PReMVOS: Proposal-generation, Refinement and Merging for the DAVIS Challenge on Video Object Segmentation 2018

Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe
Computer Vision Group
Visual Computing Institute
RWTH Aachen University, Germany
jonathon.luiten@rwth-aachen.de {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). This method involves generating coarse object proposals using a Mask R-CNN like 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, and a Re-ID feature embedding vector for each 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. Our approach surpasses all previous state-of-the-art results on the DAVIS 2017 video object segmentation benchmark and achieves first place in the DAVIS 2018 Video Object Segmentation Challenge with a mean of $\mathcal{J}$ & $\mathcal{F}$ score of 74.7.

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 for the video sequence. This can be seen as a combination and extension of both instance segmentation from single frames to videos, and multi object tracking from tracking bounding boxes to tracking pixel masks. VOS has a multitude of applications in robotics, self driving cars, and any other application which benefits from understanding the composition of a video. 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 [1, 2, 3] present a state-of-the-art testing ground for this task. In this paper we present a new paradigm for tackling the semi-supervised VOS task and evaluate the corresponding algorithm on the set of DAVIS dataset benchmarks. 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.

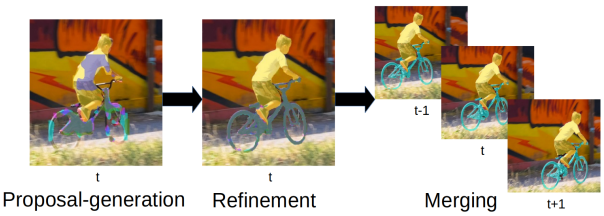

Figure 1. PReMVOS methodology overview.

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 fine-tuning. This approach, first proposed in [4], 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 [5] and [6, 7] by using semantic segmentation guidance and iterative fine-tuning, respectively. The second paradigm, used in several state-of-the-art methods [8, 9, 10, 11], 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 [9] and [11] 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. [10] im-

proves upon the mask propagation paradigm by training on a huge set of augmented images generated from the first-frame ground truth. Our method adopts parts of the ideas presented in all of the above papers, but combines them with a new paradigm for tackling the VOS task.

3. Approach

We propose PReMVOS as a new paradigm for addressing the VOS task. This approach is designed to produce more accurate and more temporally consistent pixel masks across a video. Instead of predicting object masks directly on the video pixels, as done in [4, 5, 6, 7], a key idea of our approach is to instead detect regions of interest as coarse object proposals using an object detection and mask network, and to then predict accurate object 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, especially in the multi-object VOS scenario. The methods presented in [8, 9, 10, 11] 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 through the video 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 new VOS paradigm, as shown in Figure 1, 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. Below we describe our implementation of this new VOS paradigm.

Image Augmentation. For each video we generate a set of 2500 augmented images using the first-frame ground truth. We use the method in [10] 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.

Proposal Generation. We generate coarse object proposals using a Mask R-CNN [12] network implementation by [13] with a ResNet101 [14] backbone. We adjust this network to be category agnostic by replacing the N classes with just one class for detecting generic objects. We train this network starting from pre-trained ImageNet [15] weights on both the COCO [16] and Mapillary [17] datasets jointly. We then fine-tune a separate version of this network for each video for three epochs of the 2500 augmented images. This takes around one hour on a GTX 1080 Ti GPU. 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.

Proposal Refinement. The Proposal-Refinement Network is a fully convolutional network inspired by [18] and based on the DeepLabv3+ [19] 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 [15], COCO [16], and PASCAL [20], we train this network on the Mapillary [17] dataset using random flipping, random gamma augmentations and random bounding box jitter [18] 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 each video on the 2500 augmented images. We fine-tune for five epochs of the 2500 images, which takes around twenty minutes per object in the video on a GTX 1080 Ti GPU. We then use this network to generate accurate pixel mask proposals for each of the coarse bounding box proposals previously generated by the Proposal-Generation Network.

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 [21].

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 [23]. This is based on a wide ResNet variant [24] pre-trained on ImageNet [15] and then trained on the COCO dataset [16] 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 [25]. 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. This trains this network to be able to generate a ReID vector which separates all the possible objects of interest in these three datasets from each other. We use this network to calculate

			Ours (Ens)	Ours	Lixx	Dawns	ILC_R	Apata	UIT	DyeNet [11]	MRF [22]	Lucid [10]	ReID [9]	OSVOS-S [5]	OnAVOS [6][7]	OSVOS [4]
17/18 T-C	$\mathcal{J}\&\mathcal{F}$	Mean	**74.7**	71.8	73.8	69.7	69.5	67.8	66.3	-	-	67.8	69.9	-	57.7	-
	$\mathcal{J}$	Mean	71.0	67.9	**71.9**	66.9	67.5	65.1	64.1	-	-	65.1	67.9	-	54.8	-
		Recall	**79.5**	75.9	79.4	74.1	77.0	72.5	75.0	-	-	72.5	74.6	-	60.8	-
		Decay	19.0	23.2	19.8	23.1	15.0	27.7	**11.7**	-	-	27.7	22.5	-	31.0	-
	$\mathcal{F}$	Mean	**78.4**	75.6	75.8	72.5	71.5	70.6	68.6	-	-	70.6	71.9	-	60.5	-
		Recall	**86.7**	82.9	83.0	80.3	82.2	79.8	80.7	-	-	79.8	79.1	-	67.2	-
		Decay	20.8	24.7	20.3	25.9	18.5	30.2	**13.5**	-	-	30.2	24.1	-	34.7	-
17 T-D	$\mathcal{J}\&\mathcal{F}$	Mean	**71.9**	71.6	-	-	-	-	-	68.2	67.5	66.6	66.1	57.5	56.5	50.9
	$\mathcal{J}$	Mean	**67.7**	67.5	-	-	-	-	-	65.8	64.5	63.4	64.4	52.9	52.4	47.0
		Recall	**77.1**	76.8	-	-	-	-	-	-	-	73.9	-	60.2	-	52.1
		Decay	21.0	21.7	-	-	-	-	-	-	-	19.5	-	24.1	-	**19.2**
	$\mathcal{F}$	Mean	**76.1**	75.7	-	-	-	-	-	70.5	70.5	69.9	67.8	62.1	59.6	54.8
		Recall	**84.7**	84.3	-	-	-	-	-	-	-	80.1	-	70.5	-	59.7
		Decay	19.7	20.6	-	-	-	-	-	-	-	**19.4**	-	21.9	-	19.8
17 Val	$\mathcal{J}\&\mathcal{F}$	Mean	**78.2**	**78.2**	-	-	-	-	-	74.1	70.7	-	-	68.0	67.9	60.3
	$\mathcal{J}$	Mean	**74.3**	**74.3**	-	-	-	-	-	-	67.2	-	-	64.7	64.5	56.6
		Recall	**83.5**	**83.5**	-	-	-	-	-	-	-	-	-	74.2	-	63.8
		Decay	16.0	16.0	-	-	-	-	-	-	-	-	-	**15.1**	-	26.1
	$\mathcal{F}$	Mean	**82.2**	**82.2**	-	-	-	-	-	-	74.2	-	-	71.3	71.2	63.9
		Recall	**89.6**	**89.6**	-	-	-	-	-	-	-	-	-	80.7	-	73.8
		Decay	18.5	**18.4**	-	-	-	-	-	-	-	-	-	18.5	-	27.0
16 Val	$\mathcal{J}\&\mathcal{F}$	Mean	87.0	**87.1**	-	-	-	-	-	-	-	-	-	86.5	85.5	80.2
	$\mathcal{J}$	Mean	85.5	85.5	-	-	-	-	-	**86.2**	84.2	-	-	85.6	86.1	79.8
		Recall	96.4	96.7	-	-	-	-	-	-	-	-	-	**96.8**	96.1	93.6
		Decay	10.4	9.1	-	-	-	-	-	-	-	-	-	5.5	**5.2**	14.9
	$\mathcal{F}$	Mean	88.6	**88.7**	-	-	-	-	-	-	-	-	-	87.5	84.9	80.6
		Recall	94.5	94.5	-	-	-	-	-	-	-	-	-	**95.9**	89.7	92.6
		Decay	12.4	10.8	-	-	-	-	-	-	-	-	-	8.2	**5.8**	15.0

Table 1. Our results (with and without ensembling) compared to state-of-the-art results on the four DAVIS benchmark datasets: the 2017/2018 DAVIS `test-challenge` set (17/18 T-C), the 2017 `test-dev` set (17 T-D), the 2017 `val` set (17 Val), and the 2016 val set (16 Val). Methods with citation are from the literature, methods without are the top five other competitors in the 2018 DAVIS Challenge.

a ReID embedding vector for each of our generated object proposals and also for each of the first-frame ground truth object masks.

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. The proposal with the highest track score is then 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 given by the Proposal-Generation network. The second score is a *Mask Propagation IoU* 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 timestep for this object track, warped into the current timestep using the optical flow. The third score is an *Inverse Mask Propagation IoU* score. This is calculated as the complement of the maximum *Mask Propagation IoU* score for the current mask proposal and all other object tracks except the object track of interest. The fourth score is a *ReID* score, calculated using the Euclidean distance between the first-frame ground truth ReID embedding vector and the ReID embedding vector of the current mask proposal. 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. The fifth 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 except the object track of interest. In cases where the selected proposals for the different objects within one timestep overlap, we assign the overlapping pixels to the proposal with the highest combined track score. The weighting for each of the five scores was tuned using random-search hyperparam-

eter optimization evaluated against the DAVIS 2017 `val` set. We ran the optimization for 25000 random parameter values. We present two versions of our algorithm, one using the best parameter values, and one 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 [1, 2, 3] datasets and benchmarks. Table 1 shows our results on the four DAVIS benchmarks. The DAVIS 2017 `test-challenge`, `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 [3]. We present results from our method both with and without ensembling. 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. This method was also used to win the 2018 DAVIS Challenge. 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.

5. Conclusion

In this paper we present the PReMVOS algorithm, a new paradigm for video object segmentation based on proposal-generation, refinement and merging. We show that this method produces results better than all current state-of-the-art results for multi-object semi-supervised video object segmentation on the DAVIS benchmarks, as well as getting the best score in the 2018 DAVIS Challenge.

Acknowledgments. This project was funded, in parts, by ERC Consolidator Grant DeeVise (ERC-2017-COG-773161).

References

[1] 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*, 2018.

[2] 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 preprint arXiv:1704.00675*, 2017.

[3] 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 *CVPR*, 2016.

[4] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *CVPR*, 2017.

[5] 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," *PAMI*, 2017.

[6] P. Voigtlaender and B. Leibe, "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.

[7] P. Voigtlaender and B. Leibe, "Online adaptation of convolutional neural networks for video object segmentation," in *BMVC*, 2017.

[8] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A.Sorkine-Hornung, "Learning video object segmentation from static images," in *CVPR*, 2017.

[9] 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," *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017.

[10] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele, "Lucid data dreaming for multiple object tracking," *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017.

[11] X. Li and C. C. Loy, "Video object segmentation with joint re-identification and attention-aware mask propagation," *arXiv preprint arXiv:1803.04242*, 2018.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017.

[13] Y. Wu *et al.*, "Tensorpack." `https://github.com/tensorpack/`, 2016.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[17] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017.

[18] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang, "Deep grabcut for object selection," in *BMVC*, 2017.

[19] 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.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[21] 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 *CVPR*, 2017.

[22] L. Bao, B. Wu, and W. Liu, "CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF," *arXiv preprint arXiv:1803.09453*, 2018.

[23] A. Ošep, P. Voigtlaender, J. Luiten, S. Breuers, and B. Leibe, "Large-scale object discovery and detector adaptation from unlabeled video," *arXiv preprint arXiv:1712.08832*, 2017.

[24] Z. Wu, C. Shen, and A. v. d. Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *arXiv preprint arXiv:1611.10080*, 2016.

[25] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.