Combining PReMVOS with Box-Level Tracking for the 2019 DAVIS Challenge

Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe
Computer Vision Group, RWTH Aachen University
{luiten, voigtlaender, leibe}@vision.rwth-aachen.de

Abstract

Recently a number of different approaches have been proposed for tackling the task of Video Object Segmentation (VOS). In this paper we compare and contrast two particularly powerful methods, PReMVOS (Proposal-generation, Refinement and Merging for VOS), and BoLTVOS (Box-Level Tracking for VOS). PReMVOS follows a tracking-by-detection framework in which a set of object proposals are generated per frame and are then linked into tracks over time by optical flow and appearance similarity cues. In contrast, BoLTVOS uses a Siamese architecture to directly detect the object to be tracked based on its similarity to the given first-frame object. Although BoLTVOS can outperform PReMVOS when the number of objects to be tracked is small, it does not scale as well to tracking multiple objects. Finally we develop a model which combines both BoLTVOS and PReMVOS and achieves a $\mathcal{J}\&\mathcal{F}$ score of 76.2% on the DAVIS 2017 test-challenge benchmark, resulting in a 2nd place finish in the 2019 DAVIS challenge on semi-supervised VOS.

1. Introduction

Semi-supervised Video Object Segmentation (VOS) is the task of producing segmentation masks for a set of objects in each frame of a video given a set of ground truth object masks in the first frame. In this paper we compare and contrast two powerful methods for VOS, PReMVOS [8, 7] (Proposal-generation, Refinement and Merging for VOS), and BoLTVOS [18] (Box-Level Tracking for VOS), and develop a model which combines both of these methods.

To evaluate these methods, we use the DAVIS datasets [13, 14]. This is a collection of datasets for video object segmentation with differing difficulty and different average number of objects per video (between 1 and 3). The DAVIS17 `test-challenge` dataset is the most difficult of these, and is used to evaluate the state-of-the-art video object segmentation algorithms in a yearly challenge. Our presented method, a combination of PReMVOS and BoLTVOS achieved second place in the 2019 challenge with a mean $\mathcal{J}\&\mathcal{F}$ score of 76.2%, which is only 0.5 percentage points below the winning method (*c.f.* Fig. 1).

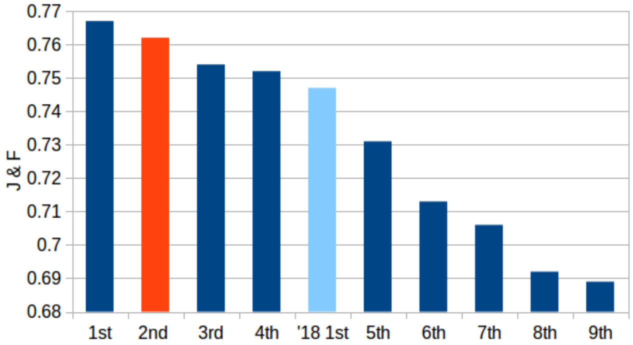

Figure 1. Semi-supervised DAVIS Challenge 2019 results. Our combination of BoLTVOS [18] and PReMVOS [8] (red) obtains 2nd place, improving 1.5 percentage points over PReMVOS alone (light blue) which won the 2018 challenge.

PReMVOS [8, 7] works in three steps which can be seen in Figure 3. First a large number of object segmentation proposals are generated from a Mask R-CNN-like [2] class-agnostic instance segmentation network. These proposals are then refined by a fully convolutional network to produce accurate segmentation masks. Finally these proposals are selected for each object in each frame using a merging algorithm that takes into account temporal consistency with optical-flow warping, visual consistency with a re-identification network, an objectness score from the proposal generation network and interactions between object tracks. The networks are all fine-tuned on a large collection of images generated from augmentations of the given first-frame using the Lucid data dreaming approach [5]. PReMVOS won both the 2018 DAVIS Challenge and the 2018 YouTube-VOS challenge.

BoLTVOS [18], as seen in Figure 2 takes an inherently different approach than PReMVOS. BoLTVOS consists of a Siamese network that directly detects the object to be tracked by conditioning the detection on the given object in the first frame. Potential objects in each frame are then re-scored using a tracklet-based temporal consistency algorithm. Finally, masks are produced by the same bounding-

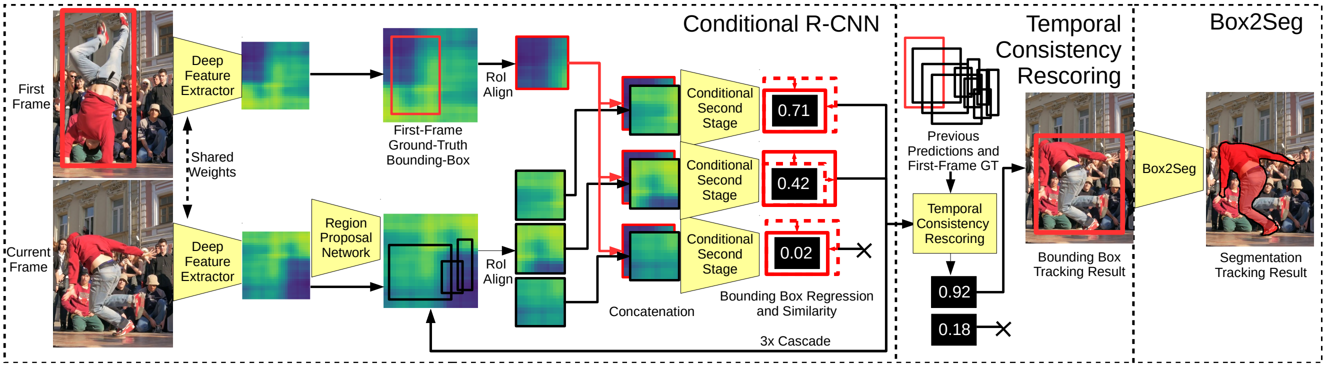

Figure 2. Overview of BoLTVOS[18]. A conditional R-CNN (left) provides detections conditioned on the first-frame bounding box, which are then rescored by a temporal consistency rescoring algorithm (center). The result are bounding box level tracks which are converted to segmentation masks by the Box2Seg network (right).

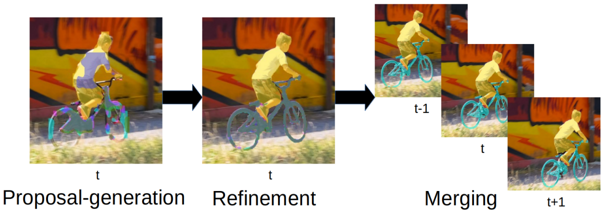

Figure 3. Overview of PReMVOS [8].

box-to-segmentation network (Box2Seg) that is also used in PReMVOS. BoLTVOS runs up to 45 times faster than PReMVOS. Moreover, it can produce accurate VOS results using only the first-frame bounding box, without using the given first-frame mask, although the first-frame mask can still be used by fine-tuning the segmentation network. As well as being an extremely strong VOS method, BoLTVOS is evaluated on the Visual Object Tracking (VOT) task, where it is currently the best-performing method on both the OTB2015 [20] and the LTB35 [9] benchmarks.

PReMVOS and BoLTVOS are currently two of the best-performing VOS algorithms. Thus it is interesting to compare how they perform across different VOS scenarios. Our experiments show that BoLTVOS can outperform PReMVOS when the number of objects to track is small, but that it is not able to perform as well when the number of objects in the video becomes much larger (see Section 3).

For our DAVIS Challenge 2019 entry, we combine both PReMVOS and BoLTVOS. We experimented with ensembles of results of different methods. For our final results we use PReMVOS as the base tracking algorithm as it deals better with tracking multiple objects. We then used the BoLTVOS conditional R-CNN to reject false positive tracking results and re-generated the given segmentation masks with BoLTVOS's Box2Seg network.

2. Method

Fine-tuning Image Augmentation. For both PReMVOS and BoLTVOS we train our network on images generated by augmenting the first-frame image with the given object masks. For each video we generate a set of 2500 augmented images using the method in [5] (only single images, not image pairs). This method removes the objects, fills in the background, randomly transforms each object and the background, and then reassembles the objects in the scene.

PReMVOS. An overview of the PReMVOS algorithm can be found in Figure 3, and more details can be found in [8, 7].

The first step of PReMVOS is to generate a number of coarse object proposals using a Mask R-CNN network. We adjust this network to be category agnostic by combining the N classes into just one class for detecting generic objects. We fine-tune a separate version of this network for each video on the augmented images. This network generates coarse mask proposals with bounding boxes and objectness scores for each image in the video sequence.

Next a Proposal-Refinement Network is used to create accurate segmentation masks for each proposal. This is a fully convolutional network which takes as input an image patch that has been cropped and resized from the bounding box of a proposal with a small padding region. This network is also fine-tuned per video on the augmented images.

After collecting a set of proposals with accurate segmentation masks, we select a proposal in each frame for each object that we wish to track. We do this using a merging algorithm that uses the objectness score from the proposal-generation network, optical-flow, and a per-proposal re-identification embedding from a Re-ID network.

We calculate the optical flow between successive image pairs using FlowNet 2.0 [4] to warp a proposed mask into the next frame and to calculate the temporal consistency between two mask proposals.

We use a triplet-loss based ReID embedding network to calculate a ReID embedding vector for each mask proposal.

In particular, we use the feature embedding network proposed in [11] and fine-tune it using the crops of each object from the generated images for each of the 90 video sequences (242 objects) in the DAVIS 2017 datasets. This trains this network to be able to generate a ReID vector which separates all the possible objects of interest from each other. We use this network to calculate a ReID embedding vector for each of our generated object proposals and also for each of the first-frame ground truth object masks.

The proposal merging algorithm works in a greedy manner. Starting from the ground truth masks in the first frame, it builds tracks for each frame by scoring each of the proposals based on their likelihood 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. The five scores are 1) an *Objectness* score from the proposal network, 2) a *Mask Propagation IoU* score from the optical-flow warped mask IoU, 3) an *Inverse Mask Propagation IoU* score which is the complement of the maximum mask-warped IoU for other objects to be tracked, 4) a *ReID* score from the Euclidean distance of a proposal's ReID embedding to that of the first frame, and 5) an *Inverse ReID* score which is the complement of the maximum ReID score for all other objects to be tracked.

BoLTVOS. As shown in Fig. 2, BoLTVOS consists of three components (see [18] for more details): 1) A conditional Siamese R-CNN detector, which detects object regions that are visually similar to the given first-frame template object. 2) An online temporal consistency rescoring algorithm that is able to choose the best detection that comes from the detector in each time step based on temporal consistency and visual similarity cues. 3) A Box2Seg network that generates a segmentation mask given a bounding box.

For the conditional detector, we base the architecture on the two-stage detection architecture of Mask R-CNN [2]. We take a pre-trained Mask R-CNN architecture, fixing the weights of the backbone and the RPN and replacing the category-specific second stage with a conditional second stage. This second stage is run for each region proposed by the RPN. To this end, we extract deep features from the proposed region and concatenate these with the deep features of the ground truth bounding box in the first-frame image, followed by a 1×1 convolution to reduce the feature dimension by half. The result is then fed into a cascaded R-CNN second stage with two output classes; either the proposed region is the object to be detected or it is not. The second stage is trained for tracking using pairs of frames from video datasets. Here, an object in one frame is used as reference and the network is trained to detect the same object in another frame. After detecting regions that are visually similar to the first-frame template with the conditional R-CNN, a temporal consistency rescoring algorithm is used to rescore the detections. This algorithm creates tracklets in an online manner by adding a detection to an existing tracklet each frame if it has an IoU with the last detection of a tracklet greater than a threshold (around 70%). Each detection that does not join an existing tracklet creates a new tracklet. The algorithm then finds the optimal set of tracklets which make up the final tracking result. It does this by scoring a number of 'track hypotheses' [12], different combinations of tracklets, using an online dynamic programming formalization. Tracklets that have overlapping time-steps cannot be composed together. Segmentation masks are then generated for BoLTVOS using the Box2Seg network, which is the same as the refinement network for PReMVOS except that fine-tuning is performed per object rather than per video.

Combining BoLTVOS and PReMVOS. We investigate two methods for combining BoLTVOS and PReMVOS. In the first method, we collect all of the proposals from both PReMVOS and BoLTVOS with their scores for each object and convert them to bounding boxes. We merge the proposals from each of the two methods if the box IoU is greater than 0.7. When merging two proposals, we add their scores together. Afterwards, we select the proposal with the highest score per object per frame and generate masks for them using the Box2Seg network. Finally, we ensemble this result with the original PReMVOS result and the original BoLTVOS result using a per pixel majority vote.

The second method uses PReMVOS first to generate an initial track estimate (as PReMVOS worked better for the multi-object tracking case). We then feed each of the selected proposals from PReMVOS into the BoLTVOS conditional R-CNN and calculate an additional score which we use to reject false positives from PReMVOS if the score from BoLTVOS is less than 0.02. Finally, we re-generate the segmentation masks using BoLTVOS's Box2Seg network. This second method performed much better than the ensemble based method.

3. Evaluation

Figure 4 compares BoLTVOS and PReMVOS across the YouTube-VOS dataset [22] and three different DAVIS datasets where the average number of objects per video varies. In benchmarks where only a single object (DAVIS16 `val`) or on average 1.8 objects per video (YouTube-VOS `val`) need to be segmented, BoLTVOS outperforms PReMVOS. When on average 2 objects per video need to be segmented (DAVIS17 `val`), PReMVOS slightly outperforms BoLTVOS. When on average 3 objects per video need to be segmented (DAVIS17 `test-challenge`), PReMVOS outperforms BoLTVOS by a large margin. This indicates that BoLTVOS, which was designed to perform well on the single object VOT task struggles to produce accurate results when multiple very similar objects need to be tracked simultaneously (for example a pod of 7 dolphins).

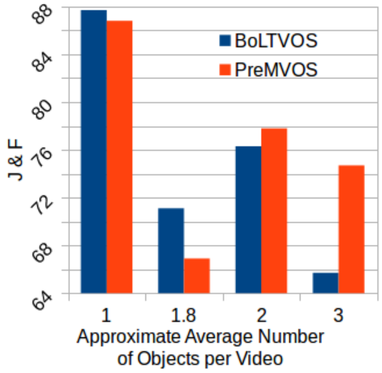

Figure 4. Comparison of BoLTVOS and PReMVOS on 4 different datasets with different average number of objects per video (1: DAVIS16 `val`, 1.8: YouTube-VOS `val`, 2: DAVIS17 `val`, 3: DAVIS17 `test-dev`).

In the following, we evaluate BoLTVOS, PReMVOS, and their combinations on the DAVIS17 `test-challenge` dataset, as part of the 2019 semi-supervised DAVIS Challenge. Our combination of BoLTVOS and PReMVOS obtains 76.2% mean $\mathcal{J}\&\mathcal{F}$ on this dataset, and reaches the second place in the challenge, as can be seen in Figure 1. This is 1.5 percentage points higher than PReMVOS alone with 74.7%, and 10.5 percentage points higher than BoLTVOS alone with 65.7%. Our alternative method of combining BoLTVOS and PReMVOS, which involved ensembling, obtained 72.6%, worse even than PReMVOS alone.

Figure 5 shows the performance and timing of BoLTVOS and PReMVOS compared to other state-of-the-art VOS methods on the DAVIS17 `val` benchmark. Both BoLTVOS (fine-tuned) and PReMVOS outperform all other methods, while BoLTVOS (fine-tuned) is 26 times faster than PReMVOS on this benchmark. Furthermore, the non fine-tuned version of BoLTVOS does not use the first-frame mask at all while still outperforming nearly all other VOS methods and also being faster.

Acknowledgements: This project has been funded, in parts, by ERC Consolidator Grant DeeViSe (ERC-2017-COG-773161) and by a Google Faculty Research Award.

References

[1] 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.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[3] Y.-T. Hu, J.-B. Huang, and A. G. Schwing. Videomatch: Matching based video object segmentation. In *ECCV*, 2018.

[4] 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.

[5] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for video object segmentation. *IJCV*, 2019.

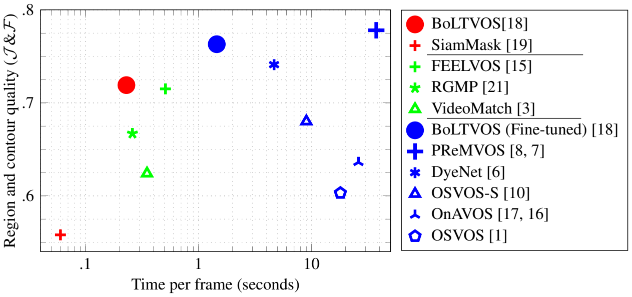

Figure 5. Quality versus timing plot comparing BoLTVOS and PReMVOS to other state-of-the-art methods on DAVIS17 `val`. Only SiamMask [19] and BoLTVOS (red) are able to work without the ground truth mask of the first frame and require just the bounding box. Methods shown in blue fine-tune on the first-frame mask.

[6] X. Li and C. Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, 2018.

[7] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS: Proposal-generation, Refinement and Merging for the DAVIS Challenge on Video Object Segmentation 2018. *CVPRW*, 2018.

[8] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018.

[9] A. Lukezic, L. C. Zajc, T. Vojír, J. Matas, and M. Kristan. Now you see me: evaluating performance in long-term visual tracking. *arXiv:1804.07056*, 2018.

[10] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. L. Taixé, and L. Van Gool. Video object segmentation without temporal information. *PAMI*, 2018.

[11] A. Ošep, P. Voigtlaender, J. Luiten, S. Breuers, and B. Leibe. Large-scale object mining for object discovery from unlabeled video. *ICRA*, 2019.

[12] A. Ošep, P. Voigtlaender, M. Weber, J. Luiten, and B. Leibe. 4D generic video object proposals. *arXiv:1901.09260*, 2019.

[13] 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.

[14] 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.

[15] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen. FEELVOS: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019.

[16] P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for the 2017 DAVIS challenge on video object segmentation. *CVPRW*, 2017.

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

[18] P. Voigtlaender, J. Luiten, and B. Leibe. BoLTVOS: Box-Level Tracking for Video Object Segmentation. *arXiv:1904.04552*, 2019.

[19] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019.

[20] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *PAMI*, 2015.

[21] S. Wug Oh, J.-Y. Lee, K. Sunkavalli, and S. Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, 2018.

[22] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *ECCV*, 2018.