# Mask Selection and Propagation for Unsupervised Video Object Segmentation

Shubhika Garg Vidit Goel
Indian Institute of Technology, Kharagpur, West Bengal, India
shubhikagarg97@gmail.com, gvidit98@gmail.com

## Abstract

*In this work we present a novel approach for Unsupervised Video Object Segmentation, that is automatically generating instance level segmentation masks for salient objects and tracking them in a video. We efficiently handle problems present in existing methods such as drift while temporal propagation, tracking and addition of new objects. To this end, we propose a novel idea of improving masks in an online manner using ensemble of criteria whose task is to inspect the quality of masks. We introduce a novel idea of assessing mask quality using a neural network called Selector Net. The proposed network is trained is such way that it is generalizes across various datasets. Our proposed method is able to limit the noise accumulated along the video, giving state of the art result on Davis 2019 Unsupervised challenge dataset with $\mathcal{J}\&\mathcal{F}$ mean 61.6%. We also tested on datasets such as FBMS and SegTrack V2 and performed better or on par compared to the other methods.*

## 1. Introduction

Video understanding has gained a lot of attention in recent years. This work focuses on unsupervised video object segmentation. In the unsupervised setting[1] there is no prior information given about the objects that need to be segmented and tracked unlike the semi-supervised scenario in which annotations are given for the first frame. The objects of interest are the ones which are likely to catch human attention[51]. Due to this loose definition, the task becomes even more challenging.

With the advent of deep learning, almost all the methods proposed recently are learning based. Though there are some classical methods which are used in deep learning pipeline such as [2, 40]. A lot of the prior work in multi-object video segmentation is done in semi-supervised setting [29, 47]. As in semi supervised setting, the masks for first frame are given, the algorithms learn good feature representation of the given objects, so that they can be used to find and track objects in further frames[36, 46]. Hence, these methods try to tackle problems such as occlusion, change in appearance of object as the video proceeds while trying to find and associate a given object. In unsupervised scenario the problem becomes even harder as the number of objects are not decided, hence, some extra objects also get detected by the algorithm. The extra objects add a lot of noise which makes it even harder to associate and track objects.

Some of the earlier works done in unsupervised video object segmentation is for single object in a video[56, 22]. These methods tried to extract foreground objects using some property which differentiates it from background. This can not work in multi object setting as we also need to differentiate between objects. Motivated by [22] we tried to learn embedding of objects but we found that embedding are not consistent across the frames hence making it difficult to track objects. Further, embedding perform very poorly when objects are small or when there are similar objects in a frame.

One of the major difference in semi-supervised scenario compared to unsupervised scenario, is the ground truth information in the first frame. So if we are able to get good annotations of first frame in some manner then it will reduce the problem to semi-supervised setting. There are many works which target object detection and segmentation[14, 6] but the problem is that the quality of masks generated is not at par with ground truth annotations. Keeping this in mind, we target for an algorithm which can improve masks and reduce noise propagation in an online manner. We also aimed to use ensemble of masks and then propagate only the best mask out of the multiple masks. To this end, we propose a method which builds upon a semi-supervised method Video Object Segmentation Using Space- Time Memory Networks(STM)[36]. 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. For getting masks in the frames we use a well know method Mask R-CNN[14]. We create an ensemble from Mask R-CNN and STM. Further, we propose a novel

[1] In this work, unlike the conventional definition, unsupervised setting refers to the task of automatically segmenting and tracking the salient objects in a video sequence without any external information about them.

selection criterion, Selector Net. The network takes input as 2 masks, and returns the relative quality scores of the masks. For any given object in a frame we get masks from STM and Mask R-CNN, then we use Selector Net and another selection criterion based on change in object shape in consecutive frames to select the best mask which is then propagated further (section 3). The only trainable component in our method is Selector Net making it highly efficient in training. The proposed method has a general structure to solve unsupervised video object segmentation problem rather than a fixed algorithm. In this work we used STM as it was the state of the art method in semi-supervised setting which uses temporal information, but in future as semi-supervised algorithms improve our accuracy should also improve.
To summarize our contributions are the following:

- We propose a novel generalizable noise resilient and modular pipeline for unsupervised video object segmentation and tracking, outperforming existing state of the art methods[2].
- Along with this, we introduce a novel idea of assessing mask quality using a neural network. The network is trained only on one dataset and we demonstrate its generalizability across different datasets.
- We evaluate our algorithm on 3 benchmark datasets for unsupervised video object segmentation and demonstrate that it can robustly handle complex scenarios with occlusions and re-identification, complex deformation, motion blur, multiple objects with similar appearance and efficiently deal with drift in long temporal propagation by online mask improvement.

## 2. Related work

### 2.1. Semi supervised video object segmentation

In semi supervised video object segmentation, we are given with the first frame ground truth annotations in the form of the masks of objects that need to be tracked throughout the video. Hence, in this we have a clear idea of the objects that need to be tracked unlike the unsupervised scenario. While there has been significant progress in this field, however, a lot of approaches[20, 37, 29, 4, 47, 55, 54] rely on online learning and fine tuning. These approaches fine tune on an augmented dataset created using the first frame annotations for every video. While they are able to achieve high accuracy using such techniques, they are not suitable for real time methods and are very slow.

Another category of these works include propagation techniques[37, 24, 20, 48] in which the segmented masks from the previous frames are propagated to the next frame using optical flow as motion cues. Such methods have an extra dependence on the optical flow methods which aren't always accurate especially in homogeneous regions and when the movement between 2 frames is very less.

Another category of these works are the memory based networks[36, 52, 46] that use temporal information by storing feature embedding of the previous frames and then do a matching of the features of the current frame with those of the stored templates. Instead of using only the previous frame, they store all the temporal information from the past as key and value vectors and a new frame is like a query vector. This query vector is then matched with the key vectors to find the results of the current frame. STM[36] is a current state of the state of art method that works on the above principle. Also it is fast, does not depend on optical flow and has a high accuracy for semi supervised video object segmentation without requiring any fine tuning and online learning. These factors make it suitable to be adapted for unsupervised video object segmentation.

### 2.2. Unsupervised video object segmentation

In unsupervised video object segmentation, there is no fixed definition of the objects that need to be segmented and tracked throughout the video. The most early works deal with foreground and background extraction in a video. The definition of the objects in these works are the most salient objects present in a video. Some of the traditional work in the area in include detecting foreground object using background subtraction[33, 11] and generating object proposals[2, 19, 60, 13, 21]. Some other weakly supervised methods include segmenting foreground object using markers[40, 3, 31]. The above methods are not robust enough to handle even a slight change in lighting conditions and are sensitive to shadows. With the rise of the deep learning era, a lot of approaches[22, 27, 59, 9, 51, 42, 58] have used deep learning methods to do the above the task. Davis 2016[38] is a common dataset that is used for such task. The above algorithms output a single binary mask for all the foreground objects, and hence, do not deal with multi-foreground object scenarios. They cannot be directly integrated with the multi object segmentation and tracking as these techniques do not have deal with some of the major problems like tracking, handling occlusion and re-identification of objects.

Another area of unsupervised video object segmentation deals with explicitly extracting moving objects as foreground objects. [1, 61, 56, 63, 26, 50] are example of such methods for single foreground mask prediction and [10] is an example of that deals multi moving foreground object segmentation and tracking. These approaches cannot be directly used, as aside from single foreground mask prediction, they focus only on moving foreground objects, which

[2]The code is available at: https://github.com/vidit98/FrameSelect

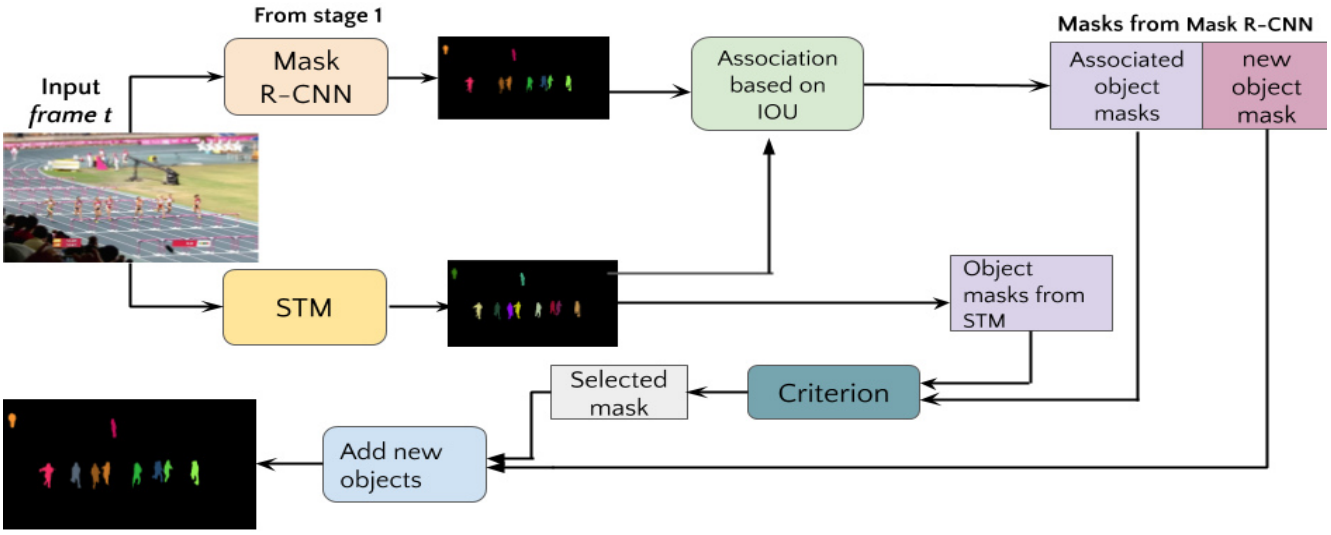


Figure 1: Block diagram of stage 2 of our algorithm. Here criterion can be either *criterion 1* or *criterion 2*. Given an input frame, two sets of mask are generated using Mask R-CNN and STM. It is followed by mask association and identifying new objects detected by Mask R-CNN. The associated pairs are sent to the criterion and best mask is selected to be propagated further. The same pipeline is followed for the two independent criteria proposed in this work. The only difference is that for one pipeline *criterion 1* is used and for second pipeline *criterion 2* is used.

might not always be the case in a generalised scenario. Also they depend on optical flow for providing motion cues, however with time the error in it accumulates. This results in inaccurate tracking due to large drift.

The area of unsupervised multi object segmentation and tracking in a video is relatively new. These[8, 30, 62, 57, 53, 45, 49, 28] are some of the works that work on the above problem. The problem was first proposed at DAVIS Unsupervised Challenge 2019[5]. In AGNN[49], a binary foreground object segmentation method is converted to multi object setting by using Mask RCNN masks to get instance level salient object masks. In UnOVOST[30] object mask proposals are only taken from Mask R-CNN frame wise and temporal information is not used for mask prediction. In VSD[57] each object is independently tracked using Siam Mask[48] and at each time step the tracked mask is replaced by the Mask R-CNN mask to avoid drift. The above process helps in handling drift in mask propagation, however, replacing the propagated mask by associated Mask R-CNN mask leads to propagation of inaccuracies of Mask R-CNN ahead and makes no use use of temporal information for mask prediction. KIS[8] is a method that is closest to our method. They use RGMP[52] for mask propagation, however only doing propagation leads to drift and accumulation of error with time. On the other hand, we handle this by using a Selector Net to chose the good masks between the propagated mask and the associated from Mask R-CNN. Further details are mentioned in Section 3.

# 3. Approach

## 3.1. Problem formulation

We address the problem of unsupervised video object segmentation with an aim to segment and track at least the objects that capture human attention. Hence, given an input of frames $[f_0, f_1, \cdots, f_{T-1}]$, we produce the following output:

1. A set of segmentation masks $[m_0, m_2, ..., m_{T-1}]$ containing non overlapping segmentation mask proposals of *N* objects. The number of objects, *N*, is not known in advance and has a max limit of 20 objects.

2. The objects are tracked throughout the video and every object is supposed to have a consistent mask id in the mask proposals generated throughout the whole video.

## 3.2. Method

Our method consists of 3 stages. In stage 1, Mask R-CNN[14] is used to generate masks for objects in a frame. This serves as the first source of masks. In stage 2, we initialize STM[36] by using masks generated by Mask R-CNN for the first frame. Then STM predicts the masks for the current frame using the previous frames stored as memory. In order to improve the mask, at each time-step, we parallelly employ 2 different independent criteria for a better quality mask selection between the current mask obtained from STM and the corresponding Mask RCNN mask

for every object. At the same time, the objects in Mask R-CNN which are not associated with any previous objects are added as new objects. In the 3rd stage, we chose the best of the 2 previously generated masks further improving the results by recovering lost objects.

**Object mask generation**

We used Mask R-CNN implementation by [32] trained on COCO[25] dataset with backbone ResNet-50[15] to get initial object masks. We set the threshold of 0.1 on confidence score given by Mask R-CNN. The low confidence threshold helps to segment objects beyond the categories Mask R-CNN is trained on. To limit the number of objects in a frame, we selected at max 10 objects in a frame ranked according to their confidence score[30]. We also filter out the objects which are very small and fragmented to reduce the noise further.

**Temporal propagation and online selection of masks and addition of new objects**

Unlike a lot of previous methods, we do not rely only on Mask R-CNN masks, but also use temporal information and improve masks on the go. In order to make use of temporal information, in stage 2, we use STM, a semi-supervised video object segmentation method. STM uses temporal and spatial information to generate masks in a current frame. As the algorithm progresses through each frame in the video, the first frame with its given annotations and some intermediate frames with predicted annotations are stored as memory frames. These memory frames along with the previous frame annotations are then used to predict instance mask of current frame. In our situation we initialize STM using the first frame mask annotations obtained in the previous step.

Using STM gives us two major benefits, first one is using temporal information to predict masks and second it helps in tracking the objects. Hence, we have a complete pipeline that handles both segmentation and tracking. However, only using STM would not suffice. This is because unlike semi supervised scenario, number of objects that need to be segmented and tracked are not fixed before hand and unsupervised scenario deals with insertion of new unknown objects in the middle of the sequence. Moreover since the annotations of the first frame are noisy compared to the ground truth annotations, the quality of masks degrades as we go progress through the video. In order to deal with the above problems, we insert modules to handle additions and online selection of masks to minimize noise propagation in the pipeline resulting in better results.

Let $M_t$ and $S_t$ be the set of masks produced by Mask R-CNN and STM for frame $t$. The two sets of masks are passed to association module (Fig. 1), where a bipartite matching is done between the object masks present in both the sets. In order to achieve this, we frame it as a optimal assignment problem. A 2D matrix is formed whose rows and columns are the objects present in $M_t$ and $S_t$ respectively and $v_{ij}$ is the IOU between the $i^{th}$ object mask in $M_t$ and $j^{th}$ object mask in $S_t$. The assignment is done 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 added in the memory as new objects. In order to limit the noise we only allow fixed number of objects to get added in the complete video sequence. Further, we only add objects whose intersection with other objects is below certain threshold compared to the area of the mask of object being added. Now, for every associated object we have two mask proposal one from STM and other from Mask R-CNN. We use two independent selection criteria to select the better of the two masks. Hence, there are two independent branches running of the above described algorithm. The better mask selected by the criterion is propagated further independently in its own branch. The difference between the two branches is the way mask frames are selected which further results in different memory frames in STM for the two branches. The complete process is explained in Algorithm 1 and block diagram for same is shown in Fig 1.

The *criterion 1* is a neural network whose task is to compare two associated masks and assign scores signifying the quality of mask. Further details for this can be found in Section 3.3. For *criterion 2* we compare the area of the object masks in frame $t$ to the corresponding object mask in frame $t-1$. We chose the mask whose change in area is less. The logic behind this criterion is that the object position and orientation does not change much between two consecutive frames, hence the mask area should also remain consistent between the frames. Using the above two independent criteria, stage 2 results in 2 mask frames for each frame. One set of masks are generated using criterion 1 as selection criteria and second set is generated using criterion 2 as selection criterion.

**Offline selection of masks**

After completion of stage 2 for the whole video, we select best masks out of the 2 generated results. Selector Net is used to chose the better mask in this stage. The objects which are present in only one of the two results is simply added as new mask. This is done because there can be situations, where one criterion might chose the wrong mask leading to incorrect propagation ahead.

### 3.3. Selector net

We propose a novel selection criterion called Selector Net. It is a neural network based approach to select the better mask from two input mask. The reason for using a neu-

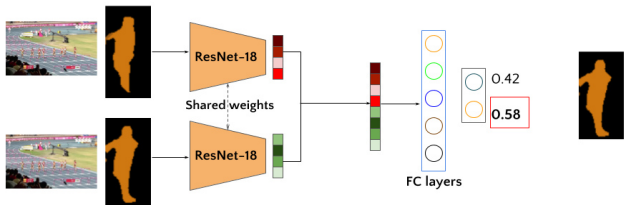


Figure 2: Architecture for the Selector Net: The 2 pairs of inputs are the original RGB image along with the corresponding Mask RCNN mask of the object and the original RGB image along with STM propagated mask. They are fed to a feature extraction module having shared weights. The features are then concatenated and fed to a network to output the mask score. The mask having a higher score is considered a better mask and propagated ahead.

ral network based approach is that there are many factors on which the quality of mask depends such as smoothness of mask, the object inside the mask, semantic consistency captured by mask. It is difficult to capture all these properties using classical formulations hence we came with a learning based technique.

The Selector Net consists of a feature extractor backbone ResNet-18[15] which is followed by fully connected(FC) layers (Fig 2). The intuition here is that feature extractor encodes the masks into feature space which captures relevant factors on which quality of mask should depend which are further processed by FC layers to make the decision. There are two inputs to the network as shown in Fig 2. Each input consists of binary mask of an object that needs to be compared along with the complete RGB image (for visual purposes we have shown colored mask). The binary mask is of the same spatial dimension that of the corresponding RGB image. The image and mask are concatenated hence making a 4 channel input. The two inputs are passed through feature extractor resulting in 1024 dimension vector after flattening and concatenating the feature vectors. This feature vector is then passed through 2 fully connected layers of output size 512 and 2. We also added a dropout layer after first FC layer with drop out probability of 0.2. The final score of the two masks is produced by passing the outputs through a softmax layer, higher score signifies better quality of mask. We used mean squared error loss for training( Eq 1).

$$L = \frac{1}{N}\sum_{n}[(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2] \tag{1}$$

Here $y_1$, $y_2$ are the ground truths of the input pair and $\hat{y}_1$, $\hat{y}_2$ are the predicted values by the network.

For training the Selector Net, we generated the dataset using training data from DAVIS 2017[39]. The STM is initialized using masks generated from Mask R-CNN. The association between the masks generated using Mask R-CNN and STM is done as explained in section 3.2. While creating the training data, the mask between STM and Mask RCNN, having higher accuracy compared to the ground truth is propagated along. During training, each object is processed independently. Let $m_t^k$ and $s_t^k$ be mask of $k^{th}$ object produced by Mask R-CNN and STM for frame at time $t$ and $g_t^k$ represent the ground truth mask for the same object. For labeling the mask pairs we use the given formulation.

$$f(m_t^k, s_t^k, g_t^k) = \begin{cases} l_m = 1, & IOU(m_t^k, g_t^k) > IOU(s_t^k, g_t^k) \\ l_m = 0, & \text{otherwise} \\ l_s = 0, & IOU(m_t^k, g_t^k) > IOU(s_t^k, g_t^k) \\ l_s = 1, & \text{otherwise} \end{cases} \tag{2}$$

**Algorithm 1:** Stage 2 algorithm

**Input :** $Frames = [f_1, \cdot, f_{T-1}], N, K, crit$
**Output:** $Masks = [m_1, ..., m_{T-1}]$

```
/* N is interval at which
   information will be stored in
   memory, K is maximum number of
   objects that can be added, crit
   is a boolean variable for
   choosing the criterion.        */
memFrames ← [(f_0, m_0)];
count ← 0;
Masks ← [];
for t ← 1 to T − 1 do
    M_t' ← MaskRCNN(f_t);
    M_t ← FilterNoise(M_t');
    S_t ← STM(f_t, memFrames);
    A_t ← Associate(S_t, M_t);
    N_t ← M_t \ A_t        // New Masks;
    F_t = []               // Final Masks;
    for (a_t^m, a_t^s) ∈ A_t do
        if crit == 0 then
            s ← SelectorNet(a_t^m, a_t^s)
        end if
        else
            s ← AreaCrit(a_t^m, a_t^s, a_{t-1}^s)
        end if
        F_t ← F_t + s;
    end for
    F_t, count ← AddObj(F_t, N_t, K, count);
    Masks ← F_t;
    if t%N == 0 then
        memFrames ← UpdateMem(F_t, f_t)
    end if
end for
return Masks;
```

| | | | Ours | TAAT[62] | UnOVOST[30] | VSD[57] | GTM[53] | KIS[8] | SiVOS[34] | RVOS[45] |
|---|---|---|---|---|---|---|---|---|---|---|
| Test C | J & F | Mean | **61.6** | 55.6 | 56.4 | 56.2 | 52.3 | 51.6 | 43.9 | - |
| | J | Mean | **58.4** | 53.1 | 53.4 | 53.5 | 50.2 | 48.7 | 40.2 | - |
| | | Recall | **65.0** | 60.0 | 60.9 | 61.3 | 57.5 | 55.1 | 45.7 | - |
| | | Decay | -1.6 | -0.5 | 1.5 | -2.1 | **-5.0** | 4.0 | -0.6 | - |
| | F | Mean | **64.7** | 58.2 | 59.4 | 59.0 | 54.4 | 54.5 | 47.5 | - |
| | | Recall | **71.1** | 62.5 | 64.1 | 63.2 | 58.9 | 59.4 | 50.1 | - |
| | | Decay | 0.5 | 1.6 | 5.8 | 0.1 | **-2.5** | 7.7 | 4.0 | - |
| Test D | J & F | Mean | 57.9 | **59.8** | 58.0 | 56.5 | 54.4 | 54.2 | - | 22.5 |
| | J | Mean | 52.9 | **56.0** | 54.0 | 51.7 | 51.4 | 50.0 | - | 17.7 |
| | | Recall | 60.4 | **65.1** | 62.9 | 59.9 | 59.9 | 58.9 | - | 16.2 |
| | | Decay | 16.7 | 7.8 | 3.5 | 21.7 | **-1.0** | 8.4 | - | 1.6 |
| | F | Mean | 63.0 | **63.7** | 62.0 | 61.4 | 57.4 | 58.3 | - | 27.3 |
| | | Recall | **69.5** | 68.4 | 66.6 | 65.7 | 61.6 | 62.1 | - | 24.8 |
| | | Decay | 20.5 | 11 | 6.6 | 15.7 | **0** | 11.4 | - | 1.8 |

Table 1: The quantitative results on DAVIS 2019 Test Challenge(Test C) and Test Dev(Test D) dataset.

Labels are generated using $f(m_t^k, s_t^k, g_t^k)$ (Eq 2). This is done for every object in each frame. This completes the process of generating the training data for Selector Net.

# 4. Experiments

## 4.1. Training details

We trained the Selector Net using the data generated from DAVIS 2017[39] as explained in previous section. Training was done on RTX 2080 GPU card. We used a batch size of 64, learning rate of 1e-4 and trained using Adam optimizer. The resulting binary classification accuracy on the held out dataset was 83%. We believe accuracy could further be improved using techniques such as data augmentation and soft labeling. In order to prevent over-fitting we randomly switched the order of input masks i.e. some times Mask R-CNN was the first mask and some times the mask generated from STM was first. We did this because if all the masks from STM would have been better and in all training samples, masks from STM were sent first then network would simply always output high score for the first mask without learning anything. Selector Net was only trained once on DAVIS 2017[39] dataset and directly used on other datasets which shows the generalizability of the proposed network. Evaluation is done using $J\&F$ metric, which is the mean of region based similarity ($J$) and contour accuracy ($F$) [38, 5]. We used the evaluation code provided by DAVIS unsupervised challenge [41] and do not penalize extra detected objects as per the evaluation criteria [5].

## 4.2. DAVIS 2019

DAVIS[5] is a dataset for unsupervised video segmentation and tracking of multiple objects. DAVIS[4] consist of 60 videos for training and validation. DAVIS provides 2 sets of data for the purpose of testing. The 2 sets are test-dev and test-challenge and each of the sets contain 30 videos. The evaluation is done through a Codalab server.

Table 1 shows the performance of our algorithm on the DAVIS test challenge and test-dev dataset respectively. In the test challenge dataset, our algorithm outperforms previous state of the art algorithm by a large margin of 5.2% resulting in $J\&F$ mean of 61.6%. We attribute this performance to the online selection of masks done using different criteria. This online selection helps to reduce drift in masks with temporal propagation and increases the accuracy. In the Test Dev dataset, we fall short only by a small margin, which shows that our algorithm can generalize well and performs well on both the datasets.

## 4.3. FBMS

FBMS[35] is also a multi object segmentation and tracking dataset that consist of 59 video sequences. However, unlike DAVIS, the objects annotated are classified as the moving objects only. Also, instead of annotating every frame, annotations are provided only for a subset of the frames and hence is sparsely labeled. We do not use FBMS data for training and only use the 30 test sequences for evaluation.

The results for FBMS dataset are presented in Table 2. All the algorithms except ours, does single binary mask prediction for salient objects in every frame and does not deal with a lot of challenges in tracking like reappearance, occlusion and arrival of new objects. Despite all of these challenges, we can see that our algorithm outperforms a lot of other algorithms and has a comparable performance with others. Also since the dataset contains videos as long as 800 frames, which is more than 8 time the ones in DAVIS, we

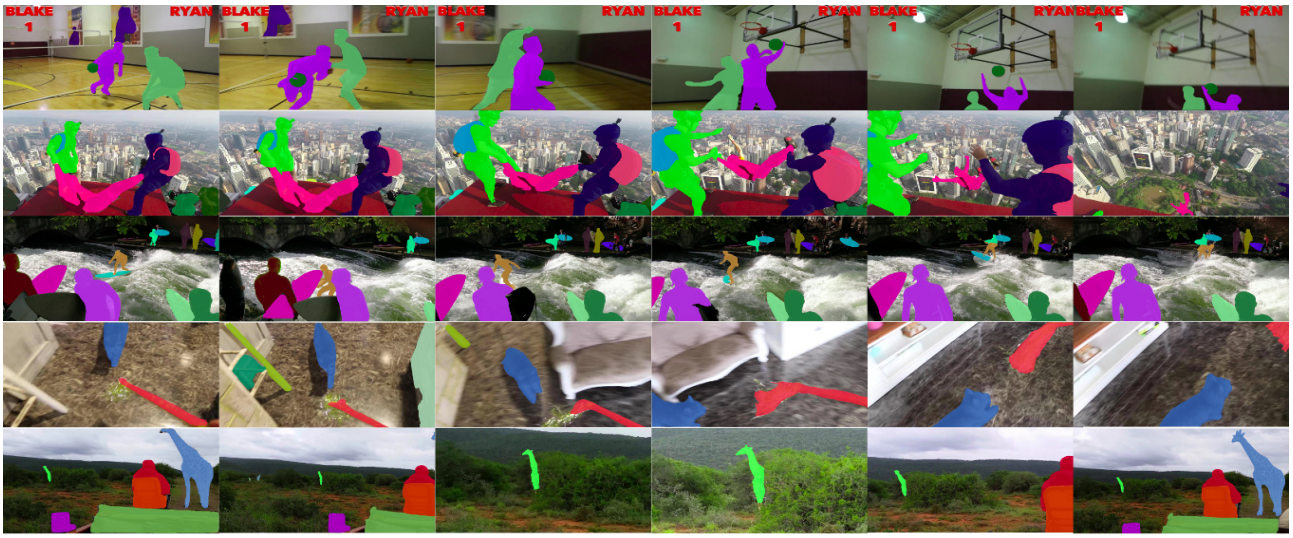

Figure 3: Qualitative results on DAVIS 19 Test Challenge and Test Dev set

| | SFL[7] | MSTP[17] | FSG[18] | IET[22] |
|---|---|---|---|---|
| MO | ✗ | ✗ | ✗ | ✗ |
| J Mean | 56.0 | 60.8 | 68.4 | 71.9 |
| | MBNM[23] | PDB[42] | COSNet[27] | Ours |
| MO | ✗ | ✗ | ✗ | ✓ |
| J Mean | 73.9 | 74.0 | **75.6** | 67.0 |

Table 2: The quantitative results on FBMS dataset. MO signifies whether the method is for multi object or single object.

| | LVO[43] | LSMO[44] | IET[22] | FSG[18] |
|---|---|---|---|---|
| MO | ✗ | ✗ | ✗ | ✗ |
| J Mean | 57.3 | 59.1 | 59.3 | 61.0 |
| | NLC[12] | STP[16] | EpO+[1] | Ours |
| MO | ✗ | ✗ | ✗ | ✓ |
| J Mean | 67.2 | 70.1 | 70.9 | **72.2** |

Table 3: The quantitative results on SegTrackV2 dataset. MO signifies whether the method is for multi object or single object.

can say that our algorithm is able to efficiently segment and track objects over a large temporal distance.

### 4.4. SegTrack V2

SegTrack V2[21] is also another multi object segmentation and tracking dataset. It consist of total of 14 videos with 24 objects over 947 annotated frames. This dataset also targets moving objects only.The video sequences are used only for evaluation. The dataset is very challenging as it contains videos with motion blur,appearance change, occlusion, complex deformation and interacting objects.

The comparison of different algorithms whose results were available on the above dataset is presented in Table 3. It can be seen that our algorithm outperforms the state of the art algorithms.

### 4.5. Qualitative Results

The qualitative results for DAVIS Test Dev and Test Challenge set are shown in Fig 3. It can be seen from sequences *dribbling*, *skydiving* and *surfer* that our algorithm can efficiently propagate temporal information and is able to segment and track them, even when they are partially occluded or extremely small in size. The sequence *cat* shows robustness to handle blurred picture cases due to fast moving camera and sequence *giraffes* shows the capability to efficiently track in cases of reappearance.

Fig 4 shows the qualitative results on Seg Track V2 and FBMS dataset. The sequence *monkey dog* again shows the capability to efficiently work on motion blurred images and the sequence *penguin* demonstrates the capability to robustly segment and track in scenarios with similar appearing objects and occlusion. In sequence *tennis*, we are able to efficiently propagate even small objects like tennis racket. Sequence *rabbits* shows the ability to handle scenarios where new objects are added in between the sequence.

## 5. Analysis

**Vanilla STM**. In order to analyse whether adding stage 2 and stage 3 led to any improvement we initialized STM as in stage 1 and used it for video object segmentation on DAVIS 2019 [5] dataset. The results are shown in table 4 and it can

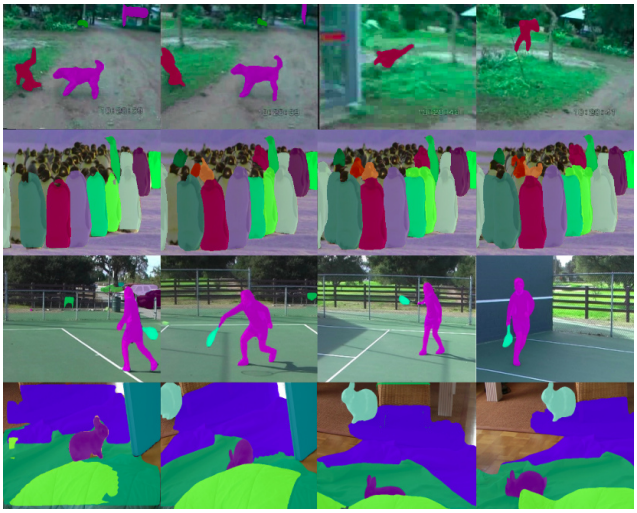

Figure 4: Qualitative results on SegTrack V2 (top 2 rows) and FBMS (bottom 2 rows).

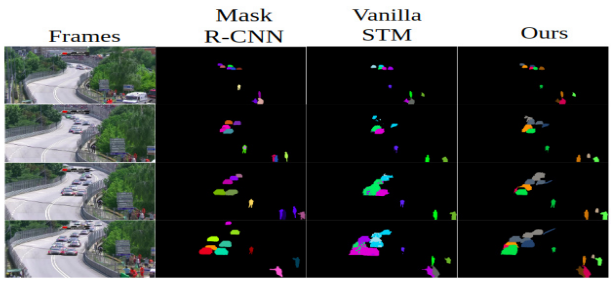


Figure 5: Comparative study of our results to vanilla STM. It can be seen that the performance of STM degrades only after a few video frames as the cars are very small and have similar visual features. Due to online selection we are able to produce better results in comparison to STM.

| Vanilla STM | Criterion 1 | Criterion 2 | Stage 3 |
|---|---|---|---|
| 0.500 | 0.564 | 0.56 | 0.579 |

Table 4: Ablation study on DAVIS 2019 test dataset

clearly seen that vanilla STM fails to perform good. The proposed method achieves $\sim 8\%$ more $J\&F$ that is 57.9% compared to STM. Qualitative results are shown in Fig. 5

**Different Stages**. In the proposed method, stage 3 is an offline stage in which masks are selected from two masks resulting from stage 2. In order to understand the importance of stage 3, we found the results after removing it. We can see there is an improvement of $\sim 2\%$ due to stage 3 (table 4). We also included stage 3 because there can be cases when one of the criterion fails, hence to achieve best of the two criteria we employed stage 3. As described in

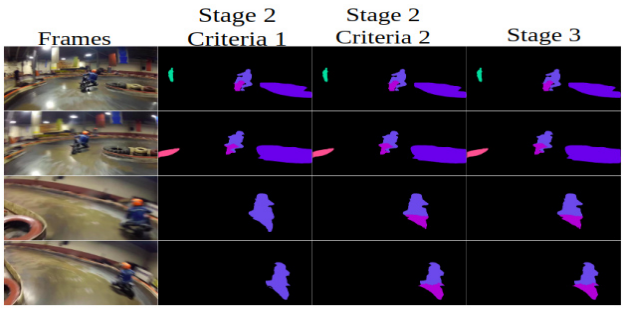


Figure 6: Results of different stages. Using them we are able to deal with recovery of objects which makes our algorithm robust to failure cases of one criterion.

| Mask R-CNN | Stage 2 C1 | Stage 2 C2 | Stage 3 |
|---|---|---|---|
| 0.09 | 0.95 | 0.48 | 0.68 |

Table 5: Runtime analysis of our algorithm during inference time. The time given is time in seconds per frame. Here *C1* stands for Criterion 1 and *C2* stands for Criterion 2

3.2 the criterion 2 calculates the change in area of an object mask compared to the predicted mask in frame $t-1$. Suppose, the mask in frame $t-1$ is very poor and only covers a small area of the actual object. Now, for the current frame consider two mask of the object resulting from Mask R-CNN and STM. If one of the mask again covers a small region of the object and the other mask covers the complete object, criterion 2 will choose the small mask which is of poor quality. Whereas our criterion 1 is a neural network and we can not always rely on it. Some of the failure cases of the criterion 1 is shown in Fig 6. The bike was not propagated when using criterion 1 and is propagated when using criterion 2.

**Timing analysis**. Table 5 presents the inference runtime analysis of our algorithm on the DAVIS Challenge dataset. The inference is done for an input image of shape $640 \times 480 \times 3$ on RTX 2080 GPU card. The processes Stage 2 Criterion 1 and Stage Criterion 2 can be done in parallel if sufficient GPU memory is available. This will lead to a total of 1.72 secs per frame. Otherwise, they can be executed serially which will lead to a time of 2.2 seconds per frame.

## 6. Conclusion

In this work we present a novel pipeline for unsupervised video object segmentation by extending semi-supervised method and achieving state of the art results. The proposed method can generalize across datasets and due to the modular structure we can replace Mask R-CNN and STM with other state of the art networks in future hence, improving the accuracy further. A promising future work could be to generate mask instead of selecting mask from available masks.

## References


[1] Ijaz Akhter, Mohsen Ali, Muhammad Faisal, and Richard Hartley. Epo-net: Exploiting geometric constraints on dense trajectories for motion saliency. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1884–1893, 2020.

[2] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014.

[3] B Basavaprasad and Ravindra S Hegadi. Improved grabcut technique for segmentation of color image. *Int. J. Comput. Appl*, 975:8887, 2014.

[4] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017.

[5] 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 preprint arXiv:1905.00737*, 2019.

[6] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4974–4983, 2019.

[7] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, pages 686–695, 2017.

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

[9] Ioana Croitoru, Simion-Vlad Bogolin, and Marius Leordeanu. Unsupervised learning of foreground object segmentation. *International Journal of Computer Vision*, 127(9):1279–1302, 2019.

[10] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[11] Fida El Baf, Thierry Bouwmans, and Bertrand Vachon. A fuzzy approach for background subtraction. In *2008 15th IEEE International Conference on Image Processing*, pages 2648–2651. IEEE, 2008.

[12] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014.

[13] Huazhu Fu, Dong Xu, Bao Zhang, Stephen Lin, and Rabab Kreidieh Ward. Object-based multiple foreground video co-segmentation via multi-state selection graph. *IEEE Transactions on Image Processing*, 24(11):3415–3424, 2015.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[15] 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.

[16] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 786–802, 2018.

[17] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 54–70, 2018.

[18] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2126. IEEE, 2017.

[19] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3442–3450, 2017.

[20] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for object tracking. In *The DAVIS Challenge on Video Object Segmentation*, 2017.

[21] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2199, 2013.

[22] Siyang Li, Bryan Seybold, Alexey Vorobyov, Alireza Fathi, Qin Huang, and C-C Jay Kuo. Instance embedding transfer to unsupervised video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6526–6535, 2018.

[23] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 207–223, 2018.

[24] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 90–105, 2018.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[26] Xinkai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. *arXiv preprint arXiv:2007.07020*, 2020.

[27] 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 computer vision and pattern recognition*, pages 3623–3632, 2019.

[28] Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David J Crandall, and Steven CH Hoi. Learning video object segmentation from unlabeled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8960–8970, 2020.

[29] 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.

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

[31] Nicolas Märki, Federico Perazzi, Oliver Wang, and Alexander Sorkine-Hornung. Bilateral space video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 743–751, 2016.

[32] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. `https://github.com/facebookresearch/maskrcnn-benchmark`, 2018. Accessed: [Insert date here].

[33] Shahrizat Shaik Mohamed, Nooritawati Md Tahir, and Ramli Adnan. Background modelling and background subtraction performance for object detection. In *2010 6th International Colloquium on Signal Processing & its Applications*, pages 1–6. IEEE, 2010.

[34] Tan-Cong Nguyen, Gia-Han Diep, Hung V Tran, and Minh-Triet Tran. Sivos: Simulated interactive video object segmentation.

[35] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[36] 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.

[37] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2663–2672, 2017.

[38] 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 Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[39] 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.

[40] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[41] Jordi Pont-Tuset Sergi Caelles and Alberto Montes. davis2017-evaluation. `https://github.com/davisvideochallenge/davis2017-evaluation`, 2018. Accessed: [Insert date here].

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

[43] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4481–4490, 2017.

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

[45] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5277–5286, 2019.

[46] 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 Computer Vision and Pattern Recognition*, pages 9481–9490, 2019.

[47] 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 *Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

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

[49] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 9236–9245, 2019.

[50] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):20–33, 2017.

[51] 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*

*computer vision and pattern recognition*, pages 3064–3074, 2019.

[52] 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 computer vision and pattern recognition*, pages 7376–7385, 2018.

[53] Xin Xiao, Changbin Cui, and Yao Lu. Global tracklet matching for unsupervised video object segmentation.

[54] Shuangjie Xu, Linchao Bao, and Pan Zhou. Class-agnostic video object segmentation without semantic re-identification. In *CVPR Workshops*, volume 1, 2018.

[55] Shuangjie Xu, Daizong Liu, Linchao Bao, Wei Liu, and Pan Zhou. Mhp-vos: Multiple hypotheses propagation for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 314–323, 2019.

[56] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2019.

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

[58] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 931–940, 2019.

[59] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4048–4056, 2017.

[60] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 628–635, 2013.

[61] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, volume 2, page 3, 2020.

[62] Tianfei Zhou, Wenguan Wang, Yazhou Yao, and Jianbing Shen. Target-aware adaptive tracking for unsupervised video object segmentation.

[63] 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.