Exploring the Combination of PReMVOS, BoLTVOS and UnOVOST for the 2019 YouTube-VOS Challenge

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

Abstract

Video Object Segmentation is the task of tracking and segmenting objects in a video given the first-frame mask of objects to be tracked. There have been a number of different successful paradigms for tackling this task, from creating object proposals and linking them in time as in PReMVOS, to detecting objects to be tracked conditioned on the given first-frame as in BoLTVOS, and creating tracklets based on motion consistency before merging these into long-term tracks as in UnOVOST. In this paper we explore how these three different approaches can be combined into a novel Video Object Segmentation algorithm. We evaluate our approach on the 2019 Youtube-VOS challenge where we obtain 6th place with an overall score of 71.5%.

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 present a method for the semi-supervised VOS track of the 2nd Large-scale Video Object Segmentation Challenge (also known as YouTube-VOS challenge), for which we achieved the 6th place.

Our method is based on ideas and components of three powerful recent methods. These are PReMVOS [10] (Proposal-generation, Refinement and Merging for VOS), BoLTVOS [16] (Box-Level Tracking for VOS), and UnOVOST [19] (Unsupervised Offline VOS and Tracking).

PReMVOS. PReMVOS [10] works in three steps which can be seen in Figure 2. First a large number of object segmentation proposals are generated from a Mask R-CNN-like [4] 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 [6]. PReMVOS won both the 2018 DAVIS Challenge [8] and the 2018 YouTube-VOS challenge [9].

BoLTVOS. BoLTVOS [16], as seen in Figure 1 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-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 [17] and the LTB35 [12] benchmarks.

UnOVOST. UnOVOST [19] addresses unsupervised video object segmentation, *i.e.*, VOS without using the first-frame ground truth masks. It first extracts mask proposals using Mask R-CNN [4]. Afterwards, it sub-selects and clips these proposals to obtain non-overlapping masks. In the next step, masks are linked together over time into short tracklets using spatio-temporal mask consistency. In a final step, UnOVOST uses a Forest Path Cutting (FPC) data association algorithm to combine tracklets into full tracks. UnOVOST achieved the first place in the unsupervised track of the DAVIS 2019 competition.

PReMVOS and BoLTVOS. PReMVOS and BoLTVOS are currently two of the best-performing semi-supervised VOS algorithms. For our second-place DAVIS Challenge

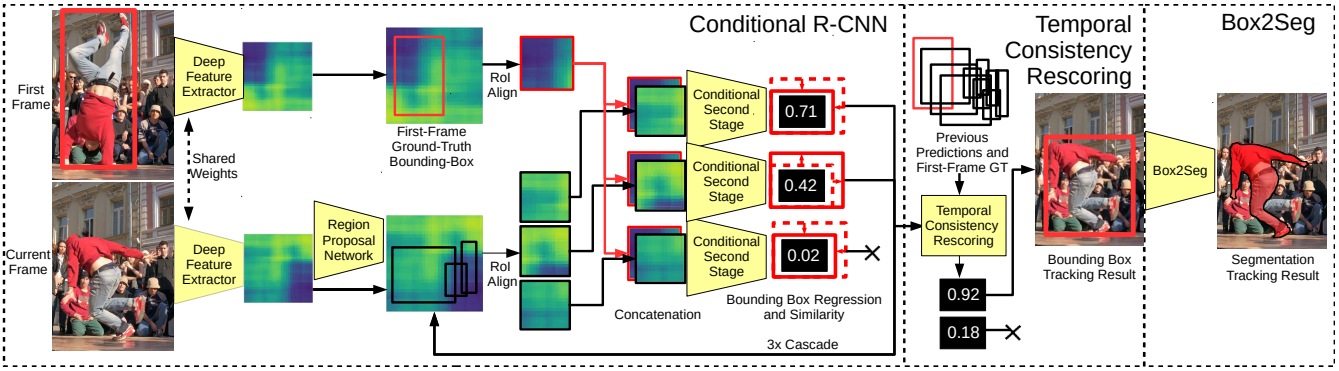

Figure 1. Overview of BoLTVOS [16]. A conditional R-CNN (left) provides detections conditioned on the first-frame bounding box, which are then rescored by a temporal consistency rescoring algorithm (center). The result are bounding box level tracks which are converted to segmentation masks by the Box2Seg network (right).

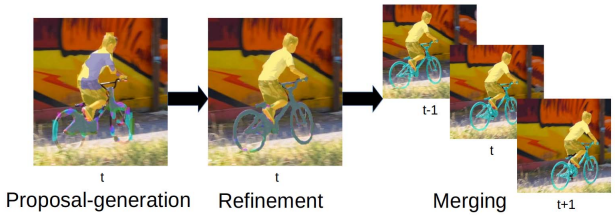

Figure 2. Overview of the three stages of PReMVOS [10].

2019 entry [11], we compared how they perform across different VOS scenarios. Our finding was 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. For the final competition entry, we combined both PReMVOS and BoLTVOS by using PReMVOS as the base tracking algorithm. We then used the BoLTVOS conditional R-CNN to reject false positive tracking results.

PReMVOS, BoLTVOS and UnOVOST. In this work, we aim to combine the best parts of each of the three approaches: we use the Conditional R-CNN detector from BoLTVOS since by conditioning its detection on the ground truth bounding box, it achieves more accurate results than the generic detectors used by PReMVOS and UnOVOST. We further adopt the tracklet-building step from UnOVOST since it proved to work very well for handling a large number of objects, unlike the temporal consistency rescoring algorithm for BoLTVOS which does not scale well to multiple objects. Finally, we merge the tracklets into full tracks using scores based on the re-identification network, similar to the successful merging algorithm of PReMVOS.

2. Method

Our proposed method works in 4 steps:

1. We use the Conditional R-CNN from BoLTVOS [16] to generate bounding box detections for each first-frame ground truth mask.
2. We use the fine-tuned Box2Seg network from BoLTVOS to convert the bounding box detections into accurate segmentation masks.
3. We link the mask-based detections into short tracklets using optical flow warping like in UnOVOST.
4. We use a re-identification-based scoring formulation similar to PReMVOS to link the tracklets into tracks.

In the following, we will describe each step in more detail.

Conditional R-CNN. For the conditional detector (Fig. 1 left) from BoLTVOS, we base the architecture on the two-stage detection architecture of Mask R-CNN [4]. 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 [1] 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.

Unlike many other VOS methods, we do not fine-tune the conditional R-CNN on the first-frame annotations. Instead, for each first-frame bounding box, we evaluate it once on each frame of the video to produce conditional detections for this object. Note that we do not use the temporal consistency rescoring algorithm from BoLTVOS, but proceed to the next step with the raw conditional detection output.

Box2Seg. In order to produce segmentation masks for the VOS task, we use the off-the-shelf bounding-box-to-

segmentation-mask network (Box2Seg) which we also used for PReMVOS and BoLTVOS. This network is a fully convolutional DeepLabV3+ [2] network with an Xception-65 [3] backbone. It has been trained on Mapillary [13] and then COCO [7] to output a segmentation mask given by the object bounding box encoded as a fourth input channel. This network runs much faster than our conditional R-CNN and is able to convert 40 bounding boxes to segmentation masks per second. For each first-frame mask, we fine-tune Box2Seg for 300 steps and then segment each conditional detection for this object. To save time, we do not use Lucid data dreaming augmentations.

Tracklets. Following UnOVOST [19], we compute optical flow between each adjacent pair of frames using PWC-Net [15]. Afterwards, for each object, the conditional detections between each pair of frames are merged in the following way. Each detection mask from $t-1$ is warped into frame t using the optical flow. Between each pair of warped detection masks from frame $t-1$ and detection masks from frame t, the Intersection-over-Union (IoU) is calculated. Afterwards, the Hungarian algorithm for bipartite matching is used to find an optimal linking between the two frames while allowing masks only to match if their IoU is at least 0.05.

Tracking. After obtaining tracklets, we then merge these together into long-term consistent objects tracks. This can be seen as adapting UnOVOST [19] to the semi-supervised task, or as adapting PReMVOS to work on tracklets as input, rather than single-frame proposals.

For this, we use object re-identification embedding vectors, which enable us to quantify the visual similarity between tracklets. We extract these re-identification vectors from a ReID network [14]. This network is trained on YouTube-VOS [18] using a triplet loss variant [5] in order to generate 128-dimensional ReID vectors which are similar for crops of the same object (in different frames), and different for crops of different objects. For each tracklet, the ReID embedding is extracted for each proposal and averaged over the whole tracklet. The L2 distance between these embeddings is then the measure of the visual dissimilarity between two tracklets.

Since each proposal is generated from the conditional R-CNN conditioned on a given first-frame mask, there exists a separate set of tracklets for each object in the video. The given first-frame masks are always assigned to a unique tracklet, this tracklet is thus always selected first to belong to the track for each object. For all other potential tracklets for an object, the ReID similarity score is calculated by taking the L2 distance of this tracklet's ReID vector to the initial tracklet's ReID vector, and then converting this distance to a similarity metric using the following equation:

Rank	Team Name	Overall	$\mathcal{J}_{seen}$	$\mathcal{J}_{unseen}$	$\mathcal{F}_{seen}$	$\mathcal{F}_{unseen}$
1	zszhou	**81.8**	**80.7**	77.3	**84.7**	84.7
2	theodoruszq	81.7	80.0	**77.9**	83.3	**85.5**
3	zxyang1996	80.4	79.4	75.9	83.3	83.1
4	swoh	80.2	78.8	75.9	82.5	83.5
5	youtube_test	79.1	77.9	74.7	81.5	82.2
6	Jono (**Ours**)	71.4	70.3	68.0	73.6	74.0
7	andr345	71.0	69.9	66.7	73.2	74.0

Table 1. Results of the 2nd Large-scale Video Object Segmentation Challenge - Track 1: Video Object Segmentation.

$$S_t = 1 - \frac{\|e_t - e_{FF}\|}{\max_i \|e_i - e_{FF}\|}$$

where S_t is the similarity score for tracklet t, e_t is the embedding for tracklet t, e_{FF} is the embedding for the tracklet containing the first-frame mask, and $\|\cdot\|$ is the L2 norm.

For sequences with more than one object, we also calculate an "interaction" score as the complement of the likelihood of each tracklet belonging to a different object in the video. High "interaction" scores indicate that this proposal is unlikely to belong to a different object, whereas a low score indicates that it is very similar to a different object and thus less likely to be the current object. This score is calculated using:

$$I_j = 1 - max_{i \neq j} S_i$$

where I is the interaction score and $max_{i \neq j} S_i$ is the maximum S score over all objects other than the one the tracklet belongs to.

The final score F_t for a tracklet t to belong to an object track is:

$$F_t = 0.8 \times S_t + 0.2 \times I_t$$

or in other words, 80% the similarity to the current object, and 20% the dissimilarity to other objects.

In order to select the best tracklets for all of the objects in a video, an iterative greedy tracklet selection method is used. Firstly, tracklets shorter than 3 frames are discarded. Then any tracklet which has a proposal in a frame where we have already selected a proposal for that object is discarded. Also tracklets for an object are discarded if they contain proposals that have greater than 50% IoU (intersection over union) with any proposal from a different object in the same video that has already been selected. From all remaining valid tracklets for all different objects in a video, we select the tracklet with the highest score F and repeat this procedure until there are no more valid tracklets for the

whole video. This gives us our final segmentation tracking results for all objects.

3. Results

For the YouTube-VOS challenge, the evaluation distinguishes between object classes which are part of the training set (seen) and those that are not (unseen). The primary evaluation measure "Overall" is the average of the $\mathcal{J}$ score and the $\mathcal{F}$ score for both seen and unseen object classes. Here, $\mathcal{J}$ measures the average IoU between the predicted masks and the ground truth masks, while $\mathcal{F}$ measures how well the boundaries of the predicted masks match the ground truth. Table 1 shows the results of the Video Object Segmentation track of the 2nd Large-scale Video Object Segmentation Challenge. Our entry achieved the 6th place with an overall score of 71.4%.

4. Discussion

Unfortunately, the results of this work were rather disappointing. We had hoped that by combining PReMVOS, BoLTVOS and UnOVOST in this way, we would see significant improvements in results, and be able to compete with the winning methods. However, there was a large gap (more than 7.5 percentage points) between our score and the 5 teams that placed higher than us. Furthermore, our results were in fact worse than that of PReMVOS alone, the winner of the 2018 YouTube-VOS Challenge (72.2 vs 71.4), although one must note that the test dataset changed slightly between years so these numbers are not directly comparable. Although the numbers for this challenge were not as we had hoped, we still believe that an approach that combines these three methods could potentially yield much better performance. Our method relies on a number of separate systems and a staged training approach. There is great potential for methods that perform all of the components of such an algorithm in an end-to-end fashion.

5. Conclusion

In this paper we present a novel method for Video Object Segmentation (VOS), which borrows ideas from three previously very successful VOS methods. Our approach detects potential objects conditioned on a given first-frame object, segments these using a separate network, links them into short tracklets using motion consistency given by optical flow, and then merges these tracklets into the final tracking results using the visual similarity of objects as defined by an object re-identification vector. We benchmark our method on the 2019 YouTube-VOS challenge where our method achieves 6th place with an overall score of 71.4%.

Acknowledgments: 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] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[5] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017.

[6] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for video object segmentation. *IJCV*, 2019.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[8] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. PReMVOS: Proposal-generation, Refinement and Merging for the DAVIS Challenge on Video Object Segmentation 2018. *CVPRW*, 2018.

[9] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. PReMVOS: Proposal-generation, refinement and merging for the YouTube-VOS challenge on video object segmentation 2018. *ECCW*, 2018.

[10] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018.

[11] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Combining PReMVOS with Box-Level Tracking for the 2019 DAVIS Challenge. *CVPRW*, 2019.

[12] Alan Lukezic, Luka Cehovin Zajc, Tomás Vojír, Jiri Matas, and Matej Kristan. Now you see me: evaluating performance in long-term visual tracking. *arXiv:1804.07056*, 2018.

[13] Gerhard Neuhold, Tobias Ollmann, S Rota Bulo, and Peter Kontschieder. The Mapillary Vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.

[14] Aljosa Osep, Paul Voigtlaender, Jonathon Luiten, Stefan Breuers, and Bastian Leibe. Large-scale object mining for object discovery from unlabeled video. In *ICRA*, 2019.

[15] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

[16] Paul Voigtlaender, Jonathon Luiten, and Bastian Leibe. BoLTVOS: Box-Level Tracking for Video Object Segmentation. *arXiv:1904.04552*, 2019.

[17] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *PAMI*, 2015.

[18] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *ECCV*, 2018.

[19] Idil Esen Zulfikar, Jonathon Luiten, and Bastian Leibe. UnOVOST: Unsupervised Offline Video Object Segmentation and Tracking for the 2019 Unsupervised DAVIS Challenge. *CVPRW*, 2019.