Flow Adaptive Video Object Segmentation

Alex Lin, Yao Chou, Tony Martinez
Brigham Young University
Brigham Young University, Provo, UT 84602
flin2@byu.edu, yaochou@byu.edu, martinez@cs.byu.edu

Abstract

We tackle the task of semi-supervised video object segmentation, i.e. pixel-level classication of the images in video sequences using only the ground truth mask for the first frame of its corresponding video. Recently introduced online adaptation of convolutional neural networks for video object segmentation (OnAVOS) has achieved excellent results by pretraining the network, fine-tuning on the first frame and training the network at test time using its approximate prediction as newly obtained ground truth. While achieving impressive performance, OnAVOS uses simple approximation of its online prediction as ground truth for online updates, which leaves signicant potential information unused. We propose Flow Adaptive Video Object Segmentation (FAVOS) that refines the generated adaptive ground truth for online updates and utilizes temporal consistency between video frames with the help of optical flow. Our experiments show that FAVOS improves the state of the art on DAVIS 2016 Challenge from a mIoU (mean intersection-over-union) of 0.861 to 0.870. For the Semi-supervised track of the 2017/2018 challenge, we improve J & F measures from 0.565 (OnAVOS) to 0.617 on the test-development set, and from 0.577 (OnAVOS) to 0.606 on the test-challenge set.

1. Introduction

As Convolutional Neural Networks (CNNs) revolutionize the field of computer vision, there has been a trend to move from tasks such as image classification [1, 2, 3, 4] to object detection [5, 6, 7, 8], and from image segmentation [9, 10, 11, 12] to video object segmentation [13, 14, 15, 17]. Video object segmentation and tracking is vital in computer vision and has many significant applications such as video editing, autonomous vehicles, robotics etc. The segmentation task for video objects involves classifying each pixel as to whether it is part of a specific object in image frames of videos. The capability to successfully segment objects in videos is a key step towards human-level understanding of the surrounding environment for machines.

Recently in video object segmentation (VOS), many have achieved good performance by pretraining on large classification datasets to help the networks learn general objectness [21, 13, 14]. Many have used optical flow to help networks learn temporal consistency [15, 17, 21, 22, 23]. Some update the networks online at test time using previous predictions in order to adapt to large changes and keep track of objects during the video sequences [14, 15, 25, 26]. There are also interesting works tackling the task of VOS using unsupervised methods [27, 28], which have great potential.

This paper focuses on the task of semi-supervised video object segmentation, testing on the recently introduced DAVIS (Densely Annotated VIdeo Segmentation) dataset [18, 19, 20], which requires a segmentation of pixels in video sequences, classifying between foreground and background (DAVIS 2016 [18]) or multiple objects (DAVIS 2017/2018 [19]) given the ground truth of the first frame. Our approach improves OnAVOS [14] by adding an adaptation algorithm that refines the adaptation masks with the utilization of optical flow. Optical flow fields are vector fields that indicate the motion of pixels from $Image_t$ to $Image_{t+1}$. We refer to our approach as Flow Adaptive Video Object Segmentation (FAVOS).

2. Related Work

Video Object Segmentation and Tracking. Recently, as large datasets and computational power become more available, convolutional neural network based approaches [13, 14, 15, 24, 29] have become the state of the art in VOS. Pretraining on large image classification datasets has been proven effective for semi-supervised VOS [13, 14]. Alternatively, Khoreva [15] performs extensive data augmentation on specific videos using the provided first frame ground truth, and argues that training on small sets of data of objects related to specific videos is sufficient to produce good results. Le [16] and Li [17] firstly detect the target objects and perform segmentation on the detected bounding boxes, their methods perform well on re-identifying

previously missing objects. The usage of optical flow is also common. Khoreva [24] uses optical flow to propagate previous masks and treats the segmentation task as a mask refinement task. In addition, [15, 17, 24] utilize temporal consistency by feeding the optical flow field as additional input to the models. While others have mostly used optical flow as additional input to variational pipelines, hoping that CNNs would automatically learn the temporal connections between frames, we propose to use precomputed optical flow directly to refine the adaptation masks for accurate new ground truth masks and update the network online.

FCNs for Semantic Segmentation. Solving the task of semantic segmentation using fully convolutional networks (FCNs) was initially proposed by Long et al. [29]. They replace the fully-connected layers in classification networks with 1x1 convolutions so the network becomes fully convolutional. In addition, they define skip connections that share features between different levels in the network to help produce detailed segmentation. Wu [30] and Zagoruyko [4] proposed shallower but wider residual network models that outperform their predecessors in image classification. Additionally, Wu introduced a slightly modified model for semantic segmentation task which also shows competitive results across multiple datasets.

We propose to improve the recently introduced online adaptation of convolutional neural networks for video object segmentation (OnAVOS) [14]. OnAVOS adopted the architecture proposed by [30], achieved first place on DAVIS 2016 Challenge and fifth place on DAVIS 2017 Challenge. OnAVOS builds on OSVOS [13], which introduces pretraining steps for the network to learn general objectness before fine-tuning on the ground truth mask of the first frame in a specific video at test time. OnAVOS claims that its predecessor lacks the ability to adapt to large changes in video sequences due to its limited knowledge based only on the first frame of videos, and adaptively introduces an approach to update the network during test time, training on previous high-confidence predictions. However, the applied method for obtaining the high-confidence prediction is rather simple, which applys a constant threshold to the foreground logits to extract the confident foreground regions, and a distance transform followed by a very large distance threshold to extract the confident background regions, which neglects large portions of effective potential training due to its simplicity in selection of the new masks for online adaptation.

3. FAVOS

We introduce Flow Adaptive Video Object Segmentation (FAVOS) that extracts its high-confidence regions for the online training with the guide of optical flow [31, 32]. Originally, OSVOS states that methods using temporal

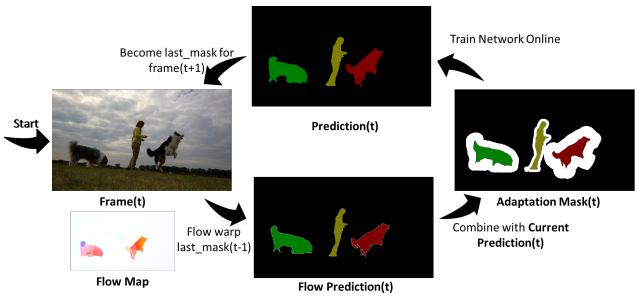

Figure 1. FAVOS online adaptation pipeline. We combine the current prediction and flow warpped estimation to create adaptation mask, which we use to finetune the network online for final prediction on each frame. **Best viewed in color.**

consistency (including optical flow) work well when the target objects transition very gradually, but fail in cases such as occlusions and abrupt motion. Building on OSVOS, OnAVOS does not utilize any temporal information either. We have also observed that the current top-performing approach [32] for obtaining the optical flow field is far from perfect and its usage could possibly lead to degradation in performance in video object segmentation. The question becomes: how can we utilize the rough optical flow field estimation correctly to improve performance?

The main idea of our approach is to obtain the confident foreground and background regions by utilizing the current prediction and the previous prediction remapped (flow warped) by the optical flow field. The adaptation algorithm used by OnAVOS [14] utilizes the temporal connection between previous and current predictions by using regions too far away from previous foreground prediction as a mask for current background predictions, which is simple yet produces good results. The intuition is that new objects entering the scenes are particularly troublesome to predict since the network has not trained on them as negative examples and therefore outputs high probabilities. By using the previous mask to help determine an approximate foreground region, the false positives far away from previous foreground can be set as background for training before the final prediction. However, OnAVOS had to set the foreground logits threshold value α and background distance threshold value d very high for safe adaptations, therefore leaving out much blank area between the confident foreground and background for useful training.

In order to obtain more informative and still accurate adaptation masks, we have performed numerous experiments which utilize additional information other than the current and previous predictions, particularly the optical flow field (obtained using FlowNet2.0 [32]). Fig. 1 illustrates our approach. Initially, we used the optical flow field in a similar way to [24], which intends to extract a rough segmentation mask of the primary object using the flow field, based on the assumption that objects tend to have

Figure 2. Optical flow fields with helpful segmentation information, consistent motion in foreground and background (first row). Optical flow fields that have various motion in foreground and background, therefore the target object's segmentation information cannot be extracted (second row).

consistent motion. This approach works well on videos where the background motion is consistent and the primary target object has a different motion than the background, but fails in most other more general cases (see fig. 2). As a result, instead of using the optical flow field for additional segmentation information, we use it as a mapping tool that warps the previous mask to produce a current estimation [15], which we refer to as flow estimation. To obtain the confident adaptation mask in each frame of the video sequence, our first step is to obtain the current foreground prediction for each object. We use distance transform on the predicted foreground object and select the region with distance values larger than an adaptive threshold determined by a percentile value ρ, such that the inner $\rho\%$ of the initial foreground is selected as confident foreground region. The second step is to produce flow estimation for each object. We then generate the confident foreground region using both the current prediction and the flow estimation, which checks the agreement between the current prediction and the flow estimation by using the *IoU* and selecting the confident foreground region accordingly. To avoid training on incorrect pixels, which can immediately lead to escalated errors in future predictions, we insert an unsure layer where no training takes place. The unsure layer is generated by applying a distance transform and a distance threshold on the confident foreground region, so pixels that are within the range of d pixels from the foreground are selected as unsure. The value of d is adaptively determined by the optical flow magnitude for the target object. Finally, the confident background region is simply the region outside of the unsure region. After obtaining the confident regions for all objects, we combine them to produce the final adaptation mask for the current frame. In general, we want to maximize the confident foreground/background region and minimize the number of incorrectly labeled pixels for the adaptation mask. As for online adaptation, we iteratively fine-tune the network by training on the adaptation mask of the current frame (for n_{cur} steps) and the ground truth of the first frame (for n_{first} steps). Re-training on the first frame is significant since the current adaptation masks can be inaccurate and the network needs to retain the knowledge of the target object. For DAVIS 2016 dataset, we perform DenseCRF [33] in the same fashion as OnAVOS. For DAVIS 2017/2018 dataset, the only post-processing required for the final prediction is using connected-components labeling for noise removal. We remove small components which fail to exceed a size threshold of $s = 5\%$ of the largest component for the corresponding object class.

4. Conclusion

In this work, we propose FAVOS for the task of semi-supervised video object segmentation. We have improved over OnAVOS, introducing a new pipeline that performs online adaptation with the utilization of optical flow and achieves better accuracy without increasing the model complexity. FAVOS improves the state of the art on DAVIS 2016 Challenge from a mIoU of 0.861 to 0.870. For the DAVIS 2018 challenge, we achieve the seventh place, improving *J & F* measures from 0.565 (OnAVOS) to 0.617 on the test-development set, and from 0.577 (OnAVOS) to 0.606 on the test-challenge set.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classication with deep convolutional neural networks. In NIPS, 2012.

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

[4] S. Zagoruyko and N. Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[6] R. Girshick. Fast R-CNN. In ICCV, 2015.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single shot multibox detector. In ECCV, 2016.

[8] J. Redmon and A. Farhadi. YOLO9000: Better, faster, stronger. In CVPR, 2017.

[9] K. Maninis, J. Pont-Tuset, P. Arbelaez, and L. Van Gool. Convolutional oriented boundaries. In ECCV, 2016.

[10] I. Kokkinos. Pushing the boundaries of boundary detection using deep learning. In ICLR, 2016.

[11] G. Bertasius, J. Shi, and L. Torresani. High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. InICCV,2015.

[12] G. Bertasius, J. Shi, and L. Torresani. Semantic segmentation with boundary neural elds. In CVPR, 2016.

[13] S.Caelles, K.-K.Maninis, J.Pont-Tuset, L.Leal-Taix, D.Cremers, and L. Van Gool. One-shot video object segmentation. In CVPR, 2017.

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

[15] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. In arXiv preprint arXiv: 1703.09554, 2017.

[16] T.-N. Le, K.-T. Nguyen, M.-H. Nguyen-Phan, T.-V. Ton, T.-A. N. (2), X.-S. Trinh, Q.-H. Dinh, V.-T. Nguyen, A.-D. Duong, A. Sugimoto, T. V. Nguyen, and M.-T. Tran. Instance re-identication ow for video object segmentation. The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops, 2017.

[17] X. Li, Y. Qi, Z. Wang, K. Chen, Z. Liu, J. Shi, P. Luo, C. C. Loy, and X. Tang. Video object segmentation with re-identication. The2017 DAVIS Challenge on Video Object Segmentation-CVPR Workshops, 2017.

[18] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. SorkineHornung. A benchmark dataset and evaluation methodology for video object segmentation. In CVPR, 2016.

[19] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbelaez, A. SorkineHornung, and L. Van Gool. The 2017 DAVIS challenge on video object segmentation. arXiv preprint arXiv:1704.00675, 2017.

[20] Jordi Pont-Tuset, Sergi Caelles, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, and Luc Van Gool. The 2018 DAVIS challenge on video object segmentation. arXiv preprint arXiv: 1803.00557, 2018.

[21] S. D. Jain, B. Xiong, and K. Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. In CVPR, 2017.

[22] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. In CVPR, 2016.

[23] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, Optical flow with semantic segmentation and localized layers, arXiv:1603.03911, 2016.

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

[25] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg. Beyond correlation lters: Learning continuous convolution operators for visual tracking. In-ECCV.Springer,2016.

[26] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In CVPR, 2016.

[27] Y. Jun Koh and C.-S. Kim. Primary object segmentation in videos based on region augmentation and reduction. In CVPR, 2017.

[28] P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. Technical report, arXiv:1704.05737, 2017.

[29] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[30] 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.

[31] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow. In CVPR, Boston, United States, June 2015.

[32] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical ow estimation with deep networks. In CVPR, 2017.

[33] P. Krhenbhl and V. Koltun. Efcient inference in fully connected CRFs with gaussian edge potentials. In NIPS, 2011.