3DCD: A Scene Independent End-to-End Spatiotemporal Feature Learning Framework for Change Detection in Unseen Videos

Murari Mandal, Vansh Dhar, Abhishek Mishra, Santosh Kumar Vipparthi, Mohamed Abdel-Mottaleb

***Abstract*—Change detection is an elementary task in computer vision and video processing applications. Recently, a number of supervised methods based on convolutional neural networks have reported high performance over the benchmark dataset. However, their success depends upon the availability of certain proportions of annotated frames from test video during training. Thus, their performance on completely unseen videos or scene independent setup is undocumented in the literature. In this work, we present a scene independent evaluation (SIE) framework to test the supervised methods in completely unseen videos to obtain generalized models for change detection. In addition, a scene dependent evaluation (SDE) is also performed to document the comparative analysis with the existing approaches. We propose a fast (speed-25 fps) and lightweight (0.13 million parameters, model size-1.16 MB) end-to-end 3D-CNN based change detection network (3DCD) with multiple spatiotemporal learning blocks. The proposed 3DCD consists of a gradual reductionist block for background estimation from past temporal history. It also enables motion saliency estimation, multi-schematic feature encoding-decoding, and finally foreground segmentation through several modular blocks. The proposed 3DCD outperforms the existing state-of-the-art approaches evaluated in both SIE and SDE setup over the benchmark CDnet 2014, LASIESTA and SBMI2015 datasets. To the best of our knowledge, this is a first attempt to present results in clearly defined SDE and SIE setups in three change detection datasets.**

***Index Terms*—Change detection, background subtraction, 3D-CNN, spatiotemporal, scene independence, deep learning**

I. Introduction

Change detection in the video has numerous applications in traffic monitoring, video synopsis, human-machine interaction, behavior analysis, action recognition, visual surveillance, anomaly detection, and object tracking. The objective of a change detection technique is to segment a video frame into the foreground and background regions corresponding to object motion. Since it is often used as the first pre-processing step, the output accuracy has an overwhelming effect on the overall performance of the subsequent tasks. Therefore, it is critical to produce an accurate motion segmentation map. However, various difficult scenarios such as fluctuation in background regions, illumination variation, shadow, variable frame rate in 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 designed background subtraction methods to model the background behavior and identify the foreground regions using various thresholding techniques. Usually, these methods are unsupervised in nature and can be grouped into parametric and non-parametric approaches. The parametric statistical models [1]–[3] have been widely adopted for background subtraction in the past two decades. Whereas, more recent research in this domain is inspired by non-parametric modeling approaches [4]–[10].

Recent advances in supervised learning [11]–[14] approaches have led to the development of various deep learning-based techniques to solve the change detection problem. Many attempts in this domain leverage off-the-shelf pre-trained convolutional neural networks (CNNs) and integrate them with hand-crafted background modeling techniques for temporal feature encoding [15]–[18]. Some methods [18]–[21] divide the video frames into patches for training. Others have modeled the pixel-wise change by presenting different variations of CNN [22]–[25] and generative adversarial networks (GAN) [26], [27] based architectures.

The supervised deep learning methods have apparently outperformed the unsupervised algorithms in the literature. However, most of these models [15]–[25] have been optimized either for one specific video or a group of similar videos. We denote such an evaluation scheme as scene dependent evaluation (SDE). In SDE, some frames from the test videos are used for training the model. This has led to bloated results over the benchmark datasets. The performance of these methods has not been evaluated on unseen videos.

In order to assess the robustness of the designed models for real-world scenarios, it is imperative to evaluate the models with videos that were not used in training. Therefore, in this paper, we introduce a scene independent evaluation (SIE) scheme to avoid any bias in the evaluation. In SIE, the training and testing sets are composed of frames originating from different videos. This ensures that no labeled ground-truth from test videos has been exposed to the network in the training phase. Thus, the performance is expected to be compared only based upon unseen test videos.

This work was supported in part by IBM with online GPU grant and in part by the Department of Science and Technology-Science and Engineering Research Board (DST-SERB) project under Grant EEQ/2017/000673.

Murari Mandal and Santosh Kumar Vipparthi are with the Vision Intelligence Lab, Department of Computer Science and Engineering, Malaviya National Institute of Technology Jaipur, 302017 India (Email: murarimandal.cv@gmail.com; skvipparthi@mnit.ac.in)

Vansh Dhar and Abhishek Mishra carried out the work as an intern in the Vision Intelligence Lab, Department of Computer Science and Engineering, Malaviya National Institute of Technology Jaipur, 302017 India (Email: vanshdhar.ai@gmail.com; mishra.abhishek.ai@gmail.com)

Mohamed Abdel-Mottaleb is with Department of Electrical & Computer Engineering, University of Miami, USA (Email: mottaleb@miami.edu)

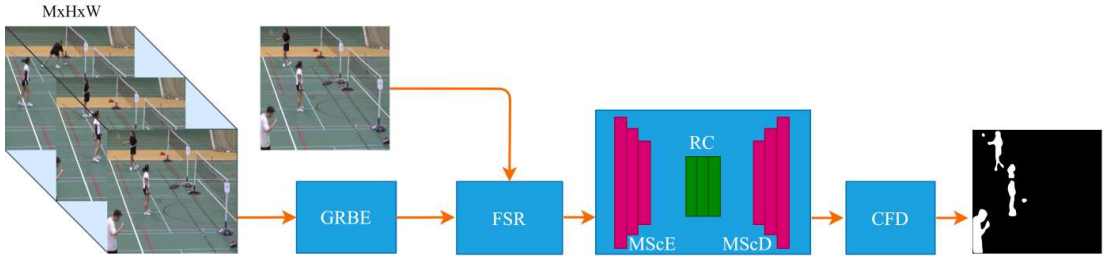

Fig. 1: The proposed 3DCD framework for change detection. GRBE: gradual reduction background estimation, FSR: foreground saliency reinforcement, MScE: multi-schematic encoder, RC: residual connector, MScD: multi-schematic decoder, CFD: compact foreground detection

In this work, we propose a fast and lightweight end-to-end spatiotemporal feature learning framework (3DCD) for change detection in unseen videos. The proposed framework enables background estimation, motion saliency estimation, multi-schematic feature encoding, decoding, and finally foreground segmentation through an end-to-end 3D-CNN network. The proposed 3DCD is demonstrated through a block diagram in Fig. 1. To summarize, this paper makes the following contributions:

1) We present a completely end-to-end spatiotemporal network 3DCD consisting of blocks for background estimation, motion saliency representation, and finally foreground segmentation map. Our online model is very fast (speed-25 fps) and lightweight (0.13 million parameters, model size-1.16 MB), making it suitable for real-time applications.
2) We designed multi-schematic architectures for features encoding and decoding with residual connectors to learn structurally diverse motion saliency features at multiple scales.
3) We present a scene independent evaluation (SIE) scheme to train and evaluate the proposed 3DCD in unseen videos in three benchmark datasets CDnet2014, LASIESTA, and SBMI2015. Evaluation over completely unseen videos in the SIE setup ensures fair evaluation of the generalization capability of the designed network. Moreover, for comparative analysis with existing approaches, scene dependent evaluation (SDE) is also conducted for CDnet2014, LASIESTA, and SBMI2015.
4) The proposed 3DCD outperforms (overall, in terms of accuracy, speed, memory, and compute efficiency) the existing state-of-the-art methods. The ablation studies and visualization of the proposed network are also discussed in the experimental section.

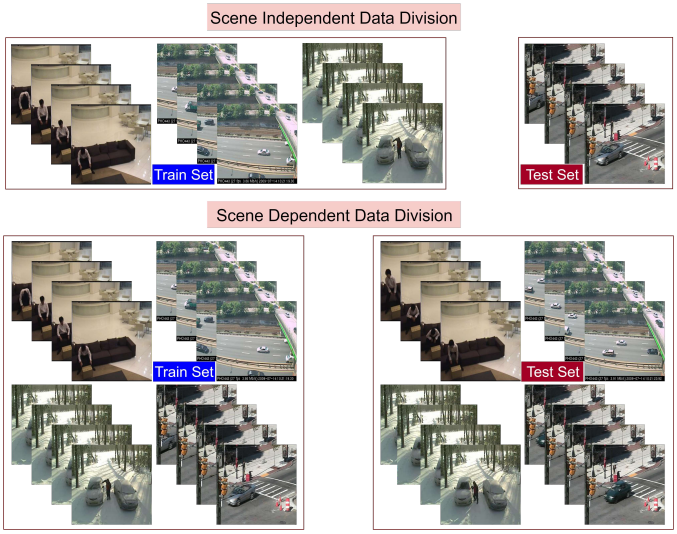

Fig. 2: Difference between the scene dependent and scene independent data division schemes. In the scene dependent setup, training, and testing frames are collected from the same video. Whereas, in the scene independent setup, only unseen videos are used for evaluation

II. Related Work

An extensive body of literature is available on change detection. We group and discuss these techniques in two categories: unsupervised methods and supervised deep learning-based methods.

A. Unsupervised Methods

The unsupervised methods for change detection are usually comprised of two crucial tasks: extraction of pertinent features from image sequences and background modeling. For feature extraction, low-level image features, i.e., grayscale, color intensity, and edge magnitudes [28], [29] are commonly used. Moreover, specific spatial and spatiotemporal local feature descriptors have also been designed for enhanced performance [8], [9], [30]. A local descriptor extracts the feature representation in a region or local neighborhood of an image. The background modelling techniques can be categorized into parametric and non-parametric methods. Stauffer and Grimson [2] developed a parametric approach using Gaussian Mixture Models (GMM). In GMM, the statistical distribution at each location is modeled and updated through a mixture of Gaussian distributions and Expectation Maximization (EM) algorithm respectively. Based on this architecture, several mechanisms were presented by designing adaptive GMM with variable parameter selection and spatial mixture of Gaussians [1], [3]. Most of the modern non-parametric methods are inspired by the consensus-based method [31] and ViBe [32]. In [31], a collection of background samples is stored at each pixel and is updated through a first-in-first-out policy. However, such an update policy doesn't necessarily reflect the background behavior in real-life video sequences. Therefore, to alleviate some of these issues, three significant background

model maintenance policies were proposed in [32]: random sample, memoryless update policy, and spatial diffusion via background sample propagation. These strategies have been widely adopted by the recent state-of-the-art methods [4], [8], [9]. The foreground detection and background model update in ViBe were performed using a manually defined threshold and static update rate. To adaptively update the decision thresholds, learning rate for foreground segmentation, and background model maintenance respectively, a Pixel-Based Adaptive Segmenter (PBAS) was proposed in [10]. Moreover, St-Charles et al. [9], [30] designed a spatiotemporal feature descriptor to simultaneously map the low-level (color intensities) as well as the local neighborhood features for robust background subtraction. They also incorporated an adaptive feedback mechanism to continuously monitor the background model fidelity and segmentation entropy to update the decision thresholds, learning rates and background samples. Furthermore, a deterministic background model update policy was proposed by Mandal et al. [29]. Bianco et al. [33] conducted multiple experiments to combine various background subtraction techniques through genetic programming for improved performance.

The traditional unsupervised methods naturally follow the SIE setup because they do not require any ground truth labels prior to the evaluation stage. This ensures fair evaluation through documentation of results in the same experimental setup. Moreover, the robustness of the algorithm is estimated in unseen videos for all the methods.

B. Supervised Deep Learning Methods

Many researchers have designed CNN models to segment video frames into foreground and background regions. Babaee et al. [20] generated the background image using proven hand-crafted approaches like SubSENSE [9] and Flux Tensor [34]. This background image and the current frame are partitioned into small patches and concatenated together to form the input layer. The motion features are learned by feeding this input to a CNN network. The final response is generated by augmenting these segmentation maps. Similarly, Nguyen et al. [19] designed a triplet CNN network to extract the relevant features for change detection. Furthermore, the feature learning capability of off-the-shelf CNN models such as VGG16 has also been successfully adapted for change detection in [15], [17], [18]. Lim et al. [16] developed a background model update policy along with the encoder-decoder network for adaptive background subtraction. The LSTM networks [35], [36] have been successfully used in the literature to model the temporal variations. Similarly, the attention-based mechanism [37] has been designed for multi-modal feature fusion. Chen et al. [38] proposed an attention ConvLSTM network to model pixel-wise changes over time. Moreover, authors in [23] temporally encoded the motion information by sampling multiple images from previous frames with increasing intervals. Patil and Murala [25] designed a compact end-to-end CNN and Akilan et al. [24] proposed a 3D-CNN LSTM based network to model pixel-wise changes over time. Several methods [39], [40] extract the multi-scale convolutional features to learn robust spatial context from the images. This approach has been used in the development of several other moving object detection algorithms as well [41], [42]. In addition, Generative Adversarial Network (cGAN) based models [26], [27] have also been designed to learn motion features for change detection.

Since, the benchmark change detection datasets do not define the train-test division. Thus, researchers have used various data division strategies for network training and evaluation. We categorize the evaluation strategies into scene independent and scene dependent setups. In SDE both train and test set contain frames from the same video whereas, only unseen videos are used for evaluation in SIE setup. In Fig. 2, we depict the difference between SDE and SIE setup.

III. Proposed End-to-End 3DCD

We give a detailed description of the proposed end-to-end 3D-CNN based change detection (3DCD) framework through its constituent blocks. The feature map visualizations for each of these blocks are qualitatively analyzed for an intuitive understanding of the proposed network. Furthermore, we discuss some insights on the strengths of the proposed 3DCD over the existing methods.

A. Gradual Reduction Background Estimation (GRBE)

The proposed 3DCD uses the spatiotemporal signals from past history to estimate the background. In addition, the current frame is also fed into the network as a reference for motion estimation. In $GRBE$ block, we estimate the background from recent history frames ($M = 50$) through a sequence of 3D convolutions and spatiotemporal average pooling layers. The motion characteristics in a video may vary for different scenarios such as slow motion, fast motion, intermittent motion, camera jitter, and dynamic background. These intrinsic challenges should be taken into account while designing a change detection algorithm. To address these challenges, different granularity of temporal depths (past history frames) can be used to describe different types of object motions. Moreover, the temporal mean value has been frequently used in the literature [29] to estimate the static regions. Motivated by the above considerations, we exploit the spatiotemporal features computed at multiple levels of temporal depths. By multi-level, we mean that the temporal features are decomposed into a single feature map by applying 3D convolutions at different granular depths. The final feature map estimates the background for subsequent processing.

To estimate the background, we apply 3D average pooling with stride 5, 2, and 5 to achieve granular spatiotemporal features with depths 10, 5, and 1, respectively. At each level of granularity, we apply 3D convolution to robustly encode the spatiotemporal patterns. The complete $GRBE$ block estimates the background through feature learning in both the spatial and temporal domains by progressive elimination of temporal movements at multiple levels of granularity. The $GRBE$ block architecture is depicted in Fig. 3. From M historical frames, the $GRBE$ block generates a single depth feature map for background representation. Let's denote the input tensor as

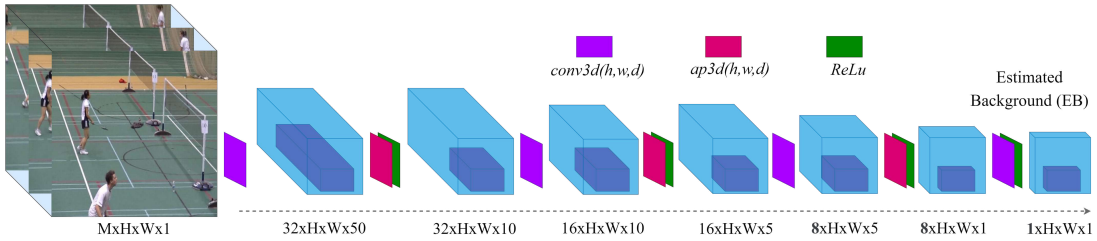

Fig. 3: Illustration of the GRBE block. It takes past temporal history as input and estimates the background (EB) through progressive reduction of the spatiotemporal features at multiple granularities. ap3d: 3d average pooling, conv3d: 3D convolution.

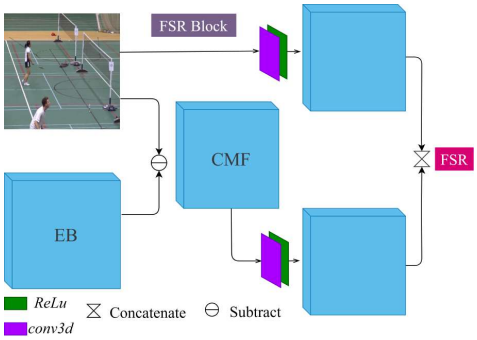

Fig. 4: The FSR block. The coarse motion features are estimated in CMF which are further refined through the FSR. CMF: coarse motion features, FSR: foreground saliency reinforcement.

V_M. We first compute the granular features GR_M at depths 10 and 5 using Eq. 1.

$$GR_M = S_l\left(V_M\right)|_{l=1}^{2} \tag{1}$$

The $S_l(\cdot)$ is computed using Eq. 2 and Eq. 3.

$$S_l\left(V_M\right) = \Re\left(ap_{h,w,d_l}\left(\kappa_{2^{l+2},h,w,d} \otimes S_{l+1}\left(V_M\right)\right)\right) \tag{2}$$

$$S_3\left(V_M\right) = \Re\left(ap_{h,w,d_3}\left(\kappa_{2^5,h,w,d} \otimes V_M\right)\right) \tag{3}$$

where $\otimes$ denote 3D convolution operation and $\kappa_{x,h,w,d}$ is convolutional kernel with parameters x, h, w, and d representing the number of kernels, height, width, and depth of the kernels, respectively. In $GRBE$ block, we use $h = w = d = 3$ and $stride = 1$. The ap_{h,w,d_l} represents 3D average pooling with parameters $d_1 = 5(stride = 5)$, $d_2 = 2(stride = 2)$, and $d_3 = 5(stride = 5)$, respectively. $\Re$ denotes the rectified linear unit (ReLu) activation function. It can be observed from Eq. 1 - Eq. 3 and Fig. 3 that the average pooling is applied spatiotemporally at multiple temporal depths which enables $GRBE$ block to estimate background with contributions from different granularities. Finally, the estimated background EB_M is computed using Eq. 4.

$$EB_M = \Re\left(\kappa_{1,h,w,1} \otimes GR_M\right) \tag{4}$$

B. Foreground Saliency Reinforcement (FSR)

After extracting the background EB_M from the input tensor V_M, the coarse motion features (CMF) are identified using a subtraction layer. The CMF for current frame I is computed using Eq. 5.

$$CMF = I - EB_M \tag{5}$$

Simply applying the CMF may lead to certain semantic shape distortions resulting in inadequate learning of foreground features. Thus, we introduce a saliency reinforcement to restore foreground semantics by assimilating features from the current frame with CMF in the foreground saliency reinforcement (FSR) block. The detailed architecture of the FSR block is depicted in Fig. 4. The FSR response is computed through Eq. 6.

$$FSR = \Re\left(\left[\kappa_{8,h,w,1} \otimes CMF, \kappa_{8,h,w,1} \otimes I\right]\right) \tag{6}$$

The kernels in FSR block and all the subsequent blocks of 3DCD are designed to learn refined spatial features from the estimated background for semantically aware foreground detection.

C. MScE and MScD with Residual Connectors

The FSR block response maps are further processed through multi-schematic encoder, residual connectors, and multi-schematic decoder to estimate a more accurate pixel-level segmentation map. Multi-scale feature representations have been successfully used in semantic segmentation applications to achieve robust performance [11], [43]–[45]. Most of the existing approaches have extracted multi-scale features in the encoder section, whereas, the decoder section usually consists of simple up sampling operations [46], fully connected layers [11], [45], shortcut connections [47], or feature fusion with encoder features [12]. However, in this paper, we designed both multi-schematic encoder ($MScE$) and multi-schematic decoder ($MScD$) architectures to learn structurally diverse features at multiple scales for robust change detection.

1) Multi-schematic Encoder: Here, the FSR map is represented through schemas at three different scales. These three schemas denoted as $MScE_1$, $MScE_2$, $MScE_3$, are computed using Eq. 7-Eq. 9.

$$MScE_1 = \Re\left(mp_{2,2,1}\left(\kappa_{16,h,w,1} \otimes mp_{2,2,1}(FSR)\right)\right) \tag{7}$$

$$MScE_2 = \Re\left(\kappa_{16,h,w,1} \otimes mp_{4,4,1}(FSR)\right) \tag{8}$$

$$\begin{aligned} MScE_3 = \Re(mp_{2,2,1}(\kappa_{16,h,w,1} \otimes \\ \Re(mp_{2,2,1}(\kappa_{16,h,w,1} \otimes FSR)))) \end{aligned} \tag{9}$$

where $mp_{h,w,d}$ represents 3D max pooling. The three schemas $MScE_1$, $MScE_2$, $MScE_3$, extract discernible features at three different scales. In $MScE_1$ and $MScE_2$, we first apply max pooling by a factor of 2 and 4, respectively over FSR. Subsequently, salient features are learned from these two

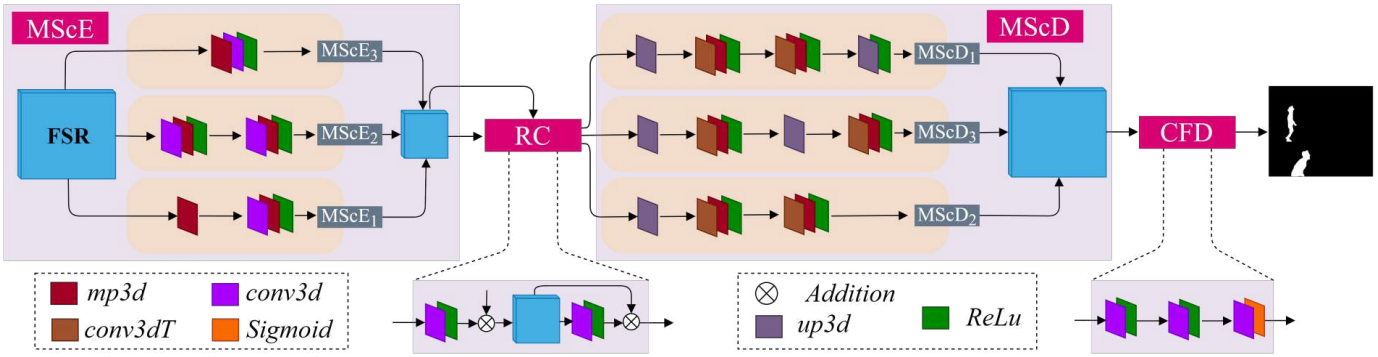

Fig. 5: Illustration of the MScE, RC, MScD, and CFD blocks. The FSR responses are encoded through three schematic streams MScE1, MScE2 and MScE3 in MScE block. Thereafter, the residual connectors (RC) are used to refine MScE features. Furthermore, the refined abstract features are decoded through three schematic streams MScD1, MScD2, and MScD3 in MScD block. The assimilated features are fed to CFD block to generate the final segmentation map. conv3d: 3D convolution, conv3dT: 3D transpose convolution, up3d: 3D up sampling, mp3d: 3D max pooling.

different levels of spatial granularity. In addition, $MScE_3$ encodes convolutional features from the original FSR. Finally, the response of the $MScE$ block is computed by combining the multi-schematic features from the three streams using Eq. 10.

$$MScE = [MScE_1, MScE_2, MScE_3] \tag{10}$$

The detailed $MScE$ architecture is shown in Fig. 5.

2) Residual Connectors: In order to fortify the high-level foreground semantics, multiple residual connectors are used to refine the $MScE$ features. The RC features, as shown in Fig. 5, is computed using Eq. 11 and Eq. 12.

$$RC = \Re\left((\kappa_{48,h,w,1} \otimes RC_1) + RC_1\right) \tag{11}$$

$$RC_1 = \Re\left((\kappa_{48,h,w,1} \otimes MScE) + MScE\right) \tag{12}$$

3) Multi-schematic Decoder: Similar to the multi-schematic encoder, we design the multi-schematic decoder in order to reproduce the original frame shaped features for pixel-level change detection. As shown in Fig. 5, we used three different schemas $MScD_1$, $MScD_2$, $MScD_3$, to capture the context information from different streams of saliency estimation. In $MScD_1$ and $MScD_2$, we first up sample the high-level encoder features by the factor of 2 and 4, respectively. Furthermore, we reconstruct the pixel-level features by learning salient patterns through convolution, max pool and up-sample layers. In $MScD_3$, we use the up-sample layers at altered position as compared to $MScD_1$. This approach of using up-sampling and convolution at alternative positions can capture saliency by extracting differentiable features from encoded maps of same resolutions at different schemas. Three multi-stream up-sampled features of $MScD$: $MScD_1$, $MScD_2$, and $MScD_3$ are computed using Eq. 13 - Eq. 15.

$$MScD_1 = \Re\left(up_{2,2,1}\left(CB_{16}\left(CB_{32}\left(up_{2,2,1}(RC)\right)\right)\right)\right) \tag{13}$$

$$MScD_2 = CB_{16}\left(CB_{32}\left(up_{4,4,1}(RC)\right)\right) \tag{14}$$

$$MScD_3 = CB_{16}\left(up_{2,2,1}\left(CB_{32}\left(up_{2,2,1}(RC)\right)\right)\right) \tag{15}$$

The $CB_j(\cdot)$ is computed using Eq. 16.

$$CB_j(x) = \Re\left(mp_{h,w,1}\left(\kappa^T_{j,h,w,1} \otimes x\right)\right) \tag{16}$$

where $up_{h,w,1}$, $\kappa^T_{h,w,1}$ denote 3D upsample, transposed convolutional kernel, respectively. The resultant multi-schematic features are assimilated using Eq. 17.

$$MScD = [MScD_1, MScD_2, MScD_3] \tag{17}$$

As the objects in a video are captured with different angle of view, have different scales, and aspect ratios. The multi-schematic features encoded and decoded in the $MScE$ and $MScD$ blocks would give more clues for foreground reconstruction for such scenarios.

D. Compact Foreground Detection

The final motion segmentation is performed by gradual feature depth reduction with the CFD block as given in Eq. 18 and Eq. 19.

$$CFD = \delta\left(\kappa_{1,h,w,1} \otimes CF_8\left(CF_{16}(MScD)\right)\right) \tag{18}$$

$$CF_j(x) = \Re\left(\kappa_{j,h,w,1} \otimes x\right)) \tag{19}$$

where $\delta(\cdot)$ denotes the sigmoid function. The final foreground segmentation map is represented through a binary image as shown in Fig. 5.

E. Strengths and Analysis of the Proposed 3DCD

One of the major advantages of the proposed 3DCD framework over the existing deep learning methods for change detection is the intuitive spatiotemporal model design. The 3DCD takes the previous 50 frames as input to estimate the background. Whereas, most of the existing methods solve the problem as a single image segmentation problem by carefully selecting the training frames. For example, FgSegNet [15] and MSFS [49] models do not use any temporal data and rather depend on the data selection strategy to optimize the results. Such models overfit the dataset and do not necessarily learn the underlying task of change detection. Our analysis is clearly supported by the significantly better performance of 3DCD over FgSegNet and MSFS in SIE setup as presented in Table III, Table V and Table VII.

The proposed 3DCD model is highly lightweight in nature. The existing methods [15], [24], [38], [49] have stacked a

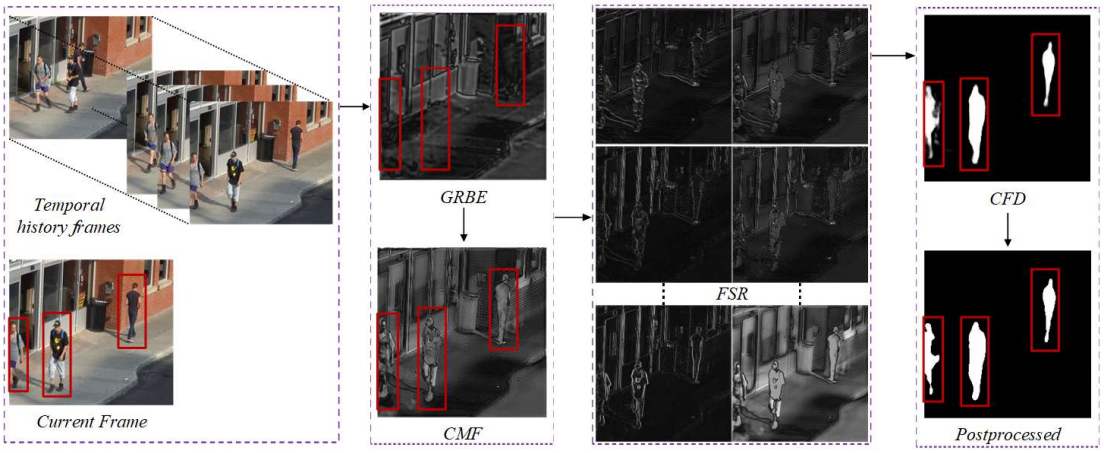

Fig. 6: Visualization of different blocks of proposed 3DCD

TABLE I
COMPARATIVE ANALYSIS OF PROPOSED 3DCD WITH EXISTING METHODS AND EVALUATION SCHEMES

Method	BE	MS	Training data selection	SIE
Babaee et al. [20]	SuBSENSE	N	Random 5% (video-wise)	N
Nguyen et al. [19]	SuBSENSE	N	Random 100 frames (video-wise)	N
Lin et al. [18]	SuBSENSE	N	LOVO	Y
Lim et al. [16]	Designed background model	N	LOVO	N
Brahman et al. [21]	IUTIS-5	N	50% (video-wise)	N
Wang et al. [48]	Frame-level segmentation	N	Selective 50/200 frames (video-wise)	N
Lim & Keles [15]	Frame-level segmentation	Y	Selective 50/200 frames (video-wise)	N
Zeng et al. [17]	Frame-level segmentation	Y	Random 150 frames (video-wise)	N
Chen et al. [38]	2D-CNN	N	50% (video-wise)	N
Yang et al. [23]	2D-CNN	N	90% (video-wise)	N
Bakkay et al. [26]	Generator (2D-CNN)	N	50% (video-wise)	N
Akilan et al. [24]	3D-CNN	N	70% (video-wise)	N
Patil et al. [25]	2D-CNN	N	Random (video-wise)	N
Proposed 3DCD	**End-to-end 3D-CNN**	**Y**	**LOVO**	**Y**

BE: Background estimation, MS: Multi-scale feature learning, SIE: Scene independent evaluation

variety of pre-trained models (VGG16, GoogleNet, ResNet50, DeepLabv3), expensive operations such as conditional random fields (CRF), semantic segmentation, etc. to obtain the results. These results come at the cost of increased space and computational complexity. Our work outperforms (overall, in terms of accuracy, speed, memory, and compute efficiency) the existing state-of-the-art methods. It makes 3DCD a highly suitable candidate for real-time applications. In addition to the network design strengths, our model is the first extensively validated model in scene independent setup across 3 different datasets. Furthermore, comparative analysis with existing methods in the same setup makes our work a valuable contribution to change detection research.

F. Visualization

In order to illustrate the intermediate responses of the constituent blocks of the 3DCD network, we also compute the feature map visualizations at $GRBE$, CMF, FSR, CFD, and depict the same in Fig. 6. Here, we can see that the background features are estimated quite robustly through the $GRBE$ block. After subtracting the estimated background with the current frame, we get a fair representation of the motion features in the CMF block. Furthermore, the refined foreground representation is achieved through the the FSR and CFD features. Thus, the proposed 3DCD can also be used as a modular framework to design and develop custom blocks for improved change detection performance.

IV. EXPERIMENTAL SETTING, RESULTS, AND DISCUSSIONS

A. Problem with Scene Dependency

In scene dependent evaluation (SDE) setup, some frames from test videos are also used in training which is not an ideal setup for deep learning model evaluation. In order to actualize a generalized model, it is essential to evaluate the trained model over unseen videos. This also makes the process of model design much more challenging and ensures better performance in real-world scenarios. Therefore, the proposed SIE setup to evaluate the designed deep networks on completely unseen data is a better evaluation strategy as compared to the SDE setup. More recent benchmark datasets for other video-based applications [50], [51] 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. Comparison with the Existing Evaluation Schemes

We compare the proposed method and evaluation scheme with existing approaches in terms of background estimation (BE), multi-scale feature learning (MS), training data selection, and scene independent evaluation (SIE) in Table I. In terms of evaluation schemes, most of the existing

methods [15]–[21], [23]–[26], [38], [48] have adopted SDE scheme in which some frames from test videos are also used in training. These SDE setups follow a variety of training data selection strategies. For example, selection of training frames through temporal division with different proportions such as 50%, 70%, 90%, 5%, 50/100/150/200 of frames, etc. Even though the authors in [18] have attempted to conduct scene independent experiments, they only used six categories in their experiments with quite low performance (68.74 F-score overall). Whereas, we propose a clearly defined SIE setup and used over 10 categories (49 videos) to ensure robust performance analysis of CNN models in unseen videos. Furthermore, in order to give a comparative analysis with reference to the existing methods, we also present a clearly defined SDE setup.

In terms of BE, the methods in [16], [18]–[21] are dependent either on statistical or non-parametric handcrafted methods to extract the temporal features. Whereas, the authors in [15], [17], [48] have just performed frame-level segmentation without considering the historical context. In [23]–[26], [38], the background features are estimated with CNN network. In the proposed 3DCD, we designed a $GRBE$ block to model temporal features from the recent history for effective BE in an end-to-end manner.

To the best of our knowledge, multi-scale features with spatiotemporal 3D-CNN layers have not been explored in the literature for change detection. The few methods [15], [17] to use multi-scale features have performed frame-wise segmentation in 2D-CNN without including any temporal context. However, in this paper we have exploited multi-schematic features through the $MScE$ and $MScD$ blocks to more robustly detect moving objects in videos having different scales, angles of view, and aspect ratios.

C. Experiment Settings and Dataset

1) Implementation Details: The entire network is implemented in Keras with Tensorflow backend. The 3DCD takes two tensors as input of shape $50 \times 256 \times 256 \times 1$ (temporal history) and $1 \times 256 \times 256 \times 1$ (current frame). We use $M = 50$ 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. We use stochastic gradient descent optimizer with binary cross-entropy loss function to train the network. The final loss is backpropagation to all the blocks of the network including GRBE for background estimation. 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.

3) Datasets: The benchmark CDnet 2014 [68], LASIESTA [69], and SBMI2015 [70] datasets are used for performance evaluation. The CDnet 2014 consists of 53 videos from a diverse set of realistic scenarios grouped into 11 different categories. Approximately 89,000 ground truth frames are available for training and evaluation. In our experiments, we exclude the PTZ category due to excessive camera motion. We used 88,882 frames for training and evaluation. The LASIESTA [69] 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 [70] dataset has 13 challenging videos. Approximately, 5,023 annotated frames are available for performance evaluation.

TABLE II
PERFORMANCE OF THE PROPOSED 3DCD IN SIE FRAMEWORK ON CDNET 2014 DATASET

Category	Scene	Prec	Rec	F-Score	PWC
BW	blizzard	0.94	0.95	0.94	0.12
BL	pedestrian	0.93	0.94	0.93	0.11
CJ	sidewalk	0.95	0.74	0.83	0.63
DB	boats	0.95	0.83	0.88	0.12
IOM	parking	0.86	0.81	0.84	2.26
LF	turnpike05fps	0.95	0.89	0.92	1.09
NV	tramStation	0.79	0.72	0.75	1.17
SD	busStation	0.73	0.84	0.79	1.58
TH	corridor	0.98	0.87	0.92	0.45
TB	turbulence1	0.91	0.74	0.82	0.09
Average		**0.90**	**0.83**	**0.86**	**0.76**

BW: bad weather, BL: baseline, CJ: camera jitter, DB: dynamic background, IOM: intermittent object motion, LF: low framerate, NV: night videos, SD: shadow, TH: thermal, TB: turbulence. Prec: Precision, Rec: Recall, PWC: Percentage of Wrong Classification

D. Quantitative Results

We conduct multiple experiments to evaluate the performance of the proposed 3DCD in both SIE and SDE setup. The performance is measured in terms of precision, recall, F-score, and percentage of wrong classification (PWC). We also perform a comparative analysis of the proposed and existing state-of-the-art deep learning and non-deep learning methods. The comparison is done based on F-score which is a comprehensive indicator of performance for change detection.

1) The problem of noncomparability in deep learning methods: We identified two glaring problems with the experimental setups of existing deep learning approaches for change detection. The problem of scene dependence is already discussed in the earlier Sections. Another issue is documentation of incomparable results even in the SDE setup. Different authors have adopted different ways to segregate training and testing data. In fact, the highest F-score is claimed [20], [48] by manually selecting a particular set of frames from a single video to train the model and then test over the same video. These video-optimized results are not directly comparable with other methods. Similarly, other methods used different data-division schemes which makes the claimed results incomparable. Therefore, we compute results with a clearly defined SDE setup and present baseline results for the same as well.

2) Quantitative results in the SIE setup: We conduct experiments on CDnet 2014, LASIESTA and SBMI2015 in the SIE setup. We also train and evaluate the existing deep learning methods FgSegNet-S, FgSegNet-M, and MSFS in the same SIE setup to present an empirical comparative analysis. All the results are presented in Table II-Table VIII.

CDnet 2014: The training and testing videos are separated using a leave-one-video-out (LOVO) strategy, i.e., one video from each category is used in evaluation and the rest are

TABLE III
COMPARATIVE F-SCORE PERFORMANCE IN SIE FRAMEWORK ON CDNET 2014 DATASET

Method	SIE	BL	PE	SW	BO	PA	TP	TS	BS	CO	T1	Avg
SuBSENSE [9]	Yes	0.85	0.95	0.81	0.69	0.48	0.85	0.86	0.86	0.91	0.79	0.81
VIBE [32]	Yes	0.53	0.90	0.30	0.22	0.26	0.60	0.67	0.67	0.75	0.58	0.55
PAWCS [52]	Yes	0.66	0.95	0.74	0.88	0.21	0.91	0.86	0.86	0.65	0.68	0.74
IUTIS-5 [33]	Yes	0.80	0.97	0.81	0.75	0.65	0.89	0.87	0.87	0.90	0.63	0.81
UBSS [28]	Yes	0.86	0.96	0.90	0.90	0.62	0.89	0.87	0.87	0.92	0.54	0.83
WeSamBe [4]	Yes	0.86	0.96	0.85	0.64	0.41	0.91	0.86	0.86	0.89	0.71	0.80
SemBGS [53]	Yes	0.84	0.98	0.85	0.98	0.69	0.88	0.92	0.92	0.82	0.30	0.82
BSUV-Net [54]	Yes	0.82	0.97	0.69	0.89	0.91	0.91	0.80	0.94	0.83	0.66	0.84
BSUV-Net+SemBGS [54]	Yes	0.82	0.97	0.71	0.91	0.91	0.91	0.80	0.96	0.82	0.65	0.85
BMN-BSN [55]	Yes	0.84	0.96	0.63	0.95	0.77	0.72	0.82	0.92	0.90	0.56	0.81
FgSegNet-S [15]	Yes	0.74	0.65	0.12	0.42	0.17	0.57	0.41	0.52	0.74	0.17	0.45
FgSegNet-M [15]	Yes	0.55	0.72	0.11	0.69	0.05	0.22	0.39	0.60	0.31	0.16	0.38
MSFS [49]	Yes	0.70	0.33	0.22	0.62	0.52	0.74	0.43	0.53	0.77	0.12	0.50
3DCD	Yes	**0.94**	**0.93**	**0.83**	**0.88**	**0.84**	**0.92**	**0.75**	**0.79**	**0.92**	**0.82**	**0.86**

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), TS: Tram Station (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 CDNET 2014 DATASET

Method	Train Data	Test data	BW	BA	CJ	DB	IOM	NV	LFR	SD	TH	TB	Avg
PAWCS [52]	0	100%	0.81	0.94	0.81	0.89	0.78	0.42	0.64	0.89	0.83	0.77	0.78
UBSS [28]	0	100%	0.79	0.89	0.87	0.78	0.76	0.53	0.62	0.78	0.79	0.47	0.73
WeSamBE [4]	0	100%	0.85	0.94	0.8	0.74	0.74	0.53	0.69	0.90	0.80	0.83	0.78
ViBe [32]	0	100%	0.77	0.88	0.45	0.72	0.47	0.40	0.33	0.83	0.55	0.61	0.60
SemBGS [53]	0	100%	0.83	0.96	0.84	0.95	0.79	0.50	0.79	0.95	0.82	0.69	0.81
IUTIS-5 [33]	0	100%	0.83	0.96	0.83	0.89	0.73	0.51	0.79	0.91	0.83	0.85	0.81
SuBSENSE [9]	0	100%	0.86	0.95	0.77	0.79	0.63	0.50	0.64	0.90	0.71	0.89	0.77
DeepBS [20]	RS	100%	0.86	0.96	0.9	0.88	0.61	0.64	0.59	0.93	0.76	0.90	0.80
MSFgNet [25]	RS	100%	0.85	0.92	0.83	0.85	0.78	0.81	0.84	0.93	0.80	0.86	0.85
SFEN(VGG) [56]	50%	100%	0.85	0.92	0.91	0.60	0.58	0.51	0.59	0.89	0.72	0.73	0.73
VGG+CRF [56]	50%	100%	0.88	0.94	0.93	0.62	0.61	0.52	0.61	0.90	0.73	0.74	0.75
VGG+PSL+CRF [56]	50%	100%	0.89	0.96	0.94	0.74	0.75	0.75	0.62	0.91	0.85	0.92	0.83
GoogLeNet+PSL+CRF [56]	50%	100%	0.8	0.86	0.89	0.66	0.65	0.60	0.59	0.80	0.77	0.76	0.74
Cascade-CNN [48]	S50	100%	0.79	0.97	0.97	0.95	0.87	0.87	0.74	0.95	0.89	0.84	0.88
EDS-CNN [16]	LOVO	100%	0.87	0.96	0.89	0.91	0.88	0.77	0.93	0.85	0.80	0.76	0.86
MCSCNNv1 [57]	S5	100%	0.88	0.94	0.68	0.81	0.77	0.77	0.65	0.89	0.88	0.80	0.81
MCSCNNv2 [57]	S5	100%	0.86	0.92	0.73	0.86	0.76	0.76	0.70	0.91	0.86	0.87	0.82
MCSCNNv3 [57]	S5	100%	0.85	0.93	0.62	0.73	0.76	0.68	0.64	0.89	0.87	0.74	0.77
MCSCNNv4 [57]	S5	100%	0.86	0.94	0.79	0.88	0.77	0.79	0.73	0.92	0.88	0.88	0.84
DPDL1 [58]	S1	100%	0.60	0.79	0.55	0.66	0.51	0.40	0.60	0.69	0.67	0.63	0.61
DPDL20 [58]	S20	100%	0.81	0.96	0.86	0.84	0.82	0.59	0.66	0.87	0.83	0.72	0.80
DPDL40 [58]	S40	100%	0.87	0.97	0.87	0.87	0.87	0.61	0.71	0.94	0.84	0.76	0.83
MSRNN [59]	NA	100%	0.89	0.96	0.92	0.91	0.87	0.56	0.84	0.95	0.85	0.80	0.85
LTDP [60]	0	100%	0.67	0.95	0.81	0.82	0.73	0.54	0.76	0.90	0.79	0.89	0.79
SBSNv1 [61]	NA	100%	0.72	0.97	0.61	0.82	NA	0.38	NA	0.56	0.66	0.58	-
SBSNv2 [61]	NA	100%	0.45	0.01	0.55	0.1	NA	0.25	NA	0.28	0.27	0.23	-
SBSNv3 [61]	NA	100%	0.74	0.98	0.42	0.61	NA	0.39	NA	0.64	0.65	0.53	-
SBSNv4 [61]	NA	100%	0.92	0.95	0.89	0.79	NA	0.77	NA	0.86	0.86	0.73	-
REDNv1 [62]	S200	100%	0.78	0.97	0.93	0.86	0.80	0.79	0.73	0.88	0.84	0.88	0.85
REDNv2 [62]	S100	100%	0.67	0.91	0.86	0.55	0.70	0.62	0.64	0.76	0.77	0.46	0.69
REDNv3 [62]	S100	100%	0.65	0.91	0.83	0.51	0.73	0.71	0.63	0.76	0.77	0.46	0.70
REDNv4 [62]	S100	100%	0.87	0.95	0.90	0.72	0.80	0.79	0.64	0.83	0.83	0.76	0.81
FgSegNet-S-51 [15]$^{\#}$	50%	100%	0.79	0.86	0.9	0.78	0.78	0.79	0.32	0.81	0.83	0.58	0.74
FgSegNet-S-55 [15]$^{\#}$	50%	50%	0.77	0.84	0.84	0.74	0.70	0.82	0.33	0.78	0.72	0.57	0.71
FgSegNet-M-51 [15]$^{\#}$	50%	100%	0.73	0.93	0.83	0.76	0.75	0.77	0.32	0.82	0.79	0.6	0.73
FgSegNet-M-55 [15]$^{\#}$	50%	50%	0.72	0.92	0.76	0.69	0.61	0.81	0.32	0.82	0.73	0.57	0.70
MSFS-51 [49]$^{\#}$	50%	100%	0.85	0.90	0.91	0.55	0.68	0.87	0.61	0.94	0.91	0.70	0.79
MSFS-55 [49]$^{\#}$	50%	50%	0.8	0.89	0.88	0.50	0.7	0.82	0.55	0.93	0.86	0.67	0.76
3DCD-51	**50%**	**100%**	**0.94**	**0.93**	**0.83**	**0.87**	**0.90**	**0.86**	**0.76**	**0.89**	**0.87**	**0.94**	**0.88**
3DCD-55	**50%**	**50%**	**0.95**	**0.91**	**0.81**	**0.85**	**0.83**	**0.87**	**0.74**	**0.88**	**0.85**	**0.92**	**0.86**

LOVO: Leave-one-video-out, RS: Random Selection, S50: Selected 50 frames; 3DCD-51: Training with 50% of frames and testing with 100% frames; 3DCD-55: Training with 50% of frames and testing with remaining 50% of frames; NA: Data not available. $^{\#}$These results are computed by training and evaluating the existing methods in the exact same SDE setup as done for the proposed 3DCD

used in training the network. The quantitative performance of the proposed 3DCD in the SIE setup is presented in

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

Method	ISI-2	ICA-2	IOC-2	IIL-2	IMB-2	IBS-2	OCL-2	ORA-2	OSN-2	OSU-2	Avg
Zivkovik [1]	0.89	0.75	0.91	0.31	0.80	0.52	0.82	0.80	0.24	0.88	0.69
Maddalena1 [63]	0.85	0.74	0.85	0.38	0.68	0.45	0.85	0.86	0.46	0.86	0.70
Maddalena2 [64]	0.94	0.87	0.95	0.23	0.85	0.40	0.88	0.86	0.71	0.88	0.76
Cuevas1 [65]	0.76	0.63	0.88	0.79	0.68	0.66	0.90	0.87	0.09	0.81	0.71
Haines [66]	0.81	0.87	0.95	0.81	0.71	0.73	0.96	0.90	0.04	0.90	0.77
Cuvas2 [67]	0.84	0.78	0.86	0.65	0.92	0.62	0.90	0.79	0.63	0.77	0.78
FgSegNet-S [15]	0.20	0.60	0.53	0.25	0.60	0.28	0.19	0.16	0.05	0.18	0.30
FgSegNet-M [15]	0.56	0.55	0.65	0.42	0.56	0.19	0.28	0.18	0.01	0.33	0.37
MSFS [49]	0.53	0.58	0.25	0.41	0.63	0.25	0.54	0.54	0.05	0.29	0.41
3DCD	**0.86**	**0.49**	**0.93**	**0.85**	**0.79**	**0.87**	**0.87**	**0.87**	**0.49**	**0.83**	**0.79**

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 VI
COMPARATIVE F-SCORE PERFORMANCE IN SDE FRAMEWORK ON LASIESTA DATASET

Method	ISI	ICA	IOC	IIL	IMB	IBS	OCL	ORA	OSN	OSU	Avg
Zivkovik [1]	0.91	0.83	0.95	0.24	0.87	0.53	0.88	0.83	0.38	0.71	0.71
Maddalena1 [63]	0.87	0.85	0.91	0.61	0.76	0.42	0.88	0.84	0.58	0.80	0.75
Maddalena2 [64]	0.95	0.86	0.95	0.21	0.91	0.40	0.87	0.85	0.81	0.88	0.77
Cuevas1 [65]	0.79	0.74	0.85	0.79	0.73	0.58	0.86	0.81	0.46	0.73	0.73
Haines [66]	0.89	0.89	0.92	0.85	0.84	0.68	0.83	0.86	0.17	0.86	0.78
Cuvas2 [67]	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 [15]	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 [15]	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 [15]	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 [15]	0.43	0.69	0.31	0.32	0.71	0.21	0.22	0.18	0.19	0.25	0.35
MSFS-51 [49]	0.44	0.60	0.30	0.32	0.50	0.22	0.31	0.24	0.28	0.38	0.36
MSFS-55 [49]	0.39	0.40	0.37	0.35	0.64	0.36	0.41	0.35	0.31	0.37	0.40
3DCD-51	**0.91**	**0.76**	**0.90**	**0.90**	**0.90**	**0.81**	**0.89**	**0.89**	**0.72**	**0.85**	**0.85**
3DCD-55	**0.87**	**0.82**	**0.91**	**0.92**	**0.89**	**0.72**	**0.87**	**0.90**	**0.69**	**0.85**	**0.84**

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

Method	Cand	CAV2	CaV	HigII	Avg
FgSegNet-S [15]	0.23	0.11	0.68	0.24	0.31
FgSegNet-M [15]	0.15	0.14	0.72	0.21	0.30
MSFS [49]	0.27	0.10	0.63	0.58	0.40
3DCD	**0.67**	**0.62**	**0.53**	**0.59**	**0.60**

Cand: Candela-m1.10, CAV2: CAVIAR2, CaV: CaVignal, HigII: HighwayII

Table II. The proposed method achieves overall precision, recall, F-score, PWC of 0.90, 0.83, 0.86, 0.76, respectively. These results reflect the robustness of the proposed network in unseen videos. We also compare our work with 38 state-of-the-art background subtraction methods as given in Table III. Since 3DCD is evaluated in scene independent setup (video-agnostic), comparing it with video-optimized or video-group-optimized algorithms (evaluated in the SDE setups) would not be fair. Thus, we only compare with the existing methods evaluated in the SIE setup. We also train and evaluate the existing networks FgSegNet-S [15], FgSegNet-M [15], MSFS [49] in the same SIE setup for comparative analysis. The proposed 3DCD outperforms FgSegNet-S, FgSegNetM, MSFS by 41%, 48%, 36%, respectively in CDnet 2014. As the model is evaluated on unseen videos, the proposed 3DCD is better generalized to handle unseen scenarios.

LASIESTA: We evaluate the model on 10 completely unseen videos in LASIESTA dataset. The result of the proposed 3DCD is compared with the existing state-of-the-art methods in Table V. Our model comfortably outperforms the existing handcrafted and deep learning approaches for change detection. More specifically, our method significantly outperforms the deep learning methods FgSegNet-S, FgSegNet-M, MSFS by 49%, 42%, 38%, respectively, which highlights the superior generalization capability of our model.

SBMI2015: We evaluate the proposed and the existing methods in 4 completely unseen videos in SBMI2015. The comparative results in the SIE setup are given in Table VII. The 3DCD outperforms the existing state-of-the-art methods. More specifically, it (0.60) achieves an overall 20% performance improvement over the MSFS (0.40).

3) Quantitative results in the SDE setup: We also conduct experiments in the SDE setup in order to present a comparative analysis of the proposed model with existing deep learning methods which prominently follow the SDE setup. For scene dependent evaluation, we temporally divide the videos with a 50:50 ratio. The training is performed with the initial 50% of frames and the evaluation is performed using the remaining 50% of frames as well as using the complete 100% frames.

CDnet 2014: The performance of the proposed and existing state-of-the-art approaches in CDnet 2014 are tabulated in Table IV. As shown in Table IV, the proposed 3DCD outperforms the best performing handcrafted method by 7%. It also outperforms the recent state-of-the-art deep learning models DeepBS [20], MSFgNet [25], SFEN(VGG) [38], VGG+CRF

TABLE VIII
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 [15]	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 [15]	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 [15]	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 [15]	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 [49]	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 [49]	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
3DCD-51	**0.85**	**0.31**	**0.81**	**0.58**	**0.55**	**0.66**	**0.63**	**0.73**	**0.79**	**0.67**	**0.74**	**0.80**	**0.74**	**0.68**
3DCD-55	**0.83**	**0.35**	**0.79**	**0.56**	**0.48**	**0.69**	**0.58**	**0.73**	**0.77**	**0.65**	**0.70**	**0.78**	**0.76**	**0.67**

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

[38], VGG+PSL+CRF [38], GoogLeNet+PSL +CRF [38], EDS-CNN [16] by 8%, 3%, 15%, 13%, 5%, 14%, 2%, respectively. The overall F-score of the proposed method is equal to Cascade-CNN [48]. However, we notice 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 each video which is not suitable for real-world applications. Whereas, the proposed 3DCD 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 [15], FgSegNet-M [15], and MSFS [49]) in the same SDE setup for a fair comparative analysis. The proposed 3DCD outperforms FgSegNet-S, FgSegNet-M, MSFS by a margin of 14%, 15%, 9%, respectively in CDnet 2014.

LASIESTA: The comparison of 3DCD with the existing methods in terms of average F-score in each video category of LASIESTA is shown in Table VI. From quantitative analysis (see in Table VI), it is evident that the proposed 3DCD outperforms in five out of ten categories of LASIESTA for foreground detection. The overall F-score of proposed 3DCD (0.85) is significantly improved from 0.79 (highest value from existing methods). Moreover, the 3DCD obtains 54%, 50%, 49% better F-score as compared to the deep learning methods FgSegNet-S, FgSegNet-M, MSFS, trained and evaluated in the same SDE setup for a fair evaluation.

SBMI2015: The SDE results for SBMI2015 is tabulated in Table VIII. The overall F-score of the proposed method (0.68) is 5% higher than the best performing existing method (0.63). All the existing deep learning methods have been trained and evaluated in the same SDE setup as the proposed method.

E. Qualitative Results

We also present a qualitative analysis through visual comparison in Fig. 7. We select videos from challenging scenarios 'night videos' (1st, 3rd, and 5th rows), 'intermittent object motion' (2nd row), and 'bad weather' (4th) for evaluation. The visual responses are compared with two deep learning methods DeepBS [20], Cascade-CNN [48], and one hand-crafted method IUTIS-5 [33]. It can be observed that 3DCD produces the best visual results in all the categories. A robust model must be able to eliminate both false positives (FP) and false negatives (FN) across different scenarios. The existing methods work well in certain categories but suffer from either higher FP or FN in other categories. In contrast, our proposed method can separate the salient foreground objects from the background and highlight them uniformly across different categories. For example, in row 2 and row 5, the other approaches are able to detect foreground but they also produce a lot of false positives which leads to performance degradation in future processing of the segmented response. The proposed 3DCD produces fewer false positives as compared to other methods which is evident from its high precision (0.90) reported in Table II. This is due to robust background estimation with spatiotemporal feature aware GRBE block and contrasting feature extraction in the FSR block. Thus, the effective model design of 3DCD enables it to exclude the background interference or noise of various types in different categories, leading to improved performance in comparison to other approaches.

TABLE IX
ABLATION STUDY OF THE PROPOSED 3DCD IN BAD WEATHER (BW) AND BASELINE (BL) CATEGORIES

	Components						
Model	**RC**	**CMF**	**FSR**	**CFD**	**Multi-Stream**	**BW**	**BL**
3DCD	Y	Y	Y	Y	Y	0.94	0.93
3DCD-v2	Y	Y	Y	Y	N	0.92	0.85
3DCD-v3	N	Y	Y	Y	Y	0.82	0.83
3DCD-v4	Y	N	Y	Y	Y	0.74	0.88
3DCD-v5	Y	Y	N	Y	Y	0.87	0.88
3DCD-v6	Y	Y	Y	N	Y	0.88	0.92
3DCD-v7	40 history frames as input					0.91	0.89
3DCD-v8	30 history frames as input					0.88	0.86
3DCD-v9	20 history frames as input					0.86	0.82
3DCD-v10	10 history frames as input					0.85	0.82

F. Ablation Studies

We investigate the influence of different components of 3DCD through ablation experiments. In order to quantify the effect of each block (MScE-MScD, RC, CMF, FSR, and CFD) in 3DCD, we conduct multiple experiments over two categories 'bad weather' and 'baseline'. We used a single stream encoder-decoder (3DCD-v2) to replace multi-schematic MScE-MScD in order to assess its importance in 3DCD. Similarly, we create variants of the proposed method by

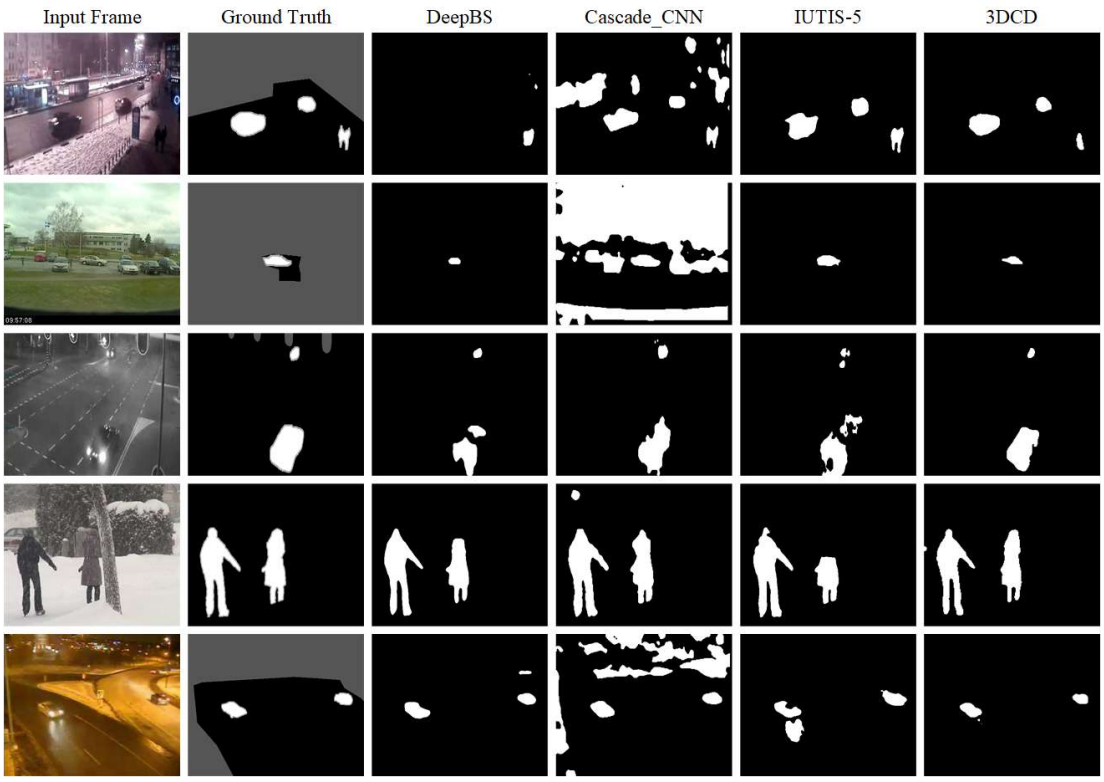

Fig. 7: Qualitative analysis of the proposed 3DCD with existing state-of-the-art approaches

TABLE X
THE PROPOSED 3DCD IS COMPARED WITH EXISTING DEEP LEARNING METHODS IN TERMS OF SPEED, MEMORY AND COMPUTATIONAL COMPLEXITY

Method	#Param	BMC	*FPS
FCSN [18]	$\sim 1.73M$	Yes	NA
FgSegNet [15]	$\sim 2.60M$	No	18
MFCN [17]	$\sim 20.83M$	No	27
EDS-CNN [16]	$\sim 18.64M$	Yes	NA
Trip-Net [19]	$\sim 0.33M$	Yes	NA
DeepBS [20]	$\sim 3.15M$	Yes	NA
Deep-ConvNet [21]	$\sim 4.40K \times P$	Yes	NA
Msednet [71]	$\sim 2.88K$	Yes	9.7
Cascade-CNN [48]	$\sim 0.25M$	No	13
VGG16 [38]	$\sim 31.92M$	No	4.9
GoogLeNet [38]	$\sim 6.02M$	No	NA
ResNet50 [38]	$\sim 23.78M$	No	NA
2D-CNN-LSTM [24]	$\sim 0.29M$	No	15
3D-CNN-LSTM [24]	$\sim 0.22M$	No	24
3DCD	$0.13M$	**No**	**25**

#Param: Number of trainable parameters, M: Millions, K: Thousands, P: Number of image patches, NA: Not Available, BMC: Separate computation cost for background modeling, *FPS is reported as given in the original papers

dropping RC (3DCD-v3), CMF (3DCD-v4), FSR (3DCD-v5), and CFD (3DCD-v6) blocks from the original network. We also conduct experiments by taking 40, 30, 20, and 10 previous frames as input to the network, denoting the models as 3DCD-v7, 3DCD-v8, 3DCD-v9, and 3DCD-v10, respectively. These changes are made separately to the original network design. The experimental results for all these variants are shown in Table IX. It is evident that removing any of the blocks from 3DCD results in lower performance in both categories. Similarly, we observe that the network (3DCD) with 50 input frames gives the best performance. The diverse results in different combination of modules also shows that certain components of 3DCD are more crucial than others in a particular scenario. The customizable nature of the proposed network is suitable for obtaining improved performance for a specific scenario by making changes at the block level. These results further justify the effectiveness of the original 3DCD model design.

G. Speed, Memory, and Computational Complexity Analysis

The proposed network consists of 0.13 million trainable parameters with a model size of 1.16 MB. The inference speed is 40 ms per frame or approximately 25 frames per second (FPS) over Titan Xp. We compare the proposed 3DCD with the existing state-of-the-art change detection techniques in terms of speed and computational complexity in Table X. From Table X, it is evident that our method is computationally more efficient than the existing approaches. Our method also achieved superior speed (25 FPS) which is the highest amongst all other methods except [17] which is computationally expensive. The memory consumption of 3DCD is much lower (only 1.16 MB), which makes it suitable for embedded devices used in real-time applications. Moreover, it can be noticed that the small and shallower networks [18], [24] including 3DCD have an overall advantage over the large/deeper networks [16], [17], [38] in terms of overall performance (accuracy, computational efficiency, and speed). Thus, an aptly designed small and shallow network such as 3DCD which performs well in all

three performance metrics is a valuable contribution for change detection applications.

V. Conclusion

This paper introduces a novel spatiotemporal end-to-end deep-learning model, 3DCD, for change detection in unseen videos. We also propose a scene independent evaluation scheme to segregate the train and test videos using a leave-one-video-out strategy. The input to 3DCD consists of the current and past history frames which enable online inference. The network is comprised of multiple customizable learning blocks to estimate the background, learn contrasting features for motion estimation, and finally refine the contrasting features through multi schematic encoder and decoder networks to learn structurally diverse features at multiple scales. These effective modules increase the generalization capacity of 3DCD to diverse change scenarios. Moreover, as opposed to the scene dependent evaluation schemes widely used in the literature, we present a scene independent data division and evaluation strategy to effectively evaluate the generalization capability of the designed network for real-world applications. We present the theoretical analysis, constituent block visualization, qualitative, and quantitative results to demonstrate the effectiveness of the proposed 3DCD. Experimental results on CDnet-2014, LASIESTA, and SBMI2015 show that 3DCD outperforms state-of-the-art algorithms in both scene independent and scene dependent setups. This shows the great potential of 3D-CNN based algorithms designed for unseen videos. Our online model is very fast (speed-25 fps) and lightweight (model size-0.16 MB), making it suitable for real-time applications.

Acknowledgement

The authors would like to thank all the members of the Vision Intelligence Lab for their support. We also thank the change detection researchers for providing the video-wise results of their papers for comparison.

References

[1] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2. IEEE, 2004, pp. 28–31.

[2] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2. IEEE, 1999, pp. 246–252.

[3] S. Varadarajan, P. Miller, and H. Zhou, "Spatial mixture of gaussians for dynamic background modelling," in *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2013, pp. 63–68.

[4] S. Jiang and X. Lu, "Wesambe: A weight-sample-based method for background subtraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2105–2115, 2017.

[5] B. Wang and P. Dudek, "A fast self-tuning background subtraction algorithm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 395–398.

[6] H. Sajid and S.-C. S. Cheung, "Background subtraction for static & moving camera," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4530–4534.

[7] G. Ramírez-Alonso and M. I. Chacón-Murguía, "Auto-adaptive parallel som architecture with a modular analysis for dynamic object segmentation in videos," *Neurocomputing*, vol. 175, pp. 990–1000, 2016.

[8] 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 Computer Vision*, vol. 13, no. 1, pp. 31–43, 2018.

[9] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, 2014.

[10] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2012, pp. 38–43.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] M. Mandal, M. Shah, P. Meena, S. Devi, and S. K. Vipparthi, "Avdnet: A small-sized vehicle detection network for aerial visual data," *IEEE Geoscience and Remote Sensing Letters*, 2019.

[14] M. Mandal, M. Shah, P. Meena, and S. K. Vipparthi, "Sssdet: Simple short and shallow network for resource efficient vehicle detection in aerial scenes," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3098–3102.

[15] L. A. Lim and H. Y. Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," *Pattern Recognition Letters*, vol. 112, pp. 256–262, 2018.

[16] K. Lim, W.-D. Jang, and C.-S. Kim, "Background subtraction using encoder-decoder structured convolutional neural network," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–6.

[17] D. Zeng and M. Zhu, "Multiscale fully convolutional network for foreground object detection in infrared videos," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 617–621, 2018.

[18] C. Lin, B. Yan, and W. Tan, "Foreground detection in surveillance video with fully convolutional semantic network," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4118–4122.

[19] 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 Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 433–446, 2018.

[20] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognition*, vol. 76, pp. 635–649, 2018.

[21] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *2016 international conference on systems, signals and image processing (IWSSIP)*. IEEE, 2016, pp. 1–4.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[23] L. Yang, J. Li, Y. Luo, Y. Zhao, H. Cheng, and J. Li, "Deep background modeling using fully convolutional network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 254–262, 2017.

[24] T. Akilan, Q. J. Wu, A. Safaei, J. Huo, and Y. Yang, "A 3d cnn-lstm-based image-to-image foreground segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[25] P. W. Patil and S. Murala, "Msfgnet: A novel compact end-to-end deep network for moving object detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4066–4077, 2018.

[26] 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 *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4018–4022.

[27] P. Patil and S. Murala, "Fggan: A cascaded unpaired learning for background estimation and foreground segmentation," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1770–1778.

[28] H. Sajid and S.-C. S. Cheung, "Universal multimode background subtraction," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3249–3260, 2017.

[29] M. Mandal, P. Saxena, S. K. Vipparthi, and S. Murala, "Candid: Robust change dynamics and deterministic update policy for dynamic background subtraction," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2468–2473.

[30] P.-L. St-Charles and G.-A. Bilodeau, "Improving background subtraction using local binary similarity patterns," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2014, pp. 509–515.

[31] H. Wang and D. Suter, "A consensus-based method for tracking: Modelling background scenario and foreground appearance," *Pattern recognition*, vol. 40, no. 3, pp. 1091–1105, 2007.

[32] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image processing*, vol. 20, no. 6, pp. 1709–1724, 2010.

[33] S. Bianco, G. Ciocca, and R. Schettini, "Combination of video change detection algorithms by genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 6, pp. 914–928, 2017.

[34] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Static and moving object detection using flux tensor with split gaussian models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 414–418.

[35] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4041–4049.

[36] H. Hu and G.-J. Qi, "State-frequency memory recurrent neural networks," in *International Conference on Machine Learning*, 2017, pp. 1568–1577.

[37] P. Gao, Z. Jiang, H. You, P. Lu, S. C. H. Hoi, X. Wang, and H. Li, "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] Y. Chen, J. Wang, B. Zhu, M. Tang, and H. Lu, "Pixel-wise deep sequence learning for moving object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[39] G.-J. Qi, "Hierarchically gated deep networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[40] Y. Hu, Y. Chen, X. Li, and J. Feng, "Dynamic feature fusion for semantic edge detection," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 782–788. [Online]. Available: https://doi.org/10.24963/ijcai.2019/110

[41] M. Mandal, L. K. Kumar, M. S. Saran, and S. K. vipparthi, "Motionrec: A unified deep framework for moving object recognition," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[42] M. Mandal, L. K. Kumar, and S. K. Vipparthi, "Mor-uav: A benchmark dataset and baselines for moving object recognition in uav videos," *arXiv preprint arXiv:2008.01699*, 2020.

[43] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 405–420.

[44] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6810–6818.

[45] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3684–3692.

[46] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[47] P. Bilinski and V. Prisacariu, "Dense decoder shortcut connections for single-pass semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6596–6605.

[48] Y. Wang, Z. Luo, and P.-M. Jodoin, "Interactive deep learning method for segmenting moving objects," *Pattern Recognition Letters*, vol. 96, pp. 66–75, 2017.

[49] L. A. Lim and H. Y. Keles, "Learning multi-scale features for foreground segmentation," *Pattern Analysis and Applications*, pp. 1–12, 2019.

[50] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[51] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[52] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "A self-adjusting approach to change detection based on background word consensus," in *2015 IEEE winter conference on applications of computer vision*. IEEE, 2015, pp. 990–997.

[53] M. Braham, S. Piérard, and M. Van Droogenbroeck, "Semantic background subtraction," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4552–4556.

[54] O. Tezcan, P. Ishwar, and J. Konrad, "Bsuv-net: A fully-convolutional neural network for background subtraction of unseen videos," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2774–2783.

[55] V. Mondéjar-Guerra, J. Rouco, J. Novo, and M. Ortega, "An end-to-end deep learning approach for simultaneous background modeling and subtraction," in *British Machine Vision Conference (BMVC), Cardiff*, 2019.

[56] Y.-Q. Chen, Z.-L. Sun, and K.-M. Lam, "An effective subsuperpixel-based approach for background subtraction," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 1, pp. 601–609, 2019.

[57] J. Liao, G. Guo, Y. Yan, and H. Wang, "Multiscale cascaded scene-specific convolutional neural networks for background subtraction," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 524–533.

[58] C. Zhao, T.-L. Cham, X. Ren, J. Cai, and H. Zhu, "Background subtraction based on deep pixel distribution learning," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[59] S. Choo, W. Seo, D.-j. Jeong, and N. I. Cho, "Learning background subtraction by video synthesis and multi-scale recurrent networks," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 357–372.

[60] M. R. Arefin, F. Makhmudkhujaev, O. Chae, and J. Kim, "Background subtraction based on fusion of color and local patterns," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 214–230.

[61] T. Minematsu, A. Shimada, and R.-i. Taniguchi, "Simple background subtraction constraint for weakly supervised background subtraction network," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[62] S. Choo, W. Seo, D.-j. Jeong, and N. I. Cho, "Multi-scale recurrent encoder-decoder network for dense temporal classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 103–108.

[63] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1168–1177, 2008.

[64] L. Maddalena and A. Petrosino, "The sobs algorithm: What are the limits?" in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 21–26.

[65] C. Cuevas and N. García, "Improved background modeling for real-time spatio-temporal non-parametric moving object detection strategies," *Image and Vision Computing*, vol. 31, no. 9, pp. 616–630, 2013.

[66] T. S. Haines and T. Xiang, "Background subtraction with dirichlet-process mixture models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 4, pp. 670–683, 2013.

[67] D. Berjón, C. Cuevas, F. Morán, and N. García, "Real-time nonparametric background subtraction with tracking-based foreground update," *Pattern Recognition*, vol. 74, pp. 156–170, 2018.

[68] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "Cdnet 2014: An expanded change detection benchmark dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 387–394.

[69] C. Cuevas, E. M. Yáñez, and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: Lasiesta," *Computer Vision and Image Understanding*, vol. 152, pp. 103–117, 2016.

[70] L. Maddalena and A. Petrosino, "Towards benchmarking scene background initialization," in *International conference on image analysis and processing*. Springer, 2015, pp. 469–476.

[71] P. Patil, S. Murala, A. Dhall, and S. Chaudhary, "Msednet: multi-scale deep saliency learning for moving object detection," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 1670–1675.