IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 September 2024, accepted 10 October 2024, date of publication 17 October 2024, date of current version 5 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3482430

RESEARCH ARTICLE

Improving Gait Recognition Through Occlusion Detection and Silhouette Sequence Reconstruction

KAMRUL HASAN[1], MD. ZASIM UDDIN[1], (Member, IEEE), AUSRUKONA RAY[1], MAHMUDUL HASAN[2], (Senior Member, IEEE), FADY ALNAJJAR[3], (Member, IEEE), AND MD ATIQUR RAHMAN AHAD[4], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, Begum Rokeya University, Rangpur 5404, Bangladesh
[2]Department of Computer Science and Engineering, Comilla University, Cumilla 3506, Bangladesh
[3]Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain, United Arab Emirates
[4]Department of Computer Science and Digital Technologies, University of East London, E16 2RD London, U.K.

Corresponding authors: Md Zasim Uddin (zasim@brur.ac.bd) and Fady Alnajjar (fady.alnajjar@uaeu.ac.ae)

This work was supported in part by the ICT Division, Government of the People's Republic of Bangladesh, under Grant 1280101-120008431-3631108.

ABSTRACT Gait recognition is an advanced biometric technology that can be used to identify individuals based on their walking patterns, even from low-spatial-resolution image sequences from security surveillance camera footage. Traditional gait recognition approaches rely on complete body information and often overlook the challenge of occlusion. In real-world scenarios, various body parts may be occluded by physical obstacles such as buildings, walls, fences, vehicles, trees, or even other individuals in crowded areas. This occlusion results in a significant portion of the human body being unobserved, causing conventional gait recognition approaches to fail to identify the person. To address this challenge, we have developed a novel framework for gait recognition in the presence of occlusion, incorporating occlusion detection and reconstruction (ODR) and feature extraction for gait recognition (FEGR) modules. The ODR module identifies the occlusion type and reconstructs the occluded portions of the human body in a silhouette sequence using three-dimensional (3D) generative adversarial networks, whereas the FEGR module extracts partwise global and local features using 3D convolutional neural networks (CNNs) and full body features on a frame-by-frame basis using two-dimensional CNNs. We validated our framework using the CASIA-B and OU-MVLP datasets with artificially added occlusions and found that it showed superior performance, with average rank-1 accuracies of 96.4%, 87.8%, and 69.2% for normal, carried object, and clothing variations on CASIA-B and 58.9% on OU-MVLP, as well as 100.0% occlusion detection accuracy. These results demonstrate the ability of our proposed framework to maintain superior gait recognition performance despite the presence of occlusions.

INDEX TERMS Deep learning, feature extraction for gait recognition, gait recognition, gait recognition against occlusion, occlusion detection and reconstruction.

I. INTRODUCTION

Gait recognition is a long-distance behavioral biometric technology that can be used to recognize an individual based on their unique walking patterns. Unlike other biometric methods such as face, fingerprint, and iris recognition, in gait recognition, samples can be captured from a distance without requiring the subject's cooperation [1]. Moreover, gait is difficult for an individual to disguise, as it is an unconscious and natural movement. Therefore, it has various potential uses, including in surveillance systems [2], digital forensics [3], and criminal investigation [4].

Despite their potential, gait recognition systems that work well in controlled laboratory environments often

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaojie Su.

struggle in real-life scenarios owing to several challenges known as covariates. Covariates may be associated with the individual of interest, for instance, carried objects [5], shoes and clothing [6], whereas others are associated with the surrounding environment, such as viewing angle deviations [7], [8], occlusions [9], walking surfaces, and shadows. These covariates have a significant negative impact on gait recognition [10]. Occlusion is considered to be the most critical and challenging covariate [9], as it causes parts of the subject to be missing from image sequences, resulting in deterioration of the quality of the observed samples even when the overall sequences appear to be of good quality. In real-life applications, occlusions are likely to be due to the presence of buildings, walls, fences, vehicles, trees, or other walking individuals, especially in crowded areas such as airports or railway stations. There are two forms of occlusion that may affect gait recognition, depending on how the occluder and the subject are positioned in an image sequence: relative dynamic occlusion and relative static occlusion [9]. In dynamic occlusion, the obscured part of the subject undergoes continuous change throughout the image sequence, whereas in relative static occlusion, the obscured part remains constant [9]. Fig. 1 shows some examples of relative static and dynamic occlusion.

Over the past two decades, research has focused on challenges in gait recognition related to viewpoint variation [7], [8], [11], carried objects [5], [12], and clothing [6] in cases where there is a mismatch of subject samples between the probe and the gallery. However, although partial occlusion of subjects is a more frequent and complex problem in real-world scenarios, it has received far less attention. This gap highlights the need for more effective methods to enhance gait recognition accuracy in cases of occlusion; such methods are essential for real-world applications. In this study, we aim to address the limitations of previous research by developing a novel approach specifically designed to tackle occlusion.

Existing approaches to occlusion can be categorized into reconstruction-free and reconstruction-based methods. Reconstruction-free methods focus on extracting robust features from silhouette sequences for gait recognition. For example, such approaches may extract energy image features over a gait cycle [13], [14], [15] using gait energy image (GEI) [14], [15] or frame difference energy image (FDEI) [13]. However, although good recognition accuracy was achieved with these approaches for a small degree of occlusion, they lose temporal features and do not work when a large portion of the body is occluded. More recently, whole-silhouette sequences have been used to extract features for gait recognition in cases of occlusion. For example, a study [16] estimated the unoccluded portion of the human body in a silhouette, followed by normalization and registration of the body. Moreover, this study introduced a pairwise masking technique to select corresponding visible regions between matching pairs of silhouettes. Finally, GaitGL [11], an existing state-of-the-art approach, was used as the backbone for feature extraction from the masked silhouette sequence and for gait recognition. Moreover, Gupta and Chellappa [17] proposed a framework to detect occlusion and generate occlusion encodings. This information could be used to learn compelling occlusion-aware discriminative features for gait recognition in cases with occlusion using existing state-of-the-art approaches [11], [18], [19] as the backbone.

By contrast, reconstruction-based approaches [9], [20], [21] focus on reconstructing occluded silhouettes and extracting features from the reconstructed silhouette sequence for gait recognition. Roy et al. [20] estimated frames that contained occlusion over a gait cycle using the Gaussian process dynamical model and reconstructed them to extract features for gait recognition; however, this approach required the gait cycle to be determined in advance, which is difficult for severely occluded sequences. To overcome the limitation, a later study [9] used a generative adversarial network (GAN) to reconstruct a silhouette sequence without the need to know the gait cycle in advance. The GEI-based approach was then used for feature extraction and gait recognition. Another study [21] used the VGG-16 model to detect occlusion and reconstruct the occluded frames in a gait sequence, using pose information as a one-hot vector with a variational autoencoder [22]. GEI was then generated using the reconstructed silhouette sequence, and, finally, the GEI was used to extract features for gait recognition using existing methods such as GEINet [23]. However, these approaches are limited by their lack of end-to-end processing; they require separate steps for reconstruction, feature extraction, and gait recognition, making the process time-consuming and complicated.

In response to these challenges, in this paper we present an end-to-end unified framework including two key modules: (a) occlusion detection and reconstruction (ODR) and (b) feature extraction for gait recognition (FEGR). A silhouette sequence with or without occlusion is provided as an input to the network. The ODR module detects whether the sequence has occlusion; if so, it reconstructs the occluded silhouette sequence. Next, the FEGR module performs feature extraction using the reconstructed silhouette sequence and gait recognition. The main contributions of our study can be summarized as follows.

- We propose an end-to-end unified framework that includes ODR and FEGR modules. The ODR module leverages a novel convolutional neural network (CNN)-based network to detect and classify the type of occlusion. Based on the identified occlusion type and position, it reconstructs the occluded portions of the silhouette sequence using a three-dimensional (3D) GAN. The FEGR module uses a 3D CNN to extract global and local spatiotemporal features and a two-dimensional (2D) CNN to perform frame-by-frame full-body feature extraction, thereby enhancing gait recognition performance.
- We demonstrate the effectiveness of our proposed framework through extensive experiments on two

FIGURE 1. Example of occlusions in real-world applications. Top: a static occlusion in which concrete benches occlude a fixed portion of the person in an image sequence; bottom: dynamic occlusion, where pillars occlude a person, and the occluded portion changes over time.

publicly available gait datasets, CASIA-B [24] and OU-MVLP [25], with artificially added occlusions. The experimental results show that our proposed framework achieves superior performance in gait recognition under conditions of occlusion, demonstrating its potential for real-world applications.

II. RELATED WORK

Various approaches have been developed for gait recognition. These can be categorized into conventional and occlusion-focused methods, as summarized in Table 1.

A. CONVENTIONAL GAIT RECOGNITION APPROACHES

1) MODEL-BASED APPROACHES

Model-based approaches focus on constructing 2D and/or 3D representations of the human body to analyze the movement of individual body parts. These methods extract both static and dynamic gait features to enable human recognition. Early approaches [26], [27], [28] fall into this category; these often used simplified physical representations such as the dynamic pendulum and stick models. For example, Yam et al. [26] introduced a method that combined a bilateral symmetry model and a coupled oscillator model inspired by pendulum motion to capture simultaneous thigh and leg movements. This combination facilitated the recognition of both human walking and running patterns. Yoo et al. [28] developed a gait recognition approach that used a neural network to create 2D stick figures from gait silhouettes; these figures were then analyzed using an artificial neural network to extract gait features.

Recent advances in model-based approaches have shifted towards the use of human pose estimation algorithms to generate skeletal data from RGB (red, green, blue) images for gait recognition [29], [30], [31]. For example, PoseGait [29] uses CNNs to estimate 3D human poses (i.e., the 3D coordinates of body joints) from RGB images, and the resulting skeletal data serve as inputs for spatiotemporal feature extraction for gait recognition. Teepe et al. [30] introduced GaitGraph, which leverages graph convolutional networks (GCNs) to extract more refined spatiotemporal features from skeletal data, as demonstrated on the CASIA-B dataset [24]. Subsequently, they developed GaitGraph2 [31], combining multibranch GCNs and residual networks to extract gait features from separate branches. Although model-based approaches are generally robust to various covariates, they can be affected by low resolution of images and inaccuracies in pose estimation, leading to low recognition accuracy.

2) APPEARANCE-BASED APPROACHES

Appearance-based approaches focus on extracting robust features using traditional handcraft or modern CNNs from silhouettes or RGB (Red, Green, Blue) images. They can be categorized into two primary streams: template-based and sequence-based approaches. The template-based approaches [32], [33], [34], [35] transform gait sequences into compact, representative template images by extracting contour features and aggregating spatiotemporal information. For example, Han and Bhanu [32] introduced GEI, representing gait features by averaging a height-normalized silhouette sequence for a complete gait cycle into a single template image. Variations of GEI were later introduced; these included gait entropy image (GEnI) [33], chrono gait image (CGI) [34], and gait flow image (GFI) [35]. These approaches primarily capture motion details, rendering them

TABLE 1. Summary of conventional and occlusion-focused gait recognition approaches.

Approach type	Category [references]	Summary	Key methods
Conventional	Model-based approaches [26]–[31]	Constructs 2D/3D human body models to analyze walking patterns and extract gait features for recognition.	Bilateral symmetry model, GCN
	Appearance-based approaches [11], [18], [19], [32]–[37]	Extracts robust features from silhouettes or RGB images; includes template-based and sequence-based methods.	GEI, GEnI, CGI, GenI, GFI, part-based model, CNN
Occlusion-focused	Reconstruction-free approaches [13], [16], [17], [38], [39]	Extracts robust features to minimize the impact of occlusions through various techniques.	Bayesian frameworks, fractal wavelet descriptors
	Reconstruction-based approaches [9], [20], [21]	Reconstructs occluded silhouettes before feature extraction using neural networks and advanced models.	GAN, Gaussian process modeling

resilient to variations in covariate conditions that affect static parts of the human body.

In recent years, sequence-based approaches [11], [18], [19], [36], [37] have become more popular than template-based approaches for gait recognition, owing to advances in 3D CNNs and the processing power of graphics processing units (GPUs). In these approaches, a silhouette sequence is given as an input to the network for feature extraction and recognition. Chao et al. [18] proposed an approach named GaitSet that considered the silhouette sequence as a set and employed 2D CNNs and 2D max-pooling layers to extract intermediate feature maps. Subsequently, it split these intermediate feature maps using horizontal pyramid mapping to achieve the final feature representation. By contrast, GaitPart [19] used a focal convolution layer that divided the silhouette of a human body into several parts. A micromotion capture module was used for each part to capture spatiotemporal features with greater discrimination. However, such approaches can only extract detailed local information and thus do not provide a sufficiently detailed global feature representation. To overcome this problem, Lin et al. [11] proposed GaitGL, which performs extraction at different levels using a global and local feature extractor (GLFE) module. Moreover, Chai et al. [36] employed Lagrange's equation to determine the significance of second-order temporal features in gait recognition. Although these sequence-based methods have demonstrated impressive performance under normal walking conditions without occlusion, they generally assume the availability of complete and unoccluded silhouette sequences, limiting their applicability in real-world scenarios, where occlusions frequently occur. This limitation may result in incomplete or missing body parts in the silhouettes, in turn leading to degraded recognition accuracy.

B. OCCLUSION-FOCUSED GAIT RECOGNITION APPROACHES

1) RECONSTRUCTION-FREE APPROACHES

Reconstruction-free approaches extract robust features in a manner that is comparatively insensitive to occlusion [13], [38], [39]. Examples include the work of Zhou et al. [38], who proposed a Bayesian framework to fit silhouettes in the presence of noises and occlusions. They introduced a simple articulated model including spatial and temporal parameters that could build substantial priors for extracting gait features. In another study, a fractal scale wavelet-based gait descriptor was introduced [39] to extract features over a complete gait cycle. Chen et al. [13] proposed an approach to mitigate the impact of silhouette incompleteness and occlusions on gait recognition. They divided a height-normalized human silhouette of a gait cycle into clusters; then, denoising was performed in each cluster to calculate the dominant energy image (DEI). Later, the FDEI was computed by combining the DEI of a cluster with the positive difference between consecutive frames. Finally, features were extracted from the FDEI representation to tackle silhouette incompleteness due to segmentation fault and occlusion. Furthermore, Kosin et al. [14] considered occluded module exclusion for handling partial occlusion. First, the GEI was divided into four separate modules. Later, occlusion detection was performed, and the corresponding occluded module was excluded for gait recognition. However, these techniques seem to be less effective in scenarios involving substantial occlusion, especially when the visible areas shared between the sample of the probe and the gallery are minimal; for example, when the upper body of a probe and the lower body of a gallery are occluded.

Recently, silhouette sequences have been used in feature extraction to address the problem of occlusion in gait recognition. Xu et al. [16] introduced a method to estimate and align the visible portions of the body, using a pairwise similarity mask to identify corresponding regions between probe and gallery samples. Similarly, Gupta and Chellappa [17] developed a framework leveraging occlusion detection to enhance discriminative feature learning, employing existing state-of-the-art methods [11], [18], [19] to form the backbone.

2) RECONSTRUCTION-BASED APPROACHES

Reconstruction-based approaches [9], [20], [21] first reconstruct occluded silhouettes and later extract features from

the reconstructed silhouette sequence for gait recognition. Hofmann et al. [40] proposed a reconstruction approach as a preprocessing step that could be used with a gait recognition method. First, they identified occluded silhouettes based on pose; then, clean silhouettes were used to replace the corrupted or occluded silhouettes. Roy et al. [20] predicted occluded frames using the Gaussian process dynamical model over a complete gait cycle; they later reconstructed the occluded silhouette to extract features for gait recognition. However, this approach was limited by the need to determine the gait cycle in advance from an occluded silhouette sequence; this is a particular problem when the silhouette sequence is severely occluded. To overcome this limitation, a conditional GAN-based approach has been proposed [9] to perform silhouette sequence reconstruction, followed by use of a CNN-based approach with GEI features for gait recognition. Moreover, Kumar et al. [21] applied the VGG-16 model to identify occlusions and reconstruct occluded silhouettes, integrating pose information as a one-hot vector within a variational autoencoder. The reconstructed silhouette sequence was subsequently used to create GEI templates for feature extraction and gait recognition. However, these approaches are not end-to-end procedures; the silhouette sequence must be reconstructed first, followed by feature extraction and gait recognition using separate approaches. This makes the process computationally expensive and labor-intensive.

III. PROPOSED METHOD

A. OVERVIEW

We propose an end-to-end unified framework including two modules: ODR and FEGR. The overview of the framework is shown in Fig. 2. First, a silhouette sequence with or without occlusion is provided to the framework. Then, ODR detects the type and position of occlusion using a simple 3D CNN-based network and reconstructs the occluded portion with a 3D GAN. Finally, the FEGR module extracts features from the reconstructed silhouette sequence and performs gait recognition.

B. NOTATION

We denote by B the batch size, and by t_1 and t_2 the numbers of consecutive silhouette frames before and after temporal downsampling, respectively. The input silhouette sequence to the proposed framework is denoted by $S \in \mathbb{R}^{c_1 \times t_1 \times w_1 \times h_1}$, where c_1 is the number of channels, and h_1 and w_1 are the height and width of each frame. The intermediate outputs after the initial 3D CNN convolution are represented as $Y^{p1} \in \mathbb{R}^{c_2 \times t_2 \times h_1 \times w_1}$ and $Y^{p2} \in \mathbb{R}^{c_4 \times t_2 \times h_3 \times w_3}$, respectively, after application of the 3D CNN and max-pool layers.

The global average pooling and the fully connected (FC) layer are denoted by $G_{avg}(.)$ and $FC(.)$, respectively. The probability of predicting the occlusion type is given by $F_{cs} \in \mathbb{R}^L$, where $\hat{L}$ represents the predicted occlusion type. The generator and discriminator networks for the ODR module are denoted by $\mathbb{G}$ and $\mathbb{D}$, respectively. The distributions of the ground truth silhouette sequence and the occluded silhouette sequence are represented as $x \sim p_{(\text{data})}$ and $z \sim p_{(z)}$, respectively. Overall, the loss function for GAN is denoted by $L_{(GAN)}$. During feature extraction by our proposed framework module, various intermediate and final features are represented as X^p, X^p_{GLFE}, X^p_{FLE}, X^{out}_{GLFE}, X^{out}_{SLE}, X^{out}, and X^{final}. The temporal pooling (TP) and generalized-mean (GeM) pooling operations are denoted by $TP(.)$ and $GeM(.)$, respectively.

C. OCCLUSION DETECTION AND RECONSTRUCTION

Occlusion type estimation uses a CNN-based network consisting of multiple 3D CNNs and max-pooling and FC layers to extract features to detect the type of occlusion. A batch of B samples, each consisting of t_1 consecutive silhouette frames, is given to the network as input; this is represented as $S \in \mathbb{R}^{c_1 \times t_1 \times w_1 \times h_1}$, where c_1 is the number of channels, and $(h_1 \times w_1)$ gives the height and width of each input frame. Initially, input silhouette sequence S is passed to two consecutive 3D CNNs with kernel sizes (3,3,3) and (3,1,1), respectively, to provide output as $Y^{p1} \in \mathbb{R}^{c_2 \times t_2 \times h_1 \times w_1}$. Next, Y^{p1} is fed through the 3D CNN and max-pooling layer twice to extract the spatiotemporal features as $Y^{p2} \in \mathbb{R}^{c_4 \times t_2 \times h_3 \times w_3}$. Then, the global average pooling and FC layer are used to classify the type of occlusion as follows:

$$F_{cs} = FC(G_{avg}(Y^{p2})) \tag{1}$$

where $G_{avg}(.)$ and $FC(.)$ are the global average pooling and FC layer respectively. $F_{cs} \in \mathbb{R}^L$ is the probability of predicting the occlusion type, and L denotes the number of the occlusion type. Finally, the maximum probability of the occlusion type is calculated as follows:

$$\hat{L} = arg\_max(F_{cs}) \tag{2}$$

where $\hat{L} \in \{0, 1, 2, \ldots, n\}$ is the predicted occlusion type. After the occlusion type has been estimated, GAN is used to reconstruct the occluded portion of the silhouette sequence based on the information of occlusion type.

Silhouette sequence reconstruction is used to reconstruct the occluded silhouette sequence after estimation of the type of occlusion. For this purpose, we devised a 3D conditional GAN model [9], [41] including generator $\mathbb{G}$ (i.e., encoder and decoder) and discriminator $\mathbb{D}$ networks, as shown in Fig. 2. The generator $\mathbb{G}$ reconstructs the occluded silhouette sequence, whereas the discriminator $\mathbb{D}$ distinguishes between the reconstructed and ground truth silhouette sequence. The overall procedure can be represented as a min-max optimization process:

$$Min_{(\mathbb{G})} Max_{(\mathbb{D})} L_{(GAN)} \tag{3}$$

$$L_{(GAN)} = \mathbb{E}_{x \sim p_{(\text{data})}}[\log \mathbb{D}(x)] + \mathbb{E}_{z \sim p_{(z)}}[\log(1 - \mathbb{D}(\mathbb{G}(z)))] \tag{4}$$

where x represents the samples from the distribution of the ground truth silhouette sequence, $p_{(\text{data})}$, and z represents those from the distribution of the occluded silhouette

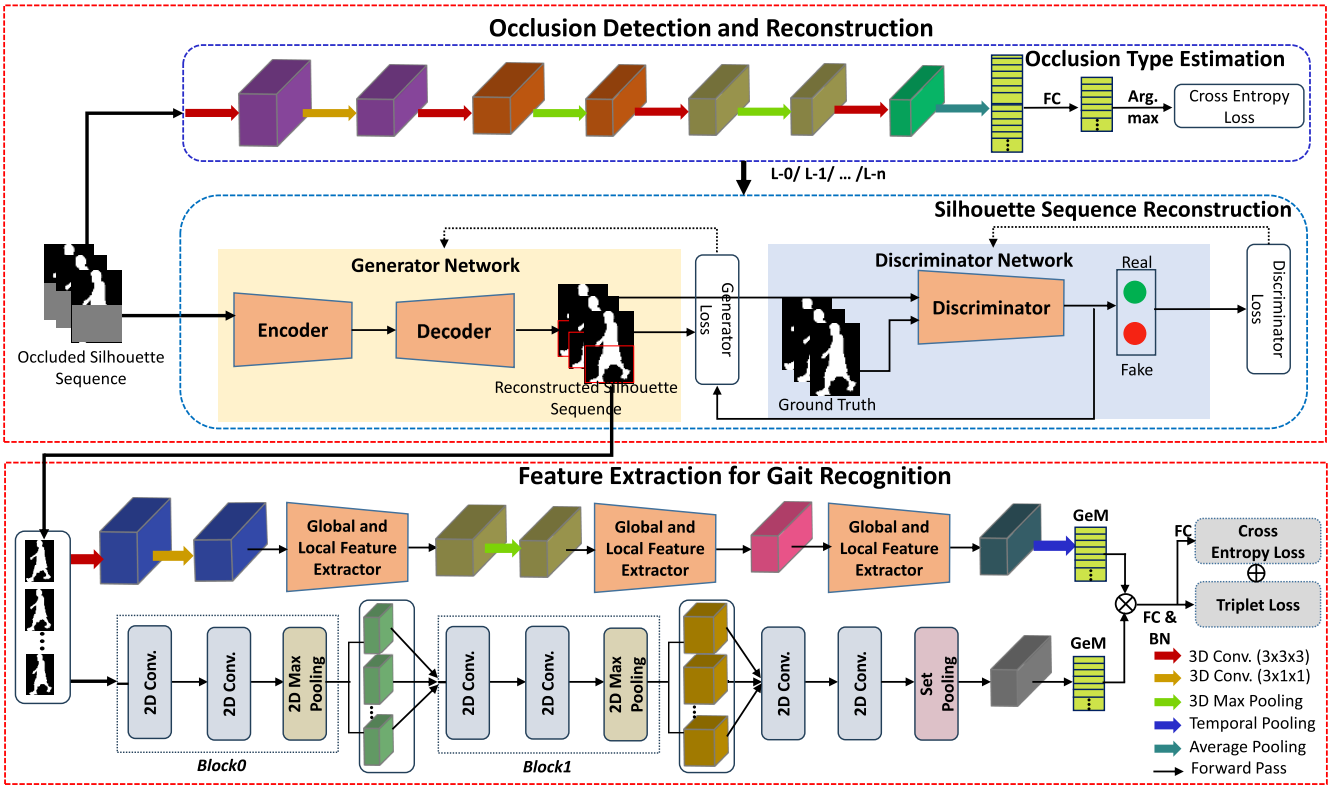

FIGURE 2. Overview of our proposed end-to-end unified framework for gait recognition based on occlusion silhouette sequences. The top module handles occlusion detection and reconstruction (ODR), whereas the bottom module is for feature extraction and gait recognition (FEGR). A silhouette sequence, with or without occlusion, is provided to the framework. The ODR module detects the type and position of occlusion and reconstructs the occluded portion using a generative adversarial network (GAN). The FEGR module then extracts features from the reconstructed silhouette sequence and performs gait recognition. Here, GeM and FC indicate the generalized mean and fully connected layer, respectively.

sequence, $p_{(z)}$. $\mathbb{E}$ is the cross-entropy of the binary classifier of the discriminator. The generator $\mathbb{G}$ aims to minimize its loss by generating the unoccluded silhouette sequence from the occluded one that cannot be discerned by the discriminator $\mathbb{D}$. This can be expressed as follows:

$$Min_{(\mathbb{G})}L_{(GAN)} = Min_{(\mathbb{G})}\mathbb{E}_{Z\sim p(Z)}[\log(1-\mathbb{D}(\mathbb{G}(z)))] \quad (5)$$

The discriminator $\mathbb{D}$ differentiates between the ground truth and reconstructed silhouette sequences and can be expressed as follows:

$$Max_{(\mathbb{D})}L_{(GAN)} = Max_{(\mathbb{D})}\Big\{\mathbb{E}_{x\sim p(\text{data})}[\log \mathbb{D}(x)] + \mathbb{E}_{z\sim p_{(z)}}[\log(1-\mathbb{D}(\mathbb{G}(z)))]\Big\} \quad (6)$$

The generator network $\mathbb{G}$ consists of an encoder and decoder network that use 3D CNNs. The occluded silhouette sequence passes through the encoder network, which down-samples its overall spatial and temporal resolutions to a latent vector. Table 2 shows the overall network architecture of the generator. Each convolution layer in the encoder has a kernel size of $4 \times 4 \times 4$, strides of $2 \times 2 \times 2$, and a ReLU activation function. Except for the first convolution layer, all the 3D CNNs in the encoder network contain batch normalization (BN) layers.

By contrast, the decoder network takes the downsampled latent vector from the encoder network and restores it to the original image space. The decoder also contains four convolution layers with similar but upstrides as the encoder. Several skip connections [42] are used in the generator network to facilitate the direct flow of low-level features to the upsampling layers, as shown in Table 2. The skip connection mitigates the problem of vanishing gradients, ensuring that the network can effectively propagate information from the earlier to later stages of the generator network.

The discriminator network $\mathbb{D}$ determines whether the reconstructed silhouette sequences differ from the ground truth silhouette sequences. As shown in Table 2, it comprises five 3D CNNs, the first four of which have kernel size of $4 \times 4 \times 4$ and stride size of $2 \times 2 \times 2$ and contain visual and motion features. From the second to fourth convolution, it also contains BN layers that reduce internal covariate shifts and improve the stability of the model. The last layer has a kernel size of $1 \times 4 \times 4$ with a stride of $1 \times 1 \times 1$, giving the output a binary classification.

D. FEATURE EXTRACTION FOR GAIT RECOGNITION

The FEGR module takes the reconstructed silhouette sequence as an input for gait recognition. It includes two separate pipelines for extracting features, as shown in Fig. 2. The first is based on GaitGL [11] and is used for global and

TABLE 2. Network structure of the generator and discriminator for reconstructing the silhouette sequence. The symbol ✓ indicates the presence of a task, whereas – indicates its absence.

Network	Layer	Activation	BN	Kernel	Stride	Padding	Input channel	Output channel	Output shape
Generator									
Encoder	Input silhouette sequence	–	–	–	–	–	–	–	$32 \times 64 \times 64$
	Conv3D (Layer 1)	LeakyReLU	–	(4,4,4)	(2,2,2)	(1,1,1)	1	64	$16 \times 32 \times 32$
	Conv3D (Layer 2)	LeakyReLU	✓	(4,4,4)	(2,2,2)	(1,1,1)	64	128	$8 \times 16 \times 16$
	Conv3D (Layer 3)	LeakyReLU	✓	(4,4,4)	(2,2,2)	(1,1,1)	128	256	$4 \times 8 \times 8$
	Conv3D (Layer 4)	LeakyReLU	✓	(4,4,4)	(2,2,2)	(1,1,1)	256	512	$2 \times 4 \times 4$
Decoder	Trans. Conv3D (Layer 5)	LeakyReLU	✓	(4,4,4)	(2,2,2)	(1,1,1)	512	256	$4 \times 8 \times 8$
	Concatenation (Layer 5, Layer 3)								
	Trans. Conv3D (Layer 6)	LeakyReLU	✓	(4,4,4)	(2,2,2)	(1,1,1)	256	128	$8 \times 16 \times 16$
	Concatenation (Layer 6, Layer 2)								
	Trans. Conv3D (Layer 7)	LeakyReLU	✓	(4,4,4)	(2,2,2)	(1,1,1)	128	64	$16 \times 32 \times 32$
	Concatenation (Layer 7, Layer 1)								
	Trans. Conv3D (Layer 8)	LeakyReLU	–	(4,4,4)	(2,2,2)	(1,1,1)	64	1	$32 \times 64 \times 64$
Discriminator									
Critic	Input silhouette sequence	–	–	–	–	–	–	–	$32 \times 64 \times 64$
	Conv3D (Layer 1)	LeakyReLU	–	(4,4,4)	(2,2,2)	(1,1,1)	1	64	$16 \times 32 \times 32$
	Conv3D (Layer 2)	LeakyReLU	✓	(4,4,4)	(2,2,2)	(1,1,1)	64	128	$8 \times 16 \times 16$
	Conv3D (Layer 3)	LeakyReLU	✓	(4,4,4)	(2,2,2)	(1,1,1)	128	256	$4 \times 8 \times 8$
	Conv3D (Layer 4)	LeakyReLU	✓	(4,4,4)	(2,2,2)	(1,1,1)	256	512	$2 \times 4 \times 4$
	Conv3D (Layer 5)	Sigmoid	–	(1,4,4)	(1,1,1)	(0,0,0)	512	1	1

part-based local features, whereas the second uses 2D CNNs and is for extracting frame-by-frame features.

To extract the spatiotemporal features, the reconstructed silhouette sequence is fed through the first pipeline. First, two 3D CNNs with different kernel sizes are applied to the reconstructed silhouette sequence and give the output X^p. Then, the GLFE block takes X^p as its input and performs two separate tasks: first, it uses a single 3D CNN to obtain global features from X^p; second, it performs a partition of X^p to obtain multiple parts. Next, 3D CNNs are applied to part to obtain part-based spatiotemporal features. Global and local part-based features are aggregated, resulting in an output of X^p_{GLFE}. In this experiment, the GLFE module was used three times: global and local features from the first two times were aggregated on the final use by pointwise addition and concatenation. When the merged global and local features have been obtained, TP is performed on X^p_{GLFE}. Then, finally, GeM pooling is used to obtain the final features from this pipeline as follows:

$$X^{out}_{GLFE} = TP(GeM(X^p_{GLFE})) \tag{7}$$

where X^{out}_{GLFE} is the final feature that is added to the features from the second pipeline and then fed through the FC layers.

Similar to the feature extraction process of the first pipeline, a copy of the reconstructed silhouette sequence is fed through the second pipeline as shown in Fig. 2. Here, we consider the 2D CNN for feature extraction based on a frame-by-frame basis; it consists of two blocks with multiple 2D CNNs and 2D max pool layers. First *Block 0* takes the reconstructed silhouette sequence and applies dual 2D CNNs with kernel sizes of (5,5) and (3,3) to each frame. Then, a 2D max pooling layer is used to extract crucial framewise spatial features. *Block 1* takes the output of *Block 0* and applies dual 2D CNNs and a single 2D max pooling layer to extract depth spatial features for each frame, with kernel sizes of (3,3) and (3,3), respectively. Dual 2D CNNs with similar kernel sizes are then used to obtain the frame-level features (FLE) as X^p_{FLE}. We use a set pooling operation to obtain set-level features as X^p_{FLE}, as well as GeM pooling to obtain the final features as follows:

$$X^{out}_{SLE} = SP(GeM(X^p_{FLE})) \tag{8}$$

Finally, the features from both pipelines are concatenated as X^{out}, and the FC layer and BN layer are applied to X^{out} to obtain the final feature for gait recognition as follows:

$$X^{final} = BN(FC(X^{out})) \tag{9}$$

E. LOSS FUNCTION

We use a combined loss function that integrates triplet loss [9] and cross-entropy loss [11] to train the proposed FEGR module efficiently. The triplet loss is used to maximize the distance between distinct classes and reduce the distance within the same class. Conversely, the cross-entropy loss has the classification objective to differentiate between various subjects. During training, the final feature X^{final} is fed into

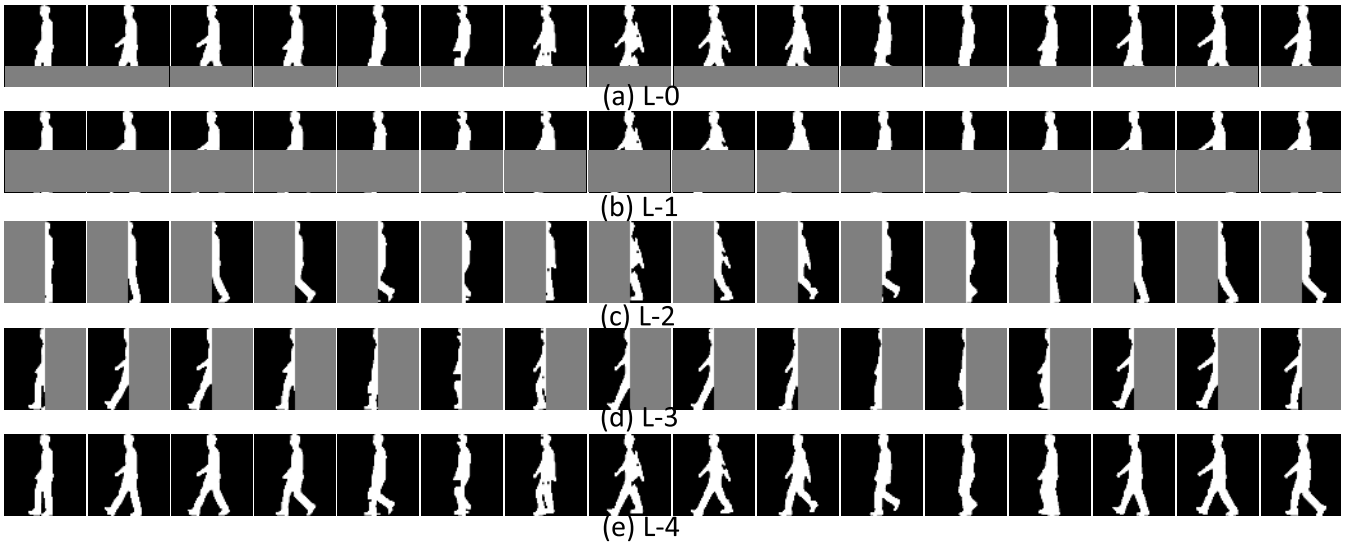

FIGURE 3. Example of simulated occlusion of a subject. The occluded portion is visualized in gray; for the experiments, the occluded area was masked with black (indicating zero values). L-0, L-1, L-2, L-3, and L-4 correspond to different levels of occlusion: L-0 represents 25% lower occlusion, L-1 represents 50% lower occlusion, L-2 represents 50% left occlusion, L-3 represents 50% right occlusion, and L-4 indicates no occlusion.

each loss function separately to compute the respective losses, and the overall combined loss is then determined as follows:

$$L_{total} = L_{triplet} + \gamma L_{cross} \tag{10}$$

where $L_{triplet}$ and L_{cross} denote the triplet and cross-entropy losses, respectively, and γ is a weighting parameter used to control the trade-off between the triplet loss $L_{triplet}$ and cross-entropy loss L_{cross}.

IV. EXPERIMENTS

A. DATASETS

There is no publicly available gait dataset containing both real-world occlusion and the corresponding ground truth. Therefore, similar to the approach used in previous studies [9], [16], [21], [43], we artificially added different occlusions to two popular and publicly available datasets: CASIA-B [24] and OU-MVLP [25]. For purposes of our experiments, we artificially simulated five types of occlusion for each subject: L-0, L-1, L-2, L-3, and L-4, where L-0 denotes masking of the lower 25%, L-1 denotes masking of the lower 50%, L-2 denotes masking of the left 50%, L-3 denotes masking of the right 50%, and L-4 denotes no masking of the silhouette sequence. Fig. 3 shows silhouette sequences with these artificial occlusions for an example subject.

CASIA-B [24] is among the most frequently used datasets in the gait recognition domain. It comprises 124 subjects with 11 views ranging from 0° to 180° with 18° intervals. Each separate view angle consists of ten sequences under three different settings: (i) six sequences collected under normal walking (NM) conditions, (ii) two sequences collected while the subject is carrying a bag (BG), and the remaining two sequences collected with clothing variations (CL). In this study, we used only the 90° view angle for all sequences to demonstrate our proposed framework. The initial 74 subjects were allocated for training; the remaining 50 were set aside for testing. During the testing phase, the first four NM sequences were used to form the gallery without any added occlusion. Occlusion was then artificially added to the last two NM sequences, as well as the two BG and two CL sequences, which were used as probes to evaluate the performance of the framework.

OU-MVLP [25] is among the largest gait datasets, comprising 10,307 subjects. Each subject has 14 different viewing angles (0°, 15°, …, 90°; 180°, 195°, …, 270°) with 15° intervals, where each view has two sequences. For our experiments, we only considered the 90° view. According to the official instructions, odd-numbered IDs were used for training, and even-numbered IDs were used for testing. In the testing phase, the second sequences were used without occlusion as the gallery, and the first sequences with artificially added occlusion were used as probes.

B. TRAINING DETAILS AND TESTING

1) TRAINING

During training, the occluded silhouette sequence is fed into the proposed framework. The ODR module detects and reconstructs the silhouette sequence; this is then passed to the FEGR module, which obtains the output features as X^{final}. The combined loss is then used to compute the loss, incorporating the Batch All (BA_+) triplet loss [11], [19]. The training batch size was set to (M, K), where M represents the number of subjects in each batch, and K represents the number of sequences for each subject.

2) TESTING

During the test phase, the occluded silhouette sequence is fed into the framework. Here, accuracy, precision, recall, and F1-score [44] were used to evaluate the detection of occlusion,

and the average L_2 distance between the reconstructed and ground truth silhouette sequence was used to evaluate the reconstruction accuracy. For evaluation of the FEGR, the overall test dataset was divided into a gallery set and the probe set. The distance between the gallery and probe samples was defined as the average Euclidean distance between the corresponding feature vectors. We used rank-1 accuracy to evaluate the performance of the proposed framework for gait recognition.

C. IMPLEMENTATION DETAILS

To preprocess the silhouette sequence, we used a the process described previously [11]; specifically, the silhouette image was normalized to a size of 64 × 64. After normalizing the silhouette sequence, we artificially added the occlusion, as described in section IV-A. During training, the frame sequence length was set to 32; if the length of the sequence was less than 16 frames, it was discarded, whereas it was repeatedly sampled if it was more than 16 frames and less than 32 frames. All experiments were performed on a single NVIDIA GeForce RTX 3090 Ti GPU using a Linux operating system with Python 3.9.2 and PyTorch 1.10.0. For both datasets, the batch size (M, K) was configured to be (16, 3), where 16 is the number of subjects and 3 is the number of sequences per subject. The whole training and test procedure was performed in an end-to-end manner; specifically, the ODR and FEGR modules were trained and tested concurrently. The learning rate (LR) was set to 0.0001 for the ODR module.

For the FEGR module, we employed different hyperparameters for the CASIA-B and OU-MVLP datasets. For the CASIA-B dataset, the FEGR module was trained for 80k iterations, and the weight decay was set to 5.0e-4. The LR was initially set to 0.0001 and reduced to 0.0000 after 70k iterations. For the OU-MVLP dataset, we trained the FEGR module for 210k iterations with an initial weight decay of 0; after 200k iterations, the weight decay was reset to 0.0005. The hyperparameter γ was set to 1.0 to optimize the loss function for the FEGR module. The LR was initially set to 0.0001 and reset to 0.00001 and 0.000001 after 150k and 200k iterations, respectively.

1) COMPUTATION TIME

Using the aforementioned hardware specification, the training process for the CASIA-B dataset was completed in approximately 3 days, whereas that for the OU-MVLP dataset took approximately 8 days. For testing, the end-to-end process of identifying a subject, including the operations of the ODR and the FEGR modules, took about 1.5 s with the CASIA-B dataset and 1.8 s with the OU-MVLP dataset.

D. ACCURACY OF OCCLUSION DETECTION AND RECONSTRUCTION

Gait recognition accuracy is highly dependent on the quality of silhouette sequence reconstruction, which in turn relies on the detection capabilities of occlusion type estimation. To thoroughly assess the performance of our framework, we evaluated the occlusion detection and silhouette reconstruction processes separately. For occlusion detection, we used accuracy, precision, recall, and F1-score metrics, as outlined in [44]. For silhouette sequence reconstruction, we measured performance by calculating the average L_2 distance between the reconstructed silhouette sequence and the ground truth. The results of these evaluations are presented in Table 3. Figs. 4, and 5 illustrate sample silhouette sequences with artificial occlusion, showing the reconstructed sequences generated by our proposed ODR module and the sVideoWGAN-hinge method [9], alongside the ground truth sequences.

Our proposed ODR module perfectly detected occlusion and obtained 100.0% scores in all cases for each of the considered metrics (accuracy, precision, recall, and F1-score) as shown in Table 3. Moreover, it achieved L_2 distances of 8.3% for the lower 25% occlusion (i.e., L-0) and 11.2% for the lower 50%. Left- and right-side occlusions (i.e., L-2 and L-3) yielded similar performance metrics, indicating that the proposed ODR module can effectively reconstruct a silhouette sequence regardless of whether the occlusion is on the left or right side. These results indicate that our proposed ODR module can perfectly a reconstructed silhouette sequence even when 50% of the human body is occluded.

E. COMPARISON WITH STATE-OF-THE-ART ALGORITHMS

We compared the recognition accuracy of our proposed FEGR module with those of state-of-the-art gait recognition methods, specifically direct matching (DM) [45], sVideoWGAN-hinge [9], GaitSet [18], GaitPart [19], and GaitGL [11]. Initially, we evaluated the gait recognition performance of these methods (GaitSet, GaitPart, and GaitGL) without reconstruction to establish a baseline accuracy under occlusion conditions. Subsequently, we assessed the recognition accuracy using the reconstructed silhouette sequences from our proposed ODR module, demonstrating the effectiveness of our approach.

1) EVALUATION USING CASIA-B

The results for the CASIA-B dataset with and without reconstruction of occluded silhouette sequences are shown in Tables 4 and Fig. 6. Each method for gait recognition showed improved accuracy when the reconstructed silhouette sequence was used, demonstrating the robustness of our approach. For lower-portion occlusion, the accuracy was slightly improved under the NM and BG conditions, whereas there were larger improvements under the CL condition for lower-25% occlusion (i.e., L-0 type occlusion). Greater improvement was seen for 50% occlusion (i.e., L-1 type occlusion). For example, the rank-1 accuracy of GaitSet improved from 2% to 14%, and that of GaitPart from 11% to 16%, whereas the accuracy of our method was enhanced from 3% to 6%.

TABLE 3. Evaluation of occlusion detection and reconstruction of silhouette sequence.

Detection of occlusion					Reconstruction of silhouette sequence
Occlusion type	Accuracy [%]	Precision [%]	Recall [%]	F1-score [%]	Avg. L_2 distance [%]
L-0	100.0	100.0	100.0	100.0	8.3
L-1	100.0	100.0	100.0	100.0	11.2
L-2	100.0	100.0	100.0	100.0	9.3
L-3	100.0	100.0	100.0	100.0	9.9
L-4	100.0	100.0	100.0	100.0	0.0

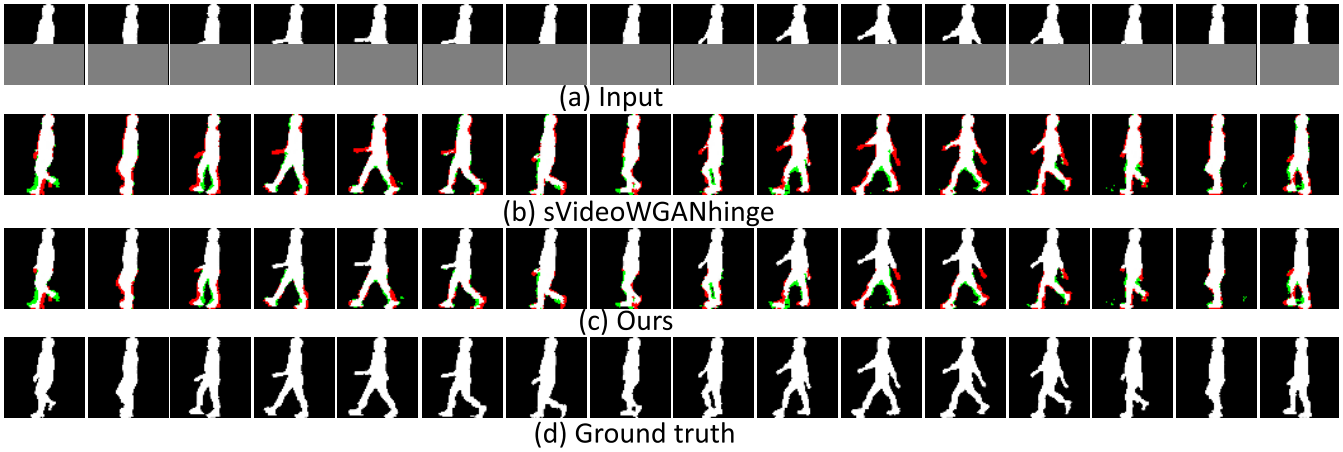

FIGURE 4. Example of a reconstructed silhouette sequence from the CASIA-B dataset. (a) Input silhouette sequence with 50% lower occlusion, (b) the occluded sequence reconstructed using the sVideoWGAN-hinge approach, (c) the occluded sequence reconstructed using the proposed approach, and (d) the ground truth silhouette sequence. Green and red denote pixels that were incorrectly reconstructed or remained unreconstructed, respectively.

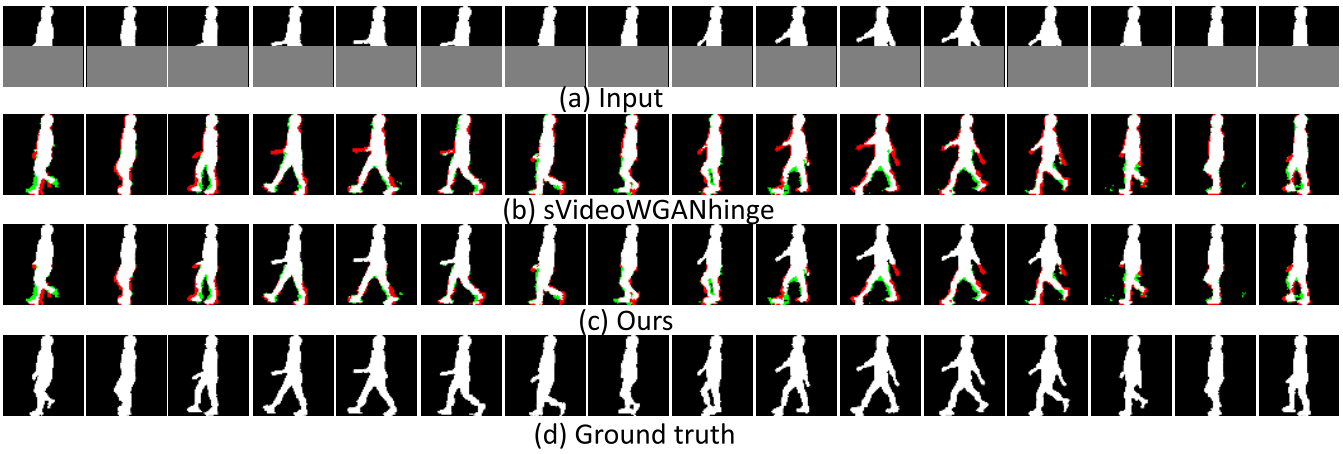

FIGURE 5. Example of a reconstructed silhouette sequence from the OU-MVLP dataset. (a) Input silhouette sequence with 50% lower occlusion, (b) the occluded sequence reconstructed using the sVideoWGAN-hinge approach, (c) the occluded sequence reconstructed using the proposed approach, and (d) the ground truth silhouette sequence. Green and red denote pixels that were incorrectly reconstructed or remained unreconstructed, respectively.

Regarding occlusions on the left and right sides, for left-side occlusions (i.e., L-2), the accuracy of the results was degraded severely when the occluded silhouette sequence was used for gait recognition; for example, the rank-1 accuracy ranged from 5% to 7% for GaitSet and from 9% to 12% for GaitPart. This may have been because subjects were walking from right to left in the CASIA-B dataset; this meant that the visible walking motion was predominantly on the left side, resulting in a negative impact on gait recognition when a subject was occluded on the left side. However, the accuracy improved substantially for each of the benchmarks (GaitSet, GaitPart, and GaitGL) when the reconstructed silhouette sequence was used. For example, the rank-1 accuracy improved from 18% to 51% for GaitSet, and from 24% to 62% for GaitPart. Furthermore, greater improvements in accuracy were observed under the CL and

TABLE 4. Rank-1 accuracy (%) for silhouette sequences from the CASIA-B dataset with different types of occlusion under normal walking (NM), carrying bags (BG), and wearing coats (CL) sequences with (W/) and without (W/O) reconstruction.

Model	Status	Occlusion type															Average
		L-0			L-1			L-2			L-3			L-4			
		NM	BG	CL	NM	BG	CL	NM	BG	CL	NM	BG	CL	NM	BG	CL	
DM [45]	W/O reconstruction	81.0	37.0	22.0	66.0	27.0	12.0	22.0	17.0	11.0	52.0	41.0	20.0	96.0	76.0	38.0	41.2
sVideoWGAN-hinge [9]	W/ reconstruction	90.0	64.0	38.0	80.0	53.0	29.0	32.0	26.0	21.0	76.0	67.0	22.0	96.0	76.0	38.0	53.9
GaitSet [18]	W/O reconstruction	97.0	75.0	34.0	70.0	50.0	21.0	7.0	7.0	5.0	45.0	39.0	18.0	100.0	100.0	94.0	50.8
	W/ reconstruction	98.0	75.0	39.0	84.0	64.0	23.0	58.0	41.0	23.0	96.0	88.0	68.0	100.0	100.0	94.0	70.1
GaitPart [19]	W/O reconstruction	97.0	79.0	31.0	82.0	57.0	20.0	12.00	9.0	12.0	49.0	40.0	23.0	100.0	100.0	98.0	53.9
	W/ reconstruction	99.0	95.0	59.0	93.0	73.0	32.0	74.0	50.0	36.0	98.0	92.0	76.0	100.0	100.0	98.0	78.3
GaitGL [11]	W/O reconstruction	99.0	98.0	69.0	93.0	79.0	34.0	33.0	22.0	19.0	74.0	64.0	34.0	100.0	100.0	99.0	67.8
	W/ reconstruction	100.0	98.0	76.0	95.0	83.0	43.0	82.0	55.0	43.0	99.0	88.0	74.0	100.0	100.0	99.0	82.5
Ours	W/O reconstruction	99.0	98.0	69.0	93.0	79.0	38.0	37.0	22.0	20.0	75.0	66.0	38.0	100.0	100.0	99.0	68.5
	W/ reconstruction	**100.0**	**99.0**	**77.0**	**96.0**	**84.0**	**44.0**	**87.0**	**58.0**	**50.0**	**99.0**	**98.0**	**76.0**	**100.0**	**100.0**	**99.0**	**84.5**

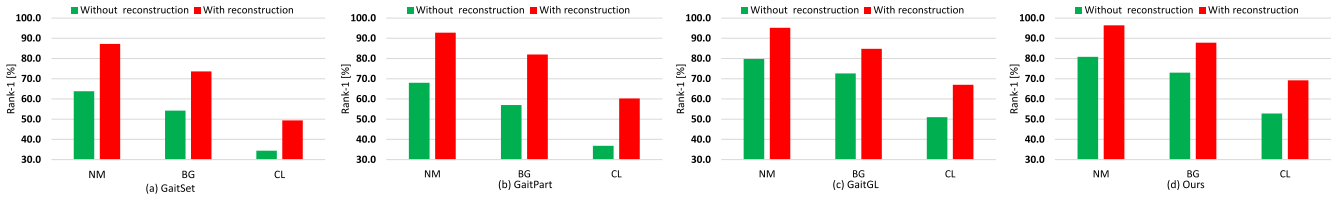

FIGURE 6. Averaged rank-1 accuracy (%) for occlusions of types L-0, L-1, L-2, L-3, and L-4 using various methods, with and without reconstruction of the silhouette sequence, on the CASIA-B dataset with subjects walking normally (NM), carrying bags (BG), and wearing coats (CL) sequences. Here, L-0, L-1, L-2, L-3, and L-4 represent different levels of occlusion in the silhouette sequences: L-0, lower 25% occluded; L-1, lower 50% occluded; L-2, left 50% occluded; L-3, right 50% occluded; L-4, no occlusion.

BG conditions compared with NM when the reconstructed silhouette sequence was used, particularly for lower and left-side occlusions; this was because carried objects could be observed in these types of occlusion, and our proposed ODR module could eliminate the effects of carried object and clothing variations.

In addition, our FEGR module achieved superior accuracy compared with various state-of-the-art methods: DM [45], sVideoWGAN-hinge [9], GaitSet [18], GaitPart [19], and GaitGL [11]. The FEGR module showed large improvements in accuracy compared with sVideoWGAN-hinge, GaitSet, and GaitPart, and its accuracy was comparable with that of GaitGL for lower-portion occlusion (i.e., L-0 and L-1 types). For example, our FEGR module surpassed the average rank-1 accuracy of GaitSet by 2%, 24%, and 38%, respectively, for the NM, BG, and CL conditions with silhouette sequence reconstruction, and it surpassed that of GaitGL by 1% for the BG and CL condition. Moreover, our FEGR module showed significantly improved performance for left- and right-side occlusion compared with other methods. For example, it surpassed the rank-1 accuracy of GaitSet by 17% to 29% and 3% to 10% for left- and right-side occlusion (i.e., L-2/L-3 type, respectively, and showed improvements of 3% to 7% and 1% to 10% compared with the baseline approach, GaitGL. Moreover, the GEI-based DM approach exhibited minimal accuracy, with an average rank-1 accuracy of 41.2% without reconstruction of the silhouette, and sVideoWGAN achieved 53.9%; our proposed method surpassed these by 43.3% and 30.6%, respectively. These results indicate that the ODR module of our proposed framework can reconstruct a silhouette sequence while preserving its discrimination ability, and the FEGR module achieves superior gait recognition accuracy.

2) EVALUATION USING OU-MVLP

The results for the OU-MVLP dataset are shown in Table 5, and Fig. 7 with and without reconstruction of the silhouette sequence. With reconstruction, the rank-1 accuracy improved from 5.6% to 30.8% for lower 25% occlusion (i.e., L-0 type) and from 21.1% to 28.4% for L-1-type occlusion. These results indicates that our proposed ODR module works better for higher occlusion levels. By contrast, for left- and right-side occlusion, the rank-1 accuracies without silhouette sequence reconstruction were 1.1% and 2.0%. However, after reconstruction of the silhouette sequence with our proposed ODR module, these accuracies showed improvements of 25.7%, 28.6%, and 34.1% for L-2-type occlusion and 23.4%, 32.9%, and 30.0% for L-3-type occlusion compared with the GaitSet, GaitPart, and GaitGL methods.

In addition, our FEGR module achieved the best accuracy among the methods tested, showing improvements of 7.0%, 5.8%, 8.1%, and 7.6% for L-0, L-1, L-2, and L-3 type occlusion, respectively, compared with GaitSet [18]. Its accuracy was comparable with that of baseline approach GaitGL [11], with improvements of 0.5%, 0.6%, 0.2%, and 0.2% for L-0, L-1, L-2, and L-3 type occlusions, respectively. These results indicate that our FEGR module extracts more profound features for gait recognition. More-

TABLE 5. Rank-1 accuracy (%) for silhouette sequences with (W/) and without (W/O) reconstruction on the OU-MVLP dataset.

Backbone	Status	Occlusion type					Average
		L-0	L-1	L-2	L-3	L-4	
DM [45]	W/O reconstruction	31.6	13.8	0.5	0.8	90.7	27.5
sVideoWGAN-hinge [9]	W/ reconstruction	62.4	34.9	20.2	20.3	90.7	45.7
GaitSet [18]	W/O reconstruction	63.9	11.0	1.6	1.2	93.5	34.2
	W/ reconstruction	78.8	39.4	27.3	24.5	93.5	52.7
GaitPart [19]	W/O reconstruction	78.6	27.3	1.1	1.5	95.9	40.9
	W/ reconstruction	84.3	51.1	29.7	34.4	95.9	59.1
GaitGL [11]	W/O reconstruction	79.7	22.1	1.1	1.9	96.0	40.2
	W/ reconstruction	85.3	44.6	35.2	31.9	96.0	58.6
Ours	W/O reconstruction	80.0	23.3	1.1	2.0	96.1	40.5
	W/ reconstruction	**85.8**	**45.2**	**35.4**	**32.1**	96.1	**58.9**

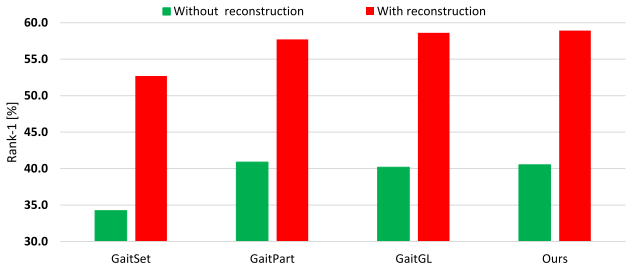

FIGURE 7. Averaged rank-1 accuracy (%) for occlusions of types L-0, L-1, L-2, L-3, and L-4 using various methods, with and without reconstruction of the silhouette sequence, on the OU-MVLP dataset. Here, L-0, L-1, L-2, L-3, and L-4 represent different levels of occlusion in the silhouette sequences: L-0, lower 25% occluded; L-1, lower 50% occluded; L-2, left 50% occluded; L-3, right 50% occluded; L-4, no occlusion.

over, our proposed approach significantly outperformed the GEI-based DM approach without reconstruction of the silhouette sequence and the sVideoWGAN-hinge approach with silhouette sequence reconstruction. Specifically, the average rank-1 accuracy improvements were 31.4% compared with DM and 13.2% compared with sVideoWGAN-hinge.

F. DISCUSSION

1) COMPARISON WITH AND WITHOUT OCCLUSION DETECTION

To evaluate the impact of occlusion detection on gait recognition accuracy, we conducted experiments on the CASIA-B dataset under different walking conditions: NM, BG, and CL, including all considered occlusion patterns. As shown in Table 6, significant improvements were achieved by incorporating occlusion detection. For instance, when using the GaitSet model [18], the rank-1 accuracy increased from 66.6% to 87.2% for the NM condition, from 57.4% to 73.6% for BG, and from 37.4% to 49.4% for CL. Similarly, the GaitPart model [19] showed improvements from 69.2% to 92.8% (NM), 59.8% to 82.0% (BG), and 40.8% to 60.2% (CL) with the addition of occlusion detection. The GaitGL model [11] also benefited significantly, with accuracy increasing from 82.4% to 95.4% (NM), 75.4% to 85.2% (BG), and 55.6% to 67.0% (CL). Notably, our proposed method outperformed all other models used in the comparison, achieving the highest rank-1 accuracies

TABLE 6. Average rank-1 accuracy (%) for all occlusion patterns on the CASIA-B dataset, with and without occlusion detection; – denotes the absence of a task, and ✓ indicates its presence.

Models	Occlusion detection	Occlusion reconstruction	NM	BG	CL
GaitSet [18]	–	✓	66.6	57.4	37.4
	✓	✓	87.2	73.6	49.4
GaitPart [19]	–	✓	69.2	59.8	40.8
	✓	✓	92.8	82.0	60.2
GaitGL [11]	–	✓	82.4	75.4	55.6
	✓	✓	95.4	85.2	67.0
Ours	–	✓	83.2	76.6	57.6
	✓	✓	96.4	87.8	69.2

of 96.4%, 87.8%, and 69.2% for NM, BG, and CL, respectively, when both occlusion detection and reconstruction were employed. These results clearly indicate that the incorporation of occlusion detection and reconstruction substantially enhances gait recognition performance across various occlusion scenarios. The consistent improvements across different models and occlusion types support the effectiveness of our unified framework in mitigating the adverse effects of occlusions on gait recognition accuracy.

2) LIMITATIONS

The proposed method, although an effective means of improving gait recognition accuracy under occluded conditions, has some limitations in practical applications. For instance, the method relies on artificially generated occlusions, which may not fully capture the complexity and variability of real-world occlusions. In practice, occlusions can vary significantly in terms of shape, size, and location, making it challenging for the proposed method to generalize to unseen occlusions. In addition, the reconstruction process, although effective for the tested datasets, might be computationally intensive and perform less well on larger datasets or in real-time applications. To improve the method, future work could focus on enhancing the robustness of the reconstruction module by training it on a more diverse set of occlusions, including those encountered in real-world scenarios. Moreover, optimizing the computational efficiency of the method, potentially by incorporating lightweight neural network architectures or advanced optimization techniques, could make it more suitable for real-time applications.

V. CONCLUSION

In this paper, we have introduced an end-to-end unified framework comprising ODR and FEGR modules to address the challenges of gait recognition under conditions of occlusion. The ODR module estimates the type of occlusion in a gait sequence, and, based on this detection, a novel video-based GAN reconstructs the occluded portions of the silhouette sequence. The FEGR module then extracts both global and local features from the entire silhouette sequence, as well as frame-by-frame features from the reconstructed sequence, to enhance gait recognition. To validate the effectiveness of our proposed framework, we conducted

experiments using artificially simulated occlusions on publicly available datasets (CASIA-B and OU-MVLP). Our approach demonstrated significant performance improvements, achieving average rank-1 accuracies of 96.4%, 87.8%, and 69.2% on CASIA-B under normal, carried object, and clothing variation conditions, respectively, along with 100.0% accuracy in occlusion type detection. On the OU-MVLP dataset, the framework achieved an average rank-1 accuracy of 58.9%, again with 100.0% occlusion type detection accuracy. These results demonstrate the ability of our framework to substantially enhance gait recognition accuracy even in the presence of occlusions. In future work, we aim to extend our evaluation by incorporating a wide variety of occlusions and collecting real-world occlusion samples. This will further validate the robustness and applicability of our proposed framework in more diverse scenarios.

ACKNOWLEDGMENT

(Kamrul Hasan and Md. Zasim Uddin are co-first authors.)

REFERENCES

[1] M. Z. Uddin, D. Muramatsu, T. Kimura, Y. Makihara, and Y. Yagi, "MultiQ: Single sensor-based multi-quality multi-modal large-scale biometric score database and its performance evaluation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, no. 1, pp. 1–25, Dec. 2017.

[2] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon, "On using gait in forensic biometrics," *J. Forensic Sci.*, vol. 56, no. 4, pp. 882–889, Jul. 2011.

[3] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi, "Gait verification system for criminal investigation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 5, no. 1, pp. 163–175, 2013.

[4] N. Lynnerup and P. K. Larsen, "Gait as evidence," *IET Biometrics*, vol. 3, no. 2, pp. 47–54, 2014.

[5] M. Mizuno, T. Fujita, Y. Kawanishi, D. Deguchi, and H. Murase, "Subjective baggage-weight estimation based on human walking behavior," *IEEE Access*, vol. 12, pp. 39390–39398, 2024.

[6] M. Altab Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," *Pattern Recognit.*, vol. 43, no. 6, pp. 2281–2291, Jun. 2010.

[7] D. Muramatsu, A. Shiraishi, Y. Makihara, Md. Z. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 140–154, Jan. 2015.

[8] Y. Makihara, A. Mansur, D. Muramatsu, Z. Uddin, and Y. Yagi, "Multi-view discriminant analysis with tensor representation and its application to cross-view gait recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–8.

[9] M. Z. Uddin, D. Muramatsu, N. Takemura, M. A. R. Ahad, and Y. Yagi, "Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion," *IPSJ Trans. Comput. Vis. Appl.*, vol. 11, no. 1, pp. 1–18, Dec. 2019.

[10] I. Rida, N. Almaadeed, and S. Almaadeed, "Robust gait recognition: A comprehensive survey," *IET Biometrics*, vol. 8, no. 1, pp. 14–28, Jan. 2019.

[11] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14628–14636.

[12] M. Z. Uddin, T. T. Ngo, Y. Makihara, N. Takemura, X. Li, D. Muramatsu, and Y. Yagi, "The OU-ISIR large population gait database with real-life carried object and its performance evaluation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–11, Dec. 2018.

[13] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes," *Pattern Recognit. Lett.*, vol. 30, no. 11, pp. 977–984, Aug. 2009.

[14] P. Nangtin, P. Kumhom, and K. Chamnongthai, "Gait identification with partial occlusion using six modules and consideration of occluded module exclusion," *J. Vis. Commun. Image Represent.*, vol. 36, pp. 107–121, Apr. 2016.

[15] J. Ortells, R. A. Mollineda, B. Mederos, and R. Martín-Félez, "Gait recognition from corrupted silhouettes: A robust statistical approach," *Mach. Vis. Appl.*, vol. 28, nos. 1–2, pp. 15–33, Feb. 2017.

[16] C. Xu, S. Tsuji, Y. Makihara, X. Li, and Y. Yagi, "Occluded gait recognition via silhouette registration guided by automated occlusion degree estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2023, pp. 3191–3201.

[17] A. Gupta and R. Chellappa, "You can run but not hide: Improving gait recognition with intrinsic occlusion type awareness," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5881–5890.

[18] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8126–8133.

[19] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14213–14221.

[20] A. Roy, S. Sural, J. Mukherjee, and G. Rigoll, "Occlusion detection and gait silhouette reconstruction from degraded scenes," *Signal, Image Video Process.*, vol. 5, no. 4, pp. 415–430, Nov. 2011.

[21] S. S. Kumar, B. Singh, P. Chattopadhyay, A. Halder, and L. Wang, "BGaitR-net: An effective neural model for occlusion reconstruction in gait sequences by exploiting the key pose information," *Expert Syst. Appl.*, vol. 246, Jul. 2024, Art. no. 123181.

[22] D. P. Kingma and M. Welling, "Stochastic gradient VB and the variational auto-encoder," in *Proc. 2nd Int. Conf. Learn. Represent.*, vol. 19, 2014, p. 121.

[23] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.

[24] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, Aug. 2006, pp. 441–444.

[25] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, p. 4, Feb. 2018.

[26] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, 2004.

[27] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 149–158, Feb. 2004.

[28] J.-H. Yoo, D. Hwang, K.-Y. Moon, and M. S. Nixon, "Automated human recognition by gait using neural network," in *Proc. 1st Workshops Image Process. Theory, Tools Appl.*, Nov. 2008, pp. 1–6.

[29] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107069.

[30] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gaitgraph: Graph convolutional network for skeleton-based gait recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2314–2318.

[31] T. Teepe, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Towards a deeper understanding of skeleton-based gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1568–1576.

[32] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[33] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention*, 2009, pp. 1–6.

[34] J. Liu and N. Zheng, "Gait history image: A novel temporal template for gait recognition," in *Proc. IEEE Multimedia Expo. Int. Conf.*, Jul. 2007, pp. 257–270.

[35] T. H. W. Lam, K. H. Cheung, and J. N. K. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recognit.*, vol. 44, no. 4, pp. 973–987, Apr. 2011.

[36] T. Chai, A. Li, S. Zhang, Z. Li, and Y. Wang, "Lagrange motion analysis and view embeddings for improved gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20217–20226.
[37] J. Chen, Z. Wang, C. Zheng, K. Zeng, Q. Zou, and L. Cui, "GaitAMR: Cross-view gait recognition via aggregated multi-feature representation," *Inf. Sci.*, vol. 636, Jul. 2023, Art. no. 118920.
[38] Z. Zhou, A. Prugel-Bennett, and R. I. Damper, "A Bayesian framework for extracting human gait using strong prior knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1738–1752, Nov. 2006.
[39] G. Zhao, L. Cui, and H. Li, "Gait recognition using fractal scale," *Pattern Anal. Appl.*, vol. 10, no. 3, pp. 235–246, Jul. 2007.
[40] M. Hofmann, D. Wolf, and G. Rigoll, "Identification and reconstruction of complete gait cycles for person identification in crowded scenes," in *Proc. Intern. Conf. Comput. Vis. Theory Appl. (VISAPP)*, 2011, pp. 1–22.
[41] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–11.
[42] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, 2015, pp. 234–241.
[43] A. Gupta and V. B. Semwal, "Occluded gait reconstruction in multi person gait environment using different numerical methods," *Multimedia Tools Appl.*, vol. 81, no. 16, pp. 23421–23448, Jul. 2022.
[44] M. Z. Uddin, M. A. Shahriar, M. N. Mahamood, F. Alnajjar, M. I. Pramanik, and M. A. R. Ahad, "Deep learning with image-based autism spectrum disorder analysis: A systematic review," *Eng. Appl. Artif. Intell.*, vol. 127, Jan. 2024, Art. no. 107185.
[45] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2019.

KAMRUL HASAN graduated in computer science and engineering from Begum Rokeya University, Rangpur, Bangladesh. He is currently an Artificial Intelligence Engineer at Next Solution Lab, Bangladesh. His research interests include computer vision and machine learning.

MD. ZASIM UDDIN (Member, IEEE) received the Ph.D. degree in gait recognition from the Institute of Scientific and Industrial Research, Osaka University, Japan. He is currently an Associate Professor with the Department of Computer Science and Engineering, Begum Rokeya University, Rangpur, Bangladesh. He has been published in IEEE TRANSACTIONS ON IMAGE PROCESSING, EAAI, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, IEEE FC, JIA, and IPSJ CVA, and has served CVPR, ECCV, ICCV, WACV, and SR. His research interests include gait recognition, computer vision, and machine learning.

AUSRUKONA RAY received the degree from the Department of Computer Science and Engineering, Begum Rokeya University, Rangpur, Bangladesh, in 2023. She is currently a Research Assistant with the Machine Vision Laboratory, Begum Rokeya University. Her research interests include computer vision and gait-based biometrics.

MAHMUDUL HASAN (Senior Member, IEEE) received the Ph.D. degree in computer vision and human robot interactions from Saitama University, Japan. He is currently an Associate Professor with the Computer Science and Engineering Department, Comilla University. His research interests include human–computer interaction, computer vision, signal processing, artificial intelligence, computers in mathematics, natural language processing, and data mining.

FADY ALNAJJAR (Member, IEEE) received the M.S. degree in artificial intelligence and the Ph.D. degree in system design engineering from the University of Fukui, Japan, in 2007 and 2010, respectively. He has been a Research Scientist with the Brain Science Institute, RIKEN, Japan. He conducted a neuro-robotics study to understand the underlying mechanisms for embodied cognition and the mind. In 2012, he started exploring the neural mechanisms of motor learning, adaptation, and recovery after brain injury from a sensory and muscle synergy perspective. His research target is to develop an advanced neuro-rehabilitation application for patients with brain injuries.

MD ATIQUR RAHMAN AHAD (Senior Member, IEEE) received the Ph.D. degree. He became a Professor with the University of Dhaka, in 2018, and a specially appointed Associate Professor with Osaka University. He is currently a Professor of artificial intelligence and machine learning (champion, research and innovation) with the Department of Computer Science and Digital Technologies, University of East London, U.K. He is also a Guest Professor with Kyutech, Japan, and a Visiting Professor with UCSI University, Malaysia. He works on pattern recognition, vision, and the IoT. He has authored 15 books and more than 200 journal/conference papers and chapters. He has been a keynote or invited speaker more than 150 times at different conferences/universities. He is a Senior Member of OPTICA. He has received the UGC Gold Medal (awarded by the Honorable President of Bangladesh) and more than 50 awards. He is also an Editorial Board Member of *Scientific Reports* and *Nature*, and serves PR, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE, IEEE TRANSACTIONS ON AUTOMATIC CONTROL, PRL, and ACM IMWUT.