IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 June 2024, accepted 8 August 2024, date of publication 19 August 2024, date of current version 29 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445415

METHODS

RGB-Based Gait Recognition With Disentangled Gait Feature Swapping

KOKI YOSHINO[1], (Graduate Student Member, IEEE), KAZUTO NAKASHIMA[2], (Member, IEEE), JEONGHO AHN[1], (Graduate Student Member, IEEE), YUMI IWASHITA[3], (Senior Member, IEEE), AND RYO KURAZUME[2], (Senior Member, IEEE)

[1]Graduate School of Information Science and Electrical Engineering, Kyushu University, Nishi-ku, Fukuoka 819-0395, Japan
[2]Faculty of Information Science and Electrical Engineering, Kyushu University, Nishi-Ku, Fukuoka, Fukuoka 819-0395, Japan
[3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA

Corresponding author: Koki Yoshino (yoshino@irvs.ait.kyushu-u.ac.jp)

This work was supported in part by JST, the establishment of university fellowships toward the creation of science technology innovation, under Grant JPMJFS2132; and in part by JSPS KAKENHI under Grant JP20H00230.

ABSTRACT Gait recognition enables the non-contact identification of individuals from a distance based on their walking patterns and body shapes. For vision-based gait recognition, covariates (e.g., clothing, baggage and background) can negatively impact identification. As a result, many existing studies extract gait features from silhouettes or skeletal information obtained through preprocessing, rather than directly from RGB image sequences. In contrast to preprocessing which relies on the fitting accuracy of models trained on different tasks, disentangled representation learning (DRL) is drawing attention as a method for directly extracting gait features from RGB image sequences. However, DRL learns to extract features of the target attribute from the differences among multiple inputs with various attributes, which means its separation performance depends on the variation and amount of the training data. In this study, aiming to enhance the variation and quantity of each subject's videos, we propose a novel data augmentation pipeline by feature swapping for RGB-based gait recognition. To expand the variety of training data, features of posture and covariates separated through DRL are paired with features extracted from different individuals, which enables the generation of images of subjects with new attributes. Dynamic gait features are extracted through temporal modeling from pose features of each frame, not only from real images but also from generated ones. The experiments demonstrate that the proposed pipeline increases both the quality of generated images and the identification accuracy. The proposed method also outperforms the RGB-based state-of-the-art method in most settings.

INDEX TERMS Biometrics, computer vision, convolutional neural networks (CNNs), disentangled representation learning (DRL), gait recognition, generative adversarial networks (GANs).

I. INTRODUCTION

Gait, i.e., walking style and body shape, can provide clues for identifying individuals. Gait is particularly useful for identifying uncooperative subjects because it can be obtained from a distance and is generally extremely difficult to disguise intentionally. The difficulty of falsification and the long-distance acquisition availability are expected to be utilized for seamless personal authentication in criminal investigation [1], [2] and forensics [3], surveillance systems, access control, and so on. There have been cases reported in which gait recognition has contributed to the arrest of criminals [1], [3].

If the target is a person with a remarkably distinctive gait, it may be possible to visually identify him or her from a distance, but for pedestrians who cannot be discriminated by

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

the human eye, an approach using a deep neural network is effective. Deep learning methods include several modalities, such as 3D point clouds [4], [5], [6], sound of footsteps [7], [8], and foot pressure [9], [10], as well as images and videos. In this study, we adopt a camera-based method, which is the most realistic and popular problem set because it can make use of surveillance cameras. Person identification using camera-based gait recognition is performed in the following flow. (i) acquisition of gait video from a camera, (ii) extraction of the gait feature by a feature extractor such as a neural network, and (iii) matching with the gait features in a database using the gait feature from the input video as a query.

In general, deep learning requires feature extractors to be pre-trained on a training dataset that is prepared separately from the evaluation dataset. In the past, it was common to use a problem setting where the evaluation dataset consisted of different videos with the same subjects as the training dataset (closed-set recognition problem) [11]. In contrast, the current mainstream is a problem set that evaluates a subject different from the training subject (open-set recognition problem) [12], [13]. The unique problem setting of the open-set recognition problem requires high generalization performance for the extractor of gait features. Besides gait information, however, gait videos contain a large amount of gait-independent information (covariates) such as clothing, background, and personal belongings, which hinder the generalization performance improvement. Two major approaches to covariates have been proposed: preprocessing removal [14], [15], [16] and end-to-end removal [17], [18]. Preprocessing methods mainly include silhouette extraction, posture estimation, and shape estimation, and by using pre-trained models for each task, covariates can be removed without supervised labels corresponding to the gait video. However, the lack of information useful for identification [17] and fitting accuracy problems [18] associated with preprocessing motivate the adoption of end-to-end covariate removal.

End-to-end covariate removal methods for camera-based gait recognition optimize covariate removal and gait-based identification simultaneously using RGB image sequences as input. The methods do not require a separate preprocessing process so they are more efficient than pre-processing-based methods. Another potential advantage of end-to-end methods is that they prevent missing information in the extraction of intermediate modalities through preprocessing. However, most end-to-end covariate elimination methods end up extracting intermediate modalities such as silhouette or skeletal information, and then performing gait feature extraction. As a result, they do not fully exploit the rich information of RGB image sequences [13]. To the best of our knowledge, only the method proposed by Zhang et al. [17] and our previous work [19] employ direct gait feature extraction from RGB images.

Zhang et al. [17] were the first to apply disentangled representation learning (DRL) to gait recognition, proposing an autoencoder-based network GaitNet that separates RGB gait image into appearance and pose features. DRL in GaitNet enables the separation and removal of appearance features, which are covariates, from the difference of two attributes of two input gait videos of the same person with different clothing and camera angles, without corresponding supervised labels. This means that the variety and amount of the training data is critical to the separation performance. The most representative dataset for gait recognition that includes RGB data is CASIA-B [20], published in 2006, and since then no dataset has been published that includes RGB videos considering clothing variations until Zhang et al. published FVG [17] in 2019. We consider that the lack of large-scale gait video datasets that include the RGB modality has been caused by issues specific to gait video datasets such as time, data volume, and privacy, and that the lack of datasets has inhibited the adoption of RGB as a source for extracting gait features. Although a few large-scale RGB gait datasets [6], [18] have been released in the past few years, including publicly unavailable ones [18], it is still difficult to fundamentally solve the problems of time, data volume, and privacy. Some datasets partially address the privacy issue by blurring the subject's face [6], [21], but most of the others discard the acquired RGB images due to privacy concerns [22], [23], [24]. Simulation data, which is a promising alternative, does not have parameters that are aware of the way people walk, and it is still very challenging to construct a dataset that includes a variety of walking styles [25]. Consequently, the difficulty of creating a large dataset specific to gait recognition may limit the possibility of direct gait feature extraction from RGB.

To unleash the potential of gait RGB images, we proposed an approach using identity-aware data augmentation in our previous work [19]. In the network we proposed in the previous study, pose feature separated by DRL from input gait image is exchanged with different image of the same person to generate a novel virtual gait image, thereby increasing the variation of the data to be trained. This method can augment the training data in proportion to the number of gait videos per person. Experiments show that data augmentation using feature swaps between gait images of the same person improves the quality of the generated images and accuracy of identification.

TABLE 1. The result of preliminary experiments. FID is a metric to evaluate the quality of image generation; the smaller the value, the higher the quality. The baseline is our reimplementation of GaitNet [17] as well as previous work [19].

Method	Mean accuracy [%] ↑			FID ↓
	NM	BG	CL	
Baseline	90.2	86.1	24.3	169.8
Previous work [19]	**94.4**	**88.8**	29.2	**118.7**
Previous work [19] w/ Cross-id	94.0	85.9	**30.0**	155.3
Ours	**95.9**	**91.1**	**30.0**	**65.2**

The above results encouraged us to consider the possibility of further improving the identification accuracy if we could train on a larger amount and variety of data. We therefore tried to extend the variation of feature exchange in data augmentation from the same person's gait videos (Self-id) [19] to different people's gait videos (Cross-id), using the pipeline of our previous method without any other modification. The results of this preliminary experiment are shown in Table 1.

Table 1 shows that simply extending the previous method's target of feature exchange from the same person to a different person results in a smaller improvement from the baseline. In addition, FID, an evaluation metric for the quality of the generated images, has degraded to a value close to the baseline value by extending the feature exchange targets. This degradation of generated image quality is possible with no additional measures, since the number of variations in feature swapping has been increased. It is also natural that data augmentation with low-quality generated images would have a negative impact on identification accuracy. Based on the results of these preliminary experiments, we hypothesized that this lack of identification accuracy was due to the degraded quality of the generated images used for data augmentation. However, since the possibility of data augmentation based on feature swapping has been demonstrated [19], it is expected that this increase in feature swapping variation could lead to further improvements in identification accuracy if the quality of the generated images, which has become quite complex, can be improved.

In this paper, we propose a novel gait recognition method that separates gait-dependent/independent features from RGB gait videos by DRL and augments the data by exchanging the separated features among different persons. The proposed method can increase the variation and quantity of training data according to the number of people in the dataset, the type of gait setting, and the number of videos per setting, which is expected to improve the performance of feature separation and generalization of the feature extractor. To be specific, in training, input RGB gait video is first separated into gait-dependent features (pose features) and appearance-related covariate features (style features) at each frame. The gait video of a different person is also separated in the same way, and by exchanging the pose features between these two gait videos, a novel image of a pedestrian is generated and added to the training data. The proposed method can increase the number of training data to the square of the total number of gait videos in the dataset. Recognition is based on the gait features extracted from the time series of pose features, and both discriminative and generative learning are optimized at the same time. During inference, dynamic gait features are extracted directly from the pose features extracted from the RGB gait video without any data augmentation, and are matched with the data in the database.

In the experiments, we quantitatively and qualitatively verify the quality of the generated images, and then evaluate the identification accuracy. In addition to comparing the proposed method with the RGB-based method, which has the same problem statement as the proposed method, the experiments also examine the change in performance when each component of the proposed method is ablated.

Our contributions can be summarized as follows.

- We present a novel framework that combines disentangled representation learning (DRL) and data augmentation with feature swapping to learn gait features from a large amount of natural training data without acquiring new data.
- Our proposed method generates a novel gait image by swapping gait features directly separated from an RGB image by DRL with those of a different person, and the generated image can be used for online augmentation of the training data. The proposed method can increase the quantity and variety of the training dataset quadratically.
- Experimental results demonstrate that the proposed method improves the quality of the generated images and the identification accuracy, and outperforms the state-of-the-art RGB-based method in most metrics.

As a preliminary step of this study, we have presented a gait recognition method that utilizes data augmentation through feature exchange between gait videos of the same person. This study extends the previous work in the following ways: (i) The amount and diversity of training data are further augmented by enhancing the feature exchange from gait videos of the same person with different gait conditions to all gait videos of different persons. (ii) We update the learning pipeline by a new process of reconstructing the original image using re-encoded features from the generated image to improve the quality of novel generation of the more complex gait images.

II. RELATED WORK

In this section, we present related work limited to camera-based methods, which belong to the same category as this study. We classify gait recognition methods into three categories according to the modality for feature extraction: silhouette-based, model-based, and RGB-based. Note that, in this paper, methods that use RGB as input but extract gait features in a different modality are considered to be methods of that modality. For example, a method that converts input RGB to silhouette in the process is introduced as a silhouette-based method, even if silhouette estimation is carried out in end-to-end manner. The strength of RGB-based methods is that they can take advantage of the rich raw information that only RGB images possess. Therefore, end-to-end preprocessing only improves the accuracy of preprocessing by fine-tuning, but does not fully utilize the information of RGB. After giving an overview of each category, we dive deeper into the studies that are highly relevant to this study from the following three perspectives: GANs, DRL, and end-to-end.

A. SILHOUETTE-BASED GAIT RECOGNITION

Silhouette-based methods use silhouette image sequences as input to neural networks, which are obtained by applying background subtraction and segmentation to a gait video. Removing all but the contour information of the person, silhouettes have consistently been the most popular modality from the past to the present because they are robust and efficient with respect to covariates such as background and clothing. Silhouette-based methods can be divided into two main categories in terms of the actual gait representation input to the network: template-based methods and sequence-based methods.

Gait templates are a lower-dimensional representation of silhouette image sequences, and were the mainstream at the time due to their compactness and ability to be handled with small computational resources prior to the era of deep learning. The most representative gait template is Gait Energy Image (GEI) [14], which is a time-averaged sequence of gait silhouettes. Since GEI has the advantage of reducing not only covariates in gait recognition but also errors in silhouette extraction by time averaging, many methods have been proposed using GEI as input [26], [27], [28], [29]. In addition to GEI, there has also been a lot of exploration of optimal gait templates for gait recognition [30], [31], [32], [33].

As computational resources have become richer, means of temporal modeling for silhouette image sequences have been explored that are richer than templates of low dimensionality. Several methods have been proposed to simultaneously extract spatio-temporal features using 3DCNNs [23], [34], [35], [36]. However, due to the efficiency of training, the combination of spatial feature extraction using 2DCNN and temporal feature extraction using RNN and attention mechanism is the mainstream. In particular, methods that model both global features of the entire body and local features of individual parts have been intensively explored [15], [18], [37], [38], [39], [40], [41]. GaitBase [41] is a baseline model robust both indoors and outdoors, developed through in-depth ablation studies of previous state-of-the-art silhouette-based methods [15], [37], [38], [40]. Silhouette-based methods remain the most dominant category in gait recognition, with a number of new approaches being investigated, such as counterfactual intervention learning utilizing causal inference [42] and dynamic aggregation in both spatial and temporal information [43].

Generative adversarial networks (GANs), an adversarial training method for generators and discriminators, were first introduced in gait recognition by Yu et al. [44]. Since GANs have the capability of generating high-quality images without any supervised data, most silhouette-based gait recognition requires a change in the camera angle of the input silhouette image [29], [33], [44], [45], [46], [47] and removal of belongings [29] have been used to manipulate covariates. Only GaitEditor [48] utilizes GAN inversion techniques [49] to perform unsupervised manipulation of multiple covariates, including not only viewing angle but also belongings and age, etc. Compared to other modalities such as RGB, silhouettes are less informative, and thus the potentially independent attributes that can be separated by DRL are also limited. Consequently, silhouette-based DRL methods are working on pioneering supervision schemes (e.g., semi-supervision [29], [50] and group-supervision [51]). End-to-end methods for identification from RGB gait videos via silhouette have become a trend in recent years. The end-to-end framework improves the identification accuracy by fine-tuning the silhouette extraction module to be optimized for gait recognition [18], [52]. In addition, taking advantage of the raw data acquired from the camera, MMGaitFormer [53] extracts skeletons in addition to silhouettes in an end-to-end manner, thus achieving complementary feature modeling.

In comparison to the aforementioned silhouette-based methods, the proposed method is fundamentally different in that it extracts features from RGB, a modality that contains much more information than silhouettes, without any intermediate representation. We also utilize image generation in the GAN framework for data augmentation, taking into account multiple covariates such as clothing and viewpoint.

B. MODEL-BASED GAIT RECOGNITION

Model-based methods utilize deformable human body models fitted to input gait images as a means of extracting gait features. Gait representation through human body modeling has the advantage of robustness to changes in appearance due to covariates such as background and belongings because it utilizes information about the human body estimated from RGB images. Although model-based methods have been suffering from a bottleneck in the accuracy of fitting physical models, they have attracted renewed attention in recent years with the development of model fitting methods based on deep learning. There are two types of human model representations used in gait recognition: skeletons and 3D human meshes.

Skeleton-based methods extract gait features by temporal modeling of posture information obtained by applying posture estimation methods [54], [55], [56], [57], [58], [59] to gait images. A number of model-based studies have adopted this approach because the extraction of the skeleton captures only the skeletal movements, one of the most important elements of gait [16], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69]. However, not only walking style but also body shape information is important for gait recognition [21], and skeleton-based methods that explicitly remove all information other than posture information, including body shape, have become difficult to improve performance further. To address the problem of missing information on body shape, multimodal methods and alternative human body models [70], [71] that take body shape into account have been proposed. Multimodal methods [53], [72], [73], [74] use

both skeleton and silhouette to compensate for the lacking information on each modality.

3D human mesh models [70], [71] have also been attracting attention in recent years as another type of human body model. Skinned Multi-Person Linear model (SMPL) [70] is one of the most well-known 3D mesh models of the human body, representing the human body with two types of vectors: shape and posture. Gait recognition uses the posture and human shape parameters as inputs to the network [75], [76], [77], [78], which are obtained through a SMPL estimation model via RGB images [70], [71]. These methods are promising alternatives to skeleton-based methods because they can acquire gait feature representations that reflect human body geometry, which is lacking in skeletons.

Due to the dependence on model fitting accuracy, attention to model-based methods has been inferior to that of silhouette-based methods, limiting the variation of the methods. While the application of GANs in model-based methods has not advanced, Yoo et al. [79] have proposed a method robust to viewpoint changes by utilizing DRLs, using as input a modified pedestrian silhouette based on the estimated posture. Employing SMPL has made the information available richer, but at the same time, estimation accuracy has again become an issue, so Li et al. [75], [76] fine-tune SMPL estimation in an end-to-end manner.

Similar to silhouette-based methods, model-based methods differ from our method in that gait features are extracted directly from RGB images. To mitigate the complexity of the RGB modality compared to other modalities, we combine GAN and DRL to improve the accuracy of covariate removal.

C. RGB-BASED GAIT RECOGNITION

Gait recognition has often been classified along two axes, appearance-based and model-based [12], [13], [80], [81], but appearance-based can be further divided into silhouette-based and RGB-based. By using RGB as input, it is possible to avoid missing information in contours due to silhouette extraction and dependence on the accuracy of model extraction. RGB is the most potential modality because it contains more raw information than silhouettes or human models, and the number of methods that use RGB as input is increasing [17], [18], [21], [52], [64], [75], [76]. However, most of them actually perform gait feature extraction via silhouettes [18], [52] or human models [64], [75], [76] as intermediate representations.

In other words, it cannot be said that the rich information in RGB is fully utilized for discriminative learning, since the pre-processing process of silhouette and human model extraction is only fine-tuning in an end-to-end manner. Therefore, we believe that RGB-based gait recognition is a method that directly extracts gait-dependent information from RGB. To the best of our knowledge, GaitNet [17], [21] and our previous study [19] are the only RGB-based methods. GaitNet is an autoencoder network that utilizes DRL to directly separate gait-dependent and gait-independent features from RGB. Our previous work [19] augments the training data with images generated by exchanging posture and style features separated by DRL between images of the same person with different covariates.

GaitNet discards the separated features that are not related to the gait, whereas our method utilizes gait-independent features for data augmentation. Compared to our previous work, the proposed method extends the combination of images for feature exchange in data augmentation, which was limited to the same person, to different persons. The proposed method can increase the variety and number of data used for training to quadratic, whereas previous work can only increase the number of training data in proportion to the number of gait videos per person.

TABLE 2. Definition of pose features and style features in our method.

Feature	Time variability	Personality dependency	Example in gait image
f_{pose}	✓	✓	posture
f_{style}			belongings, clothing, walking direction

III. METHOD

In this section, we describe the pipeline of our proposed Feature Swap GaitNet (FSGaitNet), a novel method for identity-aware data augmentation. We begin with an overview, followed by the flow of the proposed pipeline: preprocessing, encoding, generation, re-encoding, and identification. Optimization and inference are then described.

A. OVERVIEW

We propose a pipeline called FSGaitNet to improve the accuracy of gait feature extraction. Our method separates and extracts two features from the RGB gait image: pose features and style features. Table 2 details the characteristics of these two features. The goal of feature separation is to embed gait-dependent information in the pose features and gait-independent information in the style features. For example, gait-dependent information is posture, and gait-independent information is bag, attire, camera angle, and so on. In designing features, the dependence on gait can be divided into two aspects: frame-to-frame differences and video-to-video differences. Frame-to-frame differences refer to whether or not a feature changes between consecutive frames, where information that depends on the gait changes between frames, while information that does not depend on the gait remains constant between frames. Differences between videos are considered to mean that in different videos of the same person, gait-dependent information is ideally the same when time-averaged, whereas gait-independent information has no commonality. In the

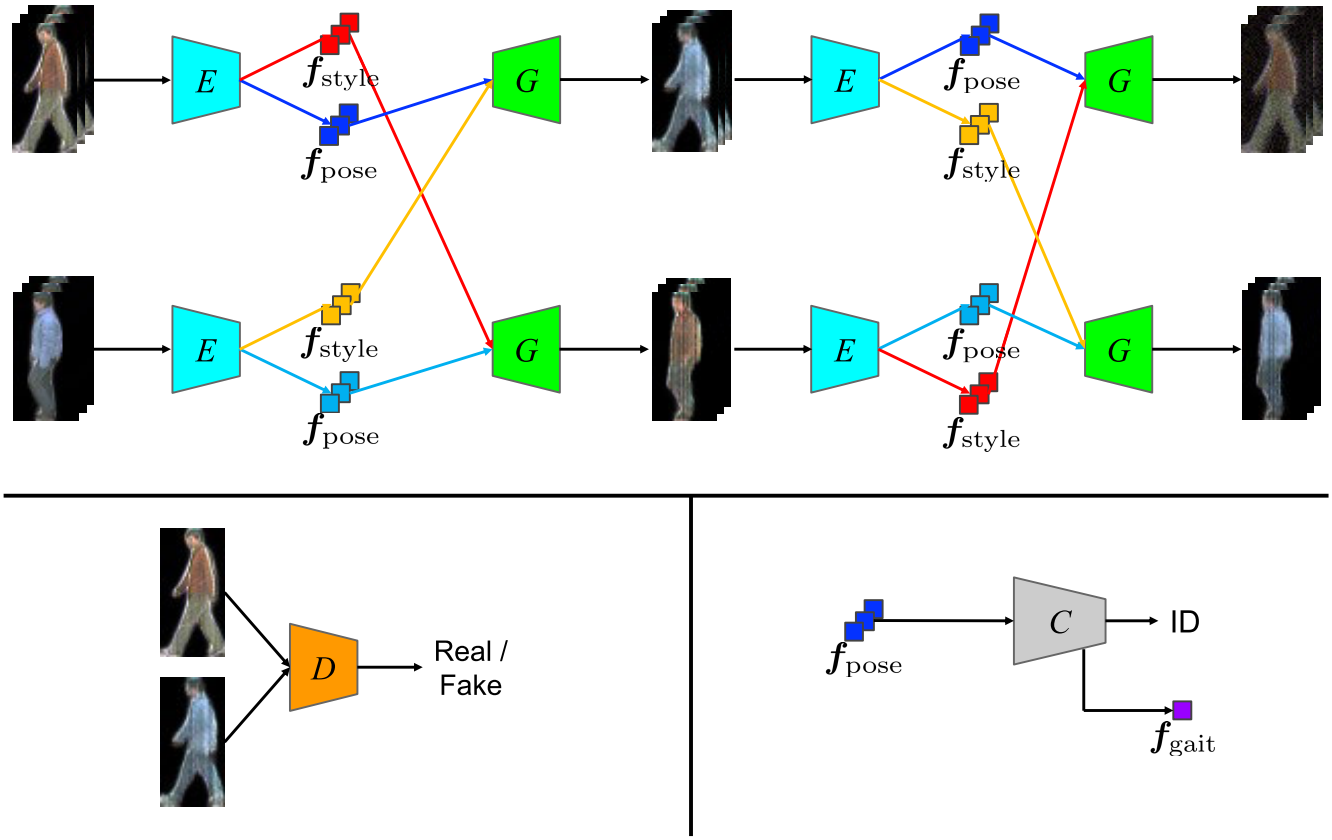

FIGURE 1. A schematic overview of the proposed Feature Swap GaitNet (FSGaitNet). The input gait video is separated into pose features f_{pose} and style features f_{style} by encoder *E*. Generator *G* generates gait images from the pose features and style features, and discriminator *D* judges the authenticity of the images. The generated images are encoded again, and classifier *C* identifies the subject ID based on the pose features separated from both the real and generated image. For evaluation, identification is conducted based on gait feature f_{gait}, which is the intermediate output of *C*.

subsequent sections, III-C and III-D, we introduce the design of loss to implement these properties.

The model architecture for our method is shown in Fig. 1. Our proposed model consists of four modules: encoder, generator, discriminator, and classifier. Encoder E separates the input gait image into pose features and style features. Generator G synthesizes a gait image from the input set of pose and style features. Discriminator D identifies whether the input image is a real image or a generated image. Classifier C extracts gait features from a time series of pose features and estimates a person ID based on the extracted features.

The following three types of gait videos are used as input in the pipeline (i) anchor, which is the main identification target; (ii) positive, which is a video of the same person as the anchor but with different gait situation settings such as clothing and walking direction; and (iii) negative, which is a gait video of a person different from the anchor. Hereafter, unless otherwise specified, we assume that we are referring to the anchor. Negatives are used only for data augmentation by exchanging features, while positives are used not only for feature exchange personnel for data augmentation, but also for adequate feature separation in the reconstruction of input images, which will be introduced in Sec. III-C

B. PREPROCESS

The actual input to the encoder is not the walk-through image in the dataset directly, but an image with the background removed. A contour of a person may contain limbs that are important to the gait, but the background outside the contour does not contain any information that depends on the gait. Therefore, the background is eliminated to simplify the problem. An instance segmentation method, Mask R-CNN, is used for object detection and person area estimation.

Although the background removal seems to be the same as the silhouette-based method, the use of the probability map output by the segmentation method is different. First, unlike silhouette-based methods such as GEI, the probability map for the person region is not binarized to 0 or 1 by thresholding, but is used as it is. Next, a soft mask image (hereinafter referred to as ''soft mask'') is obtained by multiplying the pixel values of the rectangular region of the person by the values of the probability map of the person region. The probability map value for the person area is 0 for the background and closer to the person area, the closer to 1. In other words, the image is input to the network with all areas painted black except for the person area and its surrounding area. This process reduces the dependence on the accuracy

of the person area estimation, since it is not a strict masking process like silhouette processing.

When input to the network, the soft mask is resized to 64 pixels in height and 32 pixels in width. Specifically, the height is resized to 64 pixels and the width to 32 pixels, cropped if the original soft mask is larger, or filled with zeros if it is smaller. These processes result in an image with a black background and the original image in the person area, as shown in Fig. 1.

C. SPATIAL ENCODING FROM GAIT IMAGE TO POSE AND STYLE FEATURES

The preprocessed gait video $I = \{I_1, I_2, \ldots, I_T\}$ is extracted frame by frame by the encoder for pose and style features. It is desirable to embed person-dependent information in the pose features with time-dependent degeneration, and person-independent information in the style features with time-invariance. Therefore, the encoding process concentrates on embedding person-dependent information, while time-dependent information is embedded by loss during image generation.

Anchors and positives are encoded separately, and their pose and style features are extracted. The following properties can be considered for the pose features of anchors and positives. (i) Anchors and positives are gait videos of the same person, although the walking conditions are different, so their walking styles should be highly similar. (ii) The time-series average of pose features should be close between anchors and positives, because the change of pose features is equal to the change of walking style. Based on the considerations of properties (i), (ii) and similar loss employed in GaitNet [17], the following pose similarity loss is designed using the pose features of the anchor and the positive:

$$\mathcal{L}_{\text{pose-sim}} = \left\| \frac{1}{n_1} \sum_{t=1}^{n_1} \boldsymbol{f}_{\text{pose}}^{(c_1,t)} - \frac{1}{n_2} \sum_{t=1}^{n_2} \boldsymbol{f}_{\text{pose}}^{(c_2,t)} \right\|_2^2, \tag{1}$$

where t is time.

Equation 1 computes the time-averaged difference in pose features between the anchor and the positive. By minimizing pose similarity loss $\mathcal{L}_{\text{pose-sim}}$, pose features are embedded with similar time-varying features for the same person.

D. GENERATION OF GAIT IMAGES FROM POSE AND STYLE FEATURES

As described in more detail later, the separated pose features are input to the classifier for discriminative learning. Here, equation 1 alone does not take time dependence into account, and the separation performance is insufficient. Therefore, to improve the separation performance, a gait image is generated from the separated pose and style features. Both the generated image and the source image are input to the discriminator, which identifies whether they are real or virtual images. There are two methods of generation.

- Reconstruction of the anchor image: Input style and pose features extracted from the anchor image.
- Novel generation of a virtual gait image: Input pairs of features that are different from the original image, using features extracted from the anchor image and the negative image.

1) RECONSTRUCTION

Let us first look at the reconstruction of the anchor image. While (1), as defined in the encoding process, embeds person-dependent information in the pose features, the anchor image reconstruction is designed to embed time-dependent/independent information in the pose/style features. Since style features do not change with time, the style features of any frame should be close. We exploit this time-invariance of style features to define the following loss:

$$\mathcal{L}_{\text{recon}} = \sum_{c\in\{c_1,c_2\}} \sum_{\substack{k,l\in\{1,\ldots,n\} \\ k\neq l}} \left\| G(\boldsymbol{f}_{\text{style}}^{(c,k)}, \boldsymbol{f}_{\text{pose}}^{(c,l)}) - I^{(c,l)} \right\|_1. \tag{2}$$

In (2), $G(\boldsymbol{f}_{\text{style}}^{(c,k)}, \boldsymbol{f}_{\text{pose}}^{(c,l)})$, generated from the style features extracted from the frame at time k and the pose features extracted from the frame at time l, is made similar to the original image at time l, the source of the pose features. By minimizing $\mathcal{L}_{\text{recon}}$, style features are made to embed time-invariant information, since information in the image at a time different from the source of the style feature must be extracted from the style feature.

2) SYNTHESIS

The next step is to generate a novel virtual gait image. A virtual gait image is generated from a set of pose and style features extracted from different persons. The quality of the generated image should be high because it is used for data augmentation, but unlike the reconstruction of the anchor image, there is no ground truth image in this virtual generated image, so the quality of the generated image must be improved without any teacher data. In addition, the loss in (2) results in pixel-by-pixel optimization, but does not guarantee consistency across the entire image. Therefore, adversarial loss [82] is introduced to improve the quality of the generated image.

Adversarial loss can be computed by inputting the generated image and the real image to the discriminator. In the adversarial loss optimization, the discriminator tries to detect whether the input image is a real or a generated image while the generator tries to generate an image that looks like a real image. In other words, the generator is trained to deceive the discriminator and the discriminator is trained to avoid being deceived by the generator, competing with each other. Thus, adversarial loss can improve the quality of the generated images, even in the absence of supervised data. All images, including the generated image, are input to the discriminator,

which computes the following adversarial loss:

$$\mathcal{L}_{\text{adv}} = \sum_{i,j\in\{c_1,c_2\}} \sum_{t=1}^{n} \left(\left(D(I^{(j,t)}) - \mathbb{E}\left[D\left(G(\boldsymbol{f}^{(i,t)}_{\text{style}}, \boldsymbol{f}^{(j,t)}_{\text{pose}}) \right) \right] - 1 \right)^2 + \left(D\left(G(\boldsymbol{f}^{(i,t)}_{\text{style}}, \boldsymbol{f}^{(j,t)}_{\text{pose}}) \right) - \mathbb{E}\left[D(I^{(j,t)}) \right] + 1 \right)^2 \right), \quad (3)$$

where $\mathbb{E}$ is the average within a mini-batch.

Although many derived versions of adversarial loss have been proposed since Goodfellow et al. [82] proposed it, our method employs RaLSGAN, which was the most stable in learning in preliminary experiments. RaLSGAN is defined based on LSGAN, which has better learning stability and image quality than the first definition of adversarial loss proposed by Goodfellow et al. [82]. Jolicoeur-Martineau et al. [83] point out that under the existing definition of adversarial loss, the discriminator is trained to discriminate all input images as real images. They remedy this problem by designing the discriminator to consider the discrimination results for the other type of image (i.e., discrimination results for generated images with respect to identification of real images, and discrimination results for real images with respect to identification of generated images).

By optimizing (3), the quality of the generated images can be improved even for virtual gait images with no supervisory data. In addition, while (2) is computed pixel by pixel, (3) considers the entire image, thus ensuring consistency with the source image and contributing to improved separation performance.

E. RE-ENCODING OF THE GENERATED GAIT IMAGE AND RE-GENERATION

The key to our method is online data augmentation utilizing disentangled representation learning. Hence, the generated images are used for discriminative learning as well as the real images. The generated virtual gait image is again input to encoder E and separated into pose and style features as in Sec. III-C. In our previous work [19], the following losses are employed to improve the separation performance:

$$\mathcal{L}^{\text{style}}_{\text{consis}} = \sum_{i,j\in\{c_1,c_2\}} \sum_{t=1}^{n} \left\| \boldsymbol{f}^{(i,t)}_{\text{style}} - E_{\text{style}}(G(\boldsymbol{f}^{(i,t)}_{\text{style}}, \boldsymbol{f}^{(j,t)}_{\text{pose}})) \right\|_1, \quad (4)$$

$$\mathcal{L}^{\text{pose}}_{\text{consis}} = \sum_{i,j\in\{c_1,c_2\}} \sum_{t=1}^{n} \left\| \boldsymbol{f}^{(j,t)}_{\text{pose}} - E_{\text{pose}}(G(\boldsymbol{f}^{(i,t)}_{\text{style}}, \boldsymbol{f}^{(j,t)}_{\text{pose}})) \right\|_1. \quad (5)$$

If the separation performance is high enough, encoder E should always output the same features for images with the same information. This is because the generated image should retain the features of the source image, so the features extracted from the generated image should be consistent with the features of the source image. Therefore, in order to ensure that the original image features of the separated features are accurately retained in the generated image, we designed the following losses that guarantee the consistency of the encoder E based on the difference between both pose and style features before and after re-encoding.

We focused on these losses in developing a method designed for feature exchange between the different persons (called cross-id) from a method designed for feature exchange between the same persons (called self-id) [19]. As mentioned in the introduction, preliminary experimental results show that applying the self-id method directly to cross-id yields limited improvement in identification accuracy. We attributed the problem to a decrease in the quality of the generated images due to a significant increase in the diversity of the generated images. This performance degradation encouraged us to add a regeneration flow to further improve the quality of the generated images.

In our method of exchange generation between different persons, the original image is reconstructed based on the re-encoded features. By processing two different gait images I^A, I^B in the order of encoding, feature exchange generation, and re-encoding, we obtain pose features E from I^A and style features E from I^B. Based on these two features, we define the following cycle reconstruction loss:

$$\mathcal{L}_{\text{cycle}} = \sum_{\substack{a,b\in\{A,B\} \\ a\neq b}} \left\| G\left(E_{\text{style}}\left(G\left(\boldsymbol{f}^{a}_{\text{style}}, \boldsymbol{f}^{b}_{\text{pose}} \right) \right), E_{\text{pose}}\left(G\left(\boldsymbol{f}^{b}_{\text{style}}, \boldsymbol{f}^{a}_{\text{pose}} \right) \right) \right) - I^a \right\|_1. \quad (6)$$

This loss is expected to contribute to higher quality image generation without supervised data, as reported by Zhu et al. and to improve the performance of feature separation in our proposed method.

Although it is possible here to simply apply the cycle reconstruction loss to the self-id method, we remove the feature consistency losses $\mathcal{L}^{\text{style}}_{\text{consis}}, \mathcal{L}^{\text{pose}}_{\text{consis}}$ instead of adding $\mathcal{L}_{\text{cycle}}$. The reason is that the self-id method optimizes six types of losses simultaneously, and adding new losses could make training even more difficult. Therefore, we replace these feature consistency losses $\mathcal{L}^{\text{style}}_{\text{consis}}, \mathcal{L}^{\text{pose}}_{\text{consis}}$ with cycle reconstruction loss $\mathcal{L}_{\text{cycle}}$, which only serve as supplementary functions to improve the separation performance, unlike the other losses.

F. PERSON IDENTIFICATION BY POSE FEATURES

To augment the training data, pose features extracted from both real and generated images are input to the classifier. The classifier consists of three layers of LSTM [84] followed by one fully connected layer. During training, the classifier outputs intermediate features, dynamic pose features, from the input time series of pose features through LSTM. Then, by passing it through the fully connected layer and the softmax function, the classifier outputs a probability map

of the training subject's ID. For discriminative learning, we compute the following weighted cross-entropy proposed by Zhang et al. [17]:

$$\mathcal{L}_{\mathrm{id}} = \sum_{c\in\{c_1,c_2\}} \left(\frac{1}{\sum_{t=1}^{n}\omega_t} \sum_{t=1}^{n} -\omega_t \boldsymbol{y}^{\mathrm{T}} \log\left(C(\boldsymbol{f}_{\mathrm{pose}}^{(c,1)}, \ldots, \boldsymbol{f}_{\mathrm{pose}}^{(c,t)}) \right) \right), \quad (7)$$

where $\boldsymbol{y}$ is the teacher label and ω_t is the weight to the identification result, employing $\omega_t = t^2$ as in GaitNet [17].

In (7), the general cross entropy loss is weighted according to the number of sequences of pose features input, i.e., the number of frames in the gait video to be identified. In other words, the penalty for incorrect identification results becomes larger for a gait video with a large number of frames. This is a formulation of the inductive bias that the number of frames in a video enhances the individuality of the walking style and makes it easier to discriminate.

G. SIMULTANEOUS OPTIMIZATION OF GENERATIVE AND DISCRIMINATIVE LEARNING WITH MULTI-TASK LOSS

As mentioned so far, the proposed method defines a total of five losses defined by (1), (2), (3), (6), (7). However, as mentioned in Sec. III-E, the losses described in (4), (5) are excluded to avoid instability in training. During the training phase, $\mathcal{L}_{\text{pose-sim}}$, $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{cycle}}$ contributing to generative learning, and $\mathcal{L}_{\text{id}}$ contributing to discriminative learning are simultaneously optimized. The overall loss to be optimized in the proposed method is a weighted sum of the five types of losses, which is described as follows:

$$\mathcal{L} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{pose-sim}}\mathcal{L}_{\text{pose-sim}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{consis}}^{\text{pose}}\mathcal{L}_{\text{consis}}^{\text{pose}} + \lambda_{\text{consis}}^{\text{style}}\mathcal{L}_{\text{consis}}^{\text{style}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \quad (8)$$

where λ_* is a hyperparameter that controls the effectiveness of each corresponding loss $\mathcal{L}_*$.

H. INFERENCE

Before describing the inference phase of the proposed method, we would like to review the unique problem setting of gait recognition as it relates to inference. As mentioned in the introduction, gait recognition belongs to the open-set recognition problem, which differs from the closed-set recognition problem in several respects that is employed in general classification problems in computer vision. To make the difference clearer, the pipelines during training and inference using neural networks are described for both gait recognition and classification tasks.

To begin with, for a typical classification task, the feature extractor extracts intermediate features from the input during training, and the fully connected layer and subsequent softmax function output probability maps for each class. Then, the one-hot vector of teacher labels is used to compute the error and update the network parameters. During inference, class-specific probability maps are output in the same way as during training, and the class with the largest value is used as the inference result.

In gait recognition, on the other hand, the training phase follows the same flow as in the classification problem, but the flow of inference is different. Gait recognition is an open-set recognition problem, i.e., none of the IDs used for training and evaluation are in common. As a result, fully connected layers optimized for training data cannot be adapted to evaluation data and are not suitable for use in inference.

Therefore, in gait recognition, it is common to perform a nearest neighbor search between the input probe (query) and gallery (database) based on the intermediate features output by the feature extractor. In the proposed method, as in our previous method [19], we employ the gait feature, which is the output of LSTM, as the intermediate feature and the cosine similarity as the distance metric for nearest neighbor search, respectively.

IV. EXPERIMENTS

In this section, we describe the evaluation experiments of the proposed method. First, the implementation details of the proposed method are explained and CASIA-B [20], the dataset used for the evaluation, is described. Then, the details of the experiments are described and the results are discussed. The following three experiments are conducted to verify the quality of the generated images and the discrimination accuracy of the proposed method: qualitative and quantitative evaluation of the quality of the generated images and quantitative evaluation of the identification accuracy. In the ablation study, the effectiveness of the components of the proposed method will be deeply investigated by comparing the identification accuracy with a method in which one of the components of the proposed method is removed.

A. IMPLEMENTATION DETAILS

Implementation details are described in the following order: pretreatment, network architecture, and hyperparameters. The implementation of our proposed method is built upon the implementation of our prior work [19]. This paper describes the details of the implementation, including some parameters that could not be included in the previous paper [19] due to page limitations.

1) PRETREATMENT

We begin by providing details of the pre-processing described in Sec. III-B, which was used in the experiments. As in previous work [19], the basic configuration is the same as in Zhang et al [17]. First, we use MaskR-CNN [85] to estimate the human regions from a raw RGB gait image in the dataset. MaskR-CNN uses a pre-trained model with ResNet-50-FPN as the backbone and computes a probability map of the person region for the detected rectangular region of the person. Usually, a binary hard mask is used by thresholding the probability map, but this may result in unintentional lack of information in ambiguous areas around the contour. In contrast, the softmask we use as input is a non-binary float

TABLE 3. Architecture of encoder *E*. Conv, BatchNorm and FC are abbreviations for convolution, batch normalization, fully connected, respectively.

Layer	Filter / Stride	Output Size	Activation
Input	-	3 × 64 × 32	-
Conv 1	4 × 4 / 2	64 × 32 × 16	-
BatchNorm 1	-	64 × 32 × 16	Leaky ReLU
Conv 2	4 × 4 / 2	256 × 16 × 8	-
BatchNorm 2	-	256 × 16 × 8	Leaky ReLU
Conv 3	4 × 4 / 2	512 × 8 × 4	-
BatchNorm 3	-	512 × 8 × 4	Leaky ReLU
Conv 4	4 × 4 / 2	512 × 4 × 2	-
BatchNorm 4	-	512 × 4 × 2	Leaky ReLU
FC	-	320	-
Batch Norm 5	-	320	Leaky ReLU

TABLE 4. Architecture of generator *G*. TConv stands for transposed convolution.

Layer	Filter / Stride	Output Size	Activation
Input	-	64	-
FC	-	512 × 4 × 2	-
AdaIN 0	-	512 × 4 × 2	Leaky ReLU
TConv 1	4 × 4 / 2	256 × 8 × 4	-
AdaIN 1	-	256 × 8 × 4	Leaky ReLU
TConv 2	4 × 4 / 2	128 × 16 × 8	-
AdaIN 2	-	128 × 16 × 8	Leaky ReLU
TConv 3	4 × 4 / 2	64 × 32 × 16	-
AdaIN 3	-	64 × 32 × 16	Leaky ReLU
TConv 4	4 × 4 / 2	3 × 64 × 32	Sigmoid

TABLE 5. Architecture of discriminator *D*. Spectral Norm refers to spectral normalization.

Layer	Filter / Stride	Output Size	Activation
Input	-	3 × 64 × 32	-
Conv 1	4 × 4 / 2	64 × 32 × 16	-
SpectralNorm 1	-	64 × 32 × 16	Leaky ReLU
Conv 2	4 × 4 / 2	256 × 16 × 8	-
SpectralNorm 2	-	256 × 16 × 8	Leaky ReLU
Conv 3	4 × 4 / 2	512 × 8 × 4	-
SpectralNorm 3	-	512 × 8 × 4	Leaky ReLU
Conv 4	4 × 4 / 2	512 × 4 × 2	-
SpectralNorm 4	-	512 × 4 × 2	Leaky ReLU
Conv-5	1 × 1 / 1	1 × 4 × 2	-

mask that preserves information in ambiguous areas while eliminating the influence of regions that are clearly not the person. The softmask clipped by the rectangular region of the person has a different size in each frame, so it needs to be resized before being input to the neural network. Therefore, the image padded with 0 was cropped to 64 × 32 and used as the input image.

2) NETWORK ARCHITECTURE

The network structure is the same as in the previous work [19] except for one improvement in the discriminator. Encoder E consists of four convolution layers as shown in Tab. 3, with Batch Normalization [86] and Leaky ReLU [87] following each layer. The slope of Leaky ReLU is set to 0.8 based on the DCGAN [88] discriminator. The final layer, the fully connected layer, produces a 320-dimensional vector as the output. The output vector is partitioned at a ratio of 4 : 1 and assigned to style and pose features, respectively.

Generator G has the inverse structure of encoder E, with four layers of transposed convolution layers (Tab. 4). Leaky ReLU is used as the activation function, and the slope is set to 0.8 as in encoder E. As an improvement over the baseline, Adaptive Instance Normalization (AdaIN) [89], which is widely used in style transformation tasks, was employed as the normalization process. AdaIN is computationally efficient and can apply untrained styles. The input to the G generator is a pair of style and pose features, of which the style features are used as weights and biases for AdaIN, and only the pose features are directly input to the generator. The output is a 0-1 value image tensor obtained by passing the output of the fourth transposed convolution layer through a sigmoid function.

As shown in Tab. 5, discriminator D adopted the same structure as that of encoder E, as well as DCGAN [88]. It consists of four convolution layers and uses Leaky ReLU with a slope of 0.8 as the activation function. For the normalization layer, Spectral Normalization [90] was used instead of Batch Normalization used in the encoder. Spectral Normalization is a normalization method that improves stability and generalization performance in learning GANs by guaranteeing Lipschitz continuity of the discriminator. The only difference from the previous method [19] in the network architecture is that the final layer is replaced by a convolutional layer instead of a fully connected layer that outputs a one-dimensional vector. This mechanism is based on PatchGAN's discrimnator [91], which outputs the true/false value of each patch, instead of outputting a single true/false value from the entire image. Isora et al. [91] reported that the patch-wise discrimination of authenticity values contributes to the reproduction of high-frequency components, in contrast to L1 loss $\mathcal{L}_{\text{recon}}$ defined in (2), which is concentrated on the reproduction of low-frequency components. This patch discriminator method is employed in this method with the aim to improve the quality of novel gait image generation, which is more complex than the previous method.

Classifier C has the exact same structure as Zhang et al. [17] and consists of three LSTM layers and one fully connected layer. The LSTM has 256 hidden units, and the number of units in the fully connected layer is equal to that of subjects to be identified (74 in our experiments). The sigmoid function after the fully connected layer produces a probability vector of subject IDs during training, and the LSTM produces a time series of gait features during inference.

3) HYPERPARAMETER

The other hyperparameters and other settings are exactly the same as in the previous study [19]. Adam [92] is used as

the optimizer, with a learning rate of 0.0002, beta of [0.9, 0.999], and momentum of 0.9. The batch size is 32. For the weights λ in (8), the following values are used, respectively: $\lambda_{\text{pose-sim}}$ is 10.0, λ_{recon} is 20.0, $\lambda_{\text{id}}^{\text{real}}$ is 5.0, $\lambda_{\text{id}}^{\text{fake}}$ is 0.5, λ_{adv} is 0.25, and λ_{cycle} is 10.0. All experiments were performed in the environment of Python 3.8.10 with PyTorch 1.13.0, on a single NVIDIA GeForce RTX 3090 GPU.

B. DATASET

The dataset used in the experiment is CASIA-B [20], the most widely used dataset, in order to provide a fair and comprehensive comparison. CASIA-B is a dataset consisting of 124 subjects (ID: #001-124). There are three variations of the gait condition: a normal gait setting as a reference for other settings (NM), a gait setting with a bag (BG), and a gait setting with different clothing from the reference (CL). For each subject, there are six NM videos (NM#01-06) and two each of BG and CL videos (BG#01, BG#02, CL#01, CL#02), taken from 11 different angles ranging from 0-180 degrees in 18-degree increments, for a total of 110 ($= 11 \times 10$) gait videos. In total, the dataset contains 113,640 ($= 124 \times 11 \times 10$) videos. Although CASIA-B does not provide an official evaluation protocol, this experiment follows the protocol proposed by Wu et al. [35] which has been most frequently employed in recent years, as in previous experiments [19]. All videos of 74 subjects (#001-074), in ascending order of subject ID, are used for training. The remaining subjects (#075-124) are used for evaluation.

C. COMPARED METHODS

In most gait recognition methods, a gait image is converted to another modality, such as a silhouette or a human model, by applying preprocessing. On the other hand, our method belongs to an RGB-based method that extracts gait features from RGB gait images without conversion to another modality. Since the problem statement differs between methods that extract gait features directly from RGB images and via other modalities by preprocessing, this experiment compares our method with the RGB-based method.

To the best of our knowledge, the only methods that utilize the RGB modality as a feature extraction element are the proposed work and our previous method [19] as well as GaitNet [17], [21]. GaitNet [17] is an auto-encoder with LSTM that disentangles RGB images into appearance and pose. Additionally, Zhang et al. [21] extend GaitNet by extracting gait-dependent body shape information as canonical features, which was previously included in appearance.

Due to the issue of reproducibility, the baseline is our reimplementation of GaitNet [17] with some minor improvements: changing the kernel size of the convolution and transposed convolution layers to avoid checkerboard artifacts [93], adopting mean average error in the reconstruction loss to reduce blurring [91], and changing the loss coefficients and discarding decay based on experimental results.

TABLE 6. Quantitative comparative evaluation of the quality of novel gait image generation. The smallest value is bold and underlined, and the second smallest value is underlined only.

Methods	FID ↓
Baseline	169.8
Previous work [19]	<u>118.7</u>
Ours	**<u>65.2</u>**

The previous study is a gait recognition method using feature exchange between different gait images of the same person. It differs from the proposed method in the diversity and number of data augmentation due to the huge expansion of feature exchange targets, as well as patch discriminator [91] and cycle reconstruction loss (6) to improve the quality of the generated images. For CASIA-B, which is used in this experiment, the previous work can augment the training data in proportion to the number of images per person in the dataset, and thus can learn data on the order of 110 times the original training data. Meanwhile, in the proposed method, the training data is squared to the original training data, i.e., 8,140 times (= 74 persons * 110 videos/person), since the target of feature swap is all other videos, which is much larger in both number and variety of training data than in the previous work.

D. MAIN RESULTS

To evaluate the identification accuracy, we begin with a comparison with the baseline and state-of-the-art methods. In all experiments, we use a single common model trained on all videos in all settings (NM, BG, CL) in the training dataset.

1) QUANTITATIVE EVALUATION OF GENERATED IMAGE QUALITY

For the quantitative evaluation of image generation capability, we adopt the Fréchet Inception Distance (FID), one of the most commonly used metrics for evaluating image generation models. FID is a metric that measures the distance between the distributions of both real and generated images utilizing an image recognition neural network pre-trained on a large image dataset. The smaller the FID, the closer the distribution of the generated images is to that of the real images, meaning that the quality of the generated images is higher.

In this experiment, 5,000 images are randomly selected from the dataset and used as real images. In order to generate an image, a pair of source images for a pose feature and a style feature is required. Both pairs of images are selected from the same images as the real images, and all images are used as the source images for both pose and style features, and are put together in such a way that they are not identical.

FID values calculated for each method are shown in Tab. 6. The methods selected for comparison are those that can generate RGB gait images as in this study, i.e., the baseline and our previous method [19]. The other methods introduced in Sec. IV-C are not included as comparative methods because

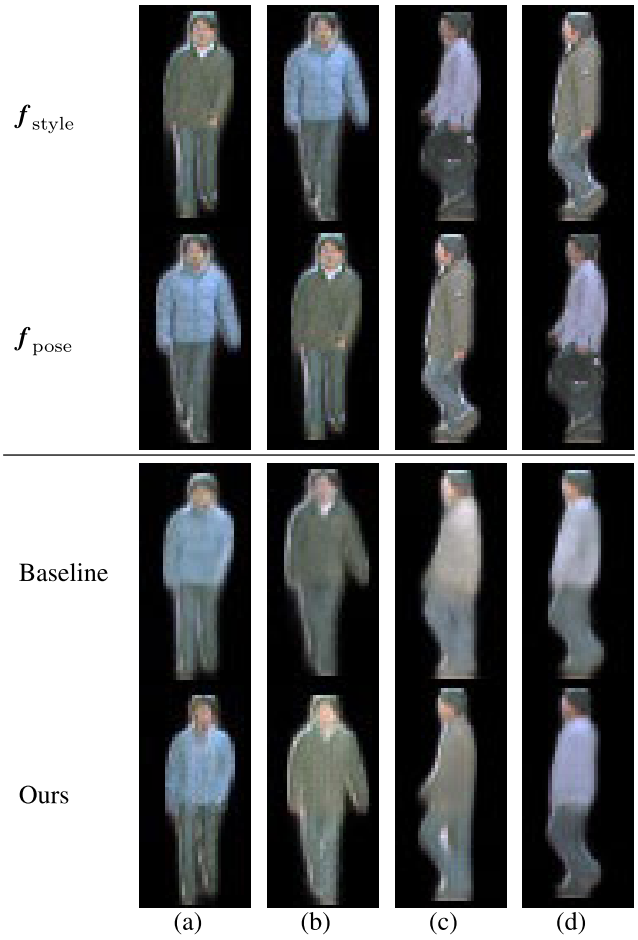

FIGURE 2. Qualitative comparative evaluation of the quality of new gait image generation. Row 1: source image of style feature f_{style}, row 2: source image of pose feature f_{pose}, rows 3 and 4: synthesized images by each method based on row 1 and 2.

they do not have a pipeline for generating novel gait images. Comparing FID of each method shows that our method has the highest quality of generated images. Since the proposed method uses the generated images as training data, this improvement in image generation quality is expected to contribute to the improvement in identification accuracy.

2) QUALITATIVE EVALUATION OF GENERATED IMAGE QUALITY

To further investigate the quality of image generation through the proposed method, a qualitative evaluation, i.e., a visual evaluation of the generated images, is performed. As shown in the results of Sec. IV-D1, the proposed method improves the quality of the generated images. However, FID measures the distribution distance between the real image and the generated image, and does not consider the quality of the image individually. In this section, visual comparison of the generated images is conducted in order to check the quality of the images at the individual level. Two gait images of different persons selected at random are used as input, and the images are generated by exchanging the combination of each other's features. Baselines, where changes are more clearly visible to the human eye, are used here as a comparison method. The generated images are shown in Fig. 2.

Fig. 2 shows the images generated using the baseline and the proposed method taking the images in the first and second rows as input. The images in the first and second rows are the base images of the generated images; the first and second rows are the source images of the style features f_{style} and pose features f_{pose}, respectively. The third and fourth rows are images generated by each method, from the top to the bottom: baseline, proposed method. This means that if the method used to generate the images performs well in image generation and feature separation, it should produce gait images such that the first row of clothing is dressed in the second row of posture. Focusing on the baseline, we see that while the two left rows (a) and (b) faithfully reproduce the clothes, the two right rows (c) and (d) cause the colors of the clothes to fade, indicating that the baseline fails to reproduce the original image. In contrast, the proposed method accurately reproduces the features of the corresponding images in all images. Therefore, it is demonstrated that the proposed method improves the quality of generated images not only at the distribution level but also at the individual level.

3) QUANTITATIVE EVALUATION OF IDENTIFICATION ACCURACY

In this section, we examine the identification accuracy, which is the final goal of the proposed method. The proposed method is unique in that it not only generates images, but also augments the training data with the generated images. Through the previous two sections IV-D1, IV-C, it has been shown that the proposed method can generate high-quality images. Here, we verify that the generated images used for training data augmentation actually contribute to the improvement of discrimination accuracy.

The results of the evaluation of the rank-1 accuracy for all settings (NM, BG, and CL) are presented in Table 7. Table 7 shows that the proposed method achieves higher accuracy than the baseline in all settings. Although the previous work also improves on the baseline performance in all settings, the proposed method has even better identification accuracy than the previous one. In addition, compared to other RGB-based methods, the proposed method outperforms them in two settings, NM and BG. On the other hand, in the CL setting, i.e., when the clothing differs between the gallery and the probe, the accuracy of the proposed method is inferior to the other methods. This is considered to be caused by the large bias in the clothing variation for each subject. NM and BG are dressed the same, and only CL is dressed differently, but for each subject there are 6 videos for NM and 2 videos each for BG and CL. That is, in each subject's videos, 80% of the total number of videos have the same clothing, and only 20% have another variation of clothing, for a total of only two types of clothing. Since the proposed method learns feature separation from differences in settings, this bias in the diversity of

TABLE 7. Rank-1 accuracy comparison among RGB-based methods. For each column, the largest value is bold and underlined, and the second largest value is underlined only.

Method	Gallery: NM #1-4											
	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
	Probe: NM #5-6											
GaitNet [17]	91.2	92.0	90.5	95.6	86.9	92.6	93.5	96.0	90.9	88.8	89.0	91.6
GaitNet [21]	93.1	92.6	90.8	92.4	87.6	95.1	94.2	95.8	92.6	90.4	90.2	92.3
Baseline†	88.5	85.8	88.5	94.1	92.1	93.6	92.3	91.7	89.7	89.9	86.1	90.2
Previous work [19]	92.4	90.8	92.8	95.3	96.1	96.5	96.7	95.8	94.7	**94.3**	**92.8**	94.4
Ours	**94.1**	**95.0**	**96.1**	**98.8**	**97.4**	**97.8**	**97.3**	**98.4**	**93.7**	93.9	92.5	**95.9**
	Probe: BG #1-2											
GaitNet [17]	83.0	87.8	88.3	93.3	82.6	74.8	89.5	91.0	86.1	81.2	85.6	85.7
GaitNet [21]	88.8	88.7	88.7	**94.3**	85.4	92.7	91.1	**92.6**	84.9	84.4	86.7	88.9
Baseline†	86.1	82.6	84.2	90.2	91.0	89.6	86.8	90.7	81.6	82.8	81.4	86.1
Previous work [19]	90.1	86.9	87.5	91.5	93.0	88.8	88.0	88.7	**87.4**	**87.8**	86.7	88.8
Ours	**94.1**	**89.2**	**90.4**	93.9	**95.5**	**92.9**	**93.6**	91.8	87.1	86.6	**87.3**	**91.1**
	Probe: CL #1-2											
GaitNet [17]	42.1	58.2	65.1	70.7	68.0	70.6	65.3	**69.4**	51.5	**50.1**	36.6	58.9
GaitNet [21]	**50.1**	**60.7**	**72.4**	**72.1**	**74.6**	**78.4**	**70.3**	68.2	**53.5**	44.1	**40.8**	**62.3**
Baseline†	31.3	27.8	25.9	28.3	25.4	25.0	23.6	24.2	18.4	20.1	17.2	24.3
Previous work [19]	38.5	37.6	34.9	32.4	31.2	27.5	26.0	26.5	22.0	23.4	21,5	29.2
Ours	46.2	42.9	38.0	31.8	26.4	27.1	26.0	25.5	22.7	21.4	21.9	30.0

† Our reimplementation of GaitNet [17].

clothing is considered to have affected feature separation. Similarly, GaitNet [17], [21] that learns feature separation from differences in settings also show lower accuracy for CL than for the other settings.[1] Although the issue of clothing variation is unique to RGB-based methods that require the separation of clothing information, the proposed method is expected to perform better if a dataset with a wide variety of clothing becomes popular. Additionally, the proposed method can handle 8,140 times more data than the original training data, but the GaitNet used as the baseline in this study is a shallow 4-layer network. If a more powerful model such as the Vision Transformer [94], though requiring a large amount of data, were employed as the backbone, the proposed method would benefit even more from data augmentation.

E. ABLATION STUDY

To better understand the effectiveness of the proposed method, we examine the performance of the proposed method with each of its components ablated one by one. Our proposed method is composed of the following five elements: patch discriminator (Patch disc.), adaptive instance normalization (AdaIN), adversarial loss ($\mathcal{L}_{\text{adv}}$), data augmentation with generated images ($\mathcal{L}_{\text{id}}^{\text{fake}}$), and cycle reconstruction loss ($\mathcal{L}_{\text{cycle}}$). We investigated the proposed method without each element one by one, and when $\mathcal{L}_{\text{consis}}$ was used instead of $\mathcal{L}_{\text{cycle}}$, which was used in the previous study [19], for a total of six patterns. For comparison, in addition to these six ablated settings, the proposed method, the baseline and the previous study, we again list the values of the previous study when the target of feature swap is changed from the same person (Self-id) to different persons (Cross-id), which is listed in Table 1. To assess both the identification accuracy and the quality of the generated images, the average rank-1 identification rates for each setting and FID are listed in Table 8. The best values among all methods are bold and underlined, and the second best values are just underlined only. Table 8 shows that the proposed method has the highest accuracy in NM and BG, and the second highest accuracy in CL, indicating that each element of the proposed method contributes to the improvement of identification accuracy. With respect to FID, the proposed method also has the third best, though not the best, quality of the generated images, hence it can be said that the proposed method is a well-balanced method in terms of both identification accuracy and quality of the generated images.

According to the relationship between identification accuracy and FID, there is a tendency that the lower the FID, the higher the identification accuracy, indicating that the improvement of the quality of the generated images is an effective and important factor in the data augmentation proposed in our method. The lowest FID is achieved when the generated images are not used for identification learning (w/o $\mathcal{L}_{\text{id}}^{\text{fake}}$), which suggests the difficulty of simultaneously optimizing both generative and discriminative learning. In particular, the highest accuracy in CL is obtained with

[1] In our experimental environment, the authors' implementation (https://github.com/ziyuanzhangtony/GaitNet-CVPR2019) could not reproduce the published values in their paper [17], [21] (mean accuracy is as follows: NM: 89.2%, BG: 86.0%, CL:34.9% for [17], NM: 92.6%, BG: 87.0%, CL: 40.0% for [21])

TABLE 8. Ablation study.

Method	Feature swap	Patch disc.	AdaIN	$\mathcal{L}_{\text{adv}}$	$\mathcal{L}_{\text{id}}^{\text{fake}}$	$\mathcal{L}_{\text{cycle}}$	Mean accuracy ↑ [%] NM	BG	CL	FID ↓
Baseline							90.2	86.1	24.3	169.8
Previous work [19]	Self-id		✓	✓	✓	$\mathcal{L}_{\text{consis}}$	94.4	88.8	29.2	118.7
Previous work [19] w/ Cross-id	Cross-id		✓	✓	✓	$\mathcal{L}_{\text{consis}}$	94.0	85.9	30.0	155.3
w/o Patch disc.	Cross-id		✓	✓	✓	✓	94.8	89.1	23.1	66.1
w/o AdaIN	Cross-id	✓		✓	✓	✓	92.6	87.7	20.8	116.0
w/o $\mathcal{L}_{\text{adv}}$	Cross-id	✓	✓		✓	✓	93.0	88.0	30.0	59.3
w/o $\mathcal{L}_{\text{id}}^{\text{fake}}$	Cross-id	✓	✓	✓		✓	95.0	89.6	**30.5**	**36.3**
w/o $\mathcal{L}_{\text{cycle}}$	Cross-id	✓	✓	✓	✓		93.3	89.6	29.1	97.6
$\mathcal{L}_{\text{cycle}} \rightarrow \mathcal{L}_{\text{consis}}$	Cross-id	✓	✓	✓	✓	$\mathcal{L}_{\text{consis}}$	94.3	89.9	26.0	80.1
Ours	Cross-id	✓	✓	✓	✓	✓	**95.9**	**91.1**	30.0	65.2

w/o $\mathcal{L}_{\text{id}}^{\text{fake}}$, suggesting that the addition of generated images to the training data reduces the accuracy of CL. As discussed in Sec. IV-D3, CL is considered to have more difficulty in extracting gait features than other settings due to the severe bias of the clothing, indicating that the images generated using CL are of lower quality than those generated using other settings for the same reason. Despite the expectation that adversarial loss improves the quality of the generated images, the second-smallest FID is in the setting where adversarial loss is eliminated. This may be due to the difficulty of learning adversarial generative networks (GANs) themselves [95], in addition to the difficulty of simultaneously optimizing generative and discriminative learning. Therefore, further improvement in the quality of generated images can be expected by employing diffusion models [96] instead, which is more stable in learning than GANs and has become the standard in image generation in recent years. Removing the patch discriminator (w/o Patch disc.) significantly worsens the accuracy of CL, which is considered to be particularly useful in the generation of complex images. In the setting excluding AdaIN (w/o AdaIN), FID degrades the most so that it plays the most important role in the generation of images in the proposed method. The settings without $\mathcal{L}_{\text{cycle}}$ (w/o $\mathcal{L}_{\text{cycle}}$) and with $\mathcal{L}_{\text{consis}}$ instead of $\mathcal{L}_{\text{cycle}}$ ($\mathcal{L}_{\text{cycle}} \rightarrow \mathcal{L}_{\text{consis}}$) result in worse all scores compared to the proposed method, which is one of the most important improvements compared to the previous work [19]. These two settings have the second largest impact on FID behind AdaIN, demonstrating that $\mathcal{L}_{\text{cycle}}$ contributes significantly to the improvement of the quality of the generated images. As mentioned in Sec. I, simply changing the target of feature exchange from the same person to a different person in the previous method [19] degraded both the identification accuracy and the quality of the generated images. However, the patch discriminator and $\mathcal{L}_{\text{cycle}}$, which are improvements in the proposed method, significantly improve the quality of the generated images, and even w/o AdaIN, which has the lowest quality among the ablated methods, has the same quality as the previous method [19]. The approach of improving the quality of the generated images used for data augmentation successfully increases the identification accuracy of the proposed method as well.

V. CONCLUSION

In this paper, we propose FSGaitNet which augments the training data by swapping gait features in order to extract gait features from RGB gait images precisely. Our proposed method is able to increase the number and variety of training data online by exchanging the gait features with those of different persons using disentangled representation learning (DRL). Experimental results demonstrate that the proposed method has the capability to generate high-quality gait images and that data augmentation with the generated images improves the discrimination accuracy. Our proposed method also outperforms the state-of-the-art method that extracts features directly from RGB in most settings. However, the huge bias in clothing in the training data indicates that challenges remain in settings where clothing differs between databases and queries. With more work addressing the issue of clothing imbalance in the dataset, we hope that the rich information in RGB images will be better explored.

REFERENCES

[1] P. K. Larsen, E. B. Simonsen, and N. Lynnerup, "Gait analysis in forensic medicine," *J. Forensic Sci.*, vol. 53, no. 5, pp. 1149–1153, 2008.
[2] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi, "Gait verification system for criminal investigation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 5, no. 1, pp. 163–175, 2013.
[3] 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, Jul. 2011.
[4] J. Ahn, K. Nakashima, K. Yoshino, Y. Iwashita, and R. Kurazume, "2 V-gait: Gait recognition using 3D LiDAR robust to changes in walking direction and measurement distance," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2022, pp. 602–607.
[5] J. Ahn, K. Nakashima, K. Yoshino, Y. Iwashita, and R. Kurazume, "Learning viewpoint-invariant features for LiDAR-based gait recognition," *IEEE Access*, vol. 11, pp. 129749–129762, 2023.
[6] C. Shen, F. Chao, W. Wu, R. Wang, G. Q. Huang, and S. Yu, "LiDARGait: Benchmarking 3D gait recognition with point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1054–1063.
[7] W. Xu, Z. Yu, Z. Wang, B. Guo, and Q. Han, "AcousticID: Gait-based human identification using acoustic signal," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–25, Sep. 2019.
[8] P. Arora, "Human recognition through gait audio," in *Proc. 8th Int. Conf. Signal Process. Commun. (ICSC)*, Dec. 2022, pp. 394–396.
[9] P. Terrier, "Gait recognition via deep learning of the center-of-pressure trajectory," *Appl. Sci.*, vol. 10, no. 3, p. 774, Jan. 2020.
[10] A. Salehi, A. Roberts, A. Phinyomark, and E. Scheme, "Feature learning networks for floor sensor-based gait recognition," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2023, pp. 1–5.

[11] M. S. Nixon, T. Tan, and R. Chellappa, *Human Identification Based on Gait*, vol. 4. Cham, Switzerland: Springer, 2010.

[12] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 264–284, Jan. 2023.

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

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

[15] 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.

[16] A. F. Bobick and A. Y. Johnson, "Gait recognition using static, activity-specific parameters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, vol. 1, Dec. 2001, pp. 1423–1430.

[17] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, and N. Wang, "Gait recognition via disentangled representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4705–4714.

[18] J. Liang, C. Fan, S. Hou, C. Shen, Y. Huang, and S. Yu, "Gaitedge: Beyond plain end-to-end gait recognition for; better practicality," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 375–390.

[19] K. Yoshino, K. Nakashima, J. Ahn, Y. Iwashita, and R. Kurazume, "Gait recognition using identity-aware adversarial data augmentation," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2022, pp. 596–601.

[20] 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. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, Aug. 2006, pp. 441–444.

[21] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On learning disentangled representations for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 345–360, Jan. 2022.

[22] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu, and J. Zhou, "Gait recognition in the wild: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, Jun. 2021, pp. 14789–14799.

[23] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3D representations and a benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20196–20205.

[24] C. Fan, S. Hou, J. Wang, Y. Huang, and S. Yu, "Learning gait representation from massive unlabelled walking videos: A benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14920–14937, Dec. 2023.

[25] P. Zhang, H. Dou, W. Zhang, Y. Zhao, Z. Qin, D. Hu, Y. Fang, and X. Li, "A large-scale synthetic gait dataset towards in-the-wild simulation and comparison study," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–23, Jan. 2023.

[26] 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.

[27] 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.

[28] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Joint intensity transformer network for gait recognition robust against clothing and carrying status," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3102–3115, Dec. 2019.

[29] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13306–13316.

[30] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes," *Pattern Recognit. Lett.*, vol. 30, no. 11, pp. 977–984, Aug. 2009.

[31] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention (ICDP)*, Dec. 2009, pp. 1–6.

[32] J. Liu and N. Zheng, "Gait history image: A novel temporal template for gait recognition," in *Proc. IEEE Multimedia Expo Int. Conf.*, Jul. 2007, pp. 257–270.

[33] 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.

[34] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 4165–4169.

[35] 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. 2017.

[36] Z. Huang, D. Xue, X. Shen, X. Tian, H. Li, J. Huang, and X.-S. Hua, "3D local convolutional neural networks for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14900–14909.

[37] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "GaitSet: Cross-view gait recognition through utilizing gait as a deep set," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3467–3478, Jul. 2022.

[38] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14213–14221.

[39] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 382–398.

[40] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14628–14636.

[41] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu, "OpenGait: Revisiting gait recognition toward better practicality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9707–9716.

[42] H. Dou, P. Zhang, W. Su, Y. Yu, Y. Lin, and X. Li, "GaitGCI: Generative counterfactual intervention for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5578–5588.

[43] K. Ma, Y. Fu, D. Zheng, C. Cao, X. Hu, and Y. Huang, "Dynamic aggregated network for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22076–22085.

[44] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 532–539.

[45] S. Yu, R. Liao, W. An, H. Chen, E. B. García, Y. Huang, and N. Poh, "GaitGANv2: Invariant gait feature extraction using generative adversarial networks," *Pattern Recognit.*, vol. 87, pp. 179–189, Mar. 2019.

[46] P. Zhang, Q. Wu, and J. Xu, "VT-GAN: View transformation GAN for gait recognition across views," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[47] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian, "Multi-view gait image generation for cross-view gait recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 3041–3055, 2021.

[48] J. Ma, D. Ye, C. Fan, and S. Yu, "Pedestrian attribute editing for gait recognition and anonymization," 2023, *arXiv:2303.05076*.

[49] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.

[50] T. Chai, X. Mei, A. Li, and Y. Wang, "Semantically-guided disentangled representation for robust gait recognition," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.

[51] L. Yao, W. Kusakunniran, P. Zhang, Q. Wu, and J. Zhang, "Improving disentangled representation learning for gait recognition using group supervision," *IEEE Trans. Multimedia*, vol. 25, pp. 4187–4198, 2022.

[52] C. Song, Y. Huang, Y. Huang, N. Jia, and L. Wang, "GaitNet: An end-to-end network for gait based human identification," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106988.

[53] Y. Cui and Y. Kang, "Multi-modal gait recognition via effective spatial–temporal feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17949–17957.

[54] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[55] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[56] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[57] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362.

[58] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7157–7173, Jun. 2023.

[59] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[60] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, May 2004.

[61] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using LSTM," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 325–330.

[62] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal–spatial network (PTSN) for gait recognition with carrying and clothing variations," in *Proc. Chin. Conf. Biometric Recognit.*, 2017, pp. 474–483.

[63] W. An, R. Liao, S. Yu, Y. Huang, and P. C. Yuen, "Improving gait recognition with 3D pose estimation," in *Proc. 13th Chin. Conf. Biometric Recognit. (CCBR)*, Aug. 2018, pp. 137–147.

[64] 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, Feb. 2020, Art. no. 107069.

[65] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gait-graph: Graph convolutional network for skeleton-based gait recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2314–2318.

[66] T. Teepe, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Towards a deeper understanding of skeleton-based gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1568–1576.

[67] E. Pinyoanuntapong, A. Ali, P. Wang, M. Lee, and C. Chen, "Gaitmixer: Skeleton-based gait representation learning via wide-spectrum multi-axial mixer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[68] C. Zhang, X. Chen, G. Han, and X. Liu, "Spatial transformer network on skeleton-based gait recognition," *Expert Syst.*, vol. 40, no. 6, p. 13244, Jul. 2023.

[69] Y. Fu, S. Meng, S. Hou, X. Hu, and Y. Huang, "GPGait: Generalized pose-based gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19595–19604.

[70] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph. (TOG)*, vol. 34, no. 6, p. 248, 2015.

[71] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10967–10977.

[72] H.-M. Hsu, Y. Wang, C.-Y. Yang, J.-N. Hwang, H. L. U. Thuc, and K.-J. Kim, "GAITTAKE: Gait recognition by temporal attention and keypoint-guided embedding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 2546–2550.

[73] J. Chen, H. Ren, F. S. Chen, S. Velipasalar, and V. V. Phoha, "Gaitpoint: A gait recognition network based on point cloud analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 1916–1920.

[74] H.-M. Hsu, Y. Wang, C.-Y. Yang, J.-N. Hwang, H. L. U. Thuc, and K.-J. Kim, "Learning temporal attention based keypoint-guided embedding for gait recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 3, pp. 1–10, Aug. 2023.

[75] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 3–20.

[76] X. Li, Y. Makihara, C. Xu, and Y. Yagi, "End-to-end model-based gait recognition using synchronized multi-view pose constraint," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 4089–4098.

[77] J. Zheng, X. Liu, X. Gu, Y. Sun, C. Gan, J. Zhang, W. Liu, and C. Yan, "Gait recognition in the wild with multi-hop temporal switch," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, p. 6136.

[78] H. Zhu, Z. Zheng, and R. Nevatia, "Gait recognition using 3-D human body shape inference," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 909–918.

[79] J.-S. Yoo and K.-H. Park, "Skeleton silhouette based disentangled feature extraction network for invariant gait recognition," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2021, pp. 687–692.

[80] P. Connor and A. Ross, "Biometric recognition by gait: A survey of modalities and features," *Comput. Vis. Image Understand.*, vol. 167, pp. 1–27, Feb. 2018.

[81] C. F. Gonçalves dos Santos, D. D. S. Oliveira, L. A. Passos, R. G. Pires, D. F. S. Santos, L. Pascotti Valem, T. P. Moreira, M. Cleison S. Santana, M. Roder, J. P. Papa, and D. Colombo, "Gait recognition based on deep learning: A survey," *ACM Comput. Surveys*, vol. 55, no. 2, pp. 1–34, 2022.

[82] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[83] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," 2018, *arXiv:1807.00734*.

[84] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.

[85] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[86] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[87] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2013, pp. 1–20.

[88] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[89] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[90] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.

[91] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[92] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[93] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, pp. 1–24, Oct. 2016.

[94] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[95] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7327–7347, Nov. 2022.

[96] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NIPS*, vol. 33, 2020, pp. 6840–6851.

KOKI YOSHINO (Graduate Student Member, IEEE) received the B.E. and M.E. degrees from Kyushu University, Japan, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Electrical Engineering. His current research interests include machine learning, computer vision, and biometrics.

KAZUTO NAKASHIMA (Member, IEEE) received the M.Eng. and Ph.D. degrees, in 2017 and 2020, respectively. From 2019 to 2020, he was a Research Fellow with Japan Society for the Promotion of Science (JSPS). He is currently an Assistant Professor with the Faculty of Information Science and Electrical Engineering, Kyushu University, Japan. His research interest includes computer vision for robotics applications.

JEONGHO AHN (Graduate Student Member, IEEE) received the B.E. degree in electronic engineering from Gachon University, South Korea, in 2019, and the M.E. degree from the Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan, in 2021, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, machine learning, and biometrics.

YUMI IWASHITA (Senior Member, IEEE) received the Ph.D. degree from the Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan. She is currently a Robotics Scientist with the Aerial and Orbital Image Analysis Group, NASA's Jet Propulsion Laboratory (JPL). Prior to joining JPL, she was an Associate Professor with Kyushu University. Her research interests include computer vision for robotics and intelligence, surveillance, and reconnaissance (ISR) applications.

RYO KURAZUME (Senior Member, IEEE) received the M.Eng. and Ph.D. degrees in mechanical engineering from Tokyo Institute of Technology, in 1989 and 1998, respectively. He was the Director of the Robotics Society of Japan (RSJ), from 2009 to 2011 and from 2014 to 2015; the Society of Instrument and Control Engineers (SICE), from 2013 to 2015; and Japan Society of Mechanical Engineers (JSME), from 2021 to 2023. He was the Chairperson of the JSME Robotics and Mechatronics Division, in 2019. He is currently a Professor with the Graduate School of Information Science and Electrical Engineering, Kyushu University. His current research interests include legged robot control, computer vision, multiple mobile robots, service robots, care technology, and biometrics. He received the JSME Robotics and Mechatronics Academic Achievement Award, in 2012; the RSJ Fellow, in 2016; the SICE System Integration Division Academic Achievement Award, in 2017; the JSME Fellow, in 2018; the SICE Fellow, in 2019; and the JSME Robotics and Mechatronics Division Robotics and Mechatronics Award, in 2021.