This CVPR 2020 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

An End-to-End Edge Aggregation Network for Moving Object Segmentation

Prashant W. Patil, Kuldeep M. Biradar, Akshay Dudhane, and Subrahmanyam Murala
CVPR Lab, Indian Institute of Technology Ropar, INDIA
2017eez0006@iitrpr.ac.in

Abstract

Moving object segmentation in videos (MOS) is a highly demanding task for security-based applications like automated outdoor video surveillance. Most of the existing techniques proposed for MOS are highly depend on fine-tuning a model on the first frame(s) of test sequence or complicated training procedure, which leads to limited practical serviceability of the algorithm. In this paper, the inherent correlation learning-based edge extraction mechanism (EEM) and dense residual block (DRB) are proposed for the discriminative foreground representation. The multi-scale EEM module provides the efficient foreground edge related information (with the help of encoder) to the decoder through skip connection at subsequent scale. Further, the response of the optical flow encoder stream and the last EEM module are embedded in the bridge network. The bridge network comprises of multi-scale residual blocks with dense connections to learn the effective and efficient foreground relevant features. Finally, to generate accurate and consistent foreground object maps, a decoder block is proposed with skip connections from respective multi-scale EEM module feature maps and the subsequent down-sampled response of previous frame output. Specifically, the proposed network does not require any pre-trained models or fine-tuning of the parameters with the initial frame(s) of the test video. The performance of the proposed network is evaluated with different configurations like disjoint, cross-data, and global training-testing techniques. The ablation study is conducted to analyse each model of the proposed network. To demonstrate the effectiveness of the proposed framework, a comprehensive analysis on four benchmark video datasets is conducted. Experimental results show that the proposed approach outperforms the state-of-the-art methods for MOS.

1. Introduction

Moving object segmentation (MOS) for video captured under the uncontrolled weather, different illumination conditions, or dynamic background is a challenging task for many computer vision applications like automated video surveillance [30], traffic monitoring [4], anomaly detection [27], etc. It aims to automatically generate precise and consistent pixel masks for foreground object(s). The accuracy achieved for indoor videos is higher as compared to outdoor videos. Because, outdoor videos suffer from several factors like poor visibility, inclement weather situations, low contrast, local motion, etc. Also, one important attention for automated video applications is that more than 70% of pixel information is redundant and irrelevant for high-level processing task [3]. This redundant information degrades the overall performance of automated applications like video surveillance, traffic monitoring, etc. Learning-based approaches gave significant performance improvement for many computer vision applications [35], [34], [14], [23], [37], [15], [26], [2], [4], [1], [40], [24]. Many approaches [26], [40], [43], [4], [1] are proposed with fine-tuning of pre-trained model using first frame(s) of test sequences. Additionally, several techniques [24], [37] achieved significant performance with high system complexity. Even-though these methods delivered impressive results, the practical serviceability of these approaches is

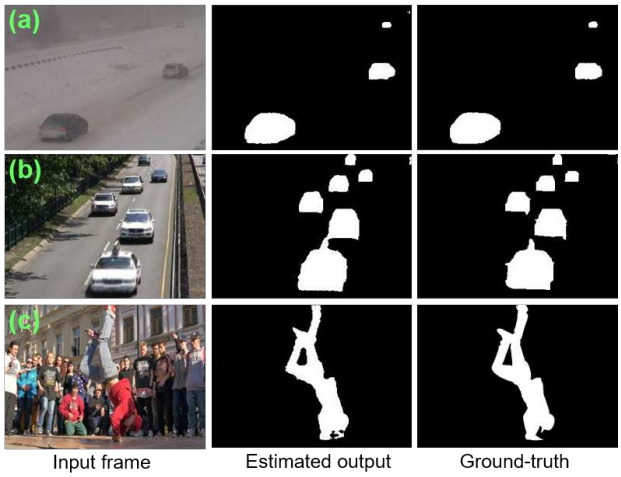

Figure 1. Sample results of the proposed framework on (a) weather degraded video, (b) traffic video with multi-objects and (c) crowded video with single object.

limited. Thus, MOS is a challenging task from several aspects in day-to-day life.

The main motivation of the proposed framework for MOS is to design a model which does not rely on fine-tuning of a pre-trained model on the first frame(s) of the test sequence. Also, the system complexity is considered for more practical serviceability *i.e. the system should be simple, fast, end-to-end, and strong*. To achieve this goal, in this work, a multi-frame multi-scale encoder-decoder adversarial learning network with edge extraction mechanism and the dense residual block is proposed for MOS. A very important and crucial step in the encoder-decoder network is that how to connect the pixel-level multi-scale encoder feature in a meaningful manner to the respective scale of the decoder. Also, while designing the network, the choice of the filter size plays an important role for better feature learning for a specific task. To do this, an inherent correlation-based edge extraction mechanism is proposed. Additionally, the predicted output of the previous frame is used with subsequent scale to provide the consistent matching among current and previous frame at the decoder for the learning of discriminative foreground representation. Some of the sample results on weather degraded, multi-object traffic, and the crowd with single object video are shown in Figure 1.

2. Related work

Existing MOS algorithms are broadly classified as unsupervised, semi-supervised, on-line, and propagation-based methods. A brief overview of existing approaches for MOS is given below.

Unsupervised video object segmentation approaches [9], [45] segment foreground-background automatically over an unconstrained video without any user annotation. Brent *et al.* [9] proposed motion and visual saliency-based approach for MOS. The forward propagation-based approach is proposed in [45] to estimate the object proposals. Wang *et al.* [35] proposed an unsupervised MOS approach with dynamic visual attention prediction and attention guided object segmentation in spatio-temporal and spatial domain respectively.

Semi-supervised video object segmentation rely on preliminary provided ground-truth masks [34], [24], [15], [44], [20], [28], [7]. Paul *et al.* [34] proposed semantic pixel-wise feature concatenation with global and local matching techniques for moving object detection. The probabilistic generative approach is proposed in [14] for the prediction of the target and background appearance. The generative appearance, backbone feature extractor and prediction modules are used for efficient feature extraction. The primary focus of existing state-of-the-art learning-based approaches is to learn the appearance and motion-based feature for frame segmentation. Along with these features, Lu *et al.* [23] proposed a co-attention mechanism to improve the discriminative foreground representations. Khoreva *et al.* [15] proposed data augmentation technique *i.e.* lucid data dreaming for semi-supervised video object segmentation (VOS). A two-stream network with a memory module is proposed in [33] to get the appearance and motion-based features. Some of the researchers used tracking-based methods to detect the region-of-interest for VOS [7]. Luiten *et al.* [24] proposed an approach with semantic proposal generation, refinement, and merging techniques for MOS. The results delivered in [24] are impressive, but the complexity of system is high as they used four different networks together with fine-tuning.

On-line learning based methods [11], [40], [26], [6], [43] are semi-supervised methods which are mainly relied on fine-tuning of pre-trained models on first frame of test sequence. Motion-guided cascaded refinement network [11] is proposed for MOS with the assumption that the foreground motion is different from the background motion. Maninis *et al.* [26] proposed an orthogonal approach without temporal information for VOS. Here, the learned features on ImageNet are used for transferring the generic semantic information for foreground-background segmentation (FBS). The spatial and temporal dependencies are encoded in [6] using CNN trained model and optical flow.Recently, generative adversarial network (GAN) based approaches shows significant improvement in various computer vision applications like image de-hazing [8], FBS [29], underwater MOS [31], etc. To capture appearance and motion cues, the temporal coherence branch with pre-training in an adversarial fashion is utilized in [40]. Based on both of the learned cues, spatial segmentation branch is proposed for accurate segmentation of objects. Akilan *et al.* proposed 3D CNN based approach with 3D transpose convolution and residual connection [2], encoder-decoder CNN technique with the help of multi-view receptive field [4], slow encoder-decoder with strided convolution and temporal median filtering-based background generation [1]. Here, authors trained their models [2], [4], [1] on baseline video and fine-tuned on frames of target video for better generalization and accurate foreground detection. The training on baseline video, fine-tuning on more number of frames, and testing on remaining video frames from target video leads to the limited practical applicability of the algorithm.

Propagation based [38], [37], [41], [39] approaches make use of previous frame(s) output for efficient and effective MOS. Along with the visual and spatial guidance, Linjie *et al.* [41] has introduced a modulator to manage the learning of intermediate layers of segmentation network. Seoung *et al.* [38] proposed identical encoder network to process the key frame and reference frame interdependently.Finally, the refinement module with residual learning is used for fast MOS. Similarly, Ziqin *et al.* [37] proposed a ranking attention technique to integrate the matching and

propagation-based encoder-decoder network for VOS. In [39] and [44], along-with input frames, optical flow [12] is used as input to guide the propagation process for foreground motion clustering.

Proposed approach overcomes the shortcomings of [2], [4], [1] with less data for training and [11], [40], [26] with no fine tuning on frame(s) from target video. Additionally, the optical flow using [22] and output of previous frame with respective scale are used to guide the propagation process in the proposed approach. The proposed work has the following key contributions:

1. An end-to-end multi-frame multi-scale encoder-decoder adversarial learning network is proposed for moving object segmentation.
2. A novel edge extraction mechanism (EEM) is proposed to integrate the multi-frame pixel-level multi-scale encoder features with respective decoder features through skip connections.
3. Bridge network with a dense residual block is proposed to embed the motion features which are extracted from optical flow encoder stream and feature maps from the last EEM module.
4. Effectiveness of the proposed approach for MOS is examined on four benchmark video databases with disjoint, global, and cross-data training-testing techniques and compared to the state-of-the-art methods.

3. Proposed system framework

Various researchers have taken the advantage of pre-trained models of convolutional neural network (CNN) [7],[6], [38], [44], [20], [28] for MOS. Also, some of the approaches fine-tuned the pre-trained network on initial frame(s) of test video for MOS [11], [40], [26], [43], [2], [4], [1]. Additionally, some methods [24], [37] achieved state-of-the-art performance, resulting in a high computational complexity. Above all factors lead the MOS towards the limited practical usability. This motivated us to design an end-to-end network for MOS, which does not rely on fine-tuning and leads towards more practical serviceability. There are two major challenges in the MOS task. **First, separation of foreground objects from background**. Based on the hypothesis of different background-foreground motion [37], we have proposed a multi-frame multi-scale encoder-decoder network for MOS. The proposed network takes video frames and optical flow as inputs to learn the inherent correlation between multi-scale encoder features of three successive frames. As multi-frame encoder gives foreground-background probability maps, learning of multi-frame multi-scale encoder features is required, and it should be propagated to the decoder network in an effective and meaningful manner. To do this, the multi-frame multi-scale edge extraction mechanism with correlation learning is proposed in this work. Also, encoded foreground edge related features using the last EEM module and encoded feature maps from optical flow encoder stream are fused using a bridge network to learn robust foreground relevant features. **Second, consistent segmentation of foreground objects across the video frames**. Based on the assumption that the previous frame foreground object(s) are not that much deviated for the current frame, we make use of estimated previous frame output with respective scale to guide the decoder network for discriminative foreground feature representation. Detailed visualization of the proposed network is given in Figure 2.

3.1. Multi-frame multi-scale encoder

The proposed approach takes RGB video frames ($I_t \in \mathbb{R}^{3\times M\times N\times 3}$) and extracted optical flow ($O_t \in \mathbb{R}^{M\times N\times 3}$) [22] as input. Here, multi-frame based encoders are used to obtain the multi-scale edge information related to foreground *i.e.* three frames are fed to three different encoder streams. Each block of encoder stream comprises of two convolution filters with a kernel size of 3×3 and 7×7 followed by a leaky rectified linear unit (ReLU) to extract the pixel-level multi-scale features. Additionally, estimated optical flow [39] between pair of frames (*t-1, t, t+1*) is given to the fourth encoder stream to learn motion features. For better visualization, the optical flow is considered as HSV representation [25]. Where, the hue and saturation represent the direction of motion and its magnitude, respectively. In this work, the only magnitude is considered and appended three times to get the three-channel image. As the performance of early or late fusion of optical flow stream features with appearance stream features is not effective [39], a mid-level fusion of motion feature from optical flow encoder stream and last EEM module features is considered in the proposed approach. An encoder block is defined as $EN_{L\_L\times f};\ [L \in (1,4),\ f = 32]\}$ where, L and (L×f) represent encoder level and number of filters in encoder respectively (*more details please refer Figure 2*).

3.2. Edge extraction mechanism module

As encoder gives foreground-background probability maps [37], effective learning of inherent correlation between multi-frame encoder with multi-scale features is required. To do this, learning based edge extraction mechanism (EEM) module is proposed. Here, EEM module is applied on each scale of the encoder network to focus on foreground relevant feature learning and to ignore the background regions. Initially, pixel-wise subtraction is performed between one scale feature of encoder and another scale feature of another encoder. All subtracted features are concatenated to get the overall response of that particular encoder level as given in Eq. (1).

$$C = \Psi\left\{X_{S\_k},\ Y_{S\_k},\ Z_{S\_k}\right\} \tag{1}$$

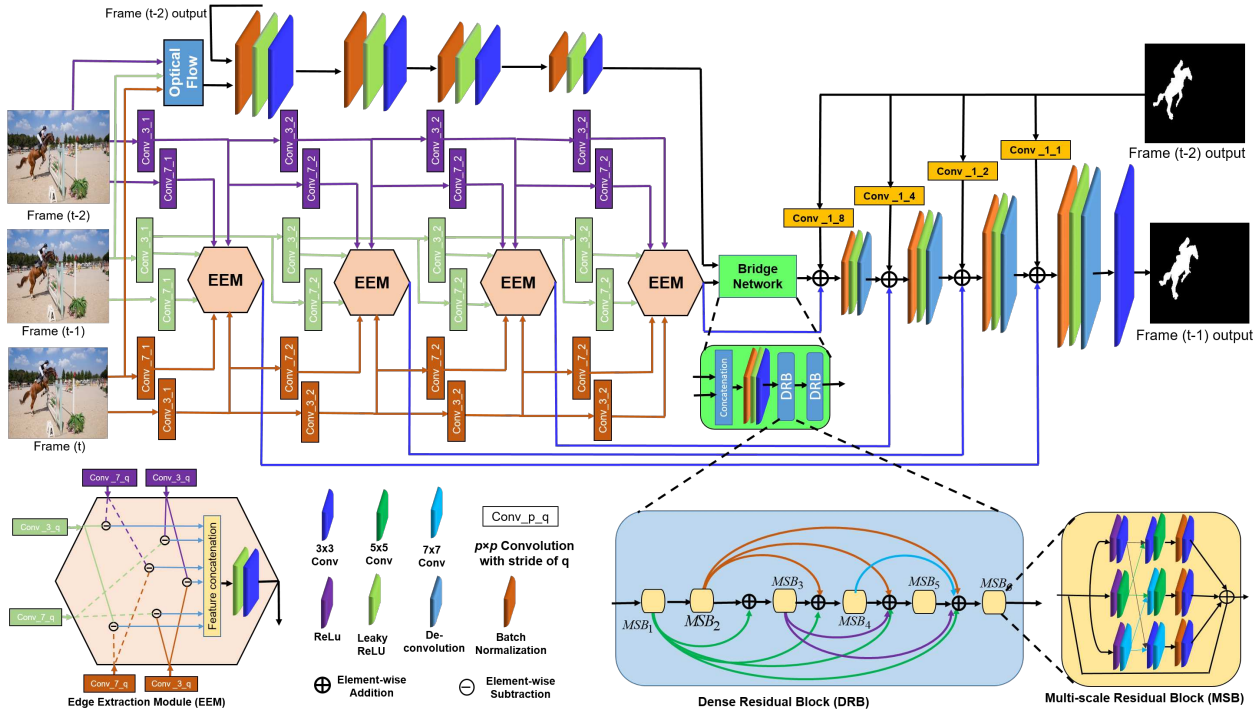

Figure 2. Overview of the proposed framework for MOS. *First, the multi-scale features related to the foreground objects are extracted from three consecutive frames with the help of the proposed edge extraction mechanism (EEM) module. Encoded feature maps from optical flow encoder stream and last EEM module are embedded to learn effective features related to the foreground. Finally, to segment current frame, the down-sampled output response of the previous frame and respective EEM module feature maps are combined in the decoder network.*

where, Ψ indicates the concatenation of subtracted features

$$X_{S\_k} = W^{(i,j)}_{(k)\_(S)} \ominus W^{(i,j)}_{(k+4)\_(S+1)} \; ; \; k = 3, \; S \in (1, 2)$$

$$Y_{S\_k} = W^{(i,j)}_{(k)\_(S)} \ominus W^{(i,j)}_{(k-4)\_(S+1)} \; ; \; k = 7, \; S \in (1, 2)$$

$$Z_{S\_k} = W^{(i,j)}_{(k)\_(S)} \ominus W^{(i,j)}_{(k)\_(S+2)} \; ; \; k \in (3, 7), S = 1$$

where, $\ominus$ is element-wise subtraction, $W^{(i,j)}_{(k)\_(S)}$ are features of S stream at location (i , j) with $k \times k$ size kernel.

The ablation study is conducted to demonstrate the impact of concatenation over addition operation for multi-scale feature extraction (*please refer Table 4*). The detailed visualisation of sample feature maps of first EEM module is given in Figure 3. Response of each EEM module is essentially preserved for segmentation and passed to the respective decoder network through skip connections for effective and meaningful foreground representations.

3.3. Bridge network

The bridge network is constructed for embedding of the motion features from optical flow encoder stream with last EEM module features of encoder. The EEM module is denoted as $\{EEM_{L\_L\times f};\ [L \in (1, 4),\ f = 32]\}$. The approaches used for automated video applications need to process large amount of data for training. The training of deeper network undergoes the vanishing gradients problem [10]. To overcome these limitations, multi-scale residual blocks (MSBs) with dense connections named as *dense residual block (DRB)* is proposed to learn prominent features related to foreground. Specifically, we conduct the ablation study to analyse the importance of DRB block in the proposed network (*please refer Table 5*). The technique for dense connections is defined as,

$$MSB_n = \sum_{i=1}^{n-1} MSB_i \; ; \; n > 1 \tag{2}$$

where, MSB_n is input to the n^{th} MSB module, MSB_i is response of i^{th} MSB module and $n \in (1, 6)$. Each MSB is having parallel convolution filters with kernel size of 3×3, 5×5 and 7×7 followed by ReLU. For effective learning, we integrate the multi-scale features with the concatenation operation followed by separate convolution block. Finally, responses of each concatenated features are added to get robust features learned by different scales with residual connection (*please refer Figure 2 for more details*).

3.4. Foreground prediction with propagation

In [37], ranking attention module is proposed to select the important features for similarity maps. Matching of current frame foreground object features with reference/first frame features [38] may fail in some of the practical scenar-

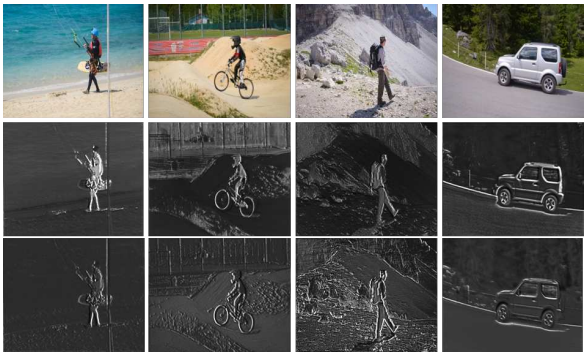

Figure 3. Visualization of two samples of feature maps of first EEM module maps.

ios like illumination, occlusion, motion blur, etc. Also, for automated video surveillance applications, reference frame object(s) may completely vanish after a few frames, and the new foreground object(s) may come in the current frame. However, the matching between current and previous frames usually referred to avoid false positive matches because motion is less. Hence, we make use of a simple mask propagation method the same as [37] *i.e.* predicted output of the previous frame is used to guide the subsequent decoder layer with respective scale to improve the potential of the proposed network for systematic foreground segmentation. Along with bridge network features and previous frame output with subsequent scale, the correlated feature from the respective EEM module is given to the decoder network for final foreground segmentation. The decoder block is defined as $\{DE_{L\_L\times f};\ [L \in (4,1),\ f = 32]\}$. Thus, proposed generator is represented as: $EN_{L\_L\times f} \rightarrow EEM_{L\_L\times f}\ ;\ [L \in (1\ ,\ 4),\ f = 32]$, $DE_{L\_L\times f};\ [L \in (4,1),\ f = 32]$

Additionally, we are able to train the proposed network in end-to-end manner for single object, multi-object and thermal data based segmentation with disjoint, global and cross-data training-testing techniques.

4. Training procedure of the proposed method

The proposed method makes use of an end-to-end adversarial training procedure and is deliberately straightforward. Because, MOS is a similar task like image-to-image translation [13] where the goal is to learn the mapping between the provided input frames and the desired response *i.e. foreground object(s)*. One advantage of the proposed framework is that it does not require a pre-training model or fine-tuning on the first frame of testing video.

We have trained the proposed network adversarially in three different configurations. (1) training and testing videos are segregated within database without any overlap [16] (**disjoint training-testing**), (2) training and testing video frames are segregated without any overlap [1] (**global training-testing**) and (3) training and testing datasets are totally different [30] (**cross-data training-testing**). Training details for each configuration are discussed in the next sub-sections. Note that the proposed network training procedure is much simpler than the existing methods [24], [26], [38], where we do not require pre-trained models or fine-tuning of the proposed network on the initial frame(s) of testing video.

4.1. Disjoint training-testing (DTT)

For disjoint training-testing (DTT), DAVIS-2016 [32] and SegTrack-v2 [18] database are used. DAVIS-2016 database is having 50 videos with different attributes like fast-motion, dynamic background, scale variation, background clutter, interacting objects, etc. From that, 30 videos (*along with respective ground truth*) are selected for training. To cover more challenging practical scenarios like slow motion, complex deformation, appearance change, background-foreground color similarity, SegTrack-v2 database is included for DTT. Out of 14 sequences, randomly 8 videos are chosen for training. Hence, total 38 (30+8) videos are used for training, and remaining (20+6) videos are used for testing similar to STCRF [16]. During training, we performed data augmentation, which includes horizontal flipping similar to [21].

4.2. Global training-testing (GTT)

For global training-testing (GTT), CDnet-2014 database [36] is considered similar to [2] and [4]. CDnet-2014 database covers a variety of practical scenarios like bad weather, camera jitter, shadow, traffic, etc. videos. In [2], [4], [1], 70% of video frames are used for training the network and rest of (30%) video frames are used to examine the effectiveness of network with video-wise fine-tuning. For the ideal case, the network should be able to give good performance on less training data, and there is no rule of thumb to pick an optimal number of frames that would lead to the best performance. In the proposed method, 50% of frames from each video are used together for training, and remaining frames are used for testing without video-wise fine-tuning. Specifically, we trained the proposed network on combined 50% frames of each video *i.e. no training on baseline video and no fine-tuning on frames of test sequence* similar to [1].

4.3. Cross-data training-testing (CTT)

CDnet-2014 database [36] and GTFD [17] are used for cross-data training and testing respectively. From CDnet-2014, the thermal video category is used for training of the proposed method. Total 5690 video frames from the thermal video category are selected. As per our knowledge, this is the first approach which uses the different database for training and different database for testing. For this technique, the optical flow encoder stream is removed from pro-

posed network *i.e. only thermal frames are used for training and testing*.

The remaining settings for all training configurations of the proposed method are similar to [13]. Weight parameters of the proposed network in all the training-testing techniques are initialized randomly and iteratively learned using stochastic gradient descent (SGD) algorithm with a learning rate of 0.0002. The weight parameters of the network are updated on NVIDIA DGX station with processor 2.2 GHz, Intel Xeon E5-2698, NVIDIA Tesla V100 16 GB GPU.

5. Network losses

In adversarial training, the objective function of generator network with discriminator (D) is defined as,

$$\mathbb{L}_{GAN}(G,D) = \mathbb{E}_{I,S}[\log D(I,S)] + \mathbb{E}_{I,Z}[\log(1 - D(I, G(I,Z))] \quad (3)$$

where, I, S and Z are input, ground-truth and random noise vector. To minimize the loss of generator network, structural similarity index metric (SSIM) and Edge losses (Sobel operator) are considered. Thus, loss function is defined as,

$$\mathbb{L}(G,D) = \arg\min_{G}\max_{D} \left(\mathbb{L}_{GAN} + \mathbb{L}_{SSIM} + \mathbb{L}_{Edge}\right) \quad (4)$$

6. Experiments

In this section, we evaluate the proposed network for MOS and multi-object segmentation on four benchmark databases, namely as DAVIS-2016 [32], SegTrack-v2 [18], CDnet-2014 [36] and GTFD [17]. Quantitative results in terms of average F-measure and visual results are evaluated and verified with the state-of-the-art methods for MOS. Further, several ablation experiments are conducted for a comprehensive understanding of the proposed method on DAVIS-2016.

6.1. Results analysis of DTT

For DTT model, the effectiveness is examined on testing set of DAVIS-2016 and SegTrack-v2 database in terms of average F-measure. We compare the proposed method results with 10 recently published methods, *i.e.* FEELVOS [34], AGAME [14], LUCID [15], CNIM [6], OSVOS [26], RANet [37], PReMVOS[24], DTNet [44], STMN [20], MGAVOS [28]. The quantitative results are given in Table 1 and Table 2 for DAVIS-2016 and SegTrack-v2 database respectively. Also, the visual results on DAVIS-2016 and SegTrack-v2 database are compared with state-of-the-art methods and given in the Figure 4 and Figure 5 respectively.

Some of the recently published work [6], [26], and [37] achieved the significant improvement in accuracy, but these models make use of pre-trained weights or require fine-tuning on a first frame(s) of test video. The DeepLabv2 VGG16 pre-trained on PASCAL VOC and is used as the initial weight parameter in [6] with VGG-Net as a backbone network. Similarly, a three-stage (base, parent, and test) network is proposed in [26]. Initially, the parent network is trained on the DAVIS training set with pre-trained

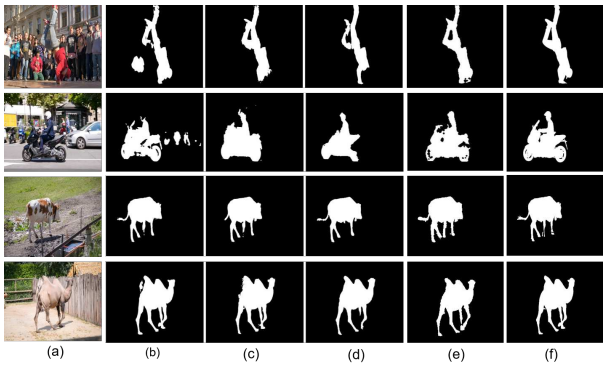

Figure 4. Visual results on DAVIS-2016 database. (a) input frames, (b) to (e) are the results from RANet-[37], OSVOS-[26], PReMVOS-[24], proposed method respectively, (f) ground-truth.

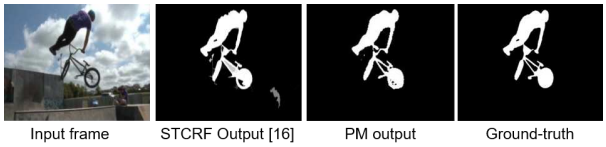

Figure 5. Visual results of proposed method (PM) and existing [16] on SegTrack-v2 database.

Methods	PT	OF	Year	F-measure
FEELVOS [34]	✓	-	CVPR-19	0.822
AGAME [14]	-	-	CVPR-19	0.822
LUCID [15]	✓	-	IJCV-19	0.820
DTNet [44]	✓	-	ICCV-19	0.835
CNIM [6]	✓	✓	CVPR-18	0.850
OSVOS [26]	✓	✓	PAMI-19	0.875
RANet [37]	-	-	ICCV-19	0.876
PReMVOS [24]	✓	-	ACCV-18	0.886
STMN [28]	-	-	ICCV-19	0.899
MGAVOS [20]	✓	-	ICCV-19	0.902
PM	-	-	-	**0.915**

Table 1. Quantitative results comparison of proposed method (PM) with existing sate-of-the-art methods on DAVIS-2016. We use "✓" to represent method with pre-training (PT) model or on-line fine-tuning (OF) .

Methods	Publications	F-measure
DSL [19]	CVPR-16	0.734
STCRF [16]	TIP-18	0.899
UOVOS [45]	TIP-19	0.643
Proposed Method	-	**0.918**

Table 2. Results comparison of proposed method and existing methods on SegTrack-v2 database.

weights of ImageNet through the base network. Further, for VOS, the trained parent network is fine-tuned on one frame along with the ground-truth of each test sequence. Similarly, the network is trained on MSRA10K, ECSSD, and HKU-IS for static image segmentation in [37]. Further, this trained model is fine-tuned on the DAVIS-2016 database for MOS. Semantic proposal generation, refinement, and merging techniques for MOS is proposed in [24]. The results delivered in [24] are impressive, but the complexity of system is high and high computational time is required as they used four different networks together with fine-tuning.

On the other hand, the proposed method achieved state-of-the-performance (*please refer Table 1 and 2*) without pre-training models or fine-tuning on the first frame of test video. The Table 1, 2 and Figure 4, 5 are evident that the proposed network outperforms the other existing state-of-the-art methods for MOS on DAVIS-2016 and SegTrack-v2.

6.2. Ablation study

To examine the effect of an individual component of the proposed network, a comprehensive ablation study is conducted on the DAVIS-2016 database.

Proposed network used three consecutive RGB frames and optical flow as inputs. Thus, the contribution of each input is to be analyzed. To do this, the effectiveness is evaluated on the presence of combined and individual inputs. Table 3 gives a quantitative comparison in terms of average F-measure and mean absolute error (MAE). The combination of input frames with optical flow contributed more as compared to individual inputs.

In the proposed approach, four inputs streams (three RGB frames and optical flow) are processed parallelly. **Does the parallel processing of three RGB frames contributed to the proposed network?** To examine this, results are obtained using three-stream (two RGB frames and optical flow) and four-stream. Also, the extracted feature from each encoder level of each scale is subtracted and concatenated in the EEM module. **How important is the feature concatenation against addition?** To evaluate the importance, results are examined with addition and concatenation operation in the EEM module. While designing the network, the filter size plays a key role for effective feature learning. Thus, accuracy is analysed by combining the 3×3 filters with 5×5 and 7×7 filters. Specifically, the combination of 3×3 and 5×5 filters in the EEM module with the additional operation is denoted as **3_5_ADD** and similarly for all other combinations. The results of all combinations is illustrated in Table 4. From Table 4, it is concluded that the parallel processing of four streams with 3×3 and 7×7 concatenation operation ***i.e.* 3_7_CONCAT** in EEM module outperform the other combinations.

The motion features from the optical flow encoder stream is combined with appearance-based features of the last EEM module using the bridge network. **How to bridge network helps the proposed approach to learn the effective foreground relevant features?** In a bridge network, DRB blocks are used for effective feature learning. Hence, we verified the performance of the proposed network without DRB and with different number of DRBs. Quantitative results with a fusion of DRBs is given in the Table 5. The proposed network with two DRBs shows improved performance compared to the other existing combinations.

In summary, we verified that how each component (**parallel processing, multi-scale filters, DRBs block**) is helping the proposed network for effective and significant feature learning of the foreground object(s) segmentation.

Input(s)	F-measure	MAE
Only optical flow (OF)	0.8648	0.0296
Only Input Frames (IFs)	0.8246	0.0395
Combination of OF and IFs	0.9149	0.0191

Table 3. Result ablation with different combination of input to the network on DAVIS-2016.

Approach with	3 Stream		4 Stream	
	F mea	MAE	F mea	MAE
3_5_ADD	0.8545	0.0258	0.8635	0.0239
3_5_CONCAT	0.8601	0.0249	0.8733	0.0265
3_7_ADD	0.8793	0.0222	0.8937	0.0215
3_7_CONCAT	0.8908	0.0219	0.9149	0.0191

Table 4. Fusion ablation of Multi-scale feature on DAVIS-2016.

Approach	F-measure	MAE
without DRB	0.8701	0.0229
with one DRB	0.8917	0.0201
with two DRB	0.9149	0.0191
with three DRB	0.8667	0.0239

Table 5. Results analysis with different number of DRBs in bridge network on DAVIS-2016 database.

6.3. Results analysis of GTT

In this experiments, the detection accuracy of the proposed method is verified on CDnet-2014 dataset using globally trained network. Video frames having spatial resolution varying from 320×420 to 720×480 and duration of videos from 900 to 7000 frames with different number of moving objects. The considered videos from different video categories are baseline (*highway, office, pedestrians, PETS2006*), bad weather (*blizzard, skating*), camera jitter (*boulevard, traffic*) and shadow (*backdoor, copyMachine, peoplenShade*). Accuracy is measured in terms of average F-measure and compared with state-of-the-art methods [2], [4] and [1], [30], [5]. Quantitative and visual results are illustrated in Table 6 and Figure 6 respectively. Some of the

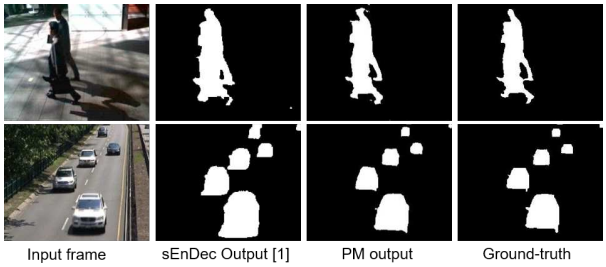

Figure 6. Visual results on CDnet-2014 database with sEnDec [1].

Methods	Publications	F-measure
MSFgNet [30]	TITS-18	0.915
DeepBs [5]	PRL-18	0.932
sEnDec [1]	TITS-19	0.961
3DLSTM [2]	TITS-19	0.964
MRCNN [4]	TVT-19	0.941
Proposed Method	-	**0.969**

Table 6. Average F-measure comparison of proposed method with existing methods for MOS on CDnet-2014 database.

approaches [2], [4] and [1] achieved promising results on CDnet-2014 database with baseline video training and fine-tuning on the some frames of target video. From Table 6 and Figure 6, it is clear that the proposed apporach outperforms the existing state-of-the-art mehods [2], [4] and [1], [30] without fine-tuning on the target video frames (only with global training) for MOS.

6.4. Results analysis of CTT

GTFD database is one of the recently published video databases for the MOS task with RGB as well as thermal data. To analyse the effectiveness of the proposed approach without optical flow, thermal data based training-testing is carried out. GTFD database comprises of 25 videos with high diversity and under different challenging situations like low illumination, etc. For result analysis purpose, each video frame is annotated manually by one person to keep a high consistency. The quantitative results of the proposed method are compared with existing state-of-the-art methods in terms of average F-measure and it is given in Table 7. The sample visual results is illustrated in Figure 7. The visual and quantitative results from Figure 7 and Table 7 are evidence that the proposed method outperforms the existing state-of-the-art methods on thermal data for MOS.

Performance analysis: Proposed method achieved state-of-the-art performance in terms of accuracy when compared to the existing end-to-end models [34], [14], [15], [44], [6], [26], [37], [24], [20], [28]. Some existing methods achieved promising results regardless of system complexity or requires fine-tuning on first frame of test video [24], [11], [40], [26], [28]. Also, the accuracy of the proposed method on weather degraded or multi-objects traffic videos is better than [2], [4]. On a single GPU of NVIDIA DGX station, we measured the average time required to process one frame is 51 msec, including optical flow time. These above observations lead the proposed method towards more practical serviceability. Finally, we observed that two scenarios in which the performance of the proposed method is limited. (1) multi-objects scenarios with moving background (2) complex motion with long-term occlusion. This could be because of the fast-moving background and long-term occlusion.

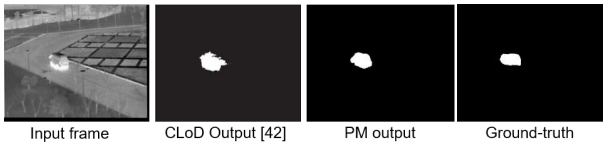

Figure 7. Visual results on GTFD database with CLoD [42].

Methods	Publications	F-measure
CLoD [42]	TCSVT-18	0.66
WELD [17]	TCSVT-17	0.67
F-WELD [17]	TCSVT-17	0.73
Proposed Method	-	**0.75**

Table 7. Quantitative results comparison of proposed method with existing sate-of-the-art methods on GTFD database.

7. Conclusion

MOS is a highly demanding and challenging task for automated outdoor video surveillance. Many methods are proposed with fruitful results for the MOS task, but some of them have limited practical usability because of complex training procedures or system complexity. At this end, we proposed an inherent correlation learning-based edge extraction mechanism (EEM) and dense residual block (DRBs) with parallel processing of RGB frames and optical flow for discriminative foreground representation. Additionally, to generate accurate and consistent foreground object mask, the decoder block is used with skip connections of subsequent multi-scale EEM features and respective down-sampled version of previous frame output. To demonstrate the effectiveness of the proposed framework, experiments are conducted on four benchmark and challenging datasets *i.e.* DAVIS-2016, SegTrack-v2, CDnet-2014 and GTFD. The experimental analysis demonstrates that the proposed network achieves favorable performance compared to the state-of-the-art methods without any pre-trained model or fine-tuning of the model on a test video frame(s) for MOS.

Acknowledgement

This work was supported by 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 Yimin Yang. Fusion-based foreground enhancement for background subtraction using multivariate multi-model gaussian distribution. *Information Sciences*, 430:414–431, 2018.

[4] 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.

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

[6] 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.

[7] 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.

[8] Akshay Dudhane, Harshjeet Singh Aulakh, and Subrahmanyam Murala. Ri-gan: An end-to-end network for single image haze removal. In *Proceedings of the IEEE Conference on CVPRW*, pages 0–0, 2019.

[9] Brent Griffin and Jason Corso. Tukey-inspired video object segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1723–1733. IEEE, 2019.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Ping Hu, Gang Wang, Xiangfei Kong, Jason Kuen, and Yap-Peng Tan. Motion-guided cascaded refinement network for video object segmentation. In *Proceedings of the IEEE Conference on CVPR*, pages 1400–1409, 2018.

[12] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[13] 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 CVPR*, pages 1125–1134, 2017.

[14] 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.

[15] 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.

[16] Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE Transactions on Image Processing*, 27(10):5002–5015, 2018.

[17] Chenglong Li, Xiao Wang, Lei Zhang, Jin Tang, Hejun Wu, and Liang Lin. Weighted low-rank decomposition for robust grayscale-thermal foreground detection. *IEEE Transactions on CSVT*, 27(4):725–738, 2017.

[18] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, pages 2192–2199, 2013.

[19] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*, pages 478–487, 2016.

[20] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7274–7283, 2019.

[21] Huaijia Lin, Xiaojuan Qi, and Jiaya Jia. Agss-vos: Attention guided single-shot video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3949–3957, 2019.

[22] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *The IEEE Conference on CVPR*, June 2019.

[23] 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.

[24] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, pages 565–580. Springer, 2018.

[25] Léo Maczyta, Patrick Bouthemy, and O Le Meur. Unsupervised motion saliency map estimation based on optical flow inpainting. pages 4469–4473, 2019.

[26] 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.

[27] Kuldeep Marotirao Biradar, Ayushi Gupta, Murari Mandal, and Santosh Kumar Vipparthi. Challenges in time-stamp aware anomaly detection in traffic videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 13–20, 2019.

[28] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9226–9235, 2019.

[29] Prashant Patil and Subrahmanyam Murala. Fggan: A cascaded unpaired learning for background estimation and fore-

ground segmentation. In *2019 IEEE WACV*, pages 1770–1778. IEEE, 2019.

[30] 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*, 20(10):4066–4077, 2019.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3978–3987, 2019.

[38] 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.

[39] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *Proceedings of the IEEE Conference on CVPR*, pages 9994–10003, 2019.

[40] 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.

[41] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on CVPR*, pages 6499–6507, 2018.

[42] Sen Yang, Bin Luo, Chenglong Li, Guizhao Wang, and Jin Tang. Fast grayscale-thermal foreground detection with collaborative low-rank decomposition. *IEEE Transactions on CSVT*, 28(10):2574–2585, 2018.

[43] Xiaohui Zeng, Renjie Liao, Li Gu, Yuwen Xiong, Sanja Fidler, and Raquel Urtasun. Dmm-net: Differentiable mask-matching network for video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3929–3938, 2019.

[44] Lu Zhang, Zhe Lin, Jianming Zhang, Huchuan Lu, and You He. Fast video object segmentation via dynamic targeting network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5582–5591, 2019.

[45] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Unsupervised online video object segmentation with motion property understanding. *IEEE Transactions on Image Processing*, 29:237–249, 2019.