Neural Computing and Applications manuscript No.
(will be inserted by the editor)

BGaitR-Net: A Neural Model for Reconstruction of Occluded Frames in a Gait Sequence by Exploiting Spatio-temporal Information

Somnath Sendhil Kumar[1] · Binit Singh[2] · Pratik Chattopadhyay[2] · Sanjay Kumar Gupta[2]

Received: date / Accepted: date

Abstract Gait recognition in the presence of occlusion is a challenging problem in real-life surveillance applications and the solutions proposed by researchers to date lack robustness and also dependent on several unrealistic constraints. We improve the state-of-the-art by developing a Deep Learning-based framework to reconstruct the occluded frames in a gait sequence by exploiting the spatio-temporal information present in adjacent frames as well as the key pose information corresponding to each frame of the sequence. Our multi-stage pipeline consists of key pose mapping, occlusion detection, and reconstruction, and finally gait recognition phases. While the key pose mapping and occlusion detection are done using existing algorithms, we propose a new model, namely, Bidirectional Gait Reconstruction Network for occlusion reconstruction by stacking a Conditional Variational Autoencoder with a Bi-Directional Long Short Time Memory. The sub-networks involved in the occlusion reconstruction model are trained using extensive synthetically occluded datasets constructed from the *CASIA-B* and *OU-ISIR LP* data. Experimental results show that our proposed model reconstructs occlusion effectively and generates frames that are temporally consistent with the periodic pattern of gait, while simultaneously preserving information about the silhouette structure of the target subject. Finally, *GEINet* feature-based classification is employed to identify the class of the subject from the reconstructed sequence. The effectiveness of our approach is evaluated using the real-occluded *TUM-IITKGP* and synthetically occluded *CASIA-B* data sets and encouraging results have been obtained. Comparative analysis with other popular occlusion handling methods in gait recognition also shows the superiority of our approach over these techniques.

Keywords Key Poses · Occlusion Reconstruction · Spatio-Temporal Model · Gait Recognition

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. 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 presence of occlusion in the scene, camera viewpoint variation, clothing changes of subjects, etc. Over the past two decades, there have been several attempts to tackle situations where the viewpoint and co-variate conditions of subjects are different in the training and test sequences, e.g., [1–3]. However, significant focus has not been given to solve the challenging problem of gait recognition in the presence of occlusion. Only a few methods [4–7] 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 approach discussed in [4] and [7] are non-Deep

· S.S. Kumar (*somnath.sendhilk.eee19@itbhu.ac.in*)
· B. Singh (*binitsingh.cse21@iitbhu.ac.in*)
· P. Chattopadhyay* (*pratik.cse@iitbhu.ac.in*)
· S.K. Gupta (*sanjaykrgupta.rs.cse17@iitbhu.ac.in*)
[1]Department of Electrical Engineering, Indian Institute of Technology (BHU), Varanasi-221005, India
[2]Pattern Recognition Laboratory, Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi-221005, India
Corresponding Author. Ph. +91-5427165311

Learning-based. While the method in [4] works by comparing the available walking poses in the training and test sequences, that in [7] uses a Gaussian Process Dynamical Model to predict the missing/occluded frames in a gait cycle and next extract features from this reconstructed cycle to perform recognition. The method in [4] fails if no matching walking poses are found in a pair of gallery and test sequences due to heavy occlusion, whereas the assumption in [7] that walking features follow a Gaussian is not an established fact and the fitted Gaussian may not be able to make good prediction about the missing frames if the input sequence is corrupted with moderate to heavy degrees of occlusion. In contrast to these two methods, that in [5] and [6] 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.

In this work, we focus on improving upon the existing solutions to occlusion handling in gait recognition and propose a new model termed as the *Bi-directional Gait Reconstruction Network* (abbreviated as *BGaitR-Net*) to reconstruct the occluded frames present in a gait sequence by exploiting the spatio-temporal information from the input sequence and the auxiliary key pose information about the gait of human being [7]. The proposed model is formed by stacking a Conditional Variational Autoencoder (*CVAE*) with a Bi-directional Long Short Term Memory (*Bi-LSTM*) and these sub-networks are trained using extensive gallery sets constructed from the *CASIA-B* data [8] and the *OU-ISIR* Large Population Dataset [9]. Performance evaluation of the proposed model has been done using both the real-occluded sequences present in the *TUM-IITKGP* [10] data and the synthetically occluded sequences generated from the *CASIA-B* data. The results in terms of reconstruction *Dice score* and gait recognition accuracy show that our model is effective enough to reconstruct sequences corrupted with moderate to heavy degrees of occlusion. The main contributions of our work are summarized as follows:

- We propose a new neural model *BGaitR-Net* to effectively reconstruct the occluded frames present in a gait sequence in a temporally consistent manner by fusing the spatio-temporal information available from the sequence and the key pose information for each frame of the sequence. Good quality reconstruction results are obtained even if 60-70% frames in a gait cycle are occluded.
- An extensive data set of synthetically occluded sequences along with the ground truth unoccluded sequences constructed form the *CASIA-B* and the *OU-ISIR LP* data form the gallery set to train the two sub-networks of our reconstruction model *BGaitR-Net* effectively, namely the *CVAE* and the *Bi-LSTM*. Suitable loss functions have been used to train both the sub-networks and an ablation study has been done to observe the effect of the different components of the proposed *BGaitR-Net* model.
- Comparative study shows that our method is superior to other occlusion handling methods in gait recognition both in terms of reconstruction quality and gait recognition accuracy.
- The resources including the synthetically occluded data and pre-trained models will be made publicly available to the research community for further comparative studies.

2 Related Work

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. The work in [11] 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 [12] 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 [13] and [4] 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 [14] 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 [15] 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 [11, 12, 14]. In [16], the *GEI* features are first projected into a lower-dimensional space using Marginal Fisher Analysis, and recognition is done using the subspace features. A viewpoint invariant gait recognition approach described in [1] 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.

With the introduction of *RGB-D* cameras such as Kinect, a few frontal-view gait recognition techniques [13, 17] 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 [18] 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.

With the advancement of Deep Learning, *CNN*-based models have also been extensively used for gait recognition. For example, in [19] and [20], 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, for example, the work in [21] 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 [22] 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 [23] 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.

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 [24], 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 [12, 13], 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 [25] 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 [11], 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 [2], 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 [3], 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 [26, 27] 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 [28] that 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 [29]. However, none of these approaches are suitable for application if occlusion is present in the gait sequences. In [30], 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.

Most of the gait recognition scenarios used in the abovementioned techniques consider a single person to be present in the field of view of a camera, and also assume that at 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 towards handling the problem of occlusion in gait recognition are discussed next. Occlusion reconstruction has been done using a Gaussian process dynamic model in [7]. 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 [10]. In [31], 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 [32], 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. Moreover, the effectiveness of Deep Neural Network-based models to predict the missing/occluded frames in a gait sequence has not been studied yet. In this work, we specifically focus on this aspect and propose a new spatio-temporal model to reconstruct the occluded frames in a gait sequence. The proposed network architecture along with the training details are explained in the following section.

3 Proposed approach

A schematic diagram explaining the steps of the proposed occlusion reconstruction approach through *BGaitR-Net* is shown in Fig. 1. With reference to the figure, standard pre-processing steps [11, 14, 22] are first applied to extract the binary silhouettes from the RGB frames and normalize these extracted frames. This step involves background subtraction using a suitable technique, cropping out the region of interest, and resizing each cropped region to a fixed height and width. These preprocessed silhouettes are next used to esti-

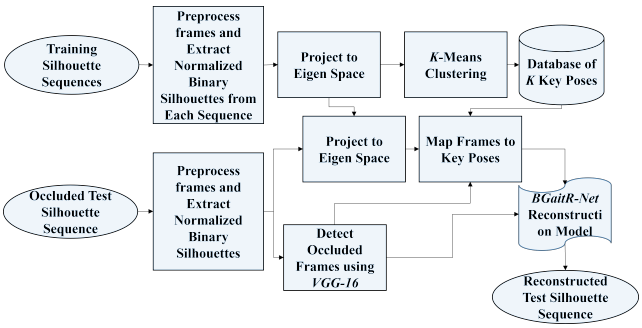

Fig. 1: A block diagram showing the pipeline of the proposed reconstruction algorithm

mate a set of key poses using a technique similar to that described in [12]. Next, given a test silhouette sequence, we carry out similar preprocessing steps to generate the normalized silhouettes, and use a *VGG-16* model [33] to predict the occluded frames followed by a graph sorting algorithm to map the unoccluded PCA-reduced frames to the appropriate key poses [7]. Finally, a sequence of preprocessed frames (which may be either occluded or un-occluded) are fed to the proposed *BGaitR-Net* occlusion reconstruction model along with a auxiliary conditional key pose vector corresponding to each frame to obtain a refined prediction about the input frames. It may be noted that, we have used existing algorithms for the background subtraction, *VGG-16*-based occlusion detection, key pose construction, and frame mapping to key poses. Hence, instead of elaborately discussing these algorithms, we provide an overview of the related approaches used in this work with proper citations. A more focus is given to explaining the steps of our proposed occlusion reconstruction algorithm including the architectural details and training of the individual sub-networks of the *BGaitR-Net* model.

3.1 Occlusion Reconstruction in a Gait Sequence

Given an occluded test sequence, standard preprocessing steps such as silhouette cropping and normalization as in [11, 14, 22] are employed to obtain clean and normalized binary silhouettes corresponding to each frame. Six frames from a sample occluded binary silhouette sequence along with the corresponding normalized frames are shown in the first and second rows of Fig. 2. While for unoccluded frames (i.e., the first and sixth frames) clean binary silhouette frames are obtained, for occluded frames (i.e., second to fifth frames) the silhouette shapes obtained are quite irregular and do not resemble human structure as can be observed from Fig. 2.

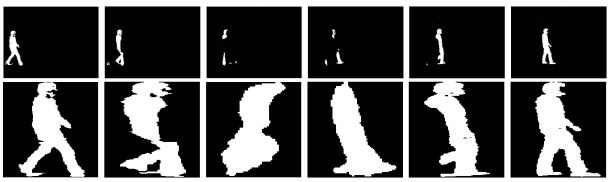

Fig. 2: The first row shows background-subtracted frames and the second row shows cropped and normalized binary silhouettes corresponding to a few occluded and unoccluded frames in a sequence from the *TUM-IITKGP* data [10]

A *VGG-16* model [33] is used to automatically identify the occluded and the unoccluded binary silhouette frames present in any normalized binary silhouette sequence. This model takes as input a normalized binary frame and classifies it as either '*Occluded*' or '*Unoccluded*'. This model is quite effective for occlusion detection and performs with a precision and recall of 99.53% and 98.72%, respectively and an overall accuracy of 98.89% on validation data.

3.1.1 Determination of Key Poses in a Gait Cycle and Mapping of Frames to the Appropriate Key Poses

Next, we determine a set of generic key poses in a gait cycle using an algorithm similar to that given [12] and map the unoccluded frames in each sequence (as predicted by the *VGG-16* model) to the appropriate key poses. A set of 50 different gait cycles extracted from the *CASIA-B* [8] and *OU-ISIR Large Population (LP)* [9] datasets have been used to compute these key poses. Similar to the work in [12], we apply constrained K-Means clustering with K=16 on these 50 gait cycles to determine the key poses and these are shown in Fig. 3. As can be seen from the figure, the set of key poses

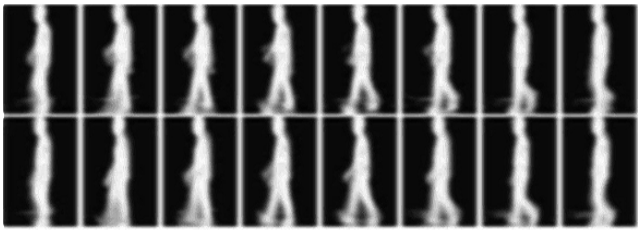

Fig. 3: 16 Key poses computed from a set of gait cycles from *CASIA-B* data and *OU-ISIR* Large Population data

preserves the temporal order of general human walking, and are not specific to any individual person. Once the generic key poses are identified, the appropriate key poses numbers for each unoccluded frame in the sequence are obtained following a frame to key pose mapping algorithm similar to that given in [7]. This algorithm classifies each unoccluded frame of an input sequence to the appropriate key pose by maintaining the temporal order of walking and also constraining the unoccluded frames to not get mapped into any key pose. Fig. 4 shows a binary silhouette sequence of 27 frames with both partial and full-body occlusions generated from the *CASIA-B* data, and the corresponding state to which each frame gets mapped to. In this figure, the symbol S_i (for i = 1, 2, ..., 16) indicates that the corresponding frame has got mapped to the i^{th} 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. 3. A detailed discussion on the key pose construction and frame to key pose mapping algorithms can be found in [7, 12].

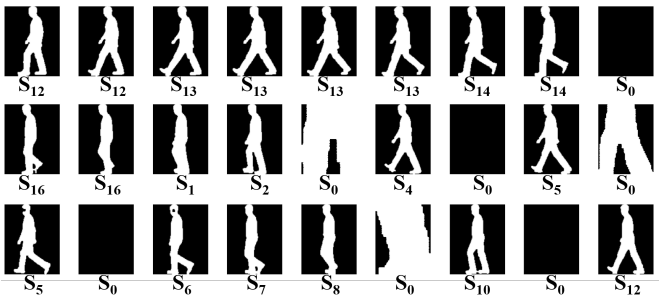

Fig. 4: An occluded frame sequence and the mapped states corresponding to each frame

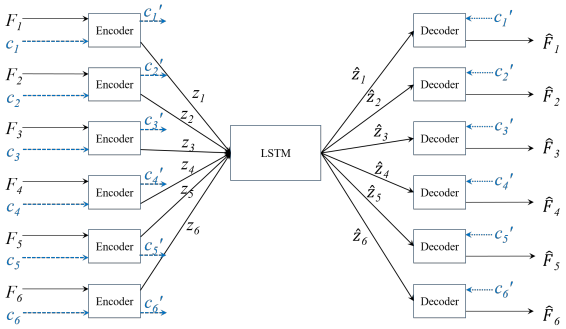

Fig. 5: An overview of the *BGaitR-Net* Model used for occlusion reconstruction

3.1.2 Architecture and Training Details of BGaitR-Net

Next, we describe in detail the architecture of *BGaitR-Net* along with the loss functions used to train the individual sub-networks present in the model. As shown in Fig. 5, the *BGaitR-Net* consists of an *Encoder-Decoder* architecture with a *Bidirectional Long-Short Term Memory* sub-network in between. Initially, an encoded vector E_i corresponding to each frame F_i is computed using a *Conditional Variational Autoencoder* that takes as input the normalized frame and a vector c obtained from one-hot encoding of the corresponding key pose number. This vector is 17-dimensional out of 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 the frame is occluded, this 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*. Six encoded vectors denoted by E_1, E_2, ...,E_6 corresponding to six frames of an input sequence, 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. Since a binary silhouette sequence corresponding to human walking can be viewed as a spatio-temporal pattern in which silhouette images form a periodic progression, the binary silhouette frame at a particular instant of time can be said to be temporally related with its neighboring frames in the sequence. Hence, the silhouette information corresponding to any occluded frame can be predicted by exploiting the spatio-temporal information contained in the neighboring frames. These reconstructed vectors are next passed through a *Decoder* network to obtain the reconstructed frames, namely, $\hat{F}_1$, $\hat{F}_2$, ...,$\hat{F}_6$. Training of the *BGaitR-Net* is done by training the two sub-networks, namely, the *Autoencoder* (i.e., the *Encoder-Decoder* architecture) and the *Bi-LSTM* separately on extensive data sets. In our work, an *Encoder-Decoder* architecture is trained to obtain the encoded representation for each image frame and convert the vector back to the image. However, for ease of explanation, in Fig. 5, six different encoders and decoders have been shown, one for each image frame. During testing, occlusion reconstruction is done by following a three-step process, shown in the figure. First, vector embedding of six consecutive frames are computed by the *Encoder*, and next each of the frames is reconstructed via the *Bi-LSTM* in the encoded space. Finally, these reconstructed vectors are decoded back to the image space using the *Decoder*. We next discuss the architecture and training details for each of the two sub-networks, namely, the *Encoder-Decoder* and the *Bi-LSTM*.

Encoder-Decoder Architecture: A *Conditional Variational Autoencoder* (*CVAE*) has been employed in this work to compute an embedding for each binary silhouette frame present in a sequence. This encoded vector has a reduced dimension compared to the original image and preserves the important characteristics of the silhouette shape at each frame while simultaneously reducing the noise and other redundant information. A detailed architecture of the *Encoder* network used in the *CVAE* is 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 vector c comprising of the key pose-related information is also com-

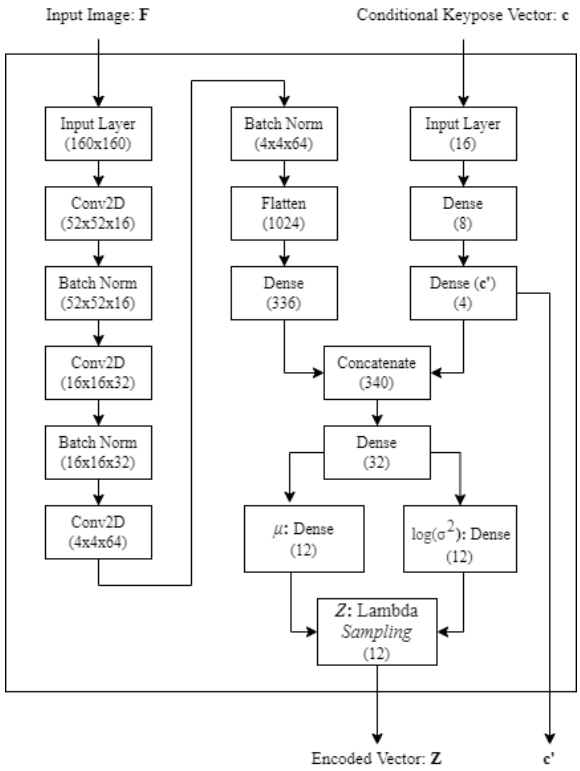

Fig. 6: Architecture of the Encoder of the *Conditional Variational Autoencoder* used for computing an embedding from each binary frame

pressed 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.

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,

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

Training of the *CVAE* is accomplished using the back-propagation algorithm by following a re-parameterization strategy [34]. Both the μ and $log(\sigma)$ vectors are also 12-dimensional, and these are combined with a random error term (ϵ) sampled from a standard normal distribution to generate the output embedded vector Z using the following expression:

$$Z = \mu + \sigma \odot \epsilon, \quad (2)$$

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) and outputs the fully reconstructed image of dimensions 160×160. During training, the *Encoder* and the *Decoder* are trained together by stacking these networks in an end-to-end manner, and hence Z is the 12-dimensional latent vector sampled from the learned normal distribution. On the other hand, during testing Z is the 12-dimensional vector output by the *LSTM*. The 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 4×4×64 dimensional feature map, which is further decoded using three transposed convolutional layers (shown in the figure as Conv2DTranspose) with dropout to obtain a 160×160×8 dimensional feature map. These feature

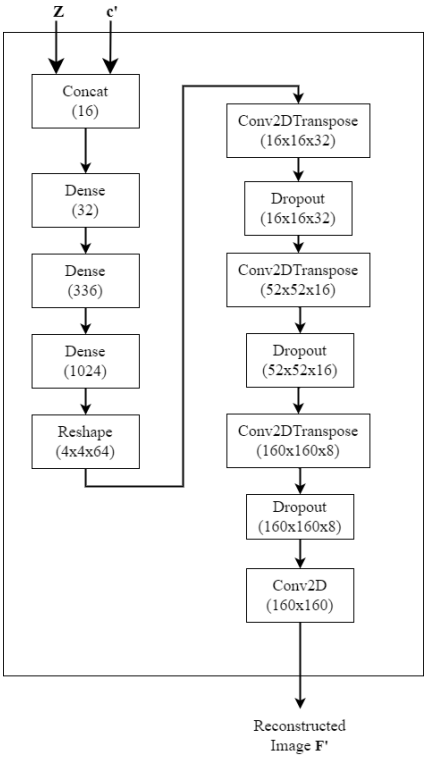

Fig. 7: Architecture of the *Decoder* of the *Conditional Variational Autoencoder* used for reconstructing an image from *LSTM*-predicted vector and conditional key pose vector

maps are next combined into a 160×160 dimensional feature map in the final convolutional layer, which is also the desired output reconstructed image. 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 Eqn. (4), is defined as the binary cross-entropy loss between the input and the reconstructed silhouettes. Mathematically,

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

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):

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

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 k^{th} image, k = 1, 2, ..., $\mathcal{M}$. 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.

LSTM-based Sequence Reconstruction: A deep time-series-to-time-series neural network, specifically a *Bi-directional Long-Short Term Memory* (*Bi-LSTM*) network [35] is next employed to reconstruct the frames of an occluded gait sequence in the encoded space. Since the gait of any person follows a temporal progression, and *Bi-LSTMs* are popularly used for time-series data filtering [36, 37], it appears that the embedding corresponding to the six binary image frames output by the *Encoder* network can be effectively filtered using the *Bi-LSTM* and these filtered vectors will preserve/reconstruct information relevant to the silhouette shape in the encoded space by eliminating the unwanted noise, occlusion, etc. Reconstruction of the filtered embedded vector through the *Decoder* network will provide us with the desired reconstructed image. A schematic diagram of the *Bi-LSTM* architecture used in this work is shown in Figure 8. As shown in the figure, this network 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 as T, then

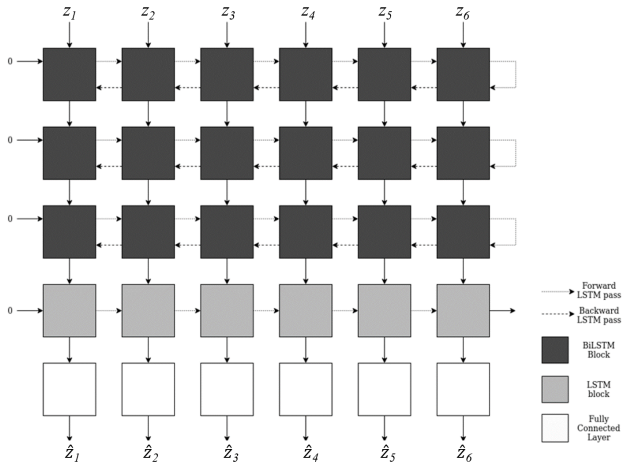

Fig. 8: Architecture of the *Bi-LSTM* model used for reconstruction

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

The model is trained using Mean Squared Error loss L_{mse} between the original latent vectors z_is and the predicted latent vectors $\hat{z}_i$, as given by (8) using an extensive gallery set constructed from synthetic occluded sequences of six frames and the corresponding ground-truth sequences.

$$L_{mse} = \frac{1}{n}\sum^{n}\sum_{i=1}^{6}(z_i - \hat{z}_i)^2\,. \tag{8}$$

3.2 *GEINet*-based Gait Recognition

For carrying out human identification from their gait signatures, first, we need to construct the database of training gait features. For this, we consider the unoccluded gallery sequences of pre-processed and normalized silhouettes for gait recognition, extract a complete cycle, and compute the *GEI* features [11] from the extracted gait cycle. Since a gait sequence typically consists of multiple gait cycles, multiple *GEI* features can

be constructed from a single sequence. Suppose, the dataset consists of $\mathcal{N}$ subjects denoted by $\mathcal{S}_1$, $\mathcal{S}_2$, .. $\mathcal{S}_\mathcal{N}$, and GC_i number of gait cycles are obtained for subject i, i= 1,2, ..., $\mathcal{N}$. Further, let the *GEI* feature for the j^{th} gait cycle of person i is denoted by $\mathcal{F}_i^j$, j, i= 1,2, ..., GC_i. The database of training gait features and corresponding labels is next constructed from the above features. If $\mathcal{D}$ denotes this training database, then $\mathcal{D}$ consists of the following labeled patterns:

Features: $\{\{\mathcal{F}_1^1\ \mathcal{F}_1^2\ ...\ \mathcal{F}_1^{GC_1}\}, \{\mathcal{F}_2^1\ \mathcal{F}_2^2\ ...\ \mathcal{F}_2^{GC_2}\}, ..., \{\mathcal{F}_\mathcal{N}^1\ \mathcal{F}_\mathcal{N}^2\ ...\ \mathcal{F}_\mathcal{N}^{GC_\mathcal{N}}\}\}$

Labels: $\{\{\mathcal{S}_1\ \mathcal{S}_1\ ...\ \mathcal{S}_1\}, \{\mathcal{S}_2\ \mathcal{S}_2\ ...\ \mathcal{S}_2\}, ..., \{\mathcal{S}_\mathcal{N}\ \mathcal{S}_\mathcal{N}\ ...\ \mathcal{S}_\mathcal{N}\}\}$.

The gallery set $\mathcal{D}$ is next used to train a *GEINet* model [22], and the same architecture of the *GEINet* as in [22] has also been used in this work. Specifically, this model has two convolutional layers with *18* and *45* kernels with max-pooling and *ReLU* activation at each layer, and a softmax classification layer. It is trained with Multi-Class-Cross-Entropy loss with Adam optimizer till convergence.

Given a test binary silhouette sequence with occluded frames, we first carry out the VGG-16-based occlusion detection, followed by determination of the key pose number for each unoccluded frame of the sequence (as explained in Section 3.1.1). The key pose number for each frame of the sequence is next converted into a one-hot encoded vector and fed to the trained *Encoder* network along with the corresponding binary frame to obtain the desired frame embedding (refer to Section 3.1.2). Each set of embedding vectors from six consecutive frames of a sequence is passed through the trained *Bi-LSTM* that outputs six corresponding refined encoded vectors. Finally, each of these vectors is passed one-by-one through the trained *Decoder* network to obtain the corresponding reconstructed image. Once the frames of a complete gait cycle of the test subject are reconstructed following the above procedure, the *GEI* features are computed and passed through the trained *GEINet* (refer to Section 3.2) to identify the class of the test subject. The above discussion assumes that the gallery sequences for gait recognition are unoccluded. In case the gallery sequences are occluded, a similar *BGaitR-Net*-based occlusion reconstruction procedure must be followed to reconstruct these sequences as well before training the *GEINet* model.

4 Experimental Setup

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 2.1GHz and GeForce GTX 1650 Ti with 4 GB RAM, 4 GB frame-buffer memory, and 128 MB BAR1 memory.

4.1 Description of the Data Sets Used in the Study

Three different gait data sets have been used in the study for training the *BGaitR-Net* or the *GEINet*, namely the *CASIA-B* [8], the *TUM-IITKGP* [10], and the *OU-ISIR Large Population* (*LP*) Data [9]. 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 data sets are briefly explained next.

The *CASIA-B* [8] data consists of walking sequences of 124 subjects under three different settings: (a) six sequences with normal walking (nm-01 to nm-06), (b) two sequences with carrying bag (bg-01 and bg-02), (c) two sequences with wearing a coat (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). Out of these, we use the sequences labeled nm-01, nm-02, nm-03, and nm-04 for training the *GEINet* model, and use the remaining two for testing after corrupting the frames in these sequences with varying levels of synthetic occlusion. On the other hand, the *OU-ISIR LP* data set [9] 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 frames in a binary silhouette sequence to be occluded and randomly select these number of 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. Next, each occluded and unoccluded frame of the sequence is preprocessed and normalized using standard techniques [11, 14, 22] before carrying out the reconstruction steps given by the block diagram of Fig. 1.

The *TUM-IITKGP* data [10] 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 present in the *TUM-IITKGP* data set consist of a large number of frames from which we segment out eight different gait cycles corresponding to each of the

35 subjects to form the gallery set for training the gait recognition model. Similarly, we segment out four sequences from each of the static and dynamic occluded videos corresponding to each subject to construct eight separate occluded test sets for evaluating the proposed gait recognition framework and also for making a comparative study with other approaches. These eight occluded test sets corresponding to the *TUM-IITKGP* data are labeled as *Set1*, *Set2*, ..., *Set8*, respectively.

4.2 Training and Evaluation of the *BGaitR-Net*-based Occlusion Reconstruction Model

The *BGaitR-Net* model is trained using the *OU-ISIR* and the *CASIA-B* data sets with synthetic occlusions introduced in the frames of the sequences following a procedure similar to that described in Section 4.1. As explained before, the two sub-networks of the *BGaitR-Net*, i.e., the *CVAE* and the *LSTM*, are trained separately. We consider 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 form the gallery set for training the *CVAE*. Out of these total 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 the Adam optimizer considering a learning rate of 0.01 for 100 epochs at which point the model converges.

The *Bi-LSTM*, on the other hand, is trained on sets of six latent vectors obtained by running 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 four unoccluded normal walking sequences, namely, nm-01, nm-02, nm-03, and nm-04 corresponding to each of the 124 subjects present in the *CASIA-B* data. This results in a total of 496 sequences of encoded vectors, from which we extract different overlapping sub-sequences of six consecutive frames to form the gallery set of 69560 sequences for training the *Bi-LSTM*. To enable the model to predict missing/occluded frames effectively, we synthetically occlude 10-70% frames in each sequence present in this gallery set by adding varying levels of synthetic occlusion using a procedure similar to that explained in Section 4.1. Out of these 69560 sequences, 65000 were used as training sequences and the remaining 4560 were used as validation sequences to evaluate the performance of the *LSTM* on unknown data. The *LSTM* is trained with Adam optimizer for 100 epochs using a learning rate of 0.01 at which point both the training and validation losses appear to converge and the training is stopped.

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. 9. using a synthetically occluded sequence generated from the *CASIA-B* data.

The first row in Fig. 9 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 [38] 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 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.

4.3 Evaluation of the Occlusion Reconstruction and Gait Recognition Framework

To verify the effectiveness of our overall approach, we evaluate the gait recognition accuracy obtained on the *BGaitR-Net*-reconstructed sequences using the trained *GEINet* model (as discussed in Section 3.2). 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. The gallery sets used for training the *GEINet* corresponding to each of these data sets and the test sets considered for the *TUM-IITKGP* data have already been discussed in Section 4.1. We experiment with nine different degrees of synthetic occlusion introduced on the test sequences of the *CASIA-B* data,

Fig. 9: The first row shows sample frames from a synthetically occluded sequence, second row corresponds to the *BGaitR-Net*-predicted frames, and the third row shows the ground-truth

namely, 0-10%, 10-20%, 20-30%, ..., 80-90%. Results are shown in terms of both Reconstruction *Dice Score* and 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

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**	**Rank 1 Acc. (%)**
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

the data set 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 can be seen that the *Dice Score* of recognition is 0.90 or higher if the degree of occlusion is 40% or less. Even for very high 90% degree of synthetic occlusion, the *Dice Score* is 0.75, which is quite impressive. From the fourth column of the table, it can be 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 very high degree of occlusion, i.e., 80-90%, the accuracy is 60.05%. 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. 10(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. 10(a) and 10(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. 10(a)), that achieved at Rank 5 for the reconstructed sets is 100%. It is further seen from Fig. 10(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 emphasizes the usefulness of our *BGaitRNet*-based occlusion reconstruction approach.

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 data sets. The training set corresponding to the *CASIA-B* data has been used

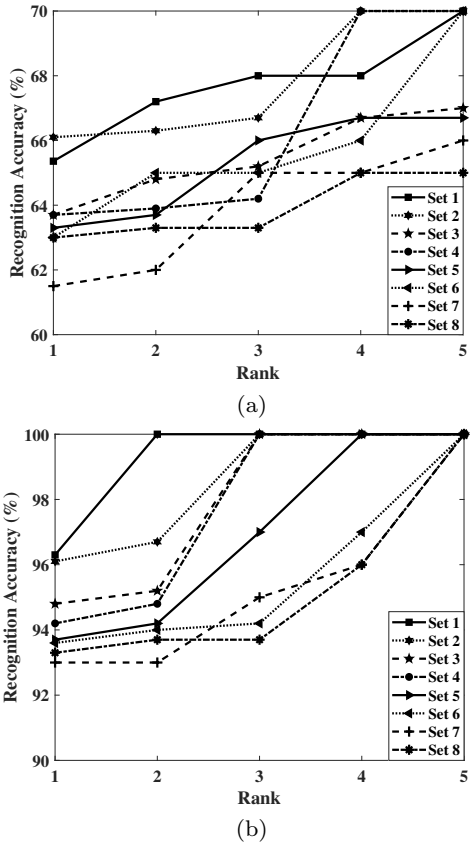

Fig. 10: CMC curves showing rank-wise improvement in recognition accuracy of *GEINet* on (a) the eight occluded test sets present in the *TUM-IITKGP* data and (b) on the same sequences after reconstruction using *BGaitR-Net*

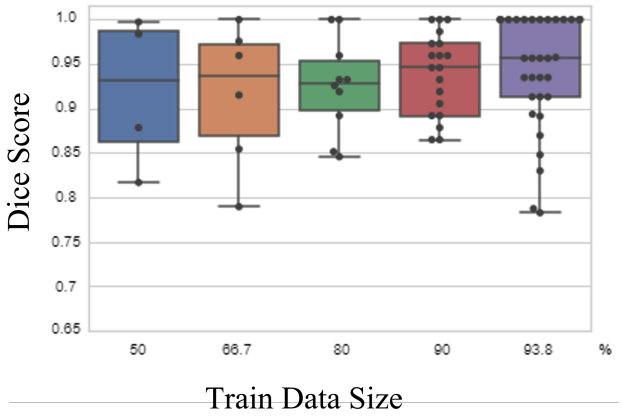

Fig. 11: Average *Dice scores* after training the *Bi-LSTM* model with $\mathcal{K}$-fold cross-validation for the following values of $\mathcal{K}$: *2*, *3*, *5*, *10*, *16*

for this experiment (refer to Section 4.2). Specifically, we use stratified $\mathcal{K}$-fold cross-validation, i.e., we partition the entire data set of 69560 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 data set of 69560 sequences, respectively. The $\mathcal{K}$ readings thus obtained are then plotted using a box plot in Fig. 11 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 data set) to 0.96 (when trained on 93.8% of the data set). 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 data sets.

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 [5–7] and also some recent video frame prediction methods that exploit the spatio-temporal information from sequences, namely [39–41]. Among the recent frame prediction methods, in [39] 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 [40], 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 [41] 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 synthetically occluded sequences generated by introducing

50% occlusion on the test set of the *CASIA-B* data has been used in this experiment. It may be noted that the approaches discussed in [5, 6] carry out reconstruction of the *GEI* features, whereas each 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 GEIs computed from the predicted frames by our method and each of [7, 39–41] with that of the reconstructed *GEI*s generated by [5, 6] in terms of average *Dice score*. Fig. 12 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

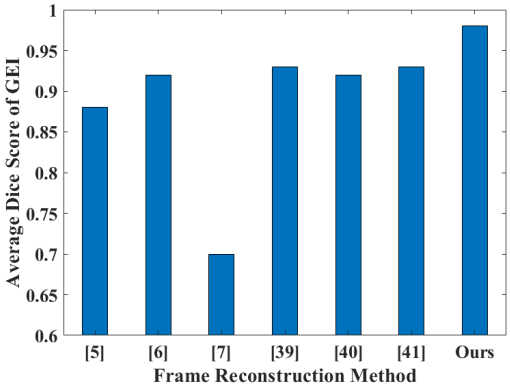

Fig. 12: Comparative study of the different reconstruction algorithms in terms of average *Dice score* of *GEI*

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 [39–41] and also [5, 6] are also relatively good and closely comparable to each other. However, the method in [7] performs poor quality reconstruction as is evident from the average *Dice score* value. This is mostly due to the fact that [7] 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 handling 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* data sets. Specifically, we categorize the existing approaches into three different groups, namely (i) gait recognition methods with occlusion handling mechanism, (ii) video frame prediction methods, and (iii) gait recognition methods without occlusion handling mechanism. The first category of methods include [5, 6] that perform recognition after reconstructing the *GEI* through a *CNN*, and [4, 42] that attempt to perform recognition without reconstructing the occlusion. In this category, we also study the accuracy given by two recent non-occlusion handling methods, namely, [15] and [24] on the sequences reconstructed by our *BGaitR-Net*. The video frame prediction methods include [39–41] for each of which the gait recognition accuracy is computed using *GEINet* [22]. Among the non-occlusion handling methods, we use some popular primitive approaches, namely, [11, 12, 14, 22, 23]. Results are shown in Table 2 in terms of *Rank 1 accuracy* for both the *TUM-IITKGP* and the *CASIA-B* datasets. Each reported accuracy

Table 2: 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 Rank 1 accuracy

Category	**Method**	**Rank 1 Acc. (%)**	
		TUM-IITKGP	**CASIA-B**
Proposed	***BGaitR-Net*+[22]**	**97.32**	**98.17**
Methods with Occl. Handling Mechanism	*BGaitR-Net*+[15]	96.37	96.77
	BGaitR-Net+[24]	95.56	97.58
	[5]	78.92	81.45
	[6]	80.00	92.74
	[42]	77.65	89.51
	[4]	85.32	89.51
	[7]	68.57	75.23
Frame Prediction Methods	[39]+[22]	47.66	91.54
	[40]+[22]	63.33	87.66
	[41]+[22]	76.66	94.36
Methods without Occl. Handling Mechanism	[23]	76.42	74.19
	[22]	76.79	63.71
	[11]	65.71	56.45
	[12]	70.23	79.83
	[14]	73.54	62.10

value in the table is obtained by training the corresponding gait recognition model on the complete training set corresponding to either the *TUM-IITKGP* data or the *CASIA-B* data, as explained in Section 4.1. 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 [15, 22, 24] results in obtaining a significantly high Rank 1 accuracy ($> 95\%$) for both synthetically and real-occluded test sets. In comparison, the accuracy values given by the other existing occlusion handling methods in gait recognition used in the comparative study, namely [4–

6, 42] on the same test sets are quite less. It may also be noted that, each of the video frame prediction methods, i.e., [39–41] combined with *GEINet* [22] shows a significantly higher recognition accuracy for the *CASIA-B* data than the *TUM-IITKGP* data. This is mostly 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 use Vanilla Autoencoder that is 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. As a result, the *Rank 1 accuracy* given by our approach on this data is *76.66%*, which improves over the best-performing video frame prediction method, i.e., [41] by more than *21%*. Also, as expected, the recognition accuracy of each of the non-occlusion handling methods [11, 12, 14, 22, 23] is quite less for occluded test sequences since these methods are designed to work well only if at least a complete gait cycle is available.

4.4 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 3. 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 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 use of the conditional vector c improves the average *Dice score* by *0.233*. Also, 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}.

Table 3: 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**

5 Conclusions and Future Work

In this work, we focus on gait recognition in the presence of occlusion. Given an occluded gait sequence we use a *VGG-16* model to detect the occluded frames in the sequence, and also a database of key poses to map the unoccluded frames of the sequence to the appropriate key poses. The novelty of the work is proposing a new and effective Deep Neural Network-based architecture, namely *BGaitR-Net*, to reconstruct the occluded frames present in a corrupted gait sequence. Although we have used the popular *GEINet* model [22] to carry out recognition from the reconstructed sequences, our *BGaitR-Net* can be conveniently integrated with any other effective gait recognition model to carry out recognition accurately. The proposed *BGaitRNet* is based on stacking of two Deep Neural Network architectures: (a) a *Convolutional Variational Autoencoder (CVAE)* and (b) a *Bidirectional Long-Short Term Memory (Bi-LSTM)*, and it has been seen to perform satisfactorily even if 60-70% frames in a gait cycle are missing/occluded. The *Encoder-Decoder* model used here, i.e., the *CVAE* employs the key walking pose corresponding to each frame of a set of sequential binary silhouettes as conditional vectors to provide compact encoded representations. The *Bi-LSTM* model is next used to filter the above-encoded sequence of frames to predict a missing/occluded frame by exploiting the spatio-temporal information available from the unoccluded frames of the sequence along with the key pose

information. Experimental results show that our overall reconstruction and recognition framework performs significantly accurately both on the synthetically occluded *CASIA-B* data and on the real static/dynamic occluded sequences present in the *TUM-IITKGP* data, and outperforms the existing techniques to occlusion handling in gait recognition by a significantly large margin. In the future, our work needs to be evaluated on more extensive real-occluded data sets.

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

1. R. T. Collins, R. Gross, and Jianbo Shi. Silhouette-Based Human Identification from Body Shape and Gait. In *Proc. of* 5^{th} *Intl. Conf. on Automatic Face Gesture Recognition*, pages 366–371, 2002.
2. Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang. Multi-task GANs for View-Specific Feature Learning in Gait Recognition. *IEEE Trans. on Information Forensics and Security*, 14(1):102–113, 2018.
3. Peng Zhang, Qiang Wu, and Jingsong Xu. VT-GAN: View Transformation GAN for Gait Recognition Across Views. In *Proc. of the Intl. Joint Conf. on Neural Networks*, pages 1–8, 2019.
4. Pratik Chattopadhyay, Shamik Sural, and Jayanta Mukherjee. Frontal Gait Recognition from Occluded Scenes. *Pattern Recognition Letters*, 63:9–15, 2015.
5. Maryam Babaee, Linwei Li, and Gerhard Rigoll. Gait Recognition from Incomplete Gait Cycle. In *Prceedings of the* 25^{th} *Intl. Conf. on Image Processing*, pages 768–772, 2018.
6. Maryam Babaee, Linwei Li, and Gerhard Rigoll. Person Identification from Partial Gait Cycle Using Fully Convolutional Neural Networks. *Neurocomputing*, 338:116–125, 2019.
7. Roy, Aditi and Sural, Shamik and Mukherjee, Jayanta and Rigoll, Gerhard. Occlusion Detection and Gait Silhouette Reconstruction from Degraded Scenes. *Signal, Image, and Video Processing*, 5(4):415–430, 2011.
8. Shiqi Yu, Daoliang Tan, and Tieniu Tan. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In *Proc. of the Intl. Conf. on Pattern Recognition*, pages 441–444, 2006.
9. Haruyuki Iwama, Mayu Okumura, Yasushi Makihara, and Yasushi Yagi. The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition. *IEEE Trans. on Information Forensics and Security*, 7(5):1511–1521, 2012.
10. Martin Hofmann, Shamik Sural, and Gerhard Rigoll. Gait Recognition in the Presence of Occlusion: A New Dataset and Baseline Algorithms. In *Proc. of the* 19^{th} *Intl. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision*, pages 99–104, 2011.
11. Ju Han and Bir Bhanu. Individual Recognition Using Gait Energy Image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006.
12. Aditi Roy, Shamik Sural, and Jayanta Mukherjee. Gait Recognition Using Pose Kinematics and Pose Energy Image. *Signal Processing*, 92(3), 2012.
13. Pratik Chattopadhyay, Aditi Roy, Shamik Sural, and Jayanta Mukhopadhyay. Pose Depth Volume Extraction from RGB-D Streams for Frontal Gait Recognition. *Journal of Visual Communication and Image Representation*, 25(1):53–63, 2014.
14. Erhu Zhang, Yongwei Zhao, and Wei Xiong. Active Energy Image Plus 2DLPP for Gait Recognition. *Signal Processing*, 90(7):2295–2302, 2010.
15. Sanjay Kumar Gupta and Pratik Chattopadhyay. Exploiting Pose Dynamics for Human Recognition from Their Gait Signatures. *Multimedia Tools and Applications*, 80(28):35903–35921, 2021.
16. Dong Xu, Shuicheng Yan, Dacheng Tao, Stephen Lin, and Hong-Jiang Zhang. Marginal Fisher Analysis and Its Variants for Human Gait Recognition and Content-Based Image Retrieval. *IEEE Trans. Image Process.*, 16(11):2811–2821, 2007.
17. Sabesan Sivapalan, Daniel Chen, Simon Denman, Sridha Sridharan, and Clinton Fookes. Gait Energy Volumes and Frontal Gait Recognition Using Depth Images. In *Proc. of the Intl. Joint Conf. on Biometrics*, pages 1–6, 2011.
18. Francesco Battistone and Alfredo Petrosino. TGLSTM: A Time Based Graph Deep Learning Approach to Gait Recognition. *Pattern Recognition Letters*, 126:132–138, 2019.
19. Hailong Hu, Yantao Li, Zhangqian Zhu, and Gang Zhou. CNNAuth: Continuous Authentication via Two-Stream Convolutional Neural Networks. In *Proc. of the Intl. Conf. on Networking, Architecture and Storage*, pages 1–9, 2018.
20. Yantao Li, Hailong Hu, Zhangqian Zhu, and Gang Zhou. SCANet: Sensor-Based Continuous Authentication With Two-Stream Convolutional Neural Networks. *ACM Trans. on Sensor Networks*, 16(3):1–27, 2020.
21. Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. On Input/Output Architectures for Convolutional Neural Network-Based Cross-View Gait Recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 29(9):2708–2719, 2017.
22. Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. GEINet: View-Invariant Gait Recognition Using a Convolutional Neural Network. In *Proc. of the Intl. Conf. on Biometrics*, pages 1–8, 2016.
23. Munif Alotaibi and Ausif Mahmood. Improved Gait Recognition Based on Specialized Deep Convolutional Neural Network. *Computer Vision and Image Understanding*, 164:103–110, 2017.
24. Sanjay Kumar Gupta and Pratik Chattopadhyay. Gait Recognition in The Presence of Co-variate Conditions. *Neurocomputing*, 454:76–87, 2021.
25. Shiqi Yu, Haifeng Chen, Edel B Garcia Reyes, and Norman Poh. GaitGAN: Invariant Gait Feature Extraction Using Generative Adversarial Networks. In *Proc. of the Conf. on Computer Vision and Pattern Recognition Workshops*, pages 30–37, 2017.
26. Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding Gait as a Set for Cross-View Gait Recognition. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 33, pages 8126–8133, 2019.
27. Hanqing Chao, Kun Wang, Yiwei He, Junping Zhang, and Jianfeng Feng. GaitSet: Cross-view Gait Recognition through Utilizing Gait as a Deep Set. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2021.
28. Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and

Zhiqiang He. Gaitpart: Temporal Part-Based Model for Gait Recognition. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 14225–14233, 2020.
29. Xu Song, Yan Huang, Caifeng Shan, Jilong Wang, and Yu Chen. Distilled Light GaitSet: Towards Scalable Gait Recognition. *Pattern Recognition Letters*, 2022.
30. Feng Han, Xuejian Li, Jian Zhao, and Furao Shen. A Unified Perspective of Classification-Based Loss and Distance-Based Loss for Cross-View Gait Recognition. *Pattern Recognition*, page 108519, 2022.
31. Wan Noorshahida Mohd Isa, Md Jahangir Alam, and Chikkanan Eswaran. Gait Recognition Using Occluded Data. In *Proc. of the Asia Pacific Conf. on Circuits and Systems*, pages 344–347, 2010.
32. Tracey K. M. Lee, Mohammed Belkhatir, and Saeid Sanei. Coping with Full Occlusion in Fronto-Normal Gait by Using Missing Data Theory. In *Proc. of the* 7^{th} *Intl. Conf. on Information, Communications and Signal Processing*, pages 1–5, 2009.
33. Dhritimaan Das, Ayush Agarwal, Pratik Chattopadhyay, and Lipo Wang. RGait-NET: An Effective Network for Recovering Missing Information from Occluded Gait Cycles. *arXiv preprint arXiv:1912.06765*, 2019.
34. Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *Stat*, 1050:10, 2014.
35. Mike Schuster and Kuldip K. Paliwal. Bidirectional Recurrent Neural Networks. *IEEE Trans. on Signal Processing*, 45(11):2673–2681, 1997.
36. Sepehr Maleki, Sasan Maleki, and Nicholas R Jennings. Unsupervised Anomaly Detection with LSTM Autoencoders Using Statistical Data-filtering. *Applied Soft Computing*, 108:107443, 2021.
37. Jose Mejia, Liliana Avelar-Sosa, Boris Mederos, Everardo Santiago Ramírez, and José David Díaz Roman. Prediction of Time Series Using an Analysis Filter Bank of LSTM Units. *Computers & Industrial Engineering*, 157:107371, 2021.
38. Aaron Carass, Snehashis Roy, Adrian Gherman, Jacob C Reinhold, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Mohsen Ghafoorian, Bram Platel, et al. Evaluating White Matter Lesion Segmentations with Refined SØRensen-Dice Analysis. *Scientific Reports*, 10(1):1–19, 2020.
39. Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3DLSTM: A Model for Video Prediction and Beyond. In *Proc. of the Intl. Conf. on Learning Representations*, 2019.
40. Vincent Le Guen and Nicolas Thome. Disentangling Physical Dynamics from Unknown Factors for Unsupervised Video Prediction. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 11474–11484, 2020.
41. Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xinguang Xiang, and Wen Gao. MAU: A Motion-Aware Unit for Video Prediction and Beyond. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Proc. of the Advances in Neural Information Processing Systems*, 2021.
42. Changhong Chen, Jimin Liang, Heng Zhao, Haihong Hu, and Jie Tian. Frame Difference Energy Image for Gait Recognition with Incomplete Silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009.