were obtained by the outdated background subtraction, there exists much noise caused by the background and clothes of subjects. Hence, we re-annotate the silhouettes of CASIA-B and denote it as *CASIA-B**. All our experiments are conducted on this newly annotated one.

TTG-200. This dataset contains 200 subjects walking in the wild, and each subject is required to walk under 6 various conditions, *i.e.* carrying, clothing, taking on the phone, and so on. For each walking process, the subject will be captured by 12 cameras located around the different viewpoints (unlabelled),

Table 1. The statistics of existing gait datasets and our collected TTG-200

Dataset	Subjects	Environment	Format	Variations	#Sequences
CASIA-B	124	Indoor	RGB	11 views, carrying, clothing	13,636
OUMVLP	10,307	Indoor	Silhouette	14 views	267,388
FVG	226	Outdoor	RGB	3 frontal views, walking speed, carrying, clothing, background	2,856
Outdoor-Gait	138	Outdoor	RGB	carrying, clothing	4,964
GREW	26,345	Outdoor	Silhouette	multiple camera	128,671
TTG-200 (our)	200	Outdoor	RGB	12 views, carrying, clothing, talking on the phone, background	14,198

which means each subject ideally owns $6 \times 12 = 72$ gait sequences. In the following experiments, we take the first 120 subjects for training and the last 80 subjects for the test. In addition, the first sequence (#1) is regarded as gallery set, and the left 5 sequences (#2-6) are regarded as probe set.

As shown in Fig. 5, compared with CASIA-B, TTG-200 has three main differences: (1) The backgrounds of TTG-200 are more complex and diverse (collected in multiple different outdoor scenes); (2) The data of TTG-200 are mostly aerial view, while data of CASIA-B are mostly horizontal view; (3) TTG-200 has better image quality. Therefore, we can treat these two datasets as different domains.

Implementation Details

Data Pre-processing. We first employ ByteTrack [27] to detect and track pedestrians from the raw RGB videos for both CASIA-B [26] and TTG-200, and then conduct the human segmentation and silhouette alignment [9] to extract the gait sequences. The obtained silhouettes are resized to 64×44 and can be taken as the input of these two-stage gait recognition methods or be the ground-truth for the pedestrian segmentation network in these end-to-end based approaches.

Pedestrian Segmentation. We use the popular U-Net [18] as our segmentation network that is supervised by Binary Cross-Entropy [10] loss L_{seg} . We set the input size as $128 \times 128 \times 3$ and the channels of U-Net as $\{3, 16, 32, 64, 128, 64, 32, 16, 1\}$ and train it via SGD [19] (batch size=960, momentum=0.9, initial learning rate=0.1, weight decay= 5×10^{-4}). For each dataset, we train the network with learning rate scaled to 1/10 two times for every 10000 iterations until convergence.

Gait Recognition. We use the latest GaitGL [16] as our recognition network and strictly follow the original paper’s settings.

Joint Training Details. In this step, the training data sampler and batch size are similar to the gait recognition network. We jointly fine-tune the segmentation and recognition networks with the joint loss $L_{joint} = \lambda L_{seg} + L_{rec}$, where L_{rec} denotes the loss of recognition network. The λ represents the loss weight of segmentation network and is set to 10. Besides, to make the joint training process converge faster, we use the trained segmentation and recognition networks parameters to initialize the end-to-end model, and accordingly, their initial learning rate is set to 10^{-5} and 10^{-4} , respectively. Moreover, we fix the first half of the segmentation network, *i.e.* U-Net, to keep the segmentation result in human shape. We jointly train the end-to-end network for a total of 20,000 iterations and reduce the learning rate by 1/10 at the 10,000th iteration.

5.2 Performance Comparison

To demonstrate the reliable cross-domain capability of GaitEdge, we conduct the single-domain and cross-domain evaluations on CASIA-B* and TTG-200, as shown in Table 2.

Table 2. The rank-1 accuracy (%) on CASIA-B* and TTG-200. The identical-view cases in CASIA-B* are excluded. The **bold** and **(bold)** numbers for the two highest accuracies of single-domain and that of cross-domain, respectively

Training Set	Method		Test Set				
			CASIA-B*				TTG-200
			NM	BG	CL	Mean	
CASIA-B*	Two-step	GaitSet [4]	92.30	86.10	73.36	83.92	40.26
		GaitPart [5]	93.14	85.99	75.05	84.72	42.23
		GaitGL [16]	94.15	89.98	81.42	88.52	(48.74)
	End2end	GaitGL-E2E	99.06	98.24	89.45	95.58	37.18
		GaitEdge	97.94	96.06	86.36	93.45	(49.12)
TTG-200	Two-step	GaitSet [4]	41.32	35.15	21.59	32.69	77.62
		GaitPart [5]	45.21	38.75	25.92	36.62	80.24
		GaitGL [16]	50.47	45.29	40.34	(45.37)	80.46
	End2end	GaitGL-E2E	51.24	45.93	27.18	41.45	90.37
		GaitEdge	54.76	49.85	38.16	(47.59)	88.66

The diagonal of Table 2 shows the single-domain performance comparisons, where these methods are trained and evaluated in the identical dataset. On the opposite, the anti-diagonal shows the cross-domain performance comparisons, where these methods are trained and evaluated in the different datasets.

Single-domain Evaluation. From the diagonal results in Table 2, we observe that the performance of traditional two-step gait recognition methods is far inferior to that of two end-to-end ones. For example, GaitGL-E2E exceeds GaitSet [4] by 11.66% for CASIA-B* and 12.75% for TTG-200, respectively. On the other hand, the accuracy of our proposed GaitEdge is slightly lower than that of GaitGL-E2E, *i.e.* -2.13% for CASIA-B* and -1.71% for TTG-200. However, we argue that GaitGL-E2E owns the higher risk of overfitting in the gait-irrelevant noises since it directly takes the float mask generated by the segmentation network as the input of the recognition network. Hence, we further conduct the cross-domain evaluation to support this notion experimentally.

Cross-domain Evaluation. If some irrelevant noises dominate the gait representations used for human identification, *i.e.* texture and color, the recognition accuracy would drop dramatically in the case of cross-domain settings since the extracted features impotently represent the relatively robust gait patterns. The anti-diagonal results in Table 2 show that all these methods have significant performance degradation compared to single-domain due to the significant difference between CASIA-B* and TTG-200. We notice that although GaitGL-E2E has the highest accuracy in single-domain, it achieves the poorest performance for crossing the domain from CASIA-B* to TTG-200. In contrast, our GaitEdge reaches the best performance than any other posted method in cross-domain evaluations, although it is about 2% lower than GaitGL-E2E in single domain.

Table 3. The rank-1 accuracy (%) on CASIA-B* across different views excluding the identical-view cases. For evaluation, the first 4 sequences (NM#01-04) are regarded as the gallery set, while the left 6 sequences are grouped into 3 probe subsets, *i.e.* NM#05-06, BG#01-02, CL#01-02. The original paper of Song GaitNet [20] does not mention the results of BG and CL

Probe	Method	Probe View										Mean	
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°		180°
NM	Song GaitNet [20]	75.6	91.3	91.2	92.9	92.5	91.0	91.8	93.8	92.9	94.1	81.9	89.9
	Zhang GaitNet [30]	93.1	92.6	90.8	92.4	87.6	95.1	94.2	95.8	92.6	90.4	90.2	92.3
	ModelGait [14]	96.9	97.1	98.5	98.4	97.7	98.2	97.6	97.6	98.0	98.4	98.6	97.9
	MvModelGait [13]	97.5	97.6	98.6	98.8	97.7	98.9	98.9	97.3	97.6	97.8	97.9	98.1
	GaitEdge	97.2	99.1	99.2	98.3	97.3	95.5	97.1	99.4	99.3	98.5	96.4	97.9
BG	Zhang GaitNet [30]	88.8	88.7	88.7	94.3	85.4	92.7	91.1	92.6	84.9	84.4	86.7	88.9
	ModelGait [14]	94.8	92.9	93.8	94.5	93.1	92.6	94.0	94.5	89.7	93.6	90.4	93.1
	MvModelGait [13]	93.9	92.5	92.9	94.1	93.4	93.4	95.0	94.7	92.9	93.1	92.1	93.4
	GaitEdge	95.3	97.4	98.4	97.6	94.3	90.6	93.1	97.8	99.1	98.0	95.0	96.1
	Zhang GaitNet [30]	50.1	60.7	72.4	72.1	74.6	78.4	70.3	68.2	53.5	44.1	40.8	62.3
CL	ModelGait [14]	78.2	81.0	82.1	82.8	80.3	76.9	75.5	77.4	72.3	73.5	74.2	77.6
	MvModelGait [13]	77.0	80.0	83.5	86.1	84.5	84.9	80.6	80.4	77.4	76.6	76.9	80.7
	GaitEdge	84.3	92.8	94.3	92.2	84.6	83.0	83.0	87.5	87.4	85.9	75.0	86.4

Hence, this cross-domain evaluation not only indicates the robustness of GaitEdge is far superior to that of GaitGL-E2E but also claims the GaitEdge is a practical and advanced framework for the end-to-end gait recognition task.

Comparison with other end-to-end methods. Last but not least, the proposed GaitEdge is compared to three previous end-to-end gait recognition methods across different views on CASIA-B*. Table 3 shows that GaitEdge reaches almost the highest accuracy on various walking conditions, especially for CL (+5.7% than MvModelGait), which reveals that GaitEdge is remarkably robust to color and texture (cloth change).

Table 4. The ablation study for the size of structuring element. The larger size for the larger edge region. The **bold** and **(bold)** numbers for the highest accuracy of single-domain and that of cross-domain, respectively

Training Set	Method	Structuring Element	Test Set				
			CASIA-B*				TTG-200
			NM	BG	CL	Mean	-
CASIA-B*	GaitEdge	3 × 3	97.94	96.06	86.36	93.45	49.12
		5 × 5	98.88	97.36	88.24	94.83	(50.98)
		7 × 7	98.97	97.90	88.36	95.08	49.15
		9 × 9	99.02	98.19	89.05	95.42	44.47
	GaitGL-E2E	-	99.06	98.24	89.45	95.58	37.18
TTG-200	GaitEdge	3 × 3	54.76	49.85	38.16	(47.59)	88.66
		5 × 5	49.21	45.22	33.71	42.71	89.62
		7 × 7	50.26	44.20	32.43	42.29	90.00
		9 × 9	48.84	41.72	27.66	39.41	90.39
	GaitGL-E2E	-	51.24	45.93	27.18	41.45	90.37



Fig. 6. (a) The original images (top) *vs.* the disturbed images (bottom). We make random pixel offset to the disturbed images, including vertical and horizontal direction. (b) The ablation study for the *GaitAlign* module. The results are reported on CASIA-B* after disturbance

5.3 Ablation Study

Impact of Edge. Table 4 shows the impact of body edge size. We extract the edges by several sizes of structuring elements—the larger the structuring element, the larger the edge area. According to the results shown in Table 4, as the size of the structuring element increases, the performance of single-domain accordingly increases, but the performance of cross-domain almost decreases at the same time. This result claims that the area of the float mask occupying the intermediate synthetic silhouette is negatively associated with the cross-domain performance for GaitEdge. Therefore, we can argue that the reason why GaitGLE2E fails in cross-domain evaluation is that it is equivalent to GaitEdge in the case of the infinite structural element. Furthermore, those non-edge regions of silhouette, *i.e.* human body and background, are unsuitable in float-coding for the end-to-end gait recognition framework.

Impact of GaitAlign. Notably, we observe that the result of pedestrian detection (upstream task) in natural scenes, is often much worse than that of the controlled environment, *i.e.* CASIA-B* and TTG-200. In order to simulate this complex situation, we first apply object detection on the videos of CASIA-B* and then perform random pixel offset along with vertical and horizontal coordinates with a probability of 0.5. As shown in Fig. 6 (a), the bottom images are disturbed, aiming to simulate the natural situation. Fig. 6 (b) shows that alignment improves the average accuracy significantly. In addition, we also notice that the accuracy of normal walking (NM) drops a little, *i.e.* -0.38%. However, we believe this is because the accuracy of NM is approaching the upper limit.

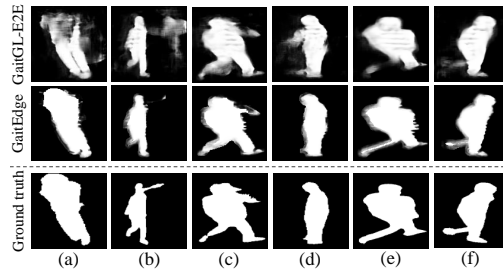


Fig. 7. The comparison between the intermediate results and ground truth, *i.e.* the first two rows *v.s.* the third row

5.4 Visualization

To better understand the performance degradation of GaitGL-E2E and the effectiveness of GaitEdge, we illustrate the intermediate results generated by GaitGL-E2E and GaitEdge respectively as well as the ground truth corresponding to the same frame, as shown in Fig. 7. Specifically, for GaitGL-E2E, the intermediate results in (a), (b), (c) and (d) capture more background and texture information, and some body parts are eliminated such as legs in (e) and (f). While for GaitEdge, the intermediate results are much more stable and reasonable making it more robust.

6 Conclusion

This paper presents a novel end-to-end gait recognition framework termed GaitEdge that can solve the performance degradation in cross-domain situation. Specifically, we design a Gait Synthesis module to mask the fixed body with tunable edges obtained by morphological operation. Besides, a differentiable alignment module named GaitAlign is proposed to solve the body position jitter caused by the upstream pedestrian detection task. We also conduct extensive and comprehensive experiments on two datasets, including CASIA-B* and our newly built TTG-200. Experimental results show that GaitEdge significantly outperforms the previous methods, indicating that GaitEdge is a more practical end-to-end paradigm that can effectively block RGB noise. Moreover, this work exposes the cross-domain problem neglected by previous studies, which provides a new perspective for future research.

Acknowledgments We would like to thank the helpful discussion with Dr. Chunshui Cao and Dr. Xu Liu. This work was supported in part by the National Natural Science Foundation of China under Grant 61976144, in part by the Stable Support Plan Program of Shenzhen Natural Science Fund under Grant 20200925155017002, in part by the National Key Research and Development Program of China under Grant 2020AAA0140002, and in part by the Shenzhen Technology Plan Program (Grant No. KQTD20170331093217368).

References

1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International conference on machine learning. pp. 173–182. PMLR (2016)
2. An, W., Yu, S., Makihara, Y., Wu, X., Xu, C., Yu, Y., Liao, R., Yagi, Y.: Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Transactions on Biometrics, Behavior, and Identity Science* **2**(4), 421–430 (2020)
3. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
4. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence. pp. 8126–8133 (2019)
5. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14225–14233 (2020)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
7. Hou, S., Cao, C., Liu, X., Huang, Y.: Gait lateral network: Learning discriminative and compact representations for gait recognition. In: European Conference on Computer Vision. pp. 382–398. Springer (2020)
8. Huang, H., Yang, W., Chen, X., Zhao, X., Huang, K., Lin, J., Huang, G., Du, D.: Eanet: Enhancing alignment for cross-domain person re-identification. arXiv preprint arXiv:1812.11369 (2018)
9. Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.: The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security* **7**(5), 1511–1521 (2012)
10. Jadon, S.: A survey of loss functions for semantic segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). pp. 1–7. IEEE (2020)
11. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018)
12. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3907–3916 (2018)
13. Li, X., Makihara, Y., Xu, C., Yagi, Y.: End-to-end model-based gait recognition using synchronized multi-view pose constraint. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4106–4115 (2021)
14. Li, X., Makihara, Y., Xu, C., Yagi, Y., Yu, S., Ren, M.: End-to-end model-based gait recognition. In: Proceedings of the Asian conference on computer vision (2020)
15. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition* **98**, 107069 (2020)
16. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14648–14656 (2021)

17. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* **34**(6), 1–16 (2015)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)
19. Ruder, S.: An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* (2016)
20. Song, C., Huang, Y., Huang, Y., Jia, N., Wang, L.: Gaitnet: An end-to-end network for gait based human identification. *Pattern recognition* **96**, 106988 (2019)
21. 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 Transactions on Computer Vision and Applications* **10**(1), 1–14 (2018)
22. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Gaitgraph: graph convolutional network for skeleton-based gait recognition. In: *2021 IEEE International Conference on Image Processing (ICIP)*. pp. 2314–2318. IEEE (2021)
23. Winter, D.A.: *Biomechanics and motor control of human gait: normal, elderly and pathological* (1991)
24. Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T.: A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE transactions on pattern analysis and machine intelligence* **39**(2), 209–226 (2016)
25. Yu, S., Tan, D., Huang, K., Tan, T.: Reducing the effect of noise on human contour in gait recognition. In: *International Conference on Biometrics*. pp. 338–346. Springer (2007)
26. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: *18th International Conference on Pattern Recognition (ICPR'06)*. vol. 4, pp. 441–444. IEEE (2006)
27. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864* (2021)
28. Zhang, Y., Huang, Y., Yu, S., Wang, L.: Cross-view gait recognition by discriminative feature learning. *IEEE Transactions on Image Processing* **29**, 1001–1015 (2019)
29. Zhang, Z., Tran, L., Liu, F., Liu, X.: On learning disentangled representations for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020)
30. Zhang, Z., Tran, L., Yin, X., Atoum, Y., Liu, X., Wan, J., Wang, N.: Gait recognition via disentangled representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4710–4719 (2019)
31. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984* (2016)
32. Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., Zhou, J.: Gait recognition in the wild: A benchmark. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 14789–14799 (2021)