联系人 | 研究海报
杨帆：FYANG.EECS@PKU.EDU.CN | CNP20982

Optical Flow-Guided Mask Generation Network For Video Segmentation

Yunyi Li, Fangping Chen, Fan Yang, Cong Ma, Yuan Li, Huizhu Jia and Don Xie

National Engineering Laboratory for Video Technology, Peking University, Beijing, China

1. Problem Proposed

Video segmentation

- 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.

- **Categories of Video segmentation methods**

- 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 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, or design additional modules to rectify the segmentation results. 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 further leverage temporal information. Graph-based methods generate object segmentation via bilateral space, supervoxel, or optical flow. Due to the powerful learning ability and the large amounts of training data, deep CNNs 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..
- **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. 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.**

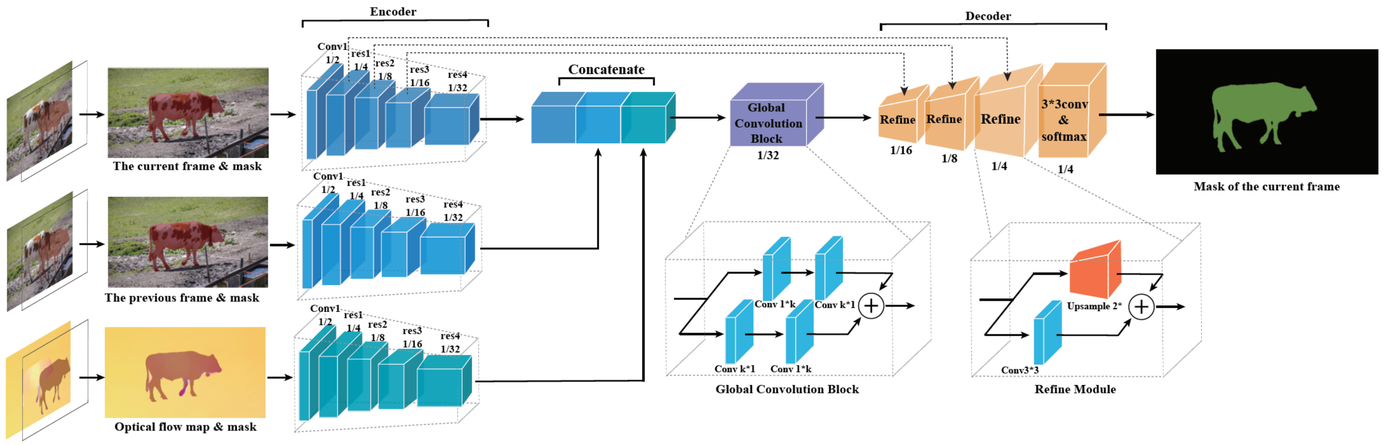

Fig. 1. Network Architechture.

2. The 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.

Encoders

- 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.
- 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.

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.

3. Experiment Results

Implement Details

The input of multiple frames heightens the memory consumption for storing layer activations. We solve these problems by employing data-parallel training on 4 NVIDIA TiTan X GPUs. The proposed method is implemented with PyTorch. With 4 GPUs, the training is 3x faster.

Qualitative Evaluation

We evaluate our method on DAVIS-2016 and DAVIS-2017, and compare its performance with other state-of-the-art methods on the same databases. The visualization results are shown in Fig. 2.

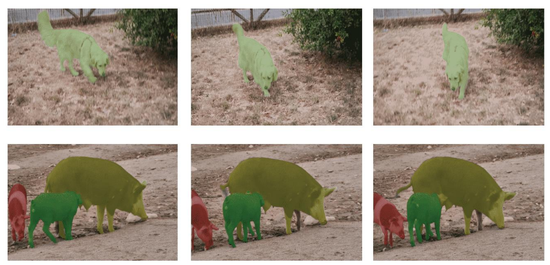

Fig. 2. The qualitative results on DAVIS-2016 and DAVIS-2017

- Among all the methods in the comparison, ours achieves comparable accuracy. We also run these two add-on studies on the DAVIS 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 and Table 2.

Quantitative Evaluation

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. 1. Quantitative evaluation on DAVIS-2016

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

Table. 2. Quantitative evaluation on DAVIS-2017