Gait Recognition using Identity-Aware Adversarial Data Augmentation

Koki Yoshino[1], Kazuto Nakashima[2], Jeongho Ahn[1], Yumi Iwashita[3] and Ryo Kurazume[2]

***Abstract*— Gait recognition is a non-contact person identification method that utilizes cameras installed at a distance. However, gait images contain person-agnostic elements (covariates) such as clothing, and the removal of covariates is important for identification with high performance. Disentanglement representation learning, which separates gait-dependent information such as posture from covariates by unsupervised learning, has been attracting attention as a method to remove covariates. However, because the amount of gait data is negligible compared to other computer vision tasks, such as image recognition, the separation performance of existing methods is insufficient. In this study, we propose a gait recognition method to improve the separation performance, which augments the training data by adversarial generation based on identity features, separated by disentanglement representation learning. The proposed method first separates gait-dependent features (pose features) and appearance-related covariate features (style features) from gait videos based on disentanglement representation learning. Then, synthesized gait images are generated by exchanging pose features between gait images of the person under different walking conditions, followed by adding them to the training data. The experiments indicate that our method can improve the separation performance, and generate high-quality gait images that are effective for data augmentation.**

I. Introduction

Person identification utilizing biometric data is more convenient and secure than PIN codes or handwriting, and is increasingly being utilized as a means of identification in smartphones and ATMs. In particular, gait videos, which comprises videos of walking people, are expected to be an effective biometric for criminal investigations and access control, because they can be obtained without contacting subjects and without asking them to make any special movements. In general, in gait recognition, feature vectors are calculated based on the subject's movement and appearance in the gait video, and then matched with features in a database. From the viewpoint of pattern recognition, gait recognition is an open set recognition problem in which a subject class to be evaluated may not be included in the training data, and in general, it requires high discrimination capability of the extracted gait features. One of the challenges is that gait videos contain a lot of information that is not specific to the gait (covariates), such as clothing and background, which makes authentication difficult.

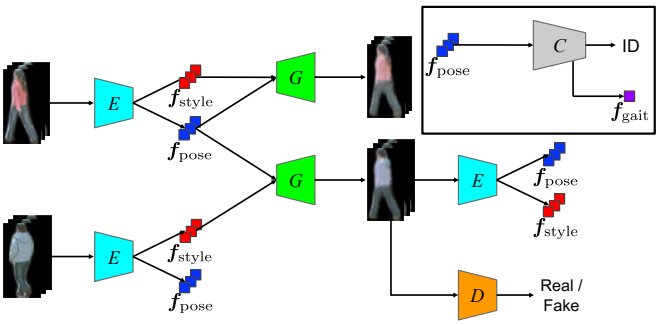

Fig. 1. A schematic overview of the proposed method. The input gait video is separated into pose features f_{pose} and style features f_{style} by encoder E. Generator G generates gait images from the pose features and style features, and discriminator D judges the authenticity of the images. The generated 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.

Recently, as a method for removing covariates, disentanglement representation learning, which separates the internal representation of a neural network according to potentially independent attributes of the data, has been actively studied. This method can learn interpretable feature representations without the need to teach attribute labels. For example, Zhang *et al.* [1] proposed a method for directly separating and removing appearance features, which is one of the covariates, by disentanglement representation learning. Thus, several conventional methods have utilized disentanglement representation learning to remove covariates. However, because the amount of gait data is negligible compared to other computer vision tasks, such as image recognition, we considered that its separation performance was limited.

To improve the separation performance of disentanglement representation learning, we propose a gait recognition method that augments the training data by adversarial generation, based on features separated by disentanglement representation learning (Fig.1). The proposed method is expected to improve the separation performances and the generalization performance of the feature extractor, because the variation and amount of training data can be increased according to the types of settings and the number of videos in each setting of the person's gait video. Specifically, the gait video is first separated into features that depend on the gait (pose feature) and covariate features related to the appearance (style feature). Then, the gait images generated by exchanging the pose features between two gait images

*This work was partially supported by JSPS KAKENHI Grant Number JP20H00230

[1]Koki Yoshino and Jeongho Ahn are with Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan. {yoshino, ahn}@irvs.ait.kyushu-u.ac.jp

[2]Kazuto Nakashima and Ryo Kurazume are with Faculty of Information Science and Electrical Engineering, Kyushu University, Japan. k_nakashima@irvs.ait.kyushu-u.ac.jp, kurazume@ait.kyushu-u.ac.jp

[3]Yumi Iwashita is with Jet Propulsion Laboratory, California Institute of Technology, USA. yumi.iwashita@jpl.nasa.gov

of the person with different settings, such as variations in clothing and walking direction, are added to the training data. Finally, the gait features extracted from the time series of the pose features are utilized for recognition of the subject. Discriminative learning and generative learning are optimized simultaneously.

In summary, our contributions are as follows:

- We propose a novel gait recognition method using identity-aware data augmentation based on disentanglement representation learning.
- Our proposed method can generate high-quality synthesized images and improve the identification accuracy.

II. Related Work

A. Deep Learning Based Gait Recognition

As with other computer vision tasks, there has been a great deal of research utilizing deep learning in gait recognition. Conventional methods have eliminated covariates by preprocessing, and can be generally classified into two categories: model-based or silhouette-based methods. Model-based gait recognition is a method that learns the movements of bones and joints specific to a person, based on skeletal information extracted by fitting a human model to gait images [2], [3]. While this method can eliminate clothing and background information that does not depend on gait, the accuracy of identification is highly dependent on the accuracy of fitting the human model, and it also removes the body shape, which is important information for identifying individuals for gait recognition.

In silhouette-based gait recognition, a human silhouette is extracted from each frame of the gait video using background subtraction or segmentation methods, and the temporal change of the silhouette contour is learned to classify the person ID. In particular, one of the most representative silhouette-based methods is a neural network that utilizes a gait energy image (GEI) [4] as input. GEI is an image obtained by averaging a sequence of silhouette images of one walking cycle over time. Because of the simplicity and efficiency of GEI, neural networks using GEI as input have been proposed. Takemura *et al.* [5] constructed a total of four types of CNNs based on metric learning, employing contrastive loss and triplet loss as the loss function, and proposed utilizing different networks, depending on the setting of the identification target. He *et al.* [6] proposed an adversarial generative network that simultaneously optimizes two tasks: estimating the observed direction, and generating a GEI transformed from the observed direction, and indicated that the proposed method is robust to differences in the observed direction between the database and query.

However, the silhouette-based method also removes the motion inside the contour, which is especially important in frontal and backward gait images where the contour shape does not alter significantly. Thus, the conventional methods of removing covariates by preprocessing have the challenge of missing information that is helpful for identification because of preprocessing.

B. Disentanglement Representation Learning for Pedestrian Image

Disentanglement representation learning has been attracting attention in recent years as a method to eliminate information that is irrelevant to gait without relying on preprocessing. Disentanglement representation learning is a method for separating and extracting potentially independent attributes in data as separate internal representations, without corresponding correct labels. By applying disentanglement representation learning to gait recognition, methods for end-to-end estimation and the removal of covariate features have been proposed. Zhang *et al.* [1] introduced disentanglement representation learning for gait recognition for the first time, and proposed a method for separating gait features and appearance features such as clothes directly from gait video. Li *et al.* [7] removed gait-independent features from GEI [4] by separating person-dependent and covariate features such as belongings and walking speed using semi-supervised disentanglement representation learning.

Disentanglement representation learning for pedestrian images has been studied not only for gait recognition, but also for the person re-identification (person re-id) task, which is an open set recognition problem for pedestrian images, as well as gait recognition. Person re-id is a task to identify a person who is captured by cameras at different locations. Person re-id does not consider changes in walking conditions between the gallery and probe, and therefore treats clothing information as the main cue for person identification. Zheng *et al.* [8] proposed a method that separates pedestrian images into appearance features (e.g., clothing color and texture) and structure features (e.g., body shape and background), and replaces the separated features with other images to simultaneously optimize data enhancement and pedestrian identification. Eom *et al.* [9] proposed a method that is robust to covariates by separating person identifying information (e.g., clothing and hair) and covariate information (e.g., posture and scale changes) from pedestrian images, and augments the data with images generated by partially exchanging the separated features with another image.

As described above, disentanglement representation learning can be utilized to directly separate and extract gait-irrelevant features from gait images without preprocessing, and explicitly remove them. However, as presented in Table I, since the dataset for gait recognition, especially the color image dataset, generally has fewer subjects than the person re-id dataset, the existing gait recognition methods based on disentanglement representation learning are considered to have inadequate separation performance, compared to the person re-id method.

III. Proposed Method

As illustrated in Fig. 1, the proposed network comprises the following four modules: encoder E that outputs pose features and style features from the gait image of each frame, generator G that generates gait images from both features output from the encoder, discriminator D that discriminates

TABLE I
COMPARISON OF REPRESENTATIVE DATASET BETWEEN GAIT RECOGNITION AND PERSON RE-ID

Dataset	Task	#IDs
CASIA-B [10]	Gait recognition	124
USF [11]	Gait recognition	122
Maket1501 [12]	Person re-id	1,501
DukeMTMC-reID [13]	Person re-id	1,812

between the real image and the generated image, and classifier C, that identifies person IDs utilizing the time-series of pose features as input.

In the proposed method, pose features are defined as time-varying and person-dependent features (e.g., body shape and joint angles) in the time series of gait images, and style features are defined as time-invariant and person-independent features (e.g., clothing and walking direction).

As a baseline for the proposed method, we adopted GaitNet [1], which is an end-to-end feature extraction method that utilizes color images. As GaitNet did in the pre-processing step, the proposed method takes gait videos as input data from which the human region is solely extracted by the segmentation method [14].

A. Encoding from Gait Image to Pose and Style Features

First, the input gait image $I^{(c,t)}$ is separated and extracted into the pose and style features by E. In $I^{(c,t)}$, the value at position c represents the walking setting, and the value at position t represents the time.

Similar to GaitNet [1], the following pose similarity loss $\mathcal{L}_{\text{pose-sim}}$ is calculated from the pose features $\boldsymbol{f}_{\text{pose}}^{(c_1,t)}$, $\boldsymbol{f}_{\text{pose}}^{(c_2,t)}$ obtained from two gait images of the same person with different walking conditions c_1 and c_2, so that the features of the gait images are properly separated and extracted:

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

The loss calculated using Eq. (1) is defined such that the time averages of the pose features of gait videos of the person in different gait conditions approach each other, based on the assumption that the gait videos of the person have a high degree of similarity in walking patterns, even if their gait conditions are different. Accordingly, the pose features are expected to be embedded with information that would cause similar temporal changes among the gait videos of the person.

B. Generation of Gait images from Pose and Style Features

Next, gait images are generated by G from the pose and style features extracted by E. The generated and original images are input to D, to determine whether they are the generated or real images. The following two patterns exist for image generation.

- Reconstruction of the original image from the style and pose features extracted from the gait image.
- Generation of a synthesized gait image by exchanging pose features between two gait images of the person under different walking conditions

To increase the quality of the generated images to a quality effective for data augmentation, we define two types of loss: reconstruction loss, which contributes to the reconstruction of the original image, and adversarial loss, which mainly contributes to the generation of synthesized gait images.

Following Zhang *et al.* [1], we separate and extract style and pose features from two randomly selected frames of the gait video at different times. The reconstruction loss $\mathcal{L}_{\text{recon}}$ is calculated from the gait image generated by inputting the style features $\boldsymbol{f}_{\text{style}}^{(c,k)}$, extracted from the frame at time k, and the pose features $\boldsymbol{f}_{\text{pose}}^{(c,l)}$ extracted from the frame at time l into G, and the frame $I^{(c,l)}$ at time l from which the pose features were extracted:

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

By minimizing Eq. (2), the image generated from the style features at time k, and the pose features at time l become closer to the real image at time l from which the pose features originate. This allows G to utilize common information across frames in the style features and unique information for each frame in the pose features when generating the image. Specifically, the time-invariant and time-variant information in the gait image are expected to be extracted as style and pose features, respectively.

In the reconstruction of the original image, the quality of the generated image can be improved by the reconstruction loss, because the original image serves as the teacher data. However, for the generation of synthesized gait images, a similar loss as reconstruction loss cannot be defined because there is no supervisory data. In addition, the reconstruction loss is the mean square error (MSE) between the original and generated image, and because the calculation is performed pixel-by-pixel, the consistency of the entire image is not guaranteed.

Therefore, in this method, we define adversarial loss [15] in addition to the reconstruction loss as the loss that contributes to generative learning. By optimizing the adversarial loss, the generator attempts to produce a realistic image, while the discriminator attempts to correctly identify whether the input image is a real or generated image. Accordingly, the generator and discriminator are trained to compete with each other, thus improving the quality of the generated image without the need for correct data from the generated image.

Both real and generated images are input to D, which computes the following adversarial loss $\mathcal{L}_{\text{adv}}$:

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

where $\mathbb{E}$ denotes the average within a mini-batch. For the definition of the adversarial loss, we adopted RaLSGAN [16],

which has the most stable learning based on the results of preliminary experiments.

By optimizing Eq. (3), even a synthesized gait image without a correct image can be generated with natural quality similar to a real image. Furthermore, unlike the reconstruction loss computed on a pixel-by-pixel basis, the optimization considers the entire image, which guarantees consistency with the real image, and is expected to improve the separation performance.

C. Re-encoding of the Generated Gait Image

To utilize the generated image for discriminative learning, it is input to E again, and separated into pose features and style features, as described in Section III-A. To improve the separation performance, E should always output the similar features for images with the similar information. The generated image should retain the features of the original image, and the features extracted from the generated image should match the features of the original image. Therefore, from the difference of the two features before and after re-encoding, the following losses $\mathcal{L}_{\text{consis}}^{\text{style}}$ and $\mathcal{L}_{\text{consis}}^{\text{pose}}$ that guarantees the consistency of the encoder E is calculated:

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

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

where $E_{\text{pose}}(*)$ refers to the pose features, and $E_{\text{style}}(*)$ refers to the style features of the output $E(*)$ from the encoder.

By optimizing Eq. (4), (5), we aim to ensure that the features of the image from which the separated features originate are accurately retained in the generated image.

D. Person Identification by Pose Features

For both real and generated images, the pose features extracted from E are input to the classifier C, and the probability distribution of the subject ID is the output. The intermediate output extracted from the layer just before the fully connected layer of C, which consists of LSTM and the fully connected layer, is defined as the gait feature $\boldsymbol{f}_{\text{gait}}$. As described in Section I, gait recognition is an open set recognition problem, and the learned fully connected layer cannot be utilized for identification. Therefore, during evaluation, the query (probe) and database (gallery) are matched by the nearest neighbor search based on the cosine similarity of the gait features. In discriminative learning, the following cross entropy is calculated from the probability distribution of the person ID output by C according to Zhang [1]:

$$\mathcal{L}_{\text{id}} = \sum_{c\in\{c_1,c_2\}} \left(\frac{1}{\sum_{t=1}^{n}\omega_t} \sum_{t=1}^{n} -\omega_t \boldsymbol{y}^{\mathrm{T}} \log\left(C(\boldsymbol{f}_{\text{pose}}^{(c,1)}, ..., \boldsymbol{f}_{\text{pose}}^{(c,t)}) \right) \right). \quad (6)$$

Note that $\boldsymbol{y}$ is the correct label, ω_t is the weight for the identification result, and $\omega_t = t^2$ was adopted according to Zhang's method [1]. The loss in Eq. (6) is weighted by ω_t according to the number of frames in the input gait video, based on the assumption that the longer the duration of the input gait video, the higher the discriminability of the gait video.

E. Simultaneous Optimization of Generative and Discriminative Learning with Multi-task Loss

As described above, six losses are defined in the proposed method. Zheng *et al.* [8] proposed a person re-id method based on data augmentation using separated features, and reported that both the quality of the generated images and the identification accuracy can be improved by simultaneously optimizing the generative and discriminative learning based on the experimental results. Therefore, our method simultaneously optimizes the above six losses in the learning process with the following multi-task loss $\mathcal{L}$:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{pose-sim}}\mathcal{L}_{\text{pose-sim}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} \\ & + \lambda_{\text{consis}}^{\text{pose}}\mathcal{L}_{\text{consis}}^{\text{pose}} + \lambda_{\text{consis}}^{\text{style}}\mathcal{L}_{\text{consis}}^{\text{style}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}},\end{aligned} \quad (7)$$

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

IV. Experiments

In the experiments, we utilized the CASIA-B [10] dataset to quantitatively and qualitatively evaluate the quality of the generated images. CASIA-B comprises three image sets: NM (normal), BG (with a bag), and CL (in a long coat). For training, we utilized the walking images of 74 subjects in the first half of the dataset. In addition, 20 consecutive frames were randomly cut from input gait video for training, and all the frames in the test dataset were used for evaluation.

Table II presents the methods adopted in the experiments. To verify the effectiveness of each component of the proposed method in detail, we performed experiments not only with the proposed method and the baseline, but also with the method in which each component was removed one by one from the proposed method. Note that the baseline utilized in this experiment was implemented by us based on the study by Zhang [1]. Adaptive instance normalization (AdaIN) [17] is a normalization process that is widely used in style transformation tasks to improve computational efficiency, and has been added to the proposed method as an improvement from the baseline. $\mathcal{L}_{\text{id}}^{\text{fake}}$ denotes only those $\mathcal{L}_{\text{id}}$ defined by (6) for the generated image (i.e., w/o $\mathcal{L}_{\text{id}}^{\text{fake}}$

TABLE II
Comparison methods

Method	$\mathcal{L}_{\text{adv}}$	$\mathcal{L}_{\text{id}}^{\text{fake}}$	$\mathcal{L}_{\text{consis}}^{\text{style}}$, $\mathcal{L}_{\text{consis}}^{\text{pose}}$	AdaIN [17]
Baseline				
Ours	✓	✓	✓	✓
w/o $\mathcal{L}_{\text{adv}}$		✓	✓	✓
w/o $\mathcal{L}_{\text{id}}^{\text{fake}}$	✓		✓	✓
w/o $\mathcal{L}_{\text{consis}}$	✓	✓		✓
w/o AdaIN	✓	✓	✓	

is the setting in which the generated image is not utilized as training data).

A. Quantitative Evaluation of Generated Image Quality

In this experiment, we randomly selected 5,000 images from all the images in the dataset and quantitatively evaluated the quality of the generated images by using Fréchet inception distance (FID) [18] for the images generated in random combinations. FID is widely used to measure the quality of generated images, and the smaller its value, the higher the quality of the generated images. Table III presents the results of the FID measurements for each method. The smallest value is bolded and underlined, and the second smallest value is underlined.

Table III indicates that the quality of the generated images is higher for the proposed method than for the baseline. Comparing the values for each method, the quality is improved only when the generated image is not used for discriminative learning (w/o $\mathcal{L}_{\mathrm{id}}^{\mathrm{fake}}$), but the quality decreases when the factors other than $\mathcal{L}_{\mathrm{id}}^{\mathrm{fake}}$ are removed from the proposed method. In particular, when the adversarial loss is removed (w/o $\mathcal{L}_{\mathrm{adv}}$), the FID is larger than the baseline, indicating that the adversarial loss contributes significantly to the improvement of the generated image quality in the proposed method.

B. Qualitative Evaluation of Generated Image Quality

The images generated by exchanging the pose features between two images of the person under different walking conditions are illustrated in Fig. 2. If the separation performance of the disentanglement representation learning is high, we expect to generate an image with the pose of the second row, but the clothing of the first row. In the second column, the leftmost two rows are images with the same shooting angle between rows 1 and 2 ($\boldsymbol{f}_{\mathrm{style}}$ and $\boldsymbol{f}_{\mathrm{pose}}$), and the rightmost two rows are images with different shooting angles.

As illustrated in Fig. 2, the outline and color of the image generated by the baseline are blurred. In addition, the details of the baseline indicated in the third column, such as the length and angle of the leg and arm, are different from those in the original image of the pose features. This indicates that the baseline can separate and extract features with blurred colors and contours when the shooting angles of the original images of the pose and style features are similar, but when the shooting angles are different, the separation performance

TABLE III
COMPARISON OF FID

Methods	FID
Baseline	169.8
Ours	<u>118.7</u>
w/o $\mathcal{L}_{\mathrm{adv}}$	205.7
w/o $\mathcal{L}_{\mathrm{id}}^{\mathrm{fake}}$	**<u>106.3</u>**
w/o $\mathcal{L}_{\mathrm{consis}}$	138.6
w/o AdaIN	130.5

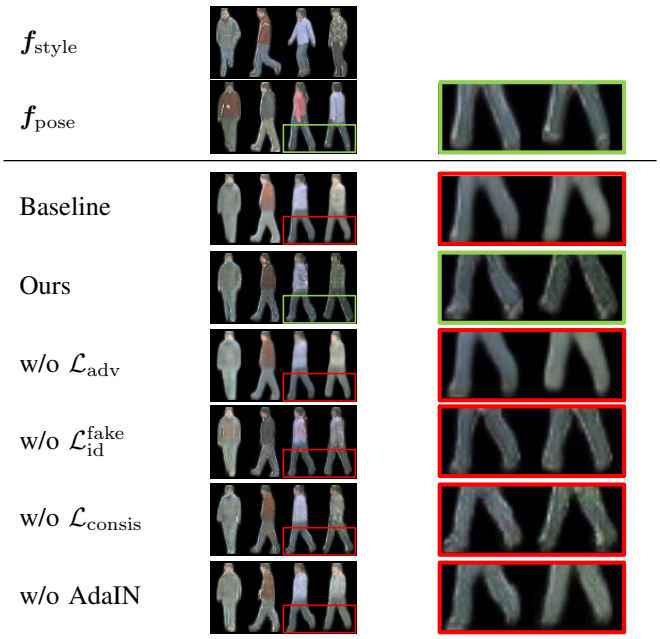

Fig. 2. Generated images by exchanging pose feature between two gait images of the person under different walking conditions. The first and second lines are the source images of the style and pose features in the generated images, respectively. The third column indicates the enlargement of a part in the second column for a detailed verification of the pose reproducibility.

is degraded. In contrast, the proposed method can separate and extract the features without blurring the contours and colors, even when the shooting angles of the source images of both features are different.

The results of the method with each element removed from the proposed method indicate that the setting with adversarial loss removed (w/o $\mathcal{L}_{\mathrm{adv}}$) results in a generated image similar to the baseline, indicating that adversarial loss contributes significantly to the improvement of the separation performance. In the setting where the generated image is not utilized for discriminative learning (w/o $\mathcal{L}_{\mathrm{id}}^{\mathrm{fake}}$), which gives the best results in the quantitative evaluation of the generated image quality, the generated image clothing is contaminated with elements of the second row image (source of pose feature), which suggests that the separation performance is worse than the proposed method.

C. Quantitative Evaluation of Identification Accuracy

In the experiments for quantitative evaluation of the discrimination accuracy, we calculated the rank-1 identification accuracy (the percentage of IDs in the nearest neighbor data in the database (gallery) that match the correct ID) for three settings: NM, BG, and CL. Following the experimental protocol as well as the baseline authors [1], we used NM gait videos of all angles as the gallery, except for the angle of the probe to be evaluated, and calculated the rank-1 identification accuracy for each angle of the query (probe). In all settings, the evaluation is performed using a common learned model.

The average values of rank-1 identification accuracy for all angles in NM, BG, and CL are presented in Table IV.

TABLE IV
RANK-1 IDENTIFICATION ACCURACY [%]

Methods	NM	BG	CL
Baseline (official announcement [1])	91.6	85.7	**58.9**
Baseline (our reimplementation)	90.2	86.1	24.3
Ours	**94.4**	**88.8**	29.2
w/o $\mathcal{L}_{\mathrm{adv}}$	91.2	87.7	27.2
w/o $\mathcal{L}_{\mathrm{id}}^{\mathrm{fake}}$	93.3	87.3	27.4
w/o $\mathcal{L}_{\mathrm{consis}}$	91.3	86.9	29.2
w/o AdaIN	92.7	86.8	29.1

The largest values are bolded and underlined, and the second largest values are underlined. The values published by Zhang [1] are also included in each table to verify the reproducibility of the baseline implementation in this experiment.

Table IV indicates that the proposed method improves the identification accuracy in all settings, and is the highest accuracy among all the methods compared. In addition, the discrimination accuracy of the proposed method is improved by each factor, especially in the settings where adversarial loss is removed (w/o $\mathcal{L}_{\mathrm{adv}}$) and where consistency loss is removed (w/o $\mathcal{L}_{\mathrm{consis}}$), the values of which are significantly less than those of the proposed method. This indicates that the contribution of adversarial and consistency losses is particularly significant in the identification accuracy of the proposed method. However, in CL, all the methods we implemented, except for the baseline indicated in [1] have an accuracy of approximately 30%, which is less than the other settings.

Based on the comparison with the published values, we believe that the reason for this is the insufficient reproducibility of the reimplemented baseline in this study. For NM and BG, the accuracy of the published values and the baseline are close, while for CL only, the baseline is approximately 30% smaller than the published values. This suggests that the reason for the lower accuracy only for CL is that our implementation of the baseline in this study is not sufficiently reproducible.

V. CONCLUSION

In this study, we present a novel gait recognition method that combines disentanglement representation learning and data augmentation to improve the separation performance of disentanglement representation learning. In the proposed method, gait images are encoded into the pose and style features. Synthesized gait images are generated by exchanging the pose features between the videos of the person under different walking conditions, and added to the training data. The experiments indicate that the proposed method improves the quality of the generated images and the performance of feature separation in disentanglement representation learning, and the synthesized images generated by the proposed method are effective for data augmentation. However, the identification accuracy is solely inadequate in the setting where the clothing is different between database and query. This is considered to be because of insufficient reproducibility of the baseline.

REFERENCES

[1] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, and N. Wang, "Gait recognition via disentangled representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4710–4719.

[2] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using lstm," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, 2016, pp. 325–330.

[3] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, p. 107069, 2020.

[4] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 28, no. 2, pp. 316–322, 2005.

[5] 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 Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 29, no. 9, pp. 2708–2719, 2017.

[6] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task gans for view-specific feature learning in gait recognition," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 102–113, 2018.

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 309–13 319.

[8] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2138–2147.

[9] C. Eom and B. Ham, "Learning disentangled representation for robust person re-identification," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 5297–5308.

[10] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, vol. 4, 2006, pp. 441–444.

[11] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 27, no. 2, pp. 162–177, 2005.

[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.

[13] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016, pp. 17–35.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, 2014, pp. 2672–2680.

[16] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," *arXiv preprint arXiv:1807.00734*, 2018.

[17] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1501–1510.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017, pp. 6629–6640.