Scene Independency Matters: An Empirical Study of Scene Dependent and Scene Independent Evaluation for CNN based Change Detection

Murari Mandal, Santosh Kumar Vipparthi

Abstract—**Visual change detection in video is one of the essential tasks in computer vision applications. Recently, a number of supervised deep learning methods have achieved top performance over the benchmark datasets for change detection. However, inconsistent training-testing data division schemes adopted by these methods have led to documentation of incomparable results. We address this crucial issue through our own propositions for benchmark comparative analysis. The existing works have evaluated the model in scene dependent evaluation setup which makes it difficult to assess the generalization capability of the model in completely unseen videos. It also leads to inflated results. Therefore, in this paper, we present a completely scene independent evaluation strategy for a comprehensive analysis of the model design for change detection. We propose well-defined scene independent and scene dependent experimental frameworks for training and evaluation over the benchmark CDnet 2014, LASIESTA and SBMI2015 datasets. A cross-data evaluation is performed with PTIS dataset to further measure the robustness of the models. We designed a fast and lightweight online end-to-end convolutional network called ChangeDet (speed-58.8 fps and model size-1.59 MB) in order to achieve robust performance in completely unseen videos. The ChangeDet estimates the background through a sequence of maximum multi-spatial receptive feature (MMSR) blocks using past temporal history. The contrasting features are produced through the assimilation of temporal median and contemporary features from the current frame. Further, these features are processed through an encoder-decoder to detect pixel-wise changes. The proposed ChangeDet outperforms the existing state-of-the-art methods in all four benchmark datasets.**

Index Terms— **Change detection, background subtraction, scene independence, video analysis, deep learning.**

I. Introduction

CHANGE detection in video is an essential computer vision task having numerous applications in anomaly detection, object tracking, traffic monitoring, human-machine interaction, behavior analysis, action recognition and visual surveillance [1, 6, 30]. The objective of change detection is to represent a video frame through foreground and background regions. However, various real-world scenarios such as fluctuation in background regions, illumination variation, shadow, variable frame rate of different cameras, weather changes, intermittent object motion, camera jitter, and variable object motion make change detection a very challenging task.

Traditional approaches for change detection have primarily used background subtraction techniques to model background behavior and identify the foreground region using various thresholding techniques [1-30]. More recently, numerous convolutional neural network (CNN) based techniques [31-47] have also been proposed in the literature. The change detection techniques in the literature can be categorized into supervised and unsupervised approaches. Moreover, in terms of experimental setups, we can further categorize the supervised methods into scene dependent (SDE) and scene independent evaluation (SIE) setup. In *'scene dependent'* setup, training and testing sets consist of frames from the same video sequences, whereas, completely unseen videos are used for testing in '*scene independent*' setup. The sample data-division strategies showcasing the difference between SDE and SIE setup is shown in Fig. 1.

The unsupervised change detection techniques [5-8, 15-17, 49-51, 66-78] do not require any labeled samples for algorithm development. These techniques naturally follow a scene independent strategy for performance evaluation. On the other hand, the supervised approaches require a certain amount of labelled training data/samples in order to learn optimal model parameters. This simple observation gives rise to the question, "what should be the data division strategy for supervised change detection?". The benchmark datasets CDnet 2014 [48], LASIESTA [52], SBMI2015 [65] do not define the training and testing partitions. Therefore, researchers [33-47, 56-63] have adopted different data division schemes to evaluate and compare their work with existing methods.

The most prominently adopted SDE strategy is to select training data from certain temporal proportions of video sequences. Since the background remains more or less similar in the entire video sequence. The training and testing data are highly similar. This would give an unfair advantage to the CNN

Murari Mandal and Santosh Kumar Vipparthi are with the Vision Intelligence Lab, Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur, India (Email: murarimandal.cv@gmail.com; skvipparthi@mnit.ac.in)

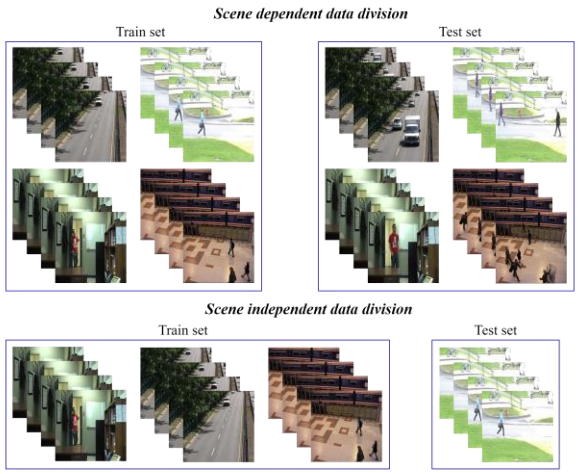

Fig. 1. Difference between scene dependent and scene independent data division schemes. In scene dependent setup, training and testing video frames share the same background which leads to high similarity between them. However, in scene independent setup, completely unseen videos are tested for evaluation.

model in evaluation in comparison to unsupervised methods. Moreover, inconsistent experimental setups for scene dependency has also led to documentation of incomparable results. The same warning is clearly mentioned in *changedetection.net* leaderboard page: "*Methods with the "supervised method" tag involve a supervised machine learning algorithm potentially trained on the ground truth data used to produce the metrics reported in this page*". Therefore, it is very important to ensure scene independence in the evaluation of supervised change detection methods. The model performance must be evaluated over completely unseen videos. In addition, there is a need for clearly defined data-division schemes for training and evaluation in both SDE and SIE setups. As of this moment, there has been no attempt to present a standard/benchmark evaluation strategy for supervised change detection methods.

In this paper, we present two well defined experimental frameworks (data-division schemes) to evaluate the effectiveness of the deep learning models. The scene independent strategy-based framework addresses the abovementioned shortcomings in the recent state-of-the-art. Thus, baseline results for the same are presented in the experiment sections. Furthermore, in order to evaluate the robustness of proposed ChangeDet over existing deep learning models, a scene dependent experimental strategy is proposed. The proposed network is designed to ensure generalized performance across both the experimental setups. The contributions of this paper can be summarized in the following points.

1) We propose a highly resource efficient end-to-end convolutional network, ChangeDet, for change detection. Our online model is very fast (speed-58.8 fps) and lightweight (model size-1.59 MB), making it suitable for real-time applications.

2) We designed DRBE block to estimate the background representation from recent temporal history and a sequence of maximum multi-spatial receptive feature (MMSR) to selectively determine the probable salient background representation. Furthermore, CFA block is designed to assimilate the contemporary features with the estimated background to produce contrasting response for effective change detection.

3) We have clearly defined the scene independent and scene dependent strategies (training-testing data division) for evaluation in three benchmark datasets CDnet2014, LASIESTA and SBMI2015. A detailed empirical study of scene independent versus scene dependent evaluations (currently being used in the literature) is also performed through multiple experimental analysis. We also conduct cross-dataset analysis over PTIS dataset. Evaluation over completely unseen videos in SIE and cross-data setup ensures fair evaluation of generalization capability of the designed network.

4) The proposed ChangeDet outperforms (overall, in terms of accuracy, speed, memory and compute efficiency) the existing state-of-the-art methods. The ablation studies of the proposed network are also discussed in the experimental section.

II. Related Work

The objective of a change detection technique is to segment the current video frame into foreground or background regions based on the past temporal behavior at each pixel location. We discuss the existing literature for change detection in two categories: traditional and deep learning-based approaches. We further discuss the training and evaluation frameworks adopted in the recent supervised change detection algorithms.

A. Traditional Approaches

The general framework for traditional change detection techniques can be described through the following stages: feature extraction from the current and previous frames, background model initialization and maintenance, foreground detection.

1) Feature extraction: The low-level image features, i.e., grayscale/color intensity [1-10, 56] and edge magnitudes [11,12] are commonly used in change detection algorithms. Superpixel based features have also been used in [13-15]. Moreover, specific spatial and spatiotemporal feature descriptors [16-18, 49-51] have been designed for enhanced performance.

2) Background model initialization and maintenance: The background modelling techniques can be loosely categorized into parametric [1-4, 12, 14] and non-parametric [5-11, 16-21] approaches. In parametric approaches, the statistical distribution at each location is modelled and updated through models such as mixture of Gaussians (MOG) [3] and Expectation- Maximization (EM) algorithms. Zivkovic [1] and Varadarajan et al. [2] improved upon the MOG with variable parameter selection, spatial mixture of Gaussians and fast initialization. The non-parametric methods are primarily inspired by the strategies based on kernel density estimation [19, 20] and the consensus-based method [21]. In a seminal work ViBe [6], three significant background model maintenance policies were proposed: random background sample replacement to represent short and long-term history, memoryless update policy and spatial diffusion via background sample propagation. These strategies have been widely adopted in recent state-of-the-art change detection techniques [5, 7, 16-18]. Adaptive update policies for decision thresholds (for foreground segmentation) and learning rates (for model update) were introduced in [5]. Furthermore, an adaptive feedback mechanism to continuously monitor background model fidelity and segmentation entropy to update these parameters was presented in [7, 17, 18,]. A deterministic policy to update background models was proposed by Mandal et al. [8]. Bianco et al. [22] conducted multiple experiments to combine various change detection techniques through genetic programming.

Numerous robust principal component analysis (RPCA) based models [23-25] have also been designed to estimate background as a low-rank component and foreground as a sparse matrix. Similarly, robust spatiotemporal subspace modelling for dynamic videos were presented in [24, 25]. Local codebook-based models [26], neural networks based self-organizing maps [27, 28, 66], semantic segmentation [29] based models, and other methods [66-78] have also been presented in the literature. A more detailed categorization of traditional change detection techniques can be found in [10].

3) Foreground detection: Threshold- based segmentation with postprocessing techniques [6, 10, 21, 29] are commonly used in the existing literature for foreground detection. Numerous policies [5, 7, 8, 17, 18, 26] have also been presented to adaptively update the foreground segmentation thresholds.

B. Deep Learning Approaches

Recently, researchers have also used CNN models for change detection [31-47, 57-64]. Many attempts have leveraged off-the-shelf pre-trained CNN to extract features and integrate with statistical/hand-crafted background modelling techniques for temporal feature encoding [33-36]. Certain researchers [33, 37-39] proposed to divide the current frame and background models into overlapping/non-overlapping patches. Thereafter, CNN features are learned from the concatenated input layer for local change detection. Braham and Droogenbroeck [39] proposed to use a single background image and current frame to feed the designed CNN model. Similarly, Babaee et al. [38] generated background models using proven hand-crafted approaches like SuBSENSE [17] and flux tensor [4]. Nguyen et al. [37] designed the triplet CNN network to learn relevant motion features. Patil et al. [40] estimated the background saliency map with temporal histogram features and designed a multiscale encoder-decoder network to estimate the foreground. Moreover, Wang et al. [41] proposed to collect selective annotations for model training and perform frame-wise segmentation for change detection. This methodology is primarily directed towards alleviating the cumbersome tasks of pixel-wise annotations for ground-truth generation. They trained a CNN model with carefully selected frames from a video and then perform binary segmentation over all video frames to generate pixel-wise estimates.

The earlier CNN based methods in the literature were dependent on statistical approaches to extract temporal features. Thus, the performance of these approaches is limited by the capability of traditional methods to estimate background. Although these techniques work very well over certain scenarios, in order to fully realize the potential of deep learning, it is important to eliminate dependence on hand-crafted features for temporal feature encoding. In this regard, end-to-end CNN models are also proposed in the literature. Chen et al. [42] designed an attention ConvLSTM to model pixel-wise changes over time. Yang et al. [43] temporally encoded the motion information by sampling multiple previous frames with increasing intervals. A compact CNN for end-to-end training was presented in [44]. Akilan et al. [45] developed 3D CNN with long short-term memory (LSTM) pipeline for foreground segmentation. Numerous other works [57-64] have also explored the CNN based designs for performance improvement. More recently, conditional generative adversarial network (cGAN) [46] and cycleGAN [47] based learning frameworks were also used for change detection.

C. Training and Evaluation Frameworks for Traditional Versus Deep Learning Approaches

The traditional methods [1-30] for change detection usually do not require (except in neural networks-based approaches [27, 28]) labeled training data. Thus, there is no need to define train-test splits. However, it is a crucial decision in supervised (deep learning) change detection techniques. The benchmark datasets CDnet 2014, LASIESTA and SBMI2015 do not define the train-test division. Therefore, researchers have used different data division strategies for network training and evaluation. Almost all existing deep learning frameworks [34-47] follow the scene dependent evaluation. To the best of the authors' knowledge, the proposed work is a first attempt to establish a benchmark evaluation framework with clearly defined data-division schemes for both SDE and SIE setups.

III. Proposed Experimental Frameworks for Change Detection

As discussed earlier, we present two standard experimental frameworks for change detection evaluation over CDnet 2014, LASIESTA and SBMI2015 datasets. We also performed cross-dataset evaluation in PTIS dataset. We discuss the need for such standardized evaluation frameworks for change detection. Furthermore, we give a detailed description of each of these frameworks in the following subsections.

A. Need for Benchmark Evaluation Frameworks for Deep Learning-based Change Detection

Various CNN models for change detection applications have been proposed in the literature. These approaches have claimed very high quantitative performance as compared to traditional techniques. However, we noticed a clear inconsistency among the results reported in the literature [33-47] in terms of evaluation frameworks adopted. For traditional unsupervised methods, the experimental setup is clearly defined [5-7, 16-18, 48] which makes these results easily comparable. However, in deep learning methods, it is essential to maintain nonoverlapping (independence) between the training and testing sets. Moreover, it is highly desired that the training and testing data should not be similar (i.e. scene dependent). These factors have not been carefully considered in the existing literature. For example, although temporal information forms the basis of change detection, the highest results claimed in [34, 41] do not even consider it in their respective CNN models. They use a carefully selected set of frames (50/200 frames) from each video to train the model and achieve more than 98% F-score over CDnet 2014 dataset. We argue that such evaluation is clearly overfitted as the training and testing data is almost the same. Moreover, as mentioned by the authors, their approach is meant for the generation of ground truth labels for video sequences. Such models are not suitable to handle the challenges in change detection in unseen videos. Therefore, these results are not comparable to other approaches. Other methods have apparently adopted different schemes to collect train data. However, the exact video frame segregation (category-wise, video-wise) details are not provided in these papers. There is a lack of standardized experimental setting for evaluation. Thus, we raise the question, *how can these methods be compared directly*? Therefore, in this manuscript, we proposed clearly defined evaluation frameworks for change detection experiments.

B. Scene Dependent Evaluation (SDE)

As discussed earlier, to establish a benchmark with detailed data segregation for scene dependent strategy, we present the SDE framework. In SDE, training and testing sets contain frames from the same video. For example, if there are 1000 frames in a video, then certain portion (i.e. 50%, the initial 500 frames) is segregated as training and the remaining 500 frames are kept for testing. In addition, the evaluation is performed over the entire video (all 1000 frames) as well. Moreover, the SDE is performed over two different experimental setups: category-wise and complete dataset training for CDnet 2014 dataset.

1) Category-wise training in SDE: In this setup, the model is trained on each individual category, i.e., bad weather, baseline, etc. Such types of models can be useful for dealing with specific applications/scenarios such as extreme weather conditions, dynamic background, night videos, etc. The 50% frames collected from the videos of the individual scene category is used in model training.

2) Complete dataset training in SDE: In this setup, the model gathers experience from multiple visual change scenarios by training on the complete dataset. The 50% frames collected from each video is used in training. This leads to better robustness as compared to category-wise trained models to handle different challenging scenarios.

In LASIESTA dataset, we divide the labeled frames in each video into 50%-50% sets. The initial 50% frames from 20 videos are used for training. Similarly, initial 50% frames from 13 videos are selected from SBMI2015 dataset.

Most of the existing deep learning-based change detection approaches have followed similar strategies as discussed above. Since the *background features* remain more or less the same for all frames in a video sequence. Thus, SDE evaluations may not necessarily reflect the actual potency of the model as the high performance is clearly due to the high similarity between the training and testing data.

C. Scene Independent Evaluation (SIE)

In SIE, the training and testing sets contain a completely different set of videos. There is no similarity between the background features of train and test videos. A leave-one-video-out (LOVO) strategy is followed to group videos into testing and training sets. For example, if a category such as *baseline* contains 4 videos, then 3 videos are selected for training and the remaining video is selected for testing. Such an experimental setup makes the model design much more challenging as compared to the SDE setup. SIE is also performed over two different setups: category-wise and complete dataset training for CDnet 2014 dataset.

1) Category-wise training in SIE: In this setup, the model is trained for each category separately. The model is evaluated to verify its effectiveness in a particular type of scenario. Such evaluation is suitable for application specific tasks. For example, the model collected from 'bad weather' category would be suitable for deployment in similar scenarios.

2) Complete dataset training in SIE: In order to validate the generalization capability of the designed network for better performance in completely unseen videos, the model is also trained over the complete dataset. The model is shown a variety of scenes to learn the robust patterns for change detection in real-world scenarios.

In LASIESTA dataset, 10 of the 20 videos are used in training. The evaluation is done over the remaining 10 completely unseen videos. Similarly, 9 of the 13 videos are for training and the remaining 4 unseen videos are used for testing in SBMI2015 dataset. In addition, we use the videos from PTIS dataset for cross-dataset evaluation to further analyze the generalization capability of deep learning models.

IV. PROPOSED CHANGEDET NETWORK

We propose a new end-to-end ChangeDet network which takes multiple inputs to assimilate both temporal and spatial features through several building blocks to achieve robust performance in scene independent evaluation. All the building blocks of ChangeDet architecture is depicted in Fig. 2. We give a detailed description of the different blocks of ChangeDet in the following subsections. Moreover, feature map visualizations of different blocks are qualitatively analyzed for a more intuitive understanding of the network. We also discussed the importance of each block through ablation analysis in the experimental section

A. Depth Reductionist Background Estimation (DRBE)

In ChangeDet, we first estimate the background from recent history frames. Thereafter, we identify the motion information by comparing estimated background with the current frame. For background estimation, we designed DRBE block which is completely trainable as a part of the end-to-end ChangeDet network as shown in Fig. 2. In DRBE, the background is learned through a sequence of maximum multi-spatial receptive feature (MMSR) blocks using recent temporal history. Each MMSR stage captures the maximum response from multiple receptive fields of size 1×1, 3×3 and 5×5. The intuition behind using multiple filter sizes stems from theoretical propositions and corresponding experimental success of algorithms presented in [5, 6, 8, 16-18, 31, 49-51]. In [5, 6, 8] pixel-based background model maintenance and update policies were proposed. In [18, 49-51], local patterns extracted from 3×3 region provided the discriminative texture features to capture background statistics. Furthermore, methods in [16, 17] rely on both spatiotemporal binary features as well as pixel-level intensities to detect

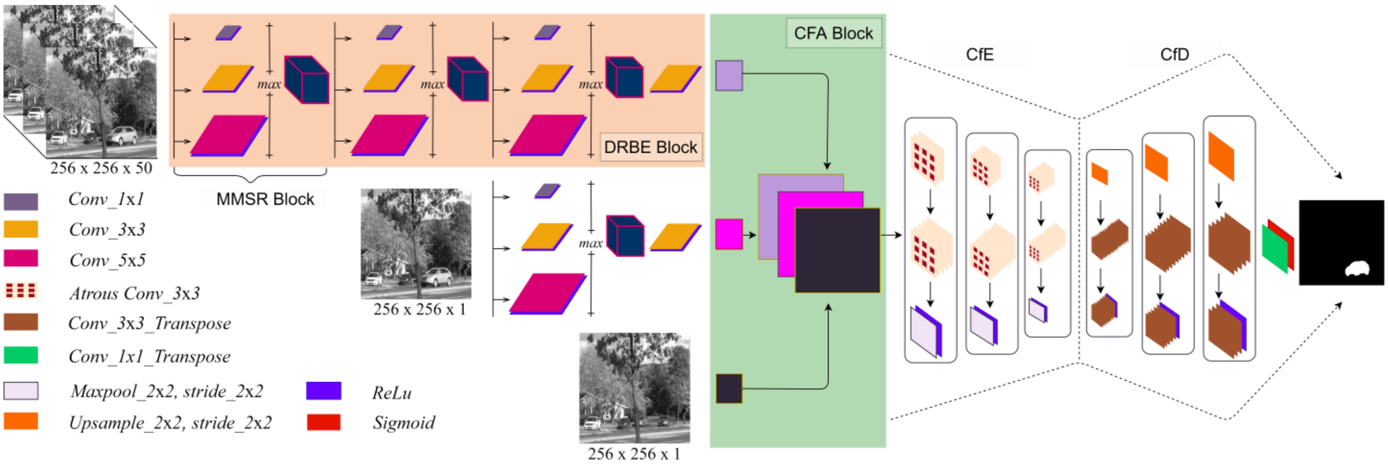

Fig. 2. The proposed ChangeDet network architecture. The DRBE block takes past temporal history as input and estimates the background through the progressive reduction of the depth using multiple MMSR blocks. Similarly, multi-receptive features are extracted from the current frame as well. A temporal median of the past 50 frames is also computed. Thereafter, in the CFA block, the contrasting features from the foreground and background features are combined and processed through an encoder-decoder (CfE-CfD) to generate the change detection map.

changes. The binary features extracted from 5×5 region and pixel-level intensities together characterize the change information more robustly in a nonparametric paradigm. Thus, in order to mimic similar features, we proposed MMSR to incorporate responses from three different levels of receptive fields. Moreover, by taking the maximum among the three responses, we ensure adaptability in the network to handle different change scenarios. These MMSR blocks with decreasing feature depths selectively determine the probable salient background representation. Finally, through these reductionist stages, we estimate a single depth background map. The sample visualizations for each MMSR block is depicted in Fig. 3.

Let's define a convolutional kernel as $\kappa_{n,h,w}$ where the parameters *h*, *w*, *n* represent the height, width and number of kernels respectively. The stack of past temporal history is denoted as P_T having height, width, and the number of frames as *H*, *W*, and *T* respectively. We compute $DRBE_T$ features using Eq. (1)-Eq. (5).

$$DRBE_T = \zeta_4(\zeta_3(\zeta_2(\zeta_1(P_T)))) \tag{1}$$

$$\zeta_1(z) = \arg\max_i \left(\Re(\kappa_{32,2i-1,2i-1} \otimes z)\right); i \in [1,3] \tag{2}$$

$$\zeta_2(z) = \arg\max_i \left(\Re(\kappa_{16,2i-1,2i-1} \otimes z)\right); i \in [1,3] \tag{3}$$

$$\zeta_3(z) = \arg\max_i \left(\Re(\kappa_{8,2i-1,2i-1} \otimes z)\right); i \in [1,3] \tag{4}$$

$$\zeta_4(z) = \Re(\kappa_{1,3,3} \otimes z) \tag{5}$$

where $\otimes$ denotes the convolution operation, z is an intermediate variable, *stride=1* and $\Re(\cdot)$ is the rectified linear unit (ReLu) activation function. We also fortify the background features by assimilating the estimated background in Eq. (1) with a pixel-wise temporal median (M_T) of P_T. This enhances the robustness of the background model estimated from DRBE block.

B. Contrasting Feature Assimilation (CFA)

The contrasting features between the background model (DRBE) and the current frame characterize the change information. In order to delineate semantically accurate foreground shape representations, we propose to assimilate current frame features with the estimated background. To this end, we first feed current frame *I* through a single MMSR block and compute contemporary features (CF) using Eq. (6).

$$CF = \zeta_4(\zeta_3(I)) \tag{6}$$

Finally, these backgrounds and contemporary features which consist of independently learned background and foreground information are combined for further processing. The CFA is computed as given in Eq. (7).

$$CFA = [DRBE, M_T, CF] \tag{7}$$

where M_T is the pixel-wise temporal median.

C. Contrasting Feature-based Encoder-Decoder (CfE-CfD)

The assimilated contrasting features are further refined through an encoder-decoder network to generate the final segmentation map. The proposed encoder (*CfE)* aims to capture foreground and background context from both low-level to high-level abstractions through multiple convolutional and max-pooling layers. It is able to efficiently gather features with atrous convolution [32] window using learnable weights. In particular, the *CfE* is built as a chain of three blocks, each consisting of two consecutive atrous convolutions and a single max-pooling layer. As shown in Fig. 2, each subsequent block takes the output of the previous max-pooling layer as input. To robustly learn the motion information, the kernel depth is doubled for every consecutive block in *CfE*. The *CfE* block can be defined through Eq. (8)-Eq. (9).

$$CfE = En_3(En_2(En_1(CFA))) \tag{8}$$

$$En_j(z) = \Re(mp_{2,2}(\kappa'_{2^{j+3},h,w} \otimes (\kappa'_{2^{j+2},h,w} \otimes z))) \tag{9}$$

where $mp_{2,2}$, $\kappa'_{n,h,w}$ denote max pooling and atrous convolution. The max pool is applied with *stride=2*. For atrous convolution, the dilation rate is set to (2, 2). For all the convolutional operators we used *h=3, w=3* with *stride=1*. The variable *j* is used to calculate the kernels depths.

The proposed decoder *(CfD)* assist in reconstructing the foreground appearance features to match the original image size. The *CfD* component consists of three blocks, each consisting of an upsample layer followed by 2 consecutive transposed convolution layers. The subsequent block takes the

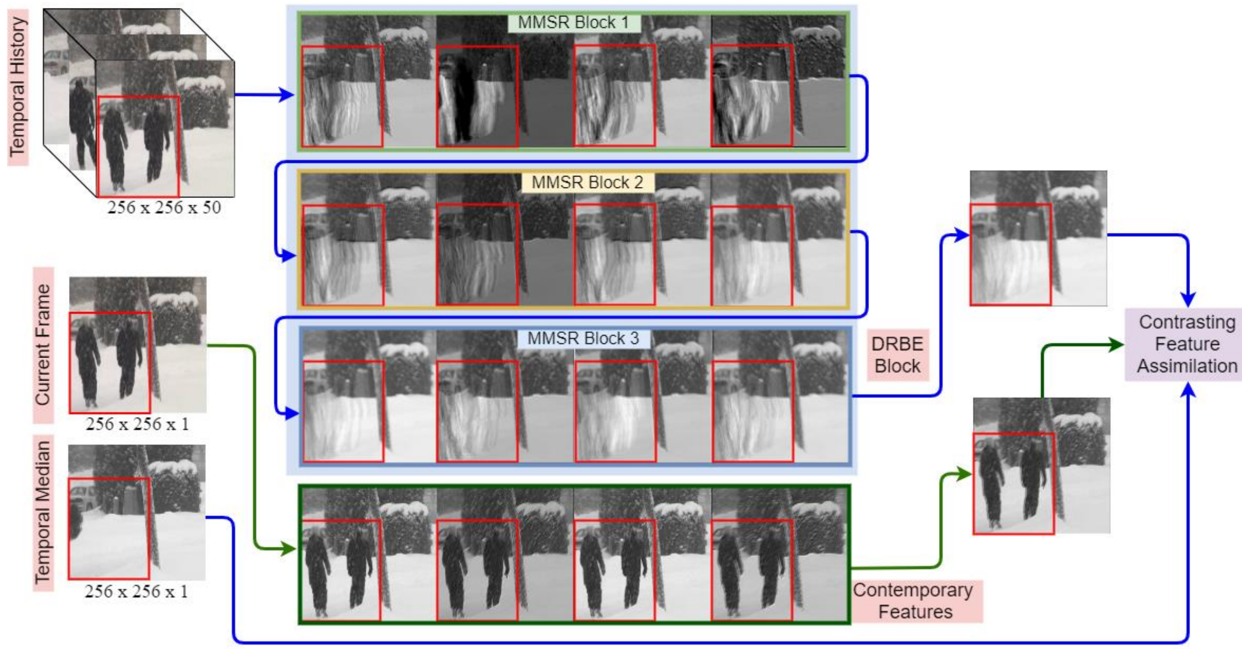

Fig. 3. Feature visualization at different stages of ChangeDet for network behavior analysis. The upper 3 rows depict the visualization of the three MMSR blocks present in the DRBE block. In each row, we show 4 representative feature maps for visual analysis. The 4th row depicts visualization of the convolutional features for the current frame (refer 2nd row in Fig. 2). It can be noticed that a gradually more refined background is estimated. The temporal median is also able to estimate a raw background although it is noisy in some regions. The last column shows that overall, quite robust contrasting features are gathered for refined change detection by the subsequent encoder-decoder blocks.

output of the previous transpose convolutional layer as input. The feature map depth is gradually reduced by using a number of kernels in reverse order as used in the *CfE* block. The *CfD* is computed as given in Eq. (10)-Eq. (11).

$$CfD = De_3(De_2(De_1(CfE))) \tag{10}$$

$$De_j(z) = \Re(\kappa^T_{2^{6-j},h,w} \otimes (\kappa^T_{2^{7-j},h,w} \otimes up_{2,2}(z)) \tag{11}$$

where $up_{2,2}$, $\kappa^T_{n,h,w}$ denote up-sampling and transposed convolution respectively. The up-sample operations are applied with the up-sample rate of 2. For all the transpose convolutional operations, we used *h=3*, *w=3* with *stride=1*. The final foreground probability map is predicted as given in Eq. (12).

$$FG = \delta(\kappa^T_{1,1,1} \otimes CfD) \tag{12}$$

where $\delta(\cdot)$ denote the sigmoid function.

The entire network can be efficiently trained in an end-to-end manner. The network is trained with a binary cross-entropy loss function. The same loss is backpropagated through *CFA* and *DRBE* blocks as well.

We also depict the feature map visualizations of the DRBE and CF blocks in Fig. 3. The successive MMSR blocks in DRBE gradually lead to enhanced background representations. The same can be verified in the first three rows of visualizations in Fig. 3. Moreover, the contemporary refined feature maps are depicted in the fourth row.

V. Empirical Study of SDE and SIE Strategies

A. *Why Scene Independency Matters in Change Detection?*

As discussed in previous sections, the deep learning models for change detection in videos can be evaluated either in SDE or SIE setups. However, the SDE setup leads to model optimization only for the same set of videos used in training. This is due to the fact that some frames from the test videos are used for training. Therefore, it is essential to evaluate the model over unseen or scene independent videos. This also makes the process of model design much more challenging in order to ensure robust performance even in real-world scenarios. Such SIE scheme ensures proper evaluation of the designed model as compared to SDE. More recent benchmark datasets for other video-based applications [54, 55] already ensure such scene independency in their evaluation schemes. Based on all these observations, our proposition is to give more importance to SIE over SDE for change detection model evaluation.

B. *Comparative Analysis of Proposed Versus Existing Models and Evaluation Schemes*

We also compare the proposed methods and evaluation schemes with existing approaches in terms of background estimation (BE), SIE, training data selection schemes and patch-based training (PP). The same is presented in Table I. In terms of the experimental setup, almost all existing methods [34-47] have adopted SDE schemes. To this end, temporal division with different proportions such as 50%, 70%, 90%, 5%, 50/100/150/200 frames, etc. have been selected for training. Lin et al. and Lim et al. [33, 36] have used a leave-one-video-out strategy to select training data. However, to increase test accuracy, the authors in [33] have selectively chosen frames containing more than 170 foreground pixels. Furthermore, authors in [36] conduct training with separate videos but present the overall results by combining the training and testing videos in their paper. They did not show results in the SIE setup. Therefore, we proposed clearly defined SIE setup to ensure robust performance analysis of CNN models.

In terms of BE, the approaches in [33, 36-40] are dependent either on statistical or non-parametric handcrafted approaches to extract the temporal features. The authors in [34, 35, 41] have

TABLE I
COMPARATIVE ANALYSIS OF PROPOSED CHANGEDET WITH EXISTING METHODS AND EVALUATION SCHEMES

Method	BE	PP	Training data Selection	SIE
Lin et al. [33]	SuBSENSE	Yes	Leave-one-video-out	Yes
Lim & Keles [34]	Frame-level segmentation	No	Selective (50/200 training frames)	No
Zeng & Zhu [35]	Frame-level segmentation	No	Random (150 training frames)	No
Lim et al. [36]	Designed background model	No	Leave-one-video-out	No
Nguyen et al. [37]	SuBSENSE	Yes	Random (100 training frames)	No
Babaee et al. [38]	SuBSENSE	Yes	Random (5% training frames)	No
Braham & Droogenbroeck [39]	IUTIS-5	Yes	50% training frames	No
Patil et al. [40]	Temporal histogram	No	Random	No
Wang et al. [41]	Frame-level segmentation	Yes	Selective (50/200 training frames)	No
Chen et al. [42]	CNN	No	50% training frames	No
Yang et al. [43]	CNN	No	90% training frames	No
Patil & Murala [44]	CNN	No	Random	No
Akilan et al. [45]	CNN	No	70% training frames	No
Bakkay et al. [46]	Generator (CNN)	No	50% training frames	No
Patil & Murala [47]	Generator (CNN)	No	Selective (200 training frames)	No
ChangeDet (SDE)	**End-to-end CNN**	**No**	**50% training frames**	**No**
ChangeDet (SIE)	**End-to-end CNN**	**No**	**Leave-one-video-out**	**Yes**

BE: Background estimation, PP: Partitioned image-based training

just performed frame-level segmentation without considering the historical context. In [42-47], background features are estimated with the CNN network. In ChangeDet, we proposed DRBE block to model temporal features from the recent history for effective BE in an end-to-end manner.

Certain methods [33, 37-39, 41] have first partitioned the video frames into patches for training the network. This requires additional preprocessing at both training and inference time. However, in this paper, we considered the complete image as input to the network as in [42-47].

VI. EXPERIMENTAL SETTINGS, RESULTS, AND DISCUSSIONS

A. Experiment Settings and Dataset

1) Implementation details: The proposed network is implemented in Keras with Tensorflow backend. The ChangeDet takes three tensors of shape $256\times256\times50$ (past temporal history), $256\times256\times1$ (current frame) and $256\times256\times1$ (temporal median) as input. We use T=50 historical frames to model the background which can be changed according to the application requirement.

2) Training configuration: Training is done with batch size=1 over Nvidia Titan Xp GPU system. We use a stochastic gradient descent optimizer with binary cross-entropy loss function to train the network. The initial learning rate is set to 0.0006 which is further decreased by 0.0002 after every 20 epochs. The minimum learning rate is set to 0.0001. We did not use any data augmentation for training.

3) Dataset: In our experiments, the benchmark CDnet 2014 [48], LASIESTA [52], SBMI2015 [65] dataset is used for performance evaluation. The CDnet dataset consists of 53 videos from a diverse set of realistic scenarios grouped into eleven different categories. Each category of videos presents a unique set of challenges. For example, dynamic changes in the background, bad weather conditions, illumination variations, shadows, and irregular object movements. In our experiments, we exclude the PTZ category due to excessive camera motion. Approximately, 89,000 frames are available for training and evaluation.

The LASIESTA consist of two different types of videos captured in indoor and outdoor scenarios. The videos are characterized with different motion type and intensity. There are 12 indoor and 8 outdoor videos. About 8,575 labeled frames are available for analysis. Similarly, SBMI2015 dataset has 13 challenging videos. Approximately, 5,023 annotated frames are available for performance evaluation. The PTIS [70] dataset consists of nine videos collected from both indoor and outdoor scenarios for change detection analysis.

B. Quantitative Analysis

The quantitative performance is measured in terms of F-score which is a comprehensive indicator of performance for change detection. Furthermore, we compare our proposed method with recent state-of-the-art (deep learning and non-deep learning) change detection methods in both SIE and SDE setups.

1) The problem of noncomparability among results of existing deep learning approaches:

Other than the problem of scene dependence as discussed earlier, we observed another issue of noncomparability among the existing deep learning-based results. Since the existing approaches have followed different strategies for training-testing data splitting, they are not comparable to each other. In fact, the highest F-score is claimed by manually selecting a particular set of frames to train the model and then test over the entire video [34, 41]. These results are not directly comparable to other methods. Therefore, we have also implemented three existing methods, FgSegNet-S [34], FgSegNet-M [34] and MSFS [64] in the same SDE setup and present baseline results for the same. We also compute the SIE results for these methods for fair comparative analysis.

2) CDnet 2014& CDnet 2012

Experimental results in the SDE framework: We conduct experiments in the SDE framework and compare the proposed method with existing state-of-the-art approaches in Table II. We also evaluate the SDE results of our model in different setups. We train the model with 50% frames and evaluate with the remaining 50% frames. For a comparative analysis with existing methods, we also evaluate these models with the complete 100% frames. Furthermore, we present the results of ChangeDet in both category-wise and complete dataset training.

As shown in Table II, the proposed ChangeDet outperforms the best performing handcrafted method by 7%. Our model also

TABLE II
COMPARATIVE F-SCORE PERFORMANCE IN SDE FRAMEWORK ON CDNET 2014 AND CDNET 2012 DATASETS

Methods	Train Data	Test data	BW	BA	CJ	DB	IOM	NV	LFR	SD	TH	TB	*Avg (cd12)*	*Avg (cd14)*
PAWCS [26]	0	100%	0.81	0.94	0.81	0.89	0.78	0.42	0.64	0.89	0.83	0.77	0.85	0.78
UBSS [10]	0	100%	0.79	0.89	0.87	0.78	0.76	0.53	0.62	0.78	0.79	0.47	0.81	0.73
WeSamBE [7]	0	100%	0.85	0.94	0.80	0.74	0.74	0.53	0.69	0.90	0.80	0.83	0.82	0.78
ViBe [6]	0	100%	0.77	0.88	0.45	0.72	0.47	0.40	0.33	0.83	0.55	0.61	0.65	0.60
SemBGS [29]	0	100%	0.83	0.96	0.84	0.95	0.79	0.50	0.79	0.95	0.82	0.69	0.88	0.81
IUTIS-5 [22]	0	100%	0.83	0.96	0.83	0.89	0.73	0.51	0.79	0.91	0.83	0.85	0.86	0.81
SuBSENSE [17]	0	100%	0.86	0.95	0.77	0.79	0.63	0.50	0.64	0.90	0.71	0.89	0.79	0.77
DeepBS [38]	RS	100%	0.86	0.96	0.90	0.88	0.61	0.64	0.59	0.93	0.76	0.90	0.84	0.80
MSFgNet [44]	RS	100%	0.85	0.92	0.83	0.85	0.78	0.81	0.84	0.93	0.80	0.86	0.85	0.85
VGG-PSL-CRF [42]	50%	100%	0.89	0.96	0.94	0.74	0.75	0.75	0.62	0.91	0.85	0.92	0.86	0.83
Cascade CNN [41]	S50	100%	0.79	0.97	0.97	0.95	0.87	0.87	0.74	0.95	0.89	0.84	0.93	0.88
EDS_CNN [36]	LOVO	100%	0.87	0.96	0.89	0.91	0.88	0.77	0.93	0.85	0.80	0.76	0.88	0.86
MCSCNN [62]	S5	100%	0.86	0.94	0.79	0.88	0.77	0.79	0.73	0.92	0.88	0.88	0.86	0.84
DPDL40 [60]	S40	100%	0.87	0.97	0.87	0.87	0.87	0.61	0.71	0.94	0.84	0.76	0.89	0.83
MSRNN [57]	NA	100%	0.89	0.96	0.92	0.91	0.87	0.56	0.84	0.95	0.85	0.80	0.91	0.85
LTDP [56]	0	100%	0.67	0.95	0.81	0.82	0.73	0.54	0.76	0.9	0.79	0.89	0.83	0.79
SBSN [58]	NA	100%	0.92	0.95	0.89	0.79	NA	0.77	NA	0.86	0.86	0.73	-	-
REDN [61]	S200	100%	0.78	0.97	0.93	0.86	0.8	0.79	0.73	0.88	0.84	0.88	0.88	0.85
FgSegNet-S_51 [34]	50%	100%	0.79	0.86	0.90	0.78	0.78	0.79	0.32	0.81	0.83	0.58	0.83	0.74
FgSegNet-S_55 [34]	50%	50%	0.77	0.84	0.84	0.74	0.70	0.82	0.33	0.78	0.72	0.57	0.77	0.71
FgSegNet-M_51 [34]	50%	100%	0.73	0.93	0.83	0.76	0.75	0.77	0.32	0.82	0.79	0.60	0.81	0.73
FgSegNet-M_55 [34]	50%	50%	0.72	0.92	0.76	0.69	0.61	0.81	0.32	0.82	0.73	0.57	0.76	0.70
MSFS_51 [64]	50%	100%	0.85	0.90	0.91	0.55	0.68	0.87	0.61	0.94	0.91	0.70	0.81	0.79
MSFS_55 [64]	50%	50%	0.80	0.89	0.88	0.50	0.70	0.82	0.55	0.93	0.86	0.67	0.79	0.76
ChangeDet_cat_51	50%	100%	0.92	0.90	0.83	0.84	0.88	0.85	0.82	0.91	0.92	0.93	0.88	0.88
ChangeDet_com_51	50%	100%	0.94	0.93	0.84	0.91	0.92	0.83	0.83	0.91	0.74	0.91	0.88	0.88
ChangeDet_cat_55	50%	50%	0.85	0.90	0.77	0.83	0.77	0.82	0.80	0.88	0.89	0.92	0.84	0.84
ChangeDet_com_55	50%	50%	0.91	0.93	0.79	0.91	0.85	0.82	0.83	0.90	0.75	0.89	0.86	0.86

LOVO: Leave-one-video-out, RS: Random Selection, SX: Selected X frames; ChangeDet_cat_51: Category-wise training with 50% frames used for training and 100% for testing; ChangeDet_com_51: Complete dataset training with 50% frames used for training and 100% for testing; ChangeDet_cat_55: Category-wise training with 50% frames used for training and remaining 50% for testing; ChangeDet_com_55: Complete dataset training with 50% frames used for training remaining 50% for testing;
NA: Data not available; cd14: CDnet 2014; cd12: CDnet 2012

outperforms recent state-of-the-art deep learning models DeepBS [38], MSFgNet [44], SFEN (VGG) [42], VGG + CRF [42], VGG + PSL + CRF [42], GoogLeNet + PSL + CRF [42], EDS_CNN [36] by 8%, 3%, 15%, 13%, 5%, 14%, 2% respectively in CDnet 2014. It also outperforms most of the existing methods in CDnet 2012 videos. The overall F-score of the proposed method is equal to Cascade CNN [41]. However, we noticed multiple issues with Cascade CNN: model is trained for each video separately, training frames are selected manually and images are processed into small-patches. In addition, the network only learns the spatial features (single image as input) without considering the temporal features (past history). Thus, the results for Cascade CNN are highly optimized for scene videos. Our proposed ChangeDet is an end-to-end network which incorporates both spatial and temporal features for decision making.

We trained the existing deep learning models (FgSegNet-S, FgSegNet-M, and MSFS) in the same SDE setup for a fair comparative analysis. The proposed ChangeDet outperforms FgSegNet-S, FgSegNet-M, MSFS by a margin of 14%, 15%, 9%, respectively in CDnet 2014. Similarly, it obtains 5%, 7%, 7% improvement over FgSegNet-S, FgSegNet-M, MSFS, respectively in CDnet 2012 videos.

Experimental results in the SIE framework: In order to evaluate the generalization capability of the proposed ChangeDet, we also present experimental results for category-wise and complete dataset training in the SIE setup. We compared our work with existing state-of-the-art methods in Table III. We also trained and evaluated the existing deep learning methods FgSegNet-S, FgSegNet-M, and MSFS in the same SIE setup to present an empirical comparative analysis. The proposed ChangeDet outperforms FgSegNet-S, FgSegNet-M, MSFS by 38%, 46%, 34%, respectively in CDnet 2014. It also achieves 38%, 41%, 32% performance improvement over FgSegNet-S, FgSegNet-M, MSFS, respectively in CDnet 2012 dataset videos.

3) LASIESTA

Experimental results in the SDE framework: The comparison of various methods in SDE setup in terms of average F-score in each video category is shown in Table IV. From quantitative analysis (see in Table IV), it is evident that proposed ChangeDet outperforms in eight out of twelve categories of LASIESTA for foreground detection. The overall F-score of proposed ChangeDet (0.87) is significantly improved from 0.79 (highest value from existing methods). Moreover, the ChangeDet obtains 54%, 52%, 47% better F-score as compared to the deep learning methods FgSegNet-S, FgSegNet-M, MSFS, trained and evaluated in the same SDE setup for fair evaluation.

Experimental results in the SIE framework: For the SIE setup, we evaluate the effectiveness of proposed ChangeDet in 12 completely unseen videos as shown in Table V. The results of the proposed ChangeDet is compared with existing state-of-the-art methods. Our model comfortably outperforms the existing handcrafted and deep learning approaches for change detection. More specifically, the proposed method outperforms the deep

TABLE III
COMPARATIVE F-SCORE PERFORMANCE IN SIE FRAMEWORK ON CDNET 2014 AND CDNET2012 DATASETS.

Method	SIE	BL	PE	SW	BO	PA	TP	WS	BS	CO	T1	Avg (cd12)	Avg (cd14)
SuBSENSE [17]	Yes	0.85	0.95	0.81	0.69	0.48	0.85	0.45	0.86	0.91	0.79	0.78	0.77
VIBE [6]	Yes	0.53	0.90	0.30	0.22	0.26	0.60	0.43	0.67	0.75	0.58	0.52	0.52
PAWCS [26]	Yes	0.66	0.95	0.74	0.88	0.21	0.91	0.46	0.86	0.65	0.68	0.72	0.70
IUTIS-5 [22]	Yes	0.80	0.97	0.81	0.75	0.65	0.89	0.60	0.87	0.90	0.63	0.82	0.79
UBSS [10]	Yes	0.86	0.96	0.90	0.90	0.62	0.89	0.49	0.87	0.92	0.54	0.86	0.80
WeSamBe [7]	Yes	0.86	0.96	0.85	0.64	0.41	0.91	0.59	0.86	0.89	0.71	0.77	0.77
SemBGS [29]	Yes	0.84	0.98	0.85	0.98	0.69	0.88	0.61	0.92	0.82	0.30	0.87	0.79
DeepBS [38]*	No	0.61	0.95	0.90	0.81	0.60	0.49	0.75	0.94	0.89	0.77	0.85	0.77
MSFgNet [44]*	No	0.85	0.94	0.73	0.80	0.87	0.91	0.67	0.89	0.80	0.76	0.84	0.82
BSUV-Net [63]	Yes	0.82	0.97	0.69	0.89	0.91	0.91	0.61	0.94	0.83	0.66	0.87	0.82
BSUV-Net+SemBGS [63]	Yes	0.82	0.97	0.71	0.91	0.91	0.91	0.62	0.96	0.82	0.65	0.88	0.83
BMN-BSN [59]	Yes	0.84	0.96	0.63	0.95	0.77	0.72	0.70	0.92	0.90	0.56	0.86	0.80
FgSegNet_S [34]	Yes	0.74	0.65	0.12	0.42	0.17	0.57	0.45	0.52	0.74	0.17	0.44	0.46
FgSegNet_M [34]	Yes	0.55	0.72	0.11	0.69	0.05	0.22	0.42	0.60	0.31	0.16	0.41	0.38
MSFS [64]	Yes	0.70	0.33	0.22	0.62	0.52	0.74	0.40	0.53	0.77	0.12	0.50	0.50
ChangeDet_cat	Yes	0.95	0.95	0.56	0.93	0.77	0.85	0.80	0.78	0.85	0.80	0.81	0.82
ChangeDet_com	Yes	0.93	0.84	0.76	0.95	0.80	0.92	0.79	0.80	0.79	0.81	0.82	0.84

ChangeDet_cat: Category-wise training in SIE setup; ChangeDet_com: Complete dataset training in SIE setup; cd14: CDnet 2014; cd12: CDnet 2012
*These video-wise results are taken from the original paper which do not follow SIE setup. These results are used just for reference to show some comparisons with deep learning techniques with our scene independent results.
BL: Blizzard (from Bad Weather), PE: Pedestrian (from Baseline), SW: Sidewalk (from Camera Jitter), PA: Parking (from Intermittent Object Motion), TP: Turnpike05fps (from Low Frame Rate), WS: Winter Street (from Night Videos), BS: BusStation (from Shadow), CO: Corridor (from Thermal), T1: Turbulence1 (from Turbulence).

TABLE IV
COMPARATIVE F-SCORE PERFORMANCE IN SDE FRAMEWORK ON LASIESTA DATASET

Method	ISI	ICA	IOC	IIL	IMB	IBS	OCL	ORA	OSN	OSU	Avg
Zovkovik [1]	0.91	0.83	0.95	0.24	0.87	0.53	0.88	0.83	0.38	0.71	0.71
Maddalena1 [66]	0.87	0.85	0.91	0.61	0.76	0.42	0.88	0.84	0.58	0.80	0.75
Maddalena2 [27]	0.95	0.86	0.95	0.21	0.91	0.40	0.87	0.85	0.81	0.88	0.77
Cuevas1 [67]	0.79	0.74	0.85	0.79	0.73	0.58	0.86	0.81	0.46	0.73	0.73
Haines [68]	0.89	0.89	0.92	0.85	0.84	0.68	0.83	0.86	0.17	0.86	0.78
Cuvas2 [69]	0.88	0.84	0.78	0.65	0.89	0.66	0.88	0.82	0.78	0.72	0.79
FgSegNet-S-51 [34]	0.32	0.57	0.37	0.33	0.64	0.21	0.17	0.10	0.08	0.27	0.31
FgSegNet-S-55 [34]	0.39	0.60	0.23	0.39	0.60	0.22	0.23	0.15	0.13	0.37	0.33
FgSegNet-M-51 [34]	0.44	0.71	0.29	0.32	0.68	0.27	0.24	0.17	0.18	0.21	0.35
FgSegNet-M-55 [34]	0.43	0.69	0.31	0.32	0.71	0.21	0.22	0.18	0.19	0.25	0.35
MSFS-51 [64]	0.44	0.60	0.30	0.32	0.50	0.22	0.31	0.24	0.28	0.38	0.36
MSFS-55 [64]	0.39	0.4	0.37	0.35	0.64	0.36	0.41	0.35	0.31	0.37	0.40
ChangeDet-51	0.91	0.74	0.92	0.92	0.89	0.88	0.92	0.87	0.46	0.83	0.83
ChangeDet-55	0.91	0.92	0.88	0.90	0.91	0.89	0.91	0.90	0.61	0.82	0.87

ISI: Simple sequences, ICA: Camouflage, IOC: Occlusions, IIL: Illumination changes, IMB: Modified background, IBS: Bootstrap, OCL: Cloudy condition, ORA: Rainy condition, OSN: Snowy condition, OSU: Sunny condition.

TABLE V
COMPARATIVE F-SCORE PERFORMANCE IN SIE FRAMEWORK ON LASIESTA DATASET

Method	ISI-02	ICA-02	IOC-02	IIL-02	IMB-02	IBS-02	OCL-02	ORA-02	OSN-02	OSU-02	Avg
Zovkovik [1]	0.89	0.75	0.91	0.31	0.80	0.52	0.82	0.80	0.24	0.88	0.69
Maddalena1 [66]	0.85	0.74	0.85	0.38	0.68	0.45	0.85	0.86	0.46	0.86	0.70
Maddalena2 [27]	0.94	0.87	0.95	0.23	0.85	0.40	0.88	0.86	0.71	0.88	0.76
Cuevas1 [67]	0.76	0.63	0.88	0.79	0.68	0.66	0.90	0.87	0.09	0.81	0.71
Haines [68]	0.81	0.87	0.95	0.81	0.71	0.73	0.96	0.90	0.04	0.90	0.77
Cuvas2 [69]	0.84	0.78	0.86	0.65	0.92	0.62	0.90	0.79	0.63	0.77	0.78
FgSegNet-S [34]	0.20	0.60	0.53	0.25	0.60	0.28	0.19	0.16	0.05	0.18	0.30
FgSegNet-M [34]	0.56	0.55	0.65	0.42	0.56	0.19	0.28	0.18	0.01	0.33	0.37
MSFS [64]	0.53	0.58	0.25	0.41	0.63	0.25	0.54	0.54	0.05	0.29	0.41
ChangeDet	0.83	0.66	0.89	0.87	0.77	0.82	0.90	0.85	0.51	0.81	0.79

learning methods FgSegNet-S, FgSegNet-M, MSFS by 49%, 42%, 38%, respectively, which highlights the superior generalization capability of our model.

4) SBMI2015

Experimental results in the SDE framework: The SDE results for SBMI2015 dataset is tabulated in Table VI. The proposed ChangeDet outperforms in six out of the thirteen videos of SBMI2015. The overall F-score of the proposed method (0.65) is 2% higher than the best performing existing methods (0.63). The existing deep learning methods have been trained and evaluated in the same SDE setup as the proposed method.

Experimental results in the SIE framework: To measure the generalization capability of the models, we evaluate the proposed and existing methods in four completely unseen

TABLE VI
COMPARATIVE F-SCORE PERFORMANCE IN SDE FRAMEWORK ON SBMI2015 DATASET

Method	Board	Cand	CAV1	CAV2	CaV	Fol	HAM	HigI	HigII	HB2	IBt2	PAF	Snel	Avg
FgSegNet-S-51 [34]	0.88	0.25	0.67	0.04	0.52	0.68	0.62	0.83	0.42	0.78	0.72	0.88	0.22	0.58
FgSegNet-S-55 [34]	0.89	0.35	0.71	0.18	0.69	0.38	0.70	0.68	0.19	0.82	0.82	0.86	0.29	0.58
FgSegNet-M-51 [34]	0.89	0.27	0.74	0.19	0.61	0.60	0.67	0.73	0.36	0.79	0.78	0.87	0.42	0.61
FgSegNet-M-55 [34]	0.89	0.21	0.70	0.05	0.57	0.91	0.71	0.75	0.31	0.83	0.83	0.90	0.52	0.63
MSFS-51 [64]	0.89	0.25	0.55	0.10	0.65	0.86	0.46	0.82	0.59	0.63	0.57	0.88	0.68	0.61
MSFS-55 [64]	0.91	0.26	0.57	0.08	0.57	0.80	0.52	0.82	0.58	0.61	0.60	0.87	0.68	0.61
ChangeDet-51	0.89	0.28	0.78	0.45	0.54	0.58	0.60	0.70	0.76	0.75	0.67	0.81	0.70	0.65
ChangeDet-55	0.88	0.41	0.82	0.38	0.53	0.41	0.63	0.71	0.77	0.73	0.59	0.91	0.71	0.65

Cand: Candela_m1.10, CAV1: CAVIAR1, CAV2: CAVIAR2, CaV: CaVignal, Fol: Foliage, HAM: HallAndMonitor, HigI: HighwayI, HigII: HighwayII, HB2: HumanBody2, IBt2: IBMtest2, PAF: PeopleAndFoliage, Sne: Snellen

TABLE VII
COMPARATIVE F-SCORE PERFORMANCE IN SIE FRAMEWORK ON SBMI2015 DATASET

Method	Cand	CAV2	CaV	HigII	Avg
FgSegNet-S [34]	0.23	0.11	0.68	0.24	0.31
FgSegNet-M [34]	0.15	0.14	0.72	0.21	0.30
MSFS [64]	0.27	0.10	0.63	0.58	0.40
ChangeDet	0.61	0.56	0.48	0.64	0.57

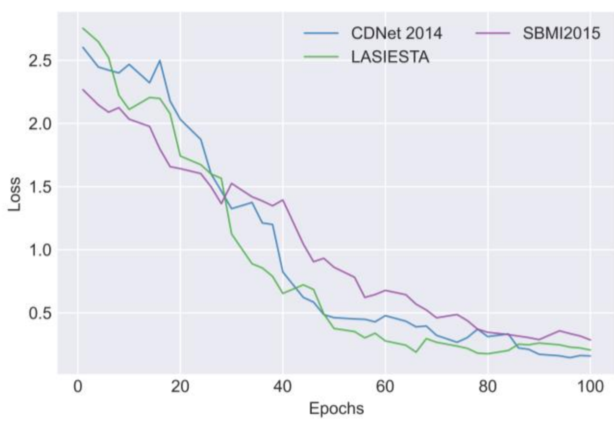

Fig. 4. The SIE training process of the proposed ChangeDet in CDnet 2014, LASIESA and SBMI2015 datasets

videos in SBMI2015. The comparative results in SIE setup is given in Table VII. The proposed ChangeDet outperforms the state-of-the-art methods in SIE setup as well. More specifically, it achieves an overall 17% performance improvement over the MSFS (0.40).

The iterative convergence of the proposed ChangeDet across the datasets CDnet 2014, LASIESTA and SBMI2015 is illustrated in Fig. 4. Similarly, the overall F-score in CDnet 2014, LASIESTA, SBMI2015 in SDE and SIE setups respectively is depicted in Fig. 5.

5) Cross-dataset Analysis

The PTIS dataset [70] is used for cross-data evaluation. The model trained on CDnet 2014 is used to test 8 videos from PTIS. Such cross-data analysis is helpful in validating the generalization capability of the deep learning models. The cross-data results of the proposed ChangeDet, existing deep learning models (FgSegNet-S, FgSegNet-M, MSFS) and other approaches are given in Table VIII. The proposed ChangeDet outperforms the existing state-of-the-art approaches. More specifically, our model outperforms FgSegNet-S, FgSegNet-M, MSFS by margin of 35%, 24%, 24% respectively.

C. Ablation Studies

1) Analysis of components in ChangeDet: In order to quantify the influence of the constituent modules (CF, M_T, CFA, and MMSR) in ChangeDet, we performed various experiments for ablation analysis in this section. We conduct experiments over two categories bad weather and baseline. We create multiple variants of the proposed method by dropping M_T (ChangeDetv2), replacing CF with the current frame (ChangeDetv3), completely removing CF (ChangeDetv4) and dropping an MMSR module from DRBE block (ChangeDetv5). The experimental results for all these variants are given in Table IX. We notice that removing either of the modules from ChangeDet results in lower performance in both categories. Moreover, different results for a single variant also shows that certain components of ChangeDet are more useful than others for different categories. Thus, the customizable nature of the proposed network is suitable for improved performance across multiple scene categories for change detection. This further justifies the original ChangeDet model design.

TABLE VIII
CROSS-DATA F-SCORE COMPARISON ON PTIS DATASET

Method	Airport	Bootstrap	Shoping Mall	Lobby	Escalator	Curtain	Water Surface	Fountain	Avg
RPCA [71]	0.71	0.68	0.73	0.78	0.64	0.82	0.66	0.72	0.69
OPRMF [72]	0.66	0.62	0.72	0.74	0.6	0.71	0.62	0.72	0.65
GRASTA [73]	0.62	0.58	0.71	0.61	0.57	0.68	0.76	0.69	0.63
PracPrProCS [74]	0.7	0.64	0.72	0.62	0.56	0.78	0.88	0.63	0.66
DECOLOR [75]	0.64	0.60	0.65	0.69	0.76	0.89	0.64	0.87	0.72
GOSUS [76]	0.66	0.62	0.72	0.81	0.84	0.68	0.84	0.72	0.71
OMoGMF+TV [77]	0.77	0.61	0.72	0.83	0.66	0.93	0.93	0.82	0.77
REDSA [78]	0.67	0.68	0.74	0.8	0.64	0.89	0.88	0.75	0.75
FgSegNet-S [34]	0.41	0.24	0.34	0.42	0.24	0.81	0.60	0.41	0.43
FgSegNet-M [34]	0.61	0.37	0.56	0.48	0.31	0.87	0.62	0.49	0.54
MSFS [64]	0.51	0.46	0.49	0.52	0.41	0.81	0.60	0.55	0.54
ChangeDet	0.84	0.68	0.75	0.74	0.69	0.90	0.94	0.70	0.78

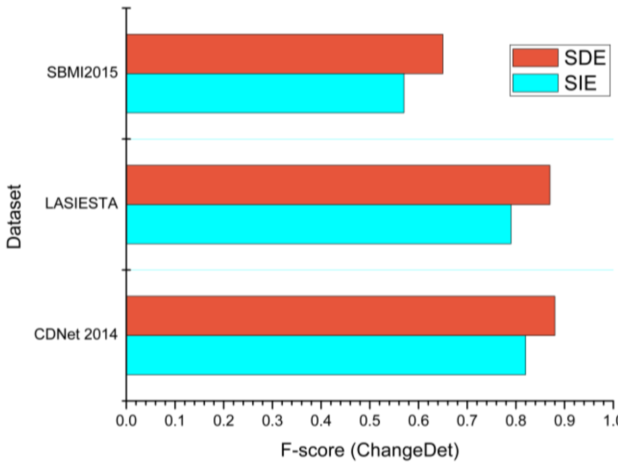

Fig. 5. The average F-score of the proposed ChangeDet for CDnet 2014, LASIESA, SBMI2015 in SDE and SIE setup respectively.

TABLE IX
ABLATION STUDIES OF PROPOSED CHANGEDET IN BAD WEATHER (BW) AND BASELINE (BL) CATEGORIES

Model	*Components*				*BW*	*BL*
	M_T	*CF*	*I*	*MMSR*		
ChangeDet	Y	Y	Y	Y	0.95	0.95
ChangeDetv2	N	Y	Y	Y	0.93	0.82
ChangeDetv3	Y	N	Y	Y	0.94	0.88
ChangeDetv4	Y	N	N	Y	0.90	0.88
ChangeDetv5	Y	Y	Y	N	0.91	0.85

M_T: temporal median, CF: contemporary features, I: current frame, MMSR: maximum multi-spatial receptive feature block

2) Memory and computation analysis: The proposed ChangeDet consist of 132.8K or 0.13M trainable parameters with a model size of 1.59 MB. The inference speed is 17 ms per frame or approximately 58.8 frames per second (FPS) over Titan Xp. We compare our ChangeDet with existing state-of-the-art change detection techniques for computational and speed comparison. The same is presented in Table X. It shows that our method is computationally efficient than most of the existing approaches. The methods in [39, 40, 44] seem to have a lower number of trainable parameters, however, our method achieved superior speed (58.8 FPS) which is highest amongst all other methods. The memory consumption of ChangeDet is much lower (only 1.59 MB) which makes it suitable for embedded devices used in real-time applications. Moreover, it can also be noticed that small/shallower networks [33, 39, 40, 44] including ChangeDet have an overall advantage over the large/deeper networks [34, 35, 36, 38, 42] in terms of overall performance (accuracy, compute efficiency and speed). Thus, an aptly designed small and shallow networks such as the proposed ChangeDet, which performs well in all three performance metrics is a valuable contribution for change detection applications.

D. Qualitative Analysis

The qualitative comparison of the visual examples selected from change detection results of the proposed ChangeDet and existing methods on multiple scene categories is depicted in Fig. 6. These video examples are selected to cover various complex visual scenarios such as intermittent object motion (1st row), night-time videos (2nd row), video with the low frame rate (3rd row), dynamic background movements (4th row) and

TABLE X
THE PROPOSED CHANGEDET IS COMPARED WITH EXISTING STATE-OF-THE-ART DEEP LEARNING-BASED CHANGE DETECTION TECHNIQUES FOR COMPUTATION AND SPEED ANALYSIS.

Method	*# Param*	*BMC*	**FPS*
FCSN [33]	~ 1.73 M	Yes	NA
FgSegNet [34]	~ 2.60 M	No	18
MFCN [35]	~ 20.83 M	No	27
EDS-CNN [36]	~ 18.64 M	Yes	NA
Trip-Net [37]	~ 0.33 M	Yes	NA
DeepBS [38]	~ 3.15 M	Yes	NA
Deep-ConvNet [39]	~ 4.40 K	Yes	NA
Msednet [40]	~ 2.88 K	Yes	9.7
Cascade CNN [41]	~ 0.25 M	No	13
VGG16 [42]	~ 31.92 M	No	4.9
GoogLeNet [42]	~ 6.02 M	No	NA
ResNet50 [42]	~ 23.78 M	No	NA
MSFgNet [44]	~ 5.40 K	No	NA
2D CNN-LSTM [45]	~ 0.29 M	No	15
3D CNN-LSTM [45]	~ 0.22 M	No	24
ChangeDet	0.13 M	No	58.8

Param: Number of trainable parameters; M: Millions; K: Thousands; NA: Not Available; BMC: Separate computation cost for background modeling.
Note: The number of parameters for existing approaches is approximated based on the information provided in the original papers respectively.
*FPS is reported as given in the original papers.

bad weather conditions (5th row). The qualitative responses of deep learning (DeepBS [38]) and non-deep learning (IUTIS-5 [22]) method is compared with the proposed ChangeDet. A robust model must be able to eliminate both false positives (FP) and false negatives (FN) across different categories. The existing approaches work well in some categories but suffer from either higher FP or FN in other. In contrast, our proposed method can separate the salient foreground object from the background and highlight them uniformly across different categories. This is due to robust background estimation under the spatiotemporal network design and contrasting feature assimilation. Therefore, the effective model design of ChangeDet enables it to exclude the background interferences or noise of various types in different categories leading to improved performance over other approaches.

VII. CONCLUSION

An empirical study of scene independent and scene dependent experimental setups for training and evaluation of CNN based change detection is presented in this paper. To standardize the evaluation of deep learning models, scene dependent and scene independent strategies are proposed. Moreover, the importance of scene independence is highlighted through experimental analysis. A fast and lightweight (speed-58.8 fps and model size-1.59 MB) end-to-end model ChangeDet is proposed which consists of intuitive modules for motion segmentation in video streams. All these modules are trained in an end-to-end manner to estimate the foreground probabilities. The significance of each constituent block in the network is analysis through ablation studies. Also, visualizations of these modules are depicted to demonstrate their feature learning capabilities. Experiments are conducted over all the proposed evaluation frameworks and comparison with existing state-of-the-art approaches are presented. For fair evaluation across different datasets, the existing deep learning methods are trained and evaluated in the same SDE, SIE and cross-dataset settings. To

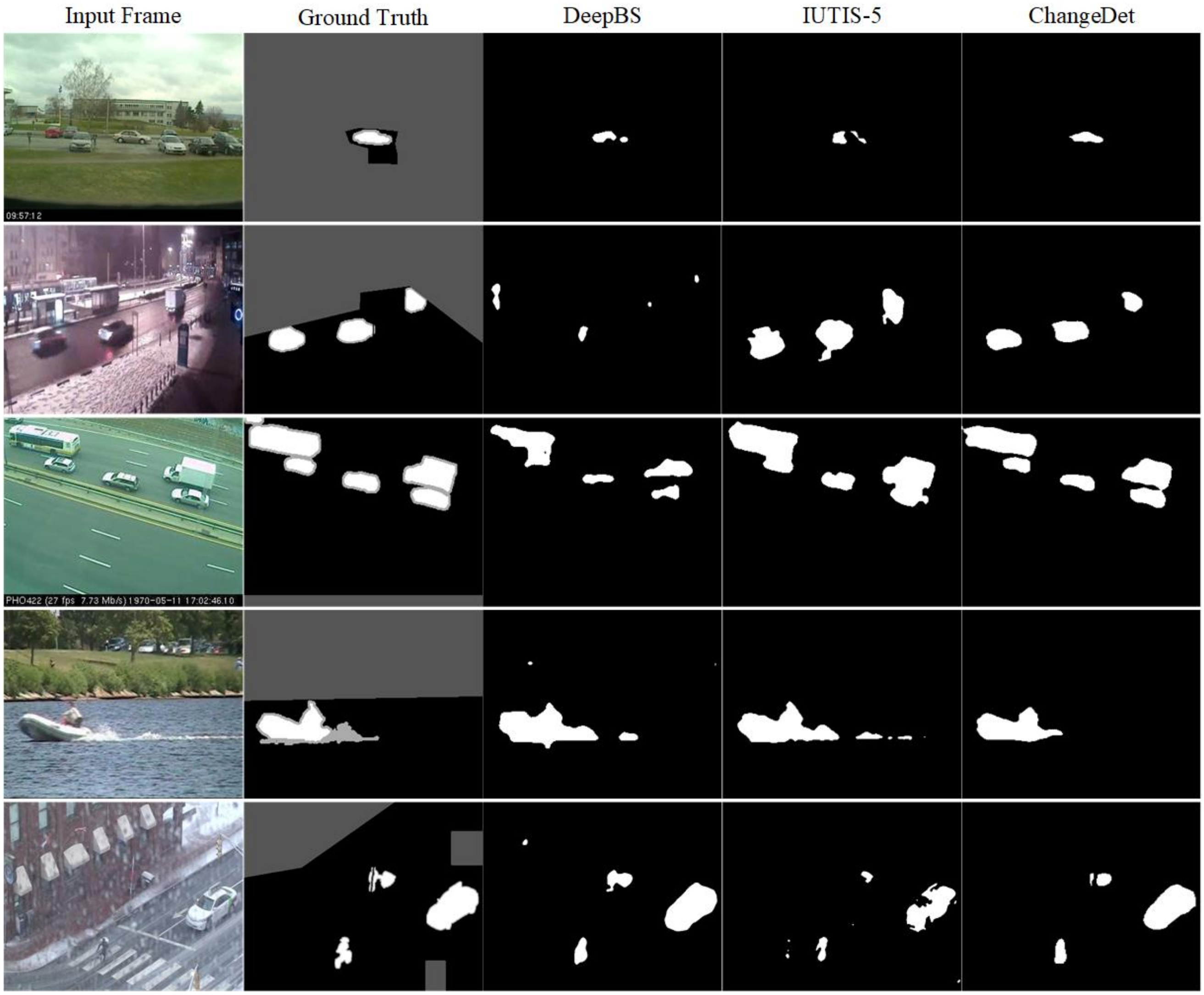

Fig. 6. Qualitative analysis of the proposed ChangeDet with existing state-of-the-art approaches.

the best of the authors' knowledge, the contribution of this paper in terms of establishing standard experimental setups for training and evaluation of supervised CNN models is a first attempt in change detection applications. In addition, the proposed ChangeDet network outperforms (overall, in terms of accuracy, speed, memory and compute efficiency) the existing state-of-the-art approaches over CDnet 2014, LASIESTA, SBMI2015 and PTIS in all the experimental setups.

Acknowledgement

The authors are highly grateful to IBM for providing with online GPU grant. The work was also supported by the DST-SERB project #EEQ/2017/000673. We thank all the members of the Vision Intelligence Lab for their support.

References

[1] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. IEEE Int. Conf. Pattern Recognit.*, pp. 28-31, 2004.

[2] S. Varadarajan, P. Miller and H. Zhou, "Spatial mixture of Gaussians for dynamic background modelling," in *Proc. IEEE Int. Conf. Adv. Video Signal Based-Surveill.*, pp. 63-68, 2013.

[3] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 246-252, 1999.

[4] R. Wang, F. Bunyak, G. Seetharaman and K. Palaniappan, "Static and Moving Object Detection Using Flux Tensor with Split Gaussian Models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 414-418, 2014.

[5] M. Hofmann, P. Tiefenbacher and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 38-43, 2012.

[6] O. Barnich and M. V. Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709-1724, 2011.

[7] S. Jiang and X. Lu, "WeSamBE: A weight-sample-based method for background subtraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2105-2115, 2018.

[8] M. Mandal, P. Saxena, S. K. Vipparthi and S. Murala, "CANDID: Robust Change Dynamics and Deterministic Update Policy for Dynamic Background Subtraction," in *Proc. IEEE Int. Conf. Pattern Recognit.*, pp. 2468-2473, 2018.

[9] J. Romero, M. Lado and A. Mendez, "A Background Modeling and Foreground Detection Algorithm using Scaling Coefficients Defined with a Color Model Called Lightness-Red-Green-Blue," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 243-1258, 2018.

[10] S. Hasan and S. C. S. Cheung, "Universal multimode background subtraction," *IEEE Trans. on Image Process.*, vol. 26, no. 7, pp. 3249-3260, 2017.

[11] K. Roy, J. Kim, M. Iqbal, F. Makhmudkhujaev, B. Ryu and O. Chae, "An adaptive fusion scheme of color and edge features for background subtraction," in *Proc. IEEE Int. Conf. Adv. Video Signal Based-Surveill.*, pp. 1-6, 2017.
[12] A. R. Rivera, M. Murhsed, J. Kim and O. Chae, "Background Modeling through Statistical Edge-Segment Distributions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1375-1387, 2013.
[13] S. Javed, S. Oh, A. Sobral, T. Bouwmans and S. Jung, "Background Subtraction via Superpixel-based Online Matrix Decomposition with Structured Foreground Constraints," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, pp. 90-98, 2015.
[14] M. Chen, Q. Yang, Q. Li, G. Wang and M. Yang, "Spatiotemporal GMM for Background Subtraction with Superpixel Hierarchy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1518-1525, 2017.
[15] Y. Chen, Z. Suna and K. Lam, "An Effective Sub-Superpixel-Based Approach for Background Subtraction," *IEEE Trans. Industrial Electronics*, to be published, doi: 10.1109/TIE.2019.2893824.
[16] P. L. St-Charles and G. A. Bilodeau, "Improving background subtraction using local binary similarity patterns," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pp. 509-515, 2014.
[17] P. L. St-Charles, G. A. Bilodeau and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359-373, 2015.
[18] M. Mandal, M. Chaudhary, S. K. Vipparthi, S. Murala, A. B. Gonde and S. K. Nagar, "ANTIC: ANTithetic Isomeric Cluster Patterns for Medical Image Retrieval and Change Detection," *IET Comput. Vis.*, vol. 13, no. 1, pp. 31-43, 2019.
[19] A. Elgammal, D. Harwood and L. Davis, "Non-parametric Model for Background Subtraction," in *Proc. European Conf. Comput. Vis.*, pp. 751-767, 2000.
[20] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1301-1306, 2010.
[21] H. Wang and D. Suter, "A consensus-based method for tracking: Modelling background scenario and foreground appearance," *Pattern Recognit.*, vol. 40, pp. 1091-1105, 2007.
[22] S. Bianco, G. Ciocca and R. Schettini, "Combination of video change detection algorithms by genetic programming," *IEEE Trans. Evolutionary Computation*, vol. 21, no. 6, pp. 914-928, 2017.
[23] S. Javed, S. H. Oh, A. Sobral, T. Bouwmans and S. K. Jung, "OR-PCA with MRF for robust foreground detection in highly dynamic backgrounds," in *Proc. Asian Conf. Comput. Vis.*, pp. 284-299, 2014.
[24] S. Javed, S. H. Oh, A. Sobral, T. Bouwmans and S. K. Jung, "Background–foreground modeling based on spatiotemporal sparse subspace clustering," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5840-5854, 2017.
[25] S. Javed, A. Mahmood, T. Bouwmans and S. K. Jung, "Spatiotemporal low-rank modeling for complex scene background initialization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1315-1329, 2016.
[26] P. L. St-Charles, G.-A. Bilodeau and R. Bergevin, "A Self-Adjusting Approach to Change Detection Based on Background Word Consensus," in *Proc. IEEE Wint. Conf. Applicat. Comput. Vis.*, pp. 6-9, 2015.
[27] L. Maddalena and A. Petrosino, "The SOBS algorithm: what are the limits?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 21 -26, 2012.
[28] G. Ramirez-Alonso and M. I. Chacon-Murguia, "Auto-Adaptive Parallel SOM architecture with a modular analysis for dynamic object segmentation in videos," *Neurocomputing*, vol. 175, pp. 990–1000, 2016.
[29] M. Braham, S. Piérard and M. V. Droogenbroeck, "Semantic background subtraction," in *Proc. IEEE Int. Conf. Image Process.*, pp. 4552-4556, 2017.
[30] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer Science Review*, vol. 11, pp. 31-66, 2014.
[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2818-2826, 2016.
[32] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, pp. 801-818, 2018.
[33] C. Lin, B. Yan and W. Tan, "Foreground Detection in Surveillance Video with Fully Convolutional Semantic Network," in *Proc. IEEE Int. Conf. Image Process.*, pp. 4118-4122, 2018.
[34] L. A. Lim and H. Y. Keles, "Foreground Segmentation Using a Triplet Convolutional Neural Network for Multiscale Feature Encoding," *Pattern Recognit. Lett.*, vol. 112, pp. 256:262, 2018.
[35] D. Zeng and M. Zhu, "Multiscale Fully Convolutional Network for Foreground Object Detection in Infrared Videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 617-621, 2018.
[36] K. Lim, W. D. Jang and C. S. Kim, "Background subtraction using encoder-decoder structured convolutional neural network," in *Proc. IEEE Int. Conf. Adv. Video Signal Based-Surveill.*, pp. 1-6, 2017.
[37] T. P. Nguyen, C. C. Pham, S. V. U. Ha and J. W. Jeon, "Change Detection by Training a Triplet Network for Motion Feature Extraction," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi:10.1109/TCSVT.2018.2795657.
[38] M. Babaee, D. T. Dinh and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognit.*, vol. 76, pp. 635-649, 2018.
[39] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *Proc. IEEE Int. Conf. Syst., Signal. Image Process.*, pp. 1-4, 2016.
[40] P. W. Patil, S. Murala, A. Dhall and S. Chaudhary, "Msednet: multi-scale deep saliency learning for moving object detection," in *Proc. IEEE Int. Conf. Systems Man Cybernet.*, pp. 1670-1675, 2018.
[41] Y. Wang, Z. Luo and P. M. Jodoin, "Interactive deep learning method for segmenting moving objects," *Pattern Recognit. Lett.*, vol. 96, pp. 66-75, 2017.
[42] Y. Chen, J. Wang, B. Zhu, M. Tang and H. Lu, "Pixel-wise deep sequence learning for moving object detection," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi:10.1109/TCSVT.2017.277031.
[43] L. Yang, J. Li, Y. Luo, Y. Zhao, H. Cheng and J. Li, "Deep Background Modeling Using Fully Convolutional Network," *IEEE Trans. Intell. Transportat. Syst.*, vol. 19, no. 1, pp. 254-262, 2018.
[44] P. W. Patil and S. Murala, "MSFgNet: A Novel Compact End-to-End Deep Network for Moving Object Detection," *IEEE Trans. Intell. Transportat. Syst.*, to be published, doi:10.1109/TITS.2018.2880096.
[45] T. Akilan, Q. J. Wu, A. Safaei, J. Huo and Y. Yang, "A 3D CNN-LSTM-Based Image-to-Image Foreground Segmentation," *IEEE Trans. Intell. Transportat. Syst.* To be published, doi:10.1109/TITS.2019.2900426.
[46] M. C. Bakkay, H. A. Rashwan, H. Salmane, L. Khoudour, D. Puigtt and Y. Ruichek, "BSCGAN: Deep Background Subtraction with Conditional Generative Adversarial Networks," in *Proc. IEEE Int. Conf. Image Process.*, pp. 4018-4022, 2018.
[47] P. W. Patil and S. Murala, "FgGAN: A Cascaded Unpaired Learning for Background Estimation and Foreground Segmentation," in *Proc. IEEE Wint. Conf. Applicat. Comput. Vis.*, pp. 1770-1778, 2019.
[48] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth and P. Ishwar, "CDnet 2014: an expanded change detection benchmark dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 387-394, 2014.
[49] M. Heikkila and M. Pietikäinen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 657-662, 2006.
[50] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1301-1306, 2010.
[51] H. Han, J. Zhu, S. Liao, Z. Lei and S. Z. Li, "Moving object detection revisited: Speed and robustness," *IEEE Trans. Cir. Syst. Video Technol.*, vol. 25, no. 6, pp. 910-921, 2014.
[52] C. Cuevas, E. M. Yez and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA," *Comput. Vis. Image Understand.*, vol. 152, pp. 103-117, 2016.
[53] P. Ochs, J. Malik, T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6., pp. 1187-1200, 2014.
[54] A. Milan, L. Leal-Taixe, I. Reid, S. Roth and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.
[55] W. Sultani, C. Chen and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 6479-6488, 2018.
[56] M. R. Arefin, F. Makhmudkhujaev, O. Chae, and J. Kim, "Background subtraction based on fusion of color and local patterns," in *Proc. Asian Conf. Comput. Vis.*, pp. 214-230, 2018.

[57] S. Choo, W. Seo, D. Jeong, and N. Ik Cho, "Learning background subtraction by video synthesis and multi-scale recurrent networks," in *Proc. Asian Conf. Comput. Vis.*, pp. 357-372, 2018.
[58] T. Minematsu, A. Shimada, and R. Taniguchi, "Simple background subtraction constraint for weakly supervised background subtraction network," in *Proc. IEEE Int. Conf. Adv. Video Signal Based-Surveill.*, pp. 1-8, 2019.
[59] V. Mondéjar-Guerra, J. Rouco, J. Novo, and M. Ortega, "An end-to-end deep learning approach for simultaneous background modelling and subtraction," in *Proc. British Mach. Vis. Conf.*, 2019.
[60] C. Zhao, T. L. Cham, X. Ren, J. Cai, and H. Zhu, "Background subtraction based on deep pixel distribution learning," in *Proc. IEEE Int. Conf. Multimed. Expo*, pp. 1-6, 2018.
[61] S. Choo, W. Seo, D. Jeong, and N. Ik Cho, "Multi-scale recurrent encoder-decoder network for dense temporal classification," in *Proc. Int. Conf. Pattern Recognit.*, pp. 103-108, 2018.
[62] J. Liao, G. Guo, Y. Yan, and H. Wang, "Multiscale cascaded scene-specific convolutional neural networks for background subtraction," in *Proc. Pacific Rim Conf. Multimed.*, pp. 524-533, 2018.
[63] M. O. Tezcan, P. Ishwar, J. Konrad, "BSUV-Net: A Fully-Convolutional Neural Network for Background Subtraction of Unseen Videos," in *Proc. IEEE Wint. Conf. Appl. Compt. Vis.*, pp. 2774-2783, 2020.
[64] L. A. Lim and H. A. Keles, "Learning multi-scale features for foreground segmentation," *Pattern Anal. Applicat.*, doi: 10.1007/s10044-019-00845-922.
[65] L. Maddalena and A. Petrosino, "Towards benchmarking scene background initialization," in *Proc. Int. Conf. Image Anal. Process.*, pp. 469–476, Springer, 2015.
[66] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1168–1177, 2008.
[67] C. Cuevas and N. García, "Improved background modeling for realtime spatio-temporal non-parametric moving object detection strategies," *Image Vis. Comput.*, vol. 31, no. 9, pp. 616–630, 2013.
[68] T. S. F. Haines and T. Xiang, "Background subtraction with Dirichlet process mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 670–683, 2013
[69] D. Berjón, C. Cuevas, F. Morán, and N. García, "Real-time nonparametric background subtraction with tracking-based foreground update," *Pattern Recognit.*, vol. 74, pp. 156–170, Feb. 2018.
[70] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, 2004.
[71] N. Wang, T. Yao, J. Wang, and D. Yeung, "A probabilistic approach to robust matrix factorization," in *Proc. Eur. Conf. Comput. Vis.*, vol. 58, no. 3, pp. 126–139, 2012.
[72] X. Guo, X. Wang, L. Yang, X. Cao, and Y. Ma, "Robust foreground detection using smoothness and arbitrariness constraints," in *Proc. Eur. Conf. Comput. Vis.*, pp. 535–550, 2014.
[73] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1568–1575, 2012.
[74] H. Guo, C. Qiu, and N. Vaswani, "An online algorithm for separating sparse and low-dimensional signal sequences from their sum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4284–4297, 2014.
[75] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.
[76] J. Xu, V. K. Ithapu, L. Mukherjee, J. M. Rehg, and V. Singh, "GOSUS: Grassmannian online subspace updates with structured-sparsity," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 3376–3383, 2013.
[77] H. Yong, D. Meng, W. Zuo, and L. Zhang, "Robust online matrix factorization for dynamic background subtraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1726–1740, 2018.
[78] K. Schindler and H. Wang, "Smooth foreground-background segmentation for video processing," in *Proc. Asian Conf. Comput. Vis.*, pp. 581–590, 2006.

Murari Mandal completed his graduation and post-graduation in Computer Science from BITS Pilani and Thapar University. He has submitted his PhD thesis in the Department of Computer Science and Engineering, Malaviya National Institute of Technology Jaipur, India. He is currently a Postdoctoral Researcher at the School of Computing, National University of Singapore. His current research is focused on deep learning, computer vision, and aerial scene analysis.

Santosh Kumar Vipparthi is an Assistant Professor at the Department of Computer Science and Engineering, MNIT Jaipur, India. He leads the Visionintellignece Lab @ MNIT with research focused on important visual perception tasks such as deep learning, object detection, aerial scene analysis, human emotion recognition, aberrant event detection, image retrieval, gesture recognition, motion analysis, etc. He received his Ph. D. and M. Tech in Systems Engineering from IIT BHU, Varanasi, India and B.E degree in Electrical and Electronics Engineering from Andhra University, India.