Uddin *et al. IPSJ Transactions on Computer Vision and Applications* (2019) 11:9
https://doi.org/10.1186/s41074-019-0061-3

IPSJ Transactions on Computer Vision and Applications

RESEARCH PAPER Open Access

Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion

Md. Zasim Uddin[1*], Daigo Muramatsu[1], Noriko Takemura[2], Md. Atiqur Rahman Ahad[1] and Yasushi Yagi[1]

Abstract

Gait-based features provide the potential for a subject to be recognized even from a low-resolution image sequence, and they can be captured at a distance without the subject's cooperation. Person recognition using gait-based features (gait recognition) is a promising real-life application. However, several body parts of the subjects are often occluded because of beams, pillars, cars and trees, or another walking person. Therefore, gait-based features are not applicable to approaches that require an unoccluded gait image sequence. Occlusion handling is a challenging but important issue for gait recognition. In this paper, we propose silhouette sequence reconstruction from an occluded sequence (sVideo) based on a conditional deep generative adversarial network (GAN). From the reconstructed sequence, we estimate the gait cycle and extract the gait features from a one gait cycle image sequence. To regularize the training of the proposed generative network, we use adversarial loss based on triplet hinge loss incorporating Wasserstein GAN (WGAN-hinge). To the best of our knowledge, WGAN-hinge is the first adversarial loss that supervises the generator network during training by incorporating pairwise similarity ranking information. The proposed approach was evaluated on multiple challenging occlusion patterns. The experimental results demonstrate that the proposed approach outperforms the existing state-of-the-art benchmarks.

Keywords: Silhouette reconstruction, Gait recognition occlusion handling, Video generation, Deep generative adversarial network, Wasserstein GAN

1 Introduction

Biometric-based person authentication is becoming increasingly important for various applications, such as access control, visual surveillance, and forensics. Gait recognition is one of the topics of active interest in the biometric research community because it provides unique advantages over other biometric features, such as the face, iris, and fingerprints. For example, it can be captured without the subject's cooperation at a distance and has discriminative capability from relatively low-resolution image sequences [36]. Recently, gait has been used as a forensic feature, and there has already been a conviction produced by gait analysis [14].

However, gait recognition has to manage some practical issues that include observation views [27, 45], clothing [13], carried object [37], and occlusion. In this study, we address the gait recognition problem against occlusion.

Occlusion for gait recognition can be one of two types based on the relative position between the occluder and the target subject in an image sequence: relative dynamic occlusion and relative static occlusion. For relative dynamic occlusion, the occluded portion of the target subject changes continuously over an image sequence, whereas, for relative static occlusion, the occluded portion does not change. An example of relative dynamic occlusion is shown in Fig. 1a and b, in which the person is occluded at different positions in each frame and the occluded portion of the person's body gradually changes in the video sequence during the person's gait cycle. For the example of relative static occlusion shown in Fig. 1c, the person is occluded at a fixed portion of the body in each frame in the video sequence during the person's gait cycle.

*Correspondence: zasim@am.sanken.osaka-u.ac.jp
[1]The Institute of Scientific and Industrial Research, Osaka University, Osaka 567-0047, Japan
Full list of author information is available at the end of the article

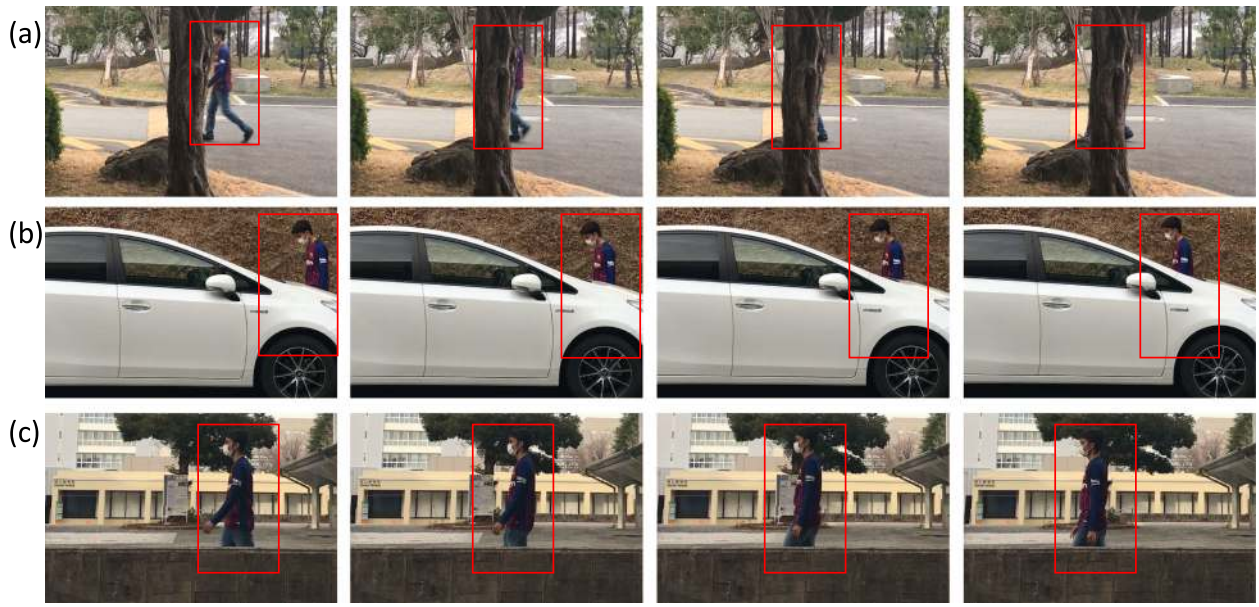

Fig. 1 Examples of occlusion in real-life applications (every fifth frame of a sequence). **a** Relative dynamic occlusion where the subject is occluded by a tree and continuously changes the occluded portion from left to right. **b** Relative dynamic occlusion where the subject is occluded by a parked car and continuously changes the occluded portion from bottom to top. **c** Relative static occlusion where the subject is occluded by wall in a fixed position

Approaches to gait recognition against occlusion can be roughly grouped into two categories. The first category is reconstruction-free approaches [5, 28, 29, 48], which focus on extracting features from a silhouette sequence of a gait cycle or an average of them, such as the gait energy image (GEI) [10]. Because gait features are extracted by considering static shape and dynamic motion information from a silhouette sequence for a gait cycle, approaches of this type can achieve good performance for a very low degree of occlusion; however, the obvious limitation of this type of approach is that it cannot be applicable to cases in which the gait cycle is difficult to estimate.

The second category is reconstruction-based approaches [12, 33]. Approaches in this category focus on reconstructing occluded silhouettes. In these approaches, occluded silhouettes are identified and a sequence is separated into occluded and unoccluded gait cycles, and then silhouettes of occluded gait cycles are reconstructed. These approaches showed good silhouette reconstruction. However, these were applied on long sequences that consisted of multiple gait cycles in which some frames were partially occluded. These approaches are difficult to apply in the case in which all frames are severely occluded in a sequence, for example, the occlusion shown in Fig. 1a and b. One of the major limitations of reconstruction-based approaches is that the reconstructed silhouette sequence sometimes deteriorates the discrimination ability of the individual after reconstruction. Therefore, it can negatively influence gait recognition performance after reconstruction [22].

With the great success of deep convolutional neural networks (CNNs) and generative adversarial networks (GANs) [8] in many research areas of computer vision and biometrics, reconstruction-based approaches have been formulated as a conditional image or video generation problem for image inpainting [15, 21, 30, 42, 44], video inpainting [20, 39], and future prediction [4, 20, 23, 38]. Although, these works have been shown to generate very good-looking realistic images, such as faces, objects, and scenes, they sometimes lost subject identity [46]. An approach that can generate not only good-looking samples but also samples with the discrimination ability of an individual is necessary for biometric-based person recognition.

We present an effective feed-forward conditional deep generative network for silhouette sequence reconstruction considering dilated convolution [15, 43] and a skip connection [32]. Dilated convolutional kernels are spread out in the spatial and temporal directions, which allows us to reconstruct each pixel by covering a large spatio-temporal input area. This is important for silhouette sequence reconstruction because each input pixel is important for reconstruction, whereas a skip connection allows us to retain unoccluded input pixels as output. The input to the encoder network that maps hidden representations is the occluded silhouette sequence, and the output of the decoder is the reconstructed silhouette

sequence. We regularize the training process of the generator network by incorporating triplet hinge loss into Wasserstein GAN (WGAN) loss [1, 9] as adversarial loss and reconstruction loss in pixel space. A triplet contains a query sequence, positive sequence, and negative sequence, where the query sequence is the reconstructed silhouette sequence, the positive sequence is the unoccluded silhouette sequences of the same subject as the query subject, and the negative sequence is of a different subject. The similarity relationship is characterized by the relative distance in the triplet.

The entire network is trained end to end with the reconstruction and proposed adversarial losses. Compared with existing inpainting or reconstruction-based approaches, one of the major advantages of our proposed approach is that it does not require occluded or inpainting position information (i.e., a mask) for reconstruction. Therefore, it can be applied to an arbitrarily structured occluded silhouette sequence during reconstruction. Because of the silhouette sequence reconstruction approach, we can evaluate gait recognition without knowing the gait cycle in advance because the gait cycle can be estimated from the reconstructed silhouette sequence.

The contributions of this paper are summarized as follows:

1. We propose to design a conditional deep generative network (sVideo) that consists of a generator with dilated convolution and a skip connection, and a critic network. It can reconstruct any type of occluded silhouette sequence.
2. We propose a novel adversarial loss based on triplet hinge loss incorporated with WGAN loss (WGAN-hinge). To the best of our knowledge, WGAN-hinge is the first adversarial loss that supervises the generator network during training by incorporating pairwise similarity ranking information.
3. We demonstrate the stability of the proposed generative network using the supervision of adversarial loss for WGAN and also propose WGAN-hinge loss during training for various experiments to reconstruct the silhouette sequence and present superior results for gait recognition compared with the state-of-the-art methods. Additionally, we also demonstrate that the proposed WGAN-hinge for a different generator network yields performance improvements over WGAN.

2 Related work

2.1 Existing approaches for gait recognition against occlusion

In this section, we review the works related to gait recognition against occlusion as two families: reconstruction-free approaches and reconstruction-based approaches.

Regarding reconstruction-free approaches, the following methods have been proposed. Zhao et al. [48] extracted features based on fractal scale wavelet analysis for each silhouette from a sequence of a gait cycle and then averaged them. They evaluated robustness against noisy data in addition to occluded data by adding a vertical bar in the silhouette sequence. Chen et al. [5] proposed an approach for an incomplete and occluded silhouette for gait recognition. They divided the silhouette sequence of a gait cycle into clusters, and the dominant energy image (DEI) was calculated by denoising each cluster. The frame difference energy image (FDEI) for a silhouette was computed as the summation of its corresponding clusters' DEI and the positive portion of the difference from the previous frame. Finally, features are extracted from the FDEI representation that mitigated the problem of spatial and temporal silhouette incompletion caused by imperfect silhouette segmentation and occlusion. In [29], a robust statistical framework was proposed that minimized the influence of silhouette defects. The authors evaluated gait recognition on GEIs and gradient histogram energy images by adding occlusion and noise into a silhouette sequence. A different technique to manage the problem of occlusion was addressed in [28], in which a GEI was separated into four modules and a module was excluded for gait recognition if occlusion was detected.

Regarding reconstruction-based approaches, Roy et al. [33] proposed a framework in which a silhouette sequence was first divided into a few subsequences of gait cycle(s) based on key poses. It also allowed the determination of whether a silhouette of a gait cycle was occluded. Occluded silhouettes were then reconstructed using a balanced Gaussian process dynamical model. Although the authors evaluated the reconstruction accuracy, they did not evaluate gait recognition using the reconstructed silhouette sequence. Hofmann et al. [12] proposed a simple approach to detect partially occluded gait subsequences from a sequence using foreground pixels. Occluded silhouettes were then replaced by similar-pose clean silhouettes from other cycles. In [26], a complete GEI was regenerated from a partially observable GEI using subspace-based method. Gait recognition was evaluated according to whether a matching pair did not share a common observable region.

We can observe that, from the above discussion, some approaches manage occlusion directly on pre-process feature GEI for a gait cycle. Thus, they assume that the gait cycle is known in advance. The remaining approaches estimate the gait cycle from the occluded silhouette sequence, which is very difficult or error prone when all frames are occluded in a sequence, for example, Fig. 1. Furthermore, they consider a large sequence where multiple gait cycles are available for gait recognition. However, there are many scenarios in real-world applications in

which only a few frames (i.e., not more than a gait cycle) are available in a sequence, and all are partially or totally occluded. In those scenarios, existing approaches are not applicable.

2.2 Deep generative approach

GAN [8] is a framework for training the generative model implemented by a system of two neural networks: generative network G and auxiliary discriminator network D. The discriminator network serves to distinguish whether content is generated by a network or is real, whereas the generator network is trained to fool the discriminator network. Specifically, G and D are trained by solving the following minimax problem :

$$\min_G \max_D \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}\left[\log D\left(x\right)\right]+\mathop{\mathbb{E}}_{G(z)\sim\mathbb{P}_g}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right], \tag{1}$$

where $\mathbb{E}(\cdot)$ denotes the expectation operator, and $\mathbb{P}_r$ and $\mathbb{P}_g$ are the real and generated data distributions, respectively. Generator G transforms input sample z to mimic a real sample. However, one of the main concerns of GAN is instability during training. Numerous works have addressed improving the training stability. Radford et al. [31] introduced deep convolutional GANs (DCGAN) that imposed empirical constraints on the architecture of the GAN and optimized the hyperparameters. Recently, Arjovsky et al. [1] introduced WGAN [9], which minimizes the Earth Mover's Distance (a.k.a Wasserstein-1) between the generator and real data distribution. Specifically, the objective function was constructed by applying the Kantorovich-Rubinstein duality:

$$\min_G \max_{D\in\mathcal{D}} \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}\left[D(x)\right]-\mathop{\mathbb{E}}_{G(z)\sim\mathbb{P}_g}\left[D\left(G(z)\right)\right], \tag{2}$$

where $\mathcal{D}$ is the set of 1-Lipschitz functions. To enforce the Lipschitz constraint on the critic function, Gulrajani et al. [9] proposed an improved version of WGAN with a gradient penalty term with respect to the input. The new objective is as follows:

$$\min_G \max_D \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}\left[D(x)\right]-\mathop{\mathbb{E}}_{G(z)\sim\mathbb{P}_g}\left[D(G(z))\right]+\lambda L_{GP}, \tag{3}$$

where $L_{GP} = \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[\left(\left\|\nabla_{\hat{x}}D(\hat{x})\right\|_2-1\right)^2\right]$, $\hat{x} = \epsilon x + (1-\epsilon)\tilde{x}$, and λ is a gradient penalty coefficient and $\epsilon \sim U[0,1]$. The authors called the auxiliary network a critic instead of discriminator because it is not a classifier. In this paper, we train our proposed approach using the framework of WGAN with the gradient penalty coefficient [9]. We present our approach in detail in Section 3.

2.3 Image and video reconstruction

A large body of literature exists for image and video reconstruction from traditional approaches to learning-based approaches (i.e., deep learning). Traditional approaches include diffusion-based [3] and patch-based techniques[7]. The diffusion-based technique propagates the image appearance around the target position, where propagation can be performed based on the isophote direction field, whereas the patch-based technique extracts patches from a source image and then pastes them into a target image. The patch-based technique is also used for video completion [40] by replacing image patches with spatio-temporal synthesis across frames. However, these types of approaches can only fill a very small and homogeneous area, and one obvious limitation is the repetition of content.

Recently, conditional GAN-based [25] approaches have emerged as promising tools for image and video completion. Regarding image completion, a context encoder (CE) [30] was the first attempt to train deep neural networks for image completion. It is trained to complete the center region using pixel-wise reconstruction and single discriminator loss. Some approaches in the literature introduce two discriminators/critics [15, 21, 44] as adversarial losses, where one discriminator/critic considers the entire image and the other focuses on a small hole area to enforce local consistency. However, the main issue for these approaches is that they assume the occluded/inpainting position is known during training and testing. The generator takes the masked image as input and outputs the generated image, and finally, it replaces pixels in the non-masked region of the generated image with the original pixels.

Regarding video completion, there are very few works in the literature. Vondrick et al. [38] first proposed a video generative network for video generation and predicted the future frame using the DCGAN model [31] and spatio-temporal three-dimensional (3D) convolutions [18, 35]. Later, Kratzwald et al. [20] improved the video generative network using WGAN with a gradient penalty critic network and applied it to multi-functional applications.

With the goal of achieving high accuracy for gait recognition in the presence of a high- or poor-quality generated silhouette sequence, we propose a conditional generative network for silhouette sequence reconstruction using spatio-temporal 3D convolution [18, 35] with a dilated kernel [43] in a bottleneck layer to increase the more receptive fields of the output neurons while maintaining a constant number of learnable weights. To regularize the generative networks, we explore triplet hinge loss incorporating WGAN with gradient penalty loss.

3 Spatio-temporal silhouette sequence reconstruction

The goal of the proposed approach is to reconstruct a silhouette sequence from an occluded sequence based on conditional GANs. An overview of our approach is shown in Fig. 2. The proposed approach uses generator G and critic D networks. A single generator network is used for the reconstruction, whereas the additional network critic is used to supervise the generator network during training to realistically reconstruct and preserve subject identity. After training, generator G can take an occluded silhouette sequence and reconstruct it.

Different from existing video generative approaches [20, 38], we propose to design a very deep architecture for the generator network considering the spatio-temporal 3D convolution with small kernels along with dilated convolution and a skip connection; we will explain the detail in Section 3.1. Regarding the critic network, we chose similar critic architecture to [20]. However, the training procedures are different; we will explain the detail in Section 3.2.

3.1 Generator network

Generator G is designed as a simple encoder-decoder pipeline. The occluded input silhouette sequence to the encoder is first mapped to hidden representations, which allows low memory and low computational cost by decreasing the spatial and temporal resolutions. Unlike a pooling layer, the encoder decreases the resolution twice using strided convolutions to avoid a blurred texture in the occluded regions. Then, the decoder takes this low-dimensional feature representation and restores it to the original spatial and temporal resolution through the convolutional layers with fractional strides [47]. Unlike [15, 38], we use convolution kernels of $3 \times 3 \times 3$ (time $\times$ width $\times$ height) and $4 \times 4 \times 4$ because it is proven that small kernels perform better in a deep 3D network [35]. An illustration of the generator network architecture is shown in Fig. 3.

We use dilated convolution [43] in the mid-layer and a skip connection [32] in the top layers. The dilated convolutional kernels are spread out in spatio-temporal directions, which allows us to compute each output pixel by considering a much larger input area, whereas the number of parameters and computational cost still remains constant. This is important for the silhouette sequence reconstruction from a partially observable occluded sequence because the spatial context and neighbor frame information are critical for reconstruction. To keep unoccluded input pixels in the reconstructed sequence, we use a U-shape-like network with skip connections (i.e., feature map of the encoder are combined with the decoder) because the decoder path is more or less symmetric to the encoder path.

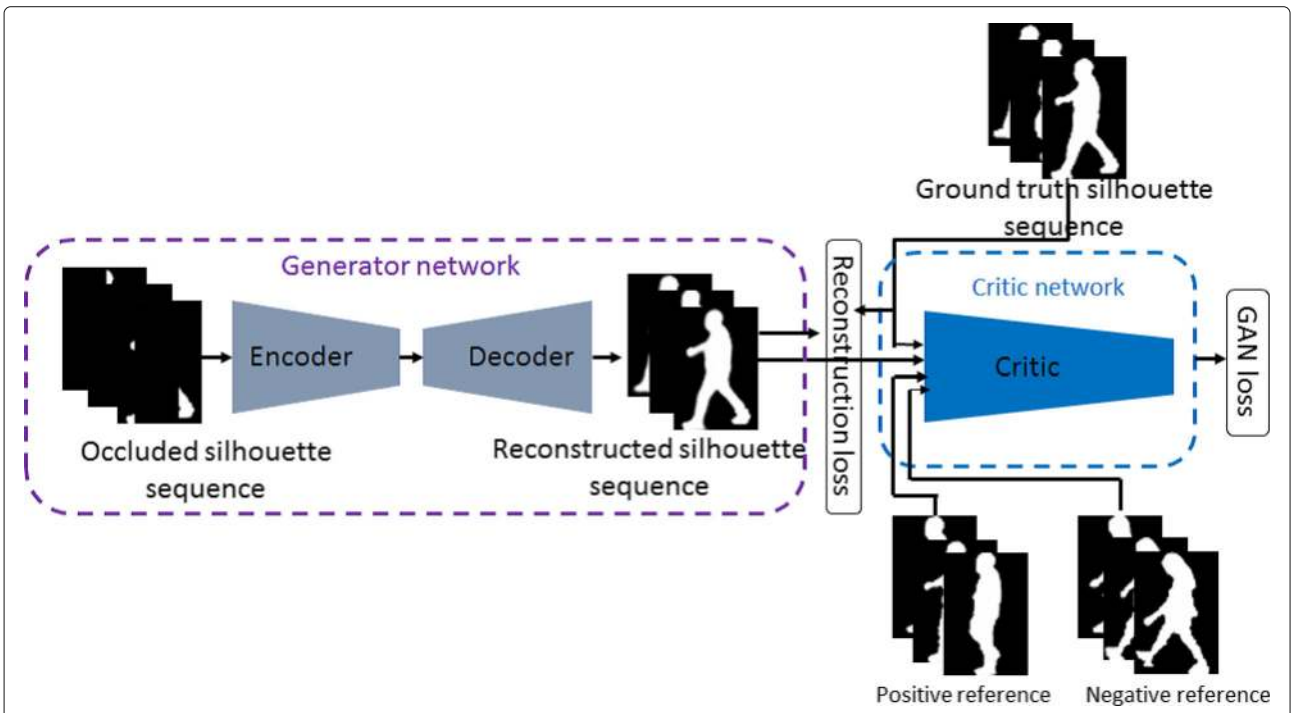

Fig. 2 Overview of our silhouette sequence reconstruction framework. It consists of a generator (encoder and decoder) and a critic network. The generator takes the occluded silhouette sequence as input and outputs the reconstructed silhouette sequence. The critic is used to supervise the generator network during training (i.e., positive reference is unnecessary for the target subject reconstruction during testing)

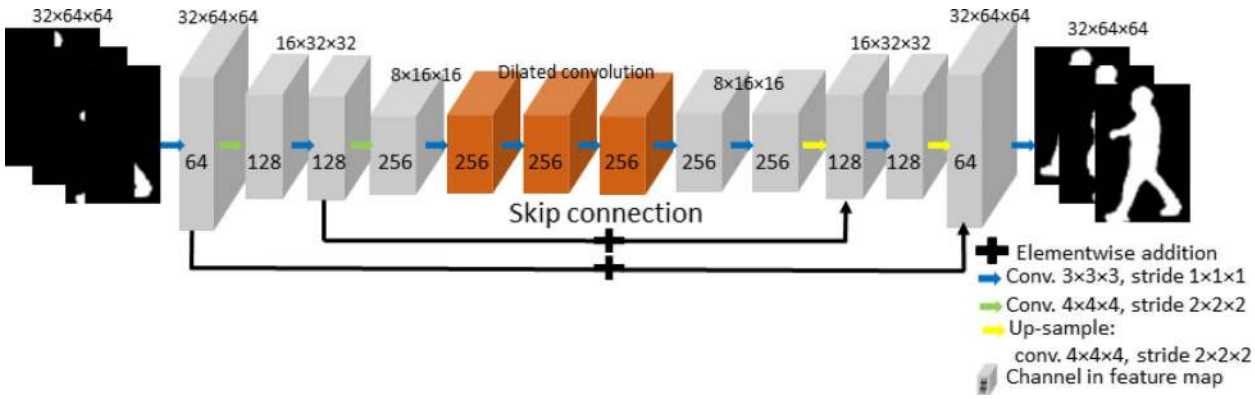

Fig. 3 Illustration of the architecture of the generator network. The silhouette sequence and feature dimensions are shown in the figure and denoted as "time × width × height"

We initialize the convolutional weights for stable training and faster convergence as [11] and perform batch normalization [16] to zero mean and unit variance followed by ReLU activation functions after each layer, except the final output layer. A hyperbolic tangent function is used in the last layer, which is beneficial for normalizing the reconstructed sequence within the range $[-1, 1]$.

3.2 Critic network

Different from existing GANs [1, 9, 20] in which a discriminator/critic distinguishes generated samples from ground truth samples and adversarial supervision of the generator network maximally fools the discriminator, in a different direction, we propose exploring an updated WGAN. Our proposed critic network, D, can distinguish a silhouette sequence of a subject from ground truth and simultaneously use the pairwise similarity ranking, in which the critic network assigns a smaller distance to a silhouette sequence of the same subject and larger distance to a different subject, and it is realized using hinge loss. Using hinge loss along with WGAN loss, we use the adversarial loss so that the generator can maximally fool the critic.

The architecture and layer settings are similar to [20]. Specifically, we use five convolutional layers, followed by a linear downsampling layer with $4 \times 4 \times 4$ kernels along with a stride of $2 \times 2 \times 2$. We set the number of output channels for the first layer to 64 and double the values as the layer gets deeper. Similar to DCGAN [31], we use LeakyReLU [41] with a threshold of 0.2. Similar to [9], we use layer normalization [2] instead of batch normalization. Because the critic is not trained to classify between the reconstructed silhouette sequence and ground truth, we exclude softmax or any other activation in the final layer and instead train the network to provide good gradient information for generator updates.

3.3 Training objective

To train our networks, we use objective functions composed of silhouette sequence reconstruction loss, WGAN loss along with hinge loss as adversarial loss. Given occluded silhouette sequences z and corresponding ground truth sequences x along with a positive reference $\bar{x}$ and a negative reference $\bar{\bar{x}}$, respectively, as the same and different subject as ground truth, our proposed approach is trained to minimize the generative loss for generator network G:

$$L_{\text{gen}} = L_{\text{adv}} + \gamma L_{\text{img}}, \tag{4}$$

where γ is a weighting parameter to control the trade-off between adversarial L_{adv} and image loss L_{img}.

Image loss L_{img} calculates the mean squared error, which attempts to minimize the pixel-wise error between the reconstructed ($\tilde{x} = G(z)$) and ground truth silhouette sequence. It is well known that stabilizing the adversarial training is a significant issue in GANs. A loss in image space is added with adversarial loss, and the loss in image space can contribute to stabilizing the training [6]. We, therefore, employed the image loss L_{img} with adversarial loss in our proposed approach, which can be defined as follows:

$$L_{\text{img}} = \mathop{\mathbb{E}}_{\tilde{x},x\sim\mathbb{P}_g,\mathbb{P}_r}\left[(\tilde{x} - x)^2\right], \tag{5}$$

where $\mathbb{P}_g$ and $\mathbb{P}_r$ represent the distributions of reconstructed silhouette sequence $\tilde{x}$ and ground truth x, respectively.

Adversarial loss L_{adv} is the generator loss in adversarial training, which is the combination of WGAN loss and triplet ranking hinge loss, which can be defined as follows:

$$L_{\text{adv}} = L_{\text{WGAN}} - \kappa L_{\text{hinge}}, \tag{6}$$

where $L_{\text{WGAN}} = -\mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}\left[D(\tilde{x})\right]$ is the WGAN loss, L_{hinge} is the hinge loss for pairwise similarity ranking, and κ is the coefficient to control the trade-off between WGAN and the proposed hinge loss. The output of critic network D is a real-valued scalar, and the hinge loss is calculated using the relative distance of the output of the reconstructed silhouette sequence with the positive reference (i.e., the silhouette sequence of same subject to the reconstructed silhouette sequence) and negative reference (i.e., the silhouette sequence of a different subject to the reconstructed silhouette sequence). Specifically, the triplet pairwise ranking hinge loss function can be defined as follows:

$$L_{\text{hinge}} = \max(\text{margin} - \mathop{\mathbb{E}}_{\tilde{x},\bar{\tilde{x}}\sim\mathbb{P}_g,\mathbb{P}_{\bar{\tilde{x}}}}\left[|D(\tilde{x}) - D(\bar{\tilde{x}})|\right] + \mathop{\mathbb{E}}_{\tilde{x},\bar{x}\sim\mathbb{P}_g,\mathbb{P}_{\bar{x}}}\left[|D(\tilde{x}) - D(\bar{x})|\right], 0), \tag{7}$$

where $\mathbb{P}_{\tilde{x}}$, $\mathbb{P}_{\bar{x}}$, and $\mathbb{P}_{\bar{\tilde{x}}}$ represent the distributions of reconstructed $\tilde{x}$, positive reference $\bar{x}$, and negative reference silhouette sequence $\bar{\tilde{x}}$, respectively.

Similar to generator network G, we train critic network D using the framework of the improved WGAN with a gradient penalty [9] together with the proposed hinge loss. Specifically, critic network D is trained to minimize the following loss function:

$$L_{\text{critic}} = \mathop{\mathbb{E}}_{\tilde{x},x\sim\mathbb{P}_g,\mathbb{P}_r}\left[D(\tilde{x}) - D(x)\right] + \lambda L_{GP} + \kappa L_{\text{hinge}}, \tag{8}$$

where $L_{GP} = \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[\left(\left\|\nabla_{\hat{x}} D(\hat{x})\right\|_2 - 1\right)^2\right]$, $\hat{x} = \epsilon x + (1-\epsilon)\tilde{x}$, and λ is a gradient penalty coefficient and $\epsilon \sim U[0,1]$. We used Adam optimization [19] to update both network G and D with a batch size of 32 and learning rate of $\alpha = 0.0001$ for fixed number of iterations n for the generator network. The other hyperparameters for the Adam optimizer were set to $\beta_1 = 0.5$ and $\beta_2 = 0.99$. Algorithm 1 shows the complete algorithm for training our proposed framework in this paper. We used default $\lambda = 10$, as suggested in [9], and $\gamma = 1000$ according to [20]. The values of the coefficients κ and *margin* were determined empirically as 20 and 3, respectively, for each experiment. All the networks were implemented in Python with the Tensorflow library, and all experiments were trained from scratch. We normalized all silhouette sequences to be in the range $[-1, 1]$.

Algorithm 1 Training of our proposed framework. We use default values $n_{critic} = 4$, $\alpha = 0.0001$, $\lambda = 10$, $margin = 3$, $\gamma = 1000$, $\kappa = 20$, $\beta_1 = 0.5$, $\beta_2 = 0.99$

Require: Batch size b, training iterations n, gradient penalty coefficient λ, number of critic iterations per generator iteration n_{critic}, coefficient κ, width W, height H of silhouette, Adam hyperparameters α, β_1, β_2

Require: Initial critic parameter W_{D_0}, initial generator parameter W_{G_0}

1: **for** $iter \leftarrow 1$ to n **do**

2: **for** $i = 1, ..., n_{critic}$ **do**

3: Sample batches for occluded silhouette sequences z, ground truth silhouette sequences x, positive reference $\bar{x}$ and negative reference $\bar{\tilde{x}}$, a random number $\epsilon \sim U[0,1]$

4: Update the weight W_D of critic network D using Eq.(8) :
$\tilde{x} = G(z)$, $\hat{x} = \epsilon x + (1-\epsilon)\tilde{x}$
$L_{Wdist} = \frac{1}{b}\sum_{j=1}^{b} D(\tilde{x_j}) - D(x_j)$,
$L_{GP} = \frac{1}{b}\sum_{j=1}^{b}(\|\nabla_{\hat{x_j}} D(\hat{x_j})\|_2 - 1)^2$,
$L_{hinge} = \max(margin - \frac{1}{b^2}\sum_{j=1}^{b}\sum_{k=1}^{b}\left|D(\tilde{x_j}) - D(\bar{\tilde{x}_k})\right| + \frac{1}{b}\sum_{j=1}^{b}\left|D(\tilde{x_j}) - D(\bar{x_j})\right|, 0)$
$W_D \leftarrow Adam(\nabla_{W_D}(L_{Wdist} + \lambda L_{GP} + \kappa L_{hinge}), W_D, \alpha, \beta_1, \beta_2)$

5: **end for**

6: Sample batches for occluded silhouette sequences z, ground truth silhouette sequences x, positive reference $\bar{x}$ and negative reference $\bar{\tilde{x}}$

7: Update the weight W_G of generator network G using Eq.(4):
$L_{img} = \frac{1}{bWH}\sum_{j=1}^{b}(\tilde{x_j} - x_j)^2$, $L_{adv} = \frac{1}{b}\sum_{j=1}^{b} -D(\tilde{x_j}) - \kappa L_{hinge}$
$W_G \leftarrow Adam(\nabla_{W_G}(L_{adv} + \gamma L_{img}), W_G, \alpha, \beta_1, \beta_2)$

8: **end for**

4 Experiments

4.1 Overview

To evaluate the accuracy of the proposed approach against a wide variety of occlusion patterns, we artificially simulated several occlusion patterns because there is no large-scale gait recognition database with occlusion variation that is publicly available, and systematic analysis for multiple occlusion patterns is necessary for evaluation. Regarding the evaluation, we used three sets of experiments to validate the proposed approach. These experiments were intended to address a variety of challenges for different occlusion patterns and different training settings that simulate multiple scenarios. We compared the results with the state-of-the-art approaches. The purposes

of these experiments were to evaluate gait recognition for the following conditions:

1. The occlusion pattern was known and the same for a matching pair (probe against gallery)
2. The occlusion pattern was known and different for a matching pair
3. The occlusion pattern was unknown for a matching pair

4.2 Dataset

We used the OU-ISIR Gait Database, Multi-View Large Population Dataset (OU-MVLP) [34], which is composed of gait image sequences with multiple views from 10,307 subjects, with a wide variety of ages and equal distribution of males and females. The image sequences were captured in a controlled environment with a green background for 25 fps using cameras placed approximately 8 m from the course at a height of 5 m. The silhouette sequence was extracted using a chroma key technique, and then the size was normalized by considering the top, bottom, and horizontal center of the silhouette regions for the subject of interest such that the height was 64 pixels and the aspect ratio of each region was maintained. Finally, 44×64 pixels silhouette images were generated. For our experiments, we choose a subset of side views and included only subjects (9001) that had at least 2 sequences. To simulate occlusion, 32 contiguous size-normalized silhouettes of a sequence were used; if a sequence had fewer than 32 samples, we repeated the last frame to make it uniform.

4.2.1 Occlusion pattern

We considered two categories of real-world occlusion that could occur in daily life, that is, relative dynamic and relative static occlusion, together with one artificial random occlusion. Regarding relative dynamic occlusion, we simulated an occlusion type in which a person walked from right to left occluded by a beam, pillar, or tree covering the entire height (e.g., Fig. 1a). As a result, we can imagine that occluder objects move from left to right in a continuous motion within the subject of interest in an image sequence if the person walked at a constant speed. To realize this pattern, we added a background rectangle mask (i.e., set to zero in the occluded position) to cover a certain area against the entire silhouette in the left-most position of first frame of a sequence, and gradually changed the position of the mask toward the end of the frame with the right-most position. Later, we refer to this type of occlusion as a relative dynamic occlusion from left to right (RDLR). Similarly, we simulated the relative dynamic occlusion from bottom to top (RDBT) when an occluder occluded a person from the bottom to the top (e.g., Fig. 1b).

Regarding relative static occlusion, we added a background mask in a fixed position for all frames in a sequence. Therefore, we simulated relative static occlusion in the bottom (RSB), top (RST), left (RSL), and right (RSR) positions. Regarding random occlusion, we added a background mask in a random position in horizontal and vertical directions across the silhouette sequence. Later, we refer to this as random occlusion horizontally (RandH) and random occlusion vertically (RandV), respectively. For each silhouette in a sequence, we added 30%, 40%, and 50% degrees of occlusion against the full area for each of type of occlusion. As a result, we simulated a total of 24 occlusion patterns. Figure 4 shows the simulated occluded silhouette sequence for a subject.

4.3 Experimental settings

We divided the total subjects randomly into three disjoint sets of approximately equal size: 3000 training, 3001 validation, and 3000 test subjects. Then, the validation and test sets were divided into two subsets: gallery set and probe set. The validation set was used to select the best iteration number n for experiments, whereas the test set was used to evaluate the accuracy of our proposed approach and other state-of-the-art approaches. Because the number of samples was large for the experiments of unknown occlusion pattern compared with the experiments of known occlusion pattern, it took more iterations to converge. We, therefore, trained the proposed approach using a validation dataset for up to 30,000 iterations for experiments for known occlusion pattern, whereas we used 60,000 iterations for unknown occlusion pattern and saved the learned parameter every 3000 iterations to select the best iteration from them for testing. We followed the same procedure for all other state-of-the-art benchmarks for a fair comparison to select the best learned model using the validation dataset.

OU-MVLP contained multiple subsequences of more or fewer than 32 silhouette frames; therefore, we selected all the subsequences of 32 silhouette frames for training to increase the training sample, and the centered subsequences of 32 samples were used for the validation and test sets where the starting pose was not the same between the probe and gallery. We padded both sides of the width with zeros for each silhouette in a sequence to make a 64×64 pixels resolution from a 44×64 pixel resolution to fit the network. After reconstructing a sequence, we padded it out to make it the original size (44×64) of the silhouette.

4.4 Evaluation method

Unlike the existing conditional video generative approaches [20, 38], those quantitatively evaluate their

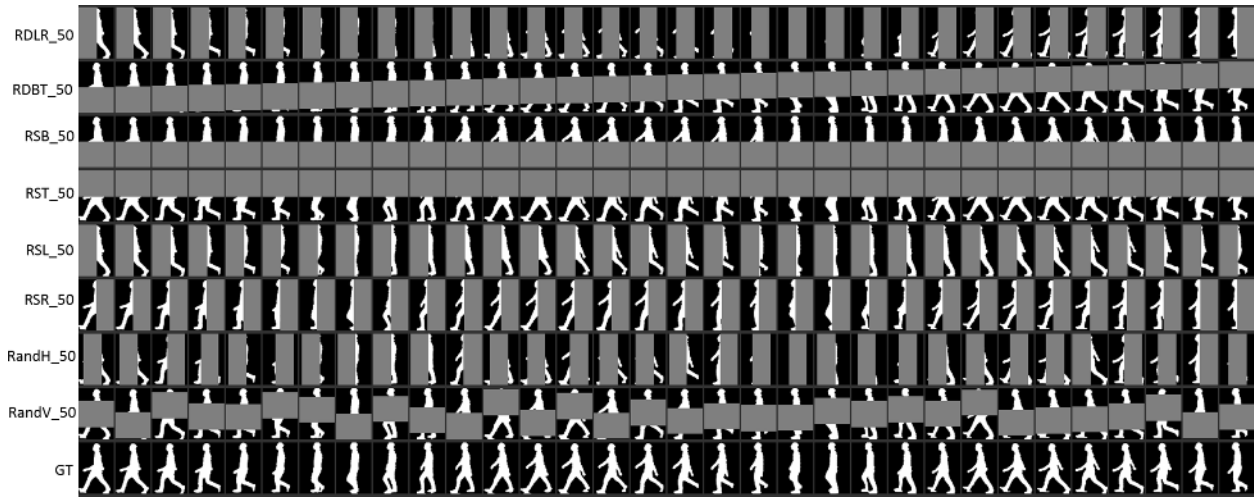

Fig. 4 Example of simulated occlusion for a subject. The left-hand side of the figure: labels for the occlusion pattern, where the first term indicates the type of occlusion and the second term shows the degree of occlusion. The occluded area is gray only for visualization purposes; in the experiment, we masked the occluded area with black, namely the values of the masked area are set to zero; this value is the same for the background

samples by rating manually, but we evaluate the accuracy of gait recognition from reconstructed silhouette sequence.

Because GEI is the most widely used feature in gait recognition, and it can achieve good recognition accuracy, we used the GEI as a gait feature. A GEI was constructed by averaging the subjects' silhouette image sequence over a gait cycle. The gait cycle was determined using normalized autocorrelation [17] of the silhouette image sequence along the temporal axis. If several gait cycles were detected, then we chose the first gait cycle. Finally, we calculated the dissimilarity using the L_2 distance between two GEIs (i.e., probe and gallery).

4.5 Evaluation criteria

We evaluated the accuracy of gait recognition using two modes: identification and verification. We plotted the cumulative matching curve (CMC) for identification and the receiver operating characteristic curve (ROC) for verification, which indicates the trade-off between the false rejection rate of genuine samples and false acceptance rate of imposter samples with varying thresholds. Moreover, we evaluated more specific measures for each evaluation mode: rank-1/5 for identification and the equal error rate (EER) for verification.

4.6 Comparison methods

In this section, we describe the three existing methods used for the evaluation of the experiments. Each of them is a state-of-the-art method for the generative approach. For the comparison, we retrained the model using our dataset from scratch to determine the best-performing model. We used the same hyperparameters as those mentioned in the original papers for the existing methods.

4.6.1 Context encoder [30]

We compared our results with those obtained from the CE, which is a state-of-the-art method for semantic image inpainting. The network architecture is similar to DCGAN [31], that is, the encoder and auxiliary discriminator architecture is similar to that of the discriminator of DCGAN, whereas the decoder is similar to the decoder of DCGAN. However, the bottleneck is 4000 instead of 100. We evaluated the CE by processing the restoration of pixels outside the occluded position for the experiment in which the occlusion pattern is known.

4.6.2 Video GAN (VideoGAN) [38]

VideoGAN is the first model for video generation from random noise. The model is also capable of predicting the future frame given a conditional input frame in the encoder network. Therefore, we adopt it as silhouette sequence reconstruction by changing its input to the occluded silhouette sequence in the encoder network. The architecture of the decoder is similar to that of DCGAN [31], except it is extended in time, whereas we added an encoder network with four strided convolutional layers followed by batch normalization for each layer and a ReLU activation function.

4.6.3 Improved video GAN (iVideoWGAN) [20]

iVideoWGAN is the improved version of VideoGAN. The major modification is that the discriminator network is replaced by a critic network and the network is trained using the framework of WGAN with gradient penalty [9].

In addition to the aforementioned existing methods, we evaluated our proposed generator network using the training of a critic network with WGAN and WGAN-hinge loss. Later, we refer to them as sVideoWGAN and sVideoWGAN-hinge, respectively. Similarly, we evaluated the proposed critic network (WGAN-hinge) with the generator networks of iVideoWGAN [20] and analyzed how the proposed critic could supervise the generator to update the parameter to reconstruct the silhouette sequence. Later, we refer to it as iVideoWGAN-hinge.

4.7 Experiment for the known and same occlusion pattern

In this section, we analyze the accuracy for gait recognition using the reconstructed silhouette sequence where the occlusion pattern is the same between a matching pair (the probe and gallery). To prepare the experiments, we selected typical occlusion patterns from artificially simulated relative dynamic-type occlusion, such as RDLR and RDBT, with the highest and lowest degrees of occlusion (i.e., 30% and 50%). We consequently prepared four subsets of occlusion patterns, denoted by RDLR_30, RDLR_50, RDBT_30, and RDBT_50, where the first and second subscripts indicate the type of occlusion and degree of occlusion, respectively. For the evaluation, the training sets for each subset were prepared in the same manner to reflect the corresponding test sets.

Figures 5 and 6 show the reconstructed silhouette sequences for the occlusion patterns of RDLR_50. From these silhouettes, we can see that sVideoWGAN-hinge, VideoWGAN-hinge, and iVideoWGAN-hinge could reconstruct the silhouette sequence well. In addition, we can also observe that the reconstructed silhouette sequence by comparing with ground truth, sVideoWGAN-hinge is similar with that of sVideoWGAN. We explain the causes in Section 4.9.1.

The results for CMC and ROC are shown in Fig. 7, and Rank-1, Rank-5, and EER are shown in Table 1. From these results, we can see that our proposed generator with the proposed critic (i.e., sVideoWGAN-hinge) outperformed the existing benchmarks in all settings. We can also observe that the proposed generator and proposed critic improved accuracy separately. For example, if we compare the proposed generator and the generator for VideoGAN [20] with the critic of WGAN, referred to as sVideoWGAN and iVideoWGAN, respectively, then accuracy improved from 80.8 to 81.9% and 6.2 to 6.1% (see Table 1) for the Rank-1 and EERs, respectively, for the occlusion pattern of RDLR_30, and 71.3 to 74.7% and 7.4 to 6.8% for RDLR_50. Similarly, the accuracy improved for the proposed generator network from 81.4 to 82.4% and 6.1 to 6.0% for Rank-1 and EERs, respectively, for RDLR_30, and 73.2 to 75.9 and 6.8 to 6.6% for RDLR_50 while the critic was trained with WGAN-hinge. By contrast, the proposed critic WGAN-hinge also (i.e., incorporating hinge loss in WGAN) improved the accuracy separately, for example, 81.9 to 82.4% and 6.1 to 6.0%, for Rank-1 and EERs, respectively, while the generator network was proposed for the type of occlusion pattern of RDLR_30.

Regarding existing benchmarks, CE reconstructed the silhouette sequence in blurred and easy-to-identify areas because it reconstructed only the occluded area frame by frame, which led to a bad recognition accuracy compared with other benchmarks, particularly for a high degree of occlusion. Although iVideoWGAN used an identical generator network to VideoGAN to reconstruct the silhouette sequence, it improved the accuracy for each experiment because the WGAN loss guided the generator network better than that of the discriminator of DCGAN.

4.8 Experiment for the known but different occlusion pattern

In this section, we analyze the accuracy of gait recognition using the reconstructed silhouette sequence where the occlusion pattern is different between a matching pair (the probe and gallery). To prepare such experiments, we selected patterns with the same occlusion type but different degrees of occlusion, and different occlusion types with different degrees of occlusion. Specifically, we compared the gait recognition accuracy of RDLR_30 against RDLR_50 and RDLR_30 against RDBT_50. For the evaluation, in the same manner as the previous experiments in which the occlusion pattern was known, the training sets for each experiment were prepared to reflect the corresponding test sets.

The results for CMC and ROC are shown in Fig. 8, and those for Rank-1, Rank-5, and EER are shown in Table 2. From these results, we can see that the recognition accuracy without reconstruction drastically changed because of the appearance change between the different occlusion patterns. However, the tendency of recognition accuracy for other benchmarks was the same as the experiment for the known and the same occlusion pattern.

4.9 Experiment for the unknown occlusion pattern

In previous sections, we analyzed the experimental results for gait recognition from the reconstructed silhouette sequence within the same and different occlusion patterns, and trained the parameters of CNN using the same occlusion pattern as a test sample. Therefore, we know the occlusion pattern in advance. However, it is difficult to collect such data in a real-world scenario because of the uncooperative and non-intrusive nature of gait biometrics. In this section, we analyze the accuracy of gait recognition when the occlusion pattern is unknown. For this purpose, we trained the parameter of our proposed

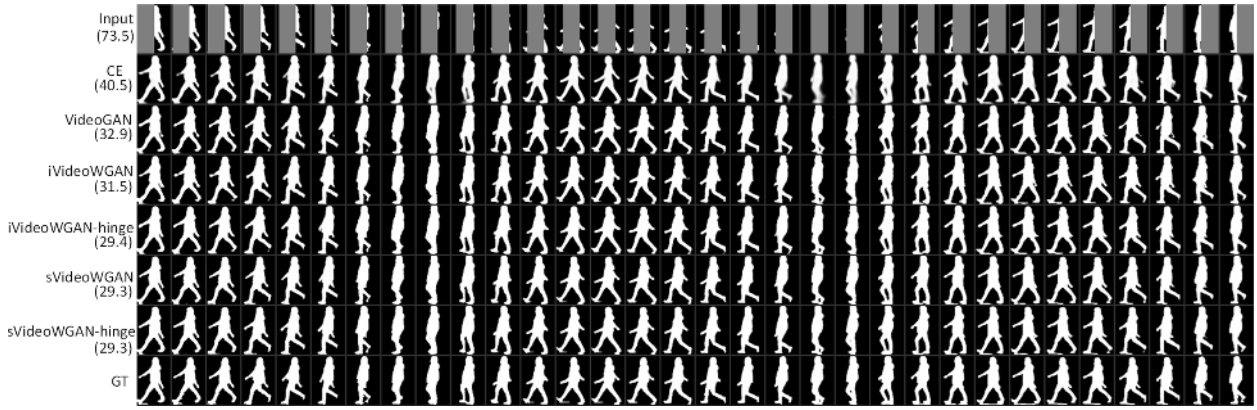

Fig. 5 Reconstructed silhouette sequence for the experiment for the known and same occlusion pattern for RDLR_50. The left-hand side of the figure: second to seventh rows show the labels for the benchmark used to reconstruct the silhouette sequence, whereas the first and last rows show the input and GT, respectively. Values in the parentheses under each label show the average L_2 distance for the reconstructed and the ground truth sequence. Occluded areas are gray only for visualization purposes; in the experiment, we masked the occluded area with black, namely the values of the masked area are set to zero; this value is the same for the background

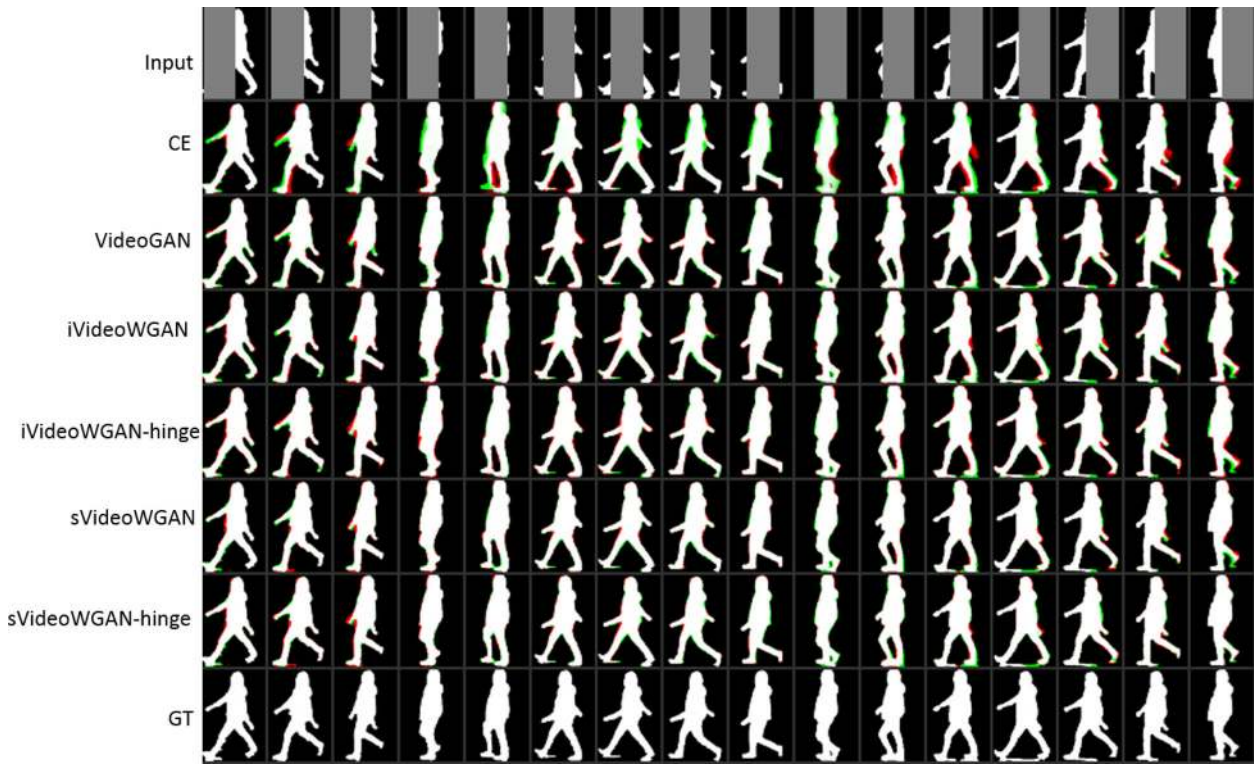

Fig. 6 Reconstructed silhouette sequence (every second frame) for the experiment for the known and same occlusion pattern for RDLR_50 to show how a benchmark can reconstruct silhouette sequence. Green and red colors indicate falsely reconstructed and falsely unreconstructed pixels, respectively, compared with GT. The left-hand side of the figure: second to seventh rows show the labels for the benchmark used to reconstruct the silhouette sequence, whereas the first and last rows show the input and GT, respectively. Occluded areas are gray only for visualization purposes; in the experiment, we masked the occluded area with black, namely the values of the masked area are set to zero; this value is the same for the background

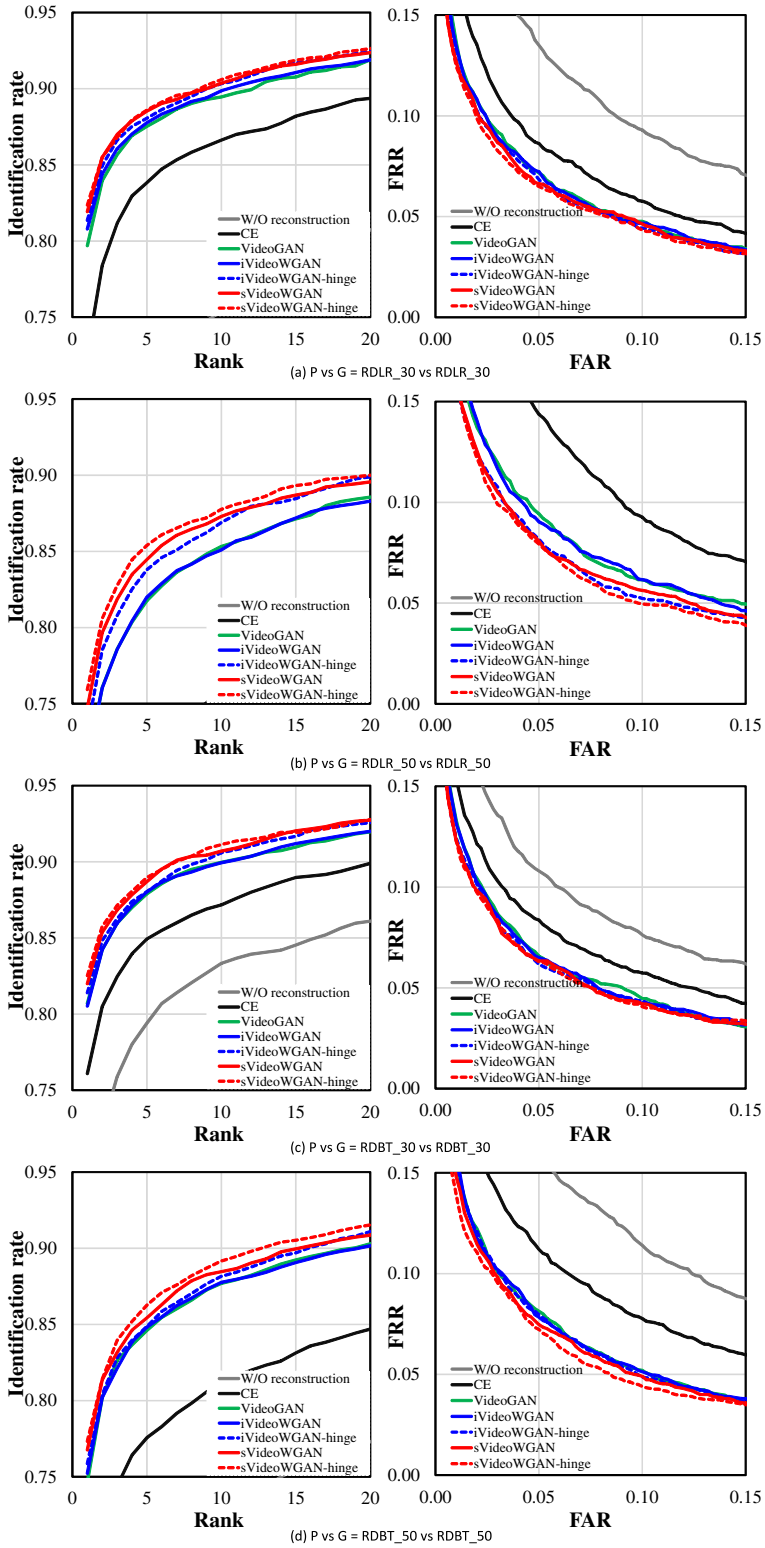

Fig. 7 CMC and ROC curves for the different experiments for the known and same occlusion pattern. The left side shows the CMC curves, and the right side shows the ROC curves; P vs G means occlusion pattern of the probe and gallery, respectively, whereas RDLR_XX and RDBT_XX indicate relative dynamic occlusion left to right and relative dynamic occlusion from bottom to top, respectively, along with the degree of occlusion (XX%). Note that some benchmarks do not provide curves. **a** P vs G = RDLR_30 vs RDLR_30. **b** P vs G = RDLR_50 vs RDLR_50. **c** P vs G = RDBT_30 vs RDBT_30. **d** P vs G = RDBT_50 vs RDBT_50

Table 1 Rank-1/5 [%] and EER [%] for the experiment for the known and same occlusion pattern

Reconstruction method	RDLR_30 vs RDLR_30			RDLR_50 vs RDLR_50			RDBT_30 vs RDBT_30			RDBT_50 vs RDBT_50		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
W/O reconstruction	57.4	70.0	9.5	30.9	39.6	18.6	66.4	79.4	8.4	52.2	66.8	10.8
CE	72.6	83.8	7.2	54.3	70.5	9.5	76.1	84.9	7.0	66.0	77.6	8.5
VideoGan	79.7	87.5	6.3	70.9	81.8	7.3	80.7	87.9	6.0	74.7	84.6	6.7
iVideoWGAN	80.8	87.8	6.2	71.3	82.0	7.4	80.5	88.0	6.0	75.2	84.8	6.7
iVideoWGAN-hinge	81.4	88.1	*6.1*	73.2	83.8	*6.8*	81.4	88.0	**5.8**	75.9	84.9	6.7
sVideoWGAN	*81.9*	*88.5*	*6.1*	*74.7*	*84.5*	*6.8*	*82.0*	*88.7*	6.0	*76.8*	*85.4*	*6.6*
sVideoWGAN-hinge	**82.4**	**88.6**	**6.0**	**75.9**	**85.4**	**6.6**	**82.5**	**89.0**	*5.9*	**77.3**	**86.3**	**6.2**

Bold and italic data indicate the best and second best accuracies throughout the work in this paper, respectively

approach and other benchmark networks by considering all the occlusion patterns using training sets to make a robust model that was capable of reconstructing any type of occlusion pattern. For testing, we used the *cooperative and uncooperative setting* and the *unknown but the same and different occlusion patterns*.

4.9.1 Cooperative and uncooperative setting

The implicit assumption of the uncooperative setting is that the occlusion pattern is inconsistent for all samples throughout the probe and gallery sets [24] (i.e., the occlusion pattern is unknown), whereas for the cooperative setting, the occlusion pattern is consistent for all

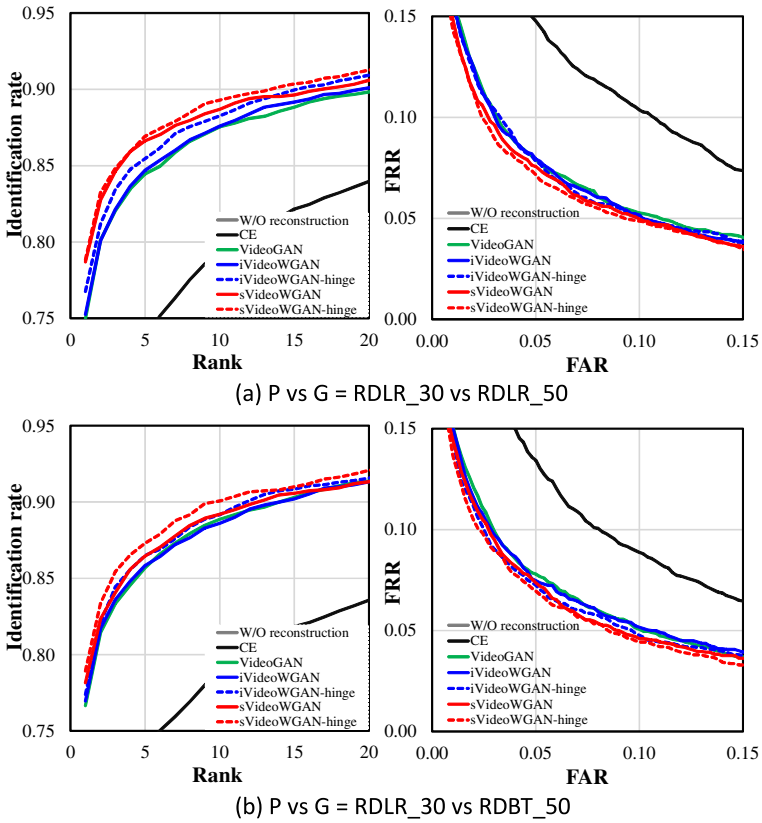

Fig. 8 CMC and ROC curves for the different experiments for the known but different occlusion pattern. The left side shows the CMC curves, and the right side shows the ROC curves; P vs G means the occlusion pattern of the probe and gallery, respectively, whereas RDLR_XX and RDBT_XX indicate the relative dynamic occlusion left to right and relative dynamic occlusion from bottom to top, respectively, along with the degree of occlusion (XX%). Note that some benchmarks do not provide curves. **a** P vs G = RDLR_30 vs RDLR_50. **b** P vs G = RDLR_30 vs RDBT_50

Table 2 Rank-1/5 [%] and EER [%] for the experiment for the known but different occlusion pattern

Reconstruction method	RDLR_30 vs RDLR_50			RDLR_30 vs RDBT_30		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
W/O reconstruction	0.8	2.2	39.6	2.0	5.5	29.2
CE	55.2	73.7	10.3	58.3	73.8	9.2
VideoGan	75.0	84.5	6.8	76.7	85.7	6.7
iVideoWGAN	75.3	84.7	6.6	77.0	85.8	6.6
iVideoWGAN-hinge	76.8	85.5	6.6	*77.4*	*86.5*	6.4
sVideoWGAN	*78.7*	*86.6*	*6.5*	*78.2*	*86.5*	*6.3*
sVideoWGAN-hinge	**78.8**	**86.9**	**6.3**	**78.9**	**87.3**	**6.2**

samples in a gallery set. To create such an uncooperative setting, occlusion patterns were randomly selected for each subject for the probe and gallery sets, whereas for the cooperative setting, ground truth samples were used in the gallery set.

The results for the cooperative and uncooperative settings for CMC and ROC are shown in Fig. 9, and Rank-1, Rank-5, and EER are shown in Table 3. From these results, we can see that the recognition accuracy for the cooperative setting was better than that for the uncooperative setting for each of the benchmarks. We can observe that the accuracy of CE degraded drastically from the cooperative to uncooperative settings compared with other benchmarks. For example, CE degraded the Rank-1 identification by 12%, whereas the maximum degradation for a benchmark was 8.2% (e.g., for iVideoWGAN-hinge). We believe that CE reconstructed the silhouette sequence frame by frame and therefore lost the motion information, particularly when a silhouette was completely occluded, as shown in Figs. 5 and 6. As a result, CE lost subject discrimination.

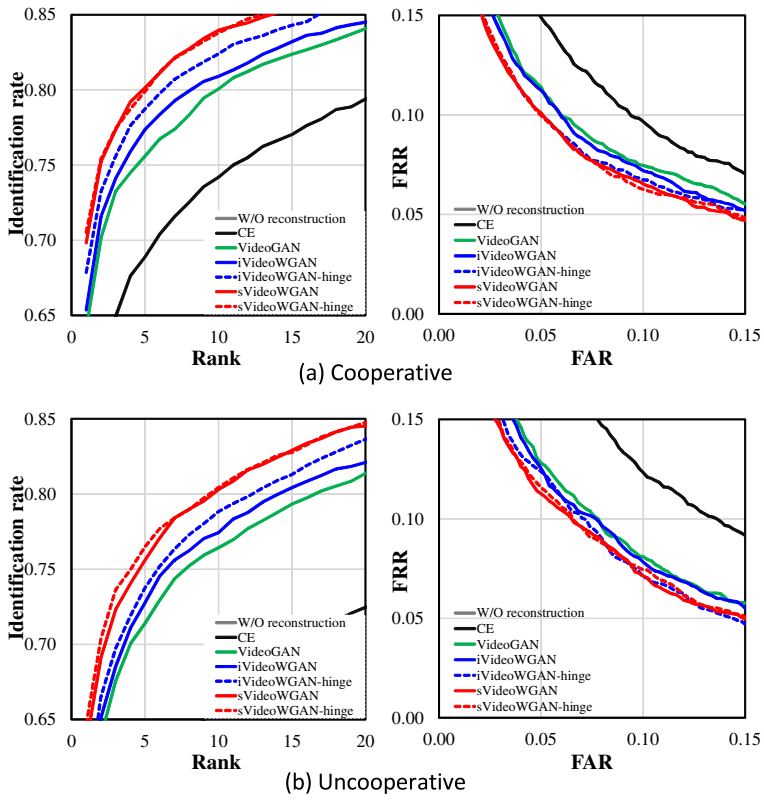

Fig. 9 CMC and ROC curves for the experiment for cooperative and uncooperative settings for the unknown occlusion pattern. The left side shows the CMC curves, and the right side shows the ROC curves. Note that some benchmarks do not provide curves. **a** Cooperative. **b** Uncooperative

Table 3 Rank-1/5 [%] and EER [%] for the experiment for cooperative and uncooperative settings for the unknown occlusion pattern

Reconstruction method	Uncooperative			Cooperative		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
W/O reconstruction	3.9	5.0	46.4	7.4	11.0	45.5
CE	42.1	59.2	11.5	54.1	68.9	9.8
VideoGan	56.8	71.4	8.9	64.1	75.6	8.3
iVideoWGAN	58.4	72.8	8.8	65.4	77.4	8.1
iVideoWGAN-hinge	59.7	73.8	**8.3**	67.9	78.7	7.7
sVideoWGAN	*63.4*	*75.6*	8.5	*69.8*	**80.1**	**7.6**
sVideoWGAN-hinge	**64.7**	**76.5**	*8.4*	**70.5**	*79.9*	**7.6**

We can also observe that sVideoWGAN-hinge did not improve the accuracy from sVideoWGAN for the cooperative setting. We think that the proposed generator network used element-wise addition of the encoder with the decoder to keep the unoccluded silhouette in the reconstructed silhouette as much as possible, and WGAN supervised the generator to reconstruct by comparing the reconstructed sequence with the ground truth sequence. However, the proposed critic (WGAN-hinge) supervised the generator by comparing not only the ground truth but also the positive and negative reference sequences. Therefore, the reconstructed silhouette sequence by comparing with ground truth, sVideoWGAN-hinge is similar or slightly worse than that of sVideoWGAN as shown in Figs. 5 and 6.

4.9.2 Unknown but the same and different occlusion pattern settings

Because the learned parameter of CNN for the experiment for the unknown occlusion pattern can reconstruct any type of occlusion pattern considered in this research, we selected the same and different occlusion patterns between the probe and gallery for evaluation. Hence, we chose the RDLR_30 occlusion pattern as the probe; two typical occlusion patterns for each type of relative dynamic occlusion, such as RDLR_30 and RDLR_50, and RDBT_30 and RDBT_50, together with the ground truth silhouette sequence as the gallery. Therefore, we could compare the accuracy of learned parameters of CNN for unknown occlusion patterns with known occlusion patterns.

The results for CMC and ROC are shown in Fig. 10, and Rank-1, Rank-5, and EER are shown in Table 4. From these results, we can see that the recognition accuracy for CE degraded for each combination when compared with that of the combination from the known occlusion pattern. For example, Rank-1 and EER were 72.6% and 7.2%, respectively, when the occlusion pattern was known for RDLR_30 vs RDLR_30, and 70.7% and 7.4% for the unknown occlusion pattern. We think that, because the occlusion pattern was unknown and we therefore did not know the occlusion position to replace the original unoccluded input pixel in the output as post processing, the reconstructed silhouette sequence for the experiment for the unknown occlusion pattern is worse than that of known occlusion pattern. Similar to the results for the experiment of cooperative setting, sVideoWGAN-hinge did not improve the accuracy from sVideoWGAN for RDLR_30 versus GT (see Table 4).

We can also see that the identification accuracy degraded for VideoGAN, iVideoWGAN, and iVideoWGAN-hinge when compared with the same combination for the known occlusion pattern; however, the verification accuracy improved. We think that those benchmarks used the same generator network of comparatively shallow architecture and therefore lost inter-subject discrimination when training the parameter for a wide variety of occlusion patterns. However, the proposed generator can manage a wide variety of occlusion patterns to train a robust model and improve accuracy.

5 Conclusion and future work

We focused on gait recognition where all frames in a sequence were occluded. For this task, we proposed an approach based on deep conditional GAN that consisted of a generator and critic networks. It allowed us to reconstruct an unoccluded image from an occluded silhouette sequence for gait recognition. We showed that triplet hinge loss along with WGAN regularized the training of the generative network and reconstructed the silhouette sequence with a high discrimination ability, which led to the better accuracy for gait recognition. To demonstrate the effectiveness of the proposed approach, we considered several occlusion patterns with relative dynamic and relative static occlusion for different degrees of occlusion that were quite common in real-world scenarios and designed a set of experiments in which the occlusion pattern between the probe and gallery was the same/different and known/unknown. The experimental results demonstrated that the reconstructed silhouette sequence of the proposed approach achieved state-of-the-art accuracy. Therefore, we conclude that the proposed approach has the potential to tackle the challenges for gait recognition in the presence of occlusion.

There are a number of limitations that need to be addressed in future work. We considered artificially simulated occlusion for a side view silhouette sequence.

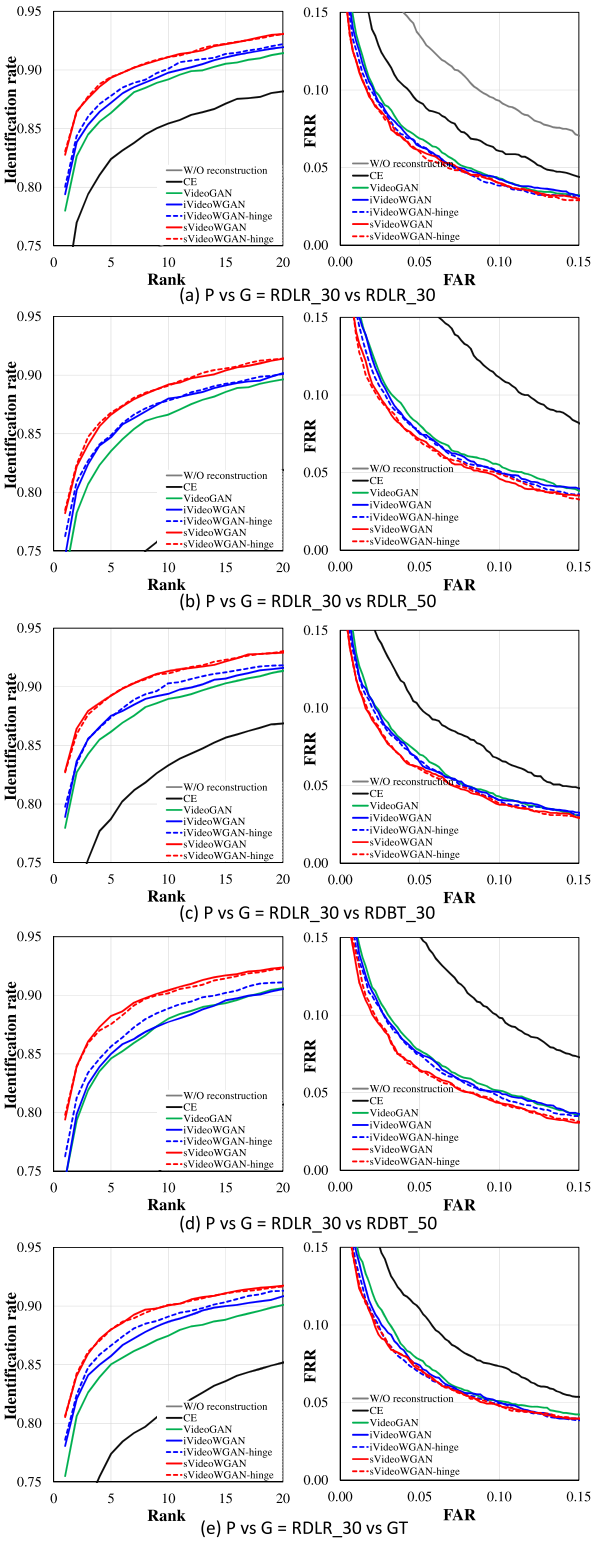

Fig. 10 CMC and ROC curves for the experiment for the unknown but same and different occlusion pattern settings. The left side shows the CMC curves, and the right side shows the ROC curves; P vs G means the occlusion pattern of the probe and gallery, respectively, whereas RDLR_XX and RDBT_XX indicate the relative dynamic occlusion left to right and relative dynamic occlusion from bottom to top, respectively, along with the degree of occlusion (XX%). Note that some benchmarks do not provide curves. **a** P vs G = RDLR_30 vs RDLR_30. **b** P vs G = RDLR_30 vs RDLR_50. **c** P vs G = RDLR_30 vs RDBT_30. **d** P vs G = RDLR_30 vs RDBT_50. **e** P vs G = RDLR_30 vs GT

Table 4 Rank-1/5 [%] and EER [%] for the experiment for the unknown but same and different occlusion pattern settings

Reconstruction method	RDLR_30 vs RDLR_30			RDLR_30 vs RDLR_50			RDLR_30 vs RDBT_30			RDLR_30 vs RDBT_50			RDLR_30 vs GT		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
W/O reconstruction	57.4	70.0	9.5	0.8	2.2	39.6	6.1	15.4	25.6	2.0	5.5	29.2	1.3	2.6	36.7
CE	70.7	82.4	7.4	51.3	70.5	10.7	66.4	78.7	8.1	52.5	69.9	9.9	62.9	77.4	8.1
VideoGan	78.0	86.3	6.2	72.9	83.5	6.8	78.0	86.2	6.2	74.3	84.6	6.7	75.5	85.0	6.5
iVideoWGAN	79.4	87.2	5.9	74.5	84.7	6.5	78.9	87.5	6.0	74.2	85.0	6.5	78.1	85.7	6.4
iVideoWGAN-hinge	80.0	87.8	*5.8*	76.3	84.9	6.5	79.8	87.4	6.0	76.3	85.6	6.3	78.6	86.6	**6.2**
sVideoWGAN	*82.8*	*89.3*	*5.8*	*78.2*	*86.6*	*6.3*	**82.8**	*89.2*	*5.8*	*79.4*	**88.2**	*6.0*	**80.6**	**88.0**	**6.2**
sVideoWGAN-hinge	**83.1**	**89.4**	**5.5**	**78.5**	**86.8**	**6.2**	*82.7*	**89.3**	**5.7**	**79.8**	*87.5*	**5.9**	*80.5*	**88.0**	**6.2**

In the future, we will use occlusion with multiple view variation.

Acknowledgements
We thank Maxine Garcia, PhD, from Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac) for editing the English text of a draft of this manuscript.

Authors' contributions
MZU contributes the most including proposing the initial research idea, generating the dataset, conducting the experiments, and writing the initial draft of the manuscript. MZU and DM analyzed and discussed the evaluated accuracy and revised the manuscript. NT and MARA are responsible for suggesting possible improvements. YY supervised the work and provided technical support. All authors reviewed and approved the final manuscript.

Funding
This work was supported by JSPS KAKENHI Grant Number 15K12037.

Availability of data and materials
The material related to this research can be publicly accessed at OU-MVLP dataset: http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitMVLP.html.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]The Institute of Scientific and Industrial Research, Osaka University, Osaka 567-0047, Japan. [2]The Institute for Datability Science, Osaka University, Osaka 567-0047, Japan.

Received: 8 March 2019 Accepted: 24 October 2019
Published online: 20 November 2019

References

1. M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN. CoRR (2017). abs/1701.07875, 1701.07875
2. L.J. Ba, R. Kiros, G.E. Hinton, Layer normalization. CoRR (2016). abs/1607.06450
3. M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, in *Proc. of the 27th Annual Conf. on Computer Graphics and Interactive Techniques. SIGGRAPH '00*. Image inpainting (ACM Press/Addison-Wesley Publishing Co., New York, 2000), pp. 417–424
4. H. Cai, C. Bai, Y. Tai, C. Tang, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14 2018, Proceedings, Part II*. Deep video generation, prediction and completion of human action sequences, (2018), pp. 374–390. https://doi.org/10.1007/978-3-030-01216-8_23
5. C. Chen, J. Liang, H. Zhao, H. Hu, J. Tian, Frame difference energy image for gait recognition with incomplete silhouettes. Pattern Recogn Lett. **30**(11), 977–984 (2009)
6. A. Dosovitskiy, T. Brox, in *Proc. of the Int. Conf. on Neural Information Processing Systems*. Generating images with perceptual similarity metrics based on deep networks (Curran Associates Inc., USA, 2016), pp. 658–666
7. A.A. Efros, T.K. Leung, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, vol 2*. Texture synthesis by non-parametric sampling, (1999), pp. 1033–1038. https://doi.org/10.1109/iccv.1999.790383
8. I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, in *Proc. of the Int. Conf. on Neural Information Processing Systems - Vol 2*. Generative adversarial nets (MIT Press, Cambridge, 2014), pp. 2672–2680
9. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. Improved training of Wasserstein GANS, (Long Beach, 2017), pp. 5769–5779
10. J. Han, B. Bhanu, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, vol 2*. Statistical feature fusion for gait-based human recognition, (2004), pp. 842–847. https://doi.org/10.1109/cvpr.2004.1315252
11. K. He, X. Zhang, S. Ren, J. Sun, in *Proc. of the IEEE Int. Conf. on Computer Vision*. Delving deep into rectifiers: surpassing human-level performance on imagenet classification (Washington, 2015), pp. 1026–1034. https://doi.org/10.1109/iccv.2015.123
12. M. Hofmann, D. Wolf, G. Rigoll, in *Proc. of the Int. Conf. on Computer Vision Theory and Applications*. Identification and reconstruction of complete gait cycles for person identification in crowded scenes, (Vilamoura, 2011), pp. 594–597. https://doi.org/10.5220/0003329305940597
13. M.A. Hossain, Y. Makihara, J. Wang, Y. Yagi, Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. Pattern Recogn. **43**(6), 2281–2291 (2010)
14. How biometrics could change security, BBC (online). available from http://news.bbc.co.uk/2/hi/programmes/click_online/7702065.stm
15. S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion. ACM Trans Graph. **36**(4), 107:1–107:14 (2017)
16. S. Ioffe, C. Szegedy, in *Proc. of the Int. Conf. on International Conference on Machine Learning - Vol 37*. Batch normalization: accelerating deep network training by reducing internal covariate shift (PMLR, Lille, 2015), pp. 448–456
17. H. Iwama, M. Okumura, Y. Makihara, Y. Yagi, The OU-ISIR Gait Database comprising the large population dataset and performance evaluation of gait recognition. IEEE Trans Inf Forensic Secur. **7**(5), 1511–1521 (2012)
18. S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell. **35**(1), 221–231 (2013)
19. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. CoRR (2014). abs/1412.6980
20. B. Kratzwald, Z. Huang, D.P. Paudel, L.V. Gool, Improving video generation for multi-functional applications. CoRR (2017). abs/1711.11453, 1711.11453
21. Y. Li, S. Liu, J. Yang, M. Yang, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. Generative face completion, (Honolulu, 2017), pp. 5892–5900. https://doi.org/10.1109/cvpr.2017.624
22. Z. Liu, S. Sarkar, Effect of silhouette quality on hard problems in gait recognition. IEEE Trans Syst Man Cybern Part B Cybern. **35**(2), 170–183 (2005)

23. C. Lu, M. Hirsch, B. Schölkopf, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. Flexible spatio-temporal networks for video prediction, (Honolulu, 2017), pp. 2137–2145. https://doi.org/10.1109/cvpr.2017.230
24. R. Martín-Félez, T. Xiang, Uncooperative gait recognition by learning to rank. Pattern Recogn. **47**(12), 3793–3806 (2014)
25. M. Mirza, S. Osindero, Conditional generative adversarial nets. CoRR (2014). abs/1411.1784
26. D. Muramatsu, Y. Makihara, Y. Yagi, in *Int. Conf. on Biometrics (ICB)*. Gait regeneration for recognition, (2015a), pp. 169–176. https://doi.org/10.1109/icb.2015.7139048
27. D. Muramatsu, A. Shiraishi, Y. Makihara, M. Uddin, Y. Yagi, Gait-based person recognition using arbitrary view transformation model. IEEE Trans Image Process. **24**(1), 140–154 (2015b)
28. P. Nangtin, P. Kumhom, K. Chamnongthai, Gait identification with partial occlusion using six modules and consideration of occluded module exclusion. J Vis Commun Image Represent. **36**, 107–121 (2016)
29. J. Ortells, R.A. Mollineda, B. Mederos, R. Martín-Félez, Gait recognition from corrupted silhouettes: a robust statistical approach. Mach Vis Appl. **28**(1), 15–33 (2017)
30. D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, A.A. Efros, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. Context encoders: feature learning by inpainting, (Las Vegas, 2016), pp. 2536–2544. https://doi.org/10.1109/cvpr.2016.278
31. A. Radford, L. Metz, S. Chintala, in *Int. Conf. on Learning Representations*. Unsupervised representation learning with deep convolutional generative adversarial networks, (San Juan, 2016)
32. O. Ronneberger, P. Fischer, Thomas Be, J. Hornegger, W.M. Wells, A.F. Frangi, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. U-net: Convolutional networks for biomedical image segmentation (Springer International Publishing, Cham, 2015), pp. 234–241
33. A. Roy, S. Sural, J. Mukherjee, G. Rigoll, Occlusion detection and gait silhouette reconstruction from degraded scenes. Signal Image Video Proc. **5**(4), 415 (2011)
34. N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Trans Comput Vis Appl. **10**(1), 4 (2018)
35. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, in *Proc. of the IEEE Int. Conf. on Computer Vision*. Learning spatiotemporal features with 3D convolutional networks, (Washington, 2015), pp. 4489–4497. https://doi.org/10.1109/iccv.2015.510
36. M. Uddin, D. Muramatsu, T. Kimura, Y. Makihara, Y. Yagi, MultiQ: single sensor-based multi-quality multi-modal large-scale biometric score database and its performance evaluation. IPSJ Trans Comput Vis Appl. **9**(1), 18 (2017)
37. M. Uddin, T.T. Ngo, Y. Makihara, N. Takemura, X. Li, D. Muramatsu, Y. Yagi, The OU-ISIR Large Population Gait Database with real-life carried object and its performance evaluation. IPSJ Trans Comput Vis Appl. **10**(1), 5 (2018)
38. C. Vondrick, H. Pirsiavash, A. Torralba, in *Advances in Neural Information Processing Systems 29: Annual Conf. on Neural Information Processing Systems 2016*. Generating videos with scene dynamics, (Barcelona, 2016), pp. 613–621
39. C. Wang, H. Huang, X. Han, J. Wang, Video inpainting by jointly learning temporal structure and spatial details. CoRR (2018). abs/1806.08482, 1806.08482. https://doi.org/10.1609/aaai.v33i01.33015232
40. Y. Wexler, E. Shechtman, M. Irani, Space-time completion of video. IEEE Trans Pattern Anal Mach Intell. **29**(3), 463–476 (2007)
41. B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network. CoRR (2015). abs/1505.00853
42. R.A. Yeh, C. Chen, T. Lim, A.G. Schwing, M. Hasegawa-Johnson, M.N. Do, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. Semantic image inpainting with deep generative models, (Honolulu, 2017), pp. 6882–6890. https://doi.org/10.1109/cvpr.2017.728
43. F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions. CoRR (2015). abs/1511.07122
44. J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. Generative image inpainting with contextual attention, (Salt Lake City, 2018), pp. 5505–5514. https://doi.org/10.1109/cvpr.2018.00577
45. S. Yu, D. Tan, T. Tan, in *Proc. of the 18th Int. Conf. on Pattern Recognition vol 4*. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, (Hong Kong, 2006), pp. 441–444. https://doi.org/10.1109/icpr.2006.67
46. S. Yu, H. Chen, E.BG. Reyes, N. Poh, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. GaitGAN: invariant gait feature extraction using generative adversarial networks, (2017), pp. 532–539. https://doi.org/10.1109/cvprw.2017.80
47. M.D. Zeiler, D. Krishnan, G.W. Taylor, R. Fergus, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. Deconvolutional networks, (2010), pp. 2528–2535. https://doi.org/10.1109/iccv.2011.6126474
48. G. Zhao, L. Cui, H. Li, Gait recognition using fractal scale. Pattern Anal Appl. **10**(3), 235–246 (2007)

Publisher's Note