The 2020 DAVIS Challenge on Video Object Segmentation - CVPR 2020 Workshops

Unsupervised Video Object Segmentation using Online Mask Selection and Space-time Memory Networks

Shubhika Garg Vidit Goel Somesh Kumar
Indian Institute of Technology, Kharagpur, West Bengal, India
shubhikagarg97@gmail.com, gvidit98@gmail.com, smsh@maths.iitkgp.ac.in

Abstract

In this work we present an approach for Unsupervised Video Object Segmentation. We build on a semi supervised method, Video Object Segmentation Using Space-Time Memory Networks(STM)[11], for unsupervised scenarios. STM stores some of the previous frames and masks as memory and uses that as temporal knowledge to predict the masks in the current frame. There is a high possibility that all the objects are not detected in the first frame hence we modify the method to add and track the newly added objects. We noticed that even if an object gets detected in a frame the mask quality of propagated objects degrades significantly sometimes. As the output of STM depends on previous masks, once the mask quality degrades it is very difficult to recover good masks. Hence, we also improve masks in an online manner. We propose a novel selection criterion, SelectorNet which evaluates the quality of masks. We evaluated our method on DAVIS 2019 Unsupervised challenge dataset and achieved the state of the art performance with $\mathcal{J}\&\mathcal{F}$ mean 61.6%.

1. Introduction

Video understanding is an important task in computer vision community with applications ranging from from autonomous surveillance systems, object tracking to sports. In this work we target the problem of unsupervised video object segmentation. In unsupervised scenario there is no information about the objects that needs to be tracked unlike the semi supervised case in which the annotations for first frame are given.

Most of the work in video object segmentation(VOS) is done in semi supervised setting[8, 19]. Unsupervised scenario is much more difficult than semi supervised scenario. There is no concrete definition of objects which need to be tracked, the objects that need to be tracked are the ones which are primary subjects in a video or in other words which are most likely to get human attention. Due to this uncertainty some extra objects also get tracked which complicates the problem further by making tracking and identification of objects more complex. Along with this the problems in VOS like occlusion, change in appearance of objects, dynamic addition of objects are still there to make the overall problem complicated.

In this work we build upon an semi-supervised method, STM [11] to make it work in an unsupervised scenario. We add and track objects dynamically. We only use Mask R-CNN[4] as an extra source of information, we also have a provision to improve mask in an online manner by using two independent selection criteria one based on neural network and other based on contour properties.

2. Related Work

2.1. Semi Supervised Video Object Segmentation

In semi-supervised setting the annotations of first frame are given and they need to segment and track those objects throughout the video. Many approaches in this field deal with online learning and fine tuning [8, 16, 19] using the ground truth information from the first frame. These types of methods show impressive performance, however they are not suitable for real time video object segmentation as they take a lot of time. Some other variations of method involve mask propagation [12, 6] using previous frames and involve feature matching [15, 11, 10] of the embedding of the current frame with the stored templates. STM [11] is a feature matching based method that stores information from all the previous frames in the memory and uses that for current frame mask prediction. Also it is fast and requires no online learning making it suitable for adaption to unsupervised scenarios.

2.2. Unsupervised Video Object Segmentation

Most of the previous work on unsupervised video object segmentation [20, 6, 7, 17] targets foreground object segmentation and outputs a single mask for all the objects. These methods cannot be directly integrated with multi object scenarios as they don't explicitly deal with

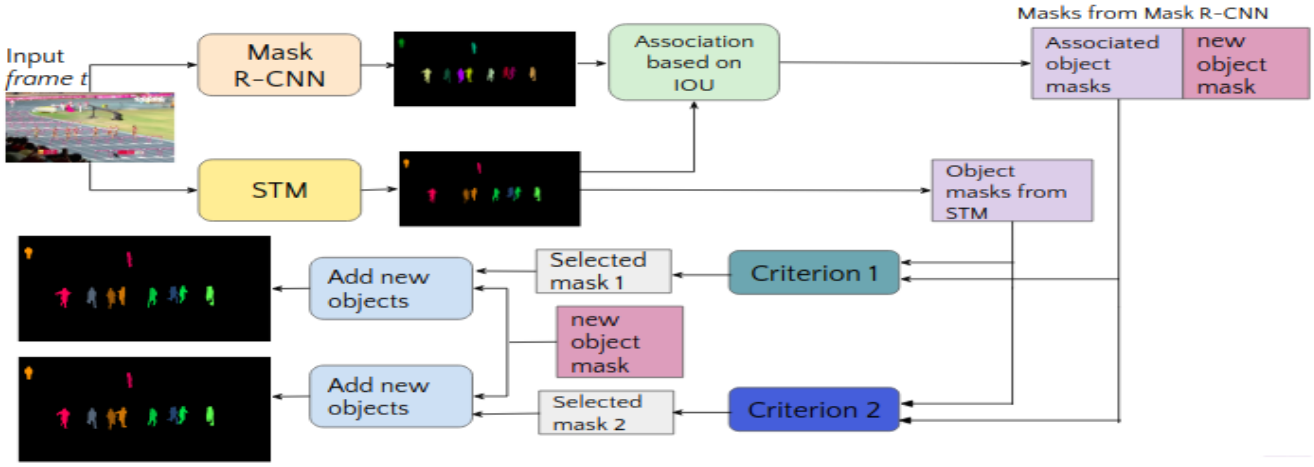

Figure 1. Block diagram of Temporal propagation and online selection of masks and addition of new objects(Stage 2)

some of the major problems like tracking, handling occlusion and re-identification of objects. The area of Unsupervised multi object segmentation and tracking of objects is very recent and not a lot of work has been done in this field. Some methods like [3, 18], focus only on moving foreground objects, which might not always be the case in a generalised scenario. These [9, 21, 2, 14] are some previous works that tackle the same problem statement as ours. UnVOST[9] proposes to run Mask RCNN[4] on every frame to find object masks and no use of temporal information is used for predicting masks. VSD[21] propagates masks for each object independently using a single object mask tracker and replaces the mask with Mask RCNN[4] mask to avoid drift. The closest method to ours is KIS[2], which uses a semi supervised method, RGMP[10] to propagate masks.The problems with these methods is that choosing only Mask RCNN[4] masks results in inaccurate masks for blur and fast moving objects and relying only on propagation methods lead to drift.

3. Method

Our framework consists of 3 stages. In the first stage, Mask R-CNN[4] is used to generate masks for each frame of the video. In the second stage, the generated Mask R-CNN[4] frames are used as the first frame input to initialize the STM[11]. In order to improve the mask, we parallelly employ 2 different independent criteria for a better quality mask selection between the current mask obtained from STM[11] and the corresponding Mask RCNN[4] mask for an object at each time-step. The objects in Mask R-CNN[4] which are not associated with any previous objects are added as new objects. In the 3rd stage, we select the best of the 2 previously generated masks further improving the results by recovering lost objects. In the following, each of the stages are explained in detail.

3.1. Object mask generation

In the first stage, we generate object masks using Mask R-CNN[4] for all the frames in the video. Mask R-CNN[4] outputs segmentation mask, object category, confidence score and bounding box for all the detected objects in a frame. For every frame, we chose at most 10 masks ranked using their confidence score and masks having confidence score below 0.1 threshold are removed.

3.2. Temporal propagation and online selection of masks and addition of new objects

It is not correct to rely only on Mask R-CNN[4] masks for every frame as it produces poor quality masks in cases where there is fast motion or blur present in the videos. Also the temporal information is completely ignored as Mask R-CNN[4] operates frame wise. In stage 2, we use STM[11], a semi-supervised video object segmentation method. Using STM[11] gives us two major benefits, first one is using temporal information to predict masks and second helps in tracking the objects. We initialize it using the first frame annotation obtained in the previous stage. In STM[11] the first frame with its given annotations and some intermediate frames with predicted annotations are stored as memory frames. The memory frames along with the previous frame are then used to predict instance mask of current frame. Let M_t and S_t be the set of masks produced by Mask R-CNN[4] and STM[11] for frame t.

Next we need to associate the masks in M_t with S_t. To achieve that, at every frame t, a bipartite matching is done between S_t and M_t using IOU based cost matrix. We do an optimal assignment using Hungarian algorithm. Object masks in M_t having a IOU higher than 0.5 are associated

to corresponding object masks in S_t and the rest of the objects are considered new objects. Now for every associated object we have two mask proposal one from STM[11] and other from Mask R-CNN[4]. We use two independent selection criteria to select the better of the two masks.

The *criterion 1* is a neural network whose task is to compare two masks and assign scores based on quality of masks. The architecture consists of feature extractor backbone followed by fully connected(FC) layers to further process the features and regress the scores. The features of two masks are extracted independently, the extracted features are then concatenated and passed through FC layers. The output of network has two heads each giving a score between 0 to 1 which indicated the quality of mask, where 0 is poor quality mask and 1 is good quality mask. The mask whose score is higher is chosen for propagation. For feature extractor we used ResNet-18[5]. The extracted features were flattened resulting in 1024 sized vector after concatenation. It is followed by 2 FC layers, with outputs of size 512 and 2. The input to the network is a four channel image(binary mask + RGB image).We used training data from DAVIS 2017 dataset [13] to train the network. Two samples of each object mask were generated one from Mask R-CNN[4] and other from STM[11]. The mask whose IOU with ground truth is greater is labeled as 1 and other mask as 0. We were able to achieve 82% accuracy on held out data. We name this network as Selector Net.

For *criterion 2* we compare the area of the object mask in frame t to the corresponding object mask in frame $t-1$. We chose the mask whose change in area is less.

Using the above two features stage 2 results in 2 mask frames resulting from two independent criterion. The complete pipeline of stage 2 is shown in Fig 1.

3.3. Offline selection of masks

After completion of stage 2 for the whole video, we then use the ensemble of the 2 previously generated results to capture the best of the 2 results. Selector Net is used to chose the better mask in this stage. Since we are dealing with an unsupervised scenario and there can be situations, where one criteria might chose the wrong mask, leading to incorrect propagation ahead. Hence, we initially use 2 independent criteria and then combine them in this stage to produce more accurate results.

4. Evaluation

We evaluate our algorithm on DAVIS 19 Unsupervised Challenge dataset[1]. The competition consisted of 2 phases, which are Test Development phase followed by Test Challenge phase. Each phase contained 30 videos and no other information regarding the objects or the masks was given. All of these videos contained multiple objects that needed to be segmented and tracked. Table 1 and table 2 show the performance of our algorithm on the test challenge and test-dev dataset respectively. The ranking for the competition is based on the test challenge dataset. Our algorithm outperforms previous state of the art algorithm by a large margin of 5.2% resulting in $\mathcal{J}\&\mathcal{F}$ mean of 61.6% and achieving first place in the competition. In table 2 we fall short by a small margin, which shows that our algorithm performs well on both the datasets. Fig. 2 shows the qual-

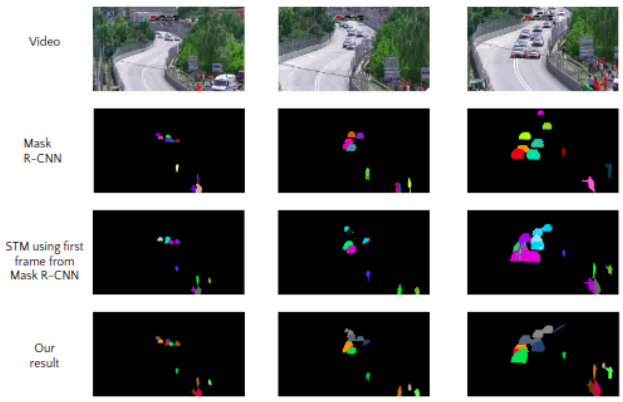

Figure 2. Comparison of our output(fourth row) with only Mask R-CNN(second row), STM using first frame from Mask R-CNN(third row) also referred as vanilla STM.

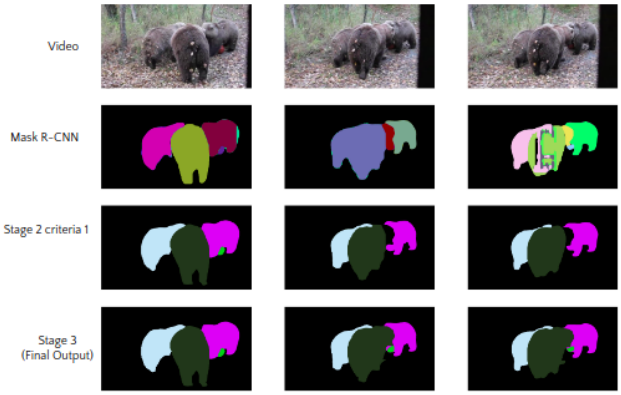

Figure 3. In this figure we show the output at different stages. The last row shows the final output which recovers the ball due to ensemble of criteria.

Table 1. Results in the test-challenge phase.

Team/Method	J & F Mean	J Mean	F Mean
Ours	**61.6**	**58.4**	**64.7**
UnVOST[9]	56.4	53.4	59.4
VSD[21]	56.2	53.5	59.0
IIAI	55.6	53.1	58.2
BLIIT	52.3	50.2	54.4
KIS[2]	51.6	48.7	54.5

Table 2. Results in the test-dev phase

Team/Method	J & F Mean	J Mean	F Mean
Ours	57.9	52.9	63.0
UnVOST[9]	58.0	54.0	62.0
VSD [21]	56.5	51.7	61.4
IIAI	**59.8**	**56.0**	**63.7**
BLIIT	51,4	51,4	57.4
KIS[2]	50.0	50.0	58.3

itative result of our algorithm on a sequence from test-dev dataset. It shows a comparison of our algorithm with the results obtained from vanilla STM[11] in row 3. The corresponding Mask R-CNN frame are shown in row 2. The performance of vanilla STM[11] degrades only after a few frame as the cars are very small and have similar visual features. Due to online selection we are able to obtain better results than vanilla STM[11]. Fig. 3 shows the result on a test challenge sequence where the Mask R-CNN[4] masks shown in row 2 are very noisy. This example shows the robustness of the Selector Net in row 3 as those noisy masks are not chosen by it. Also it shows the practicality of using stage 3 as it helps in recovering lost objects which are missed by one of the two criteria.

5. Conclusion

In this work we built upon a semi-supervised method for VOS to make it work in an unsupervised scenario. We add, track and improve mask in a online manner. We show how a semi-supervised method can be intelligently modified to solve the task in hand. We achieved the state of the art results in challenge data set and were winner of Davis challenge 2020 in unsupervised track. There are various scopes of improvements in the method like we can train the whole pipeline in an end to end manner. We leave this as our future work.

References

[1] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019. 3

[2] Donghyeon Cho, Sungeun Hong, Sungil Kang, and Jiwon Kim. Key instance selection for unsupervised video object segmentation. *arXiv preprint arXiv:1906.07851*, 2019. 2, 3, 4

[3] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting everything that moves. *arXiv preprint arXiv:1902.03715*, 2019. 2

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2, 3, 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[6] Siyang Li, Bryan Seybold, Alexey Vorobyov, Alireza Fathi, Qin Huang, and C-C Jay Kuo. Instance embedding transfer to unsupervised video object segmentation. In *CVPR*, 2018. 1

[7] 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 *CVPR*, 2019. 1

[8] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018. 1

[9] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. Unovost: Unsupervised offline video object segmentation and tracking. *arXiv preprint arXiv:2001.05425*, 2020. 2, 3, 4

[10] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, 2018. 1, 2

[11] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 1, 2, 3, 4

[12] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017. 1

[13] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 3

[14] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *CVPR*, 2019. 2

[15] 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 *CVPR*, 2019. 1

[16] Bofei Wang, Chengjian Zheng, Ning Wang, Shunfei Wang, Xiaofeng Zhang, Shaoli Liu, Si Gao, Kaidi Lu, Diankai Zhang, Lin Shen, et al. Object-based spatial similarity for semi-supervised video object segmentation. In *The 2019 DAVIS Challenge on Video Object Segmentation-CVPR Workshops*, 2019. 1

[17] 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 *CVPR*, 2019. 1

[18] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *CVPR*, 2019. 2

[19] Shuangjie Xu, Daizong Liu, Linchao Bao, Wei Liu, and Pan Zhou. Mhp-vos: Multiple hypotheses propagation for video object segmentation. In *CVPR*, 2019. 1

[20] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *CVPR*, 2019. 1

[21] Zhao Yang, Qiang Wang, Song Bai, Weiming Hu, and Philip HS Torr. Video segmentation by detection for the 2019 unsupervised davis challenge'. 2019. 2, 3, 4