Multi-frame Recurrent Adversarial Network for Moving Object Segmentation

Prashant W. Patil, Akshay Dudhane and Subrahmanyam Murala
Computer Vision and Pattern Recognition Lab, Indian Institute of Technology Ropar, India
{2017eez0006, 2017eez0001, subbumurala} @iitrpr.ac.in

Abstract

Moving object segmentation (MOS) in different practical scenarios like weather degraded, dynamic background, etc. videos is a challenging and high demanding task for various computer vision applications. Existing supervised approaches achieve remarkable performance with complicated training or extensive fine-tuning or inappropriate training-testing data distribution. Also, the generalized effect of existing works with completely unseen data is difficult to identify. In this work, the recurrent feature sharing based generative adversarial network is proposed with unseen video analysis. The proposed network comprises of dilated convolution to extract the spatial features at multiple scales. Along with the temporally sampled multiple frames, previous frame output is considered as input to the network. As the motion is very minute between the two consecutive frames, the previous frame decoder features are shared with encoder features recurrently for current frame foreground segmentation. This recurrent feature sharing of different layers helps the encoder network to learn the hierarchical interactions between the motion and appearance-based features. Also, the learning of the proposed network is concentrated in different ways, like disjoint and global training-testing for MOS. An extensive experimental analysis of the proposed network is carried out on two benchmark video datasets with seen and unseen MOS video. Qualitative and quantitative experimental study shows that the proposed network outperforms the existing methods.

1. Introduction

Computer vision applications are gaining more demand in day-to-day life. This necessitates an ample amount of data for various intelligent video processing applications. The video data of automatic traffic and surveillance system has high temporal redundancy. More than 70% pixel information from each video frame is irrelevant for a different high-level processing task. Also, the visibility of foreground objects decreases drastically in the outdoor seen like bad weather, dynamic background, etc. Due to the redundant information and poor visibility, the performance of various video processing applications with artificial intelligence is degraded. By eliminating these issues, the effective and reliable solutions for foreground localization are proposed by many researchers [22], [6], [21], [16], [14], [19], [29], [25], [12], [27], [32], [47] to improve the performance for automated video analysis applications. Thus, moving object segmentation (MOS) is an active research area in the field of video analysis.

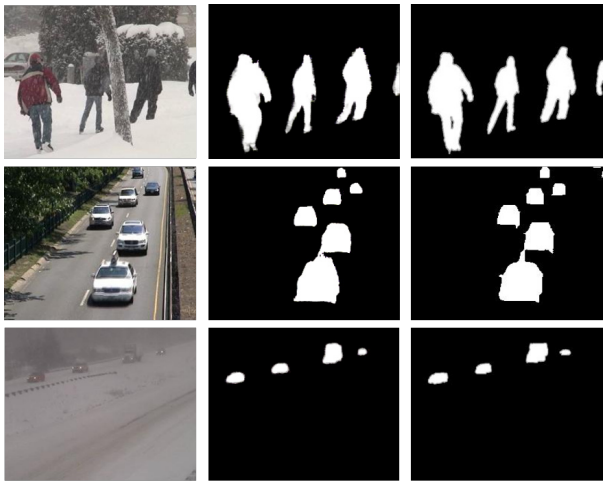

Figure 1. Foreground localization results comparison of proposed method with respective ground-truth.

Hand-crafted feature-based methods are widely used methods for MOS task [22], [6], [21], [16]. In hand-crafted methods, frame difference with appropriate threshold, region (local as well as global) level, background subtraction and saliency based methods are most widely used.

Deep learning-based methods are gaining more demand in recent time for MOS task [19], [29], [12], [47]. The learning-based techniques are more powerful for extracting low-level, mid-level and high-level features from images or video frames. However, from literature, it is observed that all CNN based methods are not able to give good performance in the presence of different outdoor environments. Also, there is a need for fine-tuning of a pre-trained model

to achieve good performance.

Generative adversarial learning (GAN) based methods give significant improved results in many computer vision applications like varicolored image dehazing [10], underwater image enhancement [11], single image depth estimation [13], image de-raining [49], foreground-background segmentation [30], video super-resolution [39], underwater [34], etc.

In this work, a novel end-to-end recurrent feature sharing based generative adversarial network is proposed with seen and unseen video analysis for MOS. The literature of the existing methods is discussed in the next section.

2. Literature Survey

The main objective of any MOS approach is to detect moving objects, for which motion is a prominent characteristic. Also, handling of diverse practical scenarios like dynamic backgrounds, degraded weather and unbalanced object motion in video frames is not possible with the help of spatial information. Thus, both spatial and temporal features are necessary for any foreground-background segmentation algorithm. Lin *et al.* [22] proposed a background subtraction based approach for MOS by utilizing the concept of the hyper-bit plane with spatial and temporal information. The existing methods for moving object detection (MOD) based on low rank and sparse decomposition are able to reduce the effect of Gaussian noise. Recently, Chen *et al.* [6] proposed unstructured regularized low-rank representation based method for MOD. Linhao *et al.* [21] proposed hierarchical background subtraction and foreground segmentation modeling approach with alternating optimization technique for MOD. Javed *et al.* [16] proposed superpixel based spatio-temporal structured sparse robust principal component analysis based technique for MOD with spatial and temporal regularization.

In recent work, the different MOD techniques successfully brought significant performance with convolutional neural networks (CNN). In [20], language referring expressions are used for MOS. Here, language specifications make the system more robust to background clutter, complex dynamic seen. Yuhua *et al.* [7] proposed pixel-wise metric learning for video frame segmentation by considering the reference frame and its segmented mask as input to the embedding network. The spatial and temporal dependencies are encoded in [5] using a trained CNN model and optical flow repectievly. Seoung *et al.* [46] proposed an identical encoder network to process the key frame and reference frame interdependently. Further, these individual features are concatenated using a global convolution network. Finally, a refinement module with residual learning is used for fast MOS.

Some of the researchers used tracking-based methods [9], [42] to detect the region-of-interest for MOS. To overcome the large deformation, occlusion, and cluttered background problem, the tracking-based method is proposed [9] for accurate online MOS. Tracking is used to find the region of interest to be segmented. Similarly, in [42] depthwise cross correlation-based approach is proposed for object tracking and segmentation task. A static and dynamic visual attention prediction approach is proposed in [43] for MOS. Paul *et al.* [41] proposed semantic pixel-wise features concatenation with global and local feature matching techniques for MOS. The combination of probabilistic generative approach and backbone feature extractor with a predication module is proposed for MOS [18]. Ping *et. al* [14] proposed dynamic identity propagation and attention network for MOS. They utilized the concept of lightweight fine-tuning on the first frame of the test video. Along with motion and appearance features, Lu *et al.* [24] has proposed a co-attention mechanism to improve the discriminative foreground representations. Mandal *et. al* proposed an approaches for MOS with the help of temporal depth reductionist based background estimation [27], [26], [28].

Propagation based encoder-decoder network with ranking attention module is proposed by Ziqin *et al.* [45] for MOS. To select and rank the effective foreground feature maps based on propagation and matching, the ranking attention module is utilized. Another approach based on motion and appearance features with a parallel processing network along with a memory module is proposed in [40]. In [40], independent object motion between the number of successive frames, object appearance and temporal consistency parameters are considered to impose additional constraints on the segmentation. Single and multiple object segmentation using lucid data dreaming is proposed by Khoreva *et al.* [19]. A fascinating approach without temporal dependencies is proposed by Maninis *et al.* [29] using the learned features on ImageNet as a pre-trained model. In [12], Griffin proposed a technique for MOS without any training or ground-truth requirement. Kai *et al.* [47] proposed two-stream network to get the effective spatial and temporal information for MOS.

Receptive field-based methods show significant improvement for various computer vision applications like fast object detection [23], video salient object detection (VSOD) [37], etc. In [23], a hand-crafted approach is proposed with a receptive field block to enhance the discriminability and robustness of features. Similarly, pyramid dilated bidirectional ConvLSTM is proposed in [37] to extract the multi-scale spatio-temporal features for VSOD. Dilated convolution helps the network to learn the spatial feature with different scales. Saliency learning with a compact encoder-decoder based approach is proposed in [31] for MOS. Yang *et al.* [48] proposed a background modeling approach with atrous convolution, residual block and deep network to estimate spatial information, to avoid degradation problems and

to capture temporal information respectively. The learning-based combination of background estimation, saliency estimation and foreground detection is proposed in [33] with small video streams obtained from the original video. Akilan *et al.* proposed different approaches for video surveillance applications with 3D transpose convolution and residual connections [2], encoder-decoder CNN technique with the help of multi-view receptive field [3] and slow encoder-decoder with strided convolution [1]. In [2], [3], and [1], authors trained the network on one baseline video and fine-tuned on target video frames with learned parameters as initial weights for accurate foreground detection. Recently, Zhou *et al.*[50] proposed a novel interleaved two-stream network architecture to learn powerful spatio-temporal features for MOS.

Tang *et al.* [38] proposed cascaded CNN based encoder-decoder approach with adversarial learning to estimate the global saliency and the residual network is used to estimate the local saliency for salient object detection. Similarly, Dudhane *et al.* [11] proposed the cycle-consistent GANs with generator of encoder-decoder architecture for underwater image enhancement. Also, Zhang *et al.* [49] proposed conditional GANs with dense connections and local-global information for image de-raining.

From the above literature, it is observed that, the existing methods have improved the performance significantly for MOS. But, it requires complicated training or extensive fine-tuning or inappropriate training-testing data distribution. Therefore, the generalized effect of this works with a completely unseen video is difficult to identify. Thus, a novel end-to-end recurrent feature sharing based generative adversarial network is proposed with seen and unseen video analysis for MOS. Figure 1 illustrates the performance comparison of the proposed method with ground-truth MOS. The major contributions of the proposed work are illustrated below:

1. An end-to-end multi-frame recurrent feature sharing adversarial learning network is proposed with seen and unseen video analysis.
2. Spatio-temporal structural dependencies are learned with the help of dilated convolution, temporally sampled successive video frame and recurrent feature sharing.
3. Multi-scale residual block with dense connections is proposed to learn prominent features related to the foreground and refinement module with the residual block is proposed for moving object segmentation.

Extensive experimental analysis of the proposed method with disjoint (**unseen video analysis**) and global (**seen video analysis**) training-testing is done on two benchmark video datasets namly DAVIS-2016 [35] and ChangeDetection.net (CDnet)-2014 [44]).

3. Proposed Framework

In literature, various researchers have taken advantage of the pre-trained CNN model to extract pixel-wise semantic features or to fine-tune the pre-trained network for the MOS task. Recently, GAN-based methods give a significant improvement in various applications like image style transfer, image enhancement, etc. Basically, GAN based approaches have two networks, namely generator and discriminator. The main aim of the discriminator network is to distinguish authentic distribution with the estimated generator output while the generator is designed to generate the fake distribution to fool the discriminator network. The methods used for any automated video processing application need to process a large amount of training data. Also, the training of a deeper network undergoes the vanishing gradients problem. To overcome these limitations, an end-to-end adversarial network with the recurrent, multi-scale residual block with dense connection and refinement module is proposed for foreground localization. The temporal information between successive video frames is captured using multi-frame selection with temporal sampling and recurrent technique. Dilated convolution at multiple scales with different sampling rates is used to extract spatial features. Thus, both the spatial and temporal features are encoded in the proposed network. The detailed information related to each block of the proposed framework is shown in Figure 2.

While designing the network, the choice of the filter size plays an important role for better feature learning for the specified task. To resolve this issue, a residual module with dense connection and multi-scale convolution filters are proposed. Finally, the refinement module with a residual block is proposed to produce the mask output. The output of previous frames along with three successive temporal sampled video frames $\{(t_{in},\ (t-3)_{in},\ (t-6)_{in},\ (t-1)_{out})\ , ...\}$ are used as input to estimate the current frame foreground objects.

The proposed network comprises of encoder block followed by dilated convolution (`DCB`), multi-scale residual block with dense connection (`MRDC`) and refinement module with residual block (`RfrB`). The encoder block (`EnC`) is defined as 3×3 convolution followed by batch Normalization with ReLU having `'m'` filters and stride `'n'` `EnCy_3_sdn-m`. Similarly, the refinement module with a residual block having 64 filters is denoted as `RfrBk-64`. Finally, decoder block (`Ded`) is used with 3×3 convolution followed by batch Normalization with ReLU having `'m'` filters and stride `'n'` is noted as `Ded3sdn-m`. Thus, the proposed network is defined as: `EnC1_3_sd1-64`, `EnC2_3_sd2-128`, `EnC3_3_sd2-192`, `EnC4_3_sd2-256`, `DCB-64`, `MRDCc-64 (c`ϵ`(1,2,3))`, `RfrB1-192`, `RfrB2-128`, `Ded3sd2-64`.

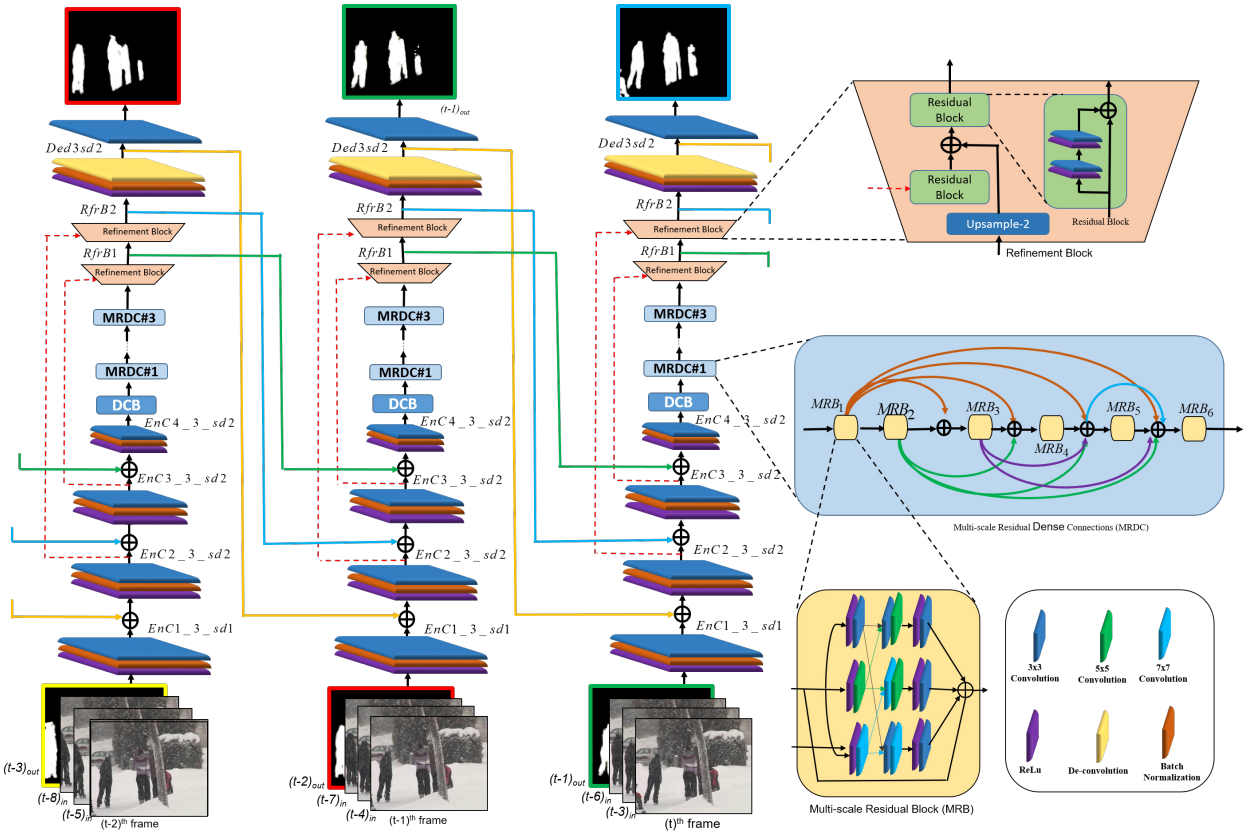

Figure 2. Proposed system framework for foreground localization.

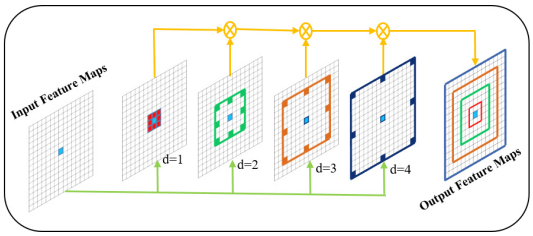

Figure 3. Visualization of dilated convolution block (DCB).

The detailed information related to each layer with mathematical expression is given below:

Convolutional layer is very important and basic operation for learning based methods. The convolution filters are used to extract the information with different kernels. In general, convolution operation is explained as,

$$\Re_{S_j}^{st_l} = b_j^l + \sum_{i \in R_j} \left(M_i^{(l-1)} \circledast \psi_{S_{i,j}}^{st_l} \right); st \in (1,2); S \in (1,3,5) \quad (1)$$

where, $M_i^{(l-1)}$ is input, $\Re_{S_j}^{st_l}$ represents j^{th} feature map of l^{th} layer, R_j represents feature map , $\psi_{S_{i,j}}^{st_l}$ is convolution kernel with stride factor (st) and filter size (S), $\circledast$ and b_j^l are convolution operation and bias factor.

Encoder block is first block of the proposed network to extract low level feature. The mathematical expression of EnC is,

$$EnC_{S_j}^{st_l} = \Phi \left(\xi^l \left(b_j^l + \sum_{i \in R_j} \left(M_i^{(l-1)} \circledast \psi_{S_{i,j}}^{st_l} \right) \right) \right) \quad (2)$$

where, ψ_S^{st} indicates convolution operation with filter size $S \times S$ (3×3) with stride factor $st \in (1,2,2,2)$ to downsample the feature maps, ξ and Φ represents batch normalization and ReLu activation function respectively. In proposed method, four encoder blocks are used with different stride factor to extract low level feature. Here, the extracted feature of decoder block from previous frame are given as input to the respective encoder block along with previous layer output. Mathematical expression for each encoder layer input is given as,

$$EnC2\_3\_sd2 = EnC1\_3\_sd1 \oplus Ded3sd2 \quad (3)$$

$$EnC3\_3\_sd2 = EnC2\_3\_sd2 \oplus RfrB2 \quad (4)$$

$$EnC4\_3\_sd2 = EnC3\_3\_sd2 \oplus RfrB1 \quad (5)$$

Recurrent make sense that the output/feature maps from previous frames acts as a feedback signal which is recursively passed through the network and network parameters

at different time instances are shared.

Dilated Convolution: After encoder blocks, dilated convolution block (DCB) is used to extract multi-scale contextual information. Dilated convolution block is specifically used for prediction due to its capability to expand the receptive field without losing resolution. The aggregated features from DCB block are fed to multi-scale residual block with dense connection (MRDC) to learn prominent feature related to foreground. Detailed visualization of DCB block is shown in Figure 3. In MRDC module, multi-scale convolution filters with residual connection are used named as multi-scale residual block (MRB). The technique used for dense connections is given as,

$$MRB_n = \sum_{i=1}^{n-1} MRB_i \;\; ; \; n > 1 \tag{6}$$

where, MRB_n is input to the n^{th} MRB module, MRB_i is response of i^{th} MRB module and $n \in (1, 6)$. Each MRB uses parallel convolution filters with kernel size of 3×3, 5×5 and 7×7 followed by ReLU. Here, to integrate the features learned by the respective convolution block with different scales, we employed feature concatenation operation followed by a convolution block. Further, these features are added to get robust features learned by different scales (*for more details, please refer MRB and MRDC block from Figure 2*).

The proposed decoder network consists of two refinement modules, a decoder block, a final convolution layer to generate the foreground object. The output from MRDC block is concatenated with target encoder stream through skip-connections and given to proposed refinement block to produce a foreground localization output. To merge the different scale feature efficiently, we used two refinement modules as the building block of proposed network. After refinement blocks, one decoder blocks is used to map the extracted features into original input size. Decoder block (Ded) is explained as,

$$Ded_{S_j}^{st_l} = \Phi\left(\xi^l\left(b_j^l + \sum_{i \in R_j}\left(M_i^{(l-1)} \circledast \overline{\psi}_{S_{i,j}}^{st_l}\right)\right)\right) \tag{7}$$

where, $\overline{\psi}_S^{st}$ indicates deconvolution operation with filter size $S \times S$ (3×3), *st* is stride factor (*st*=2) to up-sample the feature maps, ξ and Φ represents batch normalization and Relu activation function.

3.1. Network Loss

The main objective function for network is given as,

$$Y_t = \delta(X_t^i : X_{t-N}^{N-i}) \qquad N \in (0, 3, 6) \tag{8}$$

where, Y_t is estimated output of network, $X_t^i : X_{t-N}^{N-i}$ is considered input video frames. In adversarial training, the objective function of generator network with discriminator (D) is defined as,

$$\begin{aligned}\mathbb{L}_{GAN}(G, D) = \mathbb{E}_{x,\,y}[\log D(x, y)] + , \\ \mathbb{E}_{x,\,z}[\log(1 - D(x, G(x, z))]\end{aligned} \tag{9}$$

Where, $\hat{Z}$ random noise vector.

4. Training of Proposed Method

The learning of the proposed network is concentrated in different ways like disjoint (unseen) and global (seen) data based training-testing. For unseen training-testing, within database videos are separated as training-testing splits without any overlap. The video frames in a video are divided into training-testing splits for seen data training-testing. For training of the proposed network, we took an approach of disjoint (unseen) training-testing (DTT), which is performed on DAVIS-2016 [35] and global (seen) training-testing (GTT) is done on CDnet-2014 [44] database. The training details are discussed in the next sub-sections.

4.1. Disjoint (Unseen) Training-Testing

We selected the DAVIS-2016 database for DTT. This database is having 50 videos with different attributes like fast-motion, dynamic background, motion blur, scale variation, background clutter, camera shake, interacting objects, low resolution, occlusions, etc. Here, 30 videos (*along with respective ground-truths*) from DAVIS-2016 database are used similar to STCRF [45] for training. Remaining 20 videos are used for testing purpose.

4.2. Global (Seen) Training-Testing

The CDnet-2014 database is used for GTT of the proposed method similar to [2], [3] and [1]. In [38], 90% of video frames from each video are used for training and the remaining samples are used for testing. Also, in [2], [3] and [1], 70% of video frames are used for training the network and the rest of (30%) the video frames are used to test the effectiveness of the network. In the proposed method, 30% data is used for training and the remaining data is used for testing. For proposed network global training-testing, from each video 30% of initial video frames are selected from CDnet-2014 for training and the remaining frames are utilized for testing.

The discriminator network and remaining settings for the training of the proposed network are similar to [15]. The network is initialized with random weights for both the training-testing and is learned using a stochastic gradient descent algorithm with a learning rate of 0.0002. The weight parameters of the network are updated (*500 and 200 epochs for DTT and GTT respectively*) on NVIDIA DGX station with processor 2.2 GHz, Intel Xeon E5-2698 (20-Core), NVIDIA Tesla V100 4×16 GB GPU.

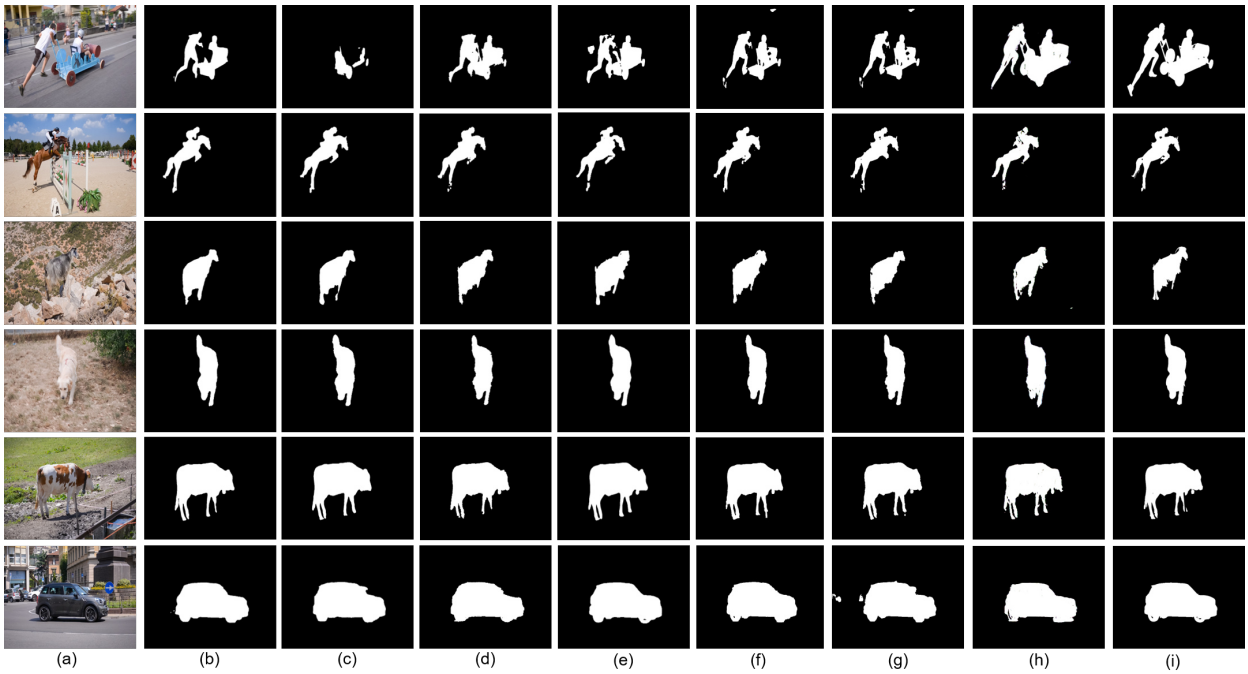

Figure 4. Qualitative result comparison of proposed method with existing state-of-the-art methods on DAVIS-2016 database, (a) input frame, output extracted using (b) FEELVOS-[41], (c) AGME-[18], (d) LUCID-[19], (e) CNIM-[5], (f) OSVOS-[29], (g) RANet-[45], (h) proposed method, (i) ground-truth.

Table 1. Quantitative results comparison of proposed method with existing sate-of-the-art methods on DAVIS-2016.

Methods	Publication	F-measure
LSMO [40]	IJCV-19	0.779
AGS [43]	CVPR-19	0.774
CoSNet [24]	CVPR-19	0.795
FEELVOS [41]	CVPR-19	0.822
RGMP [46]	CVPR-18	0.820
AGME [18]	CVPR-19	0.822
FAVOS [9]	CVPR-18	0.795
LUCID [19]	IJCV-19	0.820
CNIM [5]	CVPR-18	0.850
DIPNet [14]	WACV-20	0.864
OSVOS [29]	TPAMI-19	0.875
RANet [45]	ICCV-19	0.876
Proposed Method	-	**0.889**

5. Results and Discussion

The experimental results of the proposed method for MOS are discussed in this section. The segmentation results (qualitative and quantitative) analysis of the proposed approach are examined on two state-of-the-art databases, DAVIS-2016 [35] and CDnet-2014 [44] for MOS. For quantitative results, the average F-measure is calculated.

5.1. Results on DAVIS-2016

The DAVIS-2016 database is one of the challenging databases for foreground localization having 50 video sequences with 854×480 spatial resolution and average 70 frames per video. These videos are recorded with different attributes like fast-motion, dynamic background, motion blur, scale variation, background clutter, camera shake, interacting objects, low resolution, occlusions, etc. with 24 *frames per second*. For experimental purpose, ground truths are provided for each video frame with pixel-wise manual annotation. The segmentation accuracy is examined qualitatively as well as quantitatively and compared with respective ground-truth and existing state-of-the-art methods. Table 1 gives the quantitative results comparison with existing methods in terms of average F-measure. The qualitative results of the proposed method are compared with the existing methods given in Figure 4.

Some of the recently published work [5], [29] and [45] achieved the significant improvement in accuracy, but these models make use of pre-trained weights or require fine-tuning of the network. The DeepLabv2 VGG16 pre-trained on PASCAL VOC is used as an initial weight parameter in

Table 2. Category wise average F-measure comparison of proposed method with existing methods for MOS on CDnet-2014 database.

Methods	Publication	Baseline	DyanBG	BadWeath	CamJitt	Shadow	Avg
SGSM-BS [36]	TIP-18	0.9500	0.8600	0.8200	0.8500	0.8900	0.874
B-SSSR [17]	TIP-19	0.9600	0.9200	0.9300	0.9400	0.9300	0.9360
ResNet [8]	TCSVT-18	0.9294	0.9461	0.9518	0.8221	0.9647	0.9228
DeepBs [4]	PR-18	0.9580	0.8761	0.8301	0.8990	0.9304	0.8987
sEnDec [1]	ITIS-19	0.9591	0.9389	0.9542	0.9001	0.9293	0.9363
3DLSTM [2]	ITIS-19	0.9597	0.9487	-	0.9570	0.9352	0.9502
MRFCNN [3]	TVT-19	0.9726	0.9580	0.9655	0.9452	0.9435	0.9563
Proposed Method		**0.9498**	**0.9614**	**0.9723**	**0.9638**	**0.9573**	**0.9609**

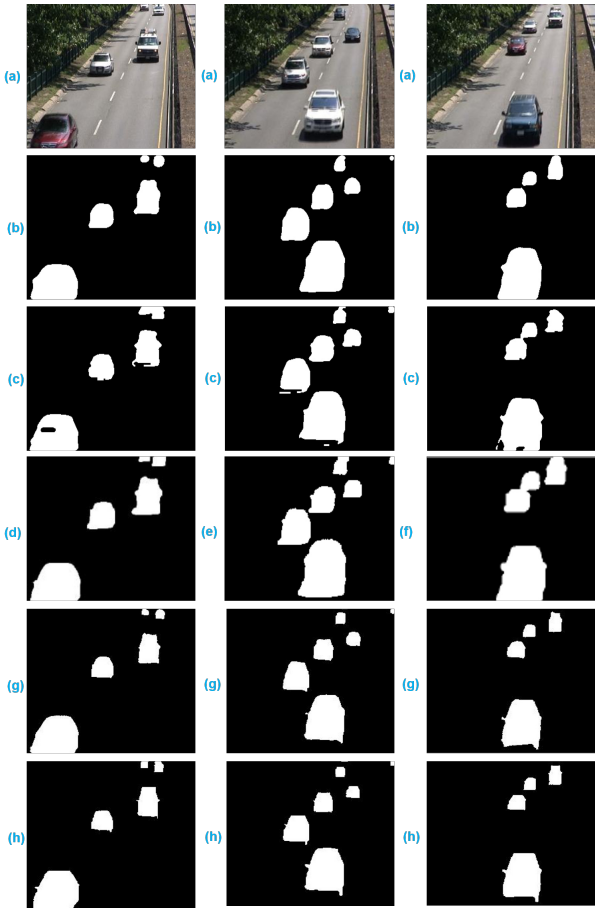

Figure 5. Visual result comparison of existing methods, (a) input frame, (b) DeepBs [4], (c) MSFgNet [33], (d) sEnDec [1], (e) 3DLSTM [2], (f) MRFCNN [3], (g) proposed method output and (h) ground-truth on CDnet-2014 database.

[5] with VGG-Net as a backbone network. After that, 60 video clips are used for training of the network for video object segmentation. Similarly, three-stage (base, parent and test) is proposed in [29]. Initially, the parent network is trained on the DAVIS-2016 training set with pre-trained weights of ImageNet through the base network. Further, for video object segmentation, the trained parent network is fine-tuned on one frame along with the ground-truth of each test sequence. In [45], initially the network is trained on MSRA10K, ECSSD and HKU-IS for static image segmentation. Further, the trained model is fine-tuned on DAVIS-2016 and DAVIS-2017 for video object segmentation. Here, the proposed method is introduced on only the DAVIS-2016 database. Without any pre-trained weights or fine-tuning of the network, the proposed method shows very close performance for foreground localization as compared to [5], [29] and [45]. We give this credit to our proposed multi-scale residual block with dense connection and recurrent technique.

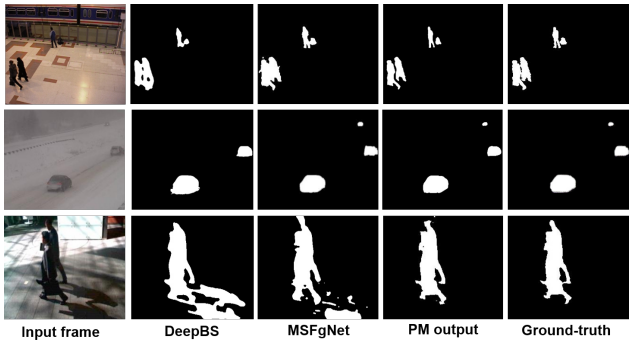

Figure 6. Visual result comparison of existing methods (DeepBs [4] and MSFgNet [33]) on CDnet-2014 database.

5.2. Results on CDnet-2014

In this experimentation, the detection accuracy of the proposed method is verified on the CDnet-2014 dataset using a globally trained network. The considered videos from different video categories are baseline (*Highway, Office, Pedestrians, PETS2006*), bad weather (*blizzard, skat-*

ing, snowFall), dynamic background (*boats, canoe, fall, overpass*), camera jitter (*badminton, boulevard, traffic*) and shadow (*backdoor, busStation, copyMachine, peoplenShade*). The Table 2, Figure 5 and Figure 6 gives the quantitative and qualitative result comparison of proposed method with existing methods respectively. In literature, [4] used 150 frames from each video to train the network. But, random selection of video frames for training-testing is questionable because in random selection frame f_t is selected for training and a temporally closest frame $f_{(t+1)}$ or $f_{(t-1)}$ may be used in testing. As FBS is a temporally processed decision-making problem, it is quite difficult to get significant results with this type of training and testing because these approaches are not able to estimate accurate motion related to the foreground. In [2], [1] and [3], 70% video frames from each video are used for training and remaining video frames are used for testing with video-wise training and fine-tuning approach. Here, all these approaches are trained on source domain video and trained weights are used to fine-tune the network for target domain video. In [1] and [3], along with gray-scale input frames, computed background frame through temporal median filtering is used as input to the network. But, estimating one background frame for the entire video is not suitable for outdoor seen video. In the proposed method, 30% video frames from each video are collectively used for training and the remaining frames are used for testing. From Table 2, it is evident that the proposed method gives significantly improved performance as compared to [2], [3] and [1] even-though less data is used for training of the proposed network.

From qualitative and quantitative results analysis, the proposed recurrent architecture with multi-scale residual dense connection is able to refine the prediction about foreground probability maps by iteratively correcting the previous mistakes.

5.3. Ablation Study of Proposed Network

The effectiveness of the proposed algorithm is depend upon the proposed recurrent feature sharing and multi-scale residual block with dense connections. **Does the proposed recurrent feature sharing will help the network for effective learning?** To do so, the accuracy is examined with (**W/i**) and without (**W/o**) recurrent feature sharing. The quantitative analysis in terms of average F-measure is given in the Table 3. From results, it is clear that the proposed recurrent feature sharing help the network to learn more robust and effective foreground features. Further, in the proposed network, the MRDC blocks are used. **How the MRDC blocks contributed to the learning of the proposed network?**. To scrutinize this, the proposed network is trained without and with MRDC blocks and accuracy is examined. The quantitative results are given in the Table 3. From Table 3, it is cleared that the proposed recurrent feature sharing from previous frame decoder features with encoder features of current frame and MRDC blocks are effective to get more robust and meaningful foreground features.

Table 3. Ablation analysis of proposed network on recurrent feature sharing and MRDC block on DAVIS-2016 (*W/i:With and W/o:Without*).

Methods	F-measure
W/o MRDC + W/o Recurrent	0.8456
W/i #1 MRDC + W/o Recurrent	0.8509
W/i #2 MRDC + W/o Recurrent	0.8596
W/i #3 MRDC + W/o Recurrent	0.8636
W/i #4 MRDC + W/o Recurrent	0.8532
W/o MRDC + W/i Recurrent	0.8587
W/i #3 MRDC + W/i Recurrent	0.8889

6. Conclusion

In this work, an end-to-end recurrent adversarial learning network is proposed for moving object segmentation with dilated convolution, multi-scale residual with dense connection module and refinement module. Along with temporally sampled successive video frames, the learned decoder features of the previous frame are shared with encoder features for current frame foreground segmentation to get enough motion information. The dilated convolution is used after the encoder block to extract spatial features at multiple scales with different sampling rates. Thus, the proposed method considers spatial and temporal features through dilated convolution and recurrent technique respectively. Also, the proposed network comprises of multi-scale residual blocks with dense connections to learn foreground related prominent features and to avoid vanishing gradient problem. Further, to produce the output, a refinement module is proposed, which takes the output from the MRDC block and also features from the target encoder block through skip-connections. Further, the network's learning is concentrated on different ways like disjoint and global training-testing for MOS. An extensive experimental analysis of the proposed method is done on two benchmark video datasets with seen and unseen data analysis for MOS. Experimental results show that the proposed network outperforms the existing state-of-the-art methods on benchmark datasets for moving object segmentation.

Acknowledgement

This work was supported by the Science and Engineering Research Board (DST-SERB), India, under Grant ECR/2018/001538.

References

[1] Thangarajah Akilan and Qingming Jonathan Wu. sendec: An improved image to image cnn for foreground localization. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[2] Thangarajah Akilan, Qingming Jonathan Wu, Amin Safaei, Jie Huo, and Yimin Yang. A 3d cnn-lstm-based image-to-image foreground segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[3] Thangarajah Akilan, QM Jonathan Wu, and Wandong Zhang. Video foreground extraction using multi-view receptive field and encoder-decoder dcnn for traffic and surveillance applications. *IEEE Transactions on Vehicular Technology*, 2019.

[4] Mohammadreza Babaee, Duc Tung Dinh, and Gerhard Rigoll. A deep convolutional neural network for video sequence background subtraction. *Pattern Recognition*, 76:635–649, 2018.

[5] Linchao Bao, Baoyuan Wu, and Wei Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *Proceedings of the IEEE Conference on CVPR*, pages 5977–5986, 2018.

[6] Lin Chen, Xue Jiang, Xingzhao Liu, Thiagalingam Kirubarajan, and Zhixin Zhou. Outlier-robust moving object and background decomposition via structured regularized low-rank representation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2019.

[7] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE Conference on CVPR*, pages 1189–1198, 2018.

[8] Yingying Chen, Jinqiao Wang, Bingke Zhu, Ming Tang, and Hanqing Lu. Pixel-wise deep sequence learning for moving object detection. *IEEE Transactions on CSVT*, 2017.

[9] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *Proceedings of the IEEE Conference on CVPR*, pages 7415–7424, 2018.

[10] Akshay Dudhane, Kuldeep M Biradar, Prashant W Patil, Praful Hambarde, and Subrahmanyam Murala. Varicolored image de-hazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4564–4573, 2020.

[11] Akshay Dudhane, Praful Hambarde, Prashant Patil, and Subrahmanyam Murala. Deep underwater image restoration and beyond. *IEEE Signal Processing Letters*, 27:675–679, 2020.

[12] Brent Griffin and Jason Corso. Tukey-inspired video object segmentation. In *2019 IEEE WACV*, pages 1723–1733. IEEE, 2019.

[13] Praful Hambarde, Akshay Dudhane, Prashant W Patil, Subrahmanyam Murala, and Abhinav Dhall. Depth estimation from single image and semantic prior. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1441–1445. IEEE, 2020.

[14] Ping Hu, Jun Liu, Gang Wang, Vitaly Ablavsky, Kate Saenko, and Stan Sclaroff. Dipnet: Dynamic identity propagation network for video object segmentation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1904–1913, 2020.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[16] Sajid Javed, Arif Mahmood, Somaya Al-Maadeed, Thierry Bouwmans, and Soon Ki Jung. Moving object detection in complex scene using spatiotemporal structured-sparse rpca. *IEEE Transactions on Image Processing*, 28(2):1007–1022, 2018.

[17] Sajid Javed, Arif Mahmood, Somaya Al-Maadeed, Thierry Bouwmans, and Soon Ki Jung. Moving object detection in complex scene using spatiotemporal structured-sparse rpca. *IEEE Transactions on Image Processing*, 28(2):1007–1022, 2018.

[18] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation. In *Proceedings of the IEEE Conference on CVPR*, pages 8953–8962, 2019.

[19] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for video object segmentation. *International Journal of Computer Vision*, 127(9):1175–1197, 2019.

[20] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pages 123–141. Springer, 2018.

[21] Linhao Li, Qinghua Hu, and Xin Li. Moving object detection in video via hierarchical modeling and alternating optimization. *IEEE Transactions on Image Processing*, 28(4):2021–2036, 2018.

[22] Chih-Yang Lin, Kahlil Muchtar, Wei-Yang Lin, and Zhi-Yao Jian. Moving object detection through image bit-planes representation without thresholding. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[23] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the ECCV*, pages 385–400, 2018.

[24] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE Conference on CVPR*, pages 3623–3632, 2019.

[25] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. Unovost: Unsupervised offline video object segmentation and tracking. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2020.

[26] Murari Mandal, Vansh Dhar, Abhishek Mishra, and Santosh Kumar Vipparthi. 3dfr: A swift 3d feature reductionist framework for scene independent change detection. *IEEE Signal Processing Letters*, 26(12):1882–1886, 2019.

[27] Murari Mandal, Lav Kush Kumar, Mahipal Singh Saran, et al. Motionrec: A unified deep framework for moving object recognition. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2734–2743, 2020.

[28] M. Mandal and S. K. Vipparthi. Scene independency matters: An empirical study of scene dependent and scene independent evaluation for cnn-based change detection. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–14, 2020.

[29] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on PAMI*, 41(6):1515–1530, 2018.

[30] Prashant Patil and Subrahmanyam Murala. Fggan: A cascaded unpaired learning for background estimation and foreground segmentation. In *2019 IEEE WACV*, pages 1770–1778. IEEE, 2019.

[31] Prashant Patil, Subrahmanyam Murala, Abhinav Dhall, and Sachin Chaudhary. Msednet: multi-scale deep saliency learning for moving object detection. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1670–1675. IEEE, 2018.

[32] Prashant W Patil, Kuldeep M Biradar, Akshay Dudhane, and Subrahmanyam Murala. An end-to-end edge aggregation network for moving object segmentation. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 8149–8158, 2020.

[33] Prashant W Patil and Subrahmanyam Murala. Msfgnet: A novel compact end-to-end deep network for moving object detection. *IEEE Transactions on Intelligent Transportation Systems*, 2018.

[34] Prashant W Patil, Omkar Thawakar, Akshay Dudhane, and Subrahmanyam Murala. Motion saliency based generative adversarial network for underwater moving object segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1565–1569. IEEE, 2019.

[35] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on CVPR*, pages 724–732, 2016.

[36] Guangming Shi, Tao Huang, Weisheng Dong, Jinjian Wu, and Xuemei Xie. Robust foreground estimation via structured gaussian scale mixture modeling. *IEEE Transactions on Image Processing*, 27(10):4810–4824, 2018.

[37] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the ECCV*, pages 715–731, 2018.

[38] Youbao Tang and Xiangqian Wu. Salient object detection using cascaded convolutional neural networks and adversarial learning. *IEEE Transactions on Multimedia*, 2019.

[39] Omkar Thawakar, Prashant W Patil, Akshay Dudhane, Subrahmanyam Murala, and Uday Kulkarni. Image and video super resolution using recurrent generative adversarial network. In *2019 16th IEEE International Conference on AVSS*, pages 1–8. IEEE, 2019.

[40] Pavel Tokmakov, Cordelia Schmid, and Karteek Alahari. Learning to segment moving objects. *International Journal of Computer Vision*, 127(3):282–301, 2019.

[41] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE Conference on CVPR*, pages 9481–9490, 2019.

[42] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE Conference on CVPR*, pages 1328–1338, 2019.

[43] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE Conference on CVPR*, pages 3064–3074, 2019.

[44] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. Cdnet 2014: an expanded change detection benchmark dataset. In *Proceedings of the IEEE conference on CVPRW*, pages 387–394, 2014.

[45] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. *International Conference on Computer Vision*, 2019.

[46] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE Conference on CVPR*, pages 7376–7385, 2018.

[47] Kai Xu, Longyin Wen, Guorong Li, Liefeng Bo, and Qingming Huang. Spatiotemporal cnn for video object segmentation. In *Proceedings of the IEEE Conference on CVPR*, pages 1379–1388, 2019.

[48] Lu Yang, Jing Li, Yuansheng Luo, Yang Zhao, Hong Cheng, and Jun Li. Deep background modeling using fully convolutional network. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):254–262, 2017.

[49] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE Transactions on CSVT*, 2019.

[50] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Transactions on Image Processing*, 29:8326–8338, 2020.