Optical Flow-Guided Mask Generation Network For Video Segmentation

Yunyi Li[2], Fangping Chen[1], Fan Yang[1], Cong Ma [1], Yuan Li [1], Huizhu Jia[1], Xiaodong Xie [1⋆]
[1]*National Engineering Laboratory for Video Technology, Peking University*
[2]*Peking University Shenzhen Graduate School*
{yunyili, cfp, fyang.eecs, cong-reeshard.ma, yuanli, hzjia, donxie}@pku.edu.cn

***Abstract*—The purpose of video segmentation is to segment foreground objects from a video sequence. In this paper, we propose a CNN based method for the semi-supervised video object segmentation, where a hybrid encoder-decoder network is designed to generate pixel-wise foreground object segmentation in use of both spatial and temporal information. In order to minimize cumulative error of the network as much as possible, we develop a two-stage training scheme: alternate training and back-propagation-through-time training. Then the performances of our method and other state-of-the-art ones are compared on two annotated video segmentation databases. Furthermore, we also run an extensive ablation study to test the effects of different components from our method.**

***Index Terms*—semi-supervised, video object segmentation, optical flow, training scheme, mask**

I. Introduction

Semi-supervised video segmentation is about partitioning specific objects in a given video sequence with annotations available in its first frame. Largely due to its wide applications in video surveillance, autonomous driving, virtual reality, etc., the subject has attracted increasing interests in recent years of the computer vision research communities. However, open challenges remain in the development of semi-supervised video object segmentation technique of which the performance is currently below the satisfactory quality level.

According to the prior information of different categories , existing methods can be broadly grouped into two categories: 1) methods using spatial cues only, 2) methods using both spatial and temporal information. Methods in the first category [1]–[3] learn the representation of a single annotated object in a reference frame, and then segment the same object in following frames at pixel-level. To handle the appearance changes of the object of interest, researchers propose online adaptation schemes [3], or design additional modules to rectify the segmentation results [2]. In lacking of temporal information within the video sequence, these methods usually have limited performances in many real tasks where multiple objects exhibit similar appearances.

Methods of the second category [4]–[10] further leverage temporal information. Graph-based methods generate object segmentation via bilateral space [6], supervoxel [7], or optical flow [5]. Due to the powerful learning ability and the large amounts of training data, deep CNNs [4], [9], [10] have achieved very good performance. To establish segmentation consistency, the mask estimated from the previous frame is regarded as a reference. However, heavy reliance on the mask of the previous frame makes these models vulnerable to the cumulative error.

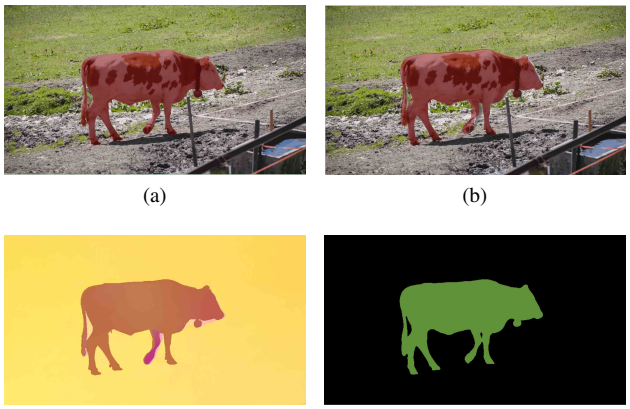


Fig. 1: Example result of our method: given one annotated frame (red), we propagate this manual labeling to the entire video sequence (green). The inputs of our network : (a) the previous frame with mask; (b) the current frame with the previous mask; (c) the optical flow map between the previous frame and the current frame with the previous mask; (d) the mask generated by our network for the current frame.

In light of the aforementioned observations, we propose a hybrid encoder-decoder network that targets at leveraging spatial and temporal information comprehensively and suppressing the influence of cumulative error. An encoder-decoder network is designed to simultaneously make use of the previous frame which specifies the target object to be detected in the current frame, the previous mask to be propagated to the current frame, and the optical flow calculates the motion of objects between two consecutive frames (as shown in Fig. 1). In addition, an efficient training strategy is adopted to minimize cumulative error. We refer to the proposed network as Optical Flow-Guided Mask Generation Network. In short, the proposed technique has three major contributions as listed:

1. We proposed an end-to-end trainable framework that uses both spatial, temporal and movement information to generate pixel-wise foreground object segmentation.

2. We developed a training method to minimize cumulative error.

978-1-7281-3320-1/20/$31.00 ©2020 IEEE

3. We demonstrated that optical flow and the previous mask are helpful to elevate segmentation result.

II. Related Work

Optical flow in segmentation. Optical flow is widely implemented in video object segmentation methods. It propagates annotated mask between consecutive frames based on graphical models. Hu et al. [11] initialized an active contour on the optical flow to roughly segment the object of interest. Tsai et al. [5] addressed video segmentation by a multi-level spatial-temporal graphical model with the optical flow and supervoxels put into use.

Many optical flow methods use track features to estimate motion in early signal processing, and then they match the correspondences between image based on optimization algorithms. Recently, learning based methods become popular [12], [13]. For the first time, Fischer et al. [12] applied CNN to optical flow prediction, and crafted two network structures: FlowNetS and FlowNetC. Ilg et al. [13] later developed [12] by stacking multiple FlowNetS and FlowNetC architectures and introducing a subnetwork specializing in small motions.

Considering the efficiency and accuracy, we implement the FlowNet 2.0 [13] to compute optical flow map, and also combine the previous mask estimation to highlight the motion information of objects.

III. Proposed Method

Moving objects change location and appearances over time. In our paper, these changes are assumed to be slow and smooth in video sequence such that it is able to calculate movement trend and instantaneous speed of the object using optical flow. To predict the mask of specific objects in the current frame, with an annotated image which is usually the first frame of the video being known, we design a hybrid encoder-decoder network. We fuse the features of objects displayed in both the previous frame and the current frame to capture the changes of location and appearance.

A. Network Architecture

The network architecture consists of three encoders, a global convolution block and a decoder, as shown in Fig. 2.

Encoders: We pick the ResNet101 as basic building element and design a multi-branch network for feature representation learning. As illustrated in Fig. 2, the designed network consists of three branches including: 1) one branch that aims to learn features from an optical flow map, and 2) two independent branches that learn features from the previous frame and the current frame. It is worth noting that we combine the previous estimated object masks with each input image or optical flow map to highlight the attentive regions. Indeed, we take the mask as an additional hot-channel which is concatenated to the image frame or optical flow map.

For the first branch, we use optical flow to extract adjacent pixels with similar motion, and filter out inactive background information. Combined with the mask, we can separate moving object instances with similar appearances.

We adopt the weights pre-trained on ImageNet to initialize the encoders, that has been proved effective in segmentation tasks since it can extract semantic features from natural image.

Global Convolution Block: The global convolution block deploys a combination of $1 \times k + k \times 1$ and $k \times 1 + 1 \times k$ convolutions to balance the classification and localization. The outputs of three encoders are concatenated to form the input of the global convolution block. The global convolution block enables densely connections within a large $k \times k$ region of the feature map, matches features between the current frame and the previous frame, and locates the target object. This module is also illustrated in Fig. 2.

Decoder: The purpose of the decoder is to refine the results and generate the mask for the current frame. To efficiently merge features in different scales, we employ three refinement block to process feature maps. It is worth noting that we also add features in the target encoder stream through skip-connections. At last, through a convolution layer with 3×3 filter and a softmax layer, we obtain a two-channel mask map.

B. Training

Training is performed on the training splits from DAVIS 2016 [14], DAVIS 2017 [15] and YouTube-VOS [16]. We want to train our network with as much data as possible. To guarantee the training quality, we present a two-stage training scheme: alternate training and back-propagation-through-time training.

Alternate training. Usually, the mask of the first frame is given, and masks of other frames in this video sequence are derived from the network. During the training process, we have two options for overlaying the mask on the current frame and the optical flow map: the ground truth and the prediction of network. The ground truth will make the training of the network lack continuity, and the result of the network will produce cumulative errors. Thus, we choose alternate training, which replacing the overlaid masks on inputs every 100 times. Tests have shown that alternate training improves the segmentation effect.

Back-propagation-through-time training. We should reduce the cumulative error during training to ensure the accuracy of each mask as much as possible. This training stage is followed by an alternate training. We take back-propagation-through-time (BPTT) to train our network. We select N consecutive frames from the entire video sequence (N=15 in our implementation), and choose the ground truth mask of the first frame for these N frames as the reference mask, then compute the train losses (mentioned in the implementation details) at each time step, and thereby update the whole network.

C. Implementation Details

FlowNet 2.0 is chosen as the optical flow network, with the original weight. At the same time, we use the pre-trained parameters in two encoders to process the current frame and the previous frame. We use Adam [17] optimizer with learning rate 1e-5, due to its supremacy over other adaptive

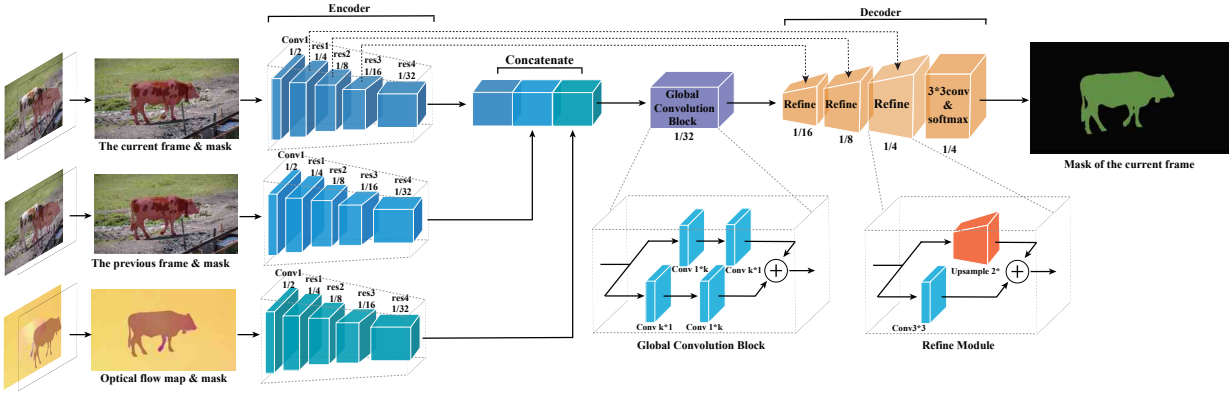


Fig. 2: Our network architecture. The network consists of three encoders, a global convolution block, a decoder. The relative spatial scales of feature maps is shown below each block.

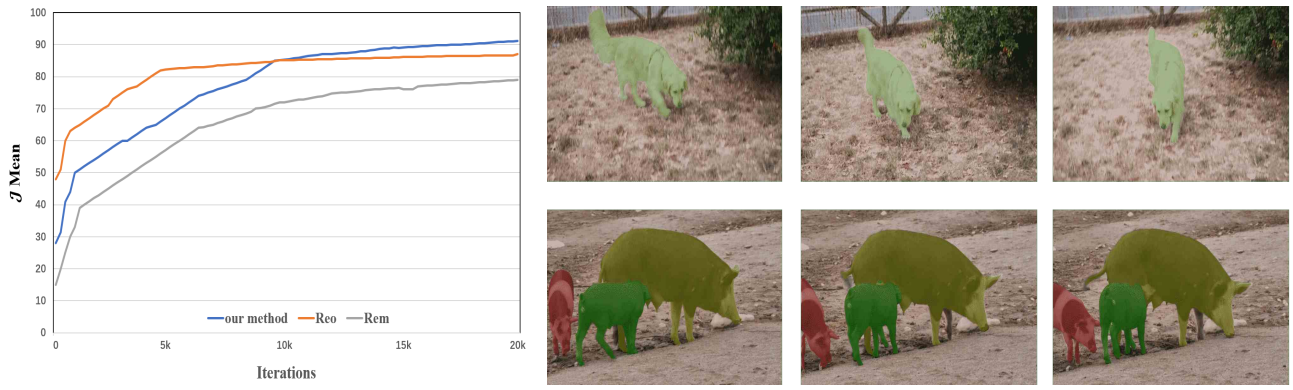


Fig. 3: (a) Performance of our method and $-Reo$, $-Rem$ on training datasets; (b) The qualitative results on DAVIS-2016 (single object) and DAVIS-2017 (multiple objects). Frames are sampled uniformly.

learning methods. We treat the mask generation as a binary classification problem, so we choose BCE Loss defined as follows:

$$loss(x_i, y_i) = -w_i[y_i \log x_i + (1 - y_i) \log(1 - x_i)] \quad (1)$$

In the case of multiple objects, we run the same model for each object independently and fuse the output masks into one overall map. Because one pixel cannot belong to multiple instances, we take the softmax aggregation [9] that combines multiple instance probabilities softly while normalizing them to be positive and to sum up to 1:

$$p_{i,m} = \sigma\left(\text{logit}\left(\hat{p}_{i,m}\right)\right) = \frac{\hat{p}_{i,m} / \left(1 - \hat{p}_{i,m}\right)}{\sum_{j=0}^{M} \hat{p}_{i,j} / \left(1 - \hat{p}_{i,j}\right)} \quad (2)$$

Where σ and logit represent the softmax and logit functions respectively, $\hat{p}_{i,m}$ is the network output probability of the instance m at the pixel location $i, m = 0$ indicates the background, and M is the number of instances. At each time step, we aggregate the network outputs of instances.

IV. EXPERIMENTS

We check our method on DAVIS-2016 [14] and DAVIS-2017 [15], and compare its performance with other state-of-the-art methods on the same databases. Then we run an extensive ablation study to demonstrate the effects of different components in our algorithm. The performance of the algorithms is evaluated by two indicators, the region similarity $\mathcal{J}$ [14] and contour accuracy $\mathcal{F}$ [14].

A. Results

In this paper, DAVIS-2016 is used for single-object segmentation, and DAVIS-2017 for multi-object segmentation.

DAVIS-2016: In Table 1, we report the results of single-object video segmentation. Among all the methods in the comparison, ours achieves comparable accuracy. We also run these two add-on studies on the DAVIS-2016 validation set, one adding the post-processing procedure which helps refine the output, and the other using online learning for adapting to the appearance of the object. The results are also shown in the Table 1.

1) Online learning

We fine-tune our model on the reference frame of a test video to make it more adaptive to the appearance of the object. The model is updated by Adam [17] optimizer with learning rate as 1e-7 and the number of iterations as 1000. The result shows online fine-tuning improves the accuracy of our model.

2) Post processing

TABLE I: Quantitative evaluation on DAVIS-2016. We highlight the features embedded in the methods: online learning (OL), post-processing (PP). We also show the experimental result of ablation study.

method	Add-on		results	
	OL	PP	$\mathcal{J}$ $Mean$	$\mathcal{F}$ $Mean$
OFL [5]			68.0	63.4
SegFlow [4]			76.1	76.0
OSVOS [1]	√	√	79.8	80.6
OSVOS^S [2]	√	√	**85.6**	**87.5**
RGMP [9]			81.5	82.0
Ours			83.7	84.0
Ours-add	√	√	**85.6**	84.5
Ours-Reo			79.8	78.0
Ours-Rem			73.4	74.6

TABLE II: Quantitative comparison among the algorithms on DAVIS-2017.

method	results	
	$\mathcal{J}$ $Mean$	$\mathcal{F}$ $Mean$
OSVOS^S [2]	52.9	**62.1**
OSVOS [1]	47.0	54.8
RGMP [9]	51.3	54.4
Ours	55.7	56.5
Ours-add (OL & PP)	**57.5**	57.0

We apply the dense CRF [8] to rectify our outputs. Compared with the refinement module used in the decoder, the method with an additional post processing unit affects the two measures differently: it improves region similarity but degrades contour accuracy since CRF smooths out object details.

The visualization results are shown in Fig. 3.

DAVIS-2017: We report the result of multi-object video segmentation in Table 2. The visualization results are shown in Fig. 3.

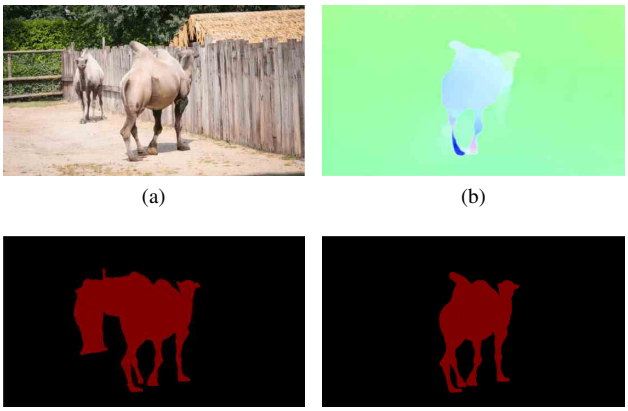
B. Ablation Study

This study evaluates the effects of the encoder that processes optical flow and previous mask by ablation. The results of ablation study are shown in Table 1.

Optical flow. One encoder processes the optical flow map to obtain the movement information of objects. To evaluate the setup, we name the model without the encoder that processes optical flow $-Reo$, and train it on DAVIS-2016. The network with optical flow spends more time to achieve the same effect as $-Reo$ at first, but after about 10k iterations, the mIoU of the original network (with the optical flow encoder) outstrips that of $-Reo$ (as shown in Fig. 3).

Fig. 4 presents some visual results. We claim that the optical flow can strongly facilitate the network to separate object instances with similar appearances from each other and from background.

Previous mask. The input of each encoder contains the previous mask in order to highlight the region of the object. If we stop feeding the previous mask, the network should target the object in the current frame without any temporal prior. To simulate this setup, we zero out to the previous mask at the target stream. We named this abridged model $-Rem$.



(a) (b) (c) (d)

Fig. 4: Comparison of visual results from proposed method and $-Reo$. (a) the 58^{th} frame, we only annotate the camel on the right; (b) the optical flow map between the 57^{th} frame and the 58^{th} frame; (c) the result of $-Reo$; (d) the result of our method.

(a) (b) (c)

Fig. 5: Comparison of visual results from proposed method and $-Rem$. (a) the 14^{th} frame; (b) the result of Rem; (c) the result of our method.

The network $-Rem$ are trained on DAVIS-2016. Compared to the network fed by the previous mask, we observe significant performance deterioration (as shown in Fig. 3).

Fig. 5 displays some visual results. We consider that the previous mask helps handling object appearance changes or target the most possible regions of object in the current frame

V. CONCLUSIONS

In this paper, we have presented an interframe information based approach for using both spatial and temporal information to propagate pixel-wise foreground object segmentation from first frame to the whole video sequence. We discusses the influence of the previous mask and optical flow map on the final segmentation result. We demonstrate that our encoder-decoder network trained by two-stage training reaches the current state-of-the-art performance in a certain index.

VI. ACKNOWLEDGMENT

This work is partially supported by the National Key Research and Development Program of China under contract No. 2016YFB0401904,Peking University Information Technology Institute (Tianjin Binhai).

REFERENCES

[1] S. Caelles, K. K. Maninis, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, and L. V. Gool, "One-shot video object segmentation," 2017.

[2] K. Maninis, S. Caelles, Y. Chen, J. Ponttuset, L. Lealtaixe, D. Cremers, and L. Van Gool, "Video object segmentation without temporal information," *arXiv: Computer Vision and Pattern Recognition*, 2017.

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

[4] J. Cheng, Y. H. Tsai, S. Wang, and M. H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," 2017.

[5] Y. H. Tsai, M. H. Yang, and M. J. Black, "Video segmentation via object flow," in *IEEE Conference on Computer Vision Pattern Recognition*, 2016.

[6] N. Mrki, F. Perazzi, O. Wang, and A. Sorkine-Hornung, "Bilateral space video segmentation," in *Computer Vision Pattern Recognition*, 2016.

[7] S. D. Jain and K. Grauman, *Supervoxel-Consistent Foreground Propagation in Video*, 2014.

[8] P. Krhenbhl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," 2012.

[9] S. W. Oh, J. Lee, K. Sunkavalli, and S. J. Kim, "Fast video object segmentation by reference-guided mask propagation," pp. 7376–7385, 2018.

[10] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkinehornung, "Learning video object segmentation from static images," 2017.

[11] P. Hu, G. Wang, X. Kong, J. Kuen, and Y. Tan, "Motion-guided cascaded refinement network for video object segmentation," pp. 1400–1409, 2018.

[12] P. Fischer, A. Dosovitskiy, E. Ilg, P. Husser, C. Hazrba, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision*, 2015.

[13] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," 2016.

[14] F. Perazzi, J. Pont-Tuset, B. Mcwilliams, L. V. Gool, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Computer Vision Pattern Recognition*, 2016.

[15] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbelez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," 2017.

[16] 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," 2018.

[17] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.