An Experimental Consideration on Gait Spoofing

Yuki Hirose[1][a], Kazuaki Nakamura[2][b], Naoko Nitta[3] and Noboru Babaguchi[4]
[1] *Graduate School of Engineering, Osaka University, Suita, Osaka, 565-0871, Japan*
[2] *Faculty of Engineering, Tokyo University of Science, Tokyo, 125-8585, Japan*
[3] *School of Human Environmental Sciences, Mukogawa Women's University, Nishinomiya, Hyogo, 663-8558, Japan*
[4] *Institute for Datability Science, Osaka University, Suita, Osaka, 565-0871, Japan*

Keywords: Gait Recognition, Spoofing Attacks, Master Gait, Masterization, Gait Spoofing, Fake Gait Silhouettes, Multimedia Generation.

Abstract: Deep learning technologies have improved the performance of biometric systems as well as increased the risk of spoofing attacks against them. So far, lots of spoofing and anti-spoofing methods were proposed for face and voice. However, for gait, there are a limited number of studies focusing on the spoofing risk. To examine the executability of gait spoofing, in this paper, we attempt to generate a sequence of fake gait silhouettes that mimics a certain target person's walking style only from his/her single photo. A feature vector extracted from such a single photo does not have full information about the target person's gait characteristics. To complement the information, we update the extracted feature so that it simultaneously contains various people's characteristics like a wolf sample. Inspired by a wolf sample or also called "master" sample, which can simultaneously pass two or more verification systems like a master key, we call the proposed process "masterization". After the masterization, we decode its resultant feature vector to a gait silhouette sequence. In our experiment, the gait recognition accuracy with the generated fake silhouette sequences is increased from 69% to 78% by the masterization, which indicates an unignorable risk of gait spoofing.

1 INTRODUCTION

Recently, deep neural networks (DNNs) have been introduced in a wide range of research fields and achieved great success. One of the most DNN-benefitted research fields is biometrics such as face identification, voice authentication, gait recognition, and so on, whose performance has been drastically improved by DNNs. On the other hand, DNNs have also accelerated the performance of multimedia generation techniques. DNNs, or more specifically generative adversarial networks (GANs), can generate highly realistic facial images, speech data, and so on that mimics an actual person's biometric characteristics (Kammoun et al., 2022; Toshpulatov et al., 2021; Tu et al., 2019). These techniques are useful in some aspects (e.g., content creation and movie production), but they bring a risk of spoofing attacks against biometrics.

The risk of spoofing attacks against face identification and voice authentication has been widely discussed in the literature (Conotter et al., 2014; Nguyen et al., 2015; Chen et al., 2022; Shiota et al., 2015; Wang et al., 2019). There are a lot of existing studies proposing anti-spoofing methods for face and voice. However, for gait, which is a relatively novel biometric clue for human identification, the risk of spoofing attacks has not been well-analyzed yet. Although the development of gait recognition is still halfway, it is advantageous in that it can be applied to low-resolution videos where facial textures are not clearly observed. Hence, gait recognition will be more widely and complementarily used with face identification in the near future society. This means that exploring the (future) risk of gait spoofing is an important issue even if the performance and the spread of gait recognition are still limited at present.

So far, a few existing studies focus on the risk of gait spoofing (Gafurov et al., 2007; Hadid et al., 2012; Hadid et al., 2013). However, they do not assume fake gait generated by multimedia generation techniques; they only assume the cases where an attacker physically mimics the target person's walking style or physically wears the same clothes as the target person. Unlike them, in this paper, we focus on the risk of

[a] https://orcid.org/0000-0003-3370-0372
[b] https://orcid.org/0000-0002-4859-4624

Hirose, Y., Nakamura, K., Nitta, N. and Babaguchi, N.
An Experimental Consideration on Gait Spoofing.
DOI: 10.5220/0011661200003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 559-566
ISBN: 978-989-758-634-7; ISSN: 2184-4321

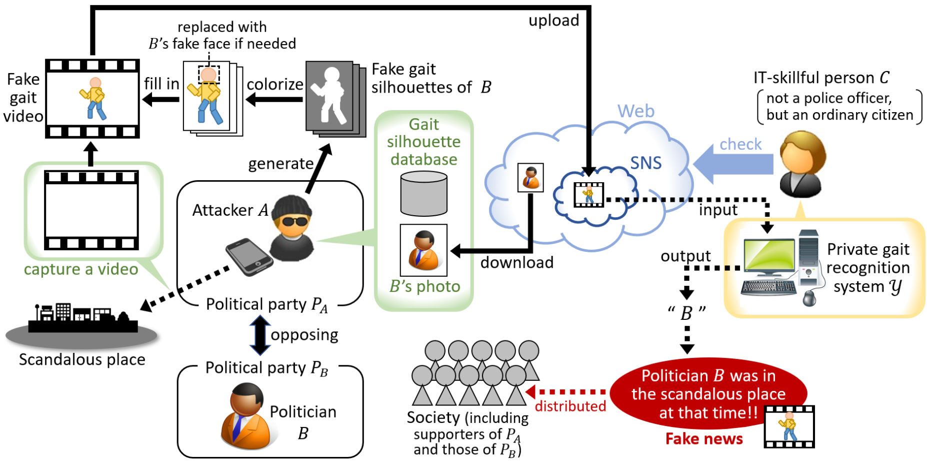

Figure 1: Assumed scenario of gait spoofing.

gait spoofing caused by DNN-based multimedia generation techniques. Specifically, in order to analyze whether a spoofing attack against gait recognition is practically possible or not, we propose a method for generating a sequence of a target individual's fake gait silhouettes using DNNs.

Figure 1 depicts a possible scenario of spoofing attacks against gait analysis. Suppose that there are two political parties P_A and P_B opposing each other. An attacker A joining the party P_A attempts to injure the reputation of a politician B in the party P_B by making his fake video. First, the attacker A captures a video of a certain scandalous place using his own device. In parallel, he generates a sequence of fake gait silhouettes that mimic or spoof B's walking style. Then, he colorizes the generated silhouettes and fills in them to the video captured above, which results in a fake gait video of B. In this step, he also generates a fake face of B and inserts it into the fake gait video if needed. This increases the reality of the fake video but is not necessarily needed when the video resolution is low. Note that the colorization process itself is not so important because human eyes cannot identify people by their body textures whereas automated systems just use silhouette information. At last, the attacker uploads the fake gait video onto the Web, particularly SNS. Nowadays, there are a lot of IT-skillful people who want to check the social behavior of politicians. For them, a gait recognition system can be a useful tool. One such person C, who is not a police officer but an ordinary citizen, checks the SNS and inputs the fake video into her private gait recognition system $\mathscr{Y}$. As a result, the politician B's behavior is fabricated and distributed as fake news even though the checker C has no malicious intent, as shown in Figure 1. The fake gait video may pass the modern deepfake detection systems because most of them are focusing only on faces. In other words, fake news fabricated with a fake face becomes more difficult to detect by combined with fake gait.

In the above scenario, we assume that the attacker A can use a single photo of the politician B as well as a large database of gait silhouettes that are not related to B nor $\mathscr{Y}$. Under this assumption, targets of the gait spoofing are not limited to politicians; not only other famous people such as celebrities and sports players but even ordinary citizens whose photo is on the Web or SNS could be a victim of this attack. The goal of the attacker is to generate a sequence of fake gait silhouettes that can be recognized as the victim by an unknown gait recognition system $\mathscr{Y}$.

The contributions of this paper are summarized as follows. First, this is the first work focusing on the method of DNN-based gait spoofing and analyzing its risk, to the best of our knowledge. Second, to achieve gait spoofing, we introduce the novel concept named "master gait", which is a master key-like gait data that can be accepted by multiple gait verification systems. The concept of master gait is utilized to complement the limited information of a single photo. We will explain its details in Section 3.

2 RELATED WORK

2.1 Spoofing Attacks Against Biometrics

Methods of spoofing attacks against face recognition and voice authentication have been actively studied for more than fifteen years. One of the simplest attack ways is a presentation attack. For a face recognition system equipped with a camera, an attacker can fool it by presenting a photo or a video of some valid user (Patel et al., 2016; Anjos et al., 2014). Similarly, for a voice authentication system equipped with a microphone, the attacker can fool it by replaying prerecorded voice of some valid user (Cheng and Roedig, 2022).

Conducting a presentation attack needs some "spoof data", i.e., a photo or pre-recorded voice of a valid user. For the face, spoof data can be retrieved from social networking services such as Facebook or Instagram in some cases (Kumar et al., 2017). In contrast, for the voice, spoof data cannot always be easily obtained. Hence, speech synthesis techniques are often exploited to make spoof data. A voice spoofing attack using such synthetic data is called a voice synthesis attack. The speech synthesis techniques used for this attack are divided into two categories: voice conversion (VC) (Liu et al., 2018) and text-to-speech (TTS) (Tu et al., 2019). VC is a technique to convert a source speaker's voice to a target speaker's voice without changing its linguistic information. TTS is a technique to convert an arbitral plaintext to spoken words with a certain target speaker's voice. Applying these techniques to the attacker's own voice or plaintext data to convert it to a valid user's voice, he can obtain spoof data (Kreuk et al., 2018; Zhang et al., 2021). Multimedia generation techniques are also exploited for face spoofing. Nowadays, not only 2D images/video but also 3D volume data of faces can be generated by GANs (Toshpulatov et al., 2021), which are at risk of being exploited for face spoofing (Galbally and Satta, 2015).

To defeat the above spoofing methods, anti-spoofing methods for face and voice authentication have also been actively studied. For instance, there is some previous work that tried to discriminate computer-generated or GAN-generated face images from real face images (Conotter et al., 2014; Nguyen et al., 2015). Recently, CNNs are often used for face anti-spoofing. For instance, Chen et al. found that the luminance component of face images is helpful to detect GAN-generated faces and proposed to use YCbCr images in addition to RGB images as input of a CNN (Chen et al., 2022). For voice, pop noise-based anti-spoofing methods are well-studied (Shiota et al., 2015; Wang et al., 2019). When a human speaks into a microphone, his/her breath sometimes reaches the microphone, which yields a pop noise. This is difficult to be naturally generated even with GANs. Therefore pop noises become a good clue for voice anti-spoofing.

As discussed above, a lot of spoofing and anti-spoofing methods have been studied for face recognition and voice authentication. In contrast, for gait recognition systems, spoofing attacks with synthetic data have not been studied yet. Thus, in this paper, we focus on gait spoofing utilizing CNN-based multimedia generation techniques.

2.2 Wolf Attacks by "Master" Samples

The purpose of spoofing attacks is to create a fake biometric sample (e.g. face) that is similar to a target person and dissimilar to all other people. However, researchers in the field of biometrics found that it is possible to create a single fake sample that is similar to two or more people. This is called a "wolf sample", and the attacks against biometric verification systems based on a wolf sample are called "wolf attacks" (Une et al., 2007). Suppose that there are two or more people who have their own biometric verification system. Each verification system is a two-class classifier that predicts whether an input biometric sample is from its owner or not. An attacker can simultaneously fool many of these systems by using a single wolf sample, where the wolf sample plays the role of a master key. Since this is a serious problem, methods of wolf attacks and their countermeasures have been studied. For instance, Ohki et al. evaluated the executability of wolf attacks against voice verification systems (Ohki et al., 2012). Nguyen et al. proposed a GAN-based method for generating a wolf sample against face recognition systems (Nguyen et al., 2020). They refer to the wolf samples generated by their method as "master faces".

Although there is no previous work focusing on wolf attacks against gait recognition systems, we believe that the characteristics of wolf samples are helpful to conduct gait spoofing. Thus, we introduce the concept of "master gait" in the proposed method, which is explained in the next section.

3 GAIT SPOOFING METHOD

3.1 Concept Definitions

In this paper, the term **"gait verification"** means a two-class classification task. A gait verification system only focuses on a single individual and predicts whether an input gait silhouette sequence $S = \{s_1, \cdots, s_m\}$ is genuinely captured from the individual or not. s_i is the i-th frame in S. Generally, a verification system first compresses S into a single feature map $f = F(S)$ by a compressor F, whose typical examples are Gait Energy Image (GEI) (Man and Bhanu, 2006) and Frequency Domain Feature (FDF)

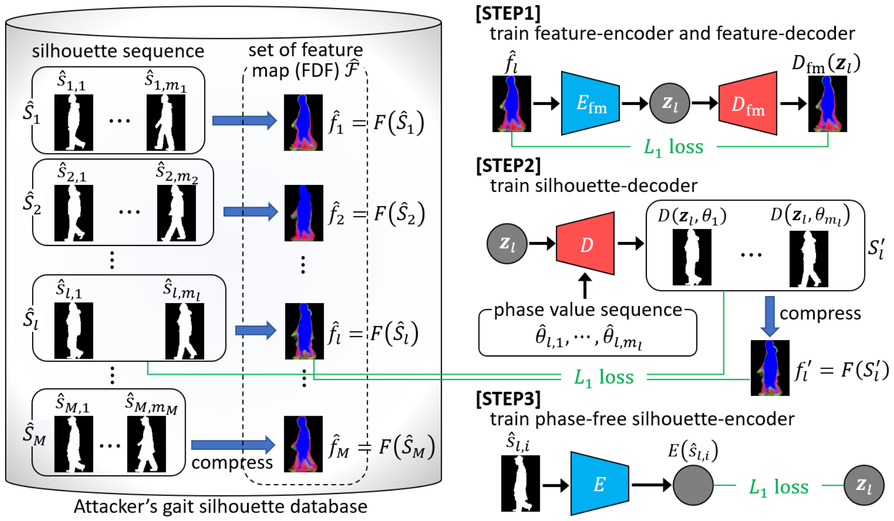
silhouette sequence
set of feature map (FDF) $\mathcal{F}$
$\hat{S}_1$ $\hat{s}_{1,1}$ … $\hat{s}_{1,m_1}$
$\hat{f}_1 = F(\hat{S}_1)$
$\hat{S}_2$ $\hat{s}_{2,1}$ … $\hat{s}_{2,m_2}$
$\hat{f}_2 = F(\hat{S}_2)$
$\hat{S}_l$ $\hat{s}_{l,1}$ $\hat{s}_{l,m_l}$
$\hat{f}_l = F(\hat{S}_l)$
$\hat{S}_M$ $\hat{s}_{M,1}$ … $\hat{s}_{M,m_M}$
compress
$\hat{f}_M = F(\hat{S}_M)$
Attacker's gait silhouette database
[STEP1]
train feature-encoder and feature-decoder
$\hat{f}_l$
E_{fm}
z_l
D_{fm}
$D_{\mathrm{fm}}(z_l)$
L_1 loss
[STEP2]
train silhouette-decoder
z_l
D
$D(z_l, \theta_1)$ … $D(z_l, \theta_{m_l})$
S_l'
phase value sequence
$\hat{\theta}_{l,1}, \cdots, \hat{\theta}_{l,m_l}$
compress
$f_l' = F(S_l')$
L_1 loss
[STEP3]
train phase-free silhouette-encoder
$\hat{s}_{l,i}$
E
$E(\hat{s}_{l,i})$
L_1 loss
z_l

Figure 2: Procedure for training gait silhouette decoder D and encoder E.

(Makihara et al., 2006). GEI is the pixel-wise average of $\{s_i\}$ while FDF is the pixel-wise Fourier coefficients calculated for $\{s_i\}$. For the feature map f, the verification system outputs a single score $\omega(f) = \omega(F(S)) \in [0,1]$, where the input S is classified as "genuine" if and only if $\omega(f) \geq 0.5$. Based on the above, we define a **"master gait"** as a feature map that has a score higher than 0.5 for two or more different individuals' gait verification systems.

In contrast, the term **"gait recognition"** means a multi-class classification task. A gait recognition system focuses on K different individuals ($K \geq 2$) and predicts which of them an input sequence S is captured from. A recognition system generally outputs K-dimensional score vector $\eta(f) = \eta(F(S)) \in [0,1]^K$. If the j-th element in $\eta(f)$ is larger than all the other elements, the system judges S is captured from the j-th individual.

3.2 Overview of the Proposed Method

The shape of a single gait silhouette s is determined by two factors: body shape (including the shape of clothes) and posture. Since walking is a periodic action, a human's posture in his/her one cycle of gait can be represented by a phase value $\theta \in [0, 2\pi]$. A human's body shape can be represented by a certain shape vector $\boldsymbol{z} \in \mathbb{R}^d$, where d is its dimensionality. Thus, a gait silhouette s can be determined by $\boldsymbol{z}$ and θ. Let D be a decoder that generates a silhouette image $s = D(\boldsymbol{z}, \theta)$ from $\boldsymbol{z}$ and θ.

The goal of gait spoofing is to obtain the optimal shape vector $\boldsymbol{z}^*$ that maximizes $\eta_b(f(\boldsymbol{z}))$, where $f(\boldsymbol{z}) = F(S(\boldsymbol{z}))$ is a feature map of a fake silhouette sequence $S(\boldsymbol{z}) = \{D(\boldsymbol{z}, \theta_i) | i = 1, \cdots, n\}$ generated by D. η is the score vector outputted by the checker C's gait recognition system $\mathscr{Y}$, and b is the ID of the spoofing target B. The phase sequence $\Theta = \{\theta_1, \cdots, \theta_n\}$ can be arbitrarily given. Note that attacker A does not know the network structure and the parameters of $\mathscr{Y}$ but he can guess F because there are only a few kinds of feature maps widely used for gait recognition. In this paper, we assume FDF as the feature map extractor F.

To obtain $\boldsymbol{z}^*$, the attacker can use a single photo of the target B, as we assumed in Section 1. Let p be the silhouette extracted from the photo. The simplest way to find $\boldsymbol{z}^*$ is training a phase-free shape encoder E that can extract $\boldsymbol{z}$ from $D(\boldsymbol{z}, \theta)$ as $\boldsymbol{z} = E(D(\boldsymbol{z}, \theta))$, by which we can get $\boldsymbol{z}^*$ as $\boldsymbol{z}^* = E(p)$. However, this strategy can hardly provide a good $\boldsymbol{z}^*$ in practice. This is because a single silhouette p does not have full information about B's gait characteristics. Hence, there is a certain extraction error $\Delta\boldsymbol{z}$ between $\tilde{\boldsymbol{z}} = E(p)$ and $\boldsymbol{z}^*$, i.e., $\tilde{\boldsymbol{z}} = \boldsymbol{z}^* + \Delta\boldsymbol{z}$. This $\Delta\boldsymbol{z}$ needs to be estimated for gait spoofing.

In summary, the process of gait spoofing is as follows, where we describe the ways to achieve steps (1) and (3) in Subsections 3.3 and 3.4, respectively.

(1) The attacker first trains D and E using his own gait silhouette database.

(2) With the trained E, he gets $\tilde{\boldsymbol{z}} = E(p)$ using a photo of the target person B.

(3) Next, he optimizes the $\tilde{\boldsymbol{z}}$ to $\boldsymbol{z}^* = \tilde{\boldsymbol{z}} - \Delta\boldsymbol{z}$ by estimating $\Delta\boldsymbol{z}$.

(4) Finally, he generates a fake silhouette sequence $\{D(\boldsymbol{z}^*, \theta_i) | i = 1, \cdots, n\}$ by the trained D and arbitrarily given $\Theta = \{\theta_1, \cdots, \theta_n\}$.

3.3 Training Process of Gait Silhouette Encoder and Decoder

Figure 2 depicts the proposed procedure for training D and E, which consists of three steps.

It is difficult to directly train the phase-free silhouette encoder E. Hence, we first train a feature map-level encoder E_{fm} and decoder D_{fm} as STEP1. To this end, for each sequence $\hat{S}_l = \{\hat{s}_{l,1}, \cdots, \hat{s}_{l,m_l}\}$ in the attacker's database, we compress it by F and obtain a feature map $\hat{f}_l = F(\hat{S}_l)$. Let $\hat{\mathcal{F}} = \{\hat{f}_l | l = 1, \cdots, M\}$ be a set of the obtained feature maps. Using $\hat{\mathcal{F}}$, we train an autoencoder, whose encoder and decoder parts are E_{fm} and D_{fm}, respectively. The loss function for STEP1 is

$$\begin{aligned}\mathrm{Loss}_1 &= \sum_l \left\| \hat{f}_l - D_{\mathrm{fm}}(\boldsymbol{z}_l) \right\| \\ &= \sum_l \left\| \hat{f}_l - D_{\mathrm{fm}}(E_{\mathrm{fm}}(\hat{f}_l)) \right\| . \qquad (1)\end{aligned}$$

Since $\boldsymbol{z}_l = E_{\mathrm{fm}}(\hat{f}_l)$ extracted from $\hat{f}_l$ by E_{fm} is phase-independent, it can be used as a phase-free shape vector of the silhouette image $\hat{s}_{l,i}$ for all $i \in \{1, \cdots, m_l\}$. Using them, we next train the silhouette-level decoder D as STEP2. The loss function for STEP2 is

$$\mathrm{Loss}_2 = \sum_l \left\{ \left\| \hat{f}_l - F(S'_l) \right\| + \frac{1}{m_l} \sum_{i=1}^{m_l} \left\| \hat{s}_{l,i} - D(\boldsymbol{z}_l, \hat{\theta}_{l,i}) \right\| \right\}, \qquad (2)$$

where $S'_l = \{D(\boldsymbol{z}_l, \hat{\theta}_{l,i}) | i = 1, \cdots, m_l\}$. The phase value $\hat{\theta}_{l,i}$ for each image $\hat{s}_{l,i}$ is calculated by our previous method (Hirose et al., 2019). Finally, we train the phase-free silhouette encoder E as STEP3, whose loss function is

$$\mathrm{Loss}_3 = \sum_l \left\| E(\hat{s}_{l,i}) - \boldsymbol{z}_l \right\| . \qquad (3)$$

3.4 Optimization of Gait Shape Vector Using Master Gait

Shape vector $\tilde{\boldsymbol{z}} = E(p)$ obtained by the encoder E includes an extraction error $\Delta\boldsymbol{z}$. Due to the error, $\tilde{\boldsymbol{z}}$ does not keep enough level of characteristics of the spoofing target B. Hence, the attacker has to emphasize the characteristics.

Here, suppose that the attacker trains a gait verification system for each individual in his database. Let $\mathcal{X}_j$ $(j = 1, \cdots, K)$ be the j-th individual's verification system and let ω_j be the output score of $\mathcal{X}_j$, where K is the number of individuals in the attacker's database. By inputting a feature map $\tilde{f} = D_{\mathrm{fm}}(\tilde{\boldsymbol{z}})$ into $\mathcal{X}_j$ for all j, the attacker can obtain a set of scores $\{\omega_j(D_{\mathrm{fm}}(\tilde{\boldsymbol{z}})) | j = 1, \cdots, K\}$. Since the target B is not any individual in the attacker's database, all of the scores are less than 0.5. However, if the database is large enough, it has some individuals somewhat similar to B. Hence, some elements in the score set are relatively larger than the other elements. This represents the target B's gait characteristics. In the proposed method, we emphasize the characteristics by perturbing $\tilde{\boldsymbol{z}}$ so that the relatively large elements become further larger and the other elements become smaller. The perturbation result is used as $\boldsymbol{z}^*$, which satisfies $\omega_j(D_{\mathrm{fm}}(\boldsymbol{z}^*)) > 0.5$ for two or more elements. This means $D_{\mathrm{fm}}(\boldsymbol{z}^*)$ behaves as a master gait, thus we refer to the above process as **"masterization"** of $\tilde{\boldsymbol{z}}$.

The concrete process of the masterization is as follows (see also Figure 3). First, the attacker trains $\mathcal{X}_1$, $\mathcal{X}_2, \cdots$, and $\mathcal{X}_K$ using his own database. Next, he inputs the shape vector $\tilde{\boldsymbol{z}} = E(p)$ into each $\mathcal{X}_j$ and obtains a score vector

$$\boldsymbol{\omega}(\tilde{\boldsymbol{z}}) = \begin{pmatrix} \omega_1(D_{\mathrm{fm}}(\tilde{\boldsymbol{z}})) \\ \vdots \\ \omega_K(D_{\mathrm{fm}}(\tilde{\boldsymbol{z}})) \end{pmatrix} \in [0,1]^K . \qquad (4)$$

Then, he finds top-N large elements in $\boldsymbol{\omega}(\tilde{\boldsymbol{z}})$ and makes a N-hot vector $\boldsymbol{h}_N = (h_{N,1} \cdots h_{N,K})^\top \in \{0,1\}^K$. Each element of $\boldsymbol{h}_N$ is set as 1 if and only if its corresponding elements in $\boldsymbol{\omega}(\tilde{\boldsymbol{z}})$ is included in the top-N ones. Other elements in $\boldsymbol{h}_N$ are set as 0. After that, he calculates the binary cross entropy between $\boldsymbol{\omega}(\tilde{\boldsymbol{z}})$ and $\boldsymbol{h}_N$, i.e.,

$$-\sum_{j=1}^{K} \Big[h_{N,j} \log\{\omega_j(D_{\mathrm{fm}}(\tilde{\boldsymbol{z}}))\} + (1 - h_{N,j}) \log\{1 - \omega_j(D_{\mathrm{fm}}(\tilde{\boldsymbol{z}}))\} \Big] , \qquad (5)$$

and minimizes it with respect to $\tilde{\boldsymbol{z}}$ to find the optimal $\boldsymbol{z}^*$. The minimization process is performed by a gradient descent algorithm. This process is equivalent to estimating $\Delta\boldsymbol{z}$ as $\Delta\boldsymbol{z} = \tilde{\boldsymbol{z}} - \boldsymbol{z}^*$.

4 EXPERIMENTS

4.1 Experimental Setup

To examine the performance of the proposed method, we conducted an experiment, where we used the OU-ISIR Gait Database (Makihara et al., 2012) as a dataset. This dataset has several different subsets, two of which called "treadmill-(A)" and "treadmill-(B)"

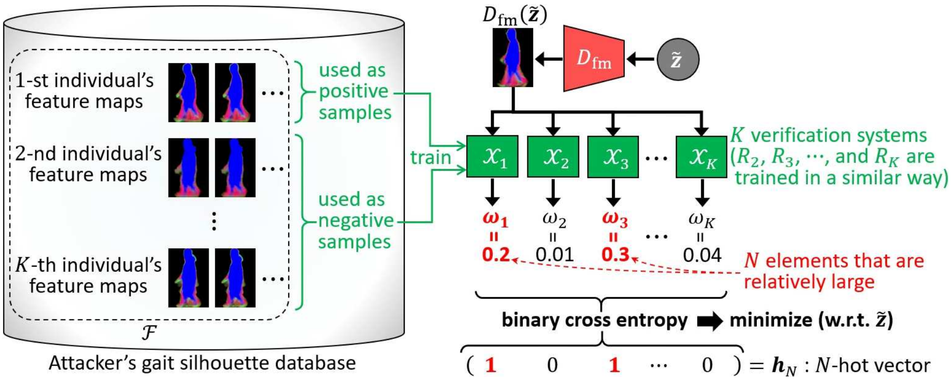
$D_{\mathrm{fm}}(\tilde{z})$
D_{fm}
$\tilde{z}$
1-st individual's feature maps
2-nd individual's feature maps
K-th individual's feature maps
used as positive samples
used as negative samples
train
$\mathcal{X}_1$
$\mathcal{X}_2$
$\mathcal{X}_3$
$\mathcal{X}_K$
K verification systems ($R_2, R_3, \cdots$, and R_K are trained in a similar way)
ω_1 = 0.2
ω_2 = 0.01
ω_3 = 0.3
ω_K = 0.04
N elements that are relatively large
binary cross entropy
minimize (w.r.t. $\tilde{z}$)
(1 0 1 ⋯ 0) = h_N : N-hot vector
$\mathcal{F}$
Attacker's gait silhouette database

Figure 3: Updating procedure of $\tilde{z}$ by masterization.

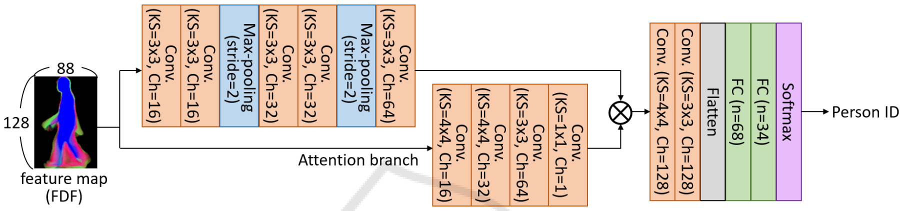

Figure 4: Network structure of gait recognition system $\mathscr{Y}$. "Conv" means a convolutional layer, where "KS" and "Ch" are its kernel size and num. of channels. "FC" means a fully-connected layer, where "n" is num. of units in it. "⊗" is pixel-wise multiplication.

were used. The treadmill-(A) consists of 612 gait silhouette sequences of 34 individuals (18 sequences per individual), while the treadmill-(B) consists of 2176 sequences of 68 individuals (32 sequences per individual). In our experiment, we used the treadmill-(A) to construct the checker C's gait recognition system $\mathscr{Y}$ as well as treated the treadmill-(B) as the attacker A's database.

We trained $\mathscr{Y}$ as a DNN, whose network structure is shown in Figure 4. After training $\mathscr{Y}$, we selected a single frame from each sequence in the treadmill-(A) and used it as the photo of the spoofing target B. From the photo, we generated a fake gait silhouette sequence and fed it into $\mathscr{Y}$ to check whether it is correctly recognized or not. We repeated this process for every frame in the treadmill-(A), and finally evaluated the recognition accuracy. Higher accuracy is desirable for the attacker since it means a high success rate of gait spoofing.

The silhouette-level encoder E and decoder D, the feature map-level encoder E_{fm} and decoder D_{fm}, and gait verification systems $\{\mathcal{X}_j\}$ were trained as a DNN with the attacker's database, namely the treadmill-(B). The network structures of these DNNs are shown in Figure 5, where we set the dimensionality of the shape vector $\boldsymbol{z} \in \mathbb{R}^d$ as 16, i.e., $d = 16$.

4.2 Results and Discussions

Figure 6 shows the result of the experiment under various settings of N. The red solid line indicates the recognition accuracy of $\mathscr{Y}$ when we fed it with the fake gait silhouette sequences generated by the proposed method. The blue dashed line indicates the recognition accuracy without the masterization. Compared to the dashed line, we obtained better accuracy when $N = 1$ and $N = 3$. This result demonstrates the effectiveness of the masterization as a technique for gait spoofing attacks. On the other hand, when $N \geq 5$, the gait recognition accuracy is seriously degraded by the masterization. The purpose of the masterization is to enlarge the relatively large elements in the score set $\{\omega_j(D_{\mathrm{fm}}(\tilde{z}))|j = 1, \cdots, K\}$. However, most of these scores are small since the spoofing target person is not any individual in the attacker's database. Therefore, even the fourth or fifth largest value in the score set is quite small, at least in this experiment. Enlarging such values is not effective for gait spoofing. Based on the above consideration, the best setting of N depends on the size of the attacker's database. We will try to find the relationship between them in our future work.

Figure 7 shows some examples of fake gait silhouettes generated by the proposed method as well as those without masterization. We can see that the generated silhouettes lose their shape when $N = 20$.

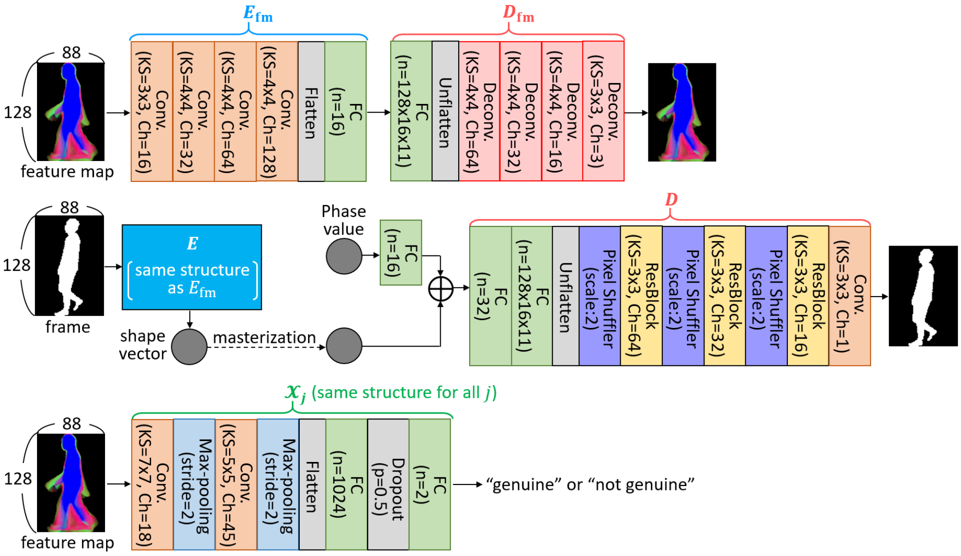

Figure 5: Network structure of E, D, E_{fm}, D_{fm}, and X_j. "Deconv" means a transposed convolutional layer. "$\oplus$" means concatenation operator.

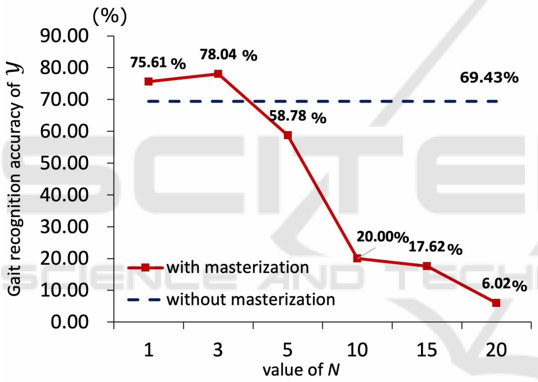

Figure 6: Recognition accuracy of gait recognition system $\mathscr{Y}$ under various settings of N.

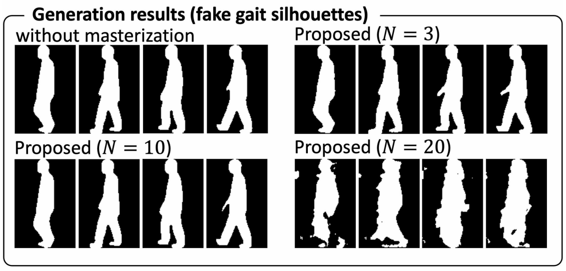

Figure 7: Examples of fake gait silhouettes generated by the proposed method.

On the other hand, the silhouettes generated with relatively small N (e.g., $N = 3$) can keep a natural appearance. These results indicate that the proposed method does not give any serious distortion to the generated fake gait silhouettes when N is appropriately set. In the case of "without masterization", arm regions in the silhouettes are not generated well. This is because the input photo does not have the information of the arm shape. Nevertheless, the proposed method with $N = 3$ can generate the arm regions more naturally. This is a reason why the proposed method achieves higher accuracy than that without masterization in Figure 6.

5 CONCLUSION

In this paper, we focused on the risk of gait spoofing and proposed a method for generating a fake gait silhouette sequence of a target person only from his/her single photo. In general, a single photo does not have full information about the gait characteristics of its owner. Hence, it is not enough for the attacker to just extract a feature vector from the photo. To solve this problem, we proposed to emphasize the gait characteristics of the target person by the masterization of the feature vector, before decoding it to a silhouette sequence. In our experiment, we found that the gait recognition accuracy with the generated fake sequences was increased from 69% to 78% by the masterization. This means the unignorable risk of gait spoofing. We will further investigate the possibility of gait spoofing as well as try to propose its countermeasure in our future work.

This work was supported by JSPS KAKENHI under Grant JP21J11069 and JST CREST under Grant JPMJCR20D3.

REFERENCES

Anjos, A., Chakka, M. M., and Marcel, S. (2014). Motion-based counter-measures to photo attacks in face recognition. *IET Biometrics*, 3(3):147–158.

Chen, B., Liu, X., Zheng, Y., Zhao, G., and Shi, Y. (2022). A robust gan-generated face detection method based on dual-color spaces and an improved xception. *IEEE Trans. on Circuits and Systems for Video Technology*, 32(6):3527–3538.

Cheng, P. and Roedig, U. (2022). Personal voice assistant security and privacy — a survey. *Proceedings of the IEEE*, 110(4):476–507.

Conotter, V., Bodnari, E., Boato, G., and Farid, H. (2014). Physiologically-based detection of computer generated faces in video. In *Proc. 21st IEEE Int'l Conf. on Image Processing*, pages 248–252.

Gafurov, D., Snekkenes, E., and Bours, P. (2007). Spoof attacks on gait authentication system. *IEEE Trans. on Information Forensics and Security*, 2(3):491–502.

Galbally, J. and Satta, R. (2015). Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models. *IET Biometrics*, 5(2):83–91.

Hadid, A., Ghahramani, M., Bustard, J., and Nixon, M. (2013). Improving gait biometrics under spoofing attacks. In *Proc. 17th IAPR Int'l Conf. on Image Analysis and Processing*, pages 1–10.

Hadid, A., Ghahramani, M., Kellokumpu, V., Pietikäinen, M., Bustard, J., and Nixon, M. (2012). Can gait biometrics be spoofed? In *Proc. 21st IAPR Int'l Conf. on Pattern Recognition*, pages 3280–3283.

He, Z., Wang, W., Dong, J., and Tan, T. (2020). Temporal sparse adversarial attack on sequence-based gait recognition. arXiv:2002.09674.

Hirose, Y., Nakamura, K., Nitta, N., and Babaguchi, N. (2019). Anonymization of gait silhouette video by perturbing its phase and shape components. In *Proc. 11th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1679–1685.

Kammoun, A., Slama, R., Tabia, H., Ouni, T., and Abid, M. (2022). Generative adversarial networks for face generation: A survey. *ACM Computing Surveys*, page 37 pages.

Kreuk, F., Adi, Y., Cisse, M., and Keshet, J. (2018). Fooling end-to-end speaker verification with adversarial examples. In *Proc. 2018 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, pages 1962–1966.

Kumar, S., Singh, S., and Kumar, J. (2017). A comparative study on face spoofing attacks. In *Proc. 2017 Int'l Conf. on Computing, Communication and Automation*, pages 1104–1108.

Liu, L., Ling, Z., Jiang, Y., Zhou, M., and Dai, L. (2018). Wavenet vocoder with limited training data for voice conversion. In *Proc. INTERSPEECH 2018*, pages 1983–1987.

Makihara, Y., Mannami, H., Tsuji, A., Hossain, M. A., Sugiura, K., Mori, A., and Yagi, Y. (2012). The ou-isir gait database comprising the treadmill dataset. *IPSJ Trans. on Computer Vision and Applications*, 4:53–62.

Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., and Yagi, Y. (2006). Gait recognition using a view transformation model in the frequency domain. In *Proc. European Conf. on Computer Vision*, pages 151–163.

Man, J. and Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(2):316–322.

Maqsood, M., Ghazanfar, M. A., Mehmood, I., Hwang, E., and Rho, S. (2022). A meta-heuristic optimization based less imperceptible adversarial attack on gait based surveillance systems. *Journal of Signal Processing Systems*, page 23 pages.

Nguyen, H., Nguyen-Son, H., Nguyen, T., and Echizen, I. (2015). Discriminating between computer-generated facial images and natural ones using smoothness property and local entropy. In *Proc. 14th Int'l Workshop on Digital Forensics and Watermarking*, pages 39–50.

Nguyen, H. H., Yamagishi, J., Echizen, I., and Marcel, S. (2020). Generating master faces for use in performing wolf attacks on face recognition systems. In *Proc. 2020 IEEE Int'l Joint Conf. on Biometrics*, page 10 pages.

Ohki, T., Hidano, S., and Takehisa, T. (2012). Evaluation of wolf attack for classified target on speaker verification systems. In *Proc. 12th Int'l Conf. on Control Automation Robotics and Vision*, pages 182–187.

Patel, K., Han, H., and Jain, A. K. (2016). Secure face unlock: Spoof detection on smartphones. *IEEE Trans. on Information Forensics and Security*, 11(10):2268–2283.

Shiota, S., Villavicencio, F., Yamagishi, J., Ono, N., Echizen, I., and Matsui, T. (2015). Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *Proc. of 16th Annual Conf. of the Int'l Speech Communication Association*, pages 239–243.

Toshpulatov, M., Lee, W., and Lee, S. (2021). Generative adversarial networks and their application to 3d face generation: A survey. *Image and Vision Computing*, 108:18 pages.

Tu, T., Chen, Y., Yeh, C., and Lee, H. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. In *Proc. INTERSPEECH 2019*, page 5 pages.

Une, M., Otsuka, A., and Imai, H. (2007). Wolf attack probability: A new security measure in biometric authentication systems. In *Proc. Int'l Conf. on Biometrics*, pages 396–406.

Wang, Q., Lin, X., Zhou, M., Chen, Y., Wang, C., Li, Q., and Luo, X. (2019). Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *Proc. IEEE Conf. on Computer Communications*, pages 2062–2070.

Zhang, Y., Jiang, F., and Duan, Z. (2021). One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28:937–941.