Chapter 6

Occlusion Handling in Video-Based Person Re-Identification

In this chapter, we consider the problem of re-identification from video data with sequential image frames corrupted with occlusion. The existing solutions to person re-identification that handle occlusion [125, 126, 154, 164–168] have been already discussed in detail in Chapter 2 and the introductory part of Chapter 5. To the best of our knowledge, to date there does not exist any method that addresses the problem of occlusion reconstruction from the sequential image frames present in a video by exploiting the spatio-temporal information contained in the adjacent frames. Although, these existing reconstruction approaches including those proposed in Chapter 5 are suitable for frame-by-frame image reconstruction in video sequences, it appears that the effectiveness of reconstruction can be improved if the spatio-temporal information extracted from the consecutive frames of a sequence is taken into consideration during the prediction. This is since in a video, the appearance of a subject in a particular frame depends on his/her appearances in the preceding frames. Here, we propose to fuse the spatio-temporal information contained in the present occluded frame and also a few preceding frames, rather than using the spatial information from the present occluded frame only, as done in Chapter 5. The extraction of the relevant spatio-temporal features has been carried out in the present chapter using a *Conv-LSTM* model. The reconstructed frames are further fine-tuned using *DCGAN*, and finally, re-identification is performed using a suitable classifier.

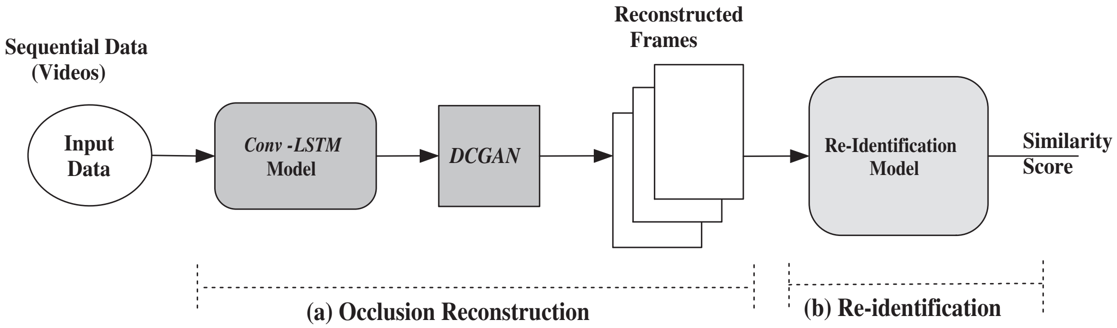

Figure 6.1: Overall framework of the proposed video-based re-identification approach in the presence of occlusion

The overall framework of the proposed multi-model approach for occlusion reconstruction and re-identification is shown in Figure 6.1. It can be seen from the figure that first, the input data (i.e., sequential frames) is fed to the *Conv-LSTM* model to perform coarse frame reconstruction, and these reconstructed frames are next individually passed through the same trained *DCGAN*, as discussed in Section 5.2.4 of Chapter 5 to fine-tune the reconstruction results further. Lastly, a Siamese network, namely, the Siamese Convolution Box (*SCB*), introduced in Chapter 3, is used to perform the re-identification.

6.1 *Conv-LSTM*-based Occlusion Reconstruction

Conv-LSTM is a specially modified version of LSTM (Long Short-Term Memory) [169] in which the convolution operation replaces the matrix multiplication inside the LSTM cell at every gate. This network is capable of capturing spatial features and learning long-term dependencies over time from sequential multi-dimensional data [170]. The equations below show the various operations involved in each *Conv-LSTM* layer. Let i_t,

f_t and o_t respectively represent input, forget and output gates, and $X_1 \ldots X_t$ represent inputs to a particular layer, and $C_1 \ldots C_t$ and $H_1 \ldots H_t$ represent cell outputs and hidden states, respectively. Further, let the symbols $*$ and $\circ$ denote the convolution operation and Hadamard Product, respectively. The symbols W and b correspond to the weight matrix and bias at particular input, and σ and $tanh$ represent the Sigmoid and Hyperbolic tangent activation function, respectively.

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \tag{6.1}$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \tag{6.2}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \tag{6.3}$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \tag{6.4}$$

$$H_t = o_t \circ tanh(C_t) \tag{6.5}$$

The *Conv-LSTM* model used in this work consists of nine convolutional layers followed by a *Conv2d* layer that outputs the reconstructed frame. An occluded frame (say, *Frame n*) is predicted by fusing the spatio-temporal information given by the same occluded frame (i.e., *Frame n*) along with the two previous frames (i.e., *Frame n-1* and *Frame n-2*). An insight view of the *Conv-LSTM* along with its input-output combination is explained using Figure 6.2. Table 6.1 presents the layer-wise configuration of the model. Each layer of the *Conv-LSTM* model except the last layer forwards the features to the next layer, and *ReLU* activation has been used across all the layers.

To train the *Conv-LSTM* model, we consider the *IIT(BHU) Re-identification* data set (refer to Section 2.4.2 of Chapter 2), and introduce varying levels of synthetic occlusion in the frames present in the gallery set of this data set using a method similar to that explained in Section 5.1 of Chapter 5. In the *IIT (BHU) Re-identification* data set, walking sequences of 41 subjects are captured, and on average, there are 47 frames

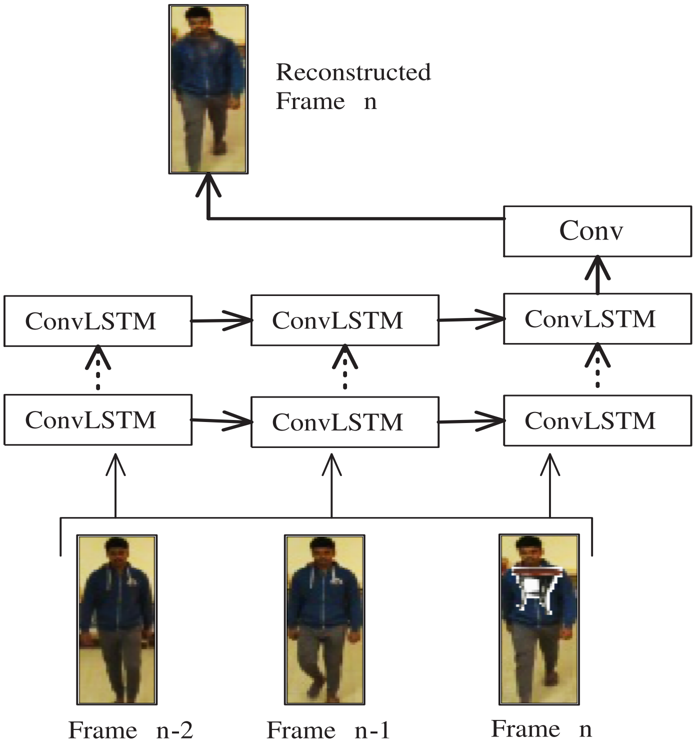

Figure 6.2: Insight of *Conv-LSTM* Model

per subject in the data set. The sequence corresponding to the first 80% frames from each subject is considered for forming the gallery set to train the occlusion reconstruction and re-identification models. The sequence corresponding to the remaining 20% frames for each subject is considered for forming the test set. On average, there are 38 sequential frames from each subject in the gallery set. Following Figure 6.2, the data set for training the *Conv-LSTM* model is formed by considering triplets of three frames from the gallery set, corrupting every third frame with synthetic occlusion, and

Table 6.1: Layer-wise specification of the *Conv-LSTM*. Here, ConvLSTM2d_i represents the i^{th} layer of the model

Network	Layer	Filter size	No. of filters
Conv-LSTM Model	ConvLSTM2d_1	5×5	128
	ConvLSTM2d_2 to ConvLSTM2d_8	3×3	128,64,64, 32,32,32,16
	ConvLSTM2d_9	1×1	3
	Conv2d	3×3	3

considering these three frames as input, and the corresponding clean third frame as the desired output. In this manner, we have created 784 such frame triplets and trained the model with Adam optimizer and binary cross-entropy loss for 150 epochs with a batch size of 64. We observe that a reconstructed frame by the *Conv-LSTM* often contains irregularities in image regions where occlusion is present in the input images, and the object edges in the generated image are not always smooth. This happens mostly due to fusing information from multiple frames with imperfectly aligned silhouettes. Hence, similar to the image-based occlusion reconstruction algorithm discussed in Chapter 5, here also we fine-tune the *Conv-LSTM*-generated images one by one by employing the same trained *DCGAN* (refer to Section 5.2.4 of Chapter 5).

6.2 Reconstruction Results

Figure 6.3 presents the reconstruction results by using only the *Conv-LSTM* model as well as by using the combination of *Conv-LSTM* and *DCGAN*, i.e., *Conv-LSTM+DCGAN*. The first two rows in the figure correspond to the unoccluded and synthetically occluded image frames from the gallery set of the *IIT (BHU) Re-identification* data set, while the third and fourth rows respectively show the *Conv-LSTM*-reconstructed frames and fine-tuned results from *Conv-LSTM+DCGAN*. It can be verified from the figure that the quality of the reconstructed outputs gets improved to a certain extent on fine-tuning the *Conv-LSTM*-generated reconstructed frames through *DCGAN*. Also, the color information present in the original frames is retained better after this fine-tuning phase.

The effectiveness of the *Conv-LSTM+DCGAN*-based occlusion reconstruction has been quantitatively evaluated using the Dice Similarity Score metric for the different margin (m) values (i.e., 5, 10, and 20). Corresponding to each of the 784 frame triplets used to train the *Conv-LSTM* model, we compute the *DSC* score between the predicted frame and the ground-truth unoccluded frame, and next average the *DSC* scores obtained

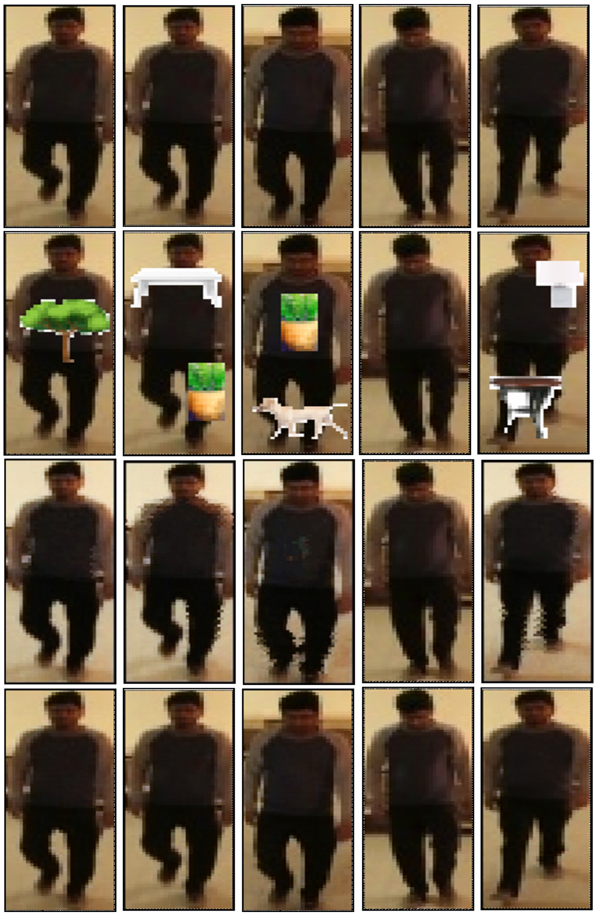

Figure 6.3: First and second rows show unoccluded and synthetically occluded frames from the *IIT (BHU)* Data set, while third and fourth rows present the reconstructed frames from the *Conv-LSTM* model and the corresponding fine-tuned frames through *DCGAN*

for all the above 784 outputs. These averaged *DSC* scores are presented in Table 6.2 for both the *Conv-LSTM* and the bi-network *Conv-LSTM+DCGAN* for the above-mentioned different m values. The results shown in the table indicate that there is a minor improvement in the reconstruction quality on applying the *DCGAN*-based fine-tuning to the *Conv-LSTM*-generated outputs.

In our next experiment, we test the stability of the *Conv-LSTM* model on unseen data by studying if the model can provide consistent performances (in terms of *DSC* score) for different random initialization of the network. For this, we train the network from scratch five different times using the same training set as already described in Section

Table 6.2: Dice score for the reconstruction model (*Conv-LSTM*) at different margin values for the *IIT (BHU) Re-identification* and *PRID2011* data set

Data set	**Network**	***m=5***	***m=10***	***m=20***
IIT (BHU) ReID	*Conv-LSTM*	0.78	0.90	0.96
	Conv-LSTM+DCGAN	0.79	0.90	0.97
PRID2011	*Conv-LSTM*	0.71	0.79	0.86
	Conv-LSTM+DCGAN	0.73	0.80	0.89

6.1. The variation in the average *DSC* scores obtained is shown by means of box and whiskers plot in Figure 6.2 using the synthetically occluded test sets from *IIT (BHU) Re-identification data* and *PRID2011 data*. To construct these synthetically occluded data sets, here also we consider the final 20% sequential frames from each subject (which are not used for training the model) and corrupt these with varying levels of synthetic occlusion to form the test sets. It can be observed from Figure 6.2 that the proposed reconstruction model, i.e., *Conv-LSTM+DCGAN* performs robustly for both the synthetically occluded video data sets used in the study for all the different margin values (i.e., 5, 10, and 20). The range of whiskers and the inter quartile range (i.e., 25

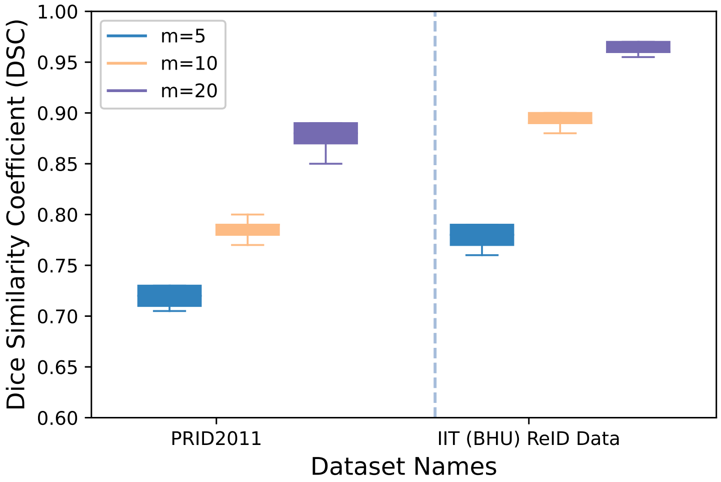

Figure 6.4: Box and whiskers plot showing the stability of the reconstruction model *Conv-LSTM+DCGAN* on the synthetically occluded test sets of *IIT (BHU) Re-identification* data and *PRID2011* data

to 75%) can be seen varying in the range of 0.01 to 0.05 and 0.1 to 0.3, respectively, which justifies the robust performance of the *Conv-LSTM+DCGAN* model for video data sets.

6.3 Person Re-identification and Experimental Evaluation

To perform re-identification using the *Conv-LSTM+DCGAN* reconstructed sequences, we study the use of both the *SCB*-based re-identification introduced in Chapter 3 and the *T-MAN*-based re-identification introduced in Chapter 4. While the first approach averages all the reconstructed image frames in a video sequence before carrying out the *SCB*-based matching, the second approach computes the matching probabilities using three different dynamic features obtained from three time-series models and next fuses these predictions to perform the matching. The same network architectures, optimizer, and loss functions discussed in Chapters 3 and 4 have also been used in this work to train the networks with the present gallery set. It may be noted that none of the existing approaches in the literature perform video-based person re-identification in the presence of occlusion. Hence, we make a comparative study with the existing approaches to perform re-identification using the reconstructed sequences generated by *Conv-LSTM+DCGAN*. We consider the test sequences from two video-based person re-identification data sets, namely, the *IIT (BHU) Re-identification data* and the *PRID2011* data, and corrupt the frames in the test sequences with varying levels of synthetic occlusion, as already discussed before. The occluded frames in each sequence are next reconstructed using *Conv-LSTM+DCGAN* and these reconstructed sequences are used for evaluating the performances of the different re-identification approaches. The methods that have been used in this comparative study include previous approaches [1–3, 63] as well as the SCB-based and *T-MAN*-based re-identification techniques discussed in Chapters 3 and 4, respectively. Before evaluation, we train each of these models with the unoccluded gallery set formed from the first 80% frames corresponding to each subject for both the data sets. Table 6.3 presents the Rank 1 re-identification accuracy of the above-mentioned techniques on the synthetically occluded test sets of the *IIT (BHU) Re-identification* data and the *PRID2011* data generated from the final 20% frames corresponding to each subject. It can be seen from the results that *T-MAN*, i.e., the video-based re-identification approach proposed in Chapter 4 performs quite accurately for both the *IIT (BHU) Re-identification* data and the *PRID2011* data. Its performance is closely comparable with that of [63] that follows a hypergraph-based matching algorithm. While for the *IIT(BHU) Re-identification* data, our *T-MAN* performs better than that of [63] by about 0.40% Rank 1 accuracy, for the *PRID2011* data, [63] outperforms our work by 2.40% in terms of Rank 1 accuracy. The results indicate that our *T-MAN* is more effective in handling sequences that contain some type of sequential activity such as walking, running, etc., as present in the IIT (BHU) Re-identification data. For video data sets containing non-sequential activities, as in the *PRID2011*, the temporal information does not play an important role in deriving effective features for re-identification, and hence, the use of *T-MAN* is not beneficial in such cases. Since videos captured in surveillance sites would typically consist of sequential activities from multiple subjects, *T-MAN* is expected to have higher applicability than [63] in such scenarios. A more rigorous comparison is required between our *T-MAN* and [63] using an extensive video-based re-identification data set (which is not available at present)

Table 6.3: Comparative study of Rank 1 accuracy of the different methods on the occlusion reconstructed images by *ConvLSTM+DCGAN* for the *IIT(BHU) Re-identification* and *PRID2011* data sets

Methods	**Re-identification Accuracy (%)**	
	IIT(BHU) ReID	***PRID2011***
SCB	88.40	70.40
T-MAN	**93.40**	82.00
Yan et al. [63]	93.00	**84.40**
Ahmed et al. [1]	74.60	69.80
Subramaniam et al. [2]	77.40	74.00
Li et al. [3]	67.50	64.80

and is scope for further study. Apart from [63], the *T-MAN* has been seen to perform with much higher accuracy than each of the other existing methods for both the data sets used in the study. The accuracy given by the *SCB*-based re-identification approach discussed in Chapter 3 is also seen to be quite good for the *IIT (BHU) Re-identification*, but its performance on the *PRID2011* data is not appreciably high. This is due to the fact that the sequences in the *IIT (BHU) Re-identification* data consist of the front view of human walking only as a result of which the contextual and shape-related person-specific information is accurately preserved in the cropped and normalized frames. On the other hand, the *PRID2011* data consists of image sequences of persons in varying poses, due to which the averaged frame input to the *SCB* contains blurred and irregular images that do not encode the shape information accurately. This, in turn, affects the performance of the *SCB*-based re-identification.

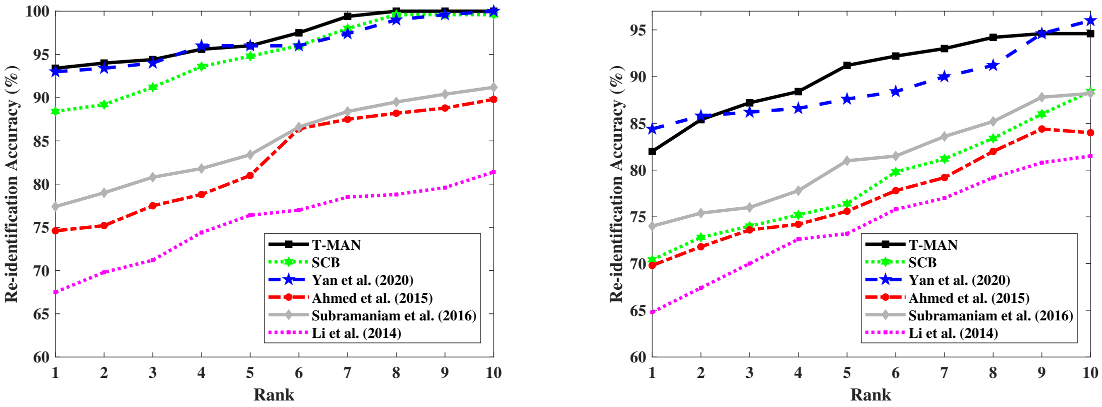

Figure 6.5: Comparative study in terms of rank-based accuracy of *T-MAN* with *SCB* and the approaches [1–3] using synthetically occluded (a) *IIT (BHU) Re-identification* data and (b) *PRID2011* data

Next, we study the rank-wise improvement in accuracy of each of the re-identification approaches used in the previous experiment through *Cumulative Match Characteristic Curves*. Figures 6.5(a) and (b) present the corresponding results for the synthetically occluded *IIT (BHU) Re-identification* data and *PRID2011* data, respectively for Ranks

1 to 10. A similar conclusion can also be derived from the results in these figures. The performances of *T-MAN* and [63] are closely comparable, and these are significantly better than the other compared approaches. While for the *IIT (BHU) Re-identification* data, *T-MAN* performs either better or equivalently to [63] for all the different rank values from 1 to 10, for the other data set, i.e., *PRID2011*, the performance of *T-MAN* is better than or equivalent to that of [63] for the ranks 3 to 9. However, the average accuracy given by *T-MAN* for the different rank values is better than [63]. While the *SCB* also performs with high accuracy in the case of the IIT(BHU) Re-identification data, its performance is not so appealing for the *PRID2011* due to a reason similar to that explained in the previous experiment. The re-identification accuracy given by each of the other video-based re-identification methods used in this study for the different rank values on the *Conv-LSTM+DCGAN*-based reconstructed sequences is also seen to be quite low for both the data sets.

Table 6.4: Comparative study of Rank 1 accuracy of *T-MAN* (with and without reconstruction) for the synthetically occluded *IIT(BHU) Re-identification* and *PRID2011* data sets

Methods	***IIT(BHU) ReID***	***PRID2011***
T-MAN (with reconstruction)	**93.40**	**82.00**
T-MAN (without reconstruction)	76.00	68.40

In our next experiment, we study the impact of the *Conv-LSTM+DCGAN*-based occlusion reconstruction model in improving the prediction accuracy of the *T-MAN* approach on occluded sequences. For this, we study the Rank 1 accuracy values given by *T-MAN (with reconstruction)* and *T-MAN (with reconstruction)* on the synthetically occluded samples of *IIT (BHU) Re-identification data* and *PRID-2011* data, and the results are shown in Table 6.4. From these results, it can be observed that the *T-MAN (with reconstruction)* improves over the accuracy given by *T-MAN (without reconstruction)* by more than 17% for the synthetically occluded *IIT (BHU) Re-identification* data and by more than 13% for the synthetically occluded *PRID2011* data. The results emphasize the need for employing the occlusion reconstruction model as an essential step before performing the re-identification. Also, it can be concluded from the extensive experimental study that the *Conv-LSTM+DCGAN*-based reconstruction model is indeed effective in reconstructing the occluded frames by exploiting the Spatio-temporal information from the preceding frames and help in obtaining an improved re-identification accuracy.

Finally, we make a thorough comparative study and thorough analyses of the best approaches determined in the individual thesis chapters on occluded data sets. The scenario considered in Chapter 3 deals with re-identification using image frames that may be sequential/non-sequential, while that in Chapter 4 considers video-based re-identification, i.e., re-identification using video data sets that are more commonly obtained in surveillance setups. Although, the approaches proposed in these two chapters have been seen to outperform several state-of-the-art re-identification approaches, these do not consist of any occlusion handling module and hence are not expected to be effective in dealing with image frames corrupted with occlusion. In the following two chapters, i.e., Chapters 5 and 6, we consider the challenging scenario of the presence of occlusion in non-sequential and sequential image frames. In Chapter 5, we propose a method for the reconstruction of occluded frames by using only the spatial information available in the frames through a Deep Neural Network-based generative model (refer to Sections 5.2.1 and 5.2.2) and next using reconstructed frames for re-identification. In Chapter 6, we consider occlusion to be present in sequential image frames, and improve upon the work in Chapter 5 by exploiting the spatio-temporal information from the adjacent sequential frames through a *Conv-LSTM* model, which is expected to make better reconstruction of the occluded regions in the frames of a video sequence. Let us use symbols $M1$, $M2$, $M3$, and $M4$ to denote the best approaches proposed in Chapters 3-6.

Tables 6.5 and 6.6 present an unified interpretation of results for all the approaches

Table 6.5: Comparative study of Rank 1 accuracy of all the proposed approaches on *iLIDS-VID*, *PRID2011*, and *IIT(BHU) Re-identification* data sets

Deg. of Occ.	**Proposed Method**	***iLIDS-VID (%)***	***PRID 2011(%)***	***IIT (BHU) ReId (%)***	***Avg. Res. Time (millisecs)***
10 %	*M1*	66.40	61.20	68.40	83.00
	M2	71.50	70.40	76.00	81.50
	M3	68.00	62.40	71.40	112.4
	M4	**76.80**	**73.40**	**90.20**	94.60
20%	*M1*	60.40	56.40	65.40	82.00
	M2	69.20	63.50	72.50	84.20
	M3	61.40	60.60	68.20	110.00
	M4	**73.40**	**70.40**	**88.40**	98.20
30%	*M1*	51.20	44.20	57.80	85.20
	M2	60.60	60.20	68.40	84.20
	M3	54.00	58.80	62.00	118.20
	M4	**71.60**	**67.50**	**85.80**	104.60

proposed in the thesis on video-based and image-based data sets corrupted with occlusion, respectively. As video data sets we consider the *iLIDS-VID*, *PRID2011*, and *IIT(BHU) Re-Identification data*, and as image data sets we consider the *CUHK_01*, *CUHK_03*, and *Market-1501* data. Each of the tables present the re-identification accuracy obtained from using our proposed approaches on the varying data sets for different degrees of occlusion, namely, 10%, 20%, and 30%, and also the average response time required by each method for each of the different degrees of occlusion.

It can be seen from the results in Table 6.5 that for the video-based data sets (sequential image frames), the approach *M4* performs the best among all the other approaches for any degree of occlusion. Although, all other approaches perform with a satisfactory level of accuracy for low degree of occlusion (i.e., 10%,) their performance degrades drastically for higher degrees of occlusion (i.e., 20% and beyond). The improved performance of $M4$ over $M1$ and $M2$ is expected since neither of the later two models incorporate any occlusion reconstruction strategy. Careful comparison of the results given by $M3$ and $M4$ in Table 6.5 for video-based data sets highlights the fact that

Table 6.6: Comparative study of Rank 1 accuracy of all the proposed approaches on *CUHK_01*, *CUHK_03*, and *Market-1501* data sets

Deg. of Occ.	**Proposed Method**	***CUHK_01 (%)***	***CUHK_03 (%)***	***Market -1501 (%)***	***Avg. Res. Time (millisecs)***
10%	*M1*	82.40	81.30	77.80	43.00
	M2	78.00	74.40	76.30	65.20
	M3	**94.80**	**94.00**	**94.60**	74.00
	M4	84.60	87.50	88.20	70.40
20%	*M1*	80.00	78.60	72.00	41.20
	M2	75.40	70.00	71.00	62.40
	M3	**94.00**	**93.20**	**94.40**	76.40
	M4	83.00	83.80	83.50	72.40
30%	*M1*	74.60	71.40	68.40	49.90
	M2	72.60	64.00	62.80	52.80
	M3	**90.20**	**89.00**	**90.80**	86.00
	M4	78.80	79.20	75.20	80.00

the use of simple spatial pixel-based information for frame reconstruction through Deep Generative Models, as done for *M3*, is not effective enough for occlusion reconstruction from sequential frames. Rather, spatio-temporal information extracted through *LSTM*, as done for *M4*, can help in achieving improved frame prediction thereby increasing the re-identification accuracy. The results presented in Table 6.6 show that instead of $M4$, $M3$ that follows *Autoencoder*-based reconstruction, provides the best re-identification accuracy on each of the image-based data sets for the varying degrees of occlusion. While the improved performance of $M3$ over both $M1$ and $M2$ (that do not incorporate any occlusion reconstruction mechanism) is justified, the improvement over $M4$ is mainly due to the absence of spatio-temporal information in the input non-sequential frames that causes the *Conv-LSTM*-based reconstruction model used for $M4$ to predict poor quality of frames.

As can be seen from Tables 6.5 and 6.6, the average response times for methods $M3$ and $M4$ are always higher than $M1$ and $M2$. This is since both $M3$ and $M4$ involve an additional occlusion reconstruction stage.

We have made further experiments to test the robustness of our proposed approaches for varying occluded test sets and also for different model initialization. First, we observe the performances on the same trained model for each of the methods $M1$, $M2$, $M3$, and $M4$, on four different subsets of 50 test ids from the complete test set of 200 test-ids corresponding to the video-based *PRID2011* data and plot the results by means of grouped bar chart in Fig. 6.6(a). In this figure, the first group of bars provide the accuracy given by $M1$ on the above four subsets of the *PRID2011* data. The next three groups of bars show the corresponding results for the methods $M2$, $M3$, and $M4$. Similar results are also presented in Fig. 6.6(b) for the synthetically occluded Market-1501 data that contains non-sequential frames. This data contains 300 test ids and hence the number of unique test ids present in each of the four subsets of the test set is 75. To compute the results for the two plots, we corrupt the frames present in each data set by introducing 20% synthetic occlusion. Once again, it can be verified from the plots that among $M1$, $M2$, $M3$, and $M4$, the method *M4* shows the best and most consistent results for all the different subsets of the *PRID2011* video-based data

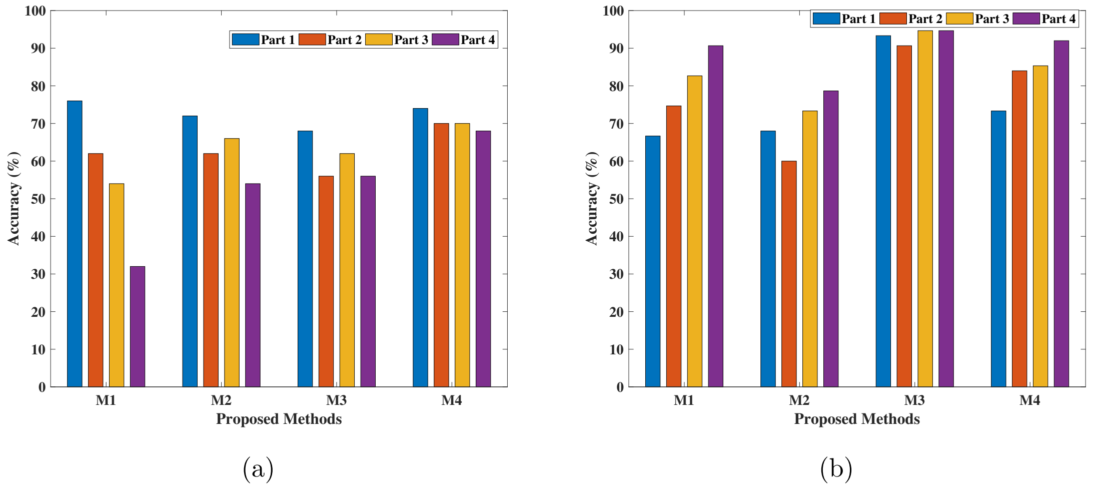

Figure 6.6: Comparison of robustness of all four proposed methods *M1*, *M2*, *M3*, and *M4* on various test sets constructed from the (i) video-based *PRID2011* data, and (ii) image-based *Market-1501* data corrupted with 20% synthetic occlusion

set while *M3* shows the best stable results for *Market-1501* image-based data set.

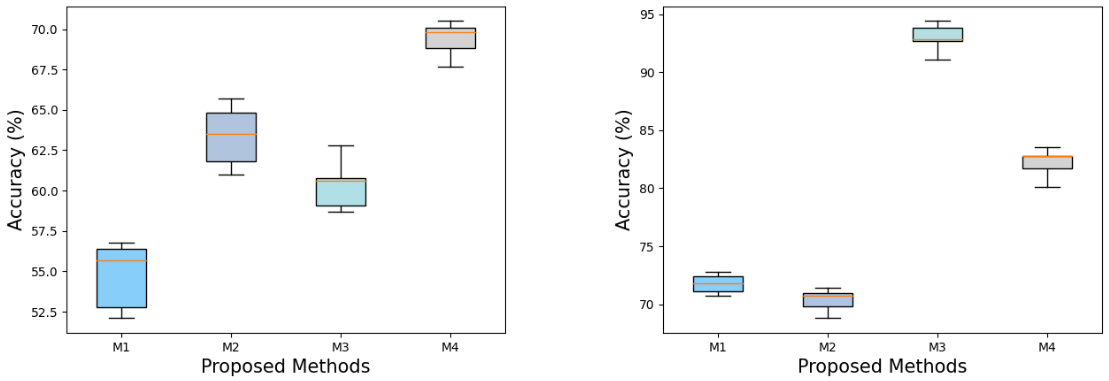

Figure 6.7: Comparison of stability of the Deep Learning models used in the four proposed methods *M1*, *M2*, *M3*, and *M4* by training the models multiple times from scratch on the same data and observing the overall re-identification accuracy of the approaches on (a) video-based *PRID2011* data, and (ii) image-based *Market-1501* data corrupted with 20% synthetic occlusion

Finally, to test the stability of the models used for the different approaches *M1*, *M2*, *M3* and *M4* proposed in the thesis, i.e., *SCB* in Chapter 3, *TMAN* in Chapter 4, *Autoencoder+DCGAN+SCB* in Chapter 5, and *(Conv-LSTM)+DCGAN+TMAN* in Chapter 6, we conduct an experiment in which we train each model from scratch five different times on the same training set with different initialization of the model parameters, and observe its performance on the test set. Here also, we use the *PRID2011* as the video-based data and *Market-1501* as the image-based data. The complete test sets from these two data sets consisting of 200 and 300 unique test ids have been used for evaluation of the model by corrupting the frames with 20% synthetic occlusion. Figs. 6.7(a) and (b) show the re-identification accuracy given by the four methods $M1$, $M2$, $M3$, and $M4$ in terms of box-plots for the synthetically occluded data sets *PRID2011* and *Market-1501*, respectively. Here also, it is observed from Fig. 6.7(a) that for video-based data sets, the the method *M4* provides the best and most consistent re-identification accuracy among all the other proposed methods which emphasizes the stability of the proposed occlusion reconstruction and re-identification approach. Similarly, for the results in Fig. 6.7(b) corresponding to the occluded image-based data sets, it is seen that the method *M3* shows the best and most consistent performance.

From the above results, it can be concluded that the approach $M4$ discussed in Chapter 6 is most suitable for automated re-identification in surveillance sites like airports, railway stations, etc., where video sequences of walking subjects are captured, and individual sequences may be corrupted due to occlusion. However, if instead of such sequential information, only snapshots of an individual from the same/different viewpoints are available that are corrupted with occlusion, then the *Conv-LSTM* based occlusion reconstruction approach discussed in Chapter 6 is likely to fail due to the unavailability of the desired sequential information. In such cases, the image-based occlusion reconstruction approach $M3$ presented in Chapter 5 would be more suitable for deployment. On the other hand, in constrained situations where occlusion is not present, the approaches $M1$ and $M2$ discussed in Chapters 3 and 4 can be conveniently deployed.

6.4 Summary

In this chapter, we present an approach for occlusion reconstruction in videos and perform re-identification using the reconstructed frames. Since video frames captured in surveillance sites usually contain some form of sequential activity of persons, we propose to reconstruct each occluded frame by exploiting the spatio-temporal information available in that frame and also a few previous frames. This is accomplished in this work by employing a *Conv-LSTM* model that takes as input a sequence of frames including the occluded frame and predicts a coarsely reconstructed frame by exploiting the spatio-temporal patterns present in the sequence. The reconstructed results are further fine-tuned by passing each reconstructed frame through a *DCGAN*. Qualitative and quantitative results show that the reconstruction results are quite good, and using these reconstructed sequences for re-identification results in a significantly improved re-identification accuracy compared to that obtained by using the occluded sequences directly for re-identification. Due to the unavailability of existing video-based re-identification approaches dealing with occlusion, we make a comparative study of our proposed re-identification methods presented in Chapters 3 and 4 with that of some popular existing approaches on the *Conv-LSTM+DCGAN*-generated reconstructed sequences. Experimental results show that our *T-MAN* (introduced in Chapter 4) performs more accurately than *SCB* (introduced in Chapter 3) on video sequences, which is expected. The *T-MAN*-based approach also performs better than each of the other compared techniques except the one in [63]. While the accuracy of *T-MAN* is better than [63] for all the different rank values in the case of the *IIT (BHU) Re-identification* data, it performs with a slightly lesser accuracy than [63] for a few ranks in the case of the *PRID2011* data. A more thorough comparative study is required to determine the best among these two on extensive video-based re-identification data sets, which is scope for future study. In Chapters 5 and 6, we have not employed any occlusion detection mechanism since the evaluation of each approach is done using synthetically occluded data sets. However, while dealing with image/video sequences in real-life surveillance sites, an occlusion detection mechanism must be incorporated before carrying out the reconstruction to determine which frames in the sequence have to be reconstructed. Standard CNN models make this prediction in an automated manner [171].