MHP-VOS: Multiple Hypotheses Propagation for Video Object Segmentation

Shuangjie Xu[1‡*] Daizong Liu[1*] Linchao Bao[2†] Wei Liu[2] Pan Zhou[1†]
[1]Huazhong University of Science and Technology [2]Tencent AI Lab
{shuangjiexu, dzliu, panzhou}@hust.edu.cn linchaobao@gmail.com wl2223@columbia.edu

Abstract

We address the problem of semi-supervised video object segmentation (VOS), where the masks of objects of interests are given in the first frame of an input video. To deal with challenging cases where objects are occluded or missing, previous work relies on greedy data association strategies that make decisions for each frame individually. In this paper, we propose a novel approach to defer the decision making for a target object in each frame, until a global view can be established with the entire video being taken into consideration. Our approach is in the same spirit as Multiple Hypotheses Tracking (MHT) methods, making several critical adaptations for the VOS problem. We employ the bounding box (bbox) hypothesis for tracking tree formation, and the multiple hypotheses are spawned by propagating the preceding bbox into the detected bbox proposals within a gated region starting from the initial object mask in the first frame. The gated region is determined by a gating scheme which takes into account a more comprehensive motion model rather than the simple Kalman filtering model in traditional MHT. To further design more customized algorithms tailored for VOS, we develop a novel mask propagation score instead of the appearance similarity score that could be brittle due to large deformations. The mask propagation score, together with the motion score, determines the affinity between the hypotheses during tree pruning. Finally, a novel mask merging strategy is employed to handle mask conflicts between objects. Extensive experiments on challenging datasets demonstrate the effectiveness of the proposed method, especially in the case of object missing.

1. Introduction

Semi-supervised Video Object Segmentation (VOS) is the task to automatically segment the objects of interests in a video given the annotations in the first frame, which is a fundamental task with wide applications in video editing, video summarization, action recognition, *etc*. Although tremendous progress has been made with semantic segmentation CNNs [24, 7, 8, 28] recently, VOS is still challenging in objects missing and association problems due to occlusions, large deformations, complex object interactions, rapid motions, *etc*., as shown in Fig. 1.

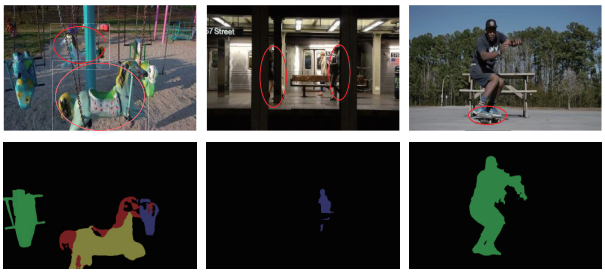


Figure 1. Challenging examples handled by previous approaches. In the first example, the front object instance is segmented as two different objects and the farther instance is missing in the result. In the second example, the occluded instance and the re-appearing instance are missing. In the last example, the smaller object near the larger object is incorrectly segmented to be the larger one.

To tackle these challenges, many recent works [22, 35, 25] resort to object proposal schemes [13, 34] to restore missing objects or re-establish objects associations. In these works, proposals of target objects are either generated individually in each frame [22, 35] by semantic detectors, or further merged with a few neighboring frames [25]. However, these approaches rely on a greedy selection of the best object proposal at each time step, for a given object, which becomes a complication with utter dependence on a reliable Re-ID network [25] that can provide accurate similarity scores. In this paper, we instead deal with this problem by employing a multiple hypotheses propagation approach, which builds up a tracking tree for different hypotheses in time steps, enabling us to defer the selection of the best object proposal for each target till a whole proposal tree along temporal domain is established. This delayed decision making provides us a global view to determine data associations in each frame by considering objects information over the entire video, provably more reliable than greedy methods.

The idea of tracking using multiple hypotheses is not new. In the seminal work by Cox and Hingorani [11], Multiple Hypotheses Tracking (MHT) was first introduced to

‡Part of the work was done during an internship at Tencent AI Lab.
*Equal contributions. †Corresponding authors.

the vision community and applied in the context of visual tracking. Unfortunately, the performance of MHT was limited by unreliable target detectors at that time and later abandoned for decades. More recently, it is again demonstrated to achieve state-of-the-art performances for multiple objects tracking when implemented with modern techniques [20]. The basic idea of MHT is to build up a tracking tree with proposals from each frame, and then prune the tree using the tracking scores until the best track left. The key ingredients for the success of MHT in [20] are the gating scheme and scoring function during the construction and pruning of the tracking trees. In the gating scheme, Kalman filtering is employed to restrict proposal children to be spawned within a certain gating area near their parent, such that the tree does not expand too quickly. The scoring function is to determine the similarity between two hypotheses using motion and appearance cues. However, the algorithm is not that reliable when it comes to VOS, especially when there are large object deformations or sudden changes of object movements (see carousel in Fig. 1 as an example). In this case, the simple motion model of Kalman filtering would break and the appearance score would be very brittle.

In this paper, we adapt MHT to VOS and propose a novel method called Multiple Hypotheses Propagation for Video Object Segmentation (MHP-VOS). Starting from the initial bounding box (bbox) of object mask in the first frame, multiple hypotheses are spawned by proposals from the class-agnostic detector within a novel motion gated region instead of Kalman filtering. We also design a novel mask propagation score instead of the appearance similarity score that could be brittle due to large deformations in challenging cases. The mask propagation score, together with motion score, determines the affinity between hypotheses during the tree pruning. After pruning the proposal tree, the final instance segmentation can be generated and propagated with a mask refinement CNN for each object of interests. And the conflicts between objects are further handled with a novel mask merging strategy. Comparing to state-of-the-art approaches, our method is much more robust and achieves the best performances on the DAVIS datasets.

Our main contributions are summarized as follows:

- We adapt a multiple hypotheses tracking method to the VOS task to build up a bbox proposal tracking tree for different objects with a new gating and pruning method, which can be regarded as a delayed decision for global consideration.
- We apply a motion model to proposal gating instead of using the Kalman filtering, and design a novel hybrid pruning score of motion and mask propagation, which are tailored for VOS tasks. We also design a novel mask merging strategy for multi-objects tasks.
- We conduct extensive experiments to show the effectiveness of our method in distinguishing similar objects, handling occluded and re-appearing objects, modeling long-term object deformations, *etc*., which are very difficult to deal with for previous approaches.

2. Related Work

In this section, we briefly summarize recent researches related to our work, including semi-supervised video object segmentation and multiple hypotheses tracking.

Matching-based Video Object Segmentation. This type of approaches generally utilize the given mask in the first frame to extract appearance information for objects of interests, which is then used to find similar objects in succeeding frames. Yoon *et al.* [42] proposed a siamese network to match the object between frames in a deep feature space. In [5], Caelles trained a parent network on still images and then finetuned the pre-trained work with one-shot online learning. To further improve the finetuning performance in [5], Khoreva *et al.* [19] synthesized more training data to enrich the appearances on the basis of the first frame. In addition, Chen *et al*. [9] and Hu *et al*. [16] used pixel-wise embeddings learned from supervision in the first frame to classify each pixel in succeeding frames. Cheng *et al.* [10] proposed to track different parts of the target object to deal with challenges like deformations and occlusions.

Propagation-based Video Object Segmentation. Different from the appearance matching methods, mask propagation methods utilize temporal information to refine segmentation masks propagated from preceding frames. MaskTracker [29] is a typical method following this line, which is trained from segmentation masks of static images with mask augmentation techniques. Hu *et al.* [15] extended MaskTracker [29] by applying active contour on optical flow to find motion cues. To overcome the problem of target missing when fast motion or occlusion occurs, methods [40, 38] combined temporal information from nearby frame to track the target. The CNN-in-MRF method [1] embeds the mask propagation step into the inference of a spatiotemporal MRF model to further improve temporal coherency. Oh *et al.* [39] applied instance detection to mask propagation using a siamese network without online finetuning for a given video. Another method [41] that does not need online learning uses Conditional Batch Normalization (CBN) to gather spatiotemporal features.

Detection-based Video Object Segmentation. Object detection has been widely used to crop out the target from a frame before sending it to a segmentation model. Li *et al*. [22] proposed VS-ReID algorithm to detect missing objects in video object segmentation. Sharir *et al.* [35] produced object proposals using Faster R-CNN [34] to gather proper bounding boxes. Luiten *et al.* [25] used Mask R-CNN [13] to detect supervised targets among the frames and crop them as the inputs of Deeplabv3+ [8]. Most works based on detections select one proposal at each time step greedily. In

contrast, we keep multiple proposals at each time step and make decisions globally for the segmentation.

Multiple Hypotheses Tracking. MHT method is widely used in the field of target tracking [3, 4]. Hypotheses tracking [11] algorithm originally evaluates its usefulness in the context of visual tracking and motion correspondence, and the MHT in [20] proposed a scoring function to prune the hypothesis space efficiently and accurately which is suited to current visual tracking context. Also, Vazquez *et al*. [36] first adopted MHT in the semantic video segmentation task without pruning. In our method, we adapt the approach to the class-agnostic video object segmentation scenario, where propagation scoring is class-agnostic with the motion rules and the propagation correspondences instead of the unreliable appearance scores.

3. Approach

The overall architecture of our proposed MHP-VOS is illustrated in Fig. 2. We first generate bbox object proposals $P^t = \{p_n^t, n = 1, \ldots, N_{\text{roi}}\}$ of image I^t from frame t with a class-agnostic detection approach in Sec. 3.1, and then apply multiple hypotheses propagation recurrently during building the hypotheses propagation tree (Sec. 3.2) with our novel gating and scoring strategies and filter out disturbing hypotheses by N-scan pruning (Sec. 3.3) to introduce long-term knowledge for hypotheses decision. To take advantage of spatial information between different objects in a sequence, the propagation trees for each object are built at the same time. After acquiring each corresponding bounding box proposal b^t associated with the best hypotheses for each object, we obtain current mask M_i for object i using a segmentation model with b^t in Sec. 3.4. At last, we merge instance masks M_i to multi-objects mask M with consideration of intra-objects conflicts in Sec. 3.5.

3.1. Proposal Generation

There are many approaches [34, 13] used to detect the target object in each video frame. In this paper, we take Mask R-CNN [13] network fine-tuned on each sequence as the base-model to generate coarse object proposals, which are the bbox around the objects. Specially, we change the category number of Mask R-CNN from N_{coco} classes to only one class to make it class-agnostic for detecting foreground objects. Note that segmentation results from the Mask branch are not used for VOS, as this branch shares the classification confidence which is not suitable for the segmentation task. With the input of each frame image, we just extract coarse object bounding box proposals with the detection confidence greater than th_p, and non-maximum suppression threshold of th_n to retain all possible proposals for the further mask proposal propagation in the next step. Here, we denote the output proposal of frame t as p_n^t, where n is the n-th proposal of all N_t proposals in detection step.

3.2. Hypotheses Tree Construction

After generating coarse object proposals, we construct the hypotheses propagation tree, whose data structures are designed as follows: each hypothesis node in the tree consists of a bounding box proposal p_k^t and its corresponding mask hypothesis $M^{p_k^t}$. For each target object, the tree starts from the ground-truth mask in the first frame, and will be extended by appending children proposals in the next frame. In this children spawning step, only proposals within a gated region are considered. And the mask hypothesis $M^{p_k^t}$ for each child proposal p_k^t is obtained using the method detailed in Sec. 3.4. This process is repeated until the final hypotheses tree is constructed completely. In addition, each proposal outside the gated region is treated as the starting node in a new tree to catch missing objects. During the tree construction, a novel mask propagation score of each node can be recorded and would be used for tree pruning later, which is more robust than the appearance score.

Gating. To build the hypotheses tree, we need to gate most closely proposals in next frame to be the child nodes, shown in Fig. 3 (a). In general, the bounding box of objects in frame t depends on two main variables: size $s_t, (w_t, h_t)$ and center point coordinate $p_t, (x_t, y_t)$. Thus, the historical movements in n frame from $t-n$ to $t-1$ are adopted as prior knowledge to predict the probability bbox in frame t. For the position prediction, the velocity v_t is estimated by

$$v_t = \frac{1}{n} \sum_{m=t-1}^{t-n} (p_m - p_{m-1}). \tag{1}$$

Then the predicted center point is obtained by $p_t = p_{t-1} + v_t$. And the corresponding average size is taken as the predicted object size $s_t = \frac{1}{n} \sum_{m=t-1}^{t-n} s_m$, since the change in size is tiny and smooth. With the estimation of p_t and s_t, it gives the bbox candidate c_t for comparison in gating.

In order to filter out disturbing proposals, we gate the candidate proposals by computing the IOU score with the bounding box c_t in the last frame as follows:

$$1_n^t = \begin{cases} 1, & \text{iou}(c_t,\ p_n^t) > th_g \\ 0, & \text{iou}(c_t,\ p_n^t) \le th_g \end{cases}, \tag{2}$$

where th_g is the threshold of gating, and 1_n^t denotes whether the candidate box p_n^t gates in or out. With proposals chosen from gating, we can build up the propagation tree to simulate multiple hypotheses proposal propagation.

Scoring. In the propagation tree, each hypotheses is associated with a class-agnostic score for further pruning. It is a recurrent process in each tree node, which is formalized as:

$$S\left(t, p_k^t\right) = w_m S_m\left(t, p_k^t\right) + w_p S_p\left(t, p_k^t\right), \tag{3}$$

where $S_m\left(t, p_k^t\right)$ and $S_p\left(t, p_k^t\right)$ denote the motion score and mask propagation score, respectively. $t = 0, 1, ..., T$ means

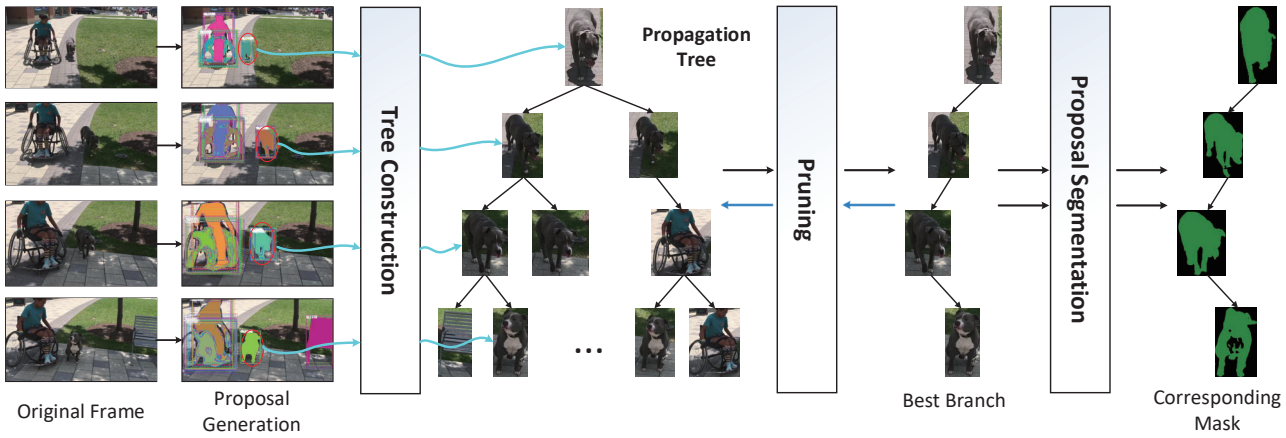


Figure 2. The pipeline of our MHP-VOS algorithm. We first obtain bounding box proposals from Mask RCNN [13], and then construct the proposal propagation tree for each object with gating and scoring strategies. To avoid calculation explosion, an N-scan pruning strategy is applied to remove branches that are far from the best hypothesis. Through this recurrent process between tree building and branches pruning, we can obtain the best propagation track, and then obtain the segmentation mask for each object by mask propagation and merging.

the current video frame number, p_k^t denotes the proposal of the k-th hypotheses track. w_m and w_p control the ratio between motion score and propagation score. There is no Re-ID score involved since it may cause ambiguity when objects of similar appearances exist.

For each bounding box proposal p_k^t of the node in the propagation tree, we define the motion score as:

$$S_m^t\left(t,p_k^t\right)=w_f\frac{p_k^t\cap p_k^{t-1}}{p_k^t\cup p_k^{t-1}}+w_n\max_{i\neq k}\left(\frac{p_k^t\cap p_i^{t-1}}{p_k^t\cup p_i^{t-1}}\right). \quad (4)$$

The motion score is composed of two parts: a) iou score between proposals of same hypotheses in continuous frames, which is positive to the decision; b) iou score between frame t proposal of k-th track and the $(t-1)$-th proposal node in other hypotheses track, and it is expected to be small.

Motion score gives a qualitative mark when the continuity of propagation track is smooth. However, the motion score will be out of order when severe occlusion occurs. In order to handle such case, the mask propagation score is proposed utilizing the quality of segmentation propagated in target proposal, which can be formalized as:

$$S_p^t\left(t,p_k^t\right)=\frac{M^{p_k^t}\cap Q^t\circ M^{p_k^{t-1}}}{M^{p_k^t}\cup Q^t\circ M^{p_k^{t-1}}}, \quad (5)$$

where $\circ$ denotes the warp operation that warps mask from last frame to current frame with optic flow Q. And $M^{p_k^t}$ denotes the single object mask segmentation obtained by method in Sec. 3.4 with the proposal p_k^t. $M^{p_k^t}$ composes the mask hypothesis with bounding box proposals: it starts from ground-truth in frame $t=0$, and forwards propagation with the construction of proposal tree (warp to next frame as priori mask for mask generation in p_k^{t+1} progressively). As for the new start tree for the missing object, the mask of tree root is obtained with blank mask as the priori mask.

At last, the final score of the long-term hypotheses can be computed recursively as:

$$S\left(t,p_k^t\right)=S\left(t-1,p_k^{t-1}\right)+S^t\left(t,p_k^t\right), \quad (6)$$

$$S^t\left(t,p_k^t\right)=\begin{cases}\ln\left(1-P_D\right), & t=0\\ w_mS_m^t+w_pS_p^t, & t\neq 0\end{cases}, \quad (7)$$

where P_D denotes the probabilities of detection.

3.3. Hypotheses Tree Pruning

During the construction of the hypotheses tree, the number of hypotheses tracks increases exponentially during propagation, which leads to the explosion of memory and computation. Thus, we have to take a pruning step to limit the size of the tree. In other words, we need to determine the most likely context propagation tracks in long term, of which the optimization can be formulated as:

$$\max_H\sum_{t=0}^{T}S\left(t,p_k^t\right), \quad (8)$$

where $H_k=\left\{p_k^i|t=0,1,\ldots,t\right\}$ means a proposal propagation hypothesis (track path from root to leaf node in the propagation tree) and $H=\left\{H_k|k=0,1,\ldots,N_h\right\}$ means hypothesis space for tracks of an object. N_h means the Hypotheses space size for the target object.

To find the best track among the kinds of propagation tracks, this task can be formulated as a Maximum Weighted Independent Set (MWIS) problem as described in [27]. For the track tree in frame t, we build an undigraph $G=(V,E)$ with each propagation hypothesis H_k taken as a node in V. The edge (l,j) in E connects the hypothesis pair (H_l,H_j) which has the same proposal at the same frame, which means the two hypotheses are conflicting and cannot coexist for the final independent set $B=\left\{b^i|i=0,\ldots,t\right\}$. With the track score described in Eq. (8) as the weight w

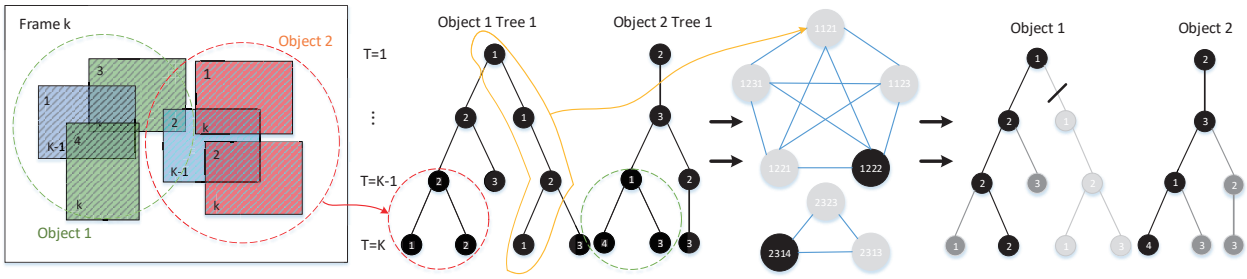


Figure 3. The illustration of MHP at time k. (a) A gating example for propagation track of two objects from frame $k-1$ to k. Bbox IOU scores between proposal from the current frame and the predicted bbox from the last frame are utilized as a gate with thresholds d_{th}. (b) The corresponding propagation trees. Each tree node is associated with a proposal observation. (c) The undigraph for the example of (b), in which each node represents a propagation path in the tree and each edge connects two tracks that are conflicted. The black nodes in graph form the Maximum Weighted Independent Set (MWIS). (d) An N-scan pruning example when $N = 2$. The dark branches denote the global hypothesis at frame k, and the oblique lines represent the pruning of this branch which is far from the global hypothesis in $k - N$.

of each track branch, we optimize the problem to find the maximum weight independent set B as follows:

$$\max_i w_i,\ i \in \{l, j\}, \forall (i, j) \in E. \tag{9}$$

We utilize the existing phased local search (PLS) algorithm [32, 33, 2] to solve the MWIS optimization problem. Also, we take the N-scan pruning method to prune the disturbing branches gradually instead of pruning the whole tree. First, we apply the Eq. (9) to choose the maximum independent set as the best hypothesis from hypothesis space H, and then track the nodes in frame k back to the node in frame $k - N$ as sub-trees. Finally, we prune the sub-trees except the independent tracks. A larger N makes a longer decision delay, which will bring an improvement in precision but take time efficiency as price. In addition, we also limit the number of branches to avoid proposal tree growing too large. If the number of branches is more than th_b at any node in any frame, we retain the top th_b branches with the propagation scores and prune the other branches.

3.4. Single Object Segmentation

We employ Deeplabv3+ [8] network with a ResNet101 [14] backbone as our segmentation module, to generate segmentation results from bounding box proposals. Similar to MaskTracker [29], the segmentation network takes an additional rough mask as input, which is warped from the mask of the previous frame to the current frame using optical flow estimated by FlowNet2 [17]. This module is used to generate mask hypothesis from proposal during the tree construction, and can produce the final segmentation result once the best proposal for an object is obtained after the tree pruning. Taking the final segmentation as an example, we crop the bounding box of a single object and its previous mask by b_i^t with margin ratio r, and then concatenate the RGB image with the warped mask Q_i^t as a fourth channel input. After obtaining the segmentation probability map Z_i^t from Deeplabv3+, we obtain the instance-specific mask M_i^t with

Algorithm 1 Multi-Instance Merging Strategy.

Require:
instance-specific masks $M_i^t, i = 1, \ldots, C$ for all objects, history mask M_i^{t-1}, segmentation probability map from Deeplabv3+ $Z_i^t, i = 1, \ldots, n$, and Gaussian map $G_i^{b^t}$.

Ensure:
set multi-instance segmentation Y^t with the object id that has the max value in M^t pixel-by-pixel;
for patch a in all overlap patches
  Ids $\Leftarrow$ all object ids sorted by value $\mathrm{sum}\left(Z_i^t[a]\right)$ from high to low;
  if $\mathrm{sum}\left(G^{b^t}_{\mathrm{Ids}[0]} * Z^t_{\mathrm{Ids}[0]}[a]\right) \cdot \lambda$
    $>\mathrm{sum}\left(G^{b^t}_{\mathrm{Ids}[1]} * Z^t_{\mathrm{Ids}[1]}[a]\right)$ **then**
    $Y^{f_t}[a] \Leftarrow \mathrm{Ids}[0]$;
  else
    obtain the warped mask $Q^t_{\mathrm{Ids}[0]}$ from $M^{t-1}_{\mathrm{Ids}[0]}$, $Q^t_{\mathrm{Ids}[1]}$ from $M^{t-1}_{\mathrm{Ids}[1]}$;
    if $\mathrm{sum}\left(G^{b^t}_{\mathrm{Ids}[0]} * Q^t_{\mathrm{Ids}[0]}[a]\right)$
      $>\mathrm{sum}\left(G^{b^t}_{\mathrm{Ids}[1]} * Q^t_{\mathrm{Ids}[1]}[a]\right)$ **then**
      $Y^t[a] \Leftarrow \mathrm{Ids}[0]$;
    else
      $Y^t[a] \Leftarrow \mathrm{Ids}[1]$;
return Y^t for the multi-instance segmentation;

threshold th_m as following:

$$M_i^t = (Z_i^t > th_m), i = 1, 2, ..., C, \tag{10}$$

where C denotes the total object number in one sequence.

3.5. Conflicts Handling for Multiple Objects

To merge the instance-specific masks M_i^t into the final multi-instance segmentation Y^t, we propose a merging strategy as shown in Algorithm 1. In general, there are two

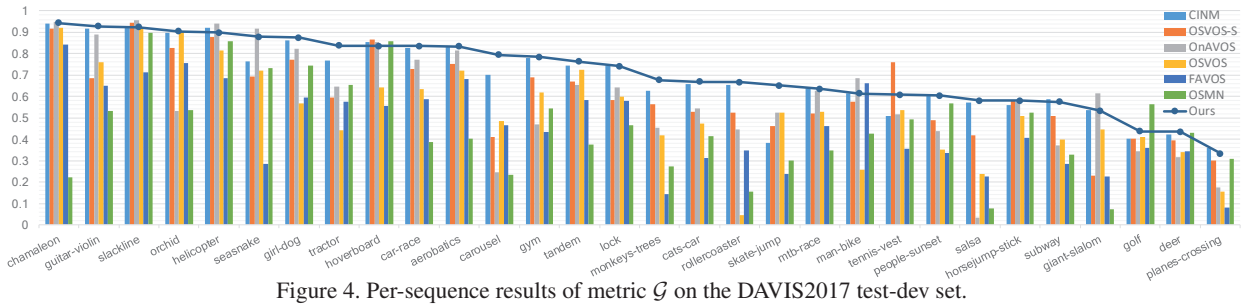


Figure 4. Per-sequence results of metric $\mathcal{G}$ on the DAVIS2017 test-dev set.

Dataset		Metric	OSMN [41]	FAVOS [10]	OSVOS [5]	OnAVOS [37]	OSVOS-S [26]	CINM [1]	Ours
validation	$\mathcal{J}$	Mean $\mathcal{M}\uparrow$	52.5	54.6	56.6	61.6	64.7	67.2	**71.8**
	$\mathcal{F}$	Mean $\mathcal{M}\uparrow$	57.1	61.8	63.9	69.1	71.3	74.0	**78.8**
	$\mathcal{G}$	Mean $\mathcal{M}\uparrow$	54.8	58.2	60.3	65.4	68.0	70.6	**75.3**
test-dev	$\mathcal{J}$	Mean $\mathcal{M}\uparrow$	37.7	42.9	47.0	49.9	52.9	64.5	**66.4**
		Recall $\mathcal{R}\uparrow$	38.9	48.1	52.1	54.3	60.2	73.8	**76.0**
		Decay $\mathcal{D}\downarrow$	19.0	18.1	19.2	23.0	24.1	20.0	**18.0**
	$\mathcal{F}$	Mean $\mathcal{M}\uparrow$	44.9	44.2	54.8	55.7	62.1	70.5	**72.7**
		Recall $\mathcal{R}\uparrow$	47.4	51.1	59.7	60.3	70.5	79.6	**82.2**
		Decay $\mathcal{D}\downarrow$	**17.4**	19.8	19.8	23.4	21.9	20.0	19.0
	$\mathcal{G}$	Mean $\mathcal{M}\uparrow$	41.3	43.6	50.9	52.8	57.5	67.5	**69.5**

Table 1. Quantitative comparison of state-of-the-art methods on the DAVIS2017 validation and test-dev sets. The up-arrow $\uparrow$ means that larger is better while the down-arrow $\downarrow$ means that smaller is better. Our algorithm achieves the best performances on both sets.

kinds of cases when we decide each pixel id in the final segmentation. For the pixel belonging to one object, we set the object id to be the same as the the corresponding pixel among the single instance masks. However, the pixel may belong to different objects at the same time when the overlap conflicts happen between multi-instance masks. To determine the object id for the overlapped region, we first take the top two possible object ids sorted by the corresponding values in the probability map from DeeplabV3+ as id candidates. We then accept the object id with higher probability only when there is a large margin between the two probability values (we use a marginal ratio $\lambda = 0.8$). Otherwise, we take temporal coherency of the warped mask in consideration when it is ambiguous to use spatial information only. Besides, a two-dimensional gaussian map G^{b^t} is generated from the proposal b^t with parameters of $\sigma_x^t = {}^{w}/_{2}$ and $\sigma_y^t = {}^{h}/_{2}$ as prior knowledge to obtain the weighted mask without noise out of the region of interests, where w and h are the width and height of proposal b^t, respectively.

4. Experiments

In this section, we investigate the performance of our method on standard benchmark datasets: DAVIS2016 [30] and DAVIS2017 [6]. We compare our model with state-of-the-art methods and perform ablation study to demonstrate the advantage of each component in MHP-VOS.

4.1. Implementation Details

To adapt the Mask R-CNN [13] network to DAVIS [30, 31, 6] task, we first train the network on COCO [23] dataset with the pre-trained ImageNet [12] weights, and then finetune it on DAVIS dataset. Before testing, we finetune the parent model weights on each sequence respectively with the corresponding $N_l = 200$ synthetic in-domain image-pairs of Lucid Dreaming [5]. Then, coarse proposals are selected with the $th_p = 0.05$ and $th_n = 0.6$.

During the training of the Deeplabv3+ [8] network with a ResNet101 [14] backbone, we crop the bbox of the four channel input by using the spatial information of the annotation with margin ratio $r = 0.15$. Then, we resize the cropped data into 512×512, jitter the image color, and then train them for 100 epochs both on COCO [23] and DAVIS [30, 31, 6] datasets. We use BCEWithLogits loss function, and set Adam [21] optimizer with lr $= 1e-5$ which reduces by power of 0.9 for every 10 epochs. In the fine-tuning, we only train the parent model on synthetic image-pairs for 50 epochs, and the lr starts from $5e-6$ and also reduces by power of 0.9 for every 10 epochs. We set $th_m = 0.3$ to get the valid mask with the corresponding probability map. At last, the instance masks are merged with $\lambda = 0.8$. In N-scan pruning phase, we set $N = 3$ and $th_b = 50$. All experiments are implemented on a single NVIDIA 1080 GPU. The code is available at https://github.com/shuangjiexu/MHP-VOS.

4.2. Datasets and Evaluation

DAVIS2016. DAVIS2016 [30] dataset is proposed recently to evaluate VOS methods and contains 50 video sequences divided into train and test parts. Each video sequence consists of a single object, and it provides each object with the corresponding mask among the sequences.

DAVIS2017. DAVIS2017 [31] dataset is extended from

Figure 5. Qualitative results from the DAVIS2017 test-dev and DAVIS2016 validation sets, where the images are sampled at the average intervals for each video. From top to bottom, the sequences are "carousel", "monkeys-trees", and "salsa" on the DAVIS2017 test-dev, "bmx-trees" and "libby" on the DAVIS2016 validation. Different objects are highlighted as different colors.

DAVIS2016, and it is more challenging in multiple objects which correspond to different targets. It provides extra test-dev data with 30 challenging videos, which contains some similar objects in the same videos and object occlusion or missing in the continues frames. Background noise is also a challenge which has similar appearance with target objects.

Evaluation. We adopt the protocols in [30] which contains two evaluation metrics, region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$. In addition, both two evaluation metrics consist of three statistics measurement: mean $\mathcal{M}$, recall $\mathcal{R}$ and decay $\mathcal{D}$. The global metric $\mathcal{G}$ is the mean of $\mathcal{J}$ and $\mathcal{F}$.

4.3. DAVIS2017

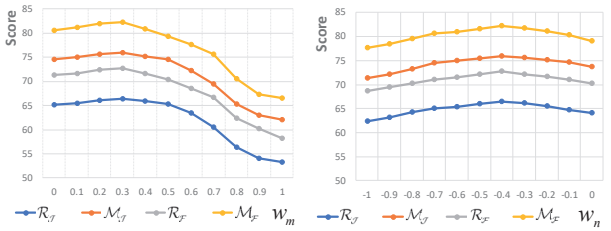
Comparison to the State-of-the-arts. Table. 1 shows the quantitative comparison on DAVIS-2017 valid and test-dev sets, where we find that MHP-VOS performs the state-of-the-art in most evaluation matrices. Especially on the validation set, MHP-VOS beats all the latest methods and achieves higher Mean value. As illustrated in Table. 1 on the more challenge test-dev set, our model also gets great results. In terms of $\mathcal{M}_\mathcal{J}$, $\mathcal{M}_\mathcal{F}$ and $\mathcal{M}_\mathcal{G}$, our method outperforms the state-of-the-art CINM [1] by 2.1%, 2.2% and 2.0% respectively, with neither CRF or MRF applied.

Improvement. Many previous works are troubled by occlusion, similar objects or fast motion. However, as shown in Fig. 5, our method handles these challenges well. In the case of similar objects like "carousel", which will be mistakenly switched identities by OSVOS [5], our propagation proposals can track different instances well and identify each object. Also, we investigate that our method is robustly enough to the issues of fast motion and small instances, especially in "monkeys-tree" sequence. For the occlusion problem, we find that the segmentation on "salsa" performs identifiable which demonstrates the strong representation power of our model. The performances on these challenge sequences can also be illustrated in Fig. 4, where we achieve the state-of-the-art on almost all the videos.

Settings					Mean $\mathcal{M}$	Boost
w_m	w_p	N	Merge	Gating		
1.0	0.0	1	×	×	47.3	-
0.3	0.7	1	×	×	52.1	4.8
0.3	0.7	3	×	×	59.7	**7.6**
0.3	0.7	3	✓	×	67.3	**7.6**
0.3	0.7	3	✓	✓	**69.5**	2.2

Table 2. Ablation study on the DAVIS2017 test-dev set.

Ablation Study. Table. 2 shows how much each presented component builds up to the final result. We start by the baseline model only with the motion score for pruning ($w_m = 1.0$,$w_p = 0.0$), and there is no no multiple hypotheses (N=1), no merge strategy (× in Merge, which means choose area with larger probability when conflict) and no traditional gating strategy [20] (× in Gating) in addition. Results show that the hybrid scoring of motion and propagation achieves 4.8 higher than the original motion score. Multiple hypotheses and the conflicts handling strategy both make the maximum improvement of performance with 7.6,



(a) $w_n = -0.4$ (b) $w_m = 0.3$

Figure 6. Segmentation qualities on DAVIS17 according to the two hyper-parameters: w_n, w_m. (a) Score versus w_m when $w_n = -0.4$. (b) Score versus w_n when $w_m = 0.3$.

respectively. At last, our gating strategy brings another improvement of 2.2 instead of using Kalman Filter [18].

In the scoring phase, four hyper-parameters (w_m, w_p, w_f and w_n) are introduced to balance the weights between the scores of motion and propagation, where $w_p = 1 - w_m$ and $w_f = 1$. We apply grid search on parameters $w_m \in [0, 1]$ and $w_n \in [-1, 0]$ with the step set as 0.1. Part of the grid search result is shown in Fig. 6. Experimental results show that MHP-VOS achieves the best result when $w_m = 0.3$ and $w_n = -0.4$. As the phase of proposal tree formation, we apply N-scan pruning with parameter N to control the delay time of proposal decision. In practice, N is an interesting parameter that makes a trades off between performance and speed. Shown as Table. 3, lager decision delay time (N) receives a performance boost, but gets the punishment in speed. We set $N = 3$ to achieve a balanced performance.

N	1	3	5
time/frame (s)	0.8	14.2	73.6
Mean $\mathcal{M}$	62.8	69.5	69.7

Table 3. Trade-off effect of N-scan pruning on DAVIS2017.

Weakness. Here we report typical examples of mistaken cases on DAVIS2017 test-dev. In the first video sequence, the segmentation of deer in the left (green) is partly missing, which is due to the similar appearance in the context pixels. The instance detector may regard the body of the deer to part of the tree and only generates the proposal of the head with the contrast background. Next in the middle sequence, we find that the racket is segmented well in previous frames but missed in the later. This is because the proposed merging strategy that classifies the identity of overlap region wrongly in the ambiguous case. In the last video, the person in yellow is gradually switched to blue which means the proposal of this person is propagated wrongly during the tree building with two overlap bounding boxes of these disturbing objects.

Figure 7. Mistaken cases on DAVIS2017 test-dev. Sequences correspond to "deer", "tennis-vest" and "people-sunset" respectively.

Method	Mean $\mathcal{M}$			Recall $\mathcal{R}$	
	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{G}$	$\mathcal{R}_\mathcal{J}$	$\mathcal{R}_\mathcal{F}$
OSMN [41]	74.0	72.9	73.5	87.6	84.0
PML [9]	75.5	79.3	77.4	89.6	93.4
MSK [29]	79.7	75.4	77.6	93.1	87.1
FAVOS [10]	82.4	79.5	81.0	96.5	89.4
RGMP [39]	81.5	82.0	81.8	91.7	90.8
CINM [1]	83.4	85.0	84.2	94.9	92.1
MoNet [40]	84.7	84.8	84.7	**96.8**	94.7
MGCRN [15]	84.4	85.7	85.1	97.1	95.2
OnAVOS [37]	**86.1**	84.9	85.5	96.1	89.7
OSVOS-S [26]	85.6	87.5	86.6	**96.8**	**95.5**
Ours	85.7	**88.1**	**86.9**	96.6	94.8

Table 4. Comparison results on the DAVIS2016 validation set.

4.4. DAVIS2016

As illustrated in Table. 4, our method achieves great progress with the $\mathcal{M}_\mathcal{J}$, $\mathcal{M}_\mathcal{F}$ and $\mathcal{M}_\mathcal{G}$ of 85.7%, 88.1% and 86.9%, which outperforms the state-of-the-art OSVOS-S [26] by 0.1%, 0.6% and 0.3% respectively. Compared to the traditional method MSK [29], our MHP-VOS improves a lot by 9.3% on the Global Mean $\mathcal{M}_\mathcal{G}$. Also, we investigate that our performance is better than many latest models, like FAVOS [10] and MoNet [40]. Although our method performs well on DAVIS2016 validation set, there are not huge improvement between ours and the state-of-the-art models, for the reason that the proposal propagation is not essential for single object tracking, and the CNN-based segmentation module is capable enough to locate the foreground instance. As shown in Fig. 5, each target object has corresponding accurate segmentation even in motion blur or occlusion cases.

按这么说，在单目标跟踪上做就没有意义了？未必吧。

5. Conclusion

In this work, we presented a novel detection based Multiple Hypotheses Propagation (MHP-VOS) method for semi-supervised video object segmentation. The key to MHP-VOS is that the decision for proposal in one frame is delayed to eliminate ambiguity with long-term information. Therefore, a hypothesis propagation tree was introduced to catch more potential proposals in each frame for tracking, with a novel class-agnostic gating and scoring strategy adapted to the VOS scenario. In addition, a novel conflicts handling method for multiple objects was proposed to transfer MHP-VOS to the multiple objects setting. Our experiments investigate performances of the pipeline and each component module, which are demonstrated to achieve significant performance gains compared against the state-of-the-arts.

References

[1] L. Bao, B. Wu, and W. Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5977–5986, 2018. 2, 6, 7, 8

[2] L. Barth, B. Niedermann, M. Nöllenburg, and D. Strash. Temporal map labeling: A new unified framework with experiments. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 23, 2016. 5

[3] S. Blackman and R. Popoli. Design and analysis of modern tracking systems (artech house radar library). *Artech house*, 1999. 3

[4] S. S. Blackman. Multiple hypothesis tracking for multiple target tracking. *IEEE Aerospace and Electronic Systems Magazine*, 19(1):5–18, 2004. 3

[5] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 6, 7

[6] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 1(2), 2018. 6

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 1

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018. 1, 2, 5, 6

[9] Y. Chen, J. Pont-Tuset, A. Montes, and L. Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1189–1198, 2018. 2, 8

[10] J. Cheng, Y.-H. Tsai, W.-C. Hung, S. Wang, and M.-H. Yang. Fast and accurate online video object segmentation via tracking parts. *arXiv preprint arXiv:1806.02323*, 2018. 2, 6, 8

[11] I. J. Cox and S. L. Hingorani. An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(2):138–150, 1996. 1, 3

[12] J. Deng, W. Dong, R. Socher, and L. J. Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 1, 2, 3, 4, 6

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. pages 770–778, 2015. 5, 6

[15] P. Hu, G. Wang, X. Kong, J. Kuen, and Y.-P. Tan. Motion-guided cascaded refinement network for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1400–1409, 2018. 2, 8

[16] Y.-T. Hu, J.-B. Huang, and A. G. Schwing. Videomatch: Matching based video object segmentation. *arXiv preprint arXiv:1809.01123*, 2018. 2

[17] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[18] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960. 8

[19] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for multiple object tracking. *arXiv preprint arXiv:1703.09554*, 2017. 2

[20] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg. Multiple hypothesis tracking revisited. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4696–4704, 2015. 2, 3, 7

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014. 6

[22] X. Li, Y. Qi, Z. Wang, K. Chen, Z. Liu, J. Shi, P. Luo, X. Tang, and C. C. Loy. Video object segmentation with re-identification. *arXiv preprint arXiv:1708.00197*, 2017. 1, 2

[23] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. 8693:740–755, 2014. 6

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[25] J. Luiten, P. Voigtlaender, and B. Leibe. Premvos: Proposal-generation, refinement and merging for the davis challenge on video object segmentation 2018, 2018. 1, 2

[26] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. Video object segmentation without temporal information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 6, 8

[27] D. J. Papageorgiou and M. R. Salpukas. The maximum weight independent set problem for data association in multiple hypothesis tracking. In *Optimization and Cooperative Control Strategies*, pages 235–255. Springer, 2009. 4

[28] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel mattersimprove semantic segmentation by global convolutional network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1743–1751, 2017. 1

[29] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 5, 8

[30] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. 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. 6, 7

[31] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 6

[32] W. Pullan. Phased local search for the maximum clique problem. *Journal of Combinatorial Optimization*, 12(3):303–323, 2006. 5

[33] W. Pullan. Optimisation of unweighted/weighted maximum independent sets and minimum vertex covers. *Discrete Optimization*, 6(2):214–219, 2009. 5

[34] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):1137–1149, 2017. 1, 2, 3

[35] G. Sharir, E. Smolyansky, and I. Friedman. Video object segmentation using tracked object proposals. *arXiv preprint arXiv:1707.06545*, 2017. 1, 2

[36] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller. Multiple hypothesis video segmentation from superpixel flows. In *Proceedings of the European Conference on Computer Vision*, pages 268–281, 2010. 3

[37] P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017. 6, 8

[38] S. Wang, Y. Zhou, J. Yan, and Z. Deng. Fully motion-aware network for video object detection. In *Proceedings of the European Conference on Computer Vision*, pages 542–557, 2018. 2

[39] S. Wug Oh, J.-Y. Lee, K. Sunkavalli, and S. 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. 2, 8

[40] H. Xiao, J. Feng, G. Lin, Y. Liu, and M. Zhang. Monet: Deep motion exploitation for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1140–1148, 2018. 2, 8

[41] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos. Efficient video object segmentation via network modulation. *algorithms*, 29:15, 2018. 2, 6, 8

[42] J. S. Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. S. Kweon. Pixel-level matching for video object segmentation using convolutional neural networks. In *IEEE International Conference on Computer Vision*, pages 2186–2195, 2017. 2