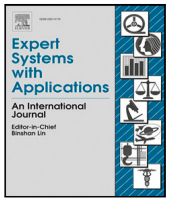
Expert Systems With Applications 246 (2024) 123181

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



BGaitR-Net: An effective neural model for occlusion reconstruction in gait sequences by exploiting the key pose information

Somnath Sendhil Kumar [a], Binit Singh [a], Pratik Chattopadhyay [a,*], Agrya Halder [a], Lipo Wang [b]

[a] Indian Institute of Technology (Banaras Hindu University), Varanasi, India
[b] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Key pose conditional vector
Occlusion reconstruction
Spatio-temporal model
Gait recognition

ABSTRACT

Gait recognition in the presence of occlusion is a challenging problem and the solutions proposed to date either lack robustness or depend on several unrealistic constraints. In this work, we propose a Deep Learning framework to detect and reconstruct the occluded frames in a gait sequence. Initially, occlusion detection is done using a VGG-16 network and for each frame the corresponding pose information is represented as a one-hot encoded vector. This vector is next fused with the corresponding spatial information using a Conditional Variational Autoencoder (*CVAE*) to obtain an effective embedding. Following this, a Bi-directional Long Short Term Memory (*Bi-LSTM*) is used to predict the occluded frames using the encoded vector sequence. A decoder next transforms these predicted frames back to the image space. Our proposed reconstruction model termed the Bidirectional Gait Reconstruction Network (*BGaitR-Net*) is formed by stacking the *CVAE*, *Bi-LSTM*, and the decoder. The *CASIA-B* and *OU-ISIR LP* datasets are used to prepare extensive gallery sets to train each of the above sub-networks and testing is done using synthetically occluded sequences from the *CASIA-B* data and real-occluded sequences from the *TUM-IITKGP* data. A thorough evaluation of our work through *Dice Score* and *GEINet*-based recognition accuracy for varying degrees of occlusion highlight the effectiveness of our model in generating frames consistent with the temporal gait pattern. Comparative study with other existing gait recognition techniques (with or without occlusion handling mechanism) and with recent Deep Learning-based video frame prediction methods emphasizes the superiority of *BGaitR-Net* over the others.

1. Introduction

Gait recognition refers to the process of identifying individuals from their walking patterns and gait is the only biometric that can be captured quite well from a distance without physical interaction with subjects (Dargan & Kumar, 2020). Due to this reason, an effective gait recognition method can be potentially used to identify suspects in surveillance zones if the gallery gait sequences of these suspects are available. An ideal gait recognition method must be able to handle all the real-life challenges including the presence of occlusion in the scene, camera viewpoint variation, clothing changes of subjects, etc.

Over the past two decades, several attempts have been made to tackle situations where the viewpoint and co-variate conditions of subjects do not match in the training and test sequences, e.g., Collins, Gross, and Shi (2002), He, Zhang, Shan, and Wang (2018), Zhang, Wu, and Xu (2019). However, significant focus has not been given to solve the challenging problem of gait recognition in the presence of occlusion. Only a few methods (Babaee, Li, & Rigoll, 2018, 2019; Chattopadhyay, Sural, & Mukherjee, 2015; Roy, Sural, Mukherjee, & Rigoll, 2011) have shown directions to approach this problem, but these methods are not effective enough to handle the variations in real-life surveillance scenarios and need further developments.

Out of the occlusion handling methods in gait recognition, the approaches discussed in Chattopadhyay et al. (2015), Roy et al. (2011) are non-Deep Learning-based. While the method in Chattopadhyay et al. (2015) works by comparing the available walking poses in the training and test sequences, that in Roy et al. (2011) uses a Gaussian Process Dynamical Model to predict the missing/occluded frames in a gait cycle and next extracts features from this reconstructed cycle to perform recognition. The method in Chattopadhyay et al. (2015) fails if no matching walking poses are found in a pair of gallery and test sequences due to heavy occlusion, whereas the assumption in Roy et al. (2011) that walking features follow a Gaussian is not a fact and the fitted Gaussian may not be able to make a good prediction about the missing frames if the input sequence is corrupted with moderate

* Corresponding author.
E-mail addresses: somnath.sendhilk.eee19@iitbhu.ac.in (S.S. Kumar), binitsingh.cse21@iitbhu.ac.in (B. Singh), pratik.cse@iitbhu.ac.in (P. Chattopadhyay), agryahalder.rs.cse21@iitbhu.ac.in (A. Halder), elpWang@ntu.edu.sg (L. Wang).

https://doi.org/10.1016/j.eswa.2024.123181
Received 23 July 2022; Received in revised form 3 November 2023; Accepted 5 January 2024
Available online 11 January 2024

to heavy degrees of occlusion. In contrast to these two methods, the work in Babaee et al. (2018, 2019) present *CNN*-based Deep Learning frameworks to reconstruct the Gait Energy Image (*GEI*) features from incomplete cycles. However, to make reliable predictions, these models also need sufficiently good *GEI* features to be provided as input, which is not possible if the degree of occlusion is high. The recent video frame prediction methods (Chang et al., 2021; Gao, Tan, Wu, & Li, 2022; Wang et al., 2022) mostly rely on recurrent networks like *LSTM* to predict the future frames after encoding the input frames of a sequence using a suitable encoder. Although these approaches can be directly applied for any occlusion reconstruction in gait sequences, it appears that for gait sequence reconstruction, information about the key pose corresponding to each frame (Roy, Sural, & Mukherjee, 2012) can be potentially fused with a suitable spatio-temporal frame generator to obtain improved prediction about the occluded/missing frames. This is since key poses are representative poses in a gait cycle and the key pose information will guide the model to predict the corresponding frame with a matching pose by preserving the temporal walking pattern.

In this work, we improve upon the existing solutions to occlusion handling in gait recognition by making use of the important frame-specific key pose information both for image encoding using *CVAE* and reconstruction using *Bi-LSTM*. Specifically, we first determine a set of key poses in walking from a large gait gallery set using an approach similar to that given in Roy et al. (2011). Next, given an occluded gait sequence occlusion detection is carried out using a pre-trained VGG-16 model (Das, Agarwal, Chattopadhyay, & Wang, 2019) following which a state transition model is used to map each frame of the sequence to the appropriate key pose. A *CVAE* is next used with the auxiliary key pose information to obtain an effective embedding for each frame and finally, a *Bi-LSTM* and *Decoder* network are employed to reconstruct the sequence from the *CVAE*-encoded frames. The effectiveness of the reconstruction has been tested based on *Dice Score* and *gait recognition accuracy* computed using *GEINet*. The main contributions of our work are summarized as follows:

- A new Deep Neural Network-based model termed *BGaitR-Net* is proposed to carry out occlusion reconstruction in gait sequences, which is formed by stacking a set of *CVAE* encoders, a *Bi-LSTM*, followed by the corresponding *CVAE* decoders in an end-to-end manner.
- A set of representative key poses in a general gait cycle is first determined and the key pose information corresponding to each frame has been used as an auxiliary input during the encoding and decoding phases of the *BGaitR-Net* as a one-hot encoded vector. This helps in generating better embedding compared to that followed in most Deep learning-based video frame prediction methods which also guides the *Bi-LSTM* to make high-quality frame reconstruction.
- We prepare an extensive gallery set of synthetically occluded sequences along with the corresponding ground truth unoccluded sequences from the *CASIA-B* and the *OU-ISIR LP* data to train the two sub-networks using suitable loss functions. This dataset can be used to train future occlusion reconstruction models and will be made publicly available along with the other pre-trained models.
- Extensive experimental evaluation and comparative study establish the effectiveness of our approach and its superiority over other occlusion handling methods in gait recognition and video frame prediction methods.

The rest of the paper is organized as follows. Section 2 presents a thorough literature survey of gait recognition in which we provide an overview of existing Non-Deep Learning and Deep Learning-based approaches to gait recognition in Sections 2.1 and 2.2, respectively. Among the Non-Deep Learning-based approaches, we have highlighted existing methods using RGB and RGB-D data (refer to Sections 2.1.1 and 2.1.2), whereas among the Deep Learning-based approaches, we have reviewed existing techniques that are extensions of traditional Deep Learning-based methods in Section 2.2.1 and those that rely on generative models to transform gait features from one domain to another in Section 2.2.2. The latter category of approaches is mostly used for performing view translation in gait recognition, or synthetically removing the effect of co-variate objects from gait features. Finally, in Section 2.3 we have elaborated on the few existing occlusion handling techniques in gait recognition along with their limitations and motivation behind the current work. Section 3 describes the architecture of the proposed *BGaitR-Net* occlusion reconstruction model along with the loss functions used to train it, and Section 4 provides a thorough experimental evaluation of the model and comparison with other existing approaches using publicly available datasets. Finally, conclusions and scopes for future research are pointed out in Section 5.

2. Related work

2.1. Non-deep learning approaches

Traditional gait recognition approaches can be classified as either appearance-based or model-based. While the appearance-based approaches extract gait features from the silhouette shape variation over a gait cycle, the model-based methods attempt to fit the kinematics of human motion in a pre-defined walking model. Appearance-based approaches have become more popular over the years due to their ease of implementation and less computational requirements and here we review only the existing appearance-based approaches in the literature.

2.1.1. Methods based on RGB data

The work in Han and Bhanu (2006) presents a feature called the Gait Energy Image (*GEI*) that computes the average of gait features over a complete gait cycle. Due to aggregating features over a gait cycle, the *GEI* cannot capture the dynamics of gait effectively. Later on, a few approaches have been developed that have made attempts to overcome the limitations of *GEI*. As an example, the work in Roy et al. (2012) introduces a pose-based feature by aggregating features from fractional parts of a gait cycle. This feature is termed Pose Energy Image (*PEI*) and it has the potential to capture the kinematics of gait at a higher resolution. A few similar fractional gait cycle-based feature extraction techniques can be seen in Chattopadhyay, Roy, Sural, and Mukhopadhyay (2014), Chattopadhyay et al. (2015) that use the RGB, depth, and skeleton streams from Kinect. However, each of these categories of approaches considers dividing a gait cycle into a fixed number of non-overlapping partitions. Another approach towards preserving the dynamic information of gait better than *GEI* is given in Zhang, Zhao, and Xiong (2010) in which a feature termed the Active Energy Image (*AEI*) is described that computes the active walking regions by subtracting the adjacent binary silhouette frames followed by averaging these difference image frames.

Instead of considering a fixed number of gait cycle partitions, in Gupta and Chattopadhyay (2021a) Gupta et al. propose using a dictionary of key pose sets, each with a different number of key poses. Next, pose-based *AEI* features are computed corresponding to each set of key poses, and the final prediction about the class of a subject is made based on the class with which the maximum number of matching key poses is observed. This approach has been seen to provide improved recognition performance over that of the previously developed features given in Han and Bhanu (2006), Roy et al. (2012), Zhang et al. (2010). In Xu, Yan, Tao, Lin, and Zhang (2007), the *GEI* features are first projected into a lower-dimensional space using Marginal Fisher Analysis, and recognition is done using the sub-space features. A viewpoint invariant gait recognition approach described in Collins et al. (2002) performs cyclic gait analysis to identify the key frames present in a walking sequence. Standard structural features such as height, width, different body-part proportions, stride length, etc., have been used for recognition via normalized correlation. All the above-mentioned approaches require a complete cycle of gait for proper functioning and hence, are not suitable for gait recognition in presence of occlusion.

2.1.2. Methods based on RGB-D data

With the introduction of *RGB-D* cameras such as Kinect, a few frontal-view gait recognition techniques (Chattopadhyay et al., 2014; Sivapalan, Chen, Denman, Sridharan, & Fookes, 2011) have also been developed. An advantage of frontal view gait recognition is that it is less prone to occlusion, as a result of which there is a higher chance of capturing clean and usable gait cycle information even from a short sequence. Since, reliable gait features cannot be extracted from frontal view binary silhouette sequences, depth streams provided by depth cameras such as Kinect have been mostly utilized in research on frontal gait recognition. The work in Battistone and Petrosino (2019) jointly exploits body structural data and temporal information from Kinect *RGB-D* streams using a spatio-temporal neural network model termed the *TGLSTM* to effectively learn long and short-term dependencies along with a graph structure. Initially, a graph is constructed from each frame containing a binary silhouette that represents the skeleton structure of the silhouette in the frame. Following this, an *LSTM* is used to capture the variation of the skeletal joint features over consecutive frames. However, the effectiveness of this method is likely to suffer if any input silhouette frame is corrupted by noise. Also, the use of depth sensors to capture gait videos in surveillance sites is not recommended due to their small depth-sensing range.

2.2. Deep learning approaches

2.2.1. Extensions of traditional approaches

With the advancement of Deep Learning, *CNN*-based models have also been extensively used for gait recognition. For example, in Hu, Li, Zhu, and Zhou (2018), Li, Hu, Zhu, and Zhou (2020), raw sensor data from the accelerometer and gyroscope of smartphones are used to monitor users' behavioral patterns. A *CNN* architecture is trained using the temporal and frequency domain data to extract an information-rich feature representation. Next, *SVM*-based classification of these features is done in the latent space to predict a person as either a legitimate user or an imposter. Recently, *CNN*s have also been used for cross-view gait recognition and also for gait recognition in the presence of co-variate objects. For example, the work in Takemura, Makihara, Muramatsu, Echigo, and Yagi (2017) describes a deep Siamese architecture-based feature comparison that works satisfactorily even for a large variation of view angles. Among the other recent Deep Learning-based gait recognition approaches, in Shiraga, Makihara, Muramatsu, Echigo, and Yagi (2016) the *GEI* features computed from a gait cycle are passed through a *CNN*-based model, termed *GEINet* to obtain deep features which are next used for classification. Since training a deep network requires tuning a large number of trainable parameters, the authors in Alotaibi and Mahmood (2017) suggest employing a small-scale *CNN* consisting of four convolutional layers (with eight features maps in each layer) and four pooling layers for gait recognition. The multi-view gait recognition framework discussed in Gul, Malik, Khan, and Shafait (2021) uses a *3D-CNN* to capture the spatio-temporal gait features after determining the best hyper-parameters of the *3D-CNN* through Bayesian optimization. In a recent approach by Ghosh (2022), the problem of co-variate condition-invariant gait recognition is approached based on stacking of a *Faster RCNN* and an *LSTM/Bi-LSTM*. Initially, object silhouettes are detected by the *Faster RCNN* model following which these are normalized and spatio-temporal gait features are extracted through the *LSTM/Bi-LSTM*. In another recent work (Lin, Zhang, Wang, Li, & Yu, 2022), a global-local based gait recognition architecture, termed *GaitGL*, is introduced that is capable of extracting global visual descriptors and local regional details through its dual-branch Global and Local Convolutional Layer (*GLCL*). The *GLCL* consists of a *GFR* extractor that extracts contextual information about the relationship among the different body parts, and a mask-based *LFR* extractor that provides intrinsic details about the posture changes of local regions.

2.2.2. Generative model-based approaches

CNN-based generative models have also been employed for handling varying co-variate conditions effectively and also for solving the challenging cross-view gait recognition problem to translate gait features from one view to a different view. For example, in Gupta and Chattopadhyay (2021b), a key pose-based gait recognition approach has been presented that can perform recognition effectively from videos with different co-variate conditions, such as wearing coat, carrying bag, etc. Here, a *GAN* model has been used to artificially transform the features with co-variate conditions to that without co-variate conditions before carrying out recognition. Additionally, in this work, the constraints of mapping frames to the different key poses, as used in other pose-based gait recognition approaches such as Chattopadhyay et al. (2014), Roy et al. (2012), have been relaxed to perform recognition effectively even if the training and test videos have different walking speeds or are captured at different frame rates. The work in Yu, Chen, Garcia Reyes, and Poh (2017) by Yu et al. focuses on developing a view-invariant and co-variate condition invariant gait recognition method based on a *GAN* framework. Given a test sequence from any view, this approach computes the *GEI* features (Han & Bhanu, 2006), and next uses a *GAN* to predict images corresponding to normal side view walking without co-variate objects. In addition to the standard *GAN* discriminator, the authors make use of an additional identification discriminator to ensure that the identity features are not lost during the view transformation process. However, this approach requires conversion of the input *GEI* features computed from any view to the corresponding side-view *GEI* features, which is expected to be time-consuming. In another similar work, namely (He et al., 2018), a new architecture termed the *Multi-Task GAN* (*MGAN*) has been introduced by He et al. that learns view-specific feature representations for transforming the gait templates across two different views. Here, the authors also present a new feature termed the Period Energy Image that preserves the temporal characteristics of gait better than the primitive *GEI* feature. However, this approach can learn the mapping between two different views only. Hence, if gait templates are available from different viewpoints, multiple such models must be trained which would make the model quite heavy. As an improvement, in Zhang et al. (2019), Zhang et al. come up with a new architecture termed the *View Transformation GAN* (*VT-GAN*) that can carry out similar view transformations across any pair of arbitrary views. Specifically, gait features in the target view are synthetically generated by conditioning on the input image from any given viewpoint and its target view indicator. An auxiliary view classifier is considered along with the standard generator and discriminator of the *GAN* to control the consistency of the generated templates. Additionally, an identity-distilling module with triplet loss is appended to the *GAN* to yield the discriminative feature embedding by retaining the identity traits.

Both the versions of the *GaitSet* model described in Chao, He, Zhang, and Feng (2019), Chao, Wang, He, Zhang, and Feng (2021) extract useful spatio-temporal information from an input sequence and integrate this information for view transformation. An improvement over the *GaitSet* model is given in Fan et al. (2020) which introduces a model termed *GaitPart* consisting of a frame-level part feature extractor that encodes the micro-motions at the different body parts followed by a temporal feature aggregator. An attempt has also been made to distill the *GaitSet* model and come up with an effective but lightweight student *CNN* model using a joint knowledge distillation algorithm in Song, Huang, Shan, Wang, and Chen (2022). However, none of these approaches are suitable for application if occlusion is present in the gait sequences. In Han, Li, Zhao, and Shen (2022), another view-invariant gait recognition approach is presented in which separable features are learned in the Cosine space through an angular softmax loss function, and simultaneously a second triplet loss function is employed to increase the separation margin among the feature vectors from different subjects. Finally, these two loss terms are optimized through batch-normalization.

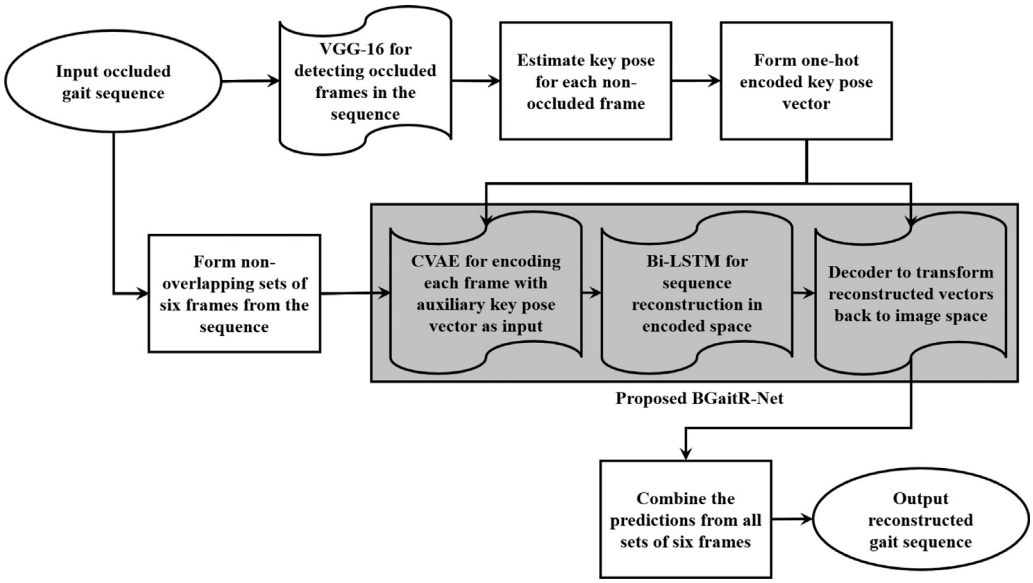


Fig. 1. A block diagram showing the pipeline of the proposed reconstruction algorithm.

2.3. Occlusion reconstruction in gait recognition

Most of the gait recognition scenarios used in the above-mentioned techniques consider a single person to be present in the field of view of a camera and also assume that at least a complete gait cycle of each individual is available. However, the presence of occlusion makes the silhouettes in the video frames noisy and hinders the capturing of a complete clean gait cycle. This affects the recognition accuracy of most traditional appearance-based approaches discussed before. Some popular approaches to handling the problem of occlusion in gait recognition are discussed next. Occlusion reconstruction has been done using a Gaussian process dynamic model in Roy et al. (2011). In this work, occluded frames in a gait sequence are first detected and next these occluded frames are reconstructed from the unoccluded frames by fitting the Gaussian model to the available set of points with the assumption that the variation of gait features over a cycle can be approximated by a Gaussian. The viability of this approach has been evaluated using the *TUM-IITKGP* data (Hofmann, Sural, & Rigoll, 2011). In Isa, Alam, and Eswaran (2010), an approach based on *SVM*-based regression is employed to reconstruct the occluded data. This reconstructed data is first projected onto the *PCA* subspace and next the projected features are classified to the appropriate class in this canonical subspace. Three different techniques for the reconstruction of missing frames have been discussed in Lee, Belkhatir, and Sanei (2009), out of which the first approach uses an interpolation of polynomials, the second one uses auto-regressive prediction, and the last one uses a method involving projection onto a convex set.

From the literature review, we observe that gait recognition in the presence of occlusion is still an emerging area of research with possibilities for significant future development. Although Deep Networks have been used to reconstruct the *GEI* features in Babaee et al. (2018, 2019), the effectiveness of these methods is likely to suffer in the presence of high degrees of occlusion. In contrast, frame-level reconstruction through deep spatio-temporal models by exploiting the useful key pose information appears to be significantly more effective for occlusion reconstruction in a gait sequence. Based on this idea, in this work we propose a new occlusion reconstruction approach in gait recognition, as detailed in the following section.

3. Proposed approach

As in any appearance-based gait recognition approach (Chao et al., 2019; Han & Bhanu, 2006; Shiraga et al., 2016), here also gait feature extraction and comparison have been done using binary silhouettes extracted from the RGB frames. Standard preprocessing techniques such as background subtraction followed by silhouette cropping and normalization are applied to obtain the binary silhouettes (Chattopadhyay et al., 2014; Shiraga et al., 2016; Zhang et al., 2010). A schematic diagram explaining the steps of the proposed occlusion reconstruction approach in binary silhouettes using our proposed model termed 'Bidirectional Gait Reconstruction Network' (*BGaitR-Net*) is shown in Fig. 1. With reference to the figure, given a gait sequence, first the occluded frames are detected using a VGG-16 model. Next, key poses in gait are estimated for the unoccluded frames and a one-hot encoded vector corresponding to the key pose of each frame is generated. Our proposed *BGaitR-Net* model accepts a contiguous sub-sequence of six frames from the complete gait sequence and regenerates these six frames through its sub-networks, namely a *CVAE*, a *Bi-LSTM*, and a *Decoder* by reconstructing the occluded frames, if any, present in the sub-sequence. To reconstruct a complete gait sequence with one or more occluded frames, first, we divide the sequence into non-overlapping sub-sequences of six frames, and then perform the reconstruction through *BGaitR-Net* and finally concatenate the predictions for each sub-sequence. The one-hot encoded key pose vector is used as an auxiliary input during each of the encoding and decoding stages that helps in preserving the temporal pattern of gait at a high resolution. It may be noted that the key pose construction, mapping, and occlusion detection stages have been performed using existing techniques and these have been explained in brief in Section 3.1. The main contribution of this work, i.e., occlusion reconstruction through *BGaitR-Net*, is discussed in detail in Section 3.2.

3.1. VGG-16-based occlusion detection and key pose estimation for unoccluded frames

3.1.1. Key pose estimation

Key poses (Roy et al., 2012) are representative poses in a gait cycle and are estimated during the training phase. These are generic and are extracted from the aligned gait cycles of a large gallery of subjects that can also include subjects from outside the gait recognition gallery set. Here, we consider a set of 50 different gait cycles extracted from the training sets corresponding to the *CASIA-B* (Yu, Tan, & Tan, 2006) and *OU-ISIR Large Population (LP)* (Iwama, Okumura, Makihara, & Yagi, 2012) data to compute the key poses through constrained K-Means clustering using an algorithm similar to that discussed in Roy et al. (2012). The optimal number of clusters determined through a rate

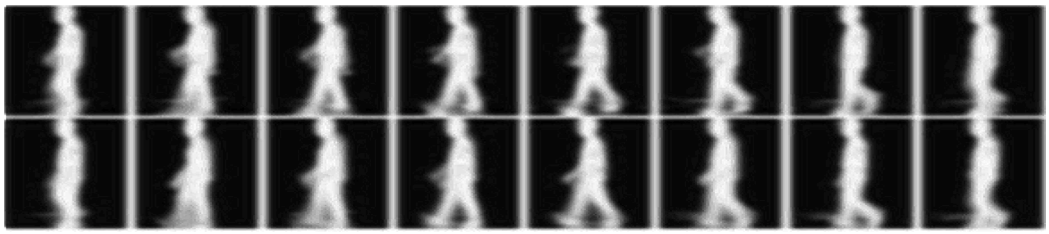


Fig. 2. 16 Key poses computed from a set of gait cycles from *CASIA-B* data and *OU-ISIR* Large Population data.

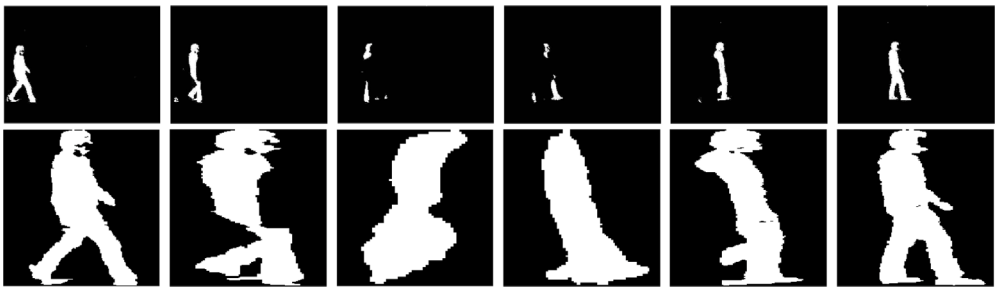


Fig. 3. The first row shows background-subtracted frames and the second row shows cropped and normalized binary silhouettes corresponding to a few occluded and unoccluded frames in a sequence from the *TUM-IITKGP* data (Hofmann et al., 2011).

distortion plot is 16, and the cluster centers obtained on termination of the clustering algorithm represent the 16 key poses in a gait cycle and are shown in Fig. 2. As observed from the figure, the set of key poses preserves the temporal order of general human walking and are not specific to any person.

3.1.2. VGG-16-based occlusion detection

Given a test gait sequence, we automatically identify the occluded and unoccluded frames present in the sequence. A few samples of occluded and unoccluded binary silhouette frames are shown in the first row of Fig. 3, and the corresponding cropped and normalized silhouettes that are used in the subsequent steps of feature extraction and reconstruction are shown in the second row of the same figure. It can be visually observed from the figure that while the first and the sixth frames are unoccluded, each of the remaining frames is occluded due to which the human body structures in the cropped and normalized silhouettes get distorted. It appears that a deep CNN model can effectively distinguish between the occluded and unoccluded silhouettes and can automate the process of classifying an input binary silhouette as either 'Òccluded' or 'Unoccluded'. Hence, we use the same pre-trained *VGG-16* model introduced in Das et al. (2019) to carry out frame-wise occlusion detection. As a gallery set, we consider an extensive dataset of 1524 silhouette images prepared from *CASIA-B* (Yu et al., 2006) and TUM-IITKGP (Hofmann et al., 2011) data with 664 images corresponding to unoccluded silhouettes and 860 images corresponding to occluded silhouettes. Binary cross-entropy loss with RMSProp optimizer is used to train the model till convergence and the trained model performs quite satisfactorily with a precision and recall of 99.53% and 98.72%, respectively and an overall accuracy of 98.89% on the training set.

3.1.3. Mapping silhouette frames to key poses

Each unoccluded frame in the above gait sequence that is predicted by the *VGG-16* model, is next mapped to the appropriate key pose. The key pose numbers for each unoccluded frame in the sequence are obtained following a frame to key pose mapping algorithm based on a state transition model similar to that given in Roy et al. (2011). This algorithm classifies each unoccluded frame of an input sequence to the appropriate key pose by maintaining the temporal order of walking and preventing the unoccluded frames already detected by the VGG-16 model from getting mapped into any key pose. Fig. 4 shows a binary silhouette sequence of 27 frames corrupted with both partial and full-body occlusions prepared from the *CASIA-B* data and the corresponding state to which each frame gets mapped using the occlusion detection and key pose mapping algorithms used in our work. In this figure, the symbol S_i (for $i = 1, 2, \ldots, 16$) indicates that the corresponding frame has got mapped to the ith key pose, and S_0 indicates that the frame is occluded. As can be seen from the figure, the occluded frames are correctly detected by the *VGG-16* model, and the key pose numbers assigned to the frames tallies with the sequence of key poses shown in Fig. 2.

We skip further discussions on the above topics since these have already been discussed in depth in the previously cited papers and focus more on describing our proposed *BGaitR-Net* occlusion reconstruction model.

3.2. Occlusion reconstruction using BGaitR-Net

We train a *Bi-LSTM* to take as input a window of encoded image frames (some of which may be occluded) along with the auxiliary key pose vector and output the corresponding reconstructed frames in the encoded space. In this work, we use a *CVAE* for frame encoding and a *Decoder* to translate the *Bi-LSTM*-predicted frames back to the image space. Further, the window size, i.e., the number of blocks in a particular layer of the *Bi-LSTM*, is considered as six. Construction of the one-hot encoded key pose vector and architecture and training details of the different sub-networks used in the *BGaitR-Net*, i.e., the *CVAE*, *Decoder*, and the *Bi-LSTM* are explained next.

3.2.1. Forming one-hot encoded key pose vector

In Section 3.1.3, we have discussed the process of obtaining the key pose numbers corresponding to the unoccluded frames. Thus, a frame is either marked as 'Occluded' or one of the key pose numbers (i.e., a number between 1 and 16) is assigned to the frame. We form a 17-dimensional one-hot encoded vector from the above information, in which the first 16 attributes correspond to the 16 key poses, and the final attribute indicates whether the frame is '*Occluded*' or not. Specifically, if a frame is occluded, the last attribute is assigned as *1* and all other attributes are assigned *0*. Otherwise, *1* is assigned to the attribute corresponding to the mapped key pose, whereas all other attributes are assigned *0*. This one-hot encoded key pose vector is used in the future feature encoding and decoding stages.

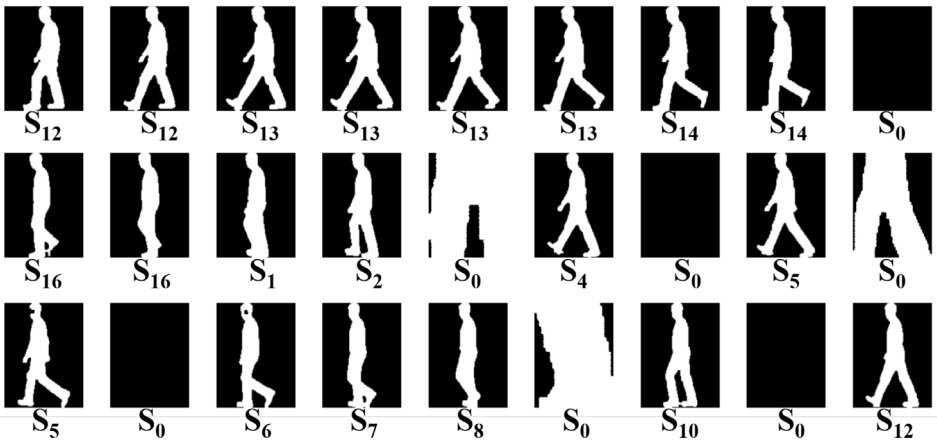


Fig. 4. An occluded frame sequence and the mapped states corresponding to each frame.

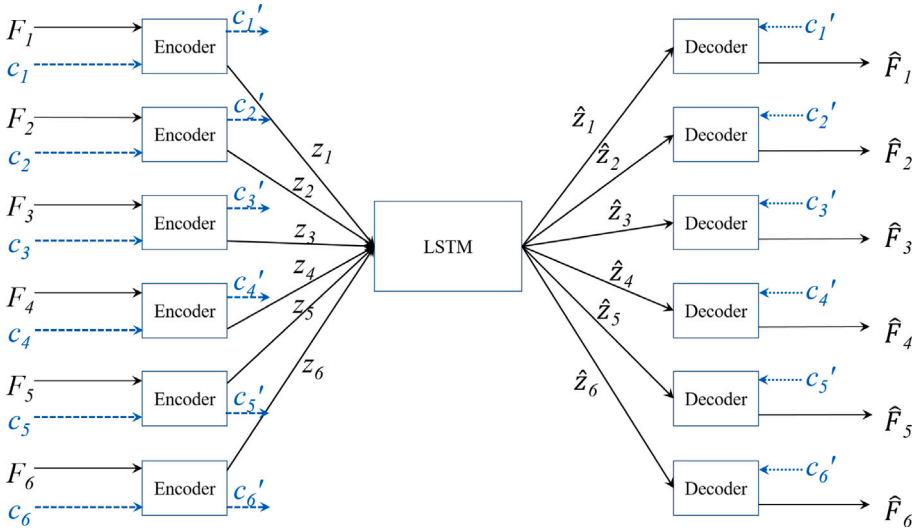


Fig. 5. An overview of the *BGaitR-Net* Model used for occlusion reconstruction.

3.2.2. Architecture and training of BGaitR-Net

Fig. 5 shows an abstract view of the frame reconstruction process using our proposed *BGaitR-Net*. Initially, an encoded vector E_i corresponding to each frame F_i is computed using a *CVAE* that takes as input a normalized frame and its corresponding one-hot encoded key pose vector denoted by c. Six encoded vectors denoted by E_1, E_2, …, E_6 corresponding to six consecutive frames in a window, namely, F_1, F_2, …, F_6, are input to a *Bi-LSTM* network that predicts the reconstructed vectors denoted by $\hat{E}_1$, $\hat{E}_2$, …, $\hat{E}_6$ for each of these six input frames. The *Bi-LSTM* next regenerates the input frames in the encoded space, following which the reconstructed vectors are passed through a *Decoder* network to obtain the reconstructed frames in the image space, namely, $\hat{F}_1$, $\hat{F}_2$, …, $\hat{F}_6$. Although a single *Encoder–Decoder* architecture has been used to obtain the encoding for each frame, for ease of explanation, in Fig. 5, six different encoders and decoders (with shared weights) have been shown, one for each image frame.

Encoder–Decoder Architecture: An insight view of the frame encoding phase along with the architecture of the *Encoder* used in the *CVAE* has been shown in Fig. 6. The figure shows rectangular blocks representing the sequence of mathematical operations that are carried out within the *Encoder* network along with the dimensions of the features that are output from each block.

With reference to the figure, the *Encoder* network fuses information from a binary silhouette frame (F) of dimensions 160 × 160 and its corresponding one-hot encoded key pose vector (c) to generate an encoded vector (Z) corresponding to the silhouette frame. The input binary image F is passed through three convolutional layers, each followed by a batch normalization operation to obtain feature maps of dimensions 4×4×64. This is next flattened into a 1024-dimensional vector and passed through a dense layer to obtain a 336-dimensional encoded representation of the input image. On the other hand, the one-hot encoded key pose vector c is also compressed through two dense layers into a 4-dimensional vector c'. These two vectors obtained from the binary image and key pose encoding are next concatenated and further passed through another dense layer to obtain a 32-dimensional feature vector. This feature vector preserves information about the input frame F as well as the key pose to which it is mapped.

During the training phase, the *Encoder* and *Decoder* networks of the *CVAE* are trained simultaneously. The *CVAE* learns to minimize the difference between the original distribution of the data from a standard normal distribution. If the function learned by the Encoder is denoted by E, then E takes as input both F and c and outputs the parameters of the fitted normal distribution, namely, the mean vector (μ) and the logarithm of the variance ($log(\sigma^2)$). Mathematically,

$$[\mu, log(\sigma^2)] = E(F, c). \tag{1}$$

Training of the *CVAE* is done through back-propagation by following a re-parameterization strategy (Kingma & Welling, 2014). Both the μ and $log(\sigma)$ vectors are also 12-dimensional, and these are combined

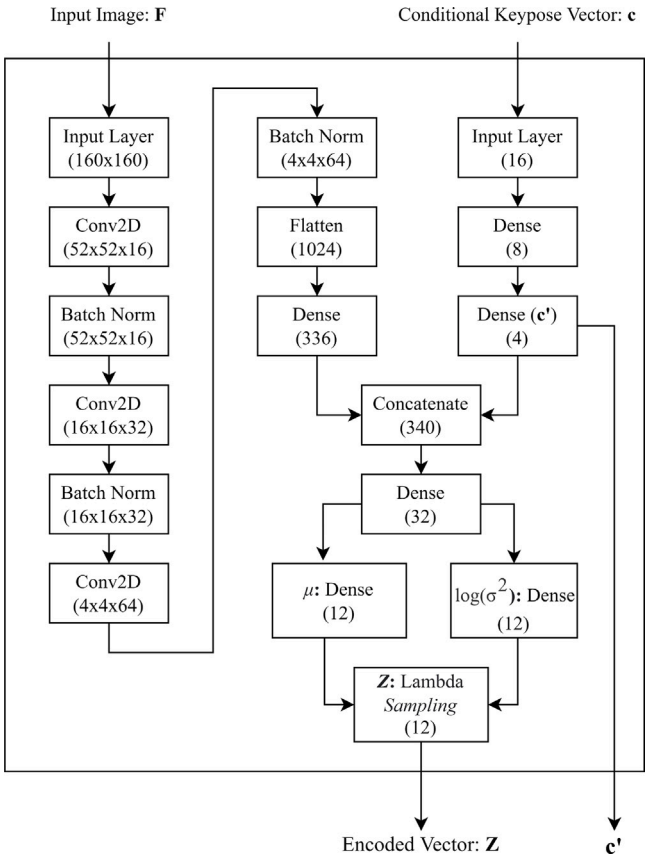


Fig. 6. Architecture of the Encoder of the *Conditional Variational Autoencoder* used for computing an embedding from each binary frame.

with a random error term (ϵ) sampled from a standard normal distribution to generate the output embedded vector Z using the following expression:

$$Z = \mu + \sigma \odot \epsilon, \tag{2}$$

where $\odot$ denotes the Hadamard product. Essentially, Z is a sample drawn from the estimated normal distribution with parameters μ and $log(\sigma)$, as discussed above, i.e., $Z \sim \mathcal{N}(\mu, \sigma)$.

The architecture of the *Decoder* network of the *CVAE* is shown in Fig. 7. As shown in the figure, this network is a fully connected convolutional network that takes as input a concatenation of a 12-dimensional vector Z and the reduced 4-dimensional key pose conditional vector c' (computed during the encoding phase). This concatenated vector $[Z\ c']$ is next uncompressed by passing it through three consecutive dense layers with 32, 336, and 1024 neurons to obtain a feature vector of dimension 1024. This resulting vector is reshaped into a $4 \times 4 \times 64$ dimensional feature map, which is further decoded using three transposed convolutional layers (shown in the figure as Conv2DTranspose) with dropout to obtain a $160 \times 160 \times 8$ dimensional feature map. These feature maps are next combined into a 160×160 dimensional feature map in the final convolutional layer, which is also the desired output reconstructed image F. The *Decoder* thus learns to generate the reconstructed frame $\hat{F}$ using information from the vectors Z and c'. Since c' is only a reduced form of the conditional key pose vector c, if the function learned by the *Decoder* network is denoted by D, then $\hat{F}$ can be represented as:

$$\hat{F} = D(z, c)\ . \tag{3}$$

The complete *Encoder–Decoder* architecture has been trained using two loss functions: the reconstruction loss and the Kullback–Leibler (KL) divergence loss. The reconstruction loss (L_{rec}), as shown in Eq. (4), is defined as the binary cross-entropy loss between the input and the reconstructed silhouettes. Mathematically,

$$L_{rec} = \frac{-1}{W\,H} \sum_{i=0}^{W} \sum_{j=0}^{H} \Big[F_{i,j} log(\hat{F}_{i,j}) + (1 - F_{i,j}) log(1 - hat F_{i,j}) \Big], \tag{4}$$

where W and H are the width and height of the input silhouette, $F_{i,j}$ denotes the intensity of the (i, j)th pixel of the input frame F, and $\hat{F}_{i,j}$ denotes the intensity of the (i, j)th pixel of the Decoder-predicted frame $\hat{F}$. The KL divergence loss (L_{kl}) for the normal probability distribution of the latent vector of the image is given by (5):

$$L_{kl} = \mu^2 + \sigma^2 - log(\sigma^2) - 1. \tag{5}$$

The incorporation of L_{kl} ensures compact and meaningful encoding of the images into the latent vector. Suppose, in total, $\mathcal{M}$ images are used in a batch while training the *CVAE*, while L^k_{rec} and L^k_{kl} respectively denote the reconstruction loss and the KL divergence loss computed for the kth image, $k = 1, 2, \ldots, \mathcal{M}$. The complete loss function (L_{cvae}) for training the *CVAE* is the weighted summation of the two losses L_{rec} and L_{kl} computed over all the $\mathcal{M}$ images and is given by (6).

$$L_{cvae} = \sum_{k=1}^{\mathcal{M}} \left(\lambda_1 L^k_{rec} + \lambda_2 L^k_{kl} \right). \tag{6}$$

In the above equation, λ_1 and λ_2 are the two user-defined constant parameters. In our experiments, the values for λ_1 and λ_2 are set to *1* and *0.5*, respectively. While during the training phase, the *Decoder* accepts the output of the *Encoder* as input, during the testing phase the *Decoder* performs reconstruction from the 12-dimensional embedding that is output by the *Bi-LSTM*.

***Bi-LSTM*-based Occlusion Reconstruction**: Since the gait of any person follows a temporal progression, and *Bi-LSTMs* are popularly used for time-series data filtering (Maleki, Maleki, & Jennings, 2021; Mejia, Avelar-Sosa, Mederos, Ramírez, & Roman, 2021), the embedding corresponding to the six binary image frames output by the *Encoder* network can be effectively filtered through the *Bi-LSTM* by maintaining the spatio-temporal relation between the adjacent frames in a gait sequence and eliminating unwanted noise, occlusion, etc. The *Bi-LSTM* network used in this work consists of three bidirectional time-distributed layers and one time-distributed *LSTM* network. It accepts a set $Z = \{ z_1, z_2, z_3, z_4, z_5, z_6 \}$ of six latent vectors from the *Encoder* as input and returns a set $\hat{Z} = \{ \hat{z_1}, \hat{z_2}, \hat{z_3}, \hat{z_4}, \hat{z_5}, \hat{z_6} \}$ of six corresponding reconstructed latent vectors as output. If the function learned by the *Bi-LSTM* is denoted by T, then

$$\hat{Z} = T(Z_{occ})\,. \tag{7}$$

The model is trained using Mean Squared Error loss (L_{mse}) between the original and predicted latent vectors given by (8) in multiple batches, each of size n. If z^j_i and $\hat{z^j_i}$ respectively denote the ith input and output latent vectors corresponding to the jth sequence in the batch, then L_{mse} is computed as:

$$L_{mse} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{6} \| z^j_i - \hat{z^j_i} \|^2. \tag{8}$$

4. Experiment and analysis

The proposed algorithm has been trained on a system with 192 GB of RAM and 16 Xeon(R) CPU E5-2609 @ 1.7 GHz and 7 GeForce GTX 1080 Ti with 11 GB RAM, 11 GB frame-buffer memory and 256 MB of BAR1 memory, and one Titan XP with 12 GB RAM, 12 GB frame-buffer memory and 256 MB BAR1 memory. Testing of the algorithm has been done on a system with 16 GB RAM and a Ryzen 5 3550H at 2.1 GHz and GeForce GTX 1650 Ti with 4 GB RAM, 4 GB frame-buffer memory, and 128 MB BAR1 memory.

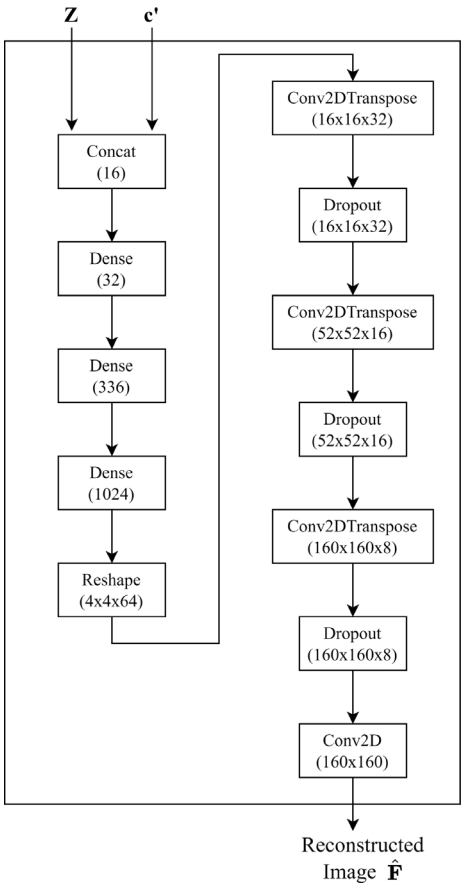


Fig. 7. Architecture of the *Decoder* of the *Conditional Variational Autoencoder* used for reconstructing an image from *LSTM*-predicted vector and conditional key pose vector.

4.1. Description of the datasets used in the study

Three different gait datasets have been used in the study for training the *BGaitR-Net*, namely the *CASIA-B* (Yu et al., 2006), the *TUM-IITKGP* (Hofmann et al., 2011), and the *OU-ISIR Large Population* (*LP*) Data (Iwama et al., 2012). Among these, both the *CASIA-B* and *OU-ISIR LP* data consist of unoccluded sequences only, whereas the *TUM-IITKGP* data consists of both unoccluded and statically/ dynamically occluded sequences. These datasets are briefly explained next.

The *CASIA-B* (Yu et al., 2006) data consists of walking sequences of 124 subjects from varying viewing angles, ranging from 0° to 180° at every 18° interval under three different settings: (a) six sequences with normal walking (in folders nm-01 to nm-06), (b) two sequences with carrying bag (in folders bg-01 and bg-02), (c) two sequences with wearing a coat (in folders cl-01 and cl-02). For conducting the experiments in the present study, we use only the normal walking sequences (i.e., sequences nm-01 to nm-06) from each viewing angle. Out of these, sequences in folders nm-01 to nm-04 corresponding to each viewing angle have been used for training purposes and the remaining two folders, namely, those in folders nm-05 and nm-06 are used for evaluation purposes after corrupting the frames in these sequences with varying levels of synthetic occlusion.

On the other hand, the *OU-ISIR LP* dataset (Iwama et al., 2012) consists of binary silhouette sequences of over 3000 subjects, and sequences from this data along with those from the *CASIA-B* data have been used to train the sub-networks of the proposed *BGaitR-Net* after corrupting these with varying levels of synthetic occlusion. Based on the chosen degree of occlusion, we decide the number of frames in a binary silhouette sequence to be occluded and randomly select these frames from the sequence for synthetic occlusion. Varying amounts of black patches are introduced on the foreground pixels of each frame that have been marked for synthetic occlusion to artificially generate partial/full occlusion in the frames. Our generated synthetic occlusion causes either a part of the silhouette or the complete silhouette to become invisible in each frame and the generated synthetic occluded frames resemble that obtained from the background subtraction of real-occluded frames. With reference to Fig. 4, the last frame in the first row, 5th, 7th, and 9th frames in the second row, and the 2nd, 6th, and 8th frames in the third row are synthetically occluded, while the rest are non-occluded. Among the occluded frames, a few frames do not have any foreground (i.e., white) pixels indicating a full occlusion scenario, whereas large white patches are seen in the other frames indicating partial occlusion. These white patches occur due to retaining the maximum-sized blob in the background-subtracted frames for silhouette normalization in the pre-processing stage (Chattopadhyay et al., 2014). The synthetically occluded CASIA-B data corrupted with varying levels of occlusion, used in our experiments, has been made available here.

It may be noted that each time our proposed *BGaitR-Net* takes as input a few consecutive binary silhouette frames along with a set of one-hot encoded vectors (as discussed in Section 3.2) and reconstructs each of these frames by exploiting the spatiotemporal information in the input sequence. The one-hot encoded vector corresponding to each frame provides the model with useful information regarding whether the frame is occluded or not. Further, if the frame is not occluded, this vector provides the corresponding key pose information. As long as a rough temporal pattern can be extracted from the input sequence, the presence of occlusion in the input sequence does not significantly affect the effectiveness of reconstruction by *BGaitR-Net*. By varying the number of frames to be occluded while generating the synthetic occluded data, it can be tested if the reconstruction model can perform effectively for the varying degrees of occlusion.

Along with the *OU-ISIR LP* dataset, we also use the *OU-ISIR MVLP* dataset (Takemura, Makihara, Muramatsu, Echigo, & Yagi, 2018) consisting of binary silhouette sequences from over 10,000 subjects to test the performance of our model for multi-view gait data. This dataset contains gait sequences for each subject from different view angles ranging from 0° to 90° with a step of 15°. We utilize these sequences along with *CASIA-B* data to train the *CVAE* sub-network of the proposed *BGaitR-Net* for the specific view angles available. A subset of sequences from the original dataset that has not been used to train the *CVAE* or the gait recognition model has been retained for the gait recognition task with multiple view angles. As this dataset does not contain any occlusion, we add synthetic occlusions using an approach similar to that used for occluding the *OU-ISIR LP*, described before.

The *TUM-IITKGP* data (Hofmann et al., 2011) consists of walking videos of 35 subjects under varying conditions and for this data also, we use the normal walking sequences to train and the statically and dynamically occluded sequences to evaluate the performance of the proposed reconstruction and recognition models. The normal walking sequences of each subject in this dataset consist of a large number of frames and from these, we segment out eight non-overlapping gait cycles for each subject to be used for training purposes. Similarly, we segment out four non-overlapping sub-sequences from each of the statically and dynamically occluded videos to construct eight separate occluded test sets for each subject to be used for evaluation purposes. These eight occluded test sets corresponding to the *TUM-IITKGP* data are labeled as *Set1*, *Set2*, . . . , and *Set8*, respectively.

The *GREW* dataset (Zhu et al., 2021) is an extensive dataset for evaluating the performances of unconstrained gait recognition algorithms. It is constructed from natural videos captured from hundreds of cameras and thousands of hours of streams in open systems. This dataset comprises 26 000 identities and 128 000 sequences, encompassing a wide range of attributes with comprehensive manual annotations. Additionally, it features a distractor set of over 233 000 sequences,

making it ideal for testing models developed for real-world applications. The dataset features a range of natural and challenging issues such as different views, distractors, varying background and carrying conditions, different dressing, occlusion, as well as varying illumination conditions, surface, speed, shoes, walking directions, etc. For evaluation of our proposed approach, we form eight different test subsets from the original dataset, labeled as *Set 1*, *Set 2*, ..., *Set 8*, each with occlusion and other challenging conditions such as different walking directions, different co-variates, etc., as mentioned above.

4.2. Training of the BGaitR-Net-based occlusion reconstruction model

The *BGaitR-Net* is trained using an extensive gallery set prepared from the *OU-ISIR*, *CASIA-B* datasets, and *OU-ISIR MVLP* datasets. The individual frames present in the sequences of randomly selected 2200 subjects from the *OU-ISIR* data and the frames corresponding to the sequences labeled nm-01, nm-02, nm-03, and nm-04 for all the 124 subjects corresponding to the *CASIA-B* data to form the gallery set for training the *CVAE*. Out of these, total of 2324 subjects in the combined data, the frames corresponding to randomly selected 2124 subjects have been used as the gallery set for training the *CVAE*, whereas the frames from the remaining 200 subjects form the validation set to evaluate the effectiveness of the model on unknown data. We train the *CVAE* with Adam optimizer for 100 epochs considering a learning rate of 0.01 at which point the model converges.

The *Bi-LSTM* is trained on sets of six latent vectors provided by the *Encoder* on the corresponding binary silhouette frames of the *CASIA-B* data. To prepare the gallery set for training the *Bi-LSTM*, we consider the unoccluded normal walking sequences corresponding to the folders nm-01, nm-02, nm-03, and nm-04 for each of the 124 subjects present in the *CASIA-B* data. Synthetic occlusion of varying amounts ranging from 10 to 70% (as explained in Section 4.1) is applied to each of these sequences and *CVAE*-based encoding is done for each frame. Next, from each sequence of encoded vectors, we extract multiple sub-sequences of six consecutive encoded frames, thereby forming a gallery of 69 560 sequences to train the *Bi-LSTM*. Out of these, 65 000 sequences are used to form the training set and the remaining 4560 are used as validation sequences to evaluate the performance of the *Bi-LSTM* on unknown data. This model is trained with Adam optimizer for 100 epochs using a learning rate of 0.01 at which point both the training and validation losses converge and the training is terminated.

4.3. Evaluation of the proposed model under varying occlusion scenarios

In our first experiment, we visually observe the quality of reconstruction of our proposed occlusion reconstruction model on sequences corrupted with occlusion. A sample result is shown in Fig. 8 using a synthetically occluded sequence generated from the *CASIA-B* data. The first row in Fig. 8 shows a set of frames from a gait cycle with several partially and fully occluded frames, whereas the second row corresponds to the reconstructed sequence after predicting the occluded frames through our *BGaitR-Net*. The third row in the figure corresponds to the ground-truth frames present in the original sequence. The good reconstruction quality of our *BGaitR-Net* is evident by comparing the second and the third rows of the figure.

Further, to quantitatively evaluate the reconstruction quality of the proposed *BGaitR-Net*, we use the Sørensen–Dice similarity score (Carass et al., 2020) as a metric to measure the degree of similarity between the predicted and ground-truth images. The test set corresponding to the *CASIA-B* data constructed from sequences labeled nm-05 and nm-06 have been used for this experiment after corrupting the sequences randomly with 10%–50% occlusion. The value of the *Dice score* lies between '*0*' and '*1*', where a value close to '*1*' indicates a high similarity between the ground-truth and the predicted frames, whereas a value close to '*0*' indicates no-similarity between the two. We observe that the average *Dice score* of the *CVAE* after convergence corresponding to the frames of the validation set of 200 subjects is *0.982*, and the average *Dice score* computed from the frames of the above-mentioned test sequences is *0.972*, which is quite good and emphasizes the fact that the proposed *BGaitR-Net* is capable of successfully handling moderately high degrees of occlusion.

Table 1
Gait recognition accuracy and *Dice score* of reconstruction for synthetically occluded test sequences generated from the *CASIA-B* data considering varying degrees of occlusion and gait recognition accuracy for real occluded sequences in the *TUM-IITKGP* data.

Dataset	Occ. Degree/ Set No.	Reconst. Dice Score	*GEINet*-based Rank 1 Accuracy (%)
CASIA-B (Synthetically Occluded)	≤10	0.99	99.83
	10 – 20%	0.98	99.53
	20 – 30%	0.95	99.32
	30 – 40%	0.90	97.16
	40 – 50%	0.86	95.00
	50 – 60%	0.86	93.21
	60 – 70%	0.82	91.22
	70 – 80%	0.78	76.65
	80 – 90%	0.75	60.05
TUM-IITKGP (Real Occluded)	Set 1	–	96.30
	Set 2	–	96.10
	Set 3	–	94.80
	Set 4	–	94.20
	Set 5	–	93.70
	Set 6	–	93.60
	Set 7	–	93.30
	Set 8	–	93.00

To verify the effectiveness of our overall approach, we evaluate the gait recognition accuracy obtained on the *BGaitR-Net*-reconstructed sequences using an existing Deep Learning-based gait recognition model, termed *GEINet* (Shiraga et al., 2016). The synthetically occluded sequences from the *CASIA-B* data and the statically/dynamically occluded sequences from the *TUM-IITKGP* data have been used for this experiment. We experiment with nine different degrees of synthetic occlusion introduced on the test sequences of the *CASIA-B* data, namely, 0%–10%, 10%–20%, 20%–30%, ..., 80%–90%. While experimenting with the *CASIA-B* data, for each subject the *GEINet* model is trained using the four GEIs computed from the training sequences present in the folders nm-01 to nm-04 (refer to Section 4.1). Similarly, while experimenting with the *TUM-IITKGP* data, the *GEINet* model is trained using the *GEI*s computed from the four normal walking sequences of each subject reserved for training purposes (refer to Section 4.1).

Results are shown in terms of both Reconstruction *Dice Score* and *GEINet*-based Rank 1 accuracy in Table 1 for the synthetically occluded sequences generated from the *CASIA-B* data. For real occluded sequences from the *TUM-IITKGP* data, we present only the Rank 1 accuracy since ground truth information is not available to compute the *Dice scores*.

In the table, the first column corresponds to the dataset name, the second column corresponds to a particular occluded gait sequence, the third column corresponds to the *Dice Score*, and the fourth column corresponds to the *Rank 1 recognition accuracy* computed from the predictions of the *GEINet* model.

From the third column, it is observed that the *Dice Score* of recognition is 0.90 or higher if the degree of occlusion is 40% or less. Even for a very high degree of synthetic occlusion (i.e., 90%), the *Dice Score* is 0.75, which is quite impressive. From the fourth column of the table, it is observed that for the synthetically occluded CASIA-B data, the Rank 1 accuracy is greater than or equal to 95% for low to moderate degrees of synthetic occlusion, i.e., when the degree of occlusion is in the range 0%–50%, whereas for 60%–70% occlusion the accuracy is 91.22%, and for a very high degree of occlusion, i.e., 80%–90%, the accuracy is 60.05%. It can be inferred from the results of synthetically occluded CASIA-B data that as the degree of occlusion is made higher, the required spatiotemporal information in

Fig. 8. The first row shows sample frames from a synthetically occluded sequence from CASIA-B data, second row corresponds to the *BGaitR-Net*-predicted frames, and the third row shows the corresponding ground-truth frames.

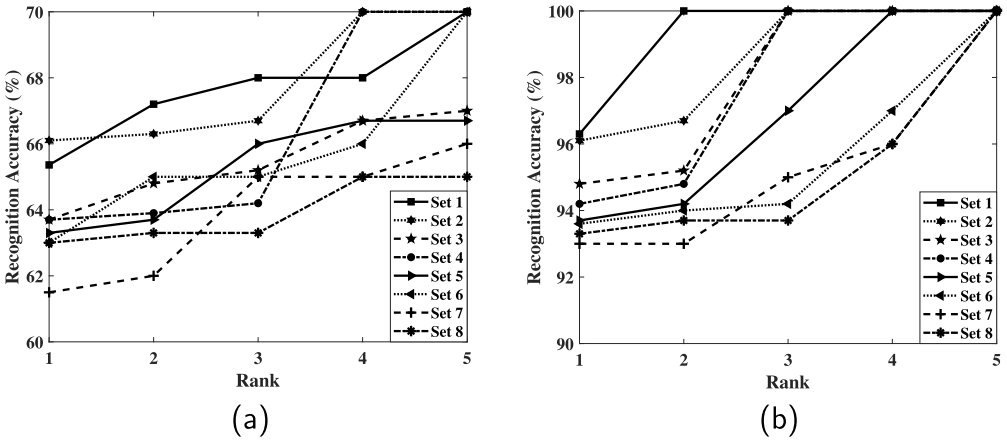


Fig. 9. CMC curves showing rank-wise improvement in recognition accuracy of *GEINet* on (a) the eight occluded test sets present in the *TUM-IITKGP* data and (b) on the same sequences after reconstruction using *BGaitR-Net*.

the sequence starts degrading, which in turn, affects its reconstruction effectiveness leading to lower Dice Score and recognition accuracy. Also, the Rank 1 accuracy obtained for each of the eight occluded test sets corresponding to the *TUM-IITKGP* data is 93% or above. The significantly high recognition accuracy of *GEINet* on each of the above-occluded test sets once again emphasizes that the reconstruction quality of our proposed *BGaitR-Net* is indeed good.

Next, we study the rank-wise performance improvement of the *GEINet* model on the eight real-occluded test sets of the *TUM-IITKGP* data and also on the *BGaitR-Net*-reconstructed sequences for the same test sets as the value of the rank is increased from *1* to *5*. Corresponding results are presented in Figs. 9(a)–(b) through Cumulative Match Characteristic (CMC) curves. In these figures, the horizontal axis represents the rank (i.e., the number of top predictions of the *GEINet* to be considered for computing the accuracy) and the vertical axis corresponds to the recognition accuracy at a particular rank (in percentage). On comparing Figs. 9(a) and 9(b), it is observed that for each test set, the recognition accuracy is significantly higher at all the ranks on using the reconstructed sequences. While the maximum accuracy achieved at Rank 5 for the occluded sets is 70% (as seen from Fig. 9(a)), that achieved at Rank 5 for the reconstructed sets is 100%. It is further seen from Fig. 9(b) that at Rank 1, all the reconstructed test sets show an accuracy greater or equal to 93%, and the corresponding accuracy at Rank 4 for all the test sets is higher than 96%. The results on the real occluded sequences of the *TUM-IITKGP* data are indeed encouraging and also emphasize the usefulness of our *BGaitR-Net*-based occlusion reconstruction approach.

4.4. Evaluating the robustness of the proposed BGaitR-Net model

Since the *Bi-LSTM* sub-network is the core framework for performing reconstruction in our proposed *BGaitR-Net*, in the next experiment we test the robustness of this model by observing how much it generalizes across different training datasets. The training set corresponding to the *CASIA-B* data has been used for this experiment (refer to Section 4.2). Specifically, we use stratified $\mathcal{K}$-fold cross-validation, i.e., we partition the entire dataset of 69 560 training sequences extracted from this data into $\mathcal{K}$ equal parts randomly, select $(\mathcal{K}-1)$ parts for training the *Bi-LSTM*, and one of the parts as the validation set to compute the average *Dice score*. The same trained model of *CVAE* as considered in the previous experiments has also been used here to transform the images into latent space and convert the *Bi-LSTM*-predicted vectors back to the image space. This process is repeated $\mathcal{K}$ different times to obtain $\mathcal{K}$ different average *Dice score* values. We consider five different values for $\mathcal{K}$, i.e., *2*, *3*, *5*, *10*, *16*, and for the above-mentioned five values of $\mathcal{K}$, the training batches are formed with 50%, 66.7%, 80%, 90%, and 93.8% samples from the complete dataset of 69 560 sequences, respectively. The $\mathcal{K}$ readings thus obtained are then plotted using a box

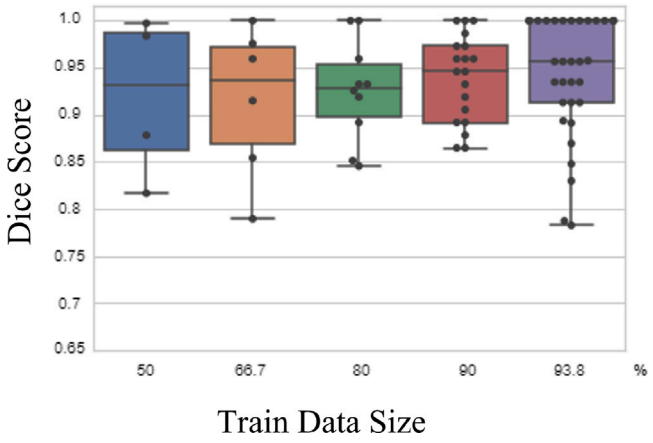


Fig. 10. Average *Dice scores* after training the *Bi-LSTM* model with $\mathcal{K}$-fold cross-validation for the following values of $\mathcal{K}$: *2*, *3*, *5*, *10*, *16*.

Table 2
Gait recognition on the *GREW* dataset with natural occlusion.

Set No.	Without BGaitR-Net		With BGaitR-Net	
	GEINet (%)	GaitGL (%)	GEINet (%)	GaitGL (%)
Set 1	58.67	63.48	64.33	72.33
Set 2	61.39	70.67	65.48	74.67
Set 3	56.78	61.33	62.56	70.24
Set 4	56.89	62.75	63.74	71.66
Set 5	62.67	72.46	69.20	78.65
Set 6	60.46	68.25	68.75	76.35
Set 7	59.33	67.67	66.67	75.33
Set 8	57.02	65.79	65.03	73.45
Avg.	**59.15**	**66.55**	**65.72**	**74.09**

plot in Fig. 10 that helps in visualizing the robustness of the *Bi-LSTM* model used in the *BGaitR-Net*. It can be seen from the plot that there is a steady increment in the average *Dice score* from 0.93 (when trained on 50% of the dataset) to 0.96 (when trained on 93.8% of the dataset). Also, the range of the average *Dice score* values obtained after training the *Bi-LSTM* $\mathcal{K}$ times for any value of $\mathcal{K}$ is quite small which highlights that *Bi-LSTM* generalizes well for varying training datasets.

4.5. Evaluation using multi-view datasets and datasets with different co-variate conditions

In this sub-section, we evaluate the performance of our proposed *BGaitR-Net* using the test sets of the *GREW* dataset and the *OU-ISIR MVLP* dataset, as explained in Section 4.1 in terms of Rank 1 accuracy given by two popular gait recognition models, namely *GaitGL* (Lin et al., 2022) and *GEINet* (Shiraga et al., 2016). Corresponding results for the *GREW* dataset are presented in Table 2 in which the individual rows correspond to the results obtained for the different test sets, namely *Set 1*, *Set 2*, . . . , *Set 8*, and the last row corresponds to the average accuracy computed from the results of the eight test sets. In the table, the first column corresponds to the set index, the second and third columns correspond to the *GEINet*-based and *GaitGL*-based recognition accuracy for the different test sets before the *BGaitR-Net*-based reconstruction, and the fourth and fifth columns correspond to the *GEINet*-based and *GaitGL*-based recognition accuracy for the different test sets after the *BGaitR-Net*-based reconstruction. From the second and third columns in the above table, it is observed that without applying reconstruction the recognition accuracy of *GEINet* ranges from 56.78–62.67% and that of the *GaitGL* ranges from 61.33–72.46%. On the other hand, a drastic improvement in the recognition accuracy is observed for both models once the *BGaitR-Net*-based reconstruction is applied on the input occluded test sequences, as can be viewed from the fourth and fifth columns of the same table. The corresponding accuracy ranges improve to 62.56–69.20% and 70.24–78.65% respectively for the *GEINet* and the *GaitGL* recognition models. The improvement in the average recognition performance using both models after applying reconstruction is evident from the last row of the table. It may be noted that the sub-networks of the proposed *BGaitR-Net* model have not been trained using silhouettes/sequences from the *GREW* dataset. Still, a satisfactory average recognition rate of 65.72% for the *GEINet* model and 74.09% for the *GaitGL* model on the test sets of the *GREW* data indicates that the proposed *BGaitR-Net* indeed has a very good reconstruction ability and it can be potentially used to reconstruct sequences effectively from any given occluded gait data. Also, as expected, in general higher gait recognition accuracy values are obtained using a more sophisticated *GaitGL* model than the *GEINet* model for gait recognition.

Next, we study the effectiveness of the *BGaitR-Net* in dealing with occluded sequences from varying view angles. For this, we consider the test sequences from the *OU-ISIR MVLP* dataset. Given any gait sequence, a view detector given in Guan, Li, and Hu (2012) is first employed to determine the walking direction and obtain the walking key poses from that direction. Table 3 presents the recognition results for the different view angles (or, walking direction) present in the *OU-ISIR MVLP* data using *GEINet* and *GaitGL* separately with the proposed *BGaitR-Net*. Also, for this experiment, we derive separate key pose sets for each of the varying walking directions and fine-tune the *CVAE* previously trained using the *CASIA-B* and the *OU-ISIR* data using silhouettes corresponding to the target walking direction.

Table 3
Gait recognition accuracy on *OU-ISIR MVLP* with 40% synthetic occlusion.

View angle (in degrees)	GEINet (%)	GaitGL (%)
0	39.88	42.75
15	42.37	44.28
30	56.48	57.99
45	70.60	72.34
60	73.49	76.75
75	87.66	89.31
90	92.48	94.64

Table 4
Gait recognition results on CASIA-B dataset(40% synthetic occlusion) for multiple view angles with varying walking conditions.

View angle (in degrees)	Normal walking		Carrying bag		Wearing coat	
	GEINet (%)	GaitGL (%)	GEINet (%)	GaitGL (%)	GEINet (%)	GaitGL (%)
0	47.70	49.67	48.25	31.75	38.25	39.75
18	64.69	63.50	52.78	46.77	54.66	58.78
36	81.33	86.25	64.24	52.67	66.89	69.17
54	90.72	92.79	76.87	73.12	83.23	86.45
72	94.78	94.33	84.33	87.98	92.54	94.69
90	96.12	97.43	88.69	90.36	96.45	98.45
108	94.69	95.63	86.89	88.26	93.12	96.23
126	91.73	93.21	74.63	74.67	82.38	84.74
144	83.48	87.37	63.98	53.18	64.77	67.36
162	56.40	61.27	51.33	48.25	51.15	54.77
180	39.74	51.74	47.77	28.54	39.69	42.67

In this table, the individual rows correspond to the different view angles, as specified in the first column. The second column presents the *Rank 1 Accuracy* obtained using *GEINet* computed from test sequences corresponding to each view angle. The third column shows similar results using the *GaitGL* model. We can observe from the table that the accuracy of gait recognition tends to decrease gradually with an increase in the view angle from 90°. Until the 45° view angle, gait recognition accuracy remains above 70%. Beyond this point, only partial information about a subject's structure is available and the observed gait kinematics gets hindered. Within the view angle range of 45°–90°, the average accuracy of *GEINet* is 81.05% and that of *GaitGL* is 83.26%. These accuracy values can be regarded as reasonably high for reconstructed sequences with 40% occlusion, once again emphasizing the superior reconstruction quality of *BGaitR-Net*.

Table 4 presents the *Rank 1 accuracy* of gait recognition on the 40% synthetically occluded test sequences of the *CASIA-B* dataset under different walking conditions and view angles. Each frame of a test sequence is first passed through the trained *Pix2Pix GAN* model given in Gupta and Chattopadhyay (2021b) to eliminate the co-variate objects like carrying bag or wearing coat. Once all the frames of a sequence are passed through the above *GAN* model, we obtain a co-variate object-free gait sequence which can, henceforth, be used for occlusion reconstruction and recognition. It may be noted that the above *GAN*-model has been specifically trained to synthetically remove co-variate conditions such as carrying bag and wearing coats from *GEI* features, whereas, in this work, we have used this model directly to remove the co-variate objects from each binary silhouette frame. The results in the table indicate that the accuracy for the different co-variate conditions is slightly lower than that of normal conditions, which may be attributed to a certain loss of silhouette structural content during *GAN*-based

co-variate object-removed frame generation. As expected, the highest accuracy rates are observed for the 90° view angle, which gradually decreases as the viewing angle is shifted towards 0° or towards 180°. In terms of gait recognition, usable view angles appear to range from 36° to 144°, with a *Rank 1 accuracy* greater than 80% for normal walking conditions. For each of the other two co-variate conditions, higher recognition accuracy is generally observed when the person is wearing a coat rather than carrying a bag. This is since for the CASIA-B data, the former co-variate condition usually retains the silhouette structural information substantially better than the latter. Thus, the resulting *GAN*-generated silhouette is also expected to have less silhouette-level distortions for the wearing coat condition than that for the carrying bag condition, which is also the main reason behind the above observation. Further, within the range of usable viewing angles, the normal walking condition has an average accuracy of 90.41% and 92.43% for *GEINet* and *GaitGL*-based classification, respectively. The corresponding values for the carrying bag condition are 77.09% and 74.32%, whereas that for the wearing coat condition are 82.77% and 85.30%, respectively. From the above results, it can be inferred that the performance of our *BGaitR-Net* model in reconstructing gait sequences is impressive, allowing the possibility of employing improved recognition models than *GEINet* or *GaitGL* to attain even higher recognition accuracy.

4.6. Comparative study of our proposed approach with other existing techniques

Next, we make a comparative study of the reconstruction quality of our proposed *BGaitR-Net* with that of the reconstruction algorithms specified in some popular occlusion handling methods in gait recognition, namely (Babaee et al., 2018, 2019; Roy et al., 2011) and also some recent video frame prediction methods that exploit the spatio-temporal information from sequences, namely (Chang et al., 2021; Gao et al., 2022; Guen & Thome, 2020; Wang, Jiang, Yang, Li, Long, & Fei-Fei, 2019; Wang et al., 2022). Among the frame prediction methods, in Wang et al. (2019) a model termed as *E3D-LSTM* is discussed that integrates 3D convolutions into RNNs, which makes local perceptrons of RNNs motion-aware and enables the memory cells to store better short-term features. For long-term relations, each memory state interacts with its historical records via a gate-controlled self-attention module. The estimated cell state and the spatio-temporal memory state are next aggregated to make the frame prediction. On the other hand, in Guen and Thome (2020), a dual-branch Deep model termed as the *PhyDNet* is presented that jointly learns the latent space to disentangle physical dynamics from residual information. The physical dynamics are modeled through *PhyCell* using a prediction correction paradigm, while the residual information is modeled using a *ConvLSTM*. The outputs from both the above units are aggregated to predict the future frame. The *MAU* model introduced in Chang et al. (2021) uses two modules, namely an attention module and a fusion module. The fusion module is utilized to aggregate the motion information from the attention module and the current spatial state to predict the next frame. The work in Gao et al. (2022) uses a video prediction model that is based on CNNs and trained with MSE loss, whereas that in Wang et al. (2022) retains long-range features by introducing a decouple loss in the ST-LSTM cells, reversed schedule sampling, and action-conditioned video prediction.

For each of the video frame prediction methods, i.e., (Chang et al., 2021; Gao et al., 2022; Guen & Thome, 2020; Wang et al., 2019, 2022), and occlusion handling methods (Babaee et al., 2018, 2019; Roy et al., 2011) we have used the trained models shared by the authors of the corresponding work publicly and re-trained the entire models till convergence with the same gallery set used to train our *BGaitR-Net*. The synthetically occluded sequences generated by introducing 50% occlusion on the test set of the *CASIA-B* data have been used in this experiment. It may be noted that the approaches discussed in Babaee et al. (2018, 2019) carry out reconstruction of the *GEI* features, whereas each

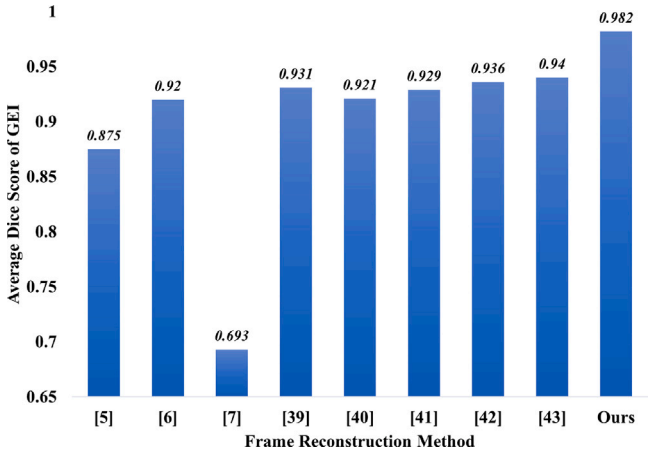


Fig. 11. Comparative study of the different reconstruction algorithms in terms of average *Dice score* of *GEI*.

of the other techniques used in this comparative study performs frame-level reconstruction. For a fair comparison, in this experiment, we compare the quality of the *GEI*s computed from the predicted frames by our method and each of Chang et al. (2021), Gao et al. (2022), Guen and Thome (2020), Roy et al. (2011), Wang et al. (2019, 2022) with that of the reconstructed *GEI*s generated by Babaee et al. (2018, 2019) in terms of average *Dice score*. Fig. 11 presents the corresponding results through a bar plot The height of each bar in the plot represents the average *Dice Score* given by the corresponding method stated along the horizontal axis. From the plot, it can be seen that the *GEI* reconstruction quality using the proposed *BGaitR-Net* is the best among all the other approaches used in the study. The reconstruction quality of the recent video frame prediction methods (Chang et al., 2021; Gao et al., 2022; Guen & Thome, 2020; Wang et al., 2019, 2022) and also Babaee et al. (2018, 2019) are also relatively good and closely comparable to each other. However, the method in Roy et al. (2011) performs poor quality reconstruction as is evident from the average *Dice score* value. This is mostly because (Roy et al., 2011) approximates the walking features over a gait cycle with a Gaussian, which cannot be used to effectively reconstruct sequences corrupted with moderately high 50% synthetic occlusion, as in the present study.

We further perform a comparative study of our work with the same approaches used in the previous experiment as well as with some other popular gait recognition techniques with and without occlusion reconstruction mechanism and observe the overall *Rank 1* gait recognition accuracy values given by these different methods on the test sets of the *TUM-IITKGP* and 50% synthetically occluded *CASIA-B* datasets. Specifically, we categorize the existing approaches into three different groups, namely (i) gait recognition methods without occlusion handling mechanism, (ii) video frame prediction methods, and (iii) gait recognition methods with occlusion handling mechanism. Among the non-occlusion handling methods, we consider some popular primitive approaches, namely those in Alotaibi and Mahmood (2017), Gupta and Chattopadhyay (2021a, 2021b), Han and Bhanu (2006), Lin et al. (2022), Roy et al. (2012), Shiraga et al. (2016), Zhang et al. (2010). Although, the method (Lin et al., 2022) has been tested on occluded gait sequences as well, it does not contain any specific occlusion reconstruction module and hence we put it in the Non-Occlusion Handling category of approaches. The video frame prediction methods include (Chang et al., 2021; Gao et al., 2022; Guen & Thome, 2020; Wang et al., 2019, 2022), for each of which the gait recognition accuracy is computed using the same trained *GEINet* (Shiraga et al., 2016), as discussed in the previous experiments. The third category of methods, namely the occlusion handling methods, includes the work in Babaee et al. (2018, 2019), each of which performs recognition after

Table 5
Comparative analysis of the proposed work with existing approaches on the real occluded sequences of the *TUM-IITKGP* data and synthetically occluded sequences of the *CASIA-B* data in terms of *GEINet*-based Rank 1 accuracy.

Type	Method	Accuracy (%)	
		TUM-IITKGP	CAS IA-B
Non-occlusion handling	Alotaibi and Mahmood (2017)	76.42	74.19
	Shiraga et al. (2016)	76.79	63.71
	Han and Bhanu (2006)	65.71	56.45
	Roy et al. (2012)	70.23	79.83
	Zhang et al. (2010)	73.54	62.10
	Gupta and Chattopadhyay (2021a)	76.79	76.77
	Gupta and Chattopadhyay (2021b)	78.36	77.58
	Lin et al. (2022)	80.23	80.54
Frame Prediction	Wang et al. (2019)+Shiraga et al. (2016)	47.66	91.54
	Guen and Thome (2020)+Shiraga et al. (2016)	63.33	87.66
	Chang et al. (2021)+Shiraga et al. (2016)	76.66	94.36
	Gao et al. (2022)+Shiraga et al. (2016)	78.41	89.54
	Wang et al. (2022)+Shiraga et al. (2016)	82.68	92.68
Occlusion Handling	*BGaitR-Net*+(Gupta & Chattopadhyay, 2021a)	96.37	96.77
	BGaitR-Net+(Gupta & Chattopadhyay, 2021b)	95.56	97.58
	Babaee et al. (2018)	78.92	81.45
	Babaee et al. (2019)	80.00	92.74
	Chen et al. (2009)	77.65	89.51
	Chattopadhyay et al. (2015)	85.32	89.51
	Roy et al. (2011)	68.57	75.23
	BGaitR-Net+(Shiraga et al., 2016)	97.32	98.17
	BGaitR-Net +(Lin et al., 2022)	**98.64**	**99.54**

reconstructing the *GEI* through a *CNN* and Chattopadhyay et al. (2015), Chen, Liang, Zhao, Hu, and Tian (2009) and perform recognition using only the available frames without occlusion reconstruction. In this category, we also study the accuracy given by four other recent non-occlusion handling methods, namely, (Gupta & Chattopadhyay, 2021a, 2021b; Lin et al., 2022; Shiraga et al., 2016) on the sequences reconstructed by our *BGaitR-Net*. The same gallery set formed from the *CASIA-B* and *OU-ISIR* datasets, as discussed in the previous experiment, has been used to train each of the occlusion reconstruction models for the approaches given in Babaee et al. (2018, 2019), Roy et al. (2011) and video frame prediction models given in Chang et al. (2021), Gao et al. (2022), Guen and Thome (2020), Wang et al. (2019, 2022) in the comparative study. Results are shown in Table 5 in terms of *Rank 1 accuracy* for both the *TUM-IITKGP* and the *CASIA-B* datasets.

The effectiveness of the proposed *BGaitR-Net*-based occlusion reconstruction method can once again be inferred from the gait recognition accuracy values shown in the table. It can be seen that fusion of *BGaitR-Net* with existing gait recognition methods, namely (Gupta & Chattopadhyay, 2021a, 2021b; Lin et al., 2022; Shiraga et al., 2016) results in a significantly high Rank 1 accuracy (> 95%) for both synthetically and real-occluded test sets. With reference to the table, the recognition accuracy provided by Gupta and Chattopadhyay (2021a, 2021b), Lin et al. (2022) on real-occluded TUM-IITKGP dataset are 76.79%, 78.36%, and 80.23, respectively and on synthetically occluded CASIA-B data are 76.77%, 77.58%, and 80.54%, respectively. The accuracy values improve to 96.37%, 95.56%, 98.64% for the *BGaitR-Net*-reconstructed TUM-IITKGP data and 96.77%, 97.58%, and 99.54% for the *BGaitR-Net*-reconstructed CASIA-B data. The highest recognition accuracy for both datasets is obtained using *GaitGL* (Lin et al., 2022) as the recognition model. This is because unlike the methods in Gupta and Chattopadhyay (2021a, 2021b), Shiraga et al. (2016) which capture global contextual features, that in Lin et al. (2022) focuses on deriving both global and local features, leading to a higher gait recognition accuracy. In comparison, the accuracy values given by the other existing occlusion handling methods in gait recognition, namely (Babaee et al., 2018, 2019; Chattopadhyay et al., 2015; Chen et al., 2009) on the same test sets are quite low. It may also be noted that each of the video frame prediction methods, i.e., Chang et al. (2021), Gao et al. (2022), Guen and Thome (2020), Wang et al. (2019, 2022) combined with *GEINet* (Shiraga et al., 2016) show a significantly higher recognition accuracy for the *CASIA-B* data than that for the *TUM-IITKGP* data. This is due to the fact that the normalized binary silhouette frames obtained from the *TUM-IITKGP* data are quite noisy and the encoding techniques used in these approaches based on Vanilla Autoencoder are not effective enough for noisy inputs. In contrast, the Variational Autoencoder along with the conditional key pose vector, as used in the proposed *BGaitR-Net* model, helps in obtaining a better embedding of the binary silhouette frames resulting in high-quality reconstruction even from the noisy *TUM-IITKGP* data. The *Rank 1 accuracy* given by our approach on the *TUM-IITKGP* data is *97.32%*, which improves over the best-performing video frame prediction method, i.e., Wang et al. (2022), by more than *14%* which is significant. Also, as expected, the recognition accuracy of each of the non-occlusion handling methods (Alotaibi & Mahmood, 2017; Han & Bhanu, 2006; Roy et al., 2012; Shiraga et al., 2016; Zhang et al., 2010) is quite low for occluded test sequences since these methods are designed to work well only if at least a complete gait cycle is available.

4.7. Ablation study

The *CVAE* component of the proposed *BGaitR-Net* reconstruction model is responsible for encoding the input frames of a gait sequence with the help of the conditional key pose vector c and decoding the predicted frames. This model is trained with a binary cross-entropy-based reconstruction loss L_{rec} (refer to (4)) and a *KL*-divergence loss L_{kl} (refer to (5)). On the other hand, the *Bi-LSTM* component of the proposed *BGaitR-Net* is trained with MSE loss L_{mse} (refer to (8)), and it is responsible for reconstructing the frames of the sequence by fusing the spatio-temporal information contained in the *CVAE-encoded* frames along with the key pose information. In the ablation study, we study the importance of the individual loss functions and the conditional key pose vector c used during training the $CVAE$. Basically, we eliminate one of the three components among L_{rec}, L_{kl}, c and train the $CVAE$, and next observe the average *Dice score* of reconstruction on the validation set of the *CASIA-B* data. Corresponding results are presented in Table 6.

In the table, the first row corresponds to the average *Dice score* obtained by eliminating the component c and training the $CVAE$ with the complete loss function given in (6). The second and third rows correspond to results obtained by retaining c but by eliminating the components L_{kl} and L_{rec}, respectively. Finally, the fourth row corresponds to the average *Dice score* on the validation set using the proposed model where each of the above components is retained during the training phase. The results presented in the table indicate that the combined loss term L_{cvae} given by (6) along with the conditional key pose vector c help in obtaining reconstructed frames of the highest quality than each of the other configurations used in the study. Without using the vector c, an average *Dice score* of only *0.749* is obtained, whereas the use of the conditional vector c improves the average *Dice score* by *0.233*. Also, the use of the combined loss term L_{cvae} results in better *Dice score* values than either of the two individual loss terms L_{kl} and L_{rec}.

5. Conclusions and future work

In this work, we propose a new neural architecture for gait sequence reconstruction in the presence of occlusion termed *BGaitR-Net*. The model is based on the stacking of two Deep Neural Network architectures, namely a *Convolutional Variational Autoencoder (CVAE)* and a *Bidirectional Long-Short Term Memory (Bi-LSTM)*. The main novelty of the work is that during the encoding and decoding phases, our model makes use of a conditional one-hot encoded key pose vector for each frame as an auxiliary input which provides useful information about the probable structure of a silhouette in the frame to be reconstructed, and thus guides the bi-LSTM, i.e., the core reconstruction model, to make a good prediction. The impact of employing this

Table 6
Ablation study to observe the effect of the individual loss terms and the conditional key pose vector while training the *CVAE* component of the proposed *BGaitR-Net*.

Model components	Avg. *Dice score*
L_{cvae} without conditional vector c (i.e., removing c)	0.749
L_{rec} with conditional vector c (i.e., removing L_{kl})	0.977
L_{kl} with conditional vector c (i.e., removing L_{rec})	0.283
L_{cvae} with conditional vector c (Proposed)	**0.982**

auxiliary key pose information can be verified from the ablation study results presented in Table 6, where it can be seen that the use of the conditional key pose information while training the *BGaitR-Net* through loss function L_{cvae} improves the Dice score by about 0.23. To the best of our knowledge, the idea of fusing key pose information to reconstruct occluded frames is new and this idea can also be applied to reconstruct frames in occluded videos consisting of activities other than walking.

It may be noted that very few occluded gait datasets are available in the public domain. A few important occluded gait datasets include the GREW and the TUM-IITKGP, which have been used for evaluation in this work. Further, for more extensive evaluation purposes, we extract test sequences from popular unoccluded gait datasets, namely CASIA-B, OU_ISIR LP, OU-ISIR MVLP datasets, and introduce varying levels of synthetic occlusion in these sequences so that the reconstruction performance of our proposed *BGaitR-Net* can be tested for varying levels of synthetic occlusion. The synthetic occlusion is applied to the binary silhouette frames to remove partial or full-body information and generate silhouette frames resembling that obtained from the background subtraction of real-occluded frames, and evaluation of the model using the synthetically occluded data helps in getting a good understanding of how the model would perform in real-occluded scenarios. We have also shared the synthetically generated occluded CASIA-B data used in our work for further comparative research studies.

Experimental results on synthetically occluded sequences from the CASIA-B data show that our model performs reconstruction quite well with Dice scores of 0.98 and 0.82 respectively if 10%–20% frames and 60%–70% frames in a gait cycle are missing/occluded. We have also observed that our reconstruction model combined with either *GEINet* or *GaitGL* always performs gait recognition with a high average accuracy (>74%) for view angles ranging from 36° to 144°. Beyond this range, the recognition accuracy tends to reduce since the binary silhouette sequences from these angles do not possess enough kinematic information for deriving effective Computer Vision-based gait features. Also, our combined *BGaitR-Net*-based reconstruction and *GEINet* or *GaitGL*-based recognition framework has been seen to outperform existing techniques to occlusion handling in gait recognition and recent video frame prediction methods both in terms of *Dice Score* and *Recognition Accuracy* on synthetically occluded CASIA-B and real-occluded TUM-IITKGP datasets. In general, the accuracy values for *GaitGL* are higher since it focuses on extracting local and global features and is a more advanced model compared to *GEINet*. The experimental results also show that the recognition accuracy for the *OU-ISIR MVLP* dataset with 40% synthetic occlusion is quite high, surpassing 70% for view angles ranging from 45° to 90°. Moreover, the average recognition accuracy using the *GaitGL* classification model for the GREW dataset with real occlusion is 66% without *BGaitR-Net*-based reconstruction and 74% with it. Therefore, despite the dataset's complicated nature, there is an average improvement of 12.12% in the recognition accuracy after performing the reconstruction, demonstrating the exemplary reconstruction capability of *BGaitR-Net*. In the future, our work can be conveniently integrated with a multi-object tracking framework to carry out gait recognition of multiple persons simultaneously.

CRediT authorship contribution statement

Somnath Sendhil Kumar: Conceptualization, Methodology, Implementation, Writing – original draft. **Binit Singh:** Methodology, Comparative Study, Implementation, Writing – original draft. **Pratik Chattopadhyay:** Conceptualization, Supervision, Writing – review & editing. **Agrya Halder:** Conceptualization, Comparative Study, Implementation, Proof reading. **Lipo Wang:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank NVIDIA for supporting their research with a Titan XP GPU and SERB, DST, Govt. of India for partially supporting this work through project grant CRG/2020/005465.

References

Alotaibi, M., & Mahmood, A. (2017). Improved gait recognition based on specialized deep convolutional neural network. *Computer Vision and Image Understanding*, *164*, 103–110.

Babaee, M., Li, L., & Rigoll, G. (2018). Gait recognition from incomplete gait cycle. In *Prceedings of the 25th international conference on image processing* (pp. 768–772).

Babaee, M., Li, L., & Rigoll, G. (2019). Person identification from partial gait cycle using fully convolutional neural networks. *Neurocomputing*, *338*, 116–125.

Battistone, F., & Petrosino, A. (2019). TGLSTM: A time based graph deep learning approach to gait recognition. *Pattern Recognition Letters*, *126*, 132–138.

Carass, A., Roy, S., Gherman, A., Reinhold, J. C., Jesson, A., Arbel, T., et al. (2020). Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis. *Scientific Reports*, *10*(1), 1–19.

Chang, Z., Zhang, X., Wang, S., Ma, S., Ye, Y., Xiang, X., et al. (2021). MAU: A motion-aware unit for video prediction and beyond. In A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), *Proceedings of the advances in neural information processing systems*.

Chao, H., He, Y., Zhang, J., & Feng, J. (2019). Gaitset: Regarding gait as a set for cross-view gait recognition. *Vol. 33*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8126–8133).

Chao, H., Wang, K., He, Y., Zhang, J., & Feng, J. (2021). GaitSet: Cross-view gait recognition through utilizing gait as a deep set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chattopadhyay, P., Roy, A., Sural, S., & Mukhopadhyay, J. (2014). Pose depth volume extraction from RGB-D streams for frontal gait recognition. *Journal of Visual Communication and Image Representation*, *25*(1), 53–63.

Chattopadhyay, P., Sural, S., & Mukherjee, J. (2015). Frontal gait recognition from occluded scenes. *Pattern Recognition Letters*, *63*, 9–15.

Chen, C., Liang, J., Zhao, H., Hu, H., & Tian, J. (2009). Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, *30*(11), 977–984.

Collins, R. T., Gross, R., & Shi, J. (2002). Silhouette-based human identification from body shape and gait. In *Proceedings of 5th international conference on automatic face gesture recognition* (pp. 366–371).

Dargan, S., & Kumar, M. (2020). A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Systems with Applications*, *143*, Article 113114.

Das, D., Agarwal, A., Chattopadhyay, P., & Wang, L. (2019). Rgait-NET: An effective network for recovering missing information from occluded gait cycles. arXiv preprint arXiv:1912.06765.

Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., et al. (2020). Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14225–14233).

Gao, Z., Tan, C., Wu, L., & Li, S. Z. (2022). SimVP: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3170–3180).

Ghosh, R. (2022). A faster R-CNN and recurrent neural network based approach of gait recognition with and without carried objects. *Expert Systems with Applications*, Article 117730.

Guan, Y., Li, C.-T., & Hu, Y. (2012). An adaptive system for gait recognition in multi-view environments. In *Proceedings of the 14th ACM multimedia and security workshop* (pp. 139–144).

Guen, V. L., & Thome, N. (2020). Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11474–11484).

Gul, S., Malik, M. I., Khan, G. M., & Shafait, F. (2021). Multi-view gait recognition system using spatio-temporal features and deep learning. *Expert Systems with Applications*, *179*, Article 115057.

Gupta, S. K., & Chattopadhyay, P. (2021a). Exploiting pose dynamics for human recognition from their gait signatures. *Multimedia Tools and Applications*, *80*(28), 35903–35921.

Gupta, S. K., & Chattopadhyay, P. (2021b). Gait recognition in the presence of co-variate conditions. *Neurocomputing*, *454*, 76–87.

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

Han, F., Li, X., Zhao, J., & Shen, F. (2022). A unified perspective of classification-based loss and distance-based loss for cross-view gait recognition. *Pattern Recognition*, Article 108519.

He, Y., Zhang, J., Shan, H., & Wang, L. (2018). Multi-task GANs for view-specific feature learning in gait recognition. *IEEE Transactions on Information Forensics and Security*, *14*(1), 102–113.

Hofmann, M., Sural, S., & Rigoll, G. (2011). Gait recognition in the presence of occlusion: A new dataset and baseline algorithms. In *Proceedings of the 19th international conference in central europe on computer graphics, visualization and computer vision* (pp. 99–104).

Hu, H., Li, Y., Zhu, Z., & Zhou, G. (2018). CNNAuth: Continuous authentication via two-stream convolutional neural networks. In *Proceedings of the international conference on networking, architecture and storage* (pp. 1–9).

Isa, W. N. M., Alam, M. J., & Eswaran, C. (2010). Gait recognition using occluded data. In *Proceedings of the asia pacific conference on circuits and systems* (pp. 344–347).

Iwama, H., Okumura, M., Makihara, Y., & Yagi, Y. (2012). The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, *7*(5), 1511–1521.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *Stat*, *1050*, 10.

Lee, T. K. M., Belkhatir, M., & Sanei, S. (2009). Coping with full occlusion in fronto-normal gait by using missing data theory. In *Proceedings of the 7th international conference on information, communications and signal processing* (pp. 1–5).

Li, Y., Hu, H., Zhu, Z., & Zhou, G. (2020). SCANet: sensor-based continuous authentication with two-stream convolutional neural networks. *ACM Transactions on Sensor Networks*, *16*, 1–27.

Lin, B., Zhang, S., Wang, M., Li, L., & Yu, X. (2022). GaitGL: Learning discriminative global-local feature representations for gait recognition. arXiv preprint arXiv:2208.01380.

Maleki, S., Maleki, S., & Jennings, N. R. (2021). Unsupervised anomaly detection with LSTM autoencoders using statistical data-filtering. *Applied Soft Computing*, *108*, Article 107443.

Mejia, J., Avelar-Sosa, L., Mederos, B., Ramírez, E. S., & Roman, J. D. D. (2021). Prediction of time series using an analysis filter bank of LSTM units. *Computers & Industrial Engineering*, *157*, Article 107371.

Roy, A., Sural, S., & Mukherjee, J. (2012). Gait recognition using pose kinematics and pose energy image. *Signal Processing*, *92*(3).

Roy, A., Sural, S., Mukherjee, J., & Rigoll, G. (2011). Occlusion detection and gait silhouette reconstruction from degraded scenes. *Signal, Image, and Video Processing*, *5*(4), 415–430.

Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., & Yagi, Y. (2016). GEINet: View-invariant gait recognition using a convolutional neural network. In *Proceedings of the international conference on biometrics* (pp. 1–8).

Sivapalan, S., Chen, D., Denman, S., Sridharan, S., & Fookes, C. (2011). Gait energy volumes and frontal gait recognition using depth images. In *Proceedings of the international joint conference on biometrics* (pp. 1–6).

Song, X., Huang, Y., Shan, C., Wang, J., & Chen, Y. (2022). Distilled light GaitSet: Towards scalable gait recognition. *Pattern Recognition Letters*.

Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., & Yagi, Y. (2017). On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(9), 2708–2719.

Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., & Yagi, Y. (2018). Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, *10*(4), 1–14.

Wang, Y., Jiang, L., Yang, M.-H., Li, L.-J., Long, M., & Fei-Fei, L. (2019). Eidetic 3DLSTM: A model for video prediction and beyond. In *Proceedings of the international conference on learning representations*.

Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Yu, P., et al. (2022). PredRNN: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, D., Yan, S., Tao, D., Lin, S., & Zhang, H.-J. (2007). Marginal Fisher analysis and its variants for human gait recognition and content- based image retrieval. *IEEE Transactions on Image Processing*, *16*(11), 2811–2821.

Yu, S., Chen, H., Garcia Reyes, E. B., & Poh, N. (2017). GaitGAN: Invariant gait feature extraction using generative adversarial networks. In *Proceedings of the conference on computer vision and pattern recognition workshops* (pp. 30–37).

Yu, S., Tan, D., & Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proceedings of the international conference on pattern recognition* (pp. 441–444).

Zhang, P., Wu, Q., & Xu, J. (2019). VT-GAN: View transformation GAN for gait recognition across views. In *Proceedings of the international joint conference on neural networks* (pp. 1–8).

Zhang, E., Zhao, Y., & Xiong, W. (2010). Active energy image plus 2DLPP for gait recognition. *Signal Processing*, *90*(7), 2295–2302.

Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., et al. (2021). Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE international conference on computer vision* (pp. 14789–14799).

Somnath Sendhil Kumar is pursuing B.Tech. from the Department of Electrical Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi, India, PIN 221005.

Binit Singh is pursuing M.Tech. from the Department of Computer Science and Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi, India, PIN 221005.

Pratik Chattopadhyay is an Assistant Professor in the Department of Computer Science and Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi, India, PIN 221005. He is also leading the Pattern Recognition Laboratory of the department.

Agrya Halder is pursuing Ph.D. from the Department of Computer Science and Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi, India, PIN 221005.

Lipo Wang, is an Associate Professor in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore PIN 639798.