Multimedia Tools and Applications (2023) 82:6769–6785
https://doi.org/10.1007/s11042-022-12981-2

Video object segmentation guided refinement on foreground-background objects

J. Sarala Devi[1] · A. Razia Sulthana[1]

Received: 8 February 2021 / Revised: 19 May 2021 / Accepted: 27 March 2022 /
Published online: 9 August 2022

Abstract

Video Object Segmentation (VOS) for separating a foreground object from a video sequence is an intricate task and relies on fine-tuning. Many recent approaches focus on pixel-wise matching of foreground objects and gives importance to balancing the relation between the pixels for identifying the foreground objects and might lead to misclassification. This paper explores mapping between the foreground and background objects in semi-supervised VOS by balancing and mutually mapping the pixels between the foreground and background objects. The proposed model makes practical and effective use of enhanced pixel and instance level matching to improve the prediction. Moreover, the framework implements ensemble learning with a Leaky-ReLU activation function that improves the segmentation process. To evaluate the results of object segmentation process, J and F scores are measured. We carry experiments broadly on popular benchmark DAVIS, in the versions 2016 and 2017. Our Model achieves a promising performance of J & F score of 82%, surpassing all the other techniques.

Keywords Video object segmentation · Ensemble learning · Pixel level mapping · Instance level matching · ResNet · Activation function

1 Introduction

Video Object Segmentation (VOS) [15] is an essential piece of work that simultaneously segments and classifies various objects from a video sequence. It aims to locate the foreground/background label for each pixel in a video image frame. Video object segmentation is widely applied to many practical applications like traffic control systems [42], recognition tasks [29, 33] object detection [45, 47] medical imaging [36], etc. Video appeals its reach to the audience as it is preferred by people of varying personalities. One of the well-known

✉ A. Razia Sulthana
razia@dubai.bits-pilani.ac.in

J. Sarala Devi
saraladevi77@gmail.com

[1] Department of Computer Science and Engineering, Birla Institute of Technology and Science, Pilani Dubai, UAE

video-sharing platforms is YouTube that brings in nearly 5 billion videos to its end-user [7]. Another platform, Facebook enables sharing videos and provides access to more than 8 billion videos for its users [26]. Every year the average amount of time spent on watching video clips over social media like Instagram is increasing by 80%.

As the volume and variety of videos are likely increasing day by day, getting images annotated according to our specifications is a challenge. These annotations are most commonly used for object recognition and scene understanding to segment image frames from whole-images [41]. Annotation is a type of manually defining or labeling the region in an image that is otherwise called tagging in general terms. The higher the quality of our image being annotated, the better our models can likely perform accurately. Semantic segmentation is where the targeted objects are grouped by associating each pixel of an image with its corresponding class label. VOS has many prominent challenges like occlusion [20], low resolution [1], non-rigid deformation [27], motion blur [15], appearance changes [8], scale variation and near-far variance [23].

Occlusion is a scenario of overlapping objects or seeming to overlap in motion; Low resolution is caused by a reduced number of pixels than obligatory, thus blurring the images; Non-rigid deformation is the change in the object structure or boundary concerning an increase/decrease in size; Motion-blur is the streaking texture of moving objects captured by the camera, for example, rainfall; Appearance change refers to any change in spatial alignment or texture of the image; Scale variation refers to the trivial difference found in the size of the object captured in correspondence with the distance from where it is captured; And finally near-far variance arises when the objects are captured at a distance, as they seem smaller than those captured at closer proximity.

Off late, the techniques [2, 18, 32] applied in VOS, fine-tune the image by comparing and embedding information from first and previous object frames to the current frame. A couple of other contemporary works like STMVOS [24] and FEELVOS [31] have inbuilt fine-tuning for VOS and compete in gaining good accuracy. STMVOS relies on large image datasets for training purposes and requires large-scale frame sequences. The image segmentation technique proposed in [44] is applied on medical images using adaptive perturbation methodology. The author in article [21] proposes new ways for feature edge detection and edge splitting operations. Deep auto encoder models [25] are applied in recommendation system for identifying images or segmenting images using multi layer architecture and matrix factorization approach. The proposed model applies segmenting images from videos and its extension can be applied on videos of purchase made from shopping malls to know the requirements and style of customers and the same can be recommended to the other similar customers.

The comparison of contemporary VOS techniques like OSMN [43], SiaMask [35], OSVOS [2], OnAVOS [32], FEELVOS, PReMVOS [23], and STMVOS approaches over the dataset 'DAVIS 2017-Semi Supervised' is shown in Fig. 1. Many of the contemporary approaches pay less attention to analyze the background information of a video frame and feature embedding approaches applied on them. It is found to be quite challenging to subtract the background features from a video frame and to extract the foreground objects from them. This observation motivates us to propose a framework and a suitable optimal embedding methodology that matches the foreground and background image in the target frame.

The reason being choosing semi-supervised VOS algorithm is that it is trained upon a combination of labeled and unlabeled data. It assumes that the points which are closer to each other are more likely to have same output label. The development of semi-supervised

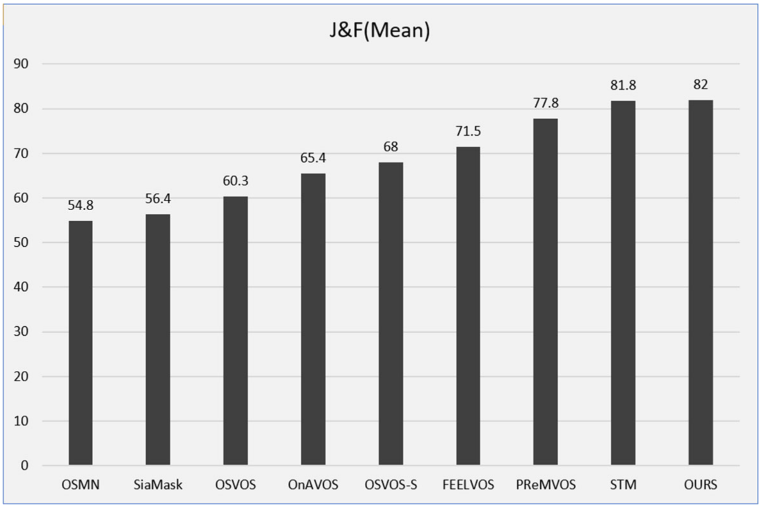

Fig. 1 Comparison of DAVIS 2017 achieved by recent methods

VOS can benefit many related tasks, such as video instance segmentation and interactive video object segmentation. Semi supervised methods target the objects in motion and identifies the image sequence in three dimensions: previous frame, current frame, first frame.

Design goals: The proposed work focuses on

1. Applying semi-supervised VOS approach that targets the objects in motion and handles the ‘Motion Blur’ problem effectively.
2. It identifies the image sequence in three dimensions: previous frame, current frame and first frame. The objects in the first frame are annotated and the collection of annotated objects is used as a training dataset or as input to the proposed system. The model is an end to end analysis of objects in video as it does not depend only on analyzing the first frame rather all the frames.
3. It applies the following embedding techniques:
   a. Enhanced Pixel level Matching b. Enhanced instance-level Matching c. Ensemble Learning
4. The scale invariance challenge persisting in contemporary work is overcome by the ensemble embedding technique. It matches the foreground and background image and segments accurately the foreground image from the background.
5. It handles non-rigid deformation challenges encountered in contemporary VOS techniques.
6. The model is simple, effective and strong producing a improved J and F value as compared to existing models.

The proposed model comes closer to FEELVOS because the FEELVOS extracts pixel-level embedding to match object information and also uses the global and local matching technique. The pixel-level matching by itself will not be sufficient to target the foreground object frames and might lead to unpredicted disturbances or noise due to the difference in pixel conditions. However, FEELVOS is limited to predict only the foreground objects as it fixes matching between the neighboring pixels only. To overcome this scale invariance, we introduce an instance-level matching [14, 37] in combination with pixel-level embedding that uses an attention mechanism to segment large scale objects. We apply enhanced pixel-wise matching, following which instance-level matching for both the foreground-background objects. Thus, the suggested framework can successfully improve the quality of matching the targeted objects in segmenting semi-supervised VOS and at the same time keeping the model simple and more effective.

The remaining part of the paper is structured as detailed. In Section 2, a complete state-of-art of existing VOS techniques, their advantages, and concerns are elucidated. In Section 3, the proposed idea and steps carried out under pixel level matching is explained. In Section 4, the architecture of the proposed system is described. In Section 5, procedure of implementation is detailed and the results are tabulated in Section 6. Section 7 concludes the research work.

2 Literature review

2.1 Models with fine-tuning

It becomes an obligatory task to relate the frames in a video sequence to end up precisely segmenting the target object. Several works are done over semi-supervised VOS to boost the object segmentation quality and these proposed techniques depend on fine-tuning to segment the target object. OSVOS and MoNet [39] predict the results from the first frame and make use of this predicted result to tune the model during the test time. Another fine-tuning model, OnAVOS widens the fine-tuning of the first frame on applying an online adaptation approach that uses heuristics-based fine-tuning policies to achieve better results. One of the characteristics of identifying the relationship between the frames is optical flow. An optic flow defines the change in velocity or variation in the brightness pattern of images in different frames. This approach is used in MaskTrack [26] that transfers the segmentation mask from one frame to another. PReMVOS conjoins four neural networks including optical flow and proposes a merging algorithm to fine-tune the image segmentation. Though the aforementioned methods strive to fine-tune segmenting the targeted images, they consume additional time, thereby slowing down the entire process.

2.2 Models without fine-tuning

A couple of research works is proposed to circumvent the problem of fine-tuning. These models also bother much to achieve respectable run time. The author in [15] proposes a model, VideoMatch that uses a soft matching layer and, in an embedding space maps the pixels of the current frame to the first frame. OSMN, ignores fine-tuning with minimal run-time by using two networks: one for making the predictions over the segmented images and another for extracting the information at the instance level. The nearest neighbor classifier is applied in PML [4] that uses pixel-wise matching. It assigns a label to every object pixel in the current and the first frame. However, it produces noisy segmentation owing to unmatching the neighboring pixels. Comparative to PML and VideoMatch methods, FEELVOS carries out fine-tuning and brings out much better speed. The information from the previous frame is stored and retrieved by STMVOS, which mirrors a kind of typical memory network. It needs extensive training with simulated images from multiple frames. Another approach RGMP [38] also demands extensive training as STMVOS. Yet, all these methods leverage to study the foreground images, overlooking the background in the image. The proposed method, however works with both foreground and background ones.

2.3 Attention mechanism in image analysis

Attention mechanisms, in general, seek to identify the pertinent part of the image that provides more information about the frame or entire frame sequence. Certain proposals in VOS chain the attention mechanism into their convolutional networks. SE-Nets [16] proposed a gated mechanism and models the channel attention in its network. OSMN uses instance-level matching and works only with foreground images. Motivated by SE-Nets and OSMN, in our proposed method, a channel-wise average pooling is applied on both the foreground and background images at the instance-level matching mechanism. Besides applying instance-level, the pixel-level matching mechanism is also applied. The proposed image matching algorithm applies average pooling. The reason behind opting for average pooling rather max pooling is the inadequacy of max pooling to identify the sharp features of the image and it goes good with dark pixels whereas the average pooling method smoothens the image and gives preference to lighter pixels too. However, on replacing average pooling by max pooling there would be a trivial change in the performance as the layers of convolution and pooling operations would handle the edge detection effectively with average pooling itself. The VOS algorithm works in linear time. Yet, a big size video might consume more execution time. As the DAVIS dataset has annotated image, additional space is not required by the algorithm during run time.

2.4 Varying activation function

ReLU (Rectified Linear Unit) is applied in several existing approaches discussed in Sections 2.2, 2.3 and 2.4. One of the downsides of applying ReLU is that for nodes with zero gradients, their weights in successive iterations will not get updated, since ReLU weighs the inactive nodes as zero value. The inactive nodes are ignored during gradient adjustment in ReLU. Further, zero gradients might slow down the complete training process. This can be alleviated by using the Leaky ReLU function [10]. The proposed approach uses Leaky ReLU as it produces a minor non-zero value for the inactive nodes, thereby tunes the weight satisfactorily during the gradient adjustment.

3 Methodology

The current methodologies in VOS focus on segmenting the foreground objects. OSMN ignores feature diversity and hence leads to coarse predictions. Though FEELVOS and PML handle the problem of feature diversity, it is drawn away by noises from the nearby pixel. On comprehensively analyzing the above-mentioned approaches, the proposed framework considers analyzing both the foreground and background objects in the frame sequence. The architecture of the proposed method is shown in Fig. 2.

As the first step of image segmentation, pixel-level matching is applied to the foreground and background objects. Pixel-level matching is a popular feature-matching technique used in semi-supervised VOS and is applied on global as well as local matching. The former matching technique uses the information of the first frame to construct the features, and the latter matching technique uses the information of the previous frame to construct the feature mapping. Successively, both the instance level and pixel-level embedding are done

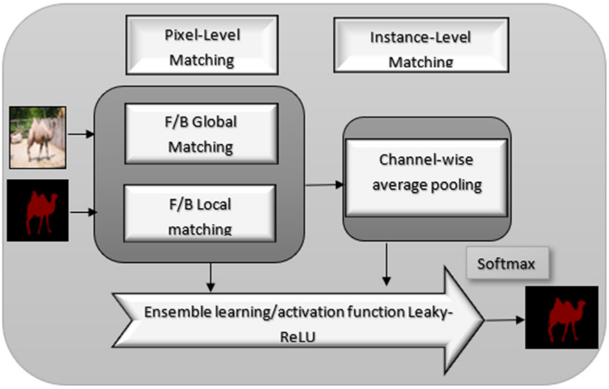

Fig. 2 Proposed framework for foreground and background (F/B) object matching

on the foreground and the background image. An attention mechanism is introduced by instance-level matching to increase the efficiency during pixel-level matching. The results of the instance and pixel matching techniques over the foreground and background images are combined using a considerable number of receptive fields thus ending up with infinite predictions. Adding ensemble learning allows large receptive field to make precise predictions.

3.1 Enhanced pixel-level matching

The following steps are carried out under enhanced pixel-level matching; Foreground and Background Matching; Global matching; Local matching.

3.1.1 Foreground- background matching

In FEELVOS, to measure the foreground pixel-level matching, a distance calculation method is applied. The distance between the pixels of the first frame and the previous frame is denoted as a and b respectively. The corresponding embedding learning for first frames and previous frames is denoted as ma and mb. The pixels that belong to the same object are considered closer in embedding space. The distance values of such pixels can be calculated using the (1) given below:

$$distance(a,b) = 1 - \frac{2}{1 - exp(0)} = 0 \tag{1}$$

The pixels that belong to distinct object are considered to be far away and the distance values of such pixels are calculated using the (2) given below:

$$distance(a,b) = 1 - \frac{2}{1 + exp(\infty)} = 1 \tag{2}$$

Equations (1) and (2) is modified to incorporate the background scenario. For a frame f, let the pixel set of background objects be bg_f and the pixel set of foreground objects be

fg_f, and the distance of pixel a and pixel b of the current frame f and frame t concerning the mapping values ma and mb is given in

$$distance_f(a,b) = 1 - \frac{2}{1 + exp(\|ma - mb\|^2 + bias_{bg})} if b \in bg_f \tag{3}$$

$$distance_f(a,b) = 1 - \frac{2}{1 + exp(\|ma - mb\|^2 + bias_{fg})} if b \in fg_f \tag{4}$$

$bias_{bg}$ represents the trainable bias of background

$bias_{fg}$ represents the trainable bias of foreground

Henceforth, from (3) and (4) the two biases for both foreground and background helps to know the variance in distances between the pixel numbers.

3.1.2 Global matching

Similar to PML, Video Match, and FEELVOS, the proposed approach considers the nearby neighboring pixels to share the feature details of the ground-truth annotated image to the current frame.

Let the pixels be denoted as A_t and the objects be denoted as Ob. Strides of 4 are taken at time t. As shown in (5) and (6), the pixels a of the first object frame otherwise known as a reference frame (t=1) and the current object frame pixels can be matched for global foreground/background matching and can be written as

The global foreground matching equation is given as

$$global_{fg}(a) = \min_{a \in b} distance(a,b) \tag{5}$$

The global background matching equation is given as

$$global_{bg}(a) = \min_{a \in b} distance(a,b) \tag{6}$$

3.1.3 Local matching

In FEELVOS, the information between pixels is shared from the first frame to the neighboring pixels of the next frame. However, it is not shared with the object frames that are far away. Also, the pixels of the objects within the nearby area suffer from scale variations and the objects will not have a similar look. Hence, the proposed approach works over the extended level of matching between the objects that suffer from the multi-scale variant issue. This kind of matching is more powerful to detect or segment objects in fast motion-blurred frames.

In addition to matching between the local and the global objects, we combine the previous frame i.e. pixel-wise feature map, with the current frame. The feature map from the first frame is extracted and the video sequence is analyzed frame-by-frame to determine the feature that matches with the current frame. Successively, we relate local and global matching to the first and previous frame and arrive at the final segmented object.

3.2 Enhanced instance-level matching

After receiving the pixel matching patterns of the first frame and previous frames, the mapped pixels are divided into pixels from the foreground and pixels from the background as given here: $(a_1, \bar{a}_1, a_{(T-1)}, \bar{a}_{(T-1)})$ based on their mask. Following this, channel-wise average pooling is applied on the pixels and an instance-level vector is obtained. This vector encompasses information of all the frames from both the foreground and background objects. Finally a module with a fully connected layer, a non-linear activation function that inputs each ResNet block is built. The instance-level matching learns complete information of the foreground and background pixels and sounds good in handling local ambiguity between them.

3.3 Ensemble learning

Following ResNets [13] and Deeplab [3, 6], the proposed ensemble learning module contains Res-blocks and ASPP [34] which undergo channel-wise average pooling and bilinear up-sampling to capture the multi-scale dimension [30] precisely. To improve the receptive field, dilated Convolutional layers are added with a well-defined gap.

Another important feature of the proposed system is that Leaky-ReLU is employed instead of ReLU as the activation function [11, 12, 46]. The difference between ReLU and Leaky-ReLU is pictured in Fig. 3. All the transposed convolution up-sampling layers follow ReLU and Leaky-ReLU non-linearity and are initialized with the fan-out initialization mode [19].

ReLU is the most used activation function in CNN, which determines the output of the network [6] and also has a considerable effect on the extraction of a feature attribute. The equation of ReLU is given below:

$$func\,(y) = \left\{ \begin{matrix} y_k\, if & y_k > 0 \\ 0\, if & y_k \leq 0 \end{matrix} \right\} \tag{7}$$

Where y_k denotes input to the ReLU $(k^{th} channel)$

func(y) denotes the output of the ReLU $(k^{th} channel)$ and (7) can also be written as in (8).

$$func(y) = max(y_k) \tag{8}$$

The ReLU activation function has a sparse activation probability and it can be achieved by zero threshold value. It accurately classifies two-class data values and does not have any

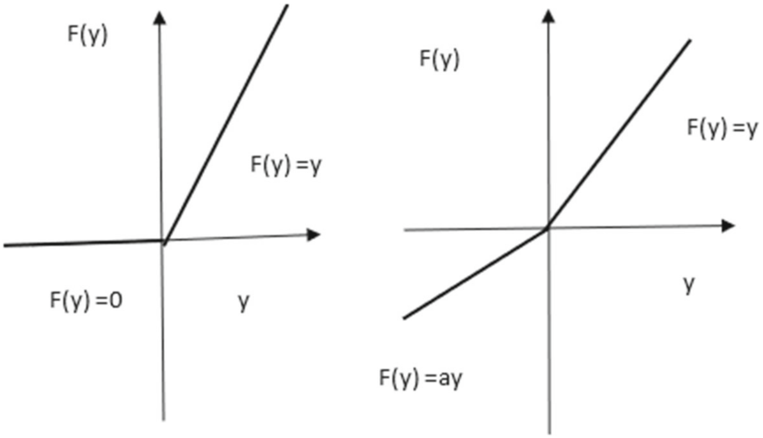

Fig. 3 ReLU vs Leaky-ReLU

gradient diffusion problem. The training process in ReLU might slow down with continuous null gradients and is considered to be one of the drawback of ReLU.

To find a solution to this, the LReLU non-linearity activation function is presented to permit a small, inactive non-zero value for the negative parts [40].

$$func\,(y) = \left\{ \begin{matrix} y_k\, if & y_k > 0 \\ \lambda y_k\, if & y_k \leq 0 \end{matrix} \right\} \tag{9}$$

Where λ denotes a predefined parameter. It takes a value of 0.01. Equation (9) can also be written as in (10).

$$func(y_k) = max(y_k, 0), \lambda min(y_k, 0) \tag{10}$$

LReLU flattens the negative part that comes up with a small, non-zero gradient value when the unit is inactive whereas ReLU does not. We chose 'fan-out' which preserves the magnitude of the variance of the weights in the backward pass, i.e., the initialization properly scales the forward signal. It is worth to have fan-out when the loss oscillates more.

4 Architecture

DeepLabv3+ architecture [5] is used in the proposed model. It is based on the dilated Resnet-101, which is the backbone for our network. ResNet helps to train extreme deep neural networks easily. ResNet architecture is made up of residual blocks which are described in Fig. 4.

The purpose of using ResNet is to decide and calculate the type of increment that is to be added to the input nodes in order to achieve the accurate output. To reduce the spatial dimension, each block in the ResNet manages to have better backpropagation and provides max-pooling operations. It also helps in training the network model smoothly. A notable feature of ResNet is that it perfectly learns to identify textures; detect edges and objects from the images. It uses less computational resources and executes in minimal time.

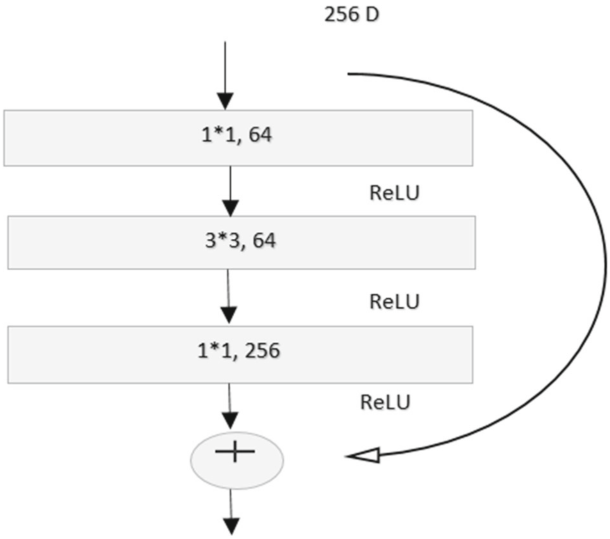

Fig. 4 ResNet 101 Residual Block

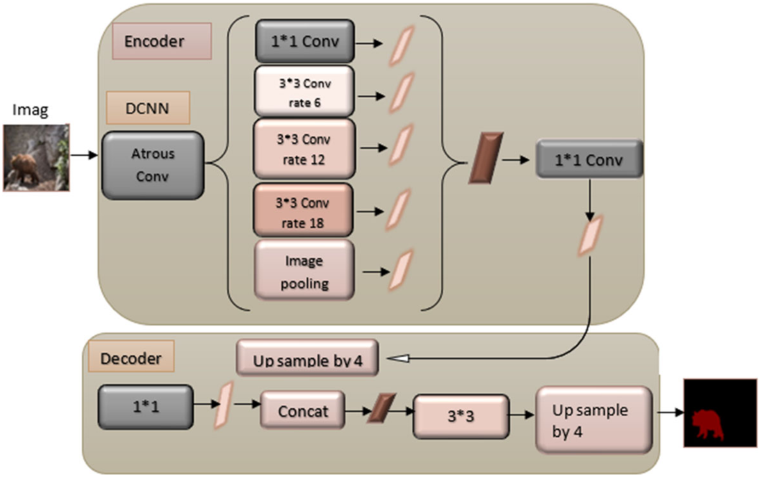

Fig. 5 Deeplabv3+ Architecture

Deeplab structure shown in Fig. 5, is considered as a specially designed architecture when compared with encoder-decoder design because it helps to perfectly extract the multi-scale features while segmenting video objects. Rather than performing regular convolution operation, the last ResNet block in Deeplab carries out atrous convolutions ASPP that uses different dilation rates to capture the multi-scale feature information.

5 Implementation details

Several datasets [9, 22] are available for video object segmentation. Yet, not all of them have been explicitly designed for identifying the pixel relationships of the foreground object from the background area.

5.1 Dataset

The proposed approach uses DAVIS (Densely Annotated Video Segmentation) dataset. DAVIS comprises of two versions: DAVIS 2016 and DAVIS 2017. It encompasses high-quality, full high definition video sequences. The number of sequences, number of frames, and number of objects of the DAVIS 2016 and 2017 datasets are tabulated in Tables 1 and 2 respectively. Table 1 shows the layout of the DAVIS 2016 dataset's structure and Table 2 shows the layout of the DAVIS 2017 dataset's structure.

Table 1 Structure of DAVIS 2016 dataset

DAVIS 2016	train	val	Total
Number of Sequences	30	20	50
Number of Frames	2079	1376	3455
Number of Objects	30	20	50

Table 2 Structure of DAVIS 2017 dataset

DAVIS 2017	Train set	Val set	Test-dev	Test challenge	Total sets
Number of Sequences	60	30	30	30	150
Number of Frames	4219	2023	2037	2180	10459
Number of Objects	138	59	89	90	376

The 2016 version provides annotations for foreground/background objects while the 2017 version provides annotations for multiple objects and instances in the foreground. Each pixel images in the video are pixel-accurate and densely annotated. Some examples of the annotations mask are shown in Fig. 6

5.2 Model analysis

In the proposed framework, we consider a randomly chosen first frame which is sampled and the images from the remaining frames; previous frames and current frames is segmented.

The existing models and their method of implementation is studied well and in the proposed system the architectural implementation and structure of CNN model is varied such that there is a notable increase in accuracy of the prediction system. The ablation study hence remodifies the entire structure and the implementation methodology as detailed below.

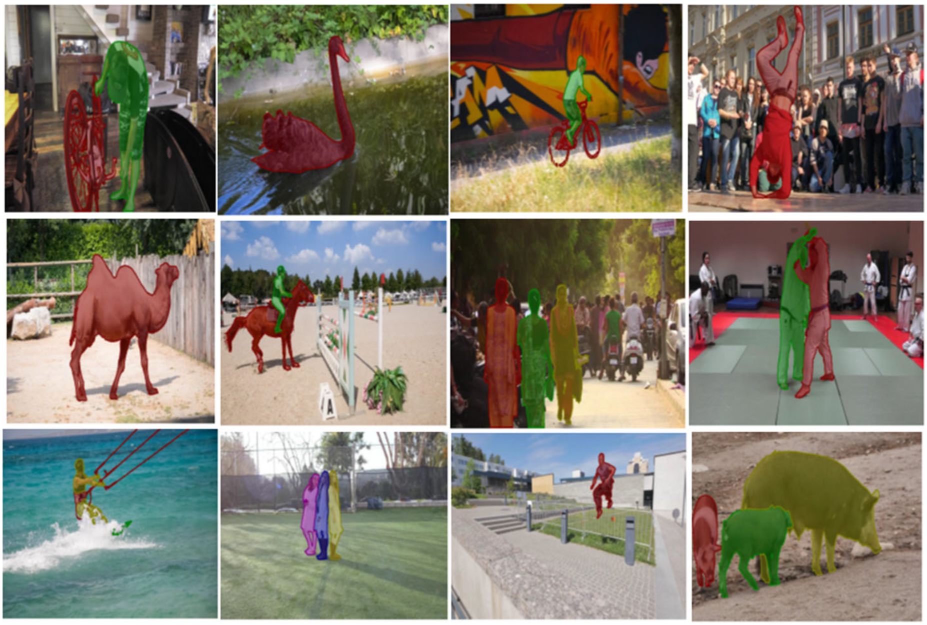

Fig. 6 First frame annotation for all the sequences in the validation subset of DAVIS. The segmented images are highlighted and shown in the image

For pixel-wise matching, we apply 3*3 convolutions that contain the batch normalization [17] and non-linearity activation function where the dimensions of an image get reduced and restored. One depth-wise separable convolution is applied with a stride of 4. The size of the strides benefits substantially faster training. Stride 16 can deal with feature objects that are four times smaller than stride 8 and can also help in producing finer segmentation results. However, it might take more training time. For local matching, we initialize bias for foreground and background $bias_{fg}$ and $bias_{bg}$ to 0. We additionally down-sample the object information to reduce its dimension with bilinear interpolation.

Group normalization is applied as an alternative to batch normalization and channel-wise average pooling with an attention mechanism is applied to improve the performance. Group Normalization computes mean and variance and it is independent of any batch sizes. Interestingly, we can see group normalization has a very lower error and comparably provides good results than batch normalization. Additionally, group normalization can be easily transferred from the pre-training process to adopt fine-tuning in video sequence segmentation.

DAVIS 2017 and DAVIS 2016 training sets are used as the training data. It has the default setting of the down-sampled 480p resolution video series. We apply stochastic gradient descent together with a learning rate of about 0.006 and a momentum of 9. Cross-entropy loss is an important cost function used to optimize segmentation methods. Here we have adopted bootstrap cross-entropy loss that only contemplates a fraction of 15%. This kind of cost function works correctly with any variants and unevenly distributed class, which are quite common for VOS. During the training stage, batch normalization parameters in the backbone are disabled.

6 Results and experiments

Later, when all the training process is done, DAVIS 2016 and the DAVIS 2017 [28] Val sets are used for estimating our framework results.

The DAVIS 2016 validation set contains 20 video sequences, and each one is an annotated single instance. We evaluate all the frames from the video sequence that are generated with our algorithm and are compared with the previous methods. Table 3 shows the results of state-of-art-methods against the proposed method.

The DAVIS 2017 dataset contains 60 training video series with multiple masks and a Val set that is extended from DAVIS 2016 which has 30 videos.

Table 3 DAVIS 2017 Val set on various models. We present the J & F score and frame per second (fps)

Models	*fps*	*J score*	*F Score*
MaskTrack [9]	0.08	79.70%	75.40%
OSVOS [16]	0.11	79.80%	80.60%
OnAVOS [17]	0.08	86.10%	84.90%
OSMN [21]	7.14	74.00%	72.90%
VideoMatch [1]	3.13	81.00%	-
RGMP [42]	7.69	81.50%	82.00%
FEELVOS [20]	2.22	81.10%	82.20%
(Ours)	Below 5	79.50%	84.60%

Table 4 J & F Mean score for our framework

Metrics	*Values*
J Mean	0.795
J Recall	0.888
J Decay	0.092
F Mean	0.846
F Recall	0.927
F Decay	0.107
J & F- Mean	.82

We assess our model on the DAVIS in both the 2016 version and the 2017 version. Table 4 tabulates the Jaccard and F-measure values of the proposed method and the proposed method obtains a means J & F value of 82% which is significantly larger than existing methods.

Intersection-Over-Union (IoU) or (Jaccard Index) is a measure of the percentage of overlapping object masks and the prediction output. It measures the pixels present across the object mask and provides a score for the segmentation prediction. The contour accuracy estimates the F-measure.

$$J\ score\ = \frac{Output\ Segmented\ Image\ \cap\ Ground\ Truth}{Output\ Segmented\ Image\ \cup\ Ground\ Truth} \tag{11}$$

$$F\ score\ = \frac{2 *\ Precision\ *\ Recall}{Precision\ +\ Recall} \tag{12}$$

We calculate our validation set results by manually created code and the test-dev result sets on the official evaluation server codalab.

In Fig. 7, the first video shows that the framework accurately detects the person and the dog though the image suffers from occlusion. In the second video, the framework succeeds in tracking the right person amidst the other similar person’s. Segmentation results of the proposed method produces results better than other approaches with a minimal inference speed.

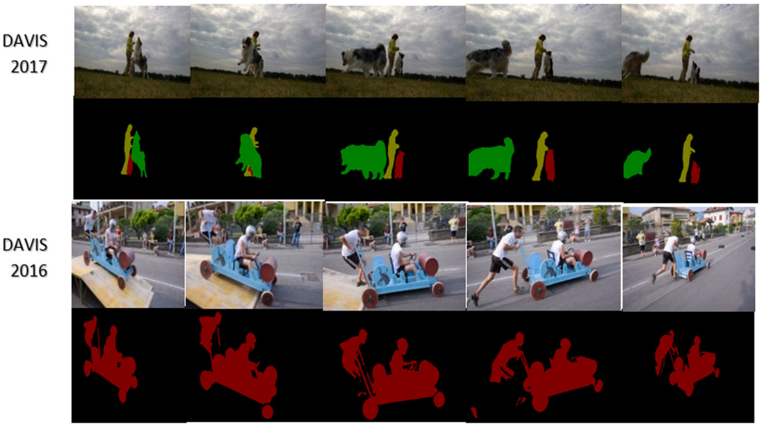

Fig. 7 Comparative performance on both 2016 and 2017 version

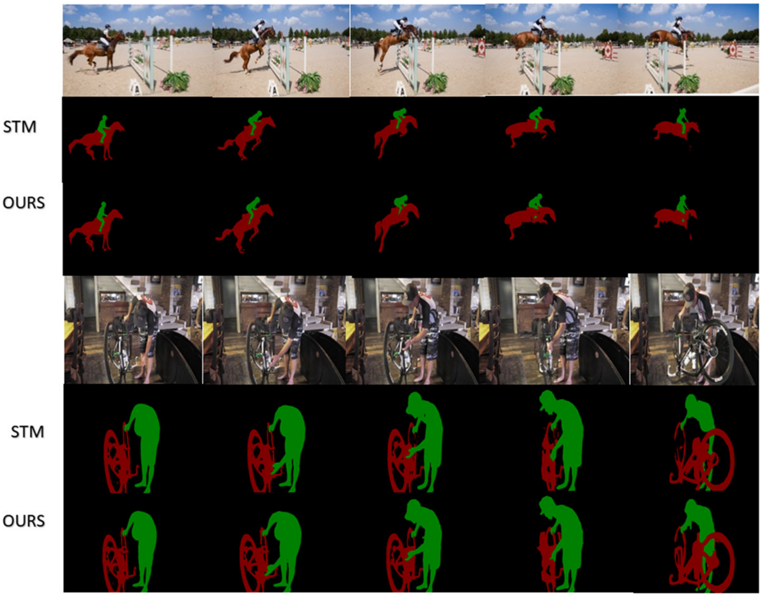

Fig. 8 Comparison with STMVOS

As it is seen from the existing models in Table 3, there is a variation found in the J score and F score. The proportionality between the J score and F score is unpredictable and depends on the type of algorithm. The proposed approach shows notable value for J and F score except for slight decrease in J score and it is because of that J score measures the intersection over union and a minimal diversity in sample image might disturb the J value. The fps (frames per second) is very important for analyzing videos and extracting frames from them. Table 3 gives the comparison of fps for different existing models. Normally for image segmentation when applied on videos, it demands a minimum range of 5-7 frames per second.

We evaluate the comprehensive analysis of the performance of our model with other models. Figure 8 shows that STMVOS overlooks to segment the back leg of the horse in the occlusion and motion-blur scenario. In the second video sequence, STMVOS fails to segment the bicycle too. The proposed framework segments both the leg and the bicycle significantly better than STMVOS.

On comparing with the results of FEELVOS, our framework makes a significantly higher score over FEELVOS (82.0% vs. 71.5%). Promisingly, if augmentation is applied to the proposed approach, at the evaluation phase, the J & F means can be further increased to more than 85%. Finally, the proposed model can able to accurately segment images in challenging situations, such as occlusion, blur, and deformations, etc.

7 Conclusion

In this paper, a new framework to handle segmentation problems in images in VOS is proposed. In particular, we solve the problem of motion blur and the scale variation that includes edge ambiguity, shape complexity, etc. Though many research efforts have been focused on estimating the variation in objects in the past decades, they yielded minimal accuracy. And most of the articles in the past focused only on foreground objects but paid little attention to the objects in the background region. We have introduced a new method to combine and process foreground and background objects and obtained the final promising results. Specifically, we introduced an approach to match and segment the images irrespective of scale invariance on multiple objects in the foreground and incorporated the background information into it using pixel level matching and instance level matching. Moreover, we combined all the modules to make our framework so powerful and to fill the gap of inadequate accuracy. To conclude, the presented method could produce better results if it can be modified with random-crop augmentation or balanced Random-crop augmentation approaches.

Declarations

Author Contributions Both the authors have equally contributed to the work

Funding NA

Availability of Data And Material The relevant data and material towards the work are available with authors

Code Availability The relevant code towards the work are available with authors

Ethics Approval The manuscript has not been submitted to any other journal nor a part of the work is published anywhere

Consent for Publication The authors would be ready for publication if this article is accepted by journal

Competing Interest

Conflicts of Interests There is no conflict of interest with any firm or person.

References

1. Bao L, Wu B, Liu W (2018) CNN in MRF Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5977–5986
2. Caelles S, Maninis K-K, Pont-Tuset J, Leal-Taixé L, Cremers D, Gool LV (2017) One-shot video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 221–230
3. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille Al (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
4. Chen Y, Pont-Tuset J, Montes A, Gool LV (2018) Blazingly fast video object segmentation with pixelwise metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1189–1198
5. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
6. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818

7. Cheng H-T, Chao C-H, Dong J-D, Wen H-K, Liu T-L, Sun M (2018) Cube padding for weakly-supervised saliency prediction in 360 videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1420–1429
8. Cheng J, Tsai Y-H, Hung W-C, Wang S, Yang M-H (2018) Fast and accurate online video object segmentation via tracking parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7415–7424
9. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255
10. Favorskaya MN, Andreev VV (2019) The study of activation functions in deep learning for pedestrian detection and tracking. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences
11. Glorot X, Bengio Y (2018) Understanding the difficulty of training deep feedforward neural networks. 2010. In: International Conference on Artificial Intelligence and Statistics
12. Griffin BA, Corso JJ (2019) Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8914–8923
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
14. Hu Y-T, Huang J-B, Schwing A (2017) Maskrnn: Instance level video object segmentation. Adv Neural Inf Process Syst 30:325–334
15. Hu Y-T, Huang J-B, Schwing AG (2018) Videomatch: Matching based video object segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 54–70
16. Hu J, Li S, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
17. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167
18. Johnander J, Danelljan M, Brissman E, Khan FS, Felsberg M (2019) A generative appearance model for end-to-end video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8953–8962
19. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
20. Li X, Loy CC (2018) Video object segmentation with joint re-identification and attention-aware mask propagation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 90–105
21. Liang Y, He F, Zeng X (2020) 3D mesh simplification with feature preservation based on Whale Optimization Algorithm and Differential Evolution. Integrated Computer-Aided Engineering Preprint, pp 1–19
22. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, pp 740–755. Springer, Cham
23. Luiten J, Voigtlaender P, Leibe B (2018) Premvos: Proposal-generation, refinement and merging for video object segmentation. In: Asian Conference on Computer Vision. Springer, Cham, pp 565–580
24. Oh SW, Lee J-Y, Xu N, Kim SJ (2019) Video object segmentation using space-time memory networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp 9226–9235
25. Pan Y, He F, Yu H (2020) Learning social representations with deep autoencoder for recommender system. World Wide Web 23(4):2259–2279
26. Perazzi F, Khoreva A, Benenson R, Schiele B, Sorkine-Hornung A (2017) Learning video object segmentation from static images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2663–2672
27. Perazzi F, Pont-Tuset J, McWilliams B, Gool LV, Gross M, Sorkine-Hornung A (2016) A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 724–732
28. Pont-Tuset J, Perazzi F, Caelles S, Arbeláez P, Sorkine-Hornung A, Van Gool L (2017) The 2017 davis challenge on video object segmentation. arXiv:1704.00675
29. Ventura C, Bellver M, Girbau A, Salvador A, Marques F, Giro-i-Nieto X (2019) Rvos: End-to-end recurrent network for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5277–5286
30. Visin F, Ciccone M, Romero A, Kastner K, Cho K, Bengio Y, Matteucci M, Courville A (2016) Reseg: A recurrent neural network-based model for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 41–48

31. Voigtlaender P, Chai Y, Schroff F, Adam H, Leibe B, Chen L-C (2019) Feelvos: Fast end-to-end embedding learning for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9481–9490
32. Voigtlaender P, Leibe B (2017) Online adaptation of convolutional neural networks for video object segmentation. arXiv:1706.09364
33. Wang W, Song H, Zhao S, Shen J, Zhao S, Hoi SCH, Ling H (2019) Learning unsupervised video object segmentation through visual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3064–3074
34. Wang L, Wang Y, Liang Z, Lin Z, Yang J, An W, Guo Y (2019) Learning parallax attention for stereo image Super-Resolution. Inproceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Long Beach, pp 16–20
35. Wang Q, Zhang L, Bertinetto L, Hu W, Torr PHS (2019) Fast online object tracking and segmentation: A unifying approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1328–1338
36. Wolf CT (2016) DIY videos on YouTube: Identity and possibility in the age of algorithms. First Monday
37. Wu Z, Shen C, Hengel Avd (2016) Bridging category level and instance-level semantic image segmentation. arXiv:1605.06885
38. Wug O, Seoung J-YL, Sunkavalli K, Kim SJ (2018) Fast video object segmentation by reference-guided mask propagation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7376–7385
39. Xiao H, Feng J, Lin G, Yu L, Zhang M (2018) Monet: Deep motion exploitation for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1140–1148
40. Xu B, Wang N, Chen T, Li M (2015) Empirical evaluation of rectified activations in convolutional network. The International Conference on Machine Learning, arXiv:1505.00853 (8 January 2019)
41. Xu N, Yang L, Fan Y, Yang J, Yue D, Liang Y, Price B, Cohen S, Huang T (2018) Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 585–601
42. Yang Z, Wang Q, Bertinetto L, Hu W, Bai S, Torr PHS (2019) Anchor diffusion for unsupervised video object segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 931–940
43. Yang L, Wang Y, Xiong X, Yang J, Aggelos K (2018) Katsaggelos: Efficient video object segmentation via network modulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6499–6507
44. Yu H, He F, Pan Y (2019) A novel segmentation model for medical images with intensity inhomogeneity based on adaptive perturbation. Multimed Tools Appl 78(9):11779–11798
45. Zhao A, Balakrishnan G, Durand F, Guttag JV, Dalca AV (2019) Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8543–8553
46. Zhu X, Dai J, Zhu X, Wei Y, Yuan L (2018) Towards high performance video object detection for mobiles. arXiv:1804.05830
47. Zhu X, Wang Y, Dai J, Yuan L, Wei Y (2017) Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp 408–417. Wang, Qiang, 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 computer vision and pattern recognition, pp 1328–1338 (2019)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.