Temporal Sparse Adversarial Attack on Sequence-based Gait Recognition

Ziwen He[a,b], Wei Wang[b], Jing Dong[b], Tieniu Tan[b]

[a] *University of Chinese Academy of Sciences, Beijing, 100049, China*
[b] *Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China*

Abstract

Gait recognition is widely used in social security applications due to its advantages in long-distance human identification. Recently, sequence-based methods have achieved high accuracy by learning abundant temporal and spatial information. However, their robustness under adversarial attacks in an open world has not been clearly explored. In this paper, we demonstrate that the state-of-the-art gait recognition model is vulnerable to such attacks. To this end, we propose a novel temporal sparse adversarial attack method. Different from previous additive noise models which add perturbations on original samples, we employ a generative adversarial network based architecture to semantically generate adversarial high-quality gait silhouettes or video frames. Moreover, by sparsely substituting or inserting a few adversarial gait silhouettes, the proposed method ensures its imperceptibility and achieves a strong attack ability. The experimental results show that if only one-fortieth of the frames are attacked, the accuracy of the target model drops dramatically.

Keywords: Adversarial attack, gait recognition, temporal sparsity.

1. Introduction

Gait recognition is designed to automatically identify people according to their way of walking. Compared to traditional biometric information such as fingerprints or irises, gaits can be obtained at long distances without the cooperation of subjects. As a result, gait recognition is widely applied in remote visual surveillance solutions. In recent years, numerous gait recognition methods [1, 2, 3, 4, 5, 6, 7] have been proposed; they have achieved a high recognition accuracy. However, the robustness of gait recognition algorithms against malicious attacks in an open world has not been thoroughly studied.

In this paper, we investigate the robustness of gait recognition models subjected to adversarial attacks [8, 9]. Different from typical spoofing attacks [10, 11] on a gait verification system, adversarial attacks aim to imperceptibly (i.e., without incurring visual cues) disable the gait recognition model. Recently, adversarial attacks have been investigated including attacks on image classification [8, 9], object detection [12], face recognition [13], etc. However, for gait recognition, to the best of our knowledge, a meaningful attempt has not been reported, yet. A very likely reason is that the popular adversarial attack methods on image classification are not suitable to directly applied to gait recognition. Firstly, for sequence-based methods that take a sequence of silhouettes segmented from the original video as input, perturbations added on the source video do not work. This is due to the signal processing these approaches require. Secondly, even if attackers have access to modify the probes, adding a norm-constrained perturbation to the original gait silhouette destroys the imperceptibility. This is illustrated in the second row of Fig. 1.

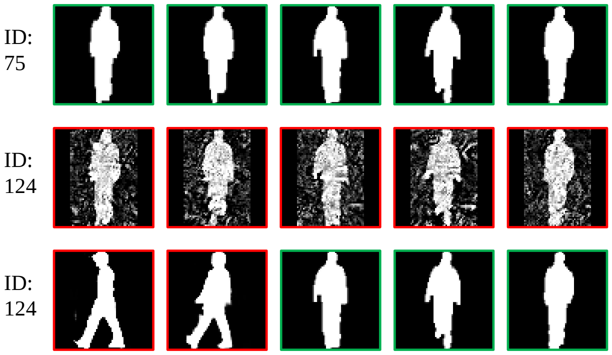


Figure 1: **Top row**: the original examples. **Middle row**: the perturbation-based adversarial examples. **Bottom row**: the temporal sparse adversarial examples. The red bounding box represents the modified example, while the green bounding box means the original example. The middle row directly transfers adversarial attack methods in image classification to gait recognition, causing all frames perturbed and imperceptibility decreased. The bottom row has only the first two frames modified. Besides, the modified frames maintain a gait appearance, so it is not easy to distinguish whether they are adversarial examples or not from human vision.

More specifically, this work focuses on the sequence-based methods [3, 5, 6]. Compared to template-based methods [1, 2], in which temporal information is difficult to preserve, sequence-based methods are better at extracting dynamic clues from silhouette frames with deep neural networks (DNNs). As a result, these methods have a higher gait recognition accuracy. However, the DNN-extracted temporal features may be vulnerable to adversarial attacks. To verify this hypothesis, we propose a novel temporal sparse adversarial attack method for the gait recognition system.

We have two primary intuitions that are illustrated in Fig. 1. Firstly, the input of gait recognition models is a sequence of silhouette frames, rather than a single image for the image classification models. Therefore, to better achieve its imperceptibility, only a few frames are modified in our attack. This ensures sparsity on the temporal domain. Secondly, motivated by unrestricted adversarial examples [14, 15], crafting an unrestricted adversarial gait silhouette via deformation better achieves imperceptibility than adding norm-bounded perturbations. Moreover, adversarial silhouettes generated by the proposed method can easily be extended to valid video frames. This enables a practical threat to gait recognition systems.

In summary, we propose a novel temporal sparse adversarial attack specifically designed to target gait recognition methods. The proposed method simultaneously achieves a high attack success rate and satisfactory imperceptibility. With the proposed method, we conduct extensive experiments to study the vulnerability of existing sequence-based gait recognition systems. The results indicate that sequence-based deep learning methods have little adversarial robustness despite their high accuracy.

2. Related work

2.1. Gait recognition

Gait recognition can generally be grouped into two categories, template-based [1, 2, 4, 16, 17] and sequence-based [3, 5, 6, 7]. The former category is composed of two main steps: template generation and matching. In the first step, human silhouettes are compressed into one template. For example, a large number of methods including GEINet [1] and GaitGAN [2] use the gait energy image (GEI) [18] as the template. In the second step, the similarity between pairs of templates is evaluated, e.g., by the Euclidean distance. The latter category directly captures dynamic clues from the sequence of silhouette frames. This category includes 3D CNN-based approaches [5], LSTM-based approaches [6], and GaitSet [3]. Currently, GaitSet achieves the state-of-the-art gait recognition results on the CASIA-B [19] dataset.

2.2. Adversarial attack

In this paper, we explore the vulnerability of gait systems under adversarial attack. Adversarial attack techniques [20, 21] have attracted increasing attention from security communities in recent years. To fool a deep neural network, attackers craft adversarial examples by maliciously adding designed perturbations to the inputs. The adversarial perturbations are typically restricted to a small norm, such as l_∞ [9], l_2 [22] or l_0 [23]. A series of methods have been proposed under this setting, such as FGSM [9], PGD [24], and MIFGSM [25].

In contrast, unrestricted adversarial examples [14, 15] are constructed entirely from scratch instead of perturbing existing data points by a small amount. Poursaeed et al. [15] manipulate stylistic and stochastic latent variables that are fed into the StyleGAN [26] to generate an unrestricted adversarial image to mislead a classification model. Similarly, we adopt a generative model to generate an adversarial high-quality gait silhouette. Here, we extend the approach to include the temporal domain. Instead of perturbing each frame, we sparsely generate adversarial frames to alert or insert into the original gait sequence.

In addition, additional approaches are available in the literature on video adversarial attacks [27, 28]. Wei et al. [27] utilize l_1 norm across frames to ensure the sparsity of adversarial perturbations on videos. A similar mask-based method is applied in our attack to control the sparsity. Chen et al. [28]

propose a new adversarial attack that appends a few dummy frames to a video clip and then adds adversarial perturbations only on these new frames. In our attack, we also explore the strategy of inserting frames into gait sequences. Both methods [27, 28] achieve a superior success rate on attacking temporal sequences. Nonetheless, their methods focus on the norm-bounded perturbations and cannot be directly transferred to the gait recognition task.

3. Methodology

3.1. Problem formulation

We have two types of adversarial attacks including *dodging attack* and *impersonation attack*. In the former attack, the attacker tries to have a gait sequence misidentified as any other arbitrary person, while in the latter attack, the attacker disguises a gait sequence as a specific authorized person. For clarity, we describe our method based on the dodging attack.

Let $\boldsymbol{X} \in \mathbb{R}^{N \times W \times H}$ denote a clean silhouette sequence, and $\boldsymbol{X}^* \in \mathbb{R}^{N \times W \times H}$ denote its adversarial sequence, where N is the number of frames, and W, H are the width and height for a specific frame, respectively.

The adversarial sequence $\boldsymbol{X}^*$ is the solution of the following objective function:

$$\arg\min_{\boldsymbol{X}^*} \lambda \mathcal{C}(\boldsymbol{X}, \boldsymbol{X}^*) + \mathcal{L}_{cos}(f(\boldsymbol{X}), f(\boldsymbol{X}^*)), \tag{1}$$

where λ is a weight that balances the two terms in the objective function and f is a gait recognition model that outputs the computed features of silhouette sequences. In addition, $\mathcal{L}_{cos}$ is the loss function to measure the cosine similarity between the ground truth sequence and the adversarial sequence,

$$\mathcal{L}_{cos}(f(\boldsymbol{X}), f(\boldsymbol{X}^*)) = \frac{f(\boldsymbol{X}) \cdot f(\boldsymbol{X}^*)}{\|f(\boldsymbol{X})\|_2 \|f(\boldsymbol{X}^*)\|_2}. \tag{2}$$

$\mathcal{C}(\boldsymbol{X}, \boldsymbol{X}^*)$ is a distortion measurement to evaluate the difference between the original sequence and its adversarial sequence. This proposed new measurement is defined as

$$\mathcal{C}(\boldsymbol{X}, \boldsymbol{X}^*) = \sum_{n \in \mathit{\Phi}} (o(\boldsymbol{X_n}) - o(\boldsymbol{X_n^*}))^2, \tag{3}$$

where $\mathit{\Phi}$ is a subset within the set of frame indices and o is an oracle to decide whether the image is a reasonable gait silhouette. $o(\cdot) = 1$ means the test example is a natural gait silhouette, and otherwise $o(\cdot) = 0$. As unrestricted adversarial examples [14, 15], the adversarial frames in our attack are expected to maintain a gait appearance even though they may have a large perturbation at the pixel-level.

We also can easily achieve an impersonation attack on a gait recognition system by modifying the objective function in Eq.(1). To craft an adversarial sequence misclassified as a target ID t is to minimize the following function,

$$\arg\min_{\boldsymbol{X}^*} \lambda \mathcal{C}(\boldsymbol{X}, \boldsymbol{X}^*) - \mathcal{L}_{cos}(f(\boldsymbol{X_t}), f(\boldsymbol{X}^*)), \tag{4}$$

where $\boldsymbol{X_t}$ is a sequence of ID t.

3.2. Temporal sparse attack

In this section, we propose a temporal sparse attack to solve the problem in Eq.(1). The pipeline of our attack is shown in Fig. 2.

To control the temporal sparsity, we denote the temporal mask as $\boldsymbol{M} \in \{\boldsymbol{0}, \boldsymbol{1}\}^{N \times W \times H}$. We let $\mathit{\Omega} = \{1, 2, ..., N\}$ be the set of frame indices, $\mathit{\Phi}$ be a subset within $\mathit{\Omega}$ having K elements randomly sampled from $\mathit{\Omega}$, and $\mathit{\Psi} = \mathit{\Omega} - \mathit{\Phi}$. The selection of $\mathit{\Phi}$ introduces randomization to make the crafted adversarial sequence more difficult to detect. If $n \in \mathit{\Phi}$, we set $\boldsymbol{M_n} = \boldsymbol{1}$, and

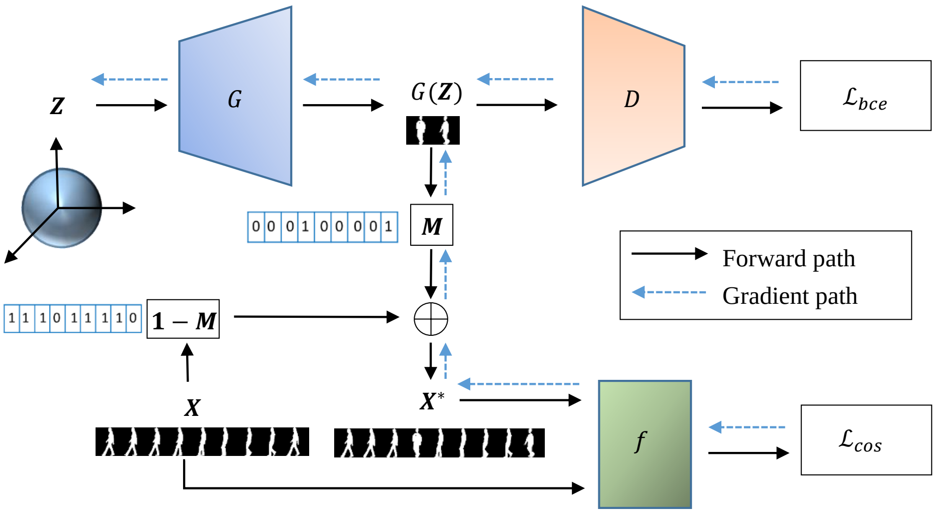


Figure 2: The pipeline of our attack approach. During attacking process, the latent vectors $\boldsymbol{Z}$ are the only parameter to be optimized. We start from a randomly sampled vector and iteratively optimize it through gradient backpropagation, shown as the blue dotted arrows. Finally, we feed the optimized latent vector into the generator G, and then put the obtained adversarial silhouettes into a source sequence to fool the target gait model f. The discriminator D is used to supervise high-quality gait silhouette image generation for imperceptibility on frame level. The mask $\boldsymbol{M}$ is used to achieve temporal sparsity for imperceptibility on the sequence level.

if $n \in \Psi$, $\boldsymbol{M_n} = \boldsymbol{0}$, where $\boldsymbol{M_n} \in \{\boldsymbol{0}, \boldsymbol{1}\}^{W \times H}$ is the n-th frame in $\boldsymbol{M}$. The sparsity is computed as $S = K/N$.

Denote the latent variable input into the generator as $\boldsymbol{z} \in \mathbb{R}^V$, where V is the dimension of each latent variable. G is a pre-trained generator on a gait silhouette dataset. Let $\mathcal{M}$ be the natural gait silhouette manifold in $\mathbb{R}^{W \times H}$. In most generative models, a simple random sample $\boldsymbol{z}$ drawn from the standard Gaussian distribution does not guarantee that $G(\boldsymbol{z}) \in \mathcal{M}$. To ensure the high quality of generated silhouettes, it must be in a region of the latent space with high probability. Inspired by Menon et al. [29], we replace the Gaussian prior on $\mathbb{R}^V$ with a uniform prior on $\sqrt{V}\mathbb{S}^{V-1}$, where $\mathbb{S}^{V-1} \subset \mathbb{R}^V$ is the unit sphere in the V dimensional Euclidean space.

For attacking a source sequence $\boldsymbol{X}$, we propose two methods to craft the corresponding adversarial sequence $\boldsymbol{X^*}$. One is the **frame-alteration attack**, which substitutes K frames in the source silhouettes with the generated adversarial ones. We first draw N vectors $\boldsymbol{Z} = [\boldsymbol{z_0}, \boldsymbol{z_1}, ..., \boldsymbol{z_{N-1}}]$ from the uniform prior $\sqrt{V}\mathbb{S}^{V-1}$ and then feed $\boldsymbol{Z}$ into the generator G. $\boldsymbol{X^*}$ is computed by the following equation,

$$\boldsymbol{X}^* = \boldsymbol{M} \cdot G(\boldsymbol{Z}) + (\boldsymbol{1} - \boldsymbol{M}) \cdot \boldsymbol{X}. \tag{5}$$

The other is the **frame-insertion attack**, which directly generate K adversarial frames and then insert them into the original sequence to obtain the adversarial sequence $\boldsymbol{X^*} \in \mathbb{R}^{(N+K)\times W \times H}$.

In our new measurement (3), the supervision of the oracle is of vital importance to improve the imperceptibility in the spatial domain. However, its binary output hinders the backpropagation and makes the objective function Eq.(1) difficult to optimize. A suboptimal solution is using a trained discriminator D to supervise the generated silhouettes. The discriminator outputs a value ranged in [0,1]. The value is close to one when the inputs are from the natural manifold; otherwise, the value near zero is output. We make sure $G(\boldsymbol{Z})$ keeps a high probability of sampling from the natural manifold by utilizing the binary cross entropy loss $\mathcal{L}_{bce}$. Thus, our objective function in

Algorithm 1 Temporal Sparse Adversarial Attack

```
Input: A gait recognition model $f$; a generator $G$; a discriminator $D$; a silhouette
    sequence $\boldsymbol{X}$; iterations $T$ and decay factor $\mu$; sparsity $S$; step size $\epsilon$; latent
    space dimension $V$; a hyper-parameter $\lambda$.
Output: An adversarial silhouette sequence $\boldsymbol{X}^*$.
$\boldsymbol{g}_0 = 0$; $\boldsymbol{X}_0^* = \boldsymbol{X}$; $\boldsymbol{Z}_0 \sim \sqrt{V}\mathbb{S}^{V-1}$, where $\mathbb{S}^{V-1}$ is the unit sphere space.
Compute the mask $\boldsymbol{M}$ according to the sparsity $S$, details are in the text;
for $t = 0$ to $T - 1$ do
    Input $\boldsymbol{Z}_t$ into the generator $G$ and obtain the images $G(\boldsymbol{Z}_t)$;
    Compute the adversarial sequence as $\boldsymbol{X}_t^* = \boldsymbol{M} \cdot G(\boldsymbol{Z}_t) + (1 - \boldsymbol{M}) \cdot \boldsymbol{X}$;
    Compute the loss $\mathcal{L}$ in Eq.(6);
    Update $\boldsymbol{g}_{(t+1)}$ by accumulating the velocity vector in the gradient direction
    as Eq.(8);
    Update $\boldsymbol{Z}_{(t+1)}$ by applying the clipped gradient as Eq.(9);
end for
return $\boldsymbol{X}^* = \boldsymbol{M} \cdot G(\boldsymbol{Z}_T) + (\mathbf{1} - \boldsymbol{M}) \cdot \boldsymbol{X}$.
```

Eq.(1) is equal to optimizing $\boldsymbol{Z}$ to maximize the following objective function:

$$\mathcal{L} = -\mathcal{L}_{cos}(f(\boldsymbol{X}), f(\boldsymbol{X^*})) - \lambda\mathcal{L}_{bce}(D(G(\boldsymbol{Z})), \mathbf{1}), \tag{6}$$

where λ is a hyper-parameter to establish a trade-off between two terms, and the binary cross entropy loss is obtained by

$$\mathcal{L}_{bce}(D(G(\boldsymbol{Z})), \mathbf{1}) = -\sum_{i=1}^{N} ln(D(G(\boldsymbol{Z}_i)). \tag{7}$$

By performing a gradient ascent in the latent variable space of the generator, the corresponding $\boldsymbol{Z}$ that maximizes the final objective function in Eq.(6) can be found. Without loss of generality, we adopt the MIFGSM [25] to attack f as follows:

$$\boldsymbol{g}_{(t+1)} = \mu \cdot \boldsymbol{g}_{(t)} + \frac{\bigtriangledown_{\boldsymbol{Z}_{(t)}} \mathcal{L}}{\| \bigtriangledown_{\boldsymbol{Z}_{(t)}} \mathcal{L} \|_1}, \tag{8}$$

$$\boldsymbol{Z}_{(t+1)} = \boldsymbol{Z}_{(t)} + \epsilon \cdot sign(\boldsymbol{g}_{(t+1)}), \tag{9}$$

where μ is the decay factor, ϵ is the step size, and t represents the t-th iteration. Algorithm 1 shows the proposed temporal sparse adversarial attack (shown as the frame-alteration attack).

3.3. Video generation

The above generation process only considers silhouette images. There are other possible attack points in a biometric recognition system, as shown in Fig.3. We hypothesize that one can access the transmission channel to repeat the previously recorded gait video on the channel. Under this hypothesis, we extend our method to the generation of a valid video. This can be regarded as a pixel-to-pixel image generation task.

Denote the source silhouette as $\boldsymbol{X} \in \mathbb{R}^{N \times W \times H}$ and the source video frame as $\boldsymbol{I} \in \mathbb{R}^{N \times W \times H \times C}$, where C denotes the color channel. We train a pix-to-pix generator G_p with paired data $(\boldsymbol{I} \times \boldsymbol{X}, \boldsymbol{X})$. In the attack process, we feed the adversarial silhouette $\boldsymbol{X^*}$ into the generator. The generated image $G_p(\boldsymbol{X}^*)$ is supposed to contain a subject; there is no background scene. We paste the generated image on the background image I_b with the formulation $I_a = I_b \times (\mathbf{1} - \boldsymbol{X}) + G_p(\boldsymbol{X}^*)$. Then, we insert the obtained adversarial frame into the source video or substitute a frame with it to generate the fake video. Though this method does not ensure spatial-temporal continuity between the adjacent real frames and adversarial frames, the modified frames are imperceptible

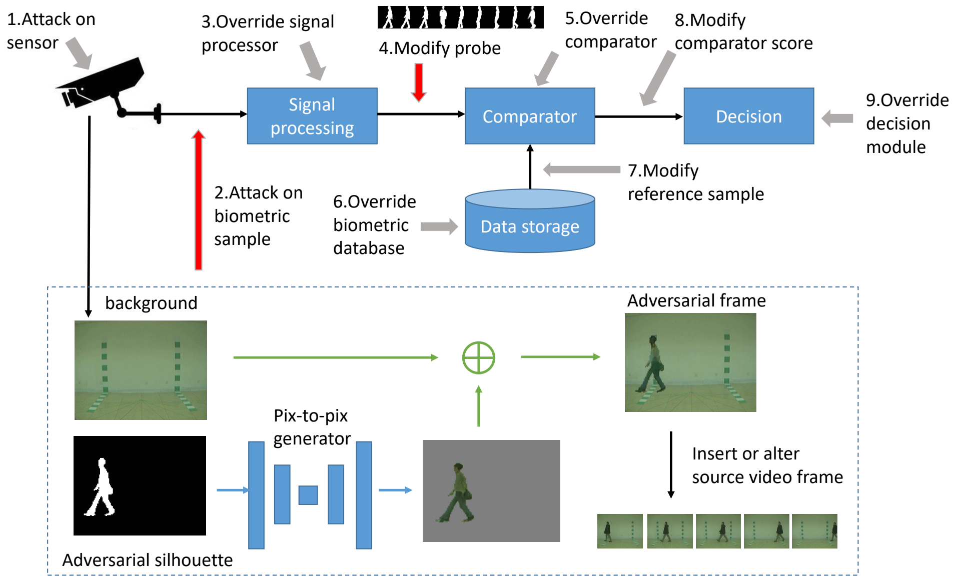

1.Attack on sensor
3.Override signal processor
4.Modify probe
5.Override comparator
8.Modify comparator score
Signal processing
Comparator
Decision
9.Override decision module
7.Modify reference sample
2.Attack on biometric sample
6.Override biometric database
Data storage
background
Adversarial frame
Pix-to-pix generator
Insert or alter source video frame
Adversarial silhouette

Figure 3: Vulnerability of a biometric recognition system. This figure is inspired by Jia et al. [10]. Our method can be applied in both the 2nd and 4th step. To attack on biometric samples, we train a generator to translate an adversarial silhouette image into a valid video frame.

due to the temporal sparsity. On the other hand, the modified frame keeps the same background, which is enough to deceive segmentation algorithms such as background difference. Moreover, some sequence-based models, like GaitSet, are flexible and capable of containing non-consecutive silhouettes in input sets. Thus, in these scenarios our temporal sparse adversarial video is not easy to detect; our video can cause a real threat to practical applications.

4. Experiments

In this section, we conduct experiments to explore the vulnerability of gait recognition models under our temporal sparse attack. First, we specify the experimental settings in Sec. 4.1. Then, we test the adversarial robustness of the state-of-the-art gait recognition model via the proposed method. The white-box attack is presented in Sec. 4.2 and a cross-dataset validation is presented in Sec. 4.4. We also perform a black-box attack to investigate whether adversarial examples of sequence-based models can transfer to template-based models in Sec. 4.3. Moreover, we provide a comparison of the proposed method with existing perturbation-based methods to demonstrate the superiority of our method in Sec. 4.5. Finally, we make a further analysis of the proposed method in Sec. 4.6.

4.1. Setup

Datasets. We conduct experiments on two datasets, CASIA-A [19] and CASIA-B [19]. CASIA-A consists of 20 subjects, and each subject has 12 image sequences, 4 sequences for each of the three directions, i.e. parallel, 45 degrees and 90 degrees to the image plane. Each sequence is labeled with 'mm_n', where 'mm' represents direction and 'n' is sequence number. For example, 4 parallel sequences are labeled with 00_1, 00_2, 00_3, 00_4, respectively. CASIA-B is a widely used gait dataset that contains 124 subjects (labeled 001-124) with 11 different viewing angles and 10 sequences per subject for each view. The 10 sequences contain three walking conditions: six sequences are in the normal walking state (NM 1-6), two sequences contain walking subject wearing coats (CL 1-2), and two sequences contain subject carrying bags (BG 1-2). We mainly use CASIA-B for evaluation and CASIA-A for cross-dataset validation.

Since our target gait models are trained on the first 74 subjects (labeled 1-74) and tested on the remaining 50 subjects of CASIA-B, we follow this

Table 1: The dataset setting.

CASIA-B	training set		ID: 001-074, nm01-nm06,bg01-bg02,cl01-cl02
	gallery set		ID: 075-124, nm01-nm04
	probe set	probeNM	ID: 075-124, nm05-nm06
		probeBG	ID: 075-124, bg01-bg02
		probeCL	ID: 075-124, cl01-cl02
CASIA-A	gallery set		ID:all, 00_4, 45_4, 90_4
	probe set	probe 0°	ID:all, 00_1, 00_2, 00_3
		probe 45°	ID:all, 45_1, 45_2, 45_3
		probe 90°	ID:all, 90_1, 90_2, 90_3

setting to attack the last 50 subjects (labeled 75-124). For each subject, the first four sequences of the NM condition (NM 1-4) are kept in the gallery to test the recognition accuracy. All of the frames in a specific view and walking condition are used as a sequence for the test. For the cross-dataset validation, we use the whole CASIA-A. The first three sequences of each angel are in the probe and the fourth sequence is in the gallery. The setting is summarized in Table 1.

Metrics. We quantitatively evaluate the vulnerability of gait systems by assessing the *accuracy* in the dodging attack and the *success rate* for an impersonation attack. We also provide some visualization results to qualitatively evaluate the imperceptibility.

Gait recognition compares the feature similarities between probe and gallery samples to identify a person. Thus, we report the average Rank-1 recognition accuracy to show the effectiveness of attacks since a strong dodging attack can largely decrease the recognition accuracy. The accuracy is averaged on all gallery views, and the identical views are excluded. For example, when testing with CASIA-B, the accuracy of the probe view 90° is averaged on 10 gallery views, excluding gallery view 90°. For cross validation on CASIA-A, the accuracy is averaged on the 3 gallery views, and the identical views are included.

For an impersonation attack, a successful attack means the probe has a highest feature similarity with the target identity among all the gallery samples. The success rate represents the proportion of successful adversarial examples targeted the gait recognition model. A higher success rate demonstrates a stronger impersonation attack.

Implementation details. If not specifically mentioned, we conduct the dodging attack. For the impersonation attack, we randomly choose the target ID and select one sequence of targeted ID as $\boldsymbol{X_t}$ in Eq.(4).

Our attack is based on MIFGSM [25], and the hyperparameters are set as follows: the iterations are 100, $\lambda = 10^{-5}$ in Eq.(6), $\mu = 1.0$ in Eq.(8), $\epsilon = 0.1$ in Eq.(9). For the mask $\boldsymbol{M}$, we let the set $\mathit{\Phi}$ be $\mathit{\Phi} = \{1, 2, ..., K\}$, which means we simply alter the first K frames with adversarial ones. K is computed according to the needed sparsity.

Generative adversarial network (GAN) [30] achieves impressive results in image synthesis and is applied in our method. WGAN-GP [31] is an important extension of GAN which improves image quality and stabilizes training. We train the WGAN-GP on CASIA-B for 16000 iterations; the trained generator G and the discriminator D are used in attacking gait silhouettes. We set the dimension of inputted latent variables as $V = 128$.

SPADE [32] is a method based on conditional normalization, and it can convert the segmentation map to a photo-realistic image. We use the SPADE in the silhouettes to video translation. We train it on CASIA-B for 670000 iterations.

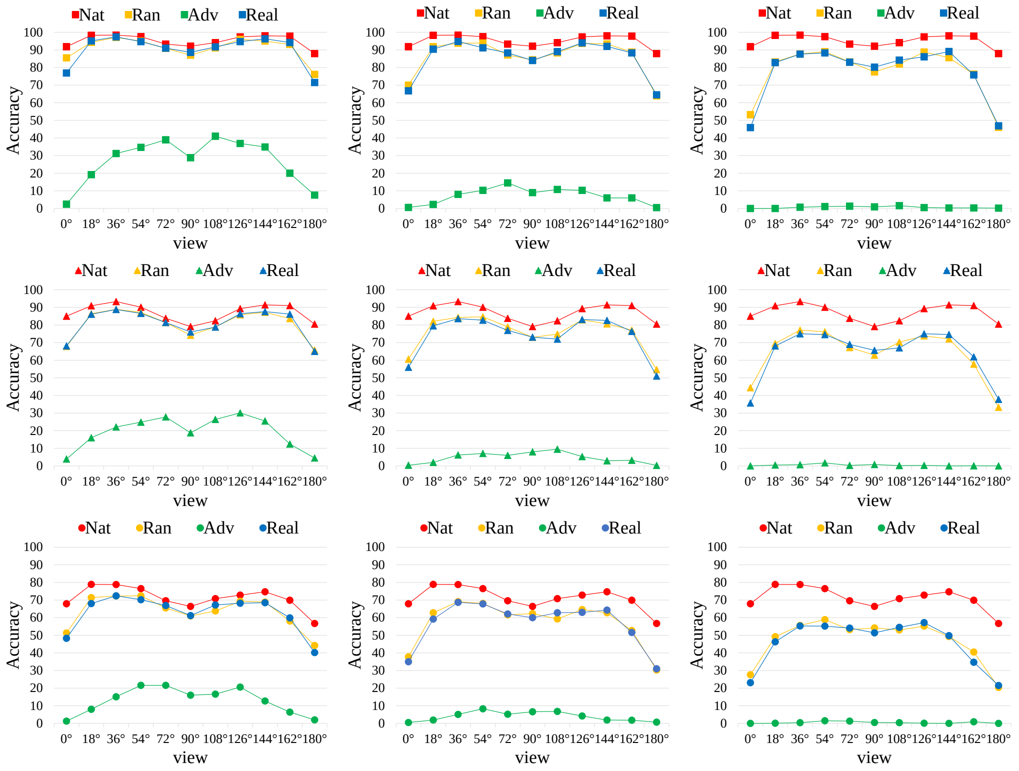
Settings for perturbation methods. For comparison, we report the experimental results under the attack setting as the 4th step in Fig. 3. For perturbation-based methods, we choose FGSM, PGD, and MIFGSM as baselines. The distortion budget is set to the value 1 for these methods, with the pixel value within [0,1]. For iterative methods, PGD and MIFGSM, the iterations are set to 20. The decay factor in MIFGSM is set as 1.0.

4.2. White-box experiments

In this subsection, we attack a gait model under the white-box protocol. This means we have the full knowledge of this target model. The attacked model is GaitSet, the state-of-the-art gait recognition model. GaitSet regards the gait as a set of gait silhouettes and utilizes a deep neural network to directly extract temporal information during training. Moreover, GaitSet is flexible since the input set can contain any number of non-consecutive silhouettes.

In the following, we evaluate the attack ability of our method. For clarity, we report the results of frame-alteration attack and omit the results of frame-insertion attack as the trends for both methods are similar.

Dodging attack results. We report the results of dodging attack in Fig.4. The natural accuracy of GaitSet is labeled with *Natural* (marked as red) and the accuracy under our attack is labeled with *Adversarial* (marked as green). We observe that our attacks with different sparsity successfully deceive GaitSet, causing low accuracy in all three walking conditions. We



(a) sparsity=1/40. (b) sparsity=1/20. (c) sparsity=1/10.

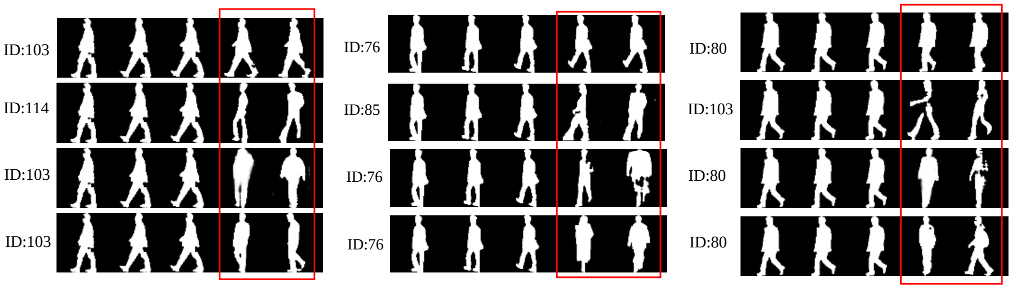
Figure 4: Results of white-box dodging attack. Each column represents a predefined sparsity, including 1/10, 1/20, and 1/40. From top to bottom are results of different walking conditions, including NM (first row), BG (second row), and CL (third row). Four settings (Natural (Nat), Random (Ran), Real and Adversary (Adv)) are labeled with different colors.

also find that a stronger attack needs more modified frames in a sequence. However, the accuracy still drops dramatically when the sparsity is 1/40.

To further prove the drop accuracy is caused by our novel attack design rather than altering some frames with randomly chosen gait silhouettes, we compare *Adversarial* with other two situations: (1) *Real* (marked as blue). Randomly selected real frames of other subjects replace some frames in the original sequence. (2) *Random* (marked as yellow). In this scenario, the latent variable is randomly sampled from a standard Gaussian distribution. Then our GAN model generates attacking frames from the randomly sampled vector. This is different from our attacking method, since in our attack the vector $\boldsymbol{Z}$ is optimized by gradient backpropagation as the blue arrows in Fig. 2, i.e., our attack is searching the optimal $\boldsymbol{Z}$ in a prior distribution rather than randomly sampling $\boldsymbol{Z}$. As shown in Fig. 4, Both *Real* and *Random* only slightly affect the accuracy of GaitSet, while *Adversarial* has more severe damage to the recognition performance. These results demonstrate the effectiveness of our attack.

One intriguing phenomenon is that altering some original frames have a more obvious effect on the accuracy when the view angle is close to 0° or 180°. Under these conditions, the proposed attack is hardly recognizable. These remain challenging cases for most of the state-of-the-art gait recognition methods. Besides 0° and 180°, the accuracy of 90° under attack is a local minimum value. Chao et al. [3] point out that both parallel and vertical perspectives lose some part of gait information. For example, stride can be observed most clearly at 90° while a left-right swinging of body or arms can be observed most clearly at 0° or 180°. For attacking case, we conclude that the parallel and vertical perspectives are more fragile when facing noises. Fig. 4 empirically proves this statement. When replacing some frames with randomly generated or real silhouettes instead of adversarial images, the accuracy of 0° or 180° still has a larger decrease than other views.

Moreover, to prove that the accuracy is not affected by the quality of inserted images, we randomly select some silhouettes and show them in Fig. 5.



(a) NM. (b) BG. (c) CL.

Figure 5: Comparison of modified images. For each subfigure, from top to bottom are a short sequence corresponding to four different settings, i.e., *Natural*, *Adversarial*, *Random* and *Real*. ID is classified by GaitSet.

Here the source sequence of NM is five consecutive frames in the fifth sequence of subject 103 under normal condition with angle view 108°, BG is the second sequence of subject 76 with angle view 144° and CL is the first sequence of subject 80 with angle view 72°. For all the three walking conditions, NM, BG, and CL, the silhouettes drawn from attacking frames achieve a competitive quality with a real silhouette. Furthermore, we observe that, although the generated silhouettes of *Random* in Fig. 5(b) are low quality and do not seem like a person with bag, GaitSet still makes a correct classification. Contrarily, the adversarial sequence successfully fools GaitSet. Therefore, GaitSet is robust to a slight disturbance but vulnerable under the proposed adversarial attack.

Impersonation attack results. We report the results of impersonation attack in Table 2. The results share some similarities with dodging attack results: the success rate is positively correlated with sparsity; the attack achieves a higher success rate when the view angle is 0° or 180°. But com-

Table 2: Results of white-box impersonation attack, shown as success rate(%).

		viewing angles: 0° - 180°											
sparsity	condition	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	average
1/40	NM	34.60	18.90	13.20	8.60	7.70	9.40	7.90	10.10	8.30	16.80	30.80	15.118
	BG	31.70	20.70	14.60	11.41	11.70	14.40	12.40	12.10	16.10	22.93	35.50	18.504
	CL	42.30	23.40	17.20	17.50	13.10	17.90	15.70	14.90	23.20	26.00	41.00	22.927
1/20	NM	61.60	43.40	33.30	30.50	25.40	29.00	22.20	27.90	37.00	40.10	63.50	37.627
	BG	63.30	47.00	39.50	34.64	33.40	37.30	33.00	36.10	41.10	47.07	62.80	43.201
	CL	65.60	49.20	39.50	37.80	35.70	36.20	36.30	38.60	47.20	50.70	63.80	45.509
1/10	NM	78.60	67.40	59.40	61.40	60.00	61.10	58.90	60.90	65.10	68.60	80.60	65.636
	BG	82.30	69.80	65.40	64.55	61.20	62.90	60.50	63.50	66.50	70.91	80.20	67.978
	CL	78.90	70.50	63.80	66.20	62.40	65.00	65.50	65.70	69.90	73.10	81.30	69.300

pared with the dodging attack, the impersonation attack is apparently more difficult, because it aims to deceive a gait recognition model with a specific subject ID rather than any one. Nonetheless, the proposed method can successfully deceive the GaitSet with a high success rate at around 65% when the sparsity is 1/10. Therefore, our method can serve as a strong benchmark of adversarial attack on gait recognition.

The goal of impersonation attack is similar to spoofing attack, which aims to gain illegitimate access to gait systems by masquerading as others. Here we compare our impersonation attack with spoofing attack proposed by Jia et al [10]. Results are shown in Table 3. Though spoofing attack achieves a high success rate, it needs to alter each frame of the source sequence, i.e., generating a fake background to substitute the original background. Our method can achieve a satisfactory success rate with a slighter modification.

Table 3: Comparison with spoofing attack, shown as success rate(%).

gallery: NM 01-04	viewing angles: 0° - 180°												
probe: NM 05-06	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	average	sparsity
spoofing attack [10]	68.0	86.0	92.0	89.0	82.0	78.0	82.0	89.0	90.0	85.0	65.0	82.0	100%
ours	85.3	80.1	75.3	74.2	74.7	76.9	76.6	75.4	77.3	78.2	84.0	78.0	20%

4.3. Black-box experiments

Our attack method is specifically aimed at sequence-based gait recognition models, and the above experimental results demonstrate their vulnerability. In this subsection, we also make a black-box attack on the template-based model, GaitGAN [2]. Different from GaitSet, which takes a gait sequence as a set and extracts its feature with a CNN, GaitGAN uses a GEI template as the gait feature. Moreover, GaitGAN takes a GAN model as a regressor to simultaneously address variations in viewpoint, clothing, and carrying conditions in gait recognition. In the black-box scenario, we cannot access any information of GaitGAN in the attack process. To perform the black-box attack, we apply the widely used transfer-based attack [33, 34]. It leverages a property of adversarial examples, i.e., transferability, which means that adversarial examples crafted on one model can successfully attack another model with different architecture and parameters. In transfer-based attack, attackers use a local substitute model to craft adversarial examples and feed them into a black-box target model to result in wrong outputs. Specifically, here we firstly attack GaitSet with Algorithm 1 to obtain the adversarial sequence, and then use it as the input of GaitGAN to test the accuracy.

The sparsity is 1/40 and the adversarial sequences are the same as Adv,

Table 4: Results of a Black-box attack on GaitGAN, shown as accuracy(%).

probe view	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	average
natural	39.4	56.0	62.3	61.1	59.3	25.8	55.8	63.6	57.3	52.9	40.7	52.2
after attack	35.9	52.8	60.0	57.6	56.1	24.4	52.2	60.4	55.2	49.5	36.2	49.1
drop ↓	3.5	3.2	2.3	3.5	3.2	1.4	3.6	3.2	2.1	3.4	4.5	3.1

NM in Fig. 4(a). We report the results of probeNM, shown in Table 4. The recognition rate of each probe view only drops a little after attacking. We conclude that GEI is more robust than the feature extracted by GaitSet under our temporal adversarial attack. Because GEI is obtained by aligning the silhouettes in the spatial space and averaging them along the temporal dimension, the perturbation of a few frames is not enough to deceive GaitGAN. Although sequence-based gait recognition has made great progress in recognition accuracy, its robustness compared to template-based methods remains limited. This is a key area for the community to focus on in the future.

4.4. Cross-dataset validation

For a more reliable performance assessment, we conduct cross-database testing using CASIA-A. In this scenario, the training set of CASIA-B is used to train the GaitSet and WGAN-GP, while the whole CASIA-A dataset is used for testing. Results are shown in Table 5. The trend is almost the same as the results of testing on CASIA-B. The recognition capability of the attacked model drops rapidly as the attack sparsity increases.

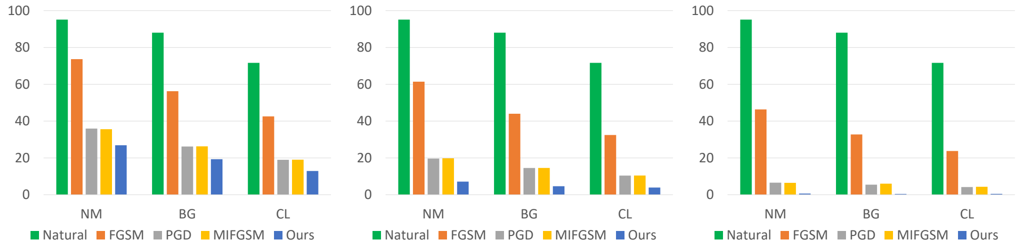
Under the dodging attack, the performance degradation could be affected by many reasons, such as domain shift or the generalization ability of the

Table 5: Results of cross-database validation with dodging attack, shown as accuracy(%).

	probe view			
sparsity	0°	45°	90°	average
0	56.67	70.00	76.67	67.78
1/40	33.33	33.33	18.33	28.33
1/20	18.33	25.00	3.33	15.55
1/10	6.67	8.33	3.33	5.11

Table 6: Results of cross-database validation with impersonation attack, shown as success rate(%).

	probe view			
sparsity	0°	45°	90°	average
1/40	11.67	8.33	31.67	17.22
1/20	26.67	20.00	61.67	36.11
1/10	56.67	63.33	85.00	68.33



(a) sparsity=1/40.

(b) sparsity=1/20.

(c) sparsity=1/10.

Figure 6: Comparison of our method with perturbation-based methods. ‘Natural’ is the original accuracy of GaitSet and others represent the accuracy under different attacks.

recognition method itself, other than attacking. For a more convincing justification, we further perform cross-database impersonation attack. Results are shown in Table 6. When the sparsity is 1/10, the success rate reaches 68.33% on average.

4.5. Comparison with perturbation-based methods

In Sec. 1, we have qualitatively demonstrated the shortcomings of the perturbation-based approaches. In this subsection, we compare our proposed method with these methods quantitatively. The perturbation-based

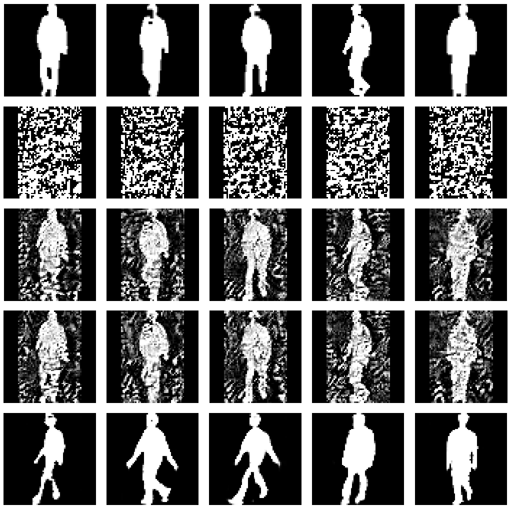


Figure 7: From top to bottom are natural examples and their corresponding adversarial examples generated by FGSM, PGD, MIFGSM, and our method.

attacks used as baselines include FGSM [9], PGD [24] and MIFGSM [25]. It is difficult to extend these attacks to video frame generation due to the signal processing in silhouette-based gait recognition. Therefore, in this study we perform attacks on the gait silhouettes. For a fair comparison, we fix the sparsity and compare the imperceptibility and the accuracy under different attacks. The distortion budgets of perturbation-based methods are all relaxed to pixel values, which means adversarial examples are not norm-bounded by a small constant for these methods. Our method is under the protocol of a frame-alteration dodging attack.

The results are shown in Fig. 6, and some crafted adversarial examples are shown in Fig. 7. Our proposed method achieves a superior attack performance and imperceptibility. In Fig. 6, we observe that while all of the attacks lower the accuracy of GaitSet, our method surpasses the perturbation-based methods by obtaining superior attack performance in all of the settings. In Fig. 7, we show that the subjects in the silhouettes retain a human posture in our method. Thus, it maintains better imperceptibility of the spatial domain than perturbation-based methods. Furthermore, it enables the transfer of these silhouettes to video frames; it makes a practical threat to the gait recognition system. However, a limitation is that the generated samples have pose changes when they are compared to their adjacent frames. Therefore, some constraints are needed to enforce the changes between the adjacent frames, which we leave for our future work. Compared with perturbation-based methods, our proposed method provides superior attack ability and imperceptibility and can serve as a stronger baseline for sequence-based gait recognition.

4.6. Analysis of the proposed method

In this section, we make a further analysis of the proposed method.

Position of frame-insertion. Firstly, we study the effects of position to insert the adversarial image. In our prior experiments, we use GaitSet as the target model. GaitSet has achieved state-of-the-art performances without modeling the temporal characteristics explicitly. In other words, GaitSet takes a set of silhouettes as input and the order of input frames does not affect the recognition. Similarly, models using gait templates, such as GaitGAN [2], aggregate temporal walking information over a sequence of silhouettes in a single map. The order of a gait sequence does not matter in these methods. Differently, some models learn from the order and relationship of frames in gait sequences, instead of aggregating them. We take SelfGait [35] as an example and perform dodging attack on it. SelfGait is a

Table 7: Study on position of inserted adversarial frames, shown as accuracy(%).

	walking condition		
position	NM	BG	CL
no inserted frames	91.079	84.735	74.261
{0,1,2,3}	17.711	12.197	10.663
{8,9,10,11}	18.895	11.193	11.212
{15,16,17,18}	18.190	11.913	11.629
{23,24,25,26}	18.176	11.420	12.689
{30,31,32,33}	19.784	12.651	11.117
{0,11,22,33}	19.607	12.784	11.723
{4,12,20,28}	20.035	14.565	11.553
{9,14,19,24}	18.406	11.572	9.583
{13,15,17,19}	18.992	11.553	10.303
random (mean±std)	19.130±0.406	12.940±0.281	10.530±0.890

self-supervised framework with spatiotemporal components to learn from the massive unlabeled gait images. Since SelfGait preserves and learns temporal representation from the order and relationship of frames in gait sequences, disrupting the order of input frames will decrease its accuracy.

In the following experiments, we randomly draw 30 sorted successive frames from a sequence as input and insert 4 adversarial frames into different positions. Results are reported in Table 7. Different positions are numbered with the index of adversarial frames in the obtained 34 frame-length sequence. For example, $\{0, 1, 2, 3\}$ represents that all the 4 adversarial frames are inserted into the start of a sequence. In Table 7, from the row of $\{0, 1, 2, 3\}$ to $\{30, 31, 32, 33\}$ are inserting all the 4 frames into two adjacent frames. From the row of $\{0, 11, 22, 33\}$ to $\{13, 15, 17, 19\}$ are insert-

Table 8: Study on number of inserted adversarial frames, shown as accuracy(%).

dynamic evaluation					
frame number	1	2	3	4	5
accuracy	47.51%	14.73%	3.71%	1.57%	0.56%
static evaluation					
accuracy			0%		
mean frame number			1.6		

ing adversarial frames into equidistant positions. The last row, 'random', means that for each original sequence, adversarial frames are inserted into randomly chosen positions. Its final result is averaged on five experiments with different random seeds (from 0 to 4). We observe different positions have a slight effect on the attacking performance. For example, for all inserting positions, the accuracy of SelfGait under condition NM is around 19%. Therefore, our method can pre-define any positions to insert the generated adversarial frames for a similar result.

Number of adversarial images. We study the minimum number of inserted adversarial frames in order to yield a satisfactory result. To make a fair comparison, we fix the length of silhouette frames in the test phase. Specifically, the sampler collects 30 sorted successive frames as input. We evaluate under two settings: (1) Static evaluation. We pre-define a number of inserted adversarial frames and test the accuracy under such a pre-defined number. (2) Dynamic evaluation. For a source gait sequence, we gradually increase the number of inserted frames and run until a successful attack. Under this setting, the evaluation metric is the mean of the inserted frame numbers. We perform dodging attack on GaitSet and obtain results in Ta-

Table 9: Ablation study on the hyper-parameter λ, shown as accuracy(%).

λ	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
NM	25.3	25.9	25.9	26.4	32.8	45.8
BG	17.9	18.3	18.5	18.6	24.8	24.4
CL	12.1	12.1	11.9	12.4	15.1	22.5

Table 10: Speed evaluation of the proposed attack, shown as the average time (seconds) to craft a adversarial sequence on CASIA-B.

iterations	40			60			80			100		
sparsity	1/40	1/20	1/10	1/40	1/20	1/10	1/40	1/20	1/10	1/40	1/20	1/10
computation time	1.43	1.58	1.65	2.29	2.32	2.54	2.98	3.03	3.44	3.42	3.54	3.80

ble 8. We can observe inserting only three adversarial frames can degrade the accuracy to to 3.71%. Furthermore, for a 30 frame-length sequence, the mean frame number to completely deceive GaitSet is only 1.6.

Hyper-parameters. We study the hyper-parameter λ in Eq.(6) since it is crucial for the trade-off between attack performance and the imperceptibility in the spatial domain. We perform the dodging attack on the GaitSet, with a sparsity of 1/40. Considering the imperceptibility is difficult to be quantitatively measured, we mainly evaluate the attack performance with the accuracy of GaitSet. Table 9 shows that the accuracy increases as the λ becomes larger.

Finally, we study the parameters in Algorithm 1 which affect the computational complexity, including the iterations T, the step size ϵ, the sparsity S and the length of the original sequence, etc. We fix the step size as $\epsilon = 0.1$

and the length of the original sequence as 40 frames to study the effects of iterations T and sparsity S. Results are shown in Table 10. We conclude the iterations has a significant influence on the speed and the sparsity has a relatively modest effect.

5. Discussion

Attack scenarios. We have mainly evaluated our attack in the digital space. For a real-world attack, one can access the transmission channel to repeat the previously prepared adversarial biometric data on the channel and avoid the sensor. This can be achieved in a number of ways, e.g. through reverse engineering the authentication protocol and directly talking to the system [36]. A predictable difficulty is that such an attack is possibly hindered by many other uncertainties in an open world. For example, the dynamic background instead of still ones can decrease the visual quality of generated videos; some frame selection process may remove the key adversarial frames.

Countermeasures. Several studies provide anti-spoofing methods to defend spoofing attacks. However, the proposed adversarial attack still lacks effective countermeasures. To improve the robustness of sequence-based gait system against our attack, one can consider adversarial training [24], i.e., augmenting the training data with our generated adversarial sequences. A shortcoming is this method will lower down the accuracy on the benign samples. Instead, one can detect the adversarial frames under the hypothesis that they have already known the attack. For example, the inter-class dissimilarity between adversarial frames and benign ones can be utilized, since the adversarial samples are all crafted by a GAN in our method.

6. Conclusion

In this paper, we propose a novel temporal sparse adversarial attack on gait recognition. Our method achieves good imperceptibility and a high attack performance. Experiments on CASIA datasets indicate that the state-of-the-art model, GaitSet, is vulnerable to our adversarial attack. This reveals a key limitation in adversarial robustness research on gait recognition that requires urgent attention. Our method also shows a potential threat in practical applications as it is flexible in either attacking on biometric samples captured by a sensor or directly modifying probes. We mainly focus on the vulnerability of sequence-based models and show template features like GEI may resist our attack. The results highlight the inherent loss of temporal and fine-grained spatial information in gait templates; consequently, they can avoid deliberate attacks on vulnerable temporal features. Therefore, we identify the need for the community to consider the robustness of sequence-based methods, which possess the benefit of high accuracy, in future research.

References

[1] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, Geinet: View-invariant gait recognition using a convolutional neural network, in: International Conference on Biometrics, 2016, pp. 1–8.

[2] S. Yu, H. Chen, E. B. G. Reyes, N. Poh, Gaitgan: Invariant gait feature extraction using generative adversarial networks, in: IEEE Conference

on Computer Vision and Pattern Recognition Workshops, 2017, pp. 532–539.

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

[4] C. Song, Y. Huang, Y. Huang, N. Jia, L. Wang, Gaitnet: An end-to-end network for gait based human identification, Pattern Recognition 96 (2019) 106988. doi:https://doi.org/10.1016/j.patcog.2019.106988.

[5] T. Wolf, M. Babaee, G. Rigoll, Multi-view gait recognition using 3d convolutional neural networks, in: IEEE International Conference on Image Processing, 2016, pp. 4165–4169.

[6] R. Liao, C. Cao, E. B. Garcia, S. Yu, Y. Huang, Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations, in: Chinese Conference on Biometric Recognition, 2017, pp. 474–483.

[7] H. Li, Y. Qiu, H. Zhao, J. Zhan, R. Chen, T. Wei, Z. Huang, Gaitslice: A gait recognition model based on spatio-temporal slice features, Pattern Recognition 124 (2022) 108453. doi:https://doi.org/10.1016/j.patcog.2021.108453.

[8] 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, 2014.

[9] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: International Conference on Learning Representations, 2015.

[10] M. Jia, H. Yang, D. Huang, Y. Wang, Attacking gait recognition systems via silhouette guided gans, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 638–646.

[11] T. Zhu, L. Fu, Q. Liu, Z. Lin, Y. Chen, T. Chen, One cycle attack: Fool sensor-based personal gait authentication with clustering, IEEE Transactions on Information Forensics and Security 16 (2021) 553–568. doi:10.1109/TIFS.2020.3016819.

[12] D. Li, J. Zhang, K. Huang, Universal adversarial perturbations against object detection, Pattern Recognition 110 (2021) 107584. doi:https://doi.org/10.1016/j.patcog.2020.107584.

[13] M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter, Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 1528–1540.

[14] Y. Song, R. Shu, N. Kushman, S. Ermon, Constructing unrestricted adversarial examples with generative models, in: Advances in Neural Information Processing Systems, 2018, pp. 8312–8323.

[15] O. Poursaeed, T. Jiang, H. Yang, S. J. Belongie, S.-N. Lim, Fine-grained synthesis of unrestricted adversarial examples, ArXiv abs/1911.09058 (2019).

[16] 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 (2016) 209–226.

[17] Y. He, J. Zhang, H. Shan, L. Wang, Multi-task gans for view-specific feature learning in gait recognition, IEEE Transactions on Information Forensics and Security 14 (2019) 102–113.

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

[19] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: International Conference on Pattern Recognition, 2006, pp. 441–444.

[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.

[21] T. Dai, Y. Feng, B. Chen, J. Lu, S.-T. Xia, Deep image prior based defense against adversarial examples, Pattern Recognition 122 (2022) 108249. doi:https://doi.org/10.1016/j.patcog.2021.108249.

[22] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: A simple and accurate method to fool deep neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582.

[23] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: IEEE European Symposium on Security and Privacy, 2016, pp. 372–387.

[24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.

[25] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9185–9193.

[26] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[27] X. Wei, J. Zhu, H. Su, Sparse adversarial perturbations for videos, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 8973–8980.

[28] Z. Chen, L. Xie, S. Pang, Y. He, Q. Tian, Appending adversarial frames for universal video attack, in: IEEE Winter Conference on Applications of Computer Vision, 2021, pp. 3198–3207.

[29] S. Menon, A. Damian, S. Hu, N. Ravi, C. Rudin, PULSE: self-supervised photo upsampling via latent space exploration of generative models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 2434–2442.

[30] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[31] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.

[32] T. Park, M. Liu, T. Wang, J. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.

[33] N. Papernot, P. D. McDaniel, I. J. Goodfellow, Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, arXiv preprint arXiv:1605.07277 (2016).

[34] Y. Liu, X. Chen, C. Liu, D. X. Song, Delving into transferable adversarial examples and black-box attacks, in: Proceedings of International Conference on Learning Representations, 2017.

[35] Y. Liu, Y. Zeng, J. Pu, H. Shan, P. He, J. Zhang, Selfgait: A spatiotemporal representation learning method for self-supervised gait recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2021, pp. 2570–2574.

[36] M. Adámek, M. Matỳsek, P. Neumann, Security of biometric systems, Procedia Engineering 100 (2015) 169–176.