Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.067743

ARTICLE

A Region-Aware Deep Learning Model for Dual-Subject Gait Recognition in Occluded Surveillance Scenarios

Zeeshan Ali[1], Jihoon Moon[2], Saira Gillani[3], Sitara Afzal[4], Maryam Bukhari[5] and Seungmin Rho[6,*]

[1]Research and Development Setups, National University of Computer and Emerging Sciences, Islamabad, 44000, Pakistan
[2]Department of Data Science, Duksung Women's University, Seoul, 01369, Republic of Korea
[3]Faculty of Information Technology and Computer Science, University of Central Punjab, Lahore, 54590, Pakistan
[4]Department of Software, Sejong University, Seoul, 05006, Republic of Korea
[5]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock, 43600, Pakistan
[6]Department of Industrial Security, Chung-Ang University, Seoul, 06974, Republic of Korea

*Corresponding Author: Seungmin Rho. Email: smrho@cau.ac.kr

Received: 11 May 2025; Accepted: 08 August 2025; Published: 31 August 2025

ABSTRACT: Surveillance systems can take various forms, but gait-based surveillance is emerging as a powerful approach due to its ability to identify individuals without requiring their cooperation. In the existing studies, several approaches have been suggested for gait recognition; nevertheless, the performance of existing systems is often degraded in real-world conditions due to covariate factors such as occlusions, clothing changes, walking speed, and varying camera viewpoints. Furthermore, most existing research focuses on single-person gait recognition; however, counting, tracking, detecting, and recognizing individuals in dual-subject settings with occlusions remains a challenging task. Therefore, this research proposed a variant of an automated gait model for occluded dual-subject walk scenarios. More precisely, in the proposed method, we have designed a deep learning (DL)-based dual-subject gait model (DSG) involving three modules. The first module handles silhouette segmentation, localization, and counting (SLC) using Mask-RCNN with MobileNetV2. The next stage uses a Convolutional block attention module (CBAM)-based Siamese network for frame-level tracking with a modified gallery setting. Following the last, gait recognition based on region-based deep learning is proposed for dual-subject gait recognition. The proposed method, tested on Shri Mata Vaishno Devi University (SMVDU)-Multi-Gait and Single-Gait datasets, shows strong performance with 94.00% segmentation, 58.36% tracking, and 63.04% gait recognition accuracy in dual-subject walk scenarios.

KEYWORDS: Dual-subject based gait recognition; covariate conditions; occlusion; deep learning; human segmentation and tracking; region-based CNN

1 Introduction

In the current days of digital realm, different security issues have brought substantial advancements to video surveillance systems to ensure safe and secure environments [1]. Video surveillance systems can be implemented in a variety of ways, in which biometric-based solutions are becoming more widespread.

More precisely, in both academic and industrial applications, biometric-based authentication systems are gaining wider acceptance since they validate individual identities using their unique attributes classified as physical and behavioral [2]. Therefore, biometrics can be defined as the assessment of biological attributes such as fingerprints of persons to determine identification [1]. Among all of these biometric traits, human gait is seen to be more practical because of its strengths [3]. More explicitly, gait recognition is the method

of identifying individuals based on their walking patterns. Other biometric systems, such as the face, iris, fingerprints, signature, etc., identify the individual based on a minimal distance margin, but gait can be acquired and recognized from a large distance, and also individual cooperation is not required throughout the authentication process [4,5]. Moreover, when other attributes, including faces and fingerprints, remain obscured, gait recognition still performs effectively [3].

In the existing studies, several methods have been proposed by different researchers for human gait recognition that were mainly classified as model-based and model-free approaches. In model-based techniques, distinct geometrical shapes are employed to design the model, and attributes of an individual's body shape are exploited to acquire gait patterns for identification [6,7]. To obtain such gait patterns, several variables, including speed, stride, and step size, are used. On the other hand, Model-free (appearance-based) methods extract statistical and spatiotemporal features from silhouettes segmented in surveillance videos [8]. These methods are more common among research studies than model-based methods. Besides that, DL models are widely used in gait-based identification systems. As input, these models use gait representations such as the gait energy image (GEI) [9], which reduces computational complexity.

DL-based gait recognition pipelines segment silhouettes from videos and compute gait features for model training [10]. In addition, pre-trained models such as AlexNet, VGG19, and DenseNet are proposed for gait recognition [11]. Although deep learning models show good performance, their accuracy drops significantly under covariate conditions like occlusion or view changes. According to research [12], many covariate factors affect the performance of gait recognition, such as carrying conditions, clothing conditions, walk speed, occlusions, etc. For instance, a long coat worn by an individual may obscure features critical for recognition [3]. Furthermore, variations in camera angle along with static and dynamic occlusions significantly affect the performance [1,13]. Clothing variations also hinder recognition by affecting the extracted feature set [14].

Following on, the primary research gaps are that most existing studies focus on single-person scenarios, where one individual's gait is captured and analyzed for identification [10,15]. However, a key challenge arises when surveillance footage contains multiple individuals walking together, i.e., in a dual-subject walk scenario. In dual-subject walk scenarios, automated gait recognition faces major challenges in accurately segmenting and tracking individuals. There exist some traditional algorithms for human segmentation [16,17], but their performance drops when environments are in uncontrolled, i.e., dynamic background conditions, static and dynamic occlusions, and lightning and shadow conditions [1]. Thirdly, occlusion conditions are one of the most difficult challenges in real applications. For example, a person's gait may be partially obscured by static objects or other people, and limited research has addressed these challenges. Existing approaches to address occlusion problems include the use of generative adversarial networks (GANs) to reconstruct missing parts of silhouette images [18]. However, when a system fails to recognize a person in an image due to occlusion, GANs fail to reconstruct silhouettes. Furthermore, in existing studies, the occlusion problem occurs directly before segmentation and tracking people across video frames; thus, the performance of gait recognition is biased if tracking of people is disrupted owing to occlusion. In addition, many existing methods treat occlusion and tracking as independent stages, which can lead to compounded errors in the complete pipeline of practical and reliable gait recognition.

To tackle these challenges, we propose a comprehensive gait-based surveillance framework called the Dual-Subject Gait Model (DSG). To the best of our knowledge, this is the first comprehensive attempt at gait recognition in a dual-subject walk context by processing raw surveillance videos through the entire pipeline, including silhouette segmentation, multi-person instance detection, tracking, and final gait recognition, rather than relying on pre-extracted or manually tracked silhouettes. To begin with, we propose a modified Mask-RCNN model with a MobileNetV2 backbone for instance-based silhouette segmentation, enabling

accurate segmentation, localization, and counting of human silhouettes within video frames. Secondly, we have designed a person-re-identification module for frame-level tracking of people with a proposed modified gallery setting. Thirdly, to handle the problems of occlusion in a dual-subject walk environment, we have proposed a region-based DL model. More precisely, in the proposed approach, a surveillance camera captures individuals or records video sequences as they walk in a dual-subject gait scenario. Unlike earlier research, our study takes a more extensive and exploratory approach, presenting a complete gait recognition technique designed specifically for dual-subject walk scenarios. Because each stage, i.e., segmentation, tracking, and recognition, has a direct influence on the performance of the next, we thoroughly examine and emphasize the limitations and problems at each level. This enables a more precise evaluation of the entire system's efficacy in both single-person and two-person (i.e., dual-subject walk scenario) video scenarios. Moreover, the proposed model is validated with an extensive set of experiments on SMVDU-Multi-Gait and SMVDU-Single-Gait datasets and exhibits good performance. Fig. 1 depicts several example images from the dataset, in which a dual-subject walk scenario is presented, where more than one person is under camera observation and an occlusion problem happens, such as some part of the individual being obscured owing to another person. The following are our point-by-point contributions:

- A novel end-to-end dual-subject gait recognition system is proposed for occluded surveillance environments.
- An SLC module is proposed for silhouette segmentation, localization, and counting in a two-person walking scenario using Mask-RCNN with MobileNetV2.
- A CBAM-based Siamese network with a proposed modified gallery setting for efficient person tracking across video frames.
- A region-based deep learning model for improved gait recognition under partial occlusions.

The rest of the paper is organized as Section 2 provides the literature review, Section 3 provides the proposed methodology, and Section 4 describes the results with analysis, followed by a conclusion and references.

Figure 1: Single gait vs Multi gait (Two persons): A Pictorial representation indicating occluded-dual subject walk scenarios leading to potential challenges in gait recognition

2 Literature Review

Studies have introduced various methods for gait recognition that use either model-based or model-free approaches. These methods include standard ML and improved DL techniques [19,20]. Furthermore, these

studies differ in the data they used, which range from gait data collected from ground sensors to image-based data [21]. All of these strategies seek to limit the influence of the covariate factor, which is a challenging problem in gait recognition.

More specifically, if we dive deeper into the literature, then, for instance, Kececi et al. [22] employ ML approaches including Random forests, Decision trees, multi-layer perceptron, and rule-based methods for human identification using gait. The gait data utilized in this study were acquired from the wearable accelerometer and gyroscope. Their proposed method shows good results in terms of an accuracy value of 99.00%. Similarly, Bari and Gavrilova [23] proposed artificial neural networks for gait recognition, in which gait data can be acquired using Kinect sensors. They proposed two novel features, i.e., joint relative cosine dissimilarity and joint relative triangle area for gait recognition, and achieved good results in terms of accuracy. These sensor-based approaches show good results; however, they are limited in terms of full automation as well as in terms of cost. To be more specific, with these approaches, a person must wear a wearable sensor or floor sensors are required to acquire an individual's gait [3].

On the other hand, by making surveillance fully automated, vision-based methods are more practical and easily deployable as no human intervention is required, and they are also less costly. For instance, human gait can be acquired from long-distance and low-resolution cameras [3]. In such a dimension of vision-based approaches to gait recognition, Liao et al. [24] designed a model-based approach to gait recognition in which pose features of persons are employed. In comparison with existing methods, their method employs 3D human poses from images using Convolutional neural networks (CNN). Teepe et al. [25] proposed graph convolution neural networks (GCN) for skeleton-based gait recognition in which residual models and higher inputs are merged for efficient gait recognition. An et al. [26] compute the 3D pose from the images, which is then employed as a feature for gait recognition. According to a research study, the 3D pose is less affected by covariate factors of viewing conditions in comparison with the 2D pose. All of these model-based approaches are good at addressing the problem of covariate conditions; however, they require high-resolution videos to accurately generate the geometric model of the human body and are also computationally expensive [1]. Moreover, they typically focus on single-person gait (one person per video) and address only individual steps, like recognition.

Model-free (appearance-based) methods use binary silhouettes, are less sensitive to texture or color, and are computationally efficient. However, their performance declines with view changes, clothing variations, or carrying conditions. To tackle these issues, researchers have proposed various cost-effective gait-based surveillance methods. For instance, Gul et al. [27] extract the spatial-temporal features from human gait images and design a 3D CNN for identification. Instead of using only silhouettes, gait energy images (GEI) are computed to acquire a more compact representation of gait that encodes both motion and visual shape features. Arshad et al. [28] proposed a deep convolutional neural network with feature selection to carry out human gait recognition. Alsaggaf et al. [29] proposed cycle-consistent generative adversarial network-based method for gait recognition to address the challenge of clothing and carrying conditions. Similarly, Alotaibi and Mahmood [10] design an improved CNN model that is more generalized and able to address the challenge of occlusions and other variations of gait recognition. Moreover, Li et al. [30] design both gait-based identification and verification approaches based on a joint intensity transformer network that is robust to challenging clothing and carrying conditions. The main challenge in these studies is their limitation to covariate conditions and reliance on preprocessed steps like silhouette segmentation and tracking.

Some methods also introduce specialized gait representations to enhance recognition performance. For instance, Yao et al. [31] proposed a skeleton gait energy image (SGEI) that is robust to unconstrained environments by designing a multi-stage deep neural network. They have also proposed a method in which

hybridization of both GEI and SGEI is carried out to lessen the weakness of model-free features. The experimental outcomes of their proposed method show good results in an environment of clothing variations. Similarly, Bashir et al. [32] proposed the gait entropy image (GeNI), which indicates the randomness of pixels in a binary silhouette image over the entire gait cycle. The motion information of an individual's walk style is encoded in GeNI, which is usually not affected by covariate conditions.

Other than covariate factors, including carrying, clothing, and camera conditions, some research studies also consider occlusion problems. For instance, Uddin et al. [18] employ GANs for the reconstruction of silhouette images. From the reconstructed videos, information about the complete gait cycle and features is retrieved. Their proposed approach shows good results over cutting-edge benchmark algorithms. Similarly, Roy et al. [33] proposed a method in which a video sequence is first divided into sub-videos reliant on key poses. This assists in determining whether the silhouette images are occluded or not. In their technique, the authors only evaluate their algorithm in terms of reconstruction performance, ignoring the fact that this reconstruction affects the performance accuracy of gait recognition systems. Hofmann et al. [34] perform the detection of occluded parts in videos with a very simple technique of foreground pixels. Occluded images were then swapped with non-occluded images in identical views from previous cycles. Muramatsu et al. [35] reconstruct the GEI from occluded GEI using the subspace method. In more recent literature of the past two years, Hasan et al. [36] address the problem of occlusion by incorporating occlusion detection and reconstruction (ODR) and feature extraction for gait recognition (FEGR) modules. A novel action detection perspective is also proposed by Huang et al. [37] for the prediction of occluded body regions to solve the challenge of occlusion. Gupta and Chellappa [38] proposed the Mimic Gait model to address occlusion problems using Correlational Knowledge Distillation. Furthermore, to handle the occlusion problem, Xu et al. [39] proposed a method that directly works on gait videos. In particular, they fit skinned multi-person linear (SMPL) based human mesh models to the input images solely, with no pre-normalization or authorization of the human body, providing a gait sequencing that only includes non-occluded body components in the images. Similarly, Gupta and Chellappa [40] address the problem of occlusion by learning the occlusion types to extract more refined features from videos.

It is evident from existing studies that most methodologies focus solely on single-user gait recognition. Although some recent approaches employ advanced techniques such as GANs, they have not adequately addressed occlusion challenges. Therefore, it is important to investigate the performance of gait-based identification in surveillance systems when individuals appear in dual-subject walking scenarios. For this, a system should first count and find the individuals present in a surveillance video, followed by their identification. This study proposes a deep learning-based automated surveillance system for dual-subject walking scenarios, aiming for practical deployment. It also addresses the dynamic occlusion problem that occurs when two individuals walk together.

3 Methodology

In this section, we discussed the proposed work in depth, including how data is obtained, annotated, and used to design this framework. Furthermore, each stage of the proposed Dual-subject Gait Model (DSG) is described, such as how videos are processed to segment, track, and count people, how occlusion is addressed, and how individuals are identified based on their gait patterns. Fig. 2 shows a complete overview of the proposed work.

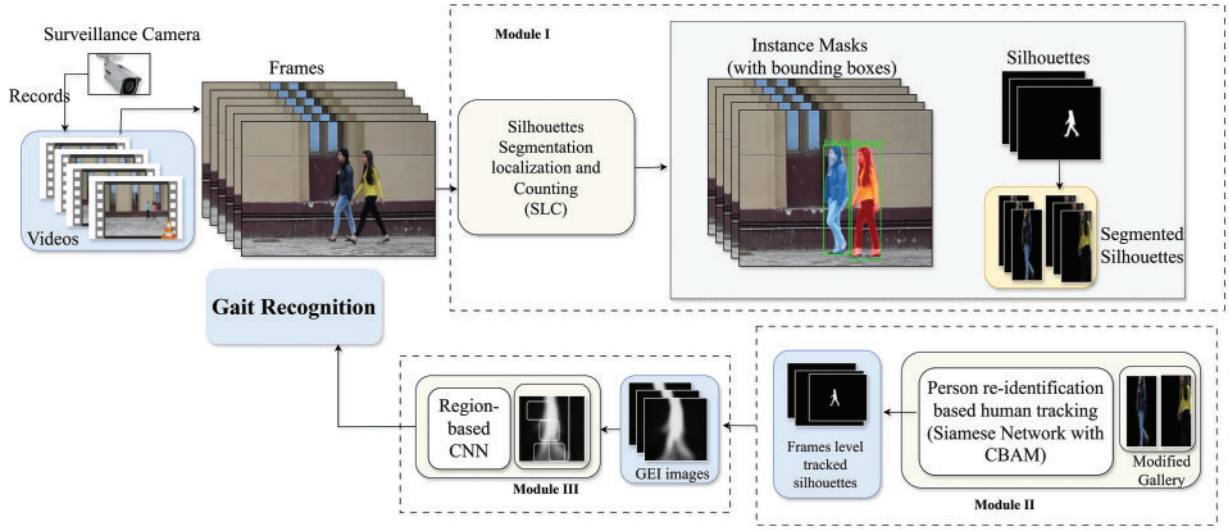

Figure 2: An illustration of the proposed Dual-subject Gait Model's (DSG) segmentation, tracking, and gait identification stages in occluded dual-subject walking scenarios

3.1 Datasets

To explore gait recognition in dual-subject-walk environments, we utilize the SMVDU-Multi-Gait and SMVDU-Single-Gait datasets, which contain videos of individuals walking in both dual-subject and single-person scenarios [41]. More specifically, in the "SMVDU-Single-Gait" dataset, videos of individuals are available in a single-walk environment, i.e., there is only one person in a single video, but "SMVDU-Multi-Gait" videos show more than one person walking in a dual-subject scenario, i.e., a maximum of two persons. When individuals walk in a dual-subject scenario, occlusion can occur, meaning that part of a person's silhouette may be obscured by the presence of another person walking alongside them. Furthermore, in both parts of the datasets, there is data from a total of 20 persons, and for each person, there are 6 sequences. Each sequence is recorded in a separate view, including left to right, right to left, lateral view from right to left, oblique view from right to left, front views, and back views. Moreover, some individuals have repeated sequences. The SMVDU-Multi-Gait dataset contains a total of 95 sequences, while the SMVDU-Single-Gait dataset includes 97 sequences. Table 1 shows a summarized description of the datasets.

Table 1: Description of Single and Multi-Gait dataset for occlusion-based gait recognition

S. No.	Dataset	Total videos	Per person videos	Views
01	SMVDU-Single-Gait	97	6	Frontal, Lateral, oblique
02	SMVDU-Multi-Gait	95	6	Frontal, Lateral, oblique

3.2 Proposed Dual-Subject Gait Model

After collecting the videos in a dual-subject walk environment of different persons, we designed a dual-subject gait model. More explicitly, the DSG is divided into three modules. The first module is known as "SLC", which is designed for human silhouette segmentation (instance-wise), localization, and counting. Following on the second module deals with person tracking with person re-identification using a modified gallery setting. The last module proposes a region-based deep-CNN module for identifying individuals based on gait, resulting in dual-subject-based testing/recognition of persons who are unaware that the system

is monitoring them (i.e., in uncontrolled contexts). The primary objective of this research is to address challenging scenarios in gait recognition while maintaining high identification accuracy for surveillance applications. These evolving gait-based systems are more realistic, better suited for real-time environments, and easier to deploy. The parameters and procedural steps of each module are detailed below.

Proposed SLC Module

In the first stage, we designed the "SLC" module for the detection of persons, i.e., how many people are in a video, and where are the individuals in the image in terms of localized bounding boxes and segmentation masks. Specifically, the SLC module performs initial tasks such as person detection, silhouette segmentation, and tracking. The output segmented silhouettes are then passed to the person re-identification module, which tracks individual identities across frames. These tracked identities are finally used by the gait recognition module for identification. In the existing studies, some traditional algorithms for person segmentation [16,17] have been proposed, but these algorithms fail to handle the adaptive background conditions. Moreover, they apply to only single-user environments, when there is one person in the image; however, what if there are many people in the video? While performing instance segmentation by classifying each pixel to its corresponding person, the system also detects the total number of individuals present in the video. Traditional semantic segmentation algorithms fall short in this task as they do not differentiate between individual person instances. To overcome this, we propose an instance segmentation model based on Mask-RCNN, capable of both segmenting and counting individuals using instance masks, supported by a transfer learning approach. The reason for utilizing Mask-RCNN is that it is a dual-task model in which person localization is accomplished using both bounding boxes and instance masks in a simple, generic, and flexible way. Mask-RCNN [42] is the extended version of the Faster-RCNN in which an additional branch is added to predict the instance segmentation mask. It has a two-stage procedure in which the first stage is based on RPN (Regional proposal network) for extracting regions of interest (ROIs), while in the second stage, ROI masks are computed for each person in the video. During training, for each sampled ROI from frames of video, a loss called "multi-task" loss is computed as defined in Eq. (1):

$$L = L_{cls} + L_{box} + L_{mask} \tag{1}$$

The classification loss L_{cls} and bounding box loss L_{box} are defined in earlier versions. However, in the Mask-RCNN, for every ROI, there is Km^2—dimensional output from the mask-branch. This branch encapsulates K binary masks with a resolution of $m \times m$, i.e., one mask for every K class. To accomplish this, a sigmoid is applied per pixel and by using mean-binary-cross entropy loss denoted by L_{mask}. In the case of an ROI, L_{mask} is computed only for k^{th} ground-mask category corresponding to the predicted class label. This formulation offers a significant advantage over traditional semantic segmentation approaches like Fully Convolutional Networks (FCNs), resulting in more accurate instance-level segmentation [43].

Mask Representations

In the Mask-RCNN, a mask of dimension $m \times m$ is predicted from every ROI with the help of FCN [43]. The specific $m \times m$ object spatial arrangement is maintained with the help of each layer in the mask branch. Fewer parameters are required for this fully convolutional formulation, and research in [42] has shown that this is a more accurate representation. To properly retain the precise per-pixel spatial relationship, a perfect alignment of the features from the ROI is necessary, which are tiny activation maps.

ROI Align Layer

ROI Pool is a typical technique that extracts a tiny feature map (for instance, 7×7) from every ROI [43]. This layer performs the quantization that leads to misalignments among the ROIs and the retrieved features.

This will cause an impact on predicted pixel-wise masks. To solve this, Mask-RCNN employs the ROI align layer that eliminates the harsh quantization and accurately aligns the features along with the input. This layer ignores the quantization of ROI borders or bins. i.e., simply using $x/16$ rather than $[x/16]$. To determine the precise values of input features at four observed places of every ROI bin, a bilinear interpolation [43] is employed. Following on, the outcomes are combined using max or average operations. It is observed from [42] that this layer brings larger improvements in the accuracy of the model.

Mask-RCNN Architecture

The architecture of Mask-RCNN is divided into two main components: the backbone network, which is responsible for feature extraction, and the network head, which performs object detection tasks such as bounding box regression and classification. In addition to this, a mask branch is added that is operated over every ROI. In existing studies, several backbone architectures are used, such as ResNet [44] variants and ResNeXt [45]. The Feature Pyramid network is also used as a backbone architecture in existing research [46]. Similarly, for the network head branch, Faster-RCNN box heads are extended from the ResNet and FPN models.

Transfer Learning Using Mask-RCNN

In this study, we have performed transfer learning using a mask-RNN architecture on our custom dataset for person detection, segmentation, and counting. For this, we need to prepare the ground truth annotations of the video frames. To accomplish this, we employed the "Label me" tool to generate the instance segmentation masks using polygon features. For instance, in a surveillance video, the frames of the video are regarded as a stack of images where each image has a corresponding target, which includes coordinates of bounding boxes, labels for every bounding class, and the segmentation masks for each of the objects in the image. The background of the image is considered as the background class, denoted as integer 0. Following on, for transfer learning and fine-tuning, we employed the pre-trained Faster-RCNN model trained on the MS-COCO dataset by modifying its backbone architecture with MobileNetV2 [47] instead of the original backbone architecture. The reason for using MobileNetV2 as a backbone architecture is that it employs inverted residual and linear bottlenecks along with depth-wise convolutions and is less complex in terms of trainable parameters, more applicable and appropriate for real-time scenarios. Later on, we set the required number of output channels according to MobileNetV2. This model is fine-tuned for some iterations, and later on, we add the instance segmentation head using Mask-RCNN with backbone ResNet50, followed by mask-classifier and Mask-RCNN predictor. The mask-RCNN predictor is replaced with a new head in which custom classes are set, i.e., a background class and a person class. During the training of this model, data augmentation of type "Flip" is also applied over images to increase the generalization power of the model as well as reduce overfitting. The resulting trained model is able to perform silhouette segmentation in the form of instance segmentation. For example, when a surveillance video is input into the model, a sequence of frames is first extracted and processed individually. For each frame, the trained model detects and segments all visible persons, generating corresponding silhouette images. The number of instance masks or bounding boxes indicates how many individuals were identified in each frame.

3.3 Person-ReID for Tracking

At this stage, a person re-identification module is designed to track individuals throughout the walking sequence until the video stream ends (i.e., when the gait cycle is completed). The person-re-identification (ReID) module is based on a Siamese neural network with CBAM attention layers to perfectly target more information features in the images. More precisely, the training of the Siamese neural network is done on positive and negative pairs of images, i.e., consider a set of images of some person having ID p_1, e.g., $\{X\} \in p_1$ and another set of images belonging to another person having ID p_2, e.g., $\{Y\} \in p_2$. We generate the positive

and negative pairs of images such as $\{X_1\} \in p_1 = \{X_2\} \in p_1$ is positive pair while $\{X_1\} \in p_1 = \{Y_2\} \in p_2$ is a negative pair. This process is repeated over a complete dataset having n persons. A Siamese neural network is a dual-branch architecture that takes two input images and determines whether they belong to the same individual. The architecture of the Siamese neural network comprises convolutional layers and pooling layers along with CBAM-based features refinement layers. More precisely, a set of images is first passed through the convolution layer. For instance, given the collection of a pair of images, $\{f_i\} \in I$ a set of convolution filters has been applied over the images to generate the activation maps denoted as $\{H_j\} \in J$. To preserve the associations among the inputs and output, a connection Table CT is built to record the input i, kernels k, and output j. The mathematical formulation of this layer is given below:

$$h_j(x) = \sum_{i,k \in CTi,j,k} (f_i * w_k)(x) \tag{2}$$

In the above Eq. (2), the symbol "$*$" denotes the operation of convolution while w_k is the weight matrices and h_j is the outcome of layer, i.e., activation maps. Following on, pooling layers are added with a pool size of 2×2. Afterward, the CBAM [48] attention layers are added for further refinement of features. This CBAM block suppresses irrelevant features across two separate image pairings, allowing greater attention to be spent on discriminative features. A 2D channel attention is inferred and denoted as $M_c \in R^{C \times 1 \times 1}$ as well as spatial attention denoted as $M_s \in R^{1 \times H \times w}$ from the activation maps resulting from Conv-Pool layers represented as $F \in R^{C \times H \times w}$. The mathematical representation of the CBAM layer is given below:

$$F' = M_c(F) \otimes F \tag{3}$$

$$F'' = M_s(F') \otimes F' \tag{4}$$

where F'' is the final refined feature. Moreover, the computation of $M_c(F)$ and $M_s(F')$ is done using Eqs. (5) and (6) given below:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))) \tag{5}$$

$$M_s(F) = \sigma f^{7\times7}([AvgPool(F); MaxPool(F)])) = \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right) \tag{6}$$

In the above equations, the symbol σ denotes sigmoid activation, where F^s_{avg} and $F^s_{\max}$ are the average and max-pool layers.

Modified Gallery Setting

After training the CBAM-based Siamese neural network, the model can effectively determine whether two input images represent the same individual. Hence, after detection results from Mask-RCNN, this trained model is utilized to track the persons across video frames. Since, in a traditional one-shot gallery setting, the given query images are ranked against the whole database, however, for accurate tracking, we rank query images against the gallery comprising the images of the person observed in the first frame of the video. Following on, these gallery images served as a database, and the rest of the frames are ranked according to this modified gallery. Fig. 3 shows the pictorial representation of person tracking using the person re-identification module. Mathematically, consider a stack of video frames of length n arrive at different time steps t, then the persons registered with different IDs according to frame received at time step 1. Later on, on the subsequent frames, the persons detected by Mask-RCNN in frame n are ranked/matched against the person images registered with different IDs at time step 1. Eq. (7) shows the mathematical formulation of this modified gallery setting.

$$(len(Q_n) - Q_1\ ranked_{against}\ (G_I)\ \epsilon\ Q_1 \tag{7}$$

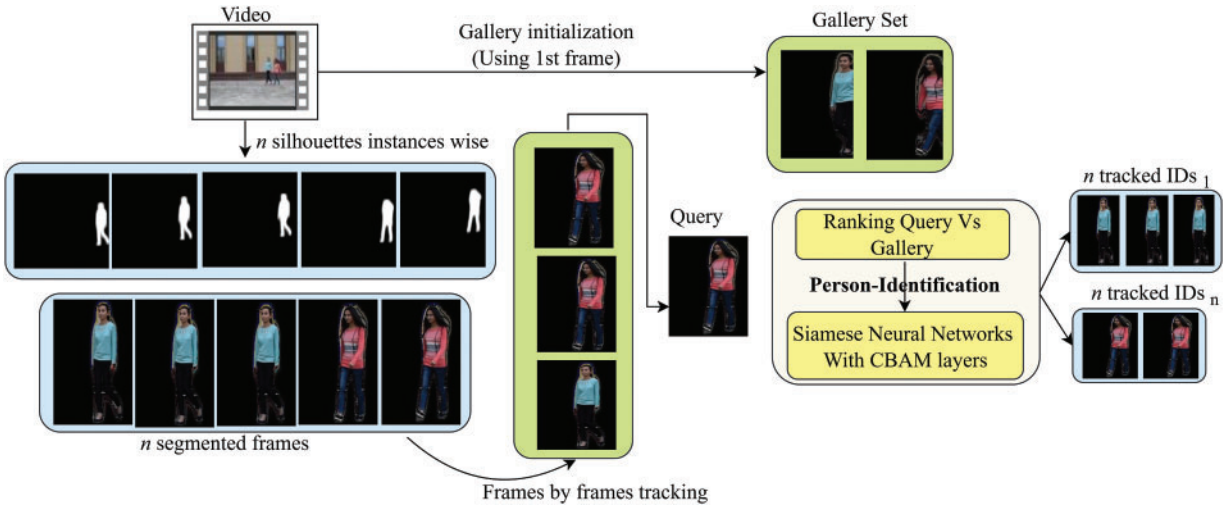

Figure 3: Frame-by-frame tracking of individuals using a modified gallery setting within the person re-identification module for improved accuracy in occluded scenarios

In the above Eq. (7), Q_1 is the first frame (i.e., query frame) registered as gallery images G_I used for tracking of person found in the subsequent frames Q_n. This improved gallery setting enhances tracking consistency by connecting identity comparisons to a stable reference set starting with the first frame of the video. Unlike classic one-shot scenarios, in which the gallery set has a full database, our method minimizes identity erroneous re-identification by using the first frame as a gallery set for person tracking. It is especially useful in obscured or low-resolution environments, allowing the system to track persons more accurately over time. This method also reduces the likelihood of a mismatch when there is very little variation in individuals' silhouettes and they appear in subsequent frames. Moreover, in this modified gallery setting, the gallery is initialized using the person images extracted from the first frame of the video and remains unchanged throughout the tracking process.

3.4 Gait Recognition

In the final stage, a region-based gait recognition module is proposed to identify individuals based on their gait patterns, as illustrated in Fig. 4. After acquiring the silhouette images with frame-level tracking, the next step is to extract the gait representation from the silhouette images, followed by person classification using gait. More explicitly, the tracked silhouette sequence of persons is processed to acquire a gait representation known as GEI.

3.4.1 Gait Energy Images (GEI)

Acquiring gait patterns directly from silhouette images and using them as a feature set is computationally as well as resource-intensive. To overcome these issues, a more realistic gait representation known as gait energy images was generated. The silhouette images generated by the trained Mask-RCNN are first tracked and aligned, and then their mean is computed to generate the gait energy image (GEI). The following is the mathematical equation for calculating GEIs:

$$GEI_{image} = G(x, y) = \frac{1}{T}\sum_{t=1}^{T} I(x, y, t) \tag{8}$$

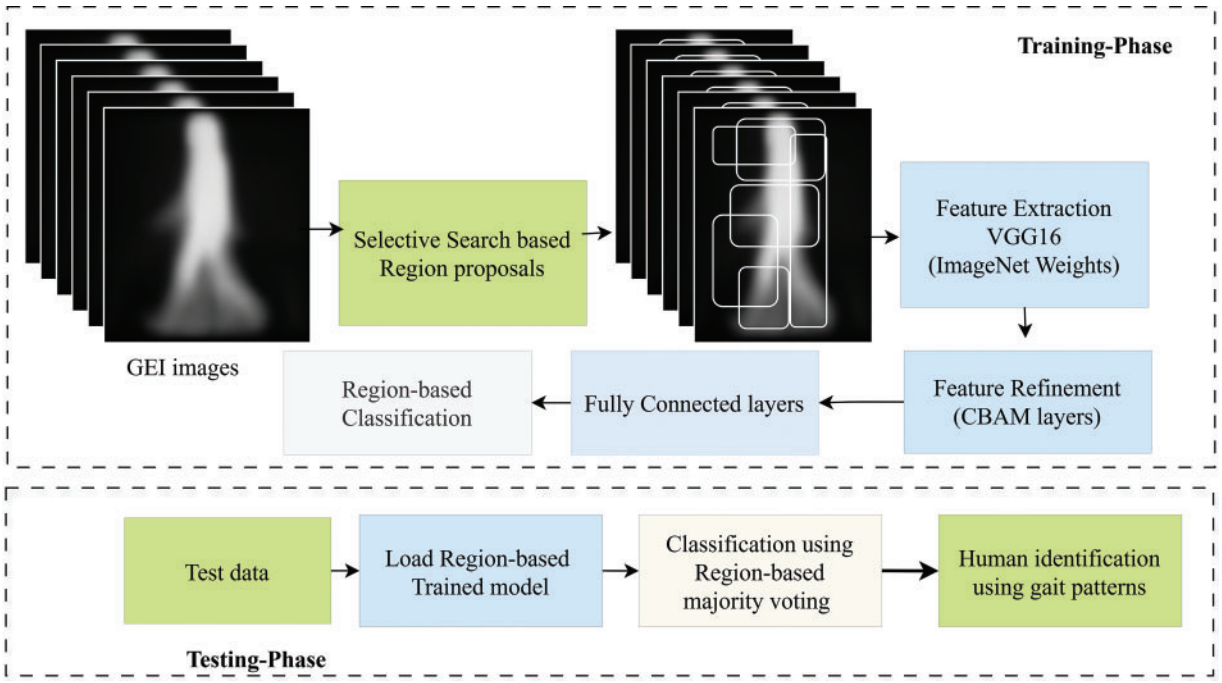

Figure 4: Region-based identification of persons based on their gait patterns in occluded dual-subject walk environments

In the above Eq. (8), $I(x, y, t)$ is the stack of frames (segmented by Mask-RCNN, i.e., silhouettes) at a different time starting from 1 to T where T is the total number of silhouette images of a particular subject in the database. The outcome of this equation is the GEI image denoted as $G(x, y)$ which holds more relevant gait features and is less affected by noise. Moreover, the motion-related information of a particular subject is represented as low gray-scale values in the GEI image, i.e., bottom regions. On the other hand, the top regions of the GEI image where the pixel intensities are high are referred to as static regions that involve the structural information of a subject.

3.4.2 Region-Based Deep Learning

The computed GEI images capture raw gait features; however, some of these features may be lost due to occlusion in dual-subject walk scenarios. Hence, recognizing persons from the complete GEI images will result in poor performance of the model. As a result, we presented a region-based CNN in which classification on different regions of GEI is done, i.e., which regions belong to which individual, and then the final class label of the person is determined by majority voting.

The rationale behind this approach is that during occlusion, some parts or regions of silhouettes are missed, and the resulting GEI is also disrupted because of the missed parts of persons in silhouettes. To address occlusion, the model is designed to recognize individuals based on the visible regions of GEI images. A selective search is applied to extract region proposals from GEIs, and the model is trained on these regions, allowing it to learn discriminative features from partial gait patterns. The architecture is based on VGG16, initialized with ImageNet weights. Each region is passed through the VGG16 for fea-ture extraction, followed

by fully connected layers with 512 ReLu-activated units for classification. Final recognition is performed using majority voting across all regions.

4 Experiments and Discussions

This section summarizes the proposed model's outcomes and discusses its strengths and drawbacks. Furthermore, the evaluation measures used to analyze the model's performance are also presented and described.

4.1 Evaluation Criteria

The evaluation criteria used to evaluate the performance of the model include different metrics. For instance, to evaluate the performance of person segmentation and localization, i.e., fine-tuned Mask-RCNN, we employed MS-COCO and PascalVoc-based metrics, comprised of average precision and recall at different IoU thresholds for both instance masks and bounding box results [42,49,50]. Likewise, to evaluate the performance of the person-reidentification module that is utilized for tracking, we have employed CMC scores, i.e., Cumulative matching scores with single-gallery shot settings [51]. Lastly, to evaluate region-based CNN, accuracy, precision, recall, and F1 Score are employed.

4.2 Experiments and Results

To evaluate the performance of the proposed model, we have employed the data described in Section 3.1. Each part of this dataset covers a different scenario, such as single gait and multi-gait (i.e., dual subject gait scenario). Delving into depth, the images of the silhouettes are the segmented results of Mask-RCNN with MobileNetV2 as a backbone feature extractor. More precisely, to train Mask-RCNN, we have first prepared the ground truth annotation using the "Label me" tool. This tool uses the frames from videos as an image, and we annotate the individuals using the polygon feature. If there are two people, we assign separate instance masks to each of them (i.e., color).

Fig. 5 illustrates the annotation results using the LabelMe tool. A subset of frames from the single-gait videos was used to train the Mask-RCNN model, while the remaining frames from the multi-gait videos were reserved for testing. To enhance the generalization capability of Mask-RCNN during training, additional images from the PennFudan dataset, specifically designed for person segmentation, were also incorporated. The trained model segments individuals in the form of mask images, assigning a unique colored mask to each person to indicate the presence of the "Person" classes. More explicitly, when provided with any video, this model will run and process each frame, detecting the person in the image with bounding boxes as well as mask images, which will subsequently be utilized as silhouette images. During training, different loss values, i.e., given in Eq. (1), are stored and shown in Fig. 6. The x-axis shows epochs or steps per epoch, while the y-axis shows scores and loss. More precisely, Fig. 6a illustrates the convergence of the classification L_{cls}, bounding box loss L_{box}, segmentation mask loss L_{mask}, and total loss $L = L_{cls} + L_{box} + L_{mask}$.

The object and RPN losses are also presented for each iteration of every epoch as shown in Fig. 6b. Similarly, Fig. 6c illustrates the average precision (AP) over IoU thresholds of 0.5–0.95. AP is the primary challenge metric for the MS COCO evaluations and is the average of 10 IoU thresholds with a difference of 0.5. The results were validated with a strict evaluation metric (i.e., an IoU threshold of 0.75). Notably, the bounding box–based results were more accurate than the pixel-wise segmentation, as indicated in Fig. 6c. Following on, Fig. 6d,e depicts the average recall over IoU thresholds of 0.5–0.95 for various maximum detections (i.e., 1 and 10). These graphs demonstrate the good convergence and results of the Mask R-CNN with MobileNetV2 as the backbone network.

Figure 5: Labelling of images in the dataset using the "Label Me" annotator tool with polygon feature

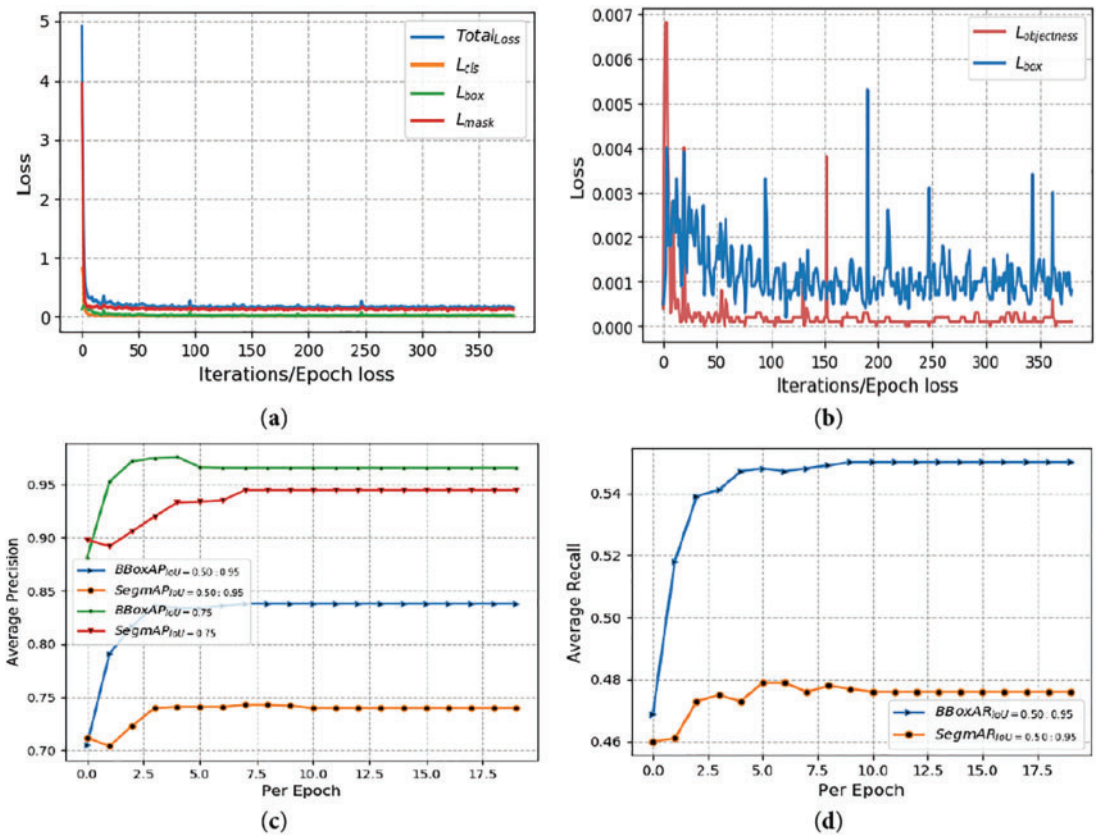

Figure 6: (Continued)

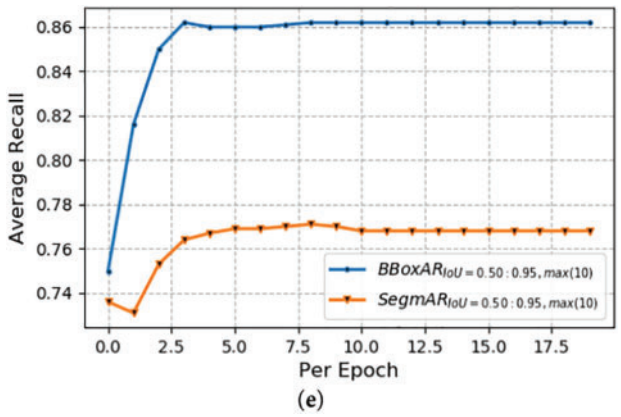

Figure 6: Results of Mask-RCNN on the custom dataset of person segmentation and counting. (**a**) Shows the loss values given in Eq. (1), i.e., classification loss, bounding box loss, mask loss, and total loss. (**b**) Loss values of RPN network (Region Proposal network) and objectness loss. (**c**) Average Precision over epochs at different IoU Thresholds. (**d**) Average Recall over epochs at different IoU Thresholds. (**e**) Average Recall over epochs at different IoU Thresholds with max-detections 10

Furthermore, the final results after all epochs of Mask-RCNN are also given in Table 2. More precisely, Table 2 first shows the values of final average precision at different thresholds for bounding detection of persons and pixel-by-pixel segmentation results, while the second part shows average recall values. Table 2 shows that Mask-RCNN produces outstanding results for person detection in surveillance videos, even with a strict metric, AP@[IOU: 0.75]. So, because the bounding box AP in this context is 0.997, and the results of pixel-wise segmentation are also 0.997. Subsequently, these detection results are utilized to acquire the silhouette representation, followed by GEI construction to perform gait recognition. The detection and segmentation results in the form of images from a video's frames to silhouette segmentation are also depicted in Fig. 7.

Table 2: IOU Scores of Mask-RCNN on bounding box localization and Instance segmentation task with Multi-Gait-SMVDU dataset

Evaluation metrics	Bounding box results	Segmentation mask	Evaluation metrics	Bounding box results	Segmentation mask
AP@[IOU:0.5:0.95]	0.838	0.740	AR@[IOU:0.5:0.95]	0.550	0.476
AP@[IOU:0.5]	0.997	0.997	AR@[IOU:0.5:0.95]	0.862	0.768
PascalVoc Metric			$area_{10}$		
AP@[IOU:0.75]	0.966	0.945	AR@[IOU:0.5:0.95]	0.800	0.700
Strict Metric			$area_{medium}$		
AP@[IOU:0.5:0.95]	0.761	0.679	AR@[IOU:0.5:0.95]	0.864	0.777
$area_{medium}$			$area_{large}$		
AP@[IOU:0.5:0.95]	0.840	0.743	AR@[IOU:0.5:0.95]	0.862	0.768
$area_{large}$			$area_{100}$		

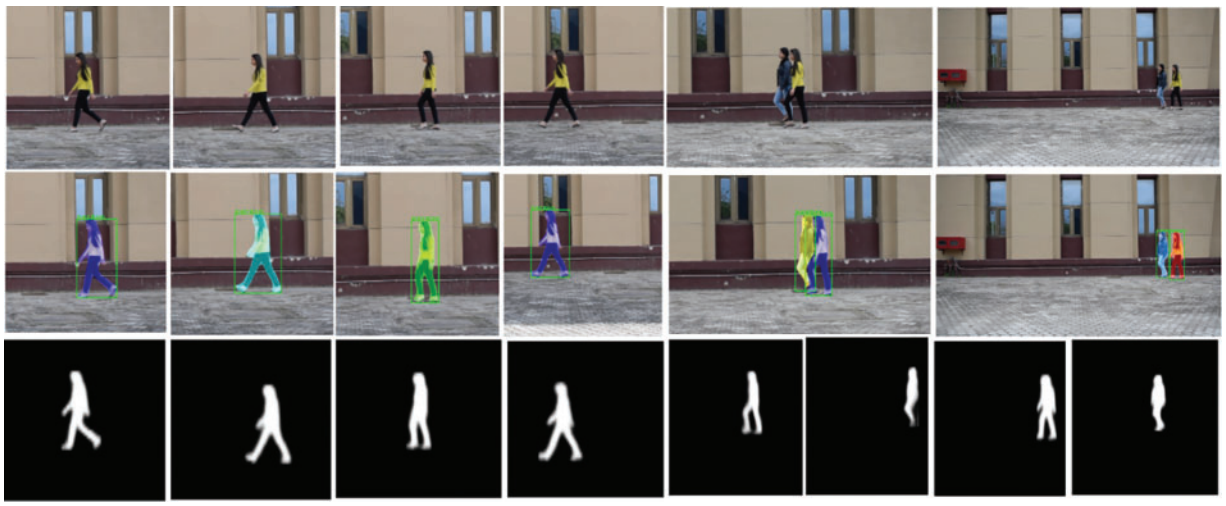

Figure 7: Results of the fine-tuned Mask R-CNN model for instance-level person segmentation and counting

Following that, for tracking, we designed the person-re-identification module with a modified gallery setting during ranking at test time. We trained the model using positive and negative image pairings on segmented images to determine whether or not the persons in the images are the same. Fig. 8 shows several examples of segmented images. After model learning, it can detect whether or not the persons in two images are the same, an approach known as binary classification. To track persons in the video frames, we registered persons detected in the very first frame by Mask-RCNN as a gallery set and ranked the remaining persons from the subsequent frames. Specifically, the remaining frames are processed individually as query images. Table 3 presents the rank scores (Rank-1 to Rank-5) achieved by the person re-identification model. As shown in Table 3, the proposed model demonstrates strong performance in ranking individuals across the gallery set.

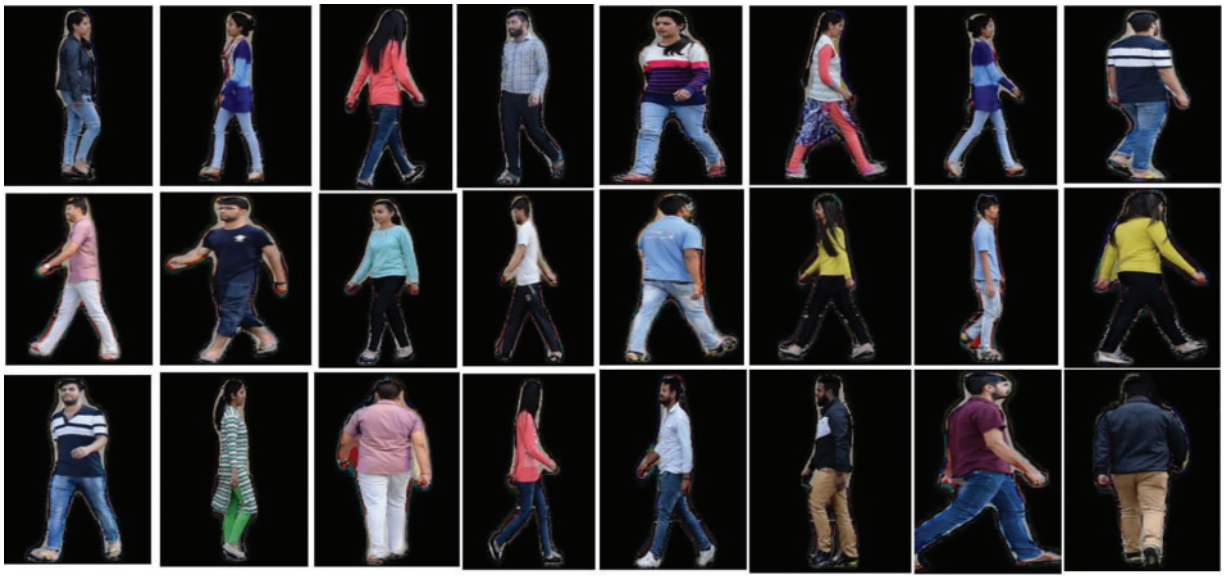

Figure 8: Results of Segmented images of different subjects using fine-tuned Mask-RCNN model for person segmentation and counting

Table 3: Performance of person re-identification utilized for tracking of persons

S. No.	Rank metric	Rank score
01	Rank@1	58.36%
02	Rank@2	71.12%
03	Rank@3	79.70%
04	Rank@4	80.01%
05	Rank@5	93.20%

Subsequently, it is observed that person tracking is a very challenging problem, especially in the presence of occlusion conditions. Several methods have been proposed in existing studies; however, when occlusion exceeds too much, then performance is also affected. This performance is also dependent upon object detection, as illustrated in Fig. 9a, it is observed that due to occlusion the other person is mostly hidden, and hence object detection fails to recognize this person as a separate person (i.e., the 2nd person is considered as part of 1st person), hence, the tracking also suffers and similarly gait features are difficult to acquire and persons cannot be properly identified. Similarly, Fig. 9b indicates that when the appearances (e.g., same color dress) of two persons are too similar, the IDs are overlapped during tracking.

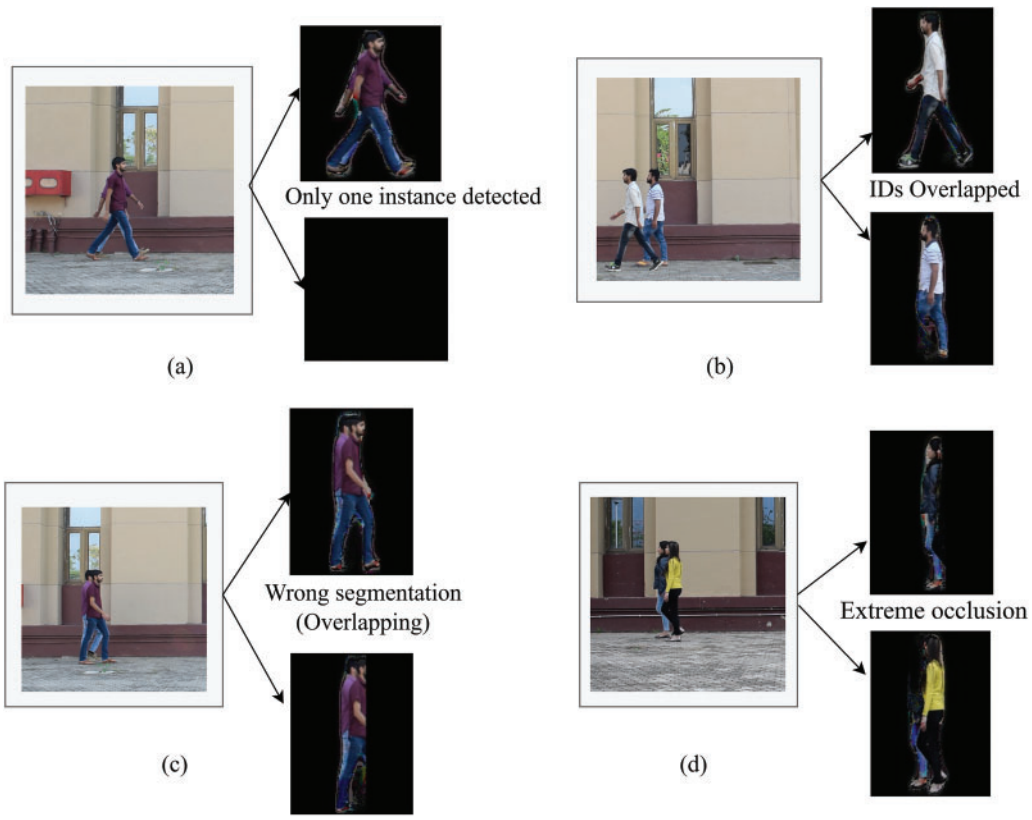

Figure 9: Examples of failure and challenging cases in person detection, segmentation, and tracking under occlusion conditions. (**a**) When only one instance is detected, (**b**) When subject IDs overlap, (**c**) Overlapped segmentation, (**d**) Extreme occlusion cases

Likewise, Fig. 9c shows an example of wrong segmentation, i.e., images of both persons overlapped during pixel-level classification. Similarly, Fig. 9d shows that most of the body part of one person is occluded due to another person walking behind him and such challenges result in a decrease in performance.

In the last, the performance of the region-based gait recognition model is assessed. For this, silhouettes are preprocessed to compute the gait representation using GEI images. The total number of images in both parts, i.e., SMVDU-Multi-Gait and SMVDU-Single-Gait, and the total number of sequences per subject is less, hence resulting in very few data of GEI images. To overcome this issue, we used data augmentation of horizontal flips to artificially increase the data. Following on, we have designed different experimental scenarios to assess the model's performance.

For example, in Table 4, the model is trained on multi-gait GEI images from one set of videos and tested on a separate set of videos, while maintaining the "multi-gait" (i.e., dual subject gait) condition throughout both training and testing. According to Table 4, the suggested model has a good accuracy of 63.04% even in stringent and challenging scenarios where human body parts are obscured owing to another person walking behind them. Furthermore, we compared the suggested model to a traditional CNN model (i.e., without region-based learning), and it was found that the region-based model outperformed the traditional CNN model by 9% with VGG19. The reason for the improvement with the proposed region-based CNN is due to the underlying occlusion scenarios, i.e., the region-based model works on each region of the image with the help of a selective search algorithm and then detects the person from each region and later decides the final label by majority voting. The advantage here is that even if certain parts of the GEI image are missing during dual-subject walking, the model can still recognize the individual based on the remaining visible parts. We strengthen the proposed model's ability to recognize the person from various regions of the GEI image.

Table 4: Performance of proposed region-based deep learning model for gait recognition under dual-subject-walk scenarios with occlusion conditions on Multi-Gait-SMVDU dataset

S. No.	Model	Train condition	Test condition	Accuracy	Precision	Recall	FScore
01	VGG16	Multi-Gait	Multi-Gait	54.34%	43%	44%	40%
02	ResNet50	Multi-Gait	Multi-Gait	39.13%	30%	29%	25%
03	VGG19	Multi-Gait	Multi-Gait	58.69%	57%	56%	53%
04	MobileNet	Multi-Gait	Multi-Gait	21.73%	21%	19%	19%
05	DenseNet121	Multi-Gait	Multi-Gait	41.30%	24%	31%	25%
	Proposed Region-based DL	**Multi-Gait**	**Multi-Gait**	**63.04%**	**55%**	**54%**	**51%**

Following on, in Table 5, there is another experimental condition (i.e., more uncooperative environment) in which the underlying model is trained on single-gait GEI images with different videos and subsequently tested on another set of videos, different from the test set, with a condition of dual subject gait. In this case, the model also shows a good accuracy of 50.1%. Similar to the previous experiment, we have also compared the performance of the proposed model with different traditional models, and it is observed that the proposed model outperforms the existing models with an improvement of 3% in this case. Furthermore, during training, the history of model accuracy and loss is recorded and displayed, as shown in Fig. 10 for both experiments. Model accuracy and loss values are shown to steadily converge to their optimal values.

Following that, it is observed that GEI images in dual subject gait scenarios are mostly disrupted region-wise. During a dual-subject walk, some areas of the individual walking are concealed, resulting in an incomplete GEI image. This is not the case for single-gait images as shown in Fig. 9a, but with multi-gait scenarios (i.e., dual-subject walk), some information is missed. Moreover, the occlusion rate is higher in the case of the Lateral view, in which either person walks in the dual-subject condition from left to right or right

to left. But when the camera view is oblique and frontal direction, then the occlusion rate is lower. However, in the presence of different occlusion rates, these problems will result in poor performance due to missing gait information. To address this problem, we presented a region-based strategy in which the model can detect people from visible, undisrupted regions of resulting GEI images.

Table 5: Performance of region-based deep learning model for gait recognition under dual-subject-walk scenarios with both single and multi-gait walking scenarios

S. No.	Model	Train condition	Test condition	Accuracy	Precision	Recall	FScore
01	VGG16	Single-Gait	Multi-Gait	47.12%	51%	49%	45%
02	ResNet50	Single-Gait	Multi-Gait	37.17%	42%	39%	34%
03	VGG19	Single-Gait	Multi-Gait	46.07%	49%	49%	44%
04	MobileNet	Single-Gait	Multi-Gait	19.37%	14%	18%	15%
05	DenseNet121	Single-Gait	Multi-Gait	29.84%	36%	29%	28%
	Proposed Region-based DL	**Single-Gait**	**Multi-Gait**	**50.10%**	**51%**	**51%**	**49%**

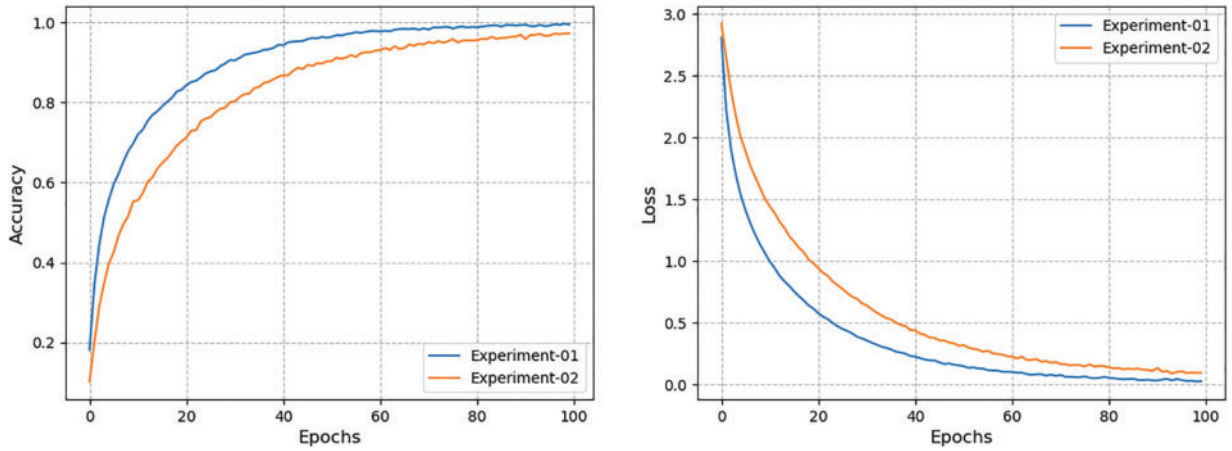

Figure 10: Accuracy and loss curves of region-based deep learning model during training for gait recognition on Multi-Gait-SMVDU dataset

4.3 Discussion and Comparisons

Vision-based human surveillance systems are gaining much importance in the current era to address several security concerns. Person identification based on gait patterns is a fascinating topic and technology since it does not require human involvement during monitoring, as in the case of other biometrics such as fingerprints, etc. In existing studies, most research studies deal with single-gait environments in which there is only one person under observation. However, one of the key goals of this study is to investigate the performance of a gait-based surveillance system in a dual-subject walk setting with two-person gaits. What, for example, if a person is walking in a dual-subject scenario and the camera captures them? The system must first determine the number of people walking, followed by detection, localization, segmentation, tracking, and finally classification based on their gait patterns. Furthermore, occlusion is a complex difficulty in dual-subject walks, as some parts of the individual silhouettes are lost owing to other individuals walking behind them. In existing studies, some very good methods have been proposed to restore occluded gait parts, such as GANs. However, these GAN-based methods fail when an object detector fails to detect a person or a

model fails to track a person. More precisely, for the reconstruction of silhouette images, silhouettes should be segmented properly as a prior step.

In comparison to previous research, we tackled the problem of multi-gait (i.e., dual-subject walk) scenarios in connection with tracking and occlusion problems in this study to design a more complete version of the gait-based surveillance system. More precisely, in this research, a dual-subject-gait model (DSG) comprised of three modules is proposed. The first module, "SLC", deals with instance-based silhouette segmentation by fine-tuning Mask-RCNN with MobileNetV2 as the backbone model. The instance segmentation model also detects the occurrences of the person in the given surveillance video, and later on person-re-identification module with a modified gallery ranking setting for person tracking is designed to track multiple people across video frames. In the last, the region-based gait recognition model is designed to classify the persons based on their gait style in the presence of missed (Occluded) parts of GEI.

The above findings demonstrate that the suggested strategy yields good results when used to exploit the system in dual-subject scenarios. The proposed "SLC" comprising Mask-RCNN shows good values of IoU score, especially in the case of PascalVoc and strict metric thresholds for both bounding boxes and segmentation masks. The performance of the gait recognition module heavily relies on this stage, as accurate human silhouette segmentation along with person detection, counting, and localization is essential for extracting meaningful features and capturing gait patterns for identification. Additionally, consistent tracking of each individual from the first to the last frame of a video plays a crucial role in ensuring reliable recognition. Likewise, after the tracking process, the region-based deep learning model is proposed to recognize persons in occluded gait scenarios. The proposed approach is further evaluated through a comparative analysis with existing studies to highlight its contributions and effectiveness. Table 6 presents a detailed comparison with previously published methods.

Table 6: Comparison of proposed Dual-subject Gait Model's (DSG) with existing studies and state-of-the-art methods

Methods	Occlusion	Gait representations	Automated silhouettes segmentation	Multi-gait scenarios	Tracking of silhouettes	Performance (Highest results)
[18]	Yes	Silhouettes images	No	No	No	CMC scores, Rank1 = 82.7%
[52]	Yes	GEI images	No	No	No	Average Acc = 97.32%
[53]	Yes	GEI images	No	No	No	Rank1 score = 93.39 (<10% occlusion) Rank1 score = 47.37 (40%–50% occlusion)
[54]	Yes	Video frames	No	No	No	Accuracy = 75.86
[55]	Yes	GEI images	No	No	No	Rank@1 = 86.00
[56]	Yes	Silhouettes images	No	No	No	Accuracy = 63.29
[36]	Yes	Silhouettes images	No	No	No	Rank@1 = 84.5%
[38]	Yes	Silhouettes images	No	No	No	Rank@1 = 16.82%
Dual-subject gait Model (DSG)	**Yes**	**GEI images**	**Yes**	**Yes**	**Yes**	**Accuracy = 63.04% For Segm = 94.50% Peron ReID Rank@1 = 58.36%**

For instance, Uddin et al. [18] proposed Wasserstein GAN with a triplet hinge loss function to reconstruct the missing gait patterns in the silhouette images. They compute the gait cycle as well as determine gait features from a single gait cycle image sequence using the reconstructed videos. Their suggested method was tested on a variety of difficult occlusion patterns and showed good results. They have used silhouette images directly for the recognition of an individual based on gait. Similarly, Kumar et al. [52] proposed

BGait-R-Net model in which occluded gait sequences are reconstructed using key pose information with Bidirectional LSTMs. In the last, the GEINet model is designed for gait-based classification using GEI images from reconstructed sequences. Subsequently, Agarwal et al. [53], Paul et al. [56], and Babaee et al. [55] also proposed DL models for the reconstruction of occluded gait information, such as LSTMs, autoencoders, and residual learning-based LSTMs. The main objective is to reconstruct the silhouette images and later on, utilizing the reconstructed images, the gait classification is carried out.

In Table 6, we can observe that in most of the studies, silhouette images are directly used from the datasets; however, in comparison with them, we have proposed an automated method of Mask-RCNN for person detection, silhouette segmentation, and localization. Another distinction is that the occlusion problem in the preceding works is handled as the restoration of silhouette images from disrupted gait sequences, e.g., GANs. In this paper, however, we treat the problem as region-based deep learning, since in real-world scenarios, a camera recording a person and the video should first go to completely detect, localize, segment, and track the person before gait-based identification. For reconstruction using GANs, the challenges highlighted in Fig. 9 must be addressed before. Specifically, an object detector should first determine where the person is, and then, if an occlusion exists, the missing regions of the silhouettes can be reconstructed. In addition, we have considered a dual-subject walk environment in connection with occlusion. From Table 6, it can be observed that this scenario is not investigated in existing studies. Moreover, different evaluation metrics are utilized in existing studies to indicate the performance of gait-recognition modules, such as rank-based metrics and classification-based metrics. Although existing occlusion-handling methods, such as guided reconstruction, show promising performance and excellent works, e.g., in [36,52,53], they typically focus on later-stage recognition improvement without emphasizing earlier processes like silhouette segmentation and person tracking, especially in complex multi-person, i.e., two-person walk in video and problems of occlusion due to dual-subject walk.

It is logically concluded from the above results and comparisons that the proposed work contributes to the literature in accessing the gait-recognition model in dual-subject walk scenarios, with occlusion as well as automated silhouette segmentation and tracking. In existing such previous steps, for example, how silhouettes are acquired, tracked, and then processed is not studied; thus, in comparison to them, this research provides a more complete and practical solution in which all steps are considered to carry out gait recognition, from beginning to end, i.e., videos, silhouettes segmentation and tracking, occlusion detection and removal to classification. On the other hand, it is critical to recognize the limitations of the research in order to inspire researchers for future work. One possible limitation in the entire framework is the silhouettes tracking phase; even if we adjusted the object tracking algorithm using a modified gallery setting and segmented images of persons, when two people walk too near enough or when their physical appearances are too similar, then the tracked IDs are switched.

Overall, the research findings reveal that occlusions are a major challenge and difficulty in multi-gait settings (i.e., dual-subject walk), and thus impede the functioning of the gait-recognition module. In this research solution is presented, however, still further improvements are required. When segmenting silhouettes with mask-RCNN, it is also found that when two people walk together and are too nearby, the instances' masks overlap, resulting in erroneous segmentation, and the pixels belonging to two people being mixed together, i.e., pixels values belong to a person 2 are classified as person 1 in instance-level segmentation. In the face of such challenges, the proposed framework "GGM" yields promising results and establishes a foundation for future researchers in occluded dual-subject walk environments. In the future, we will also address such challenges, strengthen the model, and develop new variants of DL models capable of managing multi-gait scenarios in conjunction with occlusion issues. Moreover, using advanced models for object detection and silhouette segmentation, such as Refined-Mask RCNN and feature

disentanglement-based one-stage object detectors, is one of the potential future research directions [57,58]. These models can offer improved precision in complex dual-subject-walk scenarios by addressing overlapping and occluded instances more effectively. Integrating such techniques could enhance the robustness of the proposed pipeline.

5 Conclusion

Human gait recognition is a fascinating biometric paradigm that aims to recognize a person based on how they walk. It evolves into a technology that can be used in vision-based intelligent video surveillance systems to recognize people even when in uncontrolled settings. One of the most significant benefits of using gait in video surveillance is that it can be obtained from a long distance, even with low-quality videos. In the existing literature, most of the research is focused on single individual settings while recognizing them using gait, however, the performance of this evolving technology is hidden under dual-subject-walk settings in which detection, localization, segmentation, tracking, and occlusion are one of the major problems since it weakens the gait features in a gait-cycle. To handle this, we build a dual-subject-gait model (DSG) that is able to recognize a person based on gait in case of occluded dual-subject-walk scenarios. The proposed method is divided into different stages dealing with different problems, as the SLC module deals with automated silhouette segmentation, localization, and counting of persons in a surveillance video using Mask-RCNN, and the second module deals with tracking part using a person-re-identification module with a modified gallery setting. Similarly, the last module deals with occlusion problems in a dual-subject-walk scenario by using region-based DL. The experiments are validated on SMVDU-Multi-Gait and SMVDU-Single-Gait datasets, indicating good gait recognition accuracy of 63.04% in dual-subject gait scenarios and 50.1% in single-to-multi-gait scenarios (i.e., single to dual-subject). Moreover, the performance of the segmentation model is also encouraging, with an average precision of 0.94 under strict metric IoU thresholds. Likewise, the person re-identification module achieves a rank@1 accuracy of 58.36%, respectively. Despite its strong performance, the proposed model has some limitations. In cases of severe occlusion, especially when one individual blocks another during dual-subject walking, silhouettes and Gait Energy Images (GEIs) may be partially or entirely missing, which can significantly reduce recognition accuracy.

In addition, person tracking also becomes challenging under occlusion, especially in lateral and oblique views where visibility is further compromised. Overall, the findings highlight that occlusion remains a significant challenge in dual-subject gait recognition. Additionally, the current study is limited to a single dataset; future work should consider evaluating the model on multiple datasets with diverse occlusion scenarios to enhance generalizability.

Acknowledgement: Not applicable.

Funding Statement: This research was supported by the MSIT (Ministry of Science and ICT), Republic of Korea, under the Convergence Security Core Talent Training Business Support Program (IITP-2025-RS-2023-00266605) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Zeeshan Ali and Seungmin Rho; methodology, Maryam Bukhari, Sitara Afzal and Saira Gillani; software, Zeeshan Ali and Jihoon Moon; validation, Sitara Afzal and Jihoon Moon; investigation, Saira Gillani; resources, Seungmin Rho; data curation, Zeeshan Ali and Maryam Bukhari; writing—original draft preparation, Jihoon Moon, Zeeshan Ali and Maryam Bukhari; writing—review and editing, Seungmin Rho; visualization, Saira Gillani and Sitara Afzal. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data are available on request from the authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Singh JP, Jain S, Arora S, Singh UP. Vision-based gait recognition: a survey. IEEE Access. 2018;6:70497–527. doi:10.1109/access.2018.2879896.
2. Kumar M, Singh N, Kumar R, Goel S, Kumar K. Gait recognition based on vision systems: a systematic survey. J Vis Commun Image Represent. 2021;75(6):103052. doi:10.1016/j.jvcir.2021.103052.
3. Wan C, Wang L, Phoha VV. A survey on gait recognition. ACM Comput Surv. 2019;51(5):1–35. doi:10.1145/3230633.
4. Nixon MS, Tan T, Chellappa R. Human identification based on gait. Vol. 4. Berlin/Heidelberg, Germany: Springer; 2010.
5. Shen C, Yu S, Wang J, Huang GQ, Wang L. A comprehensive survey on deep gait recognition: algorithms, datasets and challenges. arXiv:2206.13732. 2022.
6. Yang SXM, Larsen PK, Alkjær T, Simonsen EB, Lynnerup N. Variability and similarity of gait as evaluated by joint angles: implications for forensic gait analysis. J Forensic Sci. 2014;59(2):494–504. doi:10.1111/1556-4029.12322.
7. BenAbdelkader C, Cutler R, Davis L. Stride and cadence as a biometric in automatic person identification and verification. In: Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition; 2002 May 21; Washington, DC, USA. doi:10.1109/AFGR.2002.1004182.
8. Gupta SK. Reduction of covariate factors from Silhouette image for robust gait recognition. Multimed Tools Appl. 2021;80(28):36033–58. doi:10.1007/s11042-021-10941-w.
9. Han J, Bhanu B. Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell. 2006;28(2):316–22. doi:10.1109/TPAMI.2006.38.
10. Alotaibi M, Mahmood A. Improved gait recognition based on specialized deep convolutional neural network. Comput Vis Image Underst. 2017;164(13):103–10. doi:10.1016/j.cviu.2017.10.004.
11. Wu X, Yang T, Xia Z. Gait recognition based on densenet transfer learning. Int J Sci Environ. 2020;9(1):1–14.
12. Mughal AB, Khan RU, Bermak A, Rehman AU. Person recognition via gait: a review of covariate impact and challenges. Sensors. 2025;25(11):3471. doi:10.3390/s25113471.
13. Zheng S, Zhang J, Huang K, He R, Tan T. Robust view transformation model for gait recognition. In: 2011 18th IEEE International Conference on Image Processing; 2011 Sep 11–14; Brussels, Belgium. doi:10.1109/ICIP.2011.6115889.
14. Guan Y, Li CT, Hu Y. Robust clothing-invariant gait recognition. In: 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing; 2012 Jul 18–20; Piraeus-Athens, Greece. doi:10.1109/IIH-MSP.2012.84.
15. Mogan JN, Lee CP, Lim KM. Advances in vision-based gait recognition: from handcrafted to deep learning. Sensors. 2022;22(15):5682. doi:10.3390/s22155682.
16. Shao H, Wang Y, Wang Y, Hu W. A preprocessing method for gait recognition. In: International Conference of Young Computer Scientists, Engineers and Educators; 2016 Aug 20–22; Harbin, China.
17. Wang L, Tan T, Ning H, Hu W. Silhouette analysis-based gait recognition for human identification. IEEE Trans Pattern Anal Mach Intell. 2003;25(12):1505–18. doi:10.1109/TPAMI.2003.1251144.
18. Uddin MZ, Muramatsu D, Takemura N, Ahad MAR, Yagi Y. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. IPSJ Trans Comput Vis Appl. 2019;11(1):9. doi:10.1186/s41074-019-0061-3.
19. Liu F, Zheng Q, Tian X, Shu F, Jiang W, Wang M, et al. Rethinking the multi-scale feature hierarchy in object detection transformer (DETR). Appl Soft Comput. 2025;175(3):113081. doi:10.1016/j.asoc.2025.113081.
20. Mandlik SB, Labade R, Chaudhari SV, Agarkar BS. Review of gait recognition systems: approaches and challenges. Int J Electr Comput Eng. 2025;15(1):349. doi:10.11591/ijece.v15i1.pp349-355.
21. Li J, Zhang Y, Zeng Y, Ye C, Xu W, Ben X, et al. Rethinking appearance-based deep gait recognition: reviews, analysis, and insights from gait recognition evolution. IEEE Trans Neural Netw Learn Syst. 2025;36(6):9777–97. doi:10.1109/tnnls.2025.3526815.

22. Kececi A, Yildirak A, Ozyazici K, Ayluctarhan G, Agbulut O, Zincir I. Implementation of machine learning algorithms for gait recognition. Eng Sci Technol Int J. 2020;23(4):931–7. doi:10.1016/j.jestch.2020.01.005.
23. Bari ASMH, Gavrilova ML. Artificial neural network based gait recognition using kinect sensor. IEEE Access. 2019;7:162708–22. doi:10.1109/access.2019.2952065.
24. Liao R, Yu S, An W, Huang Y. A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognit. 2020;98(2):107069. doi:10.1016/j.patcog.2019.107069.
25. Teepe T, Gilg J, Herzog F, Hörmann S, Rigoll G. Towards a deeper understanding of skeleton-based gait recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2022 Jun 19–20; New Orleans, LA, USA. doi:10.1109/CVPRW56347.2022.00163.
26. An W, Liao R, Yu S, Huang Y, Yuen PC. Improving gait recognition with 3D pose estimation. In: Chinese Conference on Biometric Recognition; 2018 Aug 11–12; Urumqi, China. doi:10.1007/978-3-319-97909-0_15.
27. Gul S, Malik MI, Khan GM, Shafait F. Multi-view gait recognition system using spatio-temporal features and deep learning. Expert Syst Appl. 2021;179(1109/34):115057. doi:10.1016/j.eswa.2021.115057.
28. Arshad H, Khan MA, Sharif MI, Yasmin M, Tavares JMRS, Zhang YD, et al. A multilevel paradigm for deep convolutional neural network features selection with an application to human gait recognition. Expert Syst. 2022;39(7):e12541. doi:10.1111/exsy.12541.
29. Alsaggaf WA, Mehmood I, Khairullah EF, Alhuraiji S, Sabir MFS, Alghamdi AS, et al. A smart surveillance system for uncooperative gait recognition using cycle consistent generative adversarial networks (CCGANs). Comput Intell Neurosci. 2021;2021(1):3110416. doi:10.1155/2021/3110416.
30. Li X, Makihara Y, Xu C, Yagi Y, Ren M. Joint intensity transformer network for gait recognition robust against clothing and carrying status. IEEE Trans Inf Forensics Secur. 2019;14(12):3102–15. doi:10.1109/tifs.2019.2912577.
31. Yao L, Kusakunniran W, Wu Q, Zhang J, Tang Z, Yang W. Robust gait recognition using hybrid descriptors based on Skeleton Gait Energy Image. Pattern Recognit Lett. 2021;150(8):289–96. doi:10.1016/j.patrec.2019.05.012.
32. Bashir K, Xiang T, Gong S. Gait recognition using gait entropy image. In: Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009); 2009 Dec 3; London, UK. doi:10.1049/ic.2009.0230.
33. Roy A, Sural S, Mukherjee J, Rigoll G. Occlusion detection and gait silhouette reconstruction from degraded scenes. Signal Image Video Process. 2011;5(4):415–30. doi:10.1007/s11760-011-0245-5.
34. Hofmann M, Wolf D, Rigoll G. Identification and reconstruction of complete gait cycles for person identification in crowded scenes. In: Proceedings of the International Conference on Computer Vision Theory and Applications; 2011 Mar 5–7; Vilamoura, Algarve, Portugal. doi:10.5220/0003329305940597.
35. Muramatsu D, Makihara Y, Yagi Y. Gait regeneration for recognition. In: 2015 International Conference on Biometrics (ICB); 2015 May 19–22; Phuket, Thailand. doi:10.1109/ICB.2015.7139048.
36. Hasan K, Uddin MZ, Ray A, Hasan M, Alnajjar F, Ahad MAR. Improving gait recognition through occlusion detection and silhouette sequence reconstruction. IEEE Access. 2024;12:158597–610. doi:10.1109/access.2024.3482430.
37. Huang P, Peng Y, Hou S, Cao C, Liu X, He Z, et al. Occluded gait recognition with mixture of experts: an action detection perspective. In: Computer Vision—ECCV 2024; 2024 Sep 29–Oct 4; Milan, Italy. doi:10.1007/978-3-031-72658-3_22.
38. Gupta A, Chellappa R. MimicGait: a model agnostic approach for occluded gait recognition using correlational knowledge distillation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision; 2025 Feb 26–Mar 6; Tucson, AZ, USA. doi:10.1109/WACV61041.2025.00466.
39. Xu C, Makihara Y, Li X, Yagi Y. Occlusion-aware human mesh model-based gait recognition. IEEE Trans Inf Forensics Secur. 2023;18(6):1309–21. doi:10.1109/tifs.2023.3236181.
40. Gupta A, Chellappa R. You can Run but not hide: improving gait recognition with intrinsic occlusion type awareness. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2024 Jan 3–8; Waikoloa, HI, USA. doi:10.1109/WACV57701.2024.00579.
41. Singh JP, Jain S, Arora S, Singh UP. Dataset for human recognition under multi-gait scenario. Mendeley Data. 2019;2:349–57. doi:10.1109/icict46931.2019.8977673.

42. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. doi:10.1109/ICCV.2017.322.
43. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. doi:10.1109/CVPR.2015.7298965.
44. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv:1512.03385. 2016.
45. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. doi:10.1109/CVPR.2017.634.
46. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. doi:10.1109/CVPR.2017.106.
47. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. doi:10.1109/CVPR.2018.00474.
48. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany.
49. Hoiem D, Divvala SK, Hays JH. Pascal VOC 2008 challenge. World Lit Today. 2009;24(1):1–4.
50. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Proceedings of the 13th European Conference on Computer Vision; 2014 Sep 6–2; Zurich, Switzerland. doi:10.1007/978-3-319-10602-1_48.
51. Li W, Zhao R, Xiao T, Wang X. DeepReID: deep filter pairing neural network for person re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. doi:10.1109/CVPR.2014.27.
52. Kumara SS, Chattopadhyaya P, Wang L. BGaitR-Net: occluded gait sequence reconstruction with temporally constrained model for gait recognition. arXiv:2110.09564. 2021.
53. Das D, Agarwal A, Chattopadhyay P, Wang L. Rgait-net: an effective network for recovering missing information from occluded gait cycles. arXiv:1912.06765. 2019.
54. Chattopadhyay P, Sural S, Mukherjee J. Frontal gait recognition from occluded scenes. Pattern Recognit Lett. 2015;63(4):9–15. doi:10.1016/j.patrec.2015.06.004.
55. Babaee M, Li L, Rigoll G. Gait recognition from incomplete gait cycle. In: 2018 25th IEEE International Conference on Image Processing (ICIP); 2018 Oct 7–10; Athens, Greece. doi:10.1109/ICIP.2018.8451785.
56. Paul A, Jain MM, Jain J, Chattopadhyay P. Gait cycle reconstruction and human identification from occluded sequences. arXiv:2206.13395. 2022.
57. Zhang Y, Chu J, Leng L, Miao J. Mask-refined R-CNN: a network for refining object details in instance segmentation. Sensors. 2020;20(4):1010. doi:10.3390/s20041010.
58. Lin W, Chu J, Leng L, Miao J, Wang L. Feature disentanglement in one-stage object detection. Pattern Recognit. 2024;145(2):109878. doi:10.1016/j.patcog.2023.109878.