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

Jonathon Luiten, Paul Voigtlaender, Bastian Leibe

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

Abstract. We evaluate our PReMVOS algorithm [1][2](Proposal-generation, Refinement and Merging for Video Object Segmentation) on the new YouTube-VOS dataset [3] for the task of semi-supervised video object segmentation (VOS). This task consists of automatically generating accurate and consistent pixel masks for multiple objects in a video sequence, given the object's first-frame ground truth annotations. The new YouTube-VOS dataset and the corresponding challenge, the 1st Large-scale Video Object Segmentation Challenge, provide a much larger scale evaluation than any previous VOS benchmarks. Our method achieves the best results in the 2018 Large-scale Video Object Segmentation Challenge with a $\mathcal{J}\&\mathcal{F}$ overall mean score over both known and unknown categories of 72.2.

Our method separates this problem into two steps. We first generate a set of accurate object segmentation mask proposals for all of the objects in each frame of a video. We do this by first using a Mask R-CNN [4] like object detector to generate coarse object proposals, we then use a fully convolutional refinement network inspired by [5] and based on the DeepLabv3+ [6] architecture to produce accurate pixel masks for each proposal. Secondly we select and merge these proposals into accurate and temporally consistent pixel-wise object tracks over the video sequence. We use 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. More details on the PReMVOS method can be found in [1] and [2].

The PReMVOS algorithm also won the 2018 DAVIS Challenge on video object segmentation [7]. There are four main differences between our PReMVOS entry [2] for the DAVIS Challenge and the method presented here for the YouTube-VOS challenge. First, the use of Lucid Data Dreaming [8] to generate first-frame image aumentations was dropped as this method was far too slow to be used on the much larger Youtube-VOS dataset. This was replaced with simple random rotation, translation, flipping and brightness aumentations. Secondly, we no longer fine-tune

a seperate set of weights for each video for the refinement and proposal networks. Instead, we now fine-tune just one set of weights for each the refinment and the proposal networks on the set of all first frames in the YouTube-VOS validation and test sets jointly. Thirdly, we replace the use of FlowNet 2.0 [9] with PWC-Net [10] for optical flow calculation, using pretrained weights given by [10]. The final major change from the previous version of PReMVOS was that the set of proposals that the merging algorithm has to choose from each timestep has been expanded. As before, we use the *specific* refined proposals generated using a proposal net fine-tuned on the set of first frame images. We further add a set a *general* refined proposals generated using a proposal net that has not been fine-tuned on the validation and test set first frames. We also add a third set of object proposals at each timestep generated as the optical flow warping of the merging algorithms chosen proposals from the previous timestep. These are then refined using the refinement network and added to the set of total proposals that the merging algorithm is able to choose.

Our final results in the challenge are from an ensemble of using 11 different sets of merging algorithm weights that were chosen using hyper-parameter optimization on the DAVIS validation set.

For our optical flow network, we use PWC-Net [10] and use pre-trained weights provided by the authors. Our other three networks, our proposal, refinement and ReID networks, are pre-trained on ImageNet [11], PASCAL [12], COCO [13] and Mapillary [14], before being trained on the YouTube-VOS dataset, as well as the DAVIS [7][15][16] dataset.

In conclusion, we present results of our PReMVOS method [1][2] for video object segmentation on the new YouTube-VOS dataset and show that our method out-performs all other methods on this task by achieving the best results in the 2018 1st Large-scale Video Object Segmentation Challenge with a $\mathcal{J}\&\mathcal{F}$ overall mean score over both known and unknown categories of 72.2.

References

1. J. Luiten, P. Voigtlaender, B.L.: Premvos: Proposal-generation, refinement and merging for video object segmentation. arXiv preprint arXiv:1807.09190 (2018)
2. J. Luiten, P. Voigtlaender, B.L.: 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)
3. Xu, N., et al.: Youtube-vos. https://youtube-vos.org (2018)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV. (2017)
5. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.: Deep grabcut for object selection. In: BMVC. (2017)
6. 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)

7. 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)
8. 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)
9. 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)
10. Sun, D., Yang, X., Liu, M., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: CVPR. (2018)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. (2009)
12. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV **88**(2) (2010) 303–338
13. 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)
14. Neuhold, G., Ollmann, T., Bulo, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV. (2017)
15. 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)
16. 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)