Spatio-temporal generative adversarial network for gait anonymization

Spatio-Temporal Generative Adversarial Network for Gait Anonymization

Ngoc-Dung T. Tieu[a], Huy H. Nguyen[a], Hoang-Quoc Nguyen-Son[b], Junichi Yamagishi[a,b,c], Isao Echizen[a,b]

[a] *SOKENDAI (The Graduate University for Advanced Studies), Hayama, Kanagawa, Japan*
[b] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan*
[c] *The University of Edinburgh, Edinburgh, UK*

Abstract

Gait anonymization for protecting a person's identity against gait recognition while maintaining naturalness is a new research direction. It can be used to protect the identity of people in videos to be posted on social networks, in police videos that require redaction, and in videos obtained from surveillance systems. We have developed a spatio-temporal generative adversarial network (ST-GAN) that uses random noise synthesized in the gait distribution to generate anonymized gaits that appear natural. ST-GAN consists of a generator that uses the original gait and random noise to generate an anonymized gait and two discriminators, a spatial discriminator and a temporal discriminator, to estimate the probability that a gait is the original one and not an anonymized one. Evaluation showed that the anonymized gaits generated with the proposed method are more natural than those generated with an existing method and that the proposed method outperforms the existing method in preventing gaits from being recognized by a gait recognition system.

Keywords: Gait; biometric feature; security; gait anonymization; deep learning

1. Introduction

The human gait, i.e., the manner and pattern of walking, has become an important biometric trait because it is unique to each person and can be recognized at a distance without physical contact or the person's cooperation [1]. However, a serious privacy problem may arise if the person in a video is identified unintentionally by a gait recognition system because his/her personal information (e.g., name, address, occupation) may eventually be revealed. Therefore, it is important to anonymize a person's gait in a video (while retaining its naturalness) before making the video publically available. Gait anonymization has potential applications in many areas where personal information should be protected: uploading a selfie video to the Internet, releasing a video of a criminal suspect to be shown on television, and viewing of a video captured by a surveillance system. Gait anonymization can also provide security to women who have been victims of violence. Women who have experienced violence are often afraid to speak out against the perpetrator because they do not want to be identified. Anonymizing their gait may make them more confident about speaking out on public media channels (e.g., social networks, television).

Previous research on gait anonymization has investigated such methods as obscuring a person's body [2] and pixelating or blurring the body [3, 4, 5]. These methods generally focus on privacy protection and do not address the problem of retaining gait naturalness. As far as we know, only one reported method [1] is aimed at anonymizing gaits so that they cannot be recognized while retaining their naturalness. The model used by this method anonymizes a gait by using a deep neural network to add a noise gait to the original one. The noise gait must be different from the original gait, have a length sequence similar to that of the original gait, and should have the same viewing angle as the original gait. Since the use of a noise gait may reduce anonymization success if the gait recognition system confuses the original gait with the noise gait,

Email addresses: dungtieu@nii.ac.jp (Ngoc-Dung T. Tieu), nhhuy@nii.ac.jp (Huy H. Nguyen), nshquoc@nii.ac.jp (Hoang-Quoc Nguyen-Son), jyamagis@nii.ac.jp (Junichi Yamagishi), iechizen@nii.ac.jp (Isao Echizen)

choosing the optimal noise gait is not an easy task. In addition, because the loss function for the naturalness is not strong enough, the anonymized gait looks less natural, especially for viewing angles of 0° and 180°.

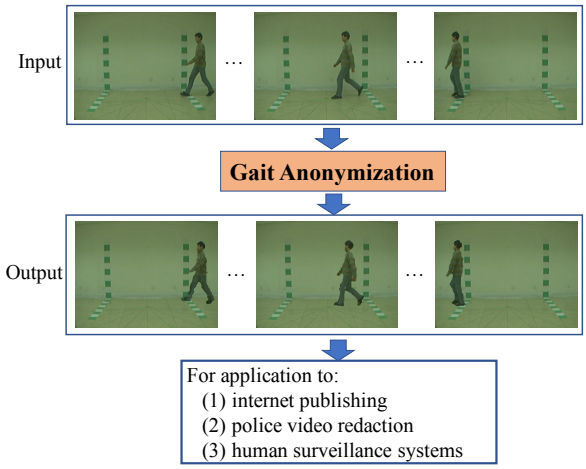


Figure 1: The scenario of gait anonymization.

Recent advances in deep neural network technologies are being applied in a wide range of areas such as image classification, image generation, and security authentication. Among the various methods for image generation, ones using a generative adversarial networks (GANs) have shown enormous potential and have become widely used. GANs of various forms have been applied to numerous tasks such as image synthesis [6, 7], image editing [8, 9], and super-resolution image generation [10, 11].

Motivated by the potential of GANs, we have developed a novel method that addresses the two problems mentioned above: confusion between original gait and noise gait and weak loss function. Instead of using a noise gait, we use random noise generated with the traditional generative adversarial network [12] from a random vector sampled from a Gaussian distribution. This network is pretrained before being assigned to the main network, a spatio-temporal generative adversarial network (ST-GAN), used to generate anonymized gaits that appear natural. The ST-GAN consists of a generator and two discriminators. The generator uses the original gait and the random noise to generate an anonymized gait while the two discriminators, a spatial discriminator and a temporal discriminator, distinguish real gaits from generated ones.

We evaluated the performance of our method by using the CASIA-B dataset [13] and two metrics suggested by Tieu et al. [1], the success rate and naturalness. We used two gait recognition systems as black-box systems for measuring the success rate, one by Zheng et al. [14] and one by Wu et al. [15]. The former was more rapid, with accuracy up to 89%; the latter was more robust, with accuracy up to 98%. We conducted two tests for measuring naturalness: subjective evaluation by human volunteers and automatic evaluation by machine. The success rate was higher and the naturalness of the anonymized gaits was higher with the proposed method.

The contributions of this work are threefold:

- The proposed ST-GAN model anonymizes gaits by using random noise obtained by mapping noise from a Gaussian distribution to the gait distribution.
- Use of the ST-GAN model substantially improves the naturalness of the anonymized gaits because it uses spatio-temporal discriminators.
- The proposed method can be applied to color video, making it much more useful.

In this paper, we overview related work in Section II, describe the proposed method in Section III, present the evaluation results in Section IV, and discussion and mention future work in Section V.

2. Related Work

2.1. Gait Recognition Systems

Each person's gait, i.e., pattern of walking, is unique [16], [17] and is thus widely used for biometric identification [18]. Gait recognition systems are aimed at recognizing an individual on the basis of his/her pattern of walking. Gait recognition approaches can be roughly divided into two categories: model-free and model-based [19, 20, 21]. The former extract features from the whole silhouette and use them to identify the individual while the latter model the gait explicitly by using body parts, such as the arms and legs.

Since the model-free approaches tend to be less sensitive to the quality of gait sequences and have lower computational cost than the model-based approaches [19, 21], we used two model-free gait recognition systems (one proposed by Zheng et al. [14] and one proposed by Wu et al. [15]) to evaluate the success rate of our proposed method. Both systems use gait energy im,ages (GEIs) as gait features. They are obtained by averaging aligned human silhouettes in gait sequences. The GEI was

Figure 2: Gait energy image (image on right) is average of human silhouettes (images on left) in gait sequence.

first defined by Han and Bhanu [22]. Fig. 2 shows an example silhouette sequence and its GEI.

2.2. Human Motion Synthesis

Human motion modeling is a key problem in two fields: computer vision and robotics. In computer vision, a common approach to motion synthesis is statistical model construction. Grochow et al. [23] used scaled Gaussian process latent variable models to produce the pose satisfying a given set of constraints while maintaining the style of the training data. Chai and Hodgins [24] regarded user-constrained motion generation as a maximum a posteriori probability problem and proposed a motion synthesis method using linear dynamic system modeling. In more recent work, Holden et al. [25] used convolutional autoencoders to learn manifold motion data and a deep feedforward neural network stacked on top of a trained autoencoder to generate a human motion sequence from control parameters. The control parameters were the trajectory of the target character over the terrain and the movement of the end effectors. These research efforts were aimed at automatically creating animations from a set of constraints while our objective is to generate a modified version of an original gait.

In robotics, Semwal et al. [26], [27] designed cellular automata rules for predicting the next gait state on the basis of the current and previous states. Another interesting approach [27], [28], [29] is to model a joint trajectory combining a vector field (VF) [28] and hybrid automata [29] in order to develop a more accurate bipedal robot. It is based on the assumption that human walking is a combination of different discrete sub-phases, and each sub-phase has a continuous dynamic state, so it can be modeled as a hybrid system. The VF is defined as a function of time and gives joint angle values for a particular joint at a given instant of time. First, a VF is designed for each person from his/her captured gait pattern, and then hybrid automata is used to generate joint trajectories for a humanoid robot, giving it a gait that is morphologically similar to the captured one. The goal of this research is to generate a bipedal gait that is similar to the captured gait.

2.3. Adversarial Examples

Adversarial examples are slightly modified versions of ordinary examples that cause unexpected recognition mistakes. A wide range of approaches has been proposed for the crafting of adversarial examples. Szegedy et al. [30] first introduced adversarial examples for deep neural networks. Their approach is based on a box-constrained optimization technique and is aimed at finding the smallest perturbation in the input space that causes the perturbed image to be classed as a predefined target label. Goodfellow et al. [31] presented a simple and computationally cheaper, yet robust, method for directly perturbing normal input by a small amount in the direction of the sign of the gradient at each pixel. Moosavi-Dezfooli et al. [32] proposed the DeepFool model, which is based on iterative linearization of the classifier and is used to generate minimal perturbations that are sufficient to change classification labels. Leveraging the DeepFool model, they developed a universal adversarial perturbation [33]. They showed the existence of a universal (image-agnostic) and very small perturbation vector that causes natural images to be misclassied with high probability. All of these approaches use a classifier in the network to generate the perturbed images. Since gait recognition systems use gait features such as body parts and GEIs, it is difficult to integrate gait recognition into an adversarial network at implementation. In addition, existing methods focus on a single image while we aim to enable application to video samples, which is more challenging.

2.4. Generative Adversarial Networks

A traditional generative adversarial network (GAN) [12] consists of two neural networks trained in opposition to one another. The input is a random noise vector, and the output is a fake image. The role of the generator is to generate images resembling real images while that of the discriminator is to distinguish real images from fake ones. The whole network is trained using a min-max objective function.

While many extensions proposed for GANs generate very natural images [10, 34, 35], few GAN-

based approaches have been proposed for video generation. Saito et al. [36] proposed a temporal generative adversarial network (TGAN) consisting of a temporal generator, an image generator, and a discriminator. The temporal generator generates a sequence of latent variables from a random variable. The image generator follows the temporal generator and produces the $t-th$ frame of the sequence. Tulyakov et al. proposed the MoCoGAN model for generating video sequences without a priming image [37]. The basic idea is to use motion and content. A content vector is sampled once from a Gaussian distribution and fixed. A recurrent neural network is used to sample and map a series of random variables for the motion subspace to a series of motions. Vondrick et al. [38] proposed a video (VGAN) model with two generators, one for background generation and the other for foreground generation. The input to both is a noise vector sampled from a Gaussian distribution. These methods are aimed at generating a video from random noise vectors while our aim is to modify a given gait so that it is incorrectly recognized by a gait recognition system.

Yan et al. [39] attempted to generate an articulated human motion sequence from a single image by using a conditional GAN. Their aim was to synthesize a video of a person from his/her skeleton and static image; therefore, it is not suitable for the gait anonymization problem, i.e., anonymizing a gait by modifying it.

3. Methodology

3.1. Definitions and Notations

Definition 1. *(Contour vector): the contour vector of a frame is a vector in which the elements are the coordinates of the pixels on the contour of the frame.*

Definition 2. *(Contour sequence): the contour sequence of a gait is the sequence of contour vectors of its frames.*

Definition 3. *(Noise contour): a noise contour is a vector that is added to the contour vector of a gait to anonymize that frame. It is generated with a noise generation network.*

Definition 4. *(Random noise): random noise is a sequence of noise contours to be added to the original gait to anonymize it.*

Definition 5. *(Random seed): a random seed is a Gaussian-distributed random vector used as input to the noise generation network to generate a noise contour.*

Notation	Meaning
x_i	$i-th$ frame of original gait
X	frame sequence of original gait, $X = [x_1, x_2, \ldots, x_t]$
y_i	contour vector of $i-th$ frame of original gait
Y	contour sequence of original gait, $Y = [y_1, y_2, \ldots, y_t]$
z_i	$i-th$ random seed
Z	sequence of random seeds, $Z = [z_1, z_2, \ldots, z_t]$
$\hat{y}_i$	contour vector of $i-th$ frame of anonymized gait
$\hat{Y}$	contour sequence of anonymized gait, $\hat{Y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_t]$
$\hat{z}_i$	$i-th$ contour vector of random noise
$\hat{Z}$	random noise, $\hat{Z} = [\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_t]$

Table 1: Notations used in paper.

The notations used throughout this paper are shown in Table 1

3.2. Overview of Proposed Method

To anonymize a gait, we need to modify its shape. Therefore, the contours of the gait's silhouette must be modified. To do this, we convert the contour of each silhouette into a vector in which the elements are the coordinates of the pixels on the contour (the contour vector) and modify this vector.

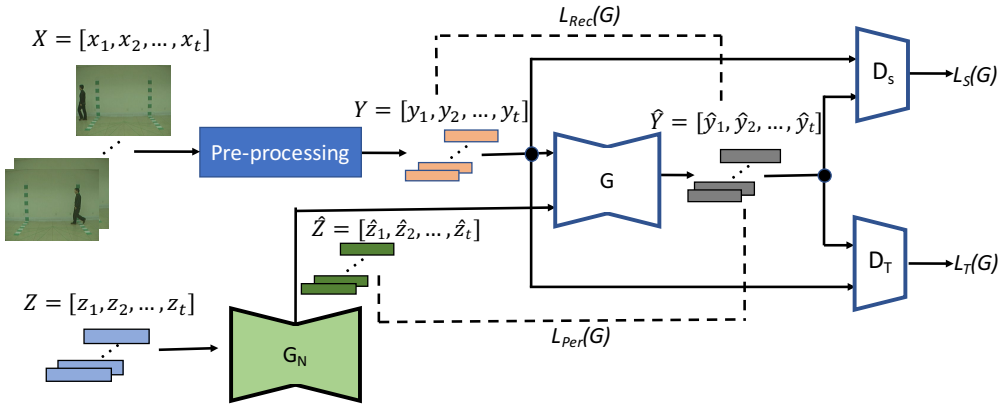
The model used comprises three steps.

Step 1 (Pre-processing): Extract the contour vectors from the frames of the gait. The length of the vectors is set to 4000, equivalent to 2000 pixels (zero padding is added at the end if needed).

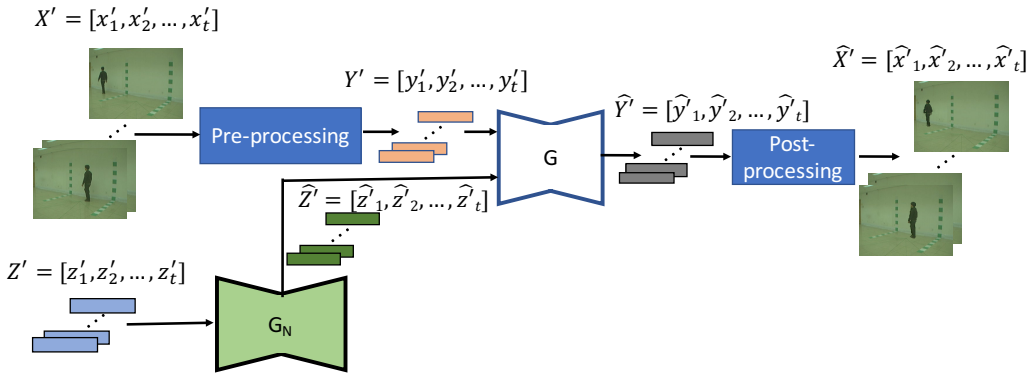
Step 2 (Contour vector modifying): Modify the contour sequence of the original gait, which is desired to anonymize.

Step 3 (Post-processing): Transfer the contour sequence of the anonymized gait to a binary anonymized gait. The binary anonymized gait is then colorized to obtain a color anonymized gait.

Step 1 corresponds to Step 1 in the method of Tieu et al. [1], but the input here is a color video instead of a binary video. Likewise, Step 3 corresponds to Step 3 in the method of Tieu et al.; however, the binary anonymized gaits are colorized to obtain color anonymized gaits, as explained in Subsection E. Our two main contributions are found



(a) Training phase: Only non-filled blocks are trained in this phase.



(b) Generation phase.

Figure 3: Overview of the proposed method.

in Step 2. The success of anonymization is improved by using *random noise* instead of a *noise gait* as previously used [1]. The naturalness of the anonymized gait is improved by using an ST-GAN containing a generator and two discriminators (a spatial discriminator and a temporal discriminator).

As illustrated in Fig.3a, the flow of the training phase is as follows. First, the original gait X is pre-processed to extract contour sequence Y. Next, random seeds Z are fed into noise generator G_N to obtain random noise $\hat{Z}$. Finally, the original contour sequence Y and random noise $\hat{Z}$ are fed into gait generator G to obtain $\hat{Y}$. Spatial discriminator D_S and temporal discriminator D_T are used to distinguish the shape and time continuity of the original and anonymized gaits, respectively.

In the generation phase (Fig.3b), contour sequence Y' of original gait X' and random noise $\hat{Z}'$ created from random seed Z' are passed through anonymized gait generator G to obtain the contour sequence of anonymized gait $\hat{Y}'$. This contour sequence is then post-processed to obtain the anonymized gait.

3.3. Noise Generation

As mentioned above, using a noise gait may reduce the success of anonymization if the gait recognition system confuses the original gait with the noise gait. In addition, it is not easy to find a noise gait that satisfies the requirements: it must be different from the original gait, its length sequence must be similar to that of the original gait, and it should have the same viewing angle as the original gait. To overcome the confusion problem and satisfy these requirements, we use the traditional GAN model [12] to generate noise in the gait distribution (*random noise*) by using a Gaussian-distributed random seed. For simplicity, here we

generate the first contour vector of random noise $\hat{z}_1$. The remaining contour vectors $(\hat{z}_2, \hat{z}_3, \ldots, \hat{z}_t)$ are copied from $\hat{z}_1$. In other words, we build a noise generation network to generate $\hat{z}_1$ from a random seed. The real gait is pre-processed before being fed into discriminator D_N as a positive example. The noise generation network is pre-trained before being assigned to the ST-GAN network.

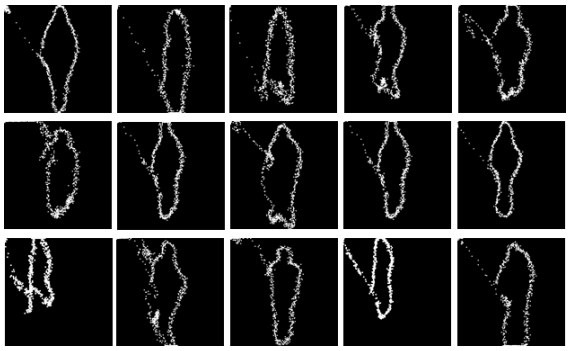


Figure 4: Example visualizations of generated noise.

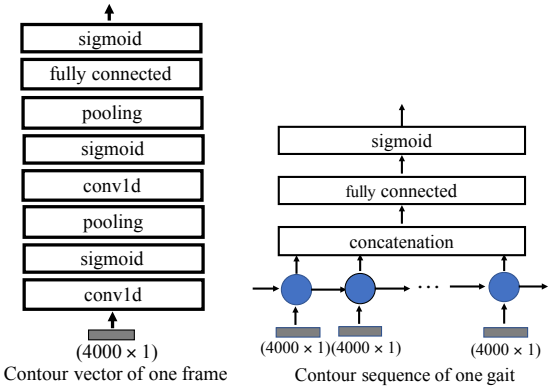
Because the random noise is generated from a Gaussian-distributed random seed and there are no constraints between the random noise and the original gait in the noise generation model, the random noise is different from the original gait. In the model of Tieu et al., the noise gait should have the same viewing angle as the original gait to maintain the naturalness of the original gait. However, in our ST-GAN model, the naturalness of the anonymized gait is adjusted using two discriminators, so the random noise does not need to have the same viewing angle as the original gait. Example visualizations of generated noise are shown in Fig. 4.

3.4. ST-GAN: Spatio-Temporal GAN

To obtain a high degree of naturalness in the anonymized gaits, we stacked two discriminators on the gait generator: a spatial discriminator network D_S and a temporal discriminator network D_T. They are schematically illustrated in Fig. 5.

The role of the spatial discriminator network is to distinguish the real frame from a generated frame. It does this by discriminating the shape of the real gait and the shape of the generated gait at each frame. The input to this network is a contour vector. The results are used to improve the naturalness of the shape of the anonymized gait. The architecture of this network includes 1-dimension convolution network followed by a fully connected layer and one sigmoid function on top.

The role of the temporal discriminator network



(a) Spatial discriminator network. (b) Temporal discriminator network.

Figure 5: Discriminator networks.

is to determine whether a generated gait moves smoothly. It does this by discriminating the temporal continuity of the real gait and that of the generated gait. To this end, we feed a contour sequence $\hat{Y}$ through a long short-term memory network. Since we want to measure the naturalness of the temporal continuity of the whole sequence, the outputs of each node are concatenated into one vector, and the sigmoid function is stacked on top of this model.

The structure of the gait anonymization generator can be designed in various ways. We used the same convolutional neural network architecture used by Tieu et al., in which the generator follows the encoder-decoder structure. There are two networks in the encoder to take the two inputs: the contour of the original gait and the random noise. These two networks are merged using a sum operator to convey a high-dimensional representation of the two inputs to the decoder.

Explicitly, in the GAN model, the training of G, D_S, and D_T is achieved by solving the min-max problem using a value function:

$$\begin{aligned}&\min\max L(G, D_S, D_T) = E_{y\sim p_y(y)}[logD_S(Y)]\\&+ E_{y\sim p_y(y), z\sim p_z(z)}[log(1 - D_S(G(Y, G_N(Z))))]\\&+ E_{y\sim p_y(y)}[logD_T(Y)]\\&+ E_{y\sim p_y(y), z\sim p_z(z)}[log(1 - D_T(G(Y, G_N(Z))))]\end{aligned} \tag{1}$$

In practice, this equation is solved by alternatively training discriminators D_S and D_T and anonymized gait generator G. In the first step, the generator network is fixed, and the two discrimina-

tors are trained by maximizing the two corresponding loss functions.

$$L(D_S) = E_{y\sim p_y(y)}[logD_S(Y)] + E_{y\sim p_y(y),z\sim p_z(z)}[log(1 - D_S(G(Y, G_N(Z))))] \quad (2)$$

$$L(D_T) = E_{y\sim p_y(y)}[logD_T(Y)] + E_{y\sim p_y(y),z\sim p_z(z)}[log(1 - D_T(G(Y, G_N(Z))))] \quad (3)$$

Next, the two discriminator networks are fixed, and the generator is trained by minimizing the two corresponding loss functions.

$$L_S(G) = E_{y\sim p_y(y),z\sim p_z(z)}[log(1 - D_S(G(Y, G_N(Z))))] \quad (4)$$

$$L_T(G) = E_{y\sim p_y(y),z\sim p_z(z)}[log(1 - D_T(G(Y, G_N(Z))))] \quad (5)$$

Because we want to retain the viewing angle and information about the action (here, "walking") of the original gait, generator G is designed to minimize the reconstruction loss by using the l_1 loss function:

$$L_{Rec}(G) = E_{y\sim p_y(y),z\sim p_z(z)}[\| Y - G(Y, G_N(Z)) \|_1] \quad (6)$$

To generate an anonymized gait that can fool a gait recognition system, we add a perturbation loss so that the generated gait is somewhat similar to the random noise.

$$L_{Per}(G) = E_{y\sim p_y(y),z\sim p_z(z)}[\| Z - G(Y, G_N(Z)) \|_1] \quad (7)$$

The gait anonymization generator is trained to minimize the four loss functions [(4), (5), (6), (7)]:

$$L(G) = L_S(G) + L_T(G) + L_{Rec}(G) + \alpha * L_{Per}(G) \quad (8)$$

where α is a hyperparameter used to control the trade-off between naturalness and success rate.

3.5. Colorizing

The methods for synthesizing colorized objects from original objects that have been reported [34], [35], [40]. However, these methods are aimed at generating static images, so they are not applicable to our work as we use video as input. The main problem is that they do not have constraints on consistency between frames or constraints on the relationships between scenes in consecutive frames. Here we present a method for colorizing binary anonymized gaits to obtain colorized anonymized gaits. For simplicity, we assume that the original video is recorded using a static camera. Let I_{Bg} be the background image, I_{Or} be the $t-th$ frame of the original gait, I_{An} be the $t-th$ frame of the color anonymized gait, and S_{Or} and S_{An} be the silhouettes of these frames, respectively. We denote the coordinate of a pixel as (i, j). Our colorizing problem is [now to compute I_{An} given I_{Bg}, I_{Or}, S_{Or}, and S_{An}.

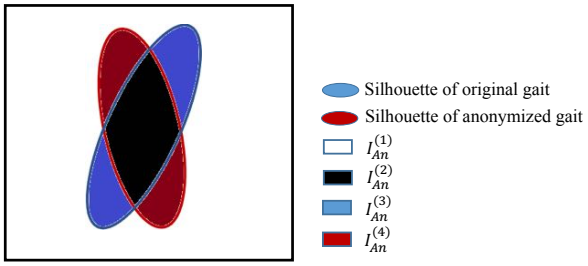


Figure 6: Four regions of color anonymized gait.

Color anonymized gait I_{An} consists of four regions, as illustrated in Fig.6. The first one, $I_{An}^{(1)}$, does not belong to the silhouette of the original gait or to the silhouette of anonymized gait $I_{An}^{(1)}$: $\{(i, j) \notin S_{Or} \cup S_{An}\}$. The second one, $I_{An}^{(2)}$, belongs to the silhouette of the original gait as well as to the silhouette of anonymized gait $I_{An}^{(2)}$: $\{(i, j) \in S_{Or} \cap S_{An}\}$. The third one, $I_{An}^{(3)}$, belongs to the silhouette of the original gait and not to the silhouette of anonymized gait $I_{An}^{(3)}$: $\{(i, j) \in S_{Or} \backslash I_{An}^{(2)}\}$. The fourth one, $I_{An}^{(4)}$, does not belong to the silhouette of the original gait and belongs to the silhouette of anonymized gait $I_{An}^{(4)}$: $\{(i, j) \in S_{An} \backslash I_{An}^{(2)}\}$. The algorithm used for colorizing the binary anonymized gait is given by

$$I_{An}(i, j) = \begin{cases} I_{Bg}(i, j), & \text{if } (i, j) \in I_{An}^{(1)} \\ I_{Or}(i, j), & \text{if } (i, j) \in I_{An}^{(2)} \\ I_{Bg}(i, j), & \text{if } (i, j) \in I_{An}^{(3)} \\ I_{Or}(i', j'), & \text{if } (i, j) \in I_{An}^{(4)} \end{cases} \quad (9)$$

where (i', j') is the pixel nearest (i, j).

Note that our colorizing algorithm is applied to each frame. Since we use the original frame for reference, the relationship between scenes in consecutive frames is preserved.

4. Evaluation

We experimentally evaluated our proposed method by using the CASIA-B gait dataset [13], in which there are 124 subjects in total, with 110 sequences (10 sequences for each of 11 viewing angles (0°,18°,...,180°)) for each subject. We divided the

124 subjects into non-overlapping groups. The first group contained 50 subjects and was used for training the two gait recognition systems. The second group contained 10 subjects (1100 sequences) and was used for training the G_N network. The third group contained 16 subjects (1760 sequences) and was used for training the ST-GAN model. Because the G_N and ST-GAN networks were not trained for each viewing angle, all sequences for each group were fed into the network. The fourth group contained 8 (880 sequences) subjects and was used for validation. The fifth group contained 40 subjects (equivalent 400 sequences for each viewing angle) and was used for testing.

We also compared the performance of the ST-GAN model with that of the model presented by Tieu et al. [1], which we used as the baseline for comparison. The evaluation metrics were success rate and naturalness for each viewing angle. We ran our model with several values of hyperparameter α and determined that when $\alpha = 0.3$ the success rate was good and the gaits looked natural. We thus used this value for the baseline comparison, which is described in Subsections A, B, and C. We discuss the effect of hyperparameter α on the success rate in Subsection D.

4.1. Generation Results

The effectiveness of the proposed method is illustrated by the visualized anonymized gaits shown in Figs. 8, 9, and 10. We further analyzed the role of the noise generation network G_N by removing it. The results are shown in Fig. 7.

From the results shown in these figures, we draw three conclusions

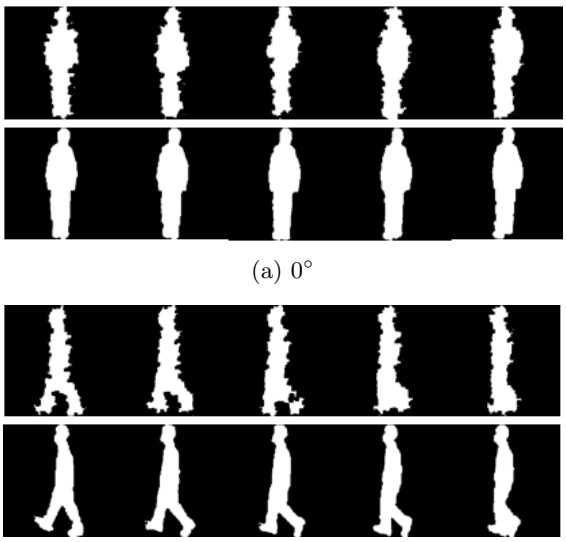
(1) The use of the ST-GAN model overcomes the problem of the anonymized gaits generated with the baseline method looking less realistic because of head distortion, especially at viewing angles of 0° and 180°, as shown in Figs. 8 and 9.

(2) The proposed method can generate colorized frames with color consistency between consecutive frames, as shown in Figs. 9 and 10.

(3) The noise generator plays an important role in making a generated gait more natural. This is evident in Fig. 7, which shows the results of gait synthesis using a Gaussian-distributed random seed, which is equivalent to removing the noise generator from the model.

4.2. Success Rate

The success rate is a measure of gait anonymization performance. It was calculated in the same way used by Tieu et al. It is typically stated as the ratio of the number of anonymized gaits that were not correctly identified and the total number of anonymized gaits. We measured it for two gait recognition systems (Zheng et al. [14]; Wu et al. (model MT) [15]) with top-1 and top-3 identification. Fig. 11a plots the success rates for the baseline and proposed methods as computed using Zheng's system while Fig. 11b plots those computed using Wu's system.



(a) 0°

(b) 90°

Figure 7: Silhouette of anonymized gaits generated using Gaussian-distributed random seed (top two rows) and using random noise in gait distribution (bottom two rows).

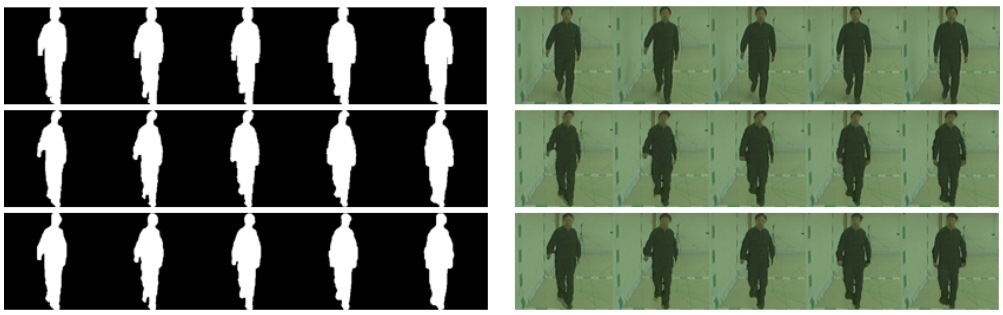
(1) The success rate with the proposed method for $\alpha = 0.3$ was higher than that for the baseline method for both gait recognition systems. This demonstrates that using random noise removes the identity information from the original gaits better than using a noise gait.

(2) The difference between the success rate for the two methods was higher for the side views because the anonymized gaits generated with the baseline method were less distorted at these viewing angles (from 72° to 108°).

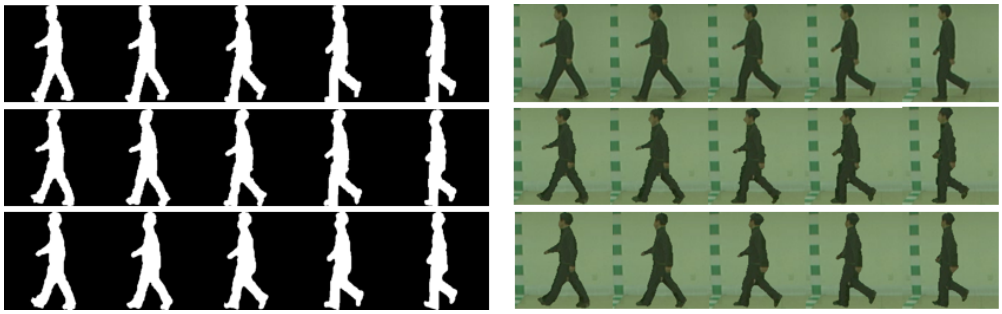
(3) The success rate with Zheng's system was higher than that with Wu's system because Wu's system is more robust.

4.3. Naturalness

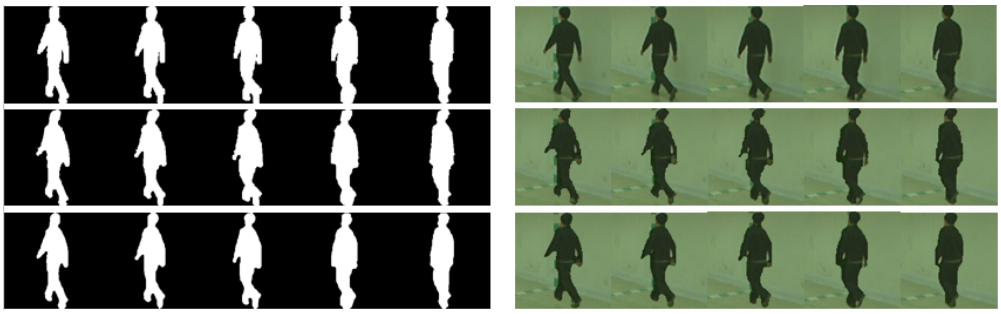
The naturalness of a gait encompasses two aspects: whether the shape of the gait looks human and whether the movement looks like a humanoid walking. To quantitatively compare the naturalness



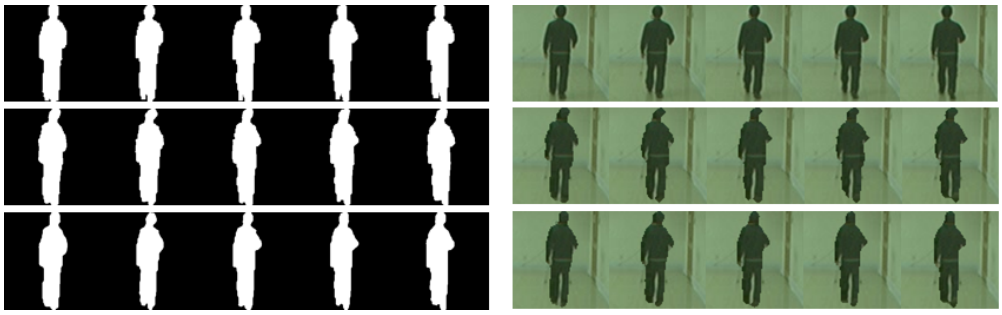
(a) 0°



(b) 90°



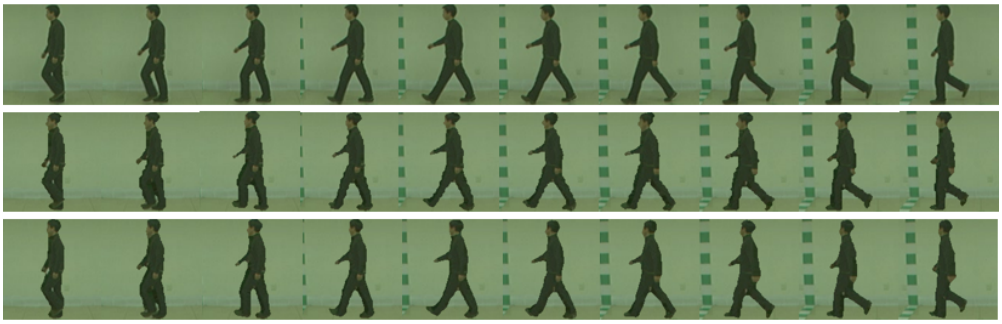
(c) 144°



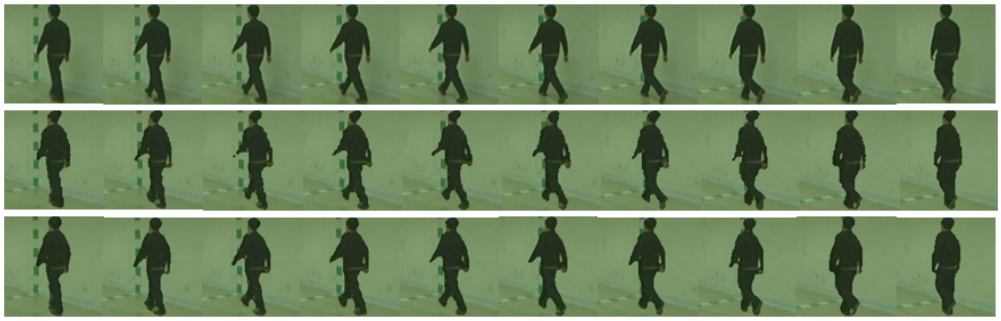
(d) 180°

Figure 8: Original and anonymized gaits generated with proposed and baseline methods for viewing angles of 0°, 90°, 144°, and 180°: top rows are original gaits, middle rows are anonymized gaits generated with baseline method, and bottom rows are anonymized gaits generated with proposed method.

(a) 36°



(b) 90°



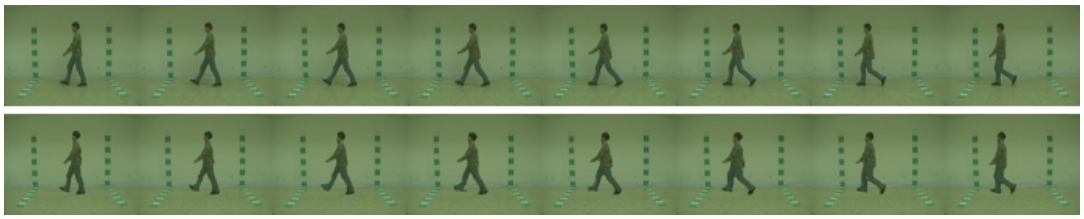
(c) 144°

Figure 9: Original and anonymized gaits generated with proposed and baseline methods for viewing angles of 36°, 90°, and 144°: top rows are original gaits, middle rows are anonymized gaits generated with baseline method, and bottom rows are anonymized gaits generated with proposed method.

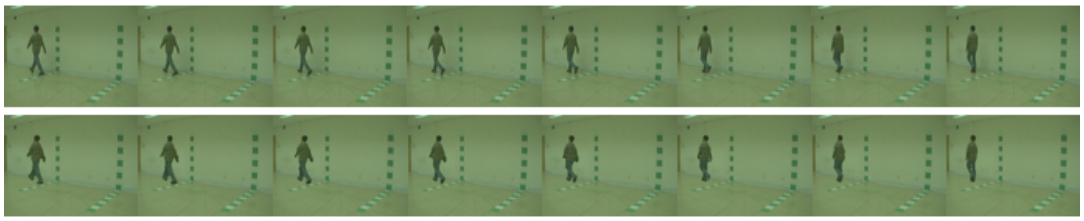
performance of our model with that of the baseline one, we conducted two kinds of evaluation on the color anonymized gait results: subjective evaluation by human volunteers and automatic evaluation by machine, which comes from the ideas of Cai et al. [41] and Walker et al. [42].

Subjective evaluation: The subjective evaluation was done using the mean opinion score (MOS), which has long been used for assessing the quality of media from the users perspective [10], [43] and which was used by Tieu et al. in their research. There were 20 volunteer evaluators with different backgrounds. Each volunteer viewed 60 color anonymized gait videos (half synthesized with the proposed method, haft synthesized with the baseline method), which were shown in random order. After watching each video, they rated the naturalness of the anonymized gait on a five-point scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). For each viewing angle, we computed the average score of each evaluator. From the results, which are plotted in Fig. 12, we draw three conclusions.

(a) 0°

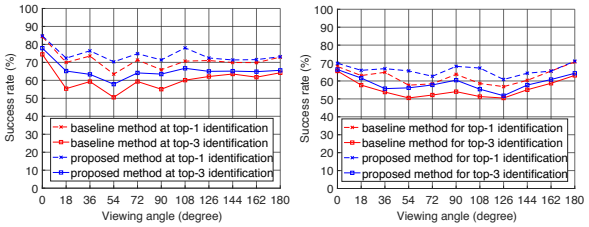


(b) 90°



(c) 126°

Figure 10: Original and anonymized gaits in original scene for three viewing angles 0°, 90°, and 126°: top rows are original gaits; bottom rows are anonymized gaits generated with proposed method.



(a) Zheng's system. (b) Wu's system.

Figure 11: Success rate comparison between proposed and baseline methods.

(1) The higher quartile1, quartile2, quartile3, and mean values for all viewing angles with the proposed method indicate that gaits anonymized with our method are more natural than those with the baseline method, from a human perspective.

(2) The greater differences at viewing angles of 0° and 180° mean that the distortion in the generated gaits at those viewing angles is mostly eliminated by our ST-GAN model.

(3) The tendency of the scores for viewing angles from 54° to 180° to be higher than those for angles from 0° to 36° indicates that an algorithm using information for the nearest pixel can colorize the body better than it can colorize the face.

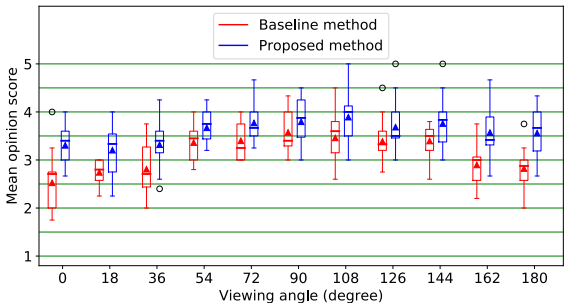


Figure 12: Mean opinion scores (triangle points represent mean values).

Automatic evaluation: Inspired by the ideas of Cai et al. [41] and Walker et al. [42], we investigated whether the anonymized gaits were sufficiently natural such that a pre-trained recognition network could recognize the object and action of the

object in the generated video. We did this by computing the probability that the recognition network recognizes the object in each frame as a "person" and the action of the object in the video sequence as "walking." The metrics representing these probabilities are the *frame score* and *video score*, respectively.

Frame score: This metric reflects the degree to which the shape of the object in each frame looks human. We used a pre-trained YOLO model (version 3) [44], an improvement of the version 2 model [45], which detects and classifies objects in an image, to compute the probability that an object in a frame is assigned to the "person" class. Fig. 13 shows the average frame score over all frames in the test set for the original gaits, the anonymized gaits generated by the baseline method, and the anonymized gaits generated by the proposed method.

Video score: This metric reflects the degree to which the movement of the gait in the video looks like a humanoid walking. We used pre-trained model ResNeXt-101 [46], which classifies human action in a video, to compute the probability that the action of the object in a video sequence is assigned to the "walking" class. Fig. 14 shows the average video score for three data gait sets: original data, anonymized gaits generated using a baseline model, and anonymized gaits generated using the proposed model.

From the results shown in these figures, we draw two conclusions.

(1) The higher frame and video scores for the proposed method than for the baseline method indicate that the anonymized gaits generated by our method in both still images (spatial domain) and video sequences (temporal domain) look more natural than those generated by the baseline method. This demonstrates the importance of the spatial and temporal discriminators in our model.

(2) In Fig. 13, we can observe that the variation of frame scores of the baseline method is not consistent with that of the original gait at the view 0° and 180° (the scores for the baseline were lower at these angles while those for the original gaits were slightly higher), while the variation of the frame scores of the proposed method is not. These demonstrate that the ST-GAN model minimizes the distortion in the gaits generated with the baseline method for the front viewing angles. This is consistent with our subjective evaluation and generation results.

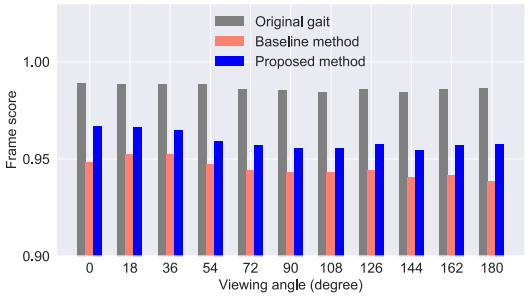


Figure 13: Frame score.

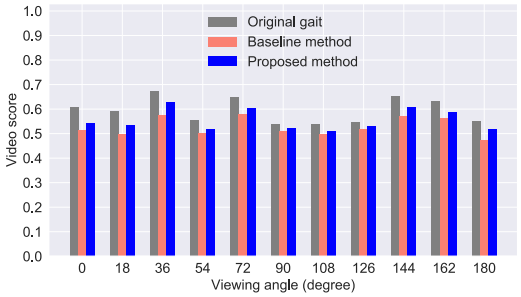
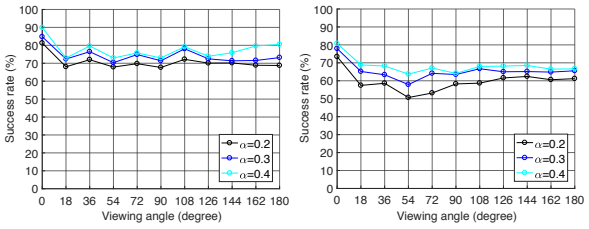


Figure 14: Video score.

4.4. *Effect of* α



(a) Success rate for top-1. (b) Success rate for top-3.

Figure 15: Effect of α on success rate against Zheng's system.

In this section, we analyzed the effect of hyperparameter α (with values of 0.2, 0.3, and 0.4) on the success rate (ratio between number of anonymized gaits not correctly identified and total number of anonymized gaits), the generation result and the naturalness of gaits generated with our method. The success rates against Zheng's and Wu's gait recognition systems are plotted in Figs. 15 and 16. The success rate increased with α for both systems.

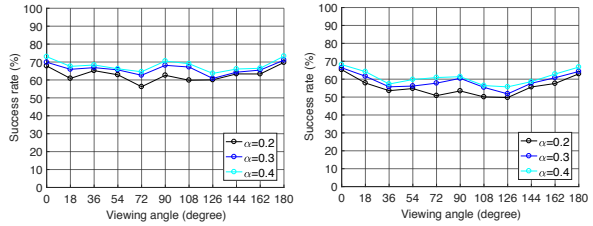
The effects of α on anonymized gaits generated with the proposed method for the three values of α are visualized in Fig. 17. Additionally, we visualized the difference between the original gait and the anonymized gaits by computing XOR images between the original gait and the anonymized gaits. As shown in Fig. 18, the difference between the

Table 2: Effect of α on frame score.

α	Viewing angle (degree)										
	0°	**18°**	**36°**	**54°**	**72°**	**90°**	**108°**	**126°**	**144°**	**162°**	**180°**
0.2	0.9764	0.9762	0.9751	0.9706	0.9699	0.9671	0.9672	0.9690	0.9679	0.9690	0.9694
0.3	0.9668	0.9664	0.9646	0.9590	0.9567	0.9556	0.9555	0.9574	0.9541	0.9568	0.9575
0.4	0.9558	0.9550	0.9533	0.9461	0.9386	0.9366	0.9363	0.9433	0.9418	0.9450	0.9464

Table 3: Effect of α on video score.

α	Viewing angle (degree)										
	0°	**18°**	**36°**	**54°**	**72°**	**90°**	**108°**	**126°**	**144°**	**162°**	**180°**
0.2	0.5738	0.5468	0.6575	0.5311	0.6268	0.5289	0.5174	0.5351	0.6239	0.6070	0.5335
0.3	0.5437	0.5353	0.6291	0.5197	0.6036	0.5225	0.5090	0.5321	0.6083	0.5887	0.5167
0.4	0.5185	0.5220	0.6016	0.4929	0.5956	0.5156	0.4969	0.5242	0.6004	0.5827	0.4983



(a) Success rate for top-1. (b) Success rate for top-3.

Figure 16: Effect of α on success rate against Wu's system.

original gait and anonymized gait increased with α (white pixels are where two silhouettes differ). This is consistent with the effect of α on the success rate.

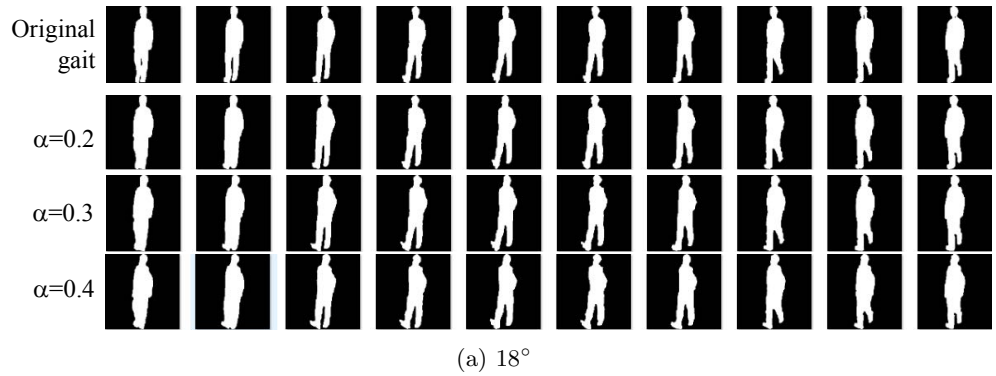
For naturalness, we computed the average frame score and average video score at each viewing angle for each value of α. As shown in Tables 2 and 3, the scores increased with a decrease in α for all viewing angles. This means that the success rate of the ST-GAN model is inversely proportional to naturalness and that changing the value of α can be used to control this trade-off.

5. Discussion and Conclusion

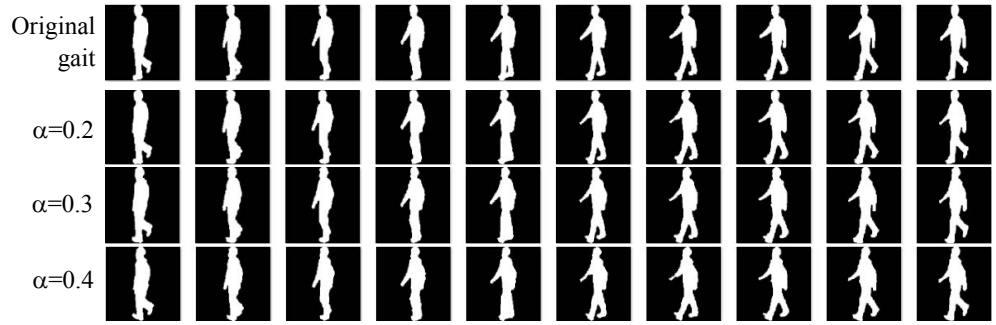
We proposed a spatio-temporal generative adversarial network model to generate anonymized gaits that appear natural as a means of preventing people in a video from being identified by a gait recognition system. The ST-GAN model generates a gait by adding random noise synthesized in the gait distribution to the original gait. Our evaluation demonstrated that this model can generate anonymized gaits with more naturalness and a higher success rate than a previous model. This means that the use of spatial and temporal discriminators in the proposed method results in spatial and temporal consistency while adding random noise to the original gait prevents the gait from being recognized by a gait recognition system. The anonymized gait maintains the originality of the action. It also retains the viewing angle by minimizing the reconstruction loss.

Two questions have been raised in anonymization research. (1) Is it possible to recover the original gait from an anonymized gait? (2) Is it possible to reproduce the identity of the original gait from an anonymized gait? Because random noise is used in our ST-GAN model, the noise added to the original gait is unknown. In addition, since our gait generator network, like Tieu's network [1], uses the non-reversible ReLU activation function, our model can be considered to be a one-way function. Moreover, the addition of random noise to the original gait means that there is no common formula for calculating the difference between the original gait and its anonymized gait. These properties make it impossible to recover the the original gait from an anonymized gait even though the properties (e.g, network architecture, network parameters) of the model are known.

As shown in Fig. 18, our method modifies all parts of the original gait, including the thigh and shank. The sizes of the thigh and shank are important features for identification using the gait pattern, especially for GEI-based recognition systems because a small modification of these features greatly changes the lower part of the body in the GEI image (see Fig. 2). Therefore, to reproduce the identity of original gait from an anonymized gait, the size of the thigh and shank must be recovered.

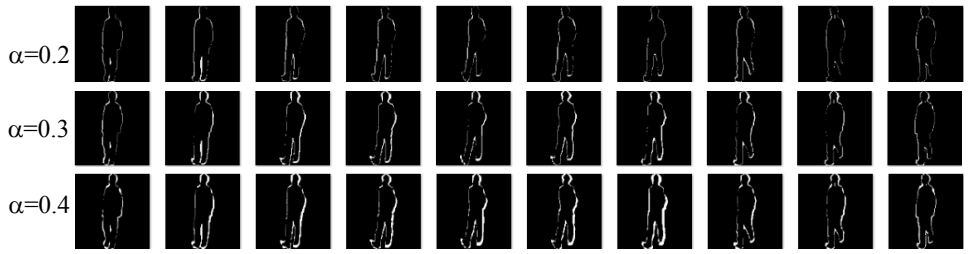


(a) 18°

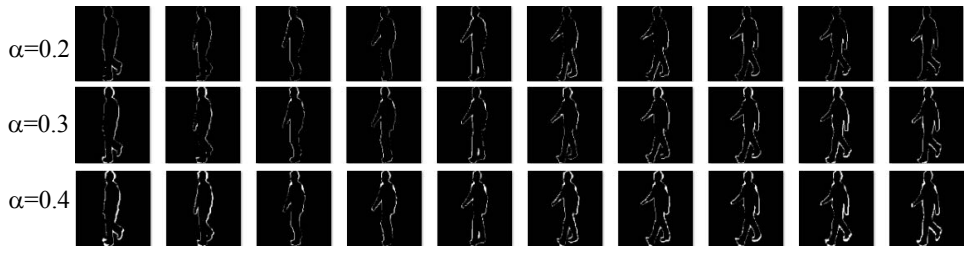


(b) 54°

Figure 17: Anonymized gaits generated for three values of α.



(a) 18°



(b) 54°

Figure 18: XOR images for three values of α.

However, as explained above, recovering the original gait from an anonymized gait is an impossible task, so reproducing the identity of the original gait from an anonymized gait is impossible.

Future work includes addressing three limitations of our method. First, the ST-GAN model generates an anonymized gait from silhouettes of the original gait, so the quality of the anonymized gait depends on the method used to extract the silhouettes. Second, our model aims to change the shape of the gait, however, the temporal information such as the speed of walking, the cycle of the gait and the po-

sitions of key joints (shoulders, hips, knees, ankles, etc.) also play an important role in gait anonymization. We believe that modifying the temporal information would increase the success rate. Third, our colorizing algorithm uses information for the nearest pixel. This works well for colorizing the body but not so well for colorizing the face because the face contains many parts (e.g., eyes, nose, and mouth).

6. Acknowledgments

This research was supported by JSPS KAKENHI Grant Number JP16H06302, JP18H04120, and JST CREST Grant Number JPMJCR18A6, Japan.

References

[1] N. D. T. Tieu, H. H. Nguyen, H. Q. Nguyen-Son, J. Yamagishi, I. Echizen, An approach for gait anonymization using deep learning, in: 2017 IEEE Workshop on Information Forensics and Security (WIFS), 2017, pp. 1–6.

[2] F. Z. Qureshi, Object-video streams for preserving privacy in video surveillance, in: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009, pp. 442–447.

[3] D. Chen, Y. Chang, R. Yan, J. Yang, Tools for protecting the privacy of specific individuals in video, EURASIP Journal on Advances in Signal Processing 2007 (1) (2007) 1–9.

[4] P. Agrawal, P. J. Narayanan, Person de-identification in videos, IEEE Transactions on Circuits and Systems for Video Technology 21 (3) (2011) 299–310.

[5] M. Ivasic-Kos, A. Iosifidis, A. Tefas, I. Pitas, Person de-identification in activity videos, in: 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, pp. 1294–1299.

[6] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, J. Clune, Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, in: Advances in Neural Information Processing Systems 29, 2016, pp. 3387–3395.

[7] C. Li, M. Wand, Precomputed real-time texture synthesis with markovian generative adversarial networks, in: The European Conference on Computer Vision (ECCV), 2016, pp. 702–716.

[8] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. DENOYER, M. A. Ranzato, Fader networks:manipulating images by sliding attributes, in: Advances in Neural Information Processing Systems 30, 2017, pp. 5967–5976.

[9] J. Zhu, P. Krähenbühl, E. Shechtman, A. A. Efros, Generative visual manipulation on the natural image manifold, in: The European Conference on Computer Vision (ECCV), 2016, pp. 597–613.

[10] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114.

[11] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, H. Li, High-resolution image inpainting using multi-scale neural patch synthesis, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4076–4084.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, 2014, pp. 2672–2680.

[13] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: 18th International Conference on Pattern Recognition (ICPR'06), Vol. 4, 2006, pp. 441–444.

[14] S. Zheng, J. Zhang, K. Huang, R. He, T. Tan, Robust view transformation model for gait recognition, in: 2011 18th IEEE International Conference on Image Processing, 2011, pp. 2073–2076.

[15] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, A comprehensive study on cross-view gait based human identification with deep cnns, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2) (2017) 209–226.

[16] M. Raj, V. B. Semwal, G. C. Nandi, Bidirectional association of joint angle trajectories for humanoid locomotion: the restricted boltzmann machine approach, Neural Computing and Applications 30 (6) (2018) 1747–1755.

[17] V. B. Semwal, J. Singha, P. Kumari, A. Chauhan, B. Behera, An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification, Multimedia Tools Appl. 76 (22) (2017) 24457–24475.

[18] V. B. Semwal, M. Raj, G. C. Nandi, Biometric gait identification based on a multilayer perceptron, Robotics and Autonomous Systems 65 (2015) 65–75.

[19] J. Wang, M. She, S. Nahavandi, A. Kouzani, A review of vision-based gait recognition methods for human identification, in: 2010 International Conference on Digital Image Computing: Techniques and Applications, 2010, pp. 320–327.

[20] C. Wan, L. Wang, V. V. Phoha, A survey on gait recognition, ACM Computing Survey 51 (5) (2018) 89:1–89:35.

[21] N. M. Bora, G. V. Molke, H. R. Munot, Understanding human gait: A survey of traits for biometrics and biomedical applications, in: 2015 International Conference on Energy Systems and Applications, 2015, pp. 723–728.

[22] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2) (2006) 316–322.

[23] K. Grochow, S. L. Martin, A. Hertzmann, Z. Popović, Style-based inverse kinematics, ACM Trans. Graph. 23 (3) (2004) 522–531.

[24] J. Chai, J. K. Hodgins, Constraint-based motion optimization using a statistical dynamic model, in: ACM SIGGRAPH 2007, 2007, pp. 1–9.

[25] D. Holden, J. Saito, T. Komura, A deep learning framework for character motion synthesis and editing, ACM Trans. Graph. 35 (4) (2016) 138:1–138:11.

[26] V. B. Semwal, N. Gaud, G. C. Nandi, Human gait state prediction using cellular automata and classification using elm, in: Machine Intelligence and Signal Analysis,

2019, pp. 135–145.
[27] V. B. Semwal, Data driven computational model for bipedal walking and push recovery, CoRR abs/1710.06548. arXiv:1710.06548.
URL http://arxiv.org/abs/1710.06548
[28] V. B. Semwal, C. Kumar, P. K. Mishra, G. C. Nandi, Design of vector field for different subphases of gait and regeneration of gait pattern, IEEE Trans. Automation Science and Engineering 15 (1) (2018) 104–110.
[29] V. B. Semwal, G. C. Nandi, Generation of joint trajectories using hybrid automate-based model: A rocking block-based approach, IEEE Sensors Journal 16 (14) (2016) 5805–5816.
[30] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations (ICLR), 2013, pp. 1–10.
[31] I. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: International Conference on Learning Representations (ICLR), 2015, pp. 1–11.
[32] S. M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: A simple and accurate method to fool deep neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582.
[33] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 86–94.
[34] J. Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251.
[35] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, in: Advances in Neural Information Processing Systems 30, 2017, pp. 465–476.
[36] M. Saito, E. Matsumoto, S. Saito, Temporal generative adversarial nets with singular value clipping, in: IEEE International Conference on Computer Vision (ICCV), 2017.
[37] S. Tulyakov, M.-Y. Liu, X. Yang, J. Kautz, Mocogan: Decomposing motion and content for video generation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[38] C. Vondrick, H. Pirsiavash, A. Torralba, Generating videos with scene dynamics, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 613–621.
[39] Y. Yan, J. Xu, B. Ni, W. Zhang, X. Yang, Skeleton-aided articulated motion generation, in: Proceedings of the 2017 ACM on Multimedia Conference, 2017, pp. 199–207.
[40] P. Isola, J. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[41] H. Cai, C. Bai, Y.-W. Tai, C.-K. Tang, Deep video generation, prediction and completion of human action sequences, in: The European Conference on Computer Vision (ECCV), 2018.
[42] J. Walker, K. Marino, A. Gupta, M. Hebert, The pose knows: Video forecasting by generating pose futures, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3352–3361.
[43] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, D. Hassabis, Parallel wavenet: Fast high-fidelity speech synthesis, in: International Conference on Machine Learning (ICML), 2018, pp. 3915–3923.
[44] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, CoRR abs/1804.02767.
[45] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
[46] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6546–6555.