Mask Propagation Network for Video Object Segmentation

Jia Sun[†], Dongdong Yu[†], Yinghong Li, Changhu Wang[*]
ByteDance AI Lab, Beijing, China
{sunjia.ring, yudongdong, liyinghong, wangchanghu}@bytedance.com

Abstract

In this work, we propose a mask propagation network to treat the video segmentation problem as a concept of the guided instance segmentation. Similar to most MaskTrack based video segmentation methods, our method takes the mask probability map of previous frame and the appearance of current frame as inputs, and predicts the mask probability map for the current frame. Specifically, we adopt the Xception backbone based DeepLab v3+ model as the probability map predictor in our prediction pipeline. Besides, instead of the full image and the original mask probability, our network takes the region of interest of the instance, and the new mask probability which warped by the optical flow between the previous and current frames as the inputs. We also ensemble the modified One-Shot Video Segmentation Network to make the final predictions in order to retrieve and segment the missing instance.

1. Introduction

In recent years, video object segmentation has attracted much attention in the computer vision community. It has wide range of applications such as video editing, video summarization, scene understanding, and autonomous driving. Given the mask of labeled objects of the first frame, video object segmentation aims to separate the labeled objects from the background region in the future frames, which can be seen as a pixel-level object tracking task requiring fine-segmented profile and shape.

Regarding the state-of-the-art works for the task, most approaches build on basis of two mainstream methods, One-Shot Video Object Segmentation(OSVOS) and MaskTrack [1, 9]. OSVOS is based on the VGG16 network which is pre-trained on ImageNet [12, 5]. At the offline stage, the network is further fine-tuned on DAVIS 2016 dataset as its parent network[10]. Finally at the online stage, for each target video, the network is fine-tuned on the parent network by the given mask of its first frame, and used to segment the rest frames. All frames are processed independently. The results are temporally coherent and stable in the scenes where there are no drastic changes between consecutive frames. Due to the lack of temporal information, its performance decreases when it comes to some complex scenes. OSVOS shows its effectiveness in single-object segmentation, but has limitation in the multi-instance segmentation task. If there exists overlapping or occlusions among the instances, it is easy to mistake or miss all or parts of them.

Another popular approach is MaskTrack, it predicts the mask probability of the current frame with the guidance of the mask probability of the previous frame and the image appearance information of current frame [9]. In this way, it pays less attention to the useless background information and helps to separate foreground objects from background more accurately. MaskTrack shows stability and superiority in the long consecutive videos, as temporal information propagated from frame to frame. Since MaskTrack depends on temporal continuity, changes like occlusions and pose variations are likely to degrade mask propagation process, which may lead to performance drop. If the model fails to track segmentation mask for an instance in current frame, it is difficult to recover this instance in the next frames.

In order to cope with the problem, we analyze the existing approaches and ensemble the two-stream ideas to make up for each other. First, we build the mask propagation network to propagate estimated mask probability of the target object. We use the estimated mask probability of the last frame as guidance to predict the segmentation result of current frame. Second, we adopt the modified OSVOS Network to retrieve and segment the missing instance. Finally, we also apply conditional random field (CRF) on the segmentation probability map to further improve the results[8].

2. Methods

We model the video object segmentation task as the mask propagation task based on the image appearance information and image motion information. Given two adjacent image frames I_i and I_j, and the estimated mask probability P_i, we aim to predict the mask probability of the current frame

[†]Equal contribution.

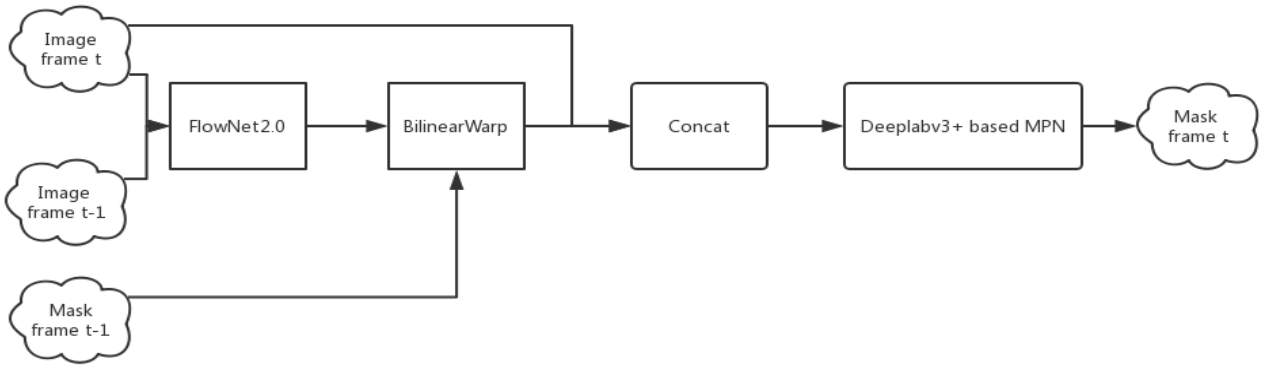

Figure 1. Network architecture of the Deeplab v3+ based mask propagation network.

j, i and j stands for two adjacent time. Inspired by the effectiveness of fully convolution network in the image classification and scene segmentation, we construct the mask propagation network to segment the identified instance by classifying each pixel into two classes: foreground instance and background. As shown in Figure 1, our network is based on the state-of-the-art scene parsing network Deeplab v3+ with the following modification: the network input is replaced with the current RGB image and the previous flow guided mask probability map, the network output is specified into two classes: foreground and background [3]. By using the Xception backbone based Deeplab v3+ pixel-wise classification network, we can obtain the powerful mask propagation result.

By training the mask propagation network, given two adjacent frames I_i and I_j, From frame i to frame j, the estimated mask in the image P_i is propagated to frame j, and the new mask P_j is computed as a propagation function N of the previous mask P_i, the new image I_j, and the optical flow $f_{i\to j}$, i.e. $P_j = N(W(f_{i\to j}, P_i), I_j)$. In this work, we first use the FlowNet 2.0 to extract the optical flow $f_{i\to j}$ from I_i and I_j [7]. The probability map P_i is warpped into $P_{i\to j}$ according to $f_{i\to j}$ by a bilinear upsampling function W. Then we crop the patches $P_{i,k}$, $f_{i\to j,k}$, and $I_{j,k}$ by using the bouding box of instance k. Rather than using the full-resolution image and flow guidance map, we feed the cropped image and flow guidance map into the mask propagation network to train the network function N. This approach can leverage any existing semantic labeling architecture, such as PSPNet, ResNet, and VGG backbone based DeepLab v2 network [13, 6, 12, 2]. In this work, we use the state-of-the-art scene parsing network DeepLab v3+ as our mask propagation network.

We also devel a modified version of OSVOS network [1]. We add skip-connections to every feature map before pooling operation and concatenate them after up-sampling to the output size in order to leverage multi-scale information. This allows direct supervision to feature maps of all scales, and the weights of supervision can be well controlled. To train this model, we follow the training procedure used in [1]. The training procedure includes two stages: the offline training stage and the online training stage at test time. At the offline stage, we first fine-tune ImageNet pretrained VGG16 network on MSRA10K dataset[12, 4], then fine-tune this network on DAVIS 2017 dataset to segment foreground objects from background. Since multiple instances may appear in the same frame, we simply merge instance labels into foreground label, and that improves model performance. At test time, for each instance we fine-tune the base network trained at offline stage on its label of the first frame to obtain the test network, then use this model to predict the segmentation masks of the whole video for this instance.

In the inference stage, we use the threshold to generate the instance ROI bounding box. We also adopt the modified OSVOS network to retrieve and segment the missing instance. As a final stage of our pipeline, we refine the generated the mask I_j using DenseCRF per frame[8]. This adjusts some image details that the network might not have captured.

3. Experiments

We evaluate our method on the DAVIS 2017 dataset, which contains 60 videos in the train set and 30 videos in the val set with pixel-level annotated object masks for all frames, and 30 videos in the test-dev set and 30 videos in the test-challenge set with the pixel-level annotated object mask for the first frame [11]. In our experiments, we use training set and validation set as training data in the training process. And predictions of test-dev set and test-challenge set are submitted to the CodaLab site of the challenge for

Table 1. The performance of different methods in DAVIS 2018 Challenge.

Team	OverAll	Region J			Boundary F		
	Mean	Mean	Recall	Decay	Mean	Recall	Decay
Jono	**74.7**	71.0	**79.5**	19.0	**78.4**	**86.7**	20.8
Lixx	73.8	**71.9**	79.4	19.8	75.8	83.0	20.3
Dawnsix	69.7	66.9	74.1	23.1	72.5	80.3	25.9
TeamILC_RIL	69.5	67.5	77.0	15.0	71.5	82.2	18.5
Apata	67.8	65.1	72.5	27.7	70.6	79.8	30.2
UIT	66.3	64.1	75.0	**11.7**	68.6	80.7	**13.5**
Alextheengineer	60.6	58.4	65.6	26.2	62.9	71.0	29.7
TeamVia(**Ours**)	60.1	57.7	64.9	27.2	62.4	71.7	28.1
Kthac	58.9	56.7	63.1	30.7	61.1	67.6	33.1

evaluation. We use Region Jaccard (J) and Boundary F measure (F) as the evaluation metrics for each instance [10]. As show in Table 1, we present the predicted segmentation results in the 2018 DAVIS Challenge and our score is 57.7% in J and 62.4% in F. The Figure 2 shows some examples of our predicted segmentation results in the 2018 DAVIS Challenge.

4. Conclusions

In this work, we propose to use the mask propagation network for video instance segmentation. We show that on the DAVIS 2017 dataset, the proposed mask propagation network achieves competitive performance for multiple instance segmentations in videos.

References

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

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[4] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[8] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[9] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *Computer Vision and Pattern Recognition*, 2017.

[10] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[11] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

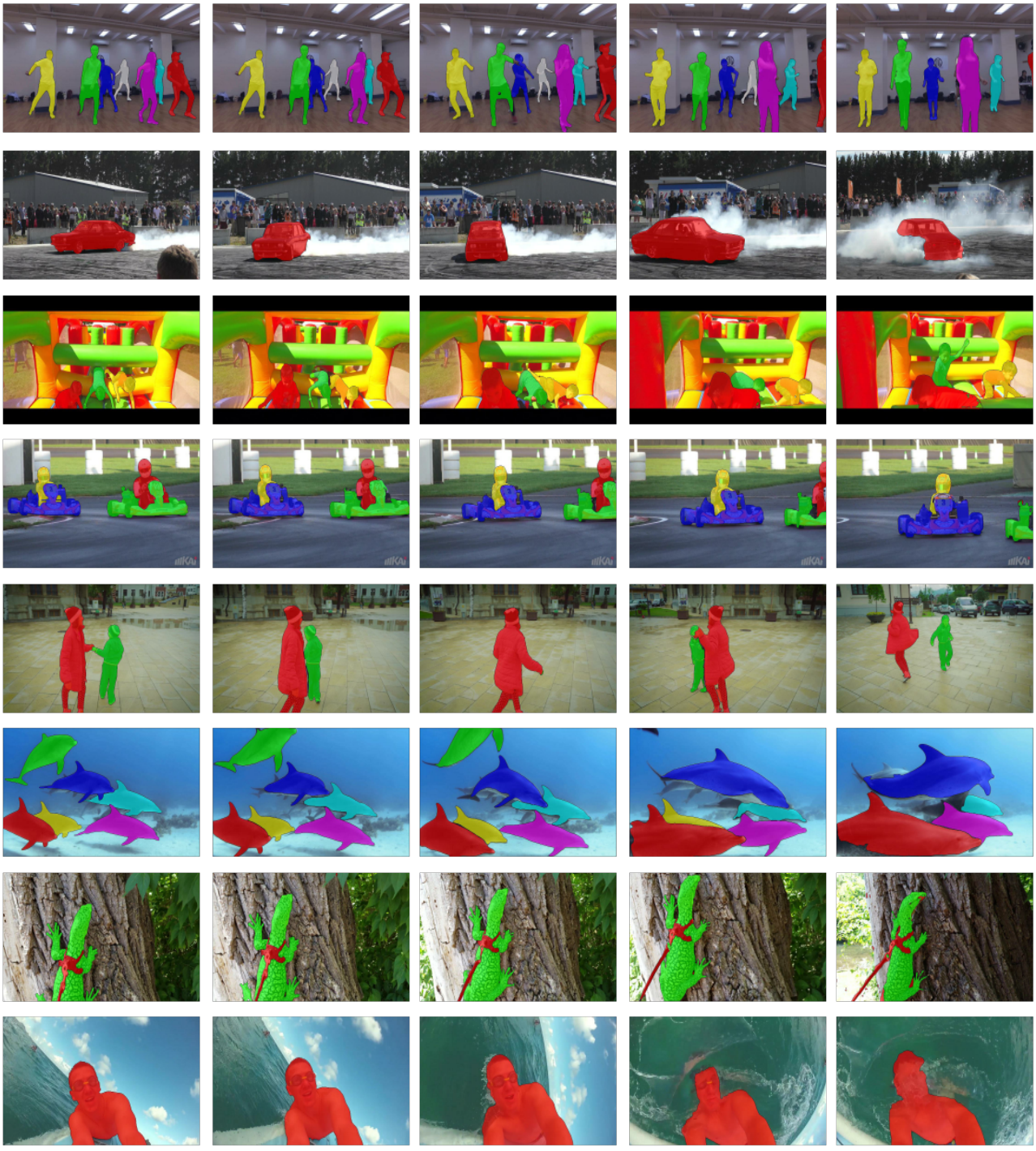

Figure 2. Qualitative results on DAVIS 2018 test-challenge set.