2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

S2Gait: RGB-based Gait Recognition with Style Feature Sampling Data Augmentation

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

***Abstract*— Gait is unique to individuals and can be acquired from a distance, making it difficult to disguise. Gait videos also contain many elements unrelated to gait, which make gait recognition challenging. Departing from common approaches that use preprocessing such as silhouette extraction, the RGB-based method extracts gait features directly from RGB gait videos. RGB-based methods leverage the difference between two inputs with different attributes to separate gait-related/unrelated features, but their separation performance depends on the diversity of the dataset. To increase the amount and diversity of training data, we focus on the latent space of gait-independent features (style features), which are usually not needed for gait recognition. In this paper, we propose S2Gait (Style feature Sampling Gait), which augments the training data online with images generated from gait-dependent features of the input images and sampled style features. Experiments demonstrate that the proposed method surpasses existing RGB-based methods on almost all metrics for both generated image quality and identification accuracy. We also explore the relationship between the amount of data augmentation and performance taking advantage of our method's flexibility to generate a wide variety of gait images.**

I. Introduction

Biometrics, such as face recognition and fingerprint recognition, are convenient means of identification that do not require a tool like an ID or password. Gait is a kind of physical characteristics, and refers to the manner of walking. Compared to other biometric data, gait is particularly advantageous in identifying uncooperative subjects because it can be acquired from a distance and is generally difficult to intentionally disguise for long time periods. Gait recognition is expected to be applied in areas such as criminal investigations and smooth entry/exit management systems, utilizing these advantages. The walking pattern is specific to each individual, but the walking situation (covariates), such as belongings, clothing, and background, vary from time to time, so the removal of covariates is important in gait recognition. In camera-based gait recognition, measures against covariates can be broadly grouped into two approaches: preprocessing removal to extract silhouettes, skeletons, etc., and end-to-end removal from RGB video. Since preprocessing can lead to missing information useful for identification [1] and dependence on estimation accuracy [2], we employ end-to-end covariate removal in this study.

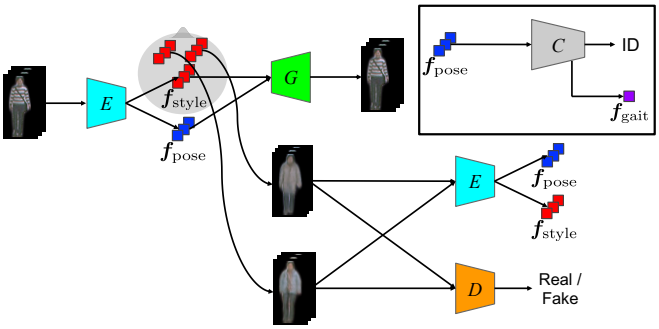

Fig. 1. A schematic overview of the proposed S2Gait. The input gait images are encoded into pose features $\boldsymbol{f}_{\text{pose}}$ and style features $\boldsymbol{f}_{\text{style}}$. Novel gait images are generated from vectors sampled from the latent space of style features along with the original pose features, and these generated images are added to the training data.

The end-to-end covariate removal approach uses RGB image sequences as input to simultaneously optimize covariate removal and gait feature extraction. Most end-to-end removal methods simply optimize the extraction of intermediate modalities such as silhouettes and human body models at the same time as discriminative learning, and thus do not prevent missing information helpful for identification. In order to take advantage of the rich information in RGB sequences, it is required to extract gait features from RGB sequences without going through intermediate representations. We refer to such approaches as RGB-based methods.

For example, Zhang *et al.* [1] proposed GaitNet, which leverages disentangled representation learning (DRL), an unsupervised learning technique to model separate feature space according to the attributes. Since their DRL approach disentangles covariate features based on the differences between videos, the separation performance depends on the diversity and quantity of the training data. To address this issue, data augmentation methods using images generated by exchanging covariate features separated by DRL with other videos have been proposed [3], [4]. Although these methods can generate new training data depending on the combination of pairs to be exchanged in the training data, the number of data generated is still limited.

In order to significantly increase the number of data generated, this paper focuses on the latent space of covariate features rather than covariate features themselves extracted from

*This work was partially supported by JST [Moonshot R&D][Grant Number JPMJMS2032], JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2132, and 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, kurazume}@ait.kyushu-u.ac.jp

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

the gait image. Covariate features are uniquely determined by the gait videos from which they are extracted, but the latent space is diverse and abundant with embedded information. The latent space of covariate features in previous methods, however, is not tractable, so features obtained by randomly sampling from the latent space do not necessarily have the similar properties as the real data.

In this paper, we propose a novel method of data augmentation of gait images by sampling covariate features independent of gait (style features) from latent space. The proposed method, named S2Gait (Style feature Sampling Gait), imposes constraints to make the latent space of style features tractable, thereby enabling effective feature sampling for data augmentation. In the proposed method, gait-dependent features (pose features) and style features are first separated from the RGB gait image through DRL. The separated style features are discarded, but new style features are sampled from the latent space of style features (Fig. 1). The sampled style features are combined with the pose features extracted from the input image to generate a novel gait image with the same pose as the input image but with different covariates. Generated images are added online to the training data, and person identification is performed based on the pose features extracted from both real and generated images. Experiments compare the quality of the generated images and the identification accuracy within the RGB-based methods. Moreover, to analyze the proposed data augmentation in depth, the relationship between the amount of data augmentation and performance is investigated.

Our contributions can be summarized as follows:

- We propose a novel gait recognition method using data augmentation that leverages the latent space of features separated by DRL.
- We demonstrate that data augmentation with images generated by sampling from the latent space of gait-independent style features improves the quality of the generated images and identification accuracy.
- We investigate the relationship between the amount of data augmentation and performance by taking advantage of the flexibility of the proposed method to generate a wide variety of gait images.

II. Related Work

A. Deep Learning Based Gait Recognition

Most gait recognition methods do not use the RGB information of the gait image directly, but identify via intermediate outputs through preprocessing such as silhouettes or human body models, so as not to be affected by the covariates in the raw gait image. Silhouette-based methods create binary silhouettes of pedestrians from the gait videos by background subtraction or segmentation, and extract gait features from the silhouettes. In the past, gait energy image (GEI) [5], which is a lower-dimensional representation of silhouette image sequence transformed by temporal averaging, was used as a gait representation [6] At present, however, silhouette-based methods, which use silhouette image sequences as input without further transformation, have been widely explored. Chao *et al.* [7] have introduced a fresh perspective by proposing GaitSet, which utilizes silhouette sequences as sets of input frames independent of their sequence order. Fan et al. [8] propose a simple silhouette-based baseline method called GaitBase that leverages insights from several previous state-of-the-art methods including GaitSet [7].

Using a human body model as a gait representation is another promising approach. In model-based methods, skeletons obtained by pose estimation have attracted attention. Teepe *et al.* [9] are the first to focus on graph convolutional networks in the field of gait recognition, and proposes GaitGraph, which handles joint positions in 3D as input. GaitTR [10] is a network that combines temporal convolution and spatial transformer.

Skeletons seem to be an efficient gait representation because they can capture only skeletal movement, but they also remove information on body shape, which is an important component of gait [11]. Alternatively, some methods [12], [13] employ a 3D mesh human body models [14], but they suffer from the same problem as skeletons in that they heavily rely on the estimation accuracy of a pretrained preprocessing models. Silhouettes similarly have a high dependence on the person area extraction model, and in addition, all information within the contour, including arms and legs, has been eliminated. Although methods that input RGB images to the network have emerged [2], [12], [13], they end up extracting gait features from intermediate representations such as silhouettes and human models output in the middle of the network, which may result in the loss of information contained in RGB. On the other hand, our proposed method directly extracts features from RGB images, and thus can take advantage of the abundant information in RGB.

B. Gait Image Generation for Identification

Despite the potential of RGB images, few RGB-based methods that directly extract gait features from RGB gait images without intermediate representations have been explored. One of the primary challenges faced in RGB-based methods is that RGB images contain numerous covariates, such as clothing textures, which can be eliminated in silhouette-based or model-based approaches. To alleviate the influence of covariates in RGB, conventional RGB-based methods have utilized the generation of gait images. GaitNet [1], [11] leverages the reconstruction of two gait image sequences with different covariates to learn the disentanglement of covariate features. Focusing on the disentangled covariate features, Yoshino *et al.* [3] propose data augmentation through the generation of a novel gait images by exchanging covariate features between two videos of the same person with different covariates. They further propose FSGaitNet [4], which extends the feature exchange pairs in its data augmentation to gait videos of different persons, improving the quality of the generated images and identification accuracy.

From the perspective of gait image generation, there exist several studies that focus on the generation of silhouettes.

Introducing generative adversarial networks (GANs) [15] for the first time in gait recognition, Yu *et al.* [16] propose a method to transform the input GEI into a canonical gait situation (right side view and without a bag). Similarly, multiple methods have been proposed that generate converted GEIs to arbitrary shooting angles and belongings, and then perform identification from the converted images [17], [18], [19]. Beyond using transformed images as a substitute for real data, different approaches have been explored that use them to augment training data. MvGGait [20] utilizes the GEIs at various viewing angles, transformed using generators trained across multiple datasets, as additional training data. GaitEditor [21] can generate anonymized silhouette images that vary in gait as well as images with the same gait and different covariates used for data augmentation.

RGB is more informative than binary silhouette and has many elements that are not related to gait. Accordingly, compared to silhouette-based methods, RGB-based methods, including the proposed method, have a greater variety of useful images for data augmentation and, as a result, are considered to have greater potential. In addition, RGB-based methods place more emphasis on the amount and diversity of training data in order to accurately remove covariates without relying on preprocessing. Contrary to methods that use two RGB videos as a pair and augment data by exchanging their features, our proposed data augmentation by sampling covariate features can generate novel gait images without requiring a pair. Therefore, the proposed method is more efficient, and increases the amount and type of training data flexibly without being bound by the way the pairs are taken.

III. Proposed Method

We describe in detail our proposed method, S2Gait, which augments data by sampling gait-independent features from the latent space. The architecture of the proposed network is shown in Fig. 1 and consists of four modules: encoder E, generator G, discriminator D, and classifier C. In the proposed method, it is necessary to separate and extract gait-dependent and gait-independent features from the input RGB gait images. Therefore, we employ GaitNet [1] as a framework for disentangling RGB images. GaitNet takes two gait videos of the same person in different situations as input and separates features from the difference between them. The gait-dependent features in the RGB gait images are features associated with posture, referred to as posture features $\boldsymbol{f}_{\text{pose}}$. On the other hand, gait-independent information includes, for example, belongings, walking direction, and clothes, which are called style features $\boldsymbol{f}_{\text{style}}$.

S2Gait's pipeline consists of four main steps: encoding, image reconstruction with different frame, data augmentation with style features sampling, and identification. There are two inputs: the main identification target gait video and a different gait video of the same person, which we call the anchor $\mathbf{I}^{\text{a}}$ and the positive $\mathbf{I}^{\text{p}}$, respectively. The positive is input in parallel with the anchor in order to allow the features to be disentangled, but all processes are performed on both anchors and positives unless otherwise noted.

A. Encoding to Extract Pose and Style Features

Initially, the input sequence of gait images $\mathbf{I} = \{I_1, I_2, ..., I_T\}$ is fed into an encoder E to extract pose and style features. The encoder performs spatial feature extraction for each frame by convolution. The extracted feature maps are output via a linear layer, separating them into pose features $\boldsymbol{f}_{\text{pose}}$ and style features $\boldsymbol{f}_{\text{style}}$.

In addition to the anchors, the positives are also input to the encoder to extract the pose and style features. For the two feature vectors to be extracted, a loss is defined to embed the properties of the pose and style features. Focusing on the person dependence, the pose features should be close to each other if they are extracted from the same person's gait videos. Therefore, using the pose features $\boldsymbol{f}^{\text{a}}_{\text{pose}}$ extracted from the anchors and $\boldsymbol{f}^{\text{p}}_{\text{pose}}$ extracted from the positives, we define the following loss:

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

Equation (1) takes the difference between the average for time t of the pose features of the anchors and the positives.

B. Reconstruction of Input Image with Different Frame Pair

The pose and style features extracted by the encoder E are input to the generator. The generator synthesizes a gait image based on the information from the two feature pairs.

Here, there are only one type of pose features extracted from the input image, while there are two types of style features in the proposed method: those extracted from the input image and those sampled from a Gaussian distribution. Style features extracted from the input image are used for feature disentanglement, and sampled style features are used for data augmentation respectively.

For feature disentanglement, (1) embeds pedestrian dependence in the features, but does not take into account variability over time. In terms of variability over time in gait videos, style features are invariant over time, whereas pose features vary with time. Taking advantage of this property, we define the following loss to reconstruct the original image of the pose features by pairing the style features extracted from different frames for the pose features in each frame.

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

where k and l denote the time and have different values.

C. Data Augmentation with Style Feature Sampling

$\mathcal{L}_{\text{pose-sim}}$ and $\mathcal{L}_{\text{recon}}$ make the posture information and style information embedded in their respective latent spaces. Our proposed method generates gait images with a variety of styles by sampling style features from the latent space and adds them to the training data. However, the latent space of style features acquired by a simple encoder is sparse, and random sampling may result in data that is far from real.

Accordingly, we constrain the latent space of style features to be tractable, in this case a Gaussian distribution $\mathcal{N}$:

$$\mathcal{L}_{\text{Gaussian}} = \sum_{t=1}^{n} D_{KL}\left(\boldsymbol{f}_{\text{style}} || \mathcal{N}(0,1)\right) \tag{3}$$

, where $D_{KL}(P||Q) = \sum P(z) \log\left(\frac{P(z)}{Q(z)}\right)$.

Optimization of $\mathcal{L}_{\text{Gaussian}}$ enables sampling of a wide variety of style features. In addition, it is shown that it also contributes to improving the performance of feature disentanglement, since it imposes constraints on the expressive capacity of the style features [22].

Style features sampled from the latent space are paired with features extracted from real data and fed into the generator to synthesize a novel gait image. In contrast to reconstruction of the source image, no image can be used as supervisory data for the synthesis of a novel gait image from sampled style features. Additionally, while $\mathcal{L}_{\text{recon}}$ emphasizes consistency on a per-pixel level, it does not take into account information over the entire image. Hence, we adopt the framework of generative adversarial networks (GANs) [15] to improve the quality of the generated images.

Both the generated and real images are input to the discriminator D, which judges whether the image is a real or a generated image. GANs are trained in such a way that the discriminator can distinguish between a real image and a generated image, and the generator tries to generate an image of such quality that the discriminator cannot distinguish the generated image from the real image. Our proposed method employs RaLSGAN [23] as an adversarial loss, which had the best quality of generated images in our preliminary experiments. The definition of loss is as follows:

$$\begin{aligned}\mathcal{L}_{\text{adv}} = \sum_{t=1}^{n} \Big(&\Big(D(I^t) - \mathbb{E}\left[D\left(G(\boldsymbol{f}^t_{\text{style}},\ \boldsymbol{f}^t_{\text{pose}})\right)\right] - 1\Big)^2 \\ &+ \Big(D\left(G(\boldsymbol{f}^t_{\text{style}},\ \boldsymbol{f}^t_{\text{pose}})\right) - \mathbb{E}\left[D(I^t)\right] + 1\Big)^2\Big),\end{aligned} \tag{4}$$

where $\mathbb{E}$ denotes the mean over mini-batch samples.

D. Gait Feature Extraction and Discriminative Learning

The pose features extracted from the real images are also input to the classifier C. The virtual gait image generated using the sampled style features is input to encoder E for training data augmentation. The pose features extracted from the virtual gait image are also input to the classifier C in the same way as the real images. The classifier C first extracts the gait feature $\boldsymbol{f}_{\text{gait}}$, which is a temporal feature of the pose features from the pose feature sequence. Then, the gait features are converted into a probability distribution of subject IDs by the fully connected layer of the classifier C. Finally the following identification loss is computed:

$$\mathcal{L}_{\text{id}} = \left(\frac{1}{\sum_{t=1}^{n} \omega_t} \sum_{t=1}^{n} -\omega_t \log\left(C(\boldsymbol{f}^1_{\text{pose}}, ..., \boldsymbol{f}^t_{\text{pose}})\right) \right). \tag{5}$$

Note that ω_t is the weight for the identification result, and $\omega_t = t^2$ was applied as same as Zhang *et al* [1]. $\mathcal{L}_{\text{id}}$ is the cross entropy loss weighted with the number of frames of the input gait video.

E. Optimization and Inference

At training phase, the following loss is minimized as the multi- task loss, which is the sum of the aforementioned five losses multiplied by their respective weights λ:

$$\begin{aligned}\mathcal{L} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{pose-sim}}\mathcal{L}_{\text{pose-sim}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} \\ + \lambda_{\text{Gaussian}}\mathcal{L}_{\text{Gaussian}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}},\end{aligned} \tag{6}$$

In the inference phase, since gait recognition is categorized as an open-set recognition problem, i.e., identifying unknown classes, the fully connected layer of a classifier C trained specific to the training subjects is not applicable. So, the nearest neighbor search is performed between probe (query) and gallery (database) based on the cosine similarity of the gait feature $\boldsymbol{f}_{\text{gait}}$.

IV. Experiments

The experiments were conducted on the CASIA-B dataset [24], which provides the RGB modality and has been the most widely used to evaluate the performance of gait recognition. CASIA-B is the dataset which consists of 124 subjects, of which 74 were used for training and the remaining 50 for evaluation. CASIA-B comprises three types of videos for analyzing covariates: NM (reference gait setting), BG (carrying a bag), and CL (dressed differently from the reference). Our proposed method is implemented based on FSGaitNet [4], which is an RGB-based method that incorporates the data augmentation as well as proposed method. In the experiments, the proposed method is compared with other methods that are included in the same RGB-based method. We compare within RGB-based methods that extract gait features from RGB gait images without going through other modalities, which is the same problem framework as this study. In order to validate the superiority of proposed method over other methods, we first conducted a quantitative and qualitative evaluation of the quality of the generated images, and a quantitative evaluation of the identification accuracy. Furthermore, for an in-depth analysis of the data augmentation in proposed method, we evaluated the correlation between the number of generated images used for data augmentation and the performance.

A. Quantitative Evaluation of Image Generation Capability

The metric for quantitative evaluation of image generation quality is the Fréchet inception distance (FID) [25], which

TABLE I
Comparison of FID

Methods	FID
Baseline	169.8
FSGaitNet-pre [3]	118.7
FSGaitNet [4]	<u>65.2</u>
Ours	**<u>25.5</u>**

Fig. 2. Generated images by style feature sampling. The leftmost images are the original images of the pose features used for the generation. The other images are generated from the pose features of the leftmost image and randomly sampled style features.

is most commonly used to evaluate generative models for images. FID measures the distance between the feature distribution of real and generated images utilizing the deep convolutional neural network that has been pretrained for the image classification task. A smaller FID indicates that the image is more faithful to the real image. For the FID measurement, we adopted the same method as FSGaitNet [4] to ensure a fair comparison with other approaches. The images used for the measurement are 5,000 randomly selected images from the dataset. Pairs of images are randomly selected. From each pair, pose features are extracted from one image and style features from the other. These extracted features are then used to generate gait images for evaluation. Although the proposed method does not require image pairs to generate images for data augmentation, this procedure was adopted for fair comparison.

The values of the FIDs for the proposed and comparative methods are shown in Table I. In all tables in this paper, the best value is bold and underlined, and the second best value is underlined only. The FID of the proposed method is the smallest among all methods, indicating that the quality of the generated images is the best. Even though the proposed method does not train image generation by feature exchange, it is able to generate images with higher fidelity than methods that concentrate on feature exchange generation [3], [4]. In addition to the benefit of data augmentation, this is considered to be due to $\mathcal{L}_{\text{Gaussian}}$, which is not employed in other methods, improving the accuracy of feature disentanglement, as described in Section III-C.

B. Qualitative Evaluation of Image Generation Capability

This experiment verifies that the proposed method can generate a variety of images by sampling style features. Pose features extracted from the input images and style features randomly sampled from a Gaussian distribution are input to the generator to generate the images. Fig. 2 demonstrates that images with the same posture and a variety of styles are generated. The images in the top row show that the position of the arms within the contour is also accurately reproduced in the generated images. Conventional methods [3], [4] can generate a variety of images, but the diversity is capped by the way the images are paired. Even if randomly sampling from the latent space of style features as in the proposed method, these methods do not guarantee that meaningful data are generated.

TABLE II
RANK-1 RECOGNITION RATE [%]

Methods	NM	BG	CL
GaitNet [1]	91.6	85.7	58.9
GaitNet [11]	92.3	88.9	**62.3**
Baseline†	90.2	86.1	24.3
FSGaitNet-pre [3]	94.4	88.8	29.2
FSGaitNet [4]	95.9	91.1	30.0
Ours	**96.1**	**92.4**	31.5

† Our reimplementation of GaitNet [1].

C. Evaluation of Identification Accuracy

The evaluation of identification accuracy is based on the rank-1 recognition rate, which is commonly used as an evaluation metric in CASIA-B [24]. The rank-1 recognition rate is the percentage of the number of cases where the ID of the data in the gallery that is closest to the probe matches the correct ID. For the three probe settings (NM, BG and CL), the NM gait videos are used in common for the gallery.

Table II shows that the proposed method improves baseline performance in all settings. Among the methods that utilize data augmentation [3], [4], the proposed method is the most accurate in all settings, which demonstrates the efficiency of the proposed data augmentation by sampling of style features. Compared to the methods that do not augment the training data, our method outperforms the performance in all settings except CL[1]. However, the proposed method is expected to achieve even higher accuracy than GaitNet [1], [11] if the same level of scores as in [1] can be reproduced at the baseline, since the proposed method improves the accuracy of CL compared to the baseline.

D. Impact of Number of Augments

In contrast to data augmentation methods based on one-to-one feature exchange within a pair [3], [4], the proposed method has no limit on the number of images to be generated for a single input image. In order to better understand the relationship between data augmentation and performance, this experiment focuses on the performance when the number of data augmentations per input image is varied.

As shown in Table III, the highest identification accuracy is achieved when two generated images are added to the training data per input image (# augments is 2), and this number is employed in the proposed method. In terms of identification accuracy, the setting with no data augmentation (# augments is 0) and the setting with a large amount of data augmentation (# augments is 16)) are less accurate than the other settings. Thus, although the proposed data augmentation is efficient enough to create an enormous

[1]We could not reproduce the score even with the authors' implementation (https://github.com/ziyuanzhangtony/GaitNet-CVPR2019).

TABLE III

PERFORMANCE DEPENDING ON THE AMOUNT OF DATA AUGMENTATION

# augments	Mean accuracy ↑ [%]			FID ↓
	NM	BG	CL	
0	92.3	87.4	25.5	**19.7**
1	94.8	90.6	29.6	26.5
2	**96.1**	**92.4**	**31.5**	25.5
4	95.7	91.9	28.2	26.0
8	95.8	**92.4**	25.0	30.1
16	91.7	86.8	25.4	123.2

number of diverse images from a single image, it is suggested that increasing the amount of data augmentation for a single image too much may cause overfitting. For the generated image quality (FID), the proposed method (# augments is 2) is the second best value, and the quality deteriorates as the amount of data augmentation increases beyond 2. FID is the best when data augmentation is not applied (# augments is 0), which is considered to be due to concentration on improving the quality of the generated images without data augmentation.

V. CONCLUSION

We proposed S2Gait, a novel gait recognition method based on data augmentation that leverages the latent space of style features. By making the latent space of style features tractable, style features can be sampled, resulting in the flexibility to add diverse data to the training data. Experimental results indicated that the proposed method improved image generation and identification capabilities, and the relationship between their performance and the amount of data augmentation. Future work includes utilizing likelihood when sampling and employing larger backbone models.

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)*, pp. 4710–4719, 2019.

[2] J. Liang, C. Fan, S. Hou, C. Shen, Y. Huang, and S. Yu, "Gaitedge: Beyond plain end-to-end gait recognition fornbsp;better practicality," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 375—-390, Springer, 2022.

[3] K. Yoshino, K. Nakashima, J. Ahn, Y. Iwashita, and R. Kurazume, "Gait recognition using identity-aware adversarial data augmentation," in *Proceedings of the IEEE/SICE International Symposium on System Integration (SII)*, pp. 596–601, IEEE, 2022.

[4] K. Yoshino, K. Nakashima, J. Ahn, Y. Iwashita, and R. Kurazume, "Rgb-based gait recognition with disentangled gait feature swapping," *IEEE Access*, vol. 12, pp. 115515–115531, 2024.

[5] 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.

[6] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in *Proceedings of the IEEE International Conference on Biometrics (ICB)*, pp. 1–8, 2016.

[7] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 8126–8133, 2019.

[8] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu, "Opengait: Revisiting gait recognition towards better practicality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9707–9716, 2023.

[9] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gaitgraph: Graph convolutional network for skeleton-based gait recognition," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 2314–2318, 2021.

[10] C. Zhang, X.-P. Chen, G.-Q. Han, and X.-J. Liu, "Spatial transformer network on skeleton-based gait recognition," *Expert Systems*, vol. 40, no. 6, p. e13244, 2023.

[11] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On learning disentangled representations for gait recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 44, no. 1, pp. 345–360, 2022.

[12] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 3–20, 2020.

[13] X. Li, Y. Makihara, C. Xu, and Y. Yagi, "End-to-end model-based gait recognition using synchronized multi-view pose constraint," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4106–4115, 2021.

[14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: a skinned multi-person linear model," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[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, pp. 2672–2680, 2014.

[16] S. Yu, H. Chen, E. B. Garcia Reyes, and N. Poh, "Gaitgan: Invariant gait feature extraction using generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 30–37, 2017.

[17] 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 (TIFS)*, vol. 14, no. 1, pp. 102–113, 2018.

[18] P. Zhang, Q. Wu, and J. Xu, "Vt-gan: View transformation gan for gait recognition across views," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[19] 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)*, pp. 13309–13319, 2020.

[20] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian, "Multi-view gait image generation for cross-view gait recognition," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 3041–3055, 2021.

[21] J. Ma, D. Ye, C. Fan, and S. Yu, "Pedestrian attribute editing for gait recognition and anonymization," *arXiv preprint arXiv:2303.05076*, 2024.

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

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

[24] 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, pp. 441–444, 2006.

[25] 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, pp. 6629–6640, 2017.