PReMVOS: Proposal-generation, Refinement and Merging for Video Object Segmentation

Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe

Computer Vision Group, RWTH Aachen University, Aachen, Germany
{luiten, voigtlaender, leibe}@vision.rwth-aachen.de

Abstract. We address semi-supervised video object segmentation, the task of automatically generating accurate and consistent pixel masks for objects in a video sequence, given the first-frame ground truth annotations. Towards this goal, we present the PReMVOS algorithm (Proposal-generation, Refinement and Merging for Video Object Segmentation). Our method separates this problem into two steps, first generating a set of accurate object segmentation mask proposals for each video frame and then selecting and merging these proposals into accurate and temporally consistent pixel-wise object tracks over a video sequence in a way which is designed to specifically tackle the difficult challenges involved with segmenting multiple objects across a video sequence. Our approach surpasses all previous state-of-the-art results on the DAVIS 2017 video object segmentation benchmark with a $\mathcal{J}\&\mathcal{F}$ mean score of 71.6 on the `test-dev` dataset, and achieves first place in both the DAVIS 2018 Video Object Segmentation Challenge and the YouTube-VOS 1st Large-scale Video Object Segmentation Challenge.

1 Introduction

Video Object Segmentation (VOS) is the task of automatically estimating the object pixel masks in a video sequence and assigning consistent object IDs to these object masks over the video sequence. This can be seen as extension of instance segmentation from single frames to videos, and also as an extension of multi object tracking from tracking bounding boxes to tracking pixel masks. This framework motivates our work in separating the VOS problem into two subproblems. The first being the instance segmentation task of generating accurate object segmentation mask proposals for all of the objects in each frame of the video, and the second being the multi object tracking task of selecting and merging these mask proposals to generate accurate and temporally consistent pixelwise object tracks throughout a video sequence. Semi-supervised Video Object Segmentation focuses on the VOS task for certain objects for which the ground truth mask for the first video frame is given. The DAVIS datasets [3],[27],[26] present a state-of-the-art testing ground for this task. In this paper we present the PReMVOS (Proposal-generation, Refinement and Merging for Video Object Segmentation) algorithm for tackling the semi-supervised VOS task. This method involves generating coarse object proposals using a Mask R-CNN like
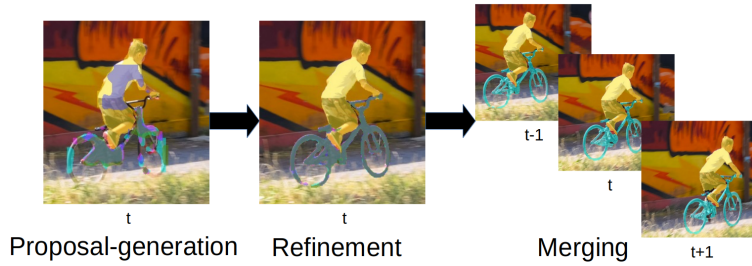
Fig. 1. PReMVOS overview. Overlay colours represent different object proposals.

object detector, followed by a refinement network that produces accurate pixel masks for each proposal. We then select and link these proposals over time using a merging algorithm that takes into account an objectness score, the optical flow warping, a Re-ID feature embedding vector, and spatial constraints for each object proposal. We adapt our networks to the target video domain by fine-tuning on a large set of augmented images generated from the first-frame ground truth. An overview of our method, PReMVOS, can be seen in Figure 1. Our method surpasses all current state-of-the-art results on all of the DAVIS benchmarks and achieves the best results in the 2018 DAVIS Video Object Segmentation Challenge [20] and the YouTube-VOS 1st Large-scale Video Object Segmentation Challenge [13].

2   Related Work

Current state-of-the-art methods for VOS fall into one of two paradigms. The first is *objectness* estimation with domain adaptation from first-frame finetuning. This approach, first proposed in [2], uses fully convolutional networks to estimate the *objectness* of each pixel by fine-tuning on the first-frame ground truth. This approach was expanded upon by [21] and [29],[30] by using semantic segmentation guidance and iterative fine-tuning, respectively. The second paradigm, used in several state-of-the-art methods [25],[18],[15],[17], involves propagating the mask from the previous frame using optical flow and then refining these estimates using a fully convolutional network. The methods proposed in [18] and [17] expand this idea by using a network to calculate a re-identification (ReID) embedding vector for proposed masks and using this to improve the object re-identification after an object has been occluded. [15] improves upon the mask propagation paradigm by training on a huge set of augmented images generated from the first-frame ground truth. Our method tackles the VOS task in an inherently different way than any of the previous papers in the literature. However, we adopt ideas presented in all of the above papers such as the use of ReID embedding vectors, optical flow proposal warping and fine-tuning on a large set of images augmented from the first frame.
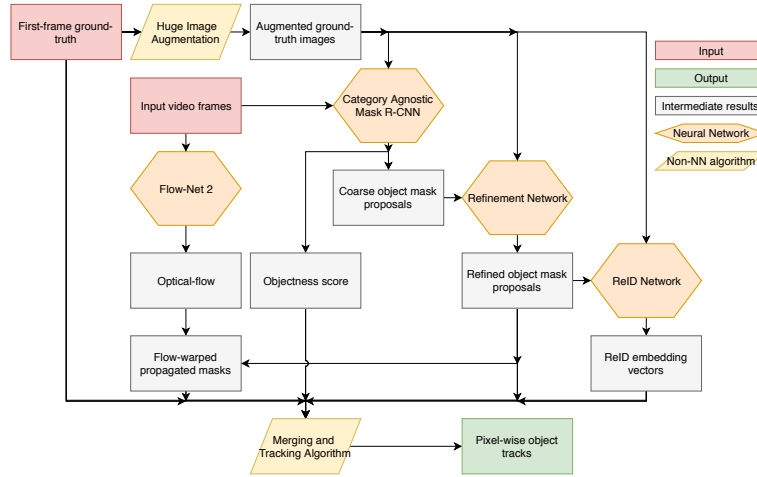
Fig. 2. Diagram showing the components of PReMVOS and their relationships.

3   Approach

We propose PReMVOS as a novel approach for addressing the VOS task. This approach is designed to produce more accurate and temporally consistent pixel masks across a video, especially in the challenging multi-object VOS task. Instead of predicting object masks directly on the video pixels, as done in [2], [21], [29] and [30], a key idea of our approach is to instead detect regions of interest as coarse object proposals using an object detection network, and to then predict accurate masks only on the cropped and resized bounding boxes. We also present a new proposal merging algorithm in order to predict more temporally consistent pixel masks. The methods presented in [25],[18],[15],[17] create temporally consistent proposals by generating their proposals directly from the previous frame's proposals warped using optical flow into the current frame. Instead, our method generates proposals independently for each frame and then selects and links these proposals using a number of cues such as optical flow based proposal warping, ReID embeddings and objectness scores, as well as taking into account the presence of other objects in the multi-object VOS scenario. This novel paradigm for solving the VOS task allows us to predict both more accurate and more temporally consistent pixel masks than all previous methods and achieves state-of-the-art results across all datasets. Figure 2 shows an overview of each of the components of the PReMVOS algorithm and how these work together to solve the VOS task.

3.1   Image Augmentation

For each video we generate a set of 2500 augmented images using the first-frame ground truth. We use the Lucid Data Dreaming method proposed in [15] but

only generate single images (not image pairs). This method removes the objects, automatically fills in the background, and then randomly transforms each object and the background before randomly reassembling the objects in the scene. Fine-tuning on this set of augmented images allows us to adapt our networks directly to the target video domain.

3.2   Proposal Generation

We generate coarse object proposals using a Mask R-CNN [9] network implementation by [31] with a ResNet101 [10] backbone. We adjust this network to be category agnostic by replacing the N classes with just one class by mapping all classes to a single foreground class for detecting generic objects. We train this network starting from pre-trained ImageNet [6] weights on both the COCO [19] and Mapillary [22] datasets jointly. We then fine-tune a separate version of this network for each video for three epochs of the 2500 augmented images. This network generates coarse mask proposals, bounding boxes, and objectness scores for each image in the video sequence. We extract proposals with a score greater than 0.05 and also perform non-maximum suppression removing proposals which have an IoU of 66% or greater with a proposal with a higher score.

3.3   Proposal Refinement

The Proposal-Refinement Network is a fully convolutional network inspired by [34] and based on the DeepLabv3+ [4] architecture. This network takes as input a 385×385 image patch that has been cropped and resized from an approximate bounding box around an object of interest. A 50 pixel (in the original image) margin is first added to the bounding box in all directions. We add a fourth channel to the input image which encodes the original bounding box as a pixel mask to the input image. Starting from weights pre-trained on ImageNet [6], COCO [19], and PASCAL [7], we train this network on the Mapillary [22] dataset using random flipping, random gamma augmentations and random bounding box jitter [34] up to 5% in each dimension, to produce an accurate object segmentation, given an object's bounding box. We then fine-tune a separate version of this network for five epochs for each video on the 2500 augmented images. We then use this network to generate accurate pixel mask proposals for each of the previously generated coarse proposals, by only taking the bounding box of these proposals as input into the Refinement network and discarding the coarse mask itself.

3.4   Mask Propagation using Optical Flow

As part of our proposal merging algorithm we use the optical flow between successive image pairs to warp a proposed mask into the next frame, to calculate the temporal consistency between two mask proposals. We calculate the Optical Flow using FlowNet 2.0 [12].
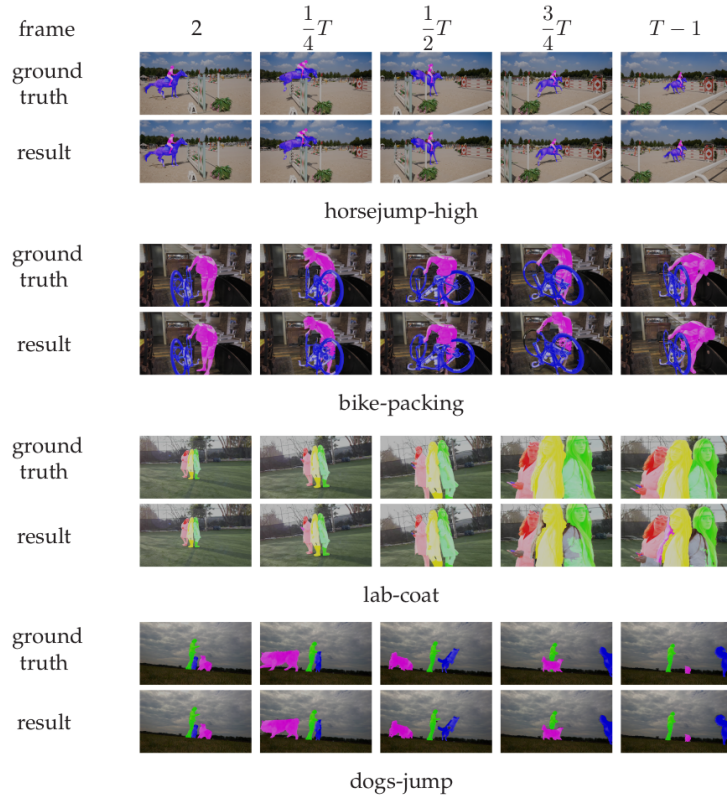
| frame | 2 | $\frac{1}{4}T$ | $\frac{1}{2}T$ | $\frac{3}{4}T$ | $T-1$ |

ground truth

result

horsejump-high

ground truth

result

bike-packing

ground truth

result

lab-coat

ground truth

result

dogs-jump

Fig. 3. Qualitative results of PReMVOS on the DAVIS 2017 `val dataset`.

3.5   ReID Embedding Vectors

We further use a triplet-loss based ReID embedding network to calculate a ReID embedding vector for each mask proposal. We use the feature embedding network proposed in [24]. This is based on a wide ResNet variant [32] pre-trained on ImageNet [6] and then trained on the COCO dataset [19] using cropped bounding boxes resized to 128×128 pixels. This uses a triplet loss to learn an embedding space in which crops of different classes are separated and crops of the same class are grouped together. It is trained using the batch-hard loss with a soft-plus margin proposed in [11]. We then fine-tune this network using the crops of each object from the generated 2500 augmented images for each of the 90 video sequences (242 objects) in the DAVIS 2017 `val`, `test-dev` and `test-challenge` datasets combined in order to have both enough positive and negative examples to train a network with a triplet-based loss. This network generates a ReID vector which differentiates all of the objects in these datasets from one other, which is used to compare the visual similarity of our generated object proposals and the first-frame ground truth object masks.

3.6   Proposal Merging

Our proposal merging algorithm works in a greedy manner. Starting from the ground truth object masks in the first-frame, it builds tracks for each frame by scoring each of the proposals on their likeliness to belong to a particular object track. We have exactly one track for each ground truth object and we make hard decisions for which proposals to add to each track each timestep in a greedy manner. The proposal with the highest track score is added to each track. This track score is calculated as an affine combination of five separate sub-scores, each with values between 0 and 1. In the following, taking the complement of a score means subtracting it from 1.

The first sub-score is the *Objectness* score. The objectness score $s_{obj,t,i,j}$ for the j-th track of the i-th proposal $c_{t,i}$ at time t is given by

$$s_{obj,t,i,j}(c_{t,i}) = MaskObj(c_{t,i}), \tag{1}$$

where $MaskObj(\cdot)$ denotes the confidence value provided by the Proposal Generation network.

The second score is a *ReID* score, calculated using the Euclidean distance between the first-frame ground truth ReID embedding vector $r(f_j)$ and the ReID embedding vector $r(c_{t,i})$ of the current mask proposal, where $r(\cdot)$ denotes applying the ReID network, $\|\cdot\|$ denotes the L2 norm, and f_j is the bounding box of the j-th ground truth object in the first frame. This distance is then normalized by dividing it by the maximum distance for all proposals in a video from the ground truth embedding vector of interest. The complement is then taken to convert from a distance into a similarity score.

$$s_{reid,t,i,j}(c_{t,i}, f_j) = 1 - \frac{\|r(c_{t,i}) - r(f_j)\|}{\max_{\tilde{t},\tilde{i}}\|r(c_{\tilde{t},\tilde{i}}) - r(f_j)\|} \tag{2}$$

The third score is a *Mask Propagation* score. This is calculated for each possible object track as the IoU between the current mask proposal and the warped proposal that was decided for in the previous time-step for this object track, warped into the current time-step using the optical flow:

$$s_{maskprop,t,i,j}(c_{t,i}, p_{t-1,j}) = IoU(c_{t,i}, warp(p_{t-1,j})), \tag{3}$$

where $p_{t-1,j}$ is is the previously selected proposal for timestep $t-1$ for object track j and $warp(\cdot)$ applies optical flow mask warping from frame $t-1$ to t.

The fourth score is an *Inverse ReID* score. This is calculated as the complement of the maximum *ReID* score for the current mask proposal and all other object tracks k except the object track of interest j:

$$s_{inv\_reid,t,i,j} = 1 - \max_{k \neq j}(s_{reid,t,i,k}). \tag{4}$$

The fifth score is an *Inverse Mask Propagation* score. This is calculated as the complement of the maximum *Mask Propagation IoU* score for the current mask proposal and all other object tracks k except the object track of interest j:

$$s_{inv\_maskprop,t,i,j} = 1 - \max_{k \neq j}(s_{maskprop,t,i,k}). \tag{5}$$

All five scores are combined together by

$$s_{comb,t,i,j} = \sum_{q \in \{objectness,reid,maskprop,inv\_reid,inv\_maskprop\}} \alpha_q s_{q,t,i,j}, \quad (6)$$

where $\sum_q \alpha_q = 1$ and all $\alpha_q \geq 0$. The greedy decisions are then made by $p_{t,j} = c_{t,k_j}$, where

$$k_j = \underset{i}{\mathrm{argmax}}\, s_{comb,t,i,j}. \quad (7)$$

In cases where the selected proposals for the different objects within one time-step overlap, we assign the overlapping pixels to the proposal with the highest combined track score. We present results with both an equal weighting for each of the five sub-score components and where the weights are tuned using random-search hyper-parameter optimisation evaluated against the DAVIS 2017 `val` set. The values of these optimised weights are shown in Table 4. We ran this optimisation for 25000 random parameter values. For the results on the 2018 DAVIS Challenge we also present results using an ensemble of the results using the top 11 sets of parameter values, using a simple pixel-wise majority vote to ensemble the results.

4 Experiments

We evaluate our algorithm on the set of DAVIS [3],[27],[26] datasets and benchmarks. Table 1 shows our results on the three DAVIS benchmarks. The DAVIS 2017 `test-dev` and `val` datasets contain multiple objects per video sequence, whereas the DAVIS 2016 `val` dataset contains a single object per sequence. The metrics of interest are the $\mathcal{J}$ score, calculated as the average IoU between the proposed masks and the ground truth mask, and the $\mathcal{F}$ score, calculated as an average boundary similarity measure between the boundary of the proposed masks and the ground truth masks. For more details on these metrics see [26].

On all of the datasets our method gives results better than all other state-of-the-art methods for both the $\mathcal{F}$ metric and the mean of the $\mathcal{J}$ and $\mathcal{F}$ score. We also produce either the best, or comparable to the best, results on the $\mathcal{J}$ metric for each dataset. These results show that the novel proposed VOS paradigm performs better than the current VOS paradigms in predicting both accurate and temporally consistent mask proposals.

Table 2 shows our results both with and without ensembling on the DAVIS 2017/2018 `test-challenge` dataset evaluated during the 2018 DAVIS Challenge compared to the top six other competing methods. Our method gives the best results and gets first place in the 2018 DAVIS Challenge.

Figure 3 shows qualitative results of our method on four video sequences from the `2017 val` dataset. These results show that our method produces both accurate and temporally consistent results across the video sequences.

			Ours	DyeNet [17]	MRF [1]	Lucid [15]	ReID [18]	OSVOS-S [21]	OnAVOS [29][30]	OSVOS [2]
	$\mathcal{J}\&\mathcal{F}$	Mean	**71.6**	68.2	67.5	66.6	66.1	57.5	56.5	50.9
17 T-D	$\mathcal{J}$	Mean	**67.5**	65.8	64.5	63.4	64.4	52.9	52.4	47.0
		Recall	**76.8**	-	-	73.9	-	60.2	-	52.1
		Decay	21.7	-	-	19.5	-	24.1	-	**19.2**
	$\mathcal{F}$	Mean	**75.7**	70.5	70.5	69.9	67.8	62.1	59.6	54.8
		Recall	**84.3**	-	-	80.1	-	70.5	-	59.7
		Decay	20.6	-	-	**19.4**	-	21.9	-	19.8
	$\mathcal{J}\&\mathcal{F}$	Mean	**77.8**	74.1	70.7	-	-	68.0	67.9	60.3
17 Val	$\mathcal{J}$	Mean	**73.9**	-	67.2	-	-	64.7	64.5	56.6
		Recall	**83.1**	-	-	-	-	74.2	-	63.8
		Decay	16.2	-	-	-	-	**15.1**	-	26.1
	$\mathcal{F}$	Mean	**81.7**	-	74.2	-	-	71.3	71.2	63.9
		Recall	**88.9**	-	-	-	-	80.7	-	73.8
		Decay	19.5	-	-	-	-	**18.5**	-	27.0
	$\mathcal{J}\&\mathcal{F}$	Mean	**86.8**	-	-	-	-	86.5	85.5	80.2
16 Val	$\mathcal{J}$	Mean	84.9	**86.2**	84.2	-	-	85.6	86.1	79.8
		Recall	96.1	-	-	-	-	**96.8**	96.1	93.6
		Decay	8.8	-	-	-	-	5.5	**5.2**	14.9
	$\mathcal{F}$	Mean	**88.6**	-	-	-	-	87.5	84.9	80.6
		Recall	94.7	-	-	-	-	**95.9**	89.7	92.6
		Decay	9.8	-	-	-	-	8.2	**5.8**	15.0
	$\mathcal{T}$	Mean	36.4	-	-	-	-	21.7	**19.0**	37.8

Table 1. Our results and other state-of-the-art results on the three DAVIS datasets: the 2017 `test-dev` set (17 T-D), the 2017 `val` set (17 Val), and the 2016 val set (16 Val). On the 17 Val and 16 Val datasets we use the naive merging component weights, whereas on the 17 T-D dataset we use the weights optimised using the 17 Val set.

4.1   Proposal Refinement

We perform an ablation study to investigate the effect of the Refinement network on the accuracy of the mask proposals. We compare the coarse proposals generated from a state-of-the-art instance segmentation method, Mask R-CNN [9], to the refined proposals generated by feeding the bounding boxes of these coarse proposals into our Refinement Network. In order to evaluate these results we calculate the best proposal for each object in each frame which maximizes the IoU score with the ground-truth mask. We then evaluate using the standard DAVIS evaluation metrics using these merged proposals. This *oracle merging* algorithm allows us to evaluate the accuracy of our proposal generation separately from our merging algorithm.

		Ours (Ens)	Ours	DyeNet [16]	ClassAgno. VOS [35]	OnlineGen. VOS [8]	Lucid [14]	ContextBased VOS [28]
$\mathcal{J}\&\mathcal{F}$	Mean	**74.7**	71.8	73.8	69.7	69.5	67.8	66.3
$\mathcal{J}$	Mean	71.0	67.9	**71.9**	66.9	67.5	65.1	64.1
	Recall	**79.5**	75.9	79.4	74.1	77.0	72.5	75.0
	Decay	19.0	23.2	19.8	23.1	15.0	27.7	**11.7**
$\mathcal{F}$	Mean	**78.4**	75.6	75.8	72.5	71.5	70.6	68.6
	Recall	**86.7**	82.9	83.0	80.3	82.2	79.8	80.7
	Decay	20.8	24.7	20.3	25.9	18.5	30.2	**13.5**

Table 2. Our results (with and without ensembling) on the DAVIS `test-challenge` dataset compared with the top five other competitors in the 2018 DAVIS Challenge.

	$\mathcal{J}$ mean	$\mathcal{F}$ mean	$\mathcal{J}\&\mathcal{F}$ mean
Without Refinement	71.2	77.3	74.2
With Refinement	**77.1**	**85.2**	**81.2**
Boost	5.9	7.9	7.0

Table 3. Quantitative results of an ablation study on the `2017 val` dataset showing the effect of the Refinement Network on the accuracy of generated mask proposals. Presented results are calculated using *oracle merging* (see Section 4.1).

Fig. 4. Qualitative results showing the effect of the Refinement Network on the mask proposal accuracy. Results are calculated using *oracle merging* (Section 4.1).

Table 3 shows the quantitative IoU ($\mathcal{J}$) and boundary measure ($\mathcal{F}$) improvement for our refined proposals over the Mask R-CNN proposals. Our method results in a 5.9% improvement in IoU and a 7.9% improvement in boundary measure over just using the Mask R-CNN proposals.

Figure 4 also visualizes the qualitative results of the improved accuracy of the refined masks over the generated Mask R-CNN masks. In all examples the refined proposals match the ground truth masks more closely than the coarse proposals and capture the boundary contours at a much higher fidelity. This is due to the refinement network extracting deep features only over the area of interest for each object, and not over the whole image as is done in the case for Mask R-CNN. The Refinement Network is also able to recover parts of objects that were completely lost in the coarse proposals, for example in the two examples of bicycles shown. This is because the Refinement Network takes as input only the bounding box of the coarse mask proposal, not the coarse mask itself, and relies only on it's trained knowledge of what is inherently an *object* in order to generate the refined segmentation masks. Also when the bounding box does not accurately cover the whole object, as is the case with the bicycle in the second column of Figure 4, the refinement network is still able to recover the accurate mask as it takes into account a 50 pixel margin around the coarse mask proposal bounding box.

4.2   Proposal Merging

We perform a further ablation study to investigate the effect of each of the merging algorithm sub-score components on the accuracy of the merging algorithm. Table 4 shows the results of this ablation study on the DAVIS `2017` `val` dataset.

We first present an upper bound baseline result for our merging algorithm. This is calculated by choosing the best proposal for each object in each frame which maximizes the IoU score with the ground-truth mask. This *oracle merging* method gives a 81.2 $\mathcal{J}\&\mathcal{F}$ mean score.

We then present the results of our merging algorithm where the weights for each of the five sub-score components were optimised using random-search hyper-parameter optimisation evaluated against the DAVIS `2017` `val` set. This optimisation was done by evaluating 25000 random affine combinations of the 5 component weights and selecting the set of weights that resulted in the best IoU score. This result gives another upper bound of 78.2 $\mathcal{J}\&\mathcal{F}$ mean score, as this was evaluated on the same dataset as the weight values were optimised on. Our greedy selection algorithm based on a combination of the 5 sub-component scores is able to reach a $\mathcal{J}\&\mathcal{F}$ mean score that is only 3% lower than the hypothetical maximum, showing that these carefully selected 5 sub-score components are sufficient to generate accurate and consistent object tracks, even in difficult cases such as multiple similar objects and large occlusions. These opimised weights used on the `2017` `test-dev` dataset presented in Table 1, all other dataset results in Table 1 and Table 4 use equal weights for the five sub-components.

The naive combination with all 5 sub-score components has a $\mathcal{J}\&\mathcal{F}$ mean score of 77.8 which is only 0.4 below that with optimised weights. This indicates that the merging algorithm is relatively robust to the exact weights and that what is more important is the presence of all five components.

The *Objectness* sub-score separates well-defined mask proposals with accurate boundary contours from proposals with boundaries that are less likely to

Num.	Merging Sub-Score Components					$\mathcal{J}\&\mathcal{F}$
Comp.	Objectness	ReID	InvReID	MaskProp	InvMaskProp	Mean
0	Oracle merging					**81.2**
5(opt.)	19%	18%	14%	22%	27%	**78.2**
5	✓	✓	✓	✓	✓	**77.8**
	✓	✓	✓	✓	-	76.7
	✓	✓	✓	-	✓	*75.5*
4	✓	✓	-	✓	✓	**76.9**
	✓	-	✓	✓	✓	76.3
	-	✓	✓	✓	✓	75.9
	✓	✓	✓	-	-	74.2
	✓	✓	-	✓	-	75.0
	✓	✓	-	-	✓	74.2
	✓	-	✓	✓	-	73.5
3	✓	-	✓	-	✓	69.6
	✓	-	-	✓	✓	71.1
	-	✓	✓	✓	-	75.8
	-	✓	✓	-	✓	*69.3*
	-	✓	-	✓	✓	**75.9**
	-	-	✓	✓	✓	74.3
	✓	✓	-	-	-	72.7
	✓	-	✓	-	-	64.7
	✓	-	-	✓	-	69.1
	✓	-	-	-	✓	57.9
	-	✓	✓	-	-	68.7
2	-	✓	-	✓	-	**74.3**
	-	✓	-	-	✓	68.8
	-	-	✓	✓	-	74.0
	-	-	✓	-	✓	*47.3*
	-	-	-	✓	✓	73.6
	✓	-	-	-	-	*29.5*
	-	✓	-	-	-	67.4
1	-	-	✓	-	-	44.3
	-	-	-	✓	-	**72.8**
	-	-	-	-	✓	34.4

Table 4. Results of an ablation study on the DAVIS `2017` `val` dataset showing the effect of each of the merging algorithm sub-score components on the accuracy of the merging algorithm. The *oracle merging* result indicates an upper bound for the merging algorithm performance (Section 4.1). The components given as percentages indicate the optimal component weights calculated using hyper-parameter optimisation on the `2017` `val` set. The components given by a checkmark (✓) have equal weights for each checked components. For each group of results with the same number of components the best result is expressed in **bold** and the worst in *italics*.

model a consistent object. It is also able to distinguish objects of interest given in the first frame from other objects in the scene, as this score was trained in the Mask R-CNN fine-tuning process to identify these objects and ignore the others. However, this sub-score component is unable to distinguish between different objects of interest if more than one is present in a video sequence.

The *ReID* sub-score is used to distinguish between objects that look visually different from each other. This works well to separate objects such as bikes and people from each other in the same video sequence, but does not work as well on sequences with multiple similar looking objects.

The *MaskProp* sub-score is used for temporal consistency. This can distinguish well between very similar objects if they are separated in the spatial domain of the image. However, this sub-score cannot deal with cases where objects heavily occlude each other, or completely disappear before later returning.

The *InvReID* and *InvMaskProp* sub-scores are used to force the selected mask proposals in each frame to be as distinguishable from each other as possible. Just using the other 3 components often results in failure cases where the same or very similar mask proposals are chosen for different objects. This occurs when similar looking objects overlap or when one object disappears. These two sub-components work by distinguishing proposals that are visually and temporally inconsistent with other objects in the video sequence, resulting in a signal of consistency with the object of interest. These components can separate well between the different objects of interest in a video sequence, but they are unable to separate these objects from possible background objects.

The results in Table 4 show that all five sub-scores are important for accurate proposal merging, as removing one of these components results in a loss of accuracy between 0.9, when only removing the *InvReID* sub-score, to 2.3, when removing the *MaskProp* sub-score. Without both the *InvReID* and *InvMaskProp*, the $\mathcal{J}\&\mathcal{F}$ mean score decreases by 1.9 points, showing that these sub-scores that were introduced to promote spatial separation of the chosen proposals are an integral part of the merging algorithm. Only using these two sub-components, however, results in a score decrease of 30.5 points. Removing the two main key components, the *ReID* and *MaskProp* sub-scores, results in a loss of 8.2 points, whereas only having these two components results in a loss of 3.5 points. When used by itself, the *MaskProp* sub-score is the strongest component, resulting in a loss of 5.0 points.

4.3   Runtime Evaluation

We perform a runtime evaluation of the different components of the PReMVOS algorithm. The complete PReMVOS algorithm with first-frame data augmentation and fine-tuning was designed in order to produce the most accurate results without regards for speed requirements. For further evaluation, we present in Table 5 three versions of PReMVOS. The original version, a fast-finetuned version, and a version without any fine-tuning. The fast-finetuned version is fine-tuned for one third of the iterations as the original method, and instead of the slow Lucid Data Dreaming [15] image augmentations, it uses simple rotations, translations,

	Augm. Gen.	Fine-tuning	Prop. Gen.	Prop. Refine.	ReID	Optic. Flow	Warping	Merging	**Total**	Av. # Prop.	Mean $\mathcal{J}\&\mathcal{F}$
Original	23.4	12.3	0.41	1.04	0.05	0.14	0.32	0.02	37.4	17.52	77.8
Fast-finetuned	0.02	3.9	0.26	0.45	0.03	0.14	0.20	0.02	5.02	9.28	73.7
Not-finetuned	0.00	0.00	0.14	0.33	0.02	0.14	0.16	0.02	0.81	6.87	65.7

Table 5. Runtime analysis of the different components of the PReMVOS algorithm. Times are in seconds per frame, averaged over the DAVIS 2017 val set. Augmentation Generation is run on 48 CPU cores, and Fine-tuning is done on 8 GPUs. Otherwise, everything is run sequentially on one GPU / CPU core.
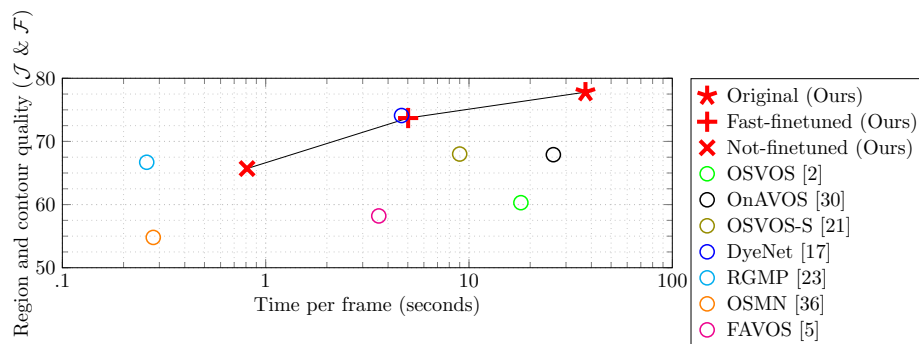

Fig. 5. Quality versus timing on the DAVIS 2017 val set. For methods that only publish runtime results on the DAVIS 2016 dataset, we take these timings as per object timings and extrapolate to the number of objects in the DAVIS 2017 val set.

flipping and brightness augmentations. Only one set of weights are fine-tuned over the whole DAVIS val set rather than different weights for each video. A combination of the *specific* proposals from the fine-tuned proposal network and the *general* proposals from a proposal network that was not fine-tuned on the validation set first-frames are used. The not-finetuned version of PReMVOS just uses the *general* proposals without any fine-tuning. Figure 5 compares the quality and runtime of PReMVOS against other methods in the literature. Across all of the presented runtime scales, our method compares to or exceeds all other state-of-the-art results.

4.4   Further Large-scale Evaluation

In Table 6 we present the results of PReMVOS on the new YouTube-VOS dataset [33], the largest VOS dataset to date (508 test video compared to 30 in DAVIS),

	Overall	$\mathcal{J}$ seen	$\mathcal{J}$ unseen	$\mathcal{F}$ seen	$\mathcal{F}$ unseen
Ours	**72.2**	**73.7**	64.8	**77.8**	72.5
Seq2Seq [33]	70.0	66.9	**66.8**	74.1	72.3
2nd	72.0	72.5	66.3	75.2	**74.1**
3rd	69.9	73.6	62.1	75.5	68.4
4th	68.4	70.6	62.3	72.8	67.7

Table 6. Results on the YouTube-VOS dataset, using the PReMVOS Fast-finetuned version of PReMVOS. These results obtained 1st place in the the 1st Large-scale Video Object Segmentation Challenge. '2nd','3rd','4th' refers to the other competitors results in this challenge with that ranking. Bold results are the best results for that metric.

our results on the test set obtained 1st place in the 1st Large-scale Video Object Segmentation Challenge. We don't run the original PReMVOS method but the PReMVOS Fast-finetuned version (see Section 4.3), which can be evaluated on this larger dataset in a more reasonable amount of time. Our results are much better than [33], the only other method that has published results on this dataset. It is also better than all other methods that submitted results to the 2017 1st Large-scale Video Object Segmentation Challenge.

5   Conclusion

In this paper we have presented a new approach for solving the video object segmentation task. Our proposed approach works by dividing this task into first generating a set of accurate object segmentation mask proposals for each video frame and then selecting and merging these proposals into accurate and temporally consistent pixel-wise object tracks over a video sequence. We have developed a novel approach for each of these sub-problems and have combined these into the PReMVOS (Proposal-generation, Refinement and Merging for Video Object Segmentation) algorithm. We show that this method is particularly well suited for the difficult multi-object video object segmentation task and that it produces results better than all current state-of-the-art results for semi-supervised video object segmentation on the DAVIS benchmarks, as well as getting the best score in the 2018 DAVIS Challenge.

References

1. Bao, L., Wu, B., Liu, W.: CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF. arXiv preprint arXiv:1803.09453 (2018)
2. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR (2017)
3. Caelles, S., Montes, A., Maninis, K.K., Chen, Y., Van Gool, L., Perazzi, F., Pont-Tuset, J.: The 2018 davis challenge on video object segmentation. arXiv preprint arXiv:1803.00557 (2018)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611 (2018)
5. Cheng, J., Tsai, Y.H., Hung, W.C., Wang, S., Yang, M.H.: Fast and accurate online video object segmentation via tracking parts. In: CVPR (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
7. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (2010)
8. Guo, P., Zhang, L., Zhang, H., Liu, X., Ren, H., Zhang, Y.: Adaptive video object segmentation with online data generation. The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2018)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
12. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR (2017)
13. J. Luiten, P. Voigtlaender, B.L.: PReMVOS: Proposal-generation, Refinement and Merging for the YouTube-VOS Challenge on Video Object Segmentation 2018. The 1st Large-scale Video Object Segmentation Challenge - ECCV Workshops (2018)
14. Khoreva, A., Benenson, R., Ilg, E., Brox, T., Schiele, B.: Lucid data dreaming for video object segmentation. The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2018)
15. Khoreva, A., Benenson, R., Ilg, E., Brox, T., Schiele, B.: Lucid data dreaming for multiple object tracking. The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2017)
16. Li, X., Loy, C.C.: Video object segmentation with joint re-identification and attention-aware mask propagation. The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2018)
17. Li, X., Change Loy, C.: Video object segmentation with joint re-identification and attention-aware mask propagation. In: ECCV (September 2018)
18. Li, X., Qi, Y., Wang, Z., Chen, K., Liu, Z., Shi, J., Luo, P., Tang, X., Loy, C.C.: Video object segmentation with re-identification. The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2017)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
20. Luiten, J., Voigtlaender, P., Leibe, B.: Premvos: Proposal-generation, refinement and merging for the davis challenge on video object segmentation 2018. The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2018)

21. Maninis, K.K., Caelles, S., Chen, Y., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: Video object segmentation without temporal information. PAMI (2017)
22. Neuhold, G., Ollmann, T., Bulo, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017)
23. Oh, S., Lee, J., Sunkavalli, K., Kim, S.: Fast video object segmentation by reference-guided mask propagation. In: CVPR (2018)
24. Ošep, A., Voigtlaender, P., Luiten, J., Breuers, S., Leibe, B.: Large-scale object discovery and detector adaptation from unlabeled video. arXiv preprint arXiv:1712.08832 (2017)
25. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., A.Sorkine-Hornung: Learning video object segmentation from static images. In: CVPR (2017)
26. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
27. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
28. Tran, M., Ton-That, V., Le, T., Nguyen, K., Ninh, T.V., T. Le, V.N., Nguyen, T.V., Do, M.N.: Context-based instance segmentation in video sequences. The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2018)
29. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for the 2017 davis challenge on video object segmentation. The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2017)
30. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. In: BMVC (2017)
31. Wu, Y., et al.: Tensorpack. https://github.com/tensorpack/ (2016)
32. Wu, Z., Shen, C., van den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. arXiv preprint arXiv:1611.10080 (2016)
33. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. arXiv preprint arXiv:1809.00461 (2018)
34. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.: Deep grabcut for object selection. In: BMVC (2017)
35. Xu, S., Bao, L., Zhou, P.: Class-agnostic video object segmentation without semantic re-identification. The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2018)
36. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.: Efficient video object segmentation via network modulation. CVPR (2018)