Video Object Segmentation based on improved OSVOS

Shizhan Hong¹,³, Tieyong Cao¹*, Shengkai Xiang¹, Zheng Fang¹, Xiaotong Deng¹ and Lei Xiang²

¹Lab of intelligent information processing, PLA Army Engineering University, Nan Jing, China
²61912, Beijing, China
³674081036@qq.com
*Corresponding author’s e-mail: cty_ice@sina.com

Abstract. Video object segmentation is a mainstream branch of current image processing direction. How to achieve deep learning from complete supervision to unsupervised is the key problem that people are trying to solve. In this process, One-Shot Video Object Segmentation (OSVOS) can successfully tackle the task of semi-supervised video object segmentation. It transfers the general semantic information learned on ImageNet to the foreground segmentation task, and then learns the mapping of a single annotated object in sequence. In this paper, based on the concept of OSVOS, an improved neural network structure with dilated convolution, multi-scale convolution fusion and skip layer is proposed. Dilated convolution can increase the size of the receiving field. Multi-scale convolution can obtain multi-scale feature maps and fuse them. Skip layers can transfer feature information from low layer to upper layer. All of them help to improve the final accuracy. The experimental results show that all indicators has been improved.

1. INTRODUCTION

As a preprocessing part of computer vision, object segmentation has received more and more attention in object relocation, scene classification and visual tracking. Video object segmentation is a mainstream developing direction of object segmentation. Unlike image object segmentation, video object segmentation has both sequential and motion properties which benefit for the perception and identification of scene.

Most of the current literature on semi-supervised video object segmentation enforces temporal consistency in video sequences to propagate the initial mask into the following frames. And in this road, some of the methods include the computation of optical flow [1,2], which considerably reduces speed. Concurrent works have also used deep learning to address Video Object Segmentation. MaskTrack [3] learns to refine the detected masks frame by frame, by using the detections of the previous frame, along with Optical Flow and post-processing with CRFs. In [4], the authors combine training of a CNN with ideas of bilateral filtering. Different from those approaches, OSVOS[5] is a simpler pipeline which segments each frame independently, and produces more accurate results, while also being significantly faster. OSVOS is one of the best semi-supervised learning method. Its contribution can be summarized as the following three points. Firstly, the contribution of the it is to adapt the CNN to a particular object instance given a single annotated image; secondly, it processes each frame of a video independently, obtaining temporal consistency as a by-product rather than as the
result of an explicitly imposed, expensive constraint; thirdly, OSVOS can make different tradeoffs between speed and accuracy.

But, due to the lack of sequential and motion information, it has the same common problem as image object segmentation. When the target is in a complex environment, there will be problems such as misrecognition or non-recognition. In the test results of OSVOS, we can find these problems. In order to improve the result without sequential and motion information, making full use of feature maps from all layers and fully extract the feature map of each layer will make the final result more accurate.

So, based on the concept of OSVOS, a new model is proposed by adding dilated convolution, multi-scale convolution and skip layer. Compared with OSVOS, this model has three advantages. Firstly, with the same parameters, dilated convolution can make the fixed-size convolution core see a larger area, so that the information can be fully utilized. Secondly, using multiple convolution cores of different sizes to obtain different scales of features, the combination of these features is often better than using a single convolution core. Thirdly, skipping layer transfers feature information from the bottom layer to the top layer, which can make full use of the feature maps of all layers.

2. OUR APPROACH

Fig 1 shows the overview of our video segmentation model. In this model, we propose a module called inception-x, as shown in Fig 2. Firstly, the basic feature map is obtained from the last convolution layer of each layer in VGG-16. Then, multi-scale feature maps are obtained by multi-scale dilated convolution, and they are fused into a feature map by convolution. Finally, the feature mapping is transferred to the upper layer by skip layer. Like OSVOS, training is divided into three stages. (1) A pre-trained base CNN is used on ImageNet labeling image. (2) The network is trained in DAVIS 2016 to improve the segmentation results, which has not yet focused on specific objects. (3) Fine-tuning the network for the specific target object in a single frame. The specific targets will be well segmented.
2.1 Multi-scale fusion
It is well known that the larger the convolution kernel, the larger the receptive field. And the more image information you can see, the better the features you will get. OSVOS uses small convolution kernel of the same size, which we think can be improved. Traditional networks use a stack of convolution layers, each layer uses the same size of convolution kernel, such as a large number of 3*3 convolution layers used in the VGG structure. But kernels with the same size can’t take full advantage of all the information. In fact, the feature map of one layer can use multiple convolution kernels of different sizes to obtain different scale features, and then fuse these features to obtain better features. In our model, four sizes of kernels are used to generate features of different scales. But there are two serious problems with this structure. Firstly, its parameters are much more than convolution kernels with same size. Such a large amount of calculation will make the model inefficient, so we introduce dilated convolution. Secondly, we do not combine the feature information obtained by all layers, so skip layer is introduced.

2.2 Dilated convolution
In the field of image segmentation method such as FCN, it will convolute and pooling the image to reduce the size of the image and increase the receptive field. Then the results are up-sampled to the original image size to complete the pixel prediction. In the process of reducing and increasing image size, some information will be lost. Dilated convolution can increase the receptive field without loss of pooling and up-sampling information. When the parameters are the same, the dilated convolution can achieve greater reception capacity.

2.3 Skip layer
In the process of encoding or convolution, the process of extracting features is image restoration. When decoding, even if the image is enlarged to its original size, a lot of information will be lost. The skip layer technology is a method of information preservation which combines the output layer with the upper layer. In the FCN model, the pooling layer in the encoder is added to the current convolution layer, and the result is used as the input of the next layer. These connections allow the network to use multi-resolution information, so the network can make more accurate segmentation decisions.

2.4 Algorithm
Hypothesis the feature from previous convolution layer is \( f_{\text{pre}} \), the feature from previous deconvolution layer is \( f_{\text{pred}} \), the convolution is \( D(\cdot, \ell) \) and the dilated convolution is \( d(\cdot, \varphi) \). The \( \varphi \) is the value of rate and \( \{\varphi_1, \varphi_2, \varphi_3, \varphi_4\} \) is the set of rates. The \( \ell \) is the parameter of deconvolution, \( \ell_u \) is the parameter of upper layer and \( \ell_m \) is the parameter of this layer.

\[
f = D(f_{\text{pre}} + f_{\text{pred}}, \varphi_1) + D(f_{\text{pre}} + f_{\text{pred}}, \varphi_2) + D(f_{\text{pre}} + f_{\text{pred}}, \varphi_3) + D(f_{\text{pre}} + f_{\text{pred}}, \varphi_4) + \cdots \cdots \cdots \cdots (1)
\]
\( f \) is the feature after the convolution at \( 1 \times 1 \times 1 \). \( f_{up} \) is the feature transfer to upper layer, \( f_{cu} \) is the final feature at this layer.

\[
\begin{align*}
\text{so } f_{up} &= D(f, \ell_{up}) \cdots \cdots \cdots (2) \\
\text{and } f_{cu} &= D(f, \ell_{cu}) \cdots \cdots \cdots (3)
\end{align*}
\]

3. EXPERIMENTS

In this paper, we use the code of OSVOS in Tensorflow. Its code is different from the code in Caffe, their results are different, yet their architecture is the same. If our improved model behaves better than OSVOS in Tensorflow, it will be as good as in Caffe. We train our model with many state-of-the-art methods on DAVIS database and use the normal evaluations to measure them. We compare to a large set of state-of-the-art methods, including two very recent semi-supervised techniques, OF, BVS, as well as the methods originally compared on the DAVIS benchmark: FCP [6], JMP [7], HVS [1], SEA [2], and TSP [8]. We also add the unsupervised techniques for comparison: FST [9], SAL [10], KEY [11], MSG [12], TRC [13], CVOS [14], and NLC.

3.1 dataset

The main part of our experiments is done on the recently-released DAVIS database [15], which consists of 50 full-HD video sequences with all of their frames segmented with pixel-level accuracy. It contains multiple video target segmentation challenges such as occlusion, motion blur and appearance changes. Each video is a dense annotation, pixel-level precision and frame-by-frame truth segmentation. To ensure adequate content diversity (which is necessary to fully evaluate the performance of different algorithms), the dataset spans four evenly distributed classes (human, animal, vehicle, object) and multiple actions.

3.2 evaluations

In this paper, we use three measures which are widely used: region similarity in terms of intersection over union (\( J \)), contour accuracy (\( F \)), and temporal instability of the masks (\( T \)). All evaluation results are computed on the validation set of DAVIS.

3.2.1 region similarity

The region similarity is the intersection over union between the mask \( M \) and the true value \( G \).

\[
J = \frac{|M \cap G|}{|M \cup G|}
\]

3.2.2 contour accuracy

Think of mask as a set of closed contours and calculate the contour-based \( F \) value which is the function of accuracy and recall. That is, contour accuracy is an \( F \) metric for contour-based accuracy and recall.

\[
F = \frac{2PR}{P + R}
\]

3.2.3 temporal instability

Intuitively, the region similarity measures the number of wrong pixels, while the contour accuracy measures the accuracy of the segmentation boundary. However, the temporal stability of the results is an important aspect of video object segmentation. Since the evolution of object shape is an important clue to recognition and jitter, unstable boundaries are unacceptable in video editing applications. Therefore, we have also introduced a time stability measurement method to punish this undesired effect. The key challenge is to distinguish between acceptable motion and unwanted instability and jitter. To this end, we estimate the deformation required to convert a frame mask to the next frame. Intuitively, if the conversion is smooth and accurate, the result can be considered stable.
3.3 results

Fig 3 and Fig 4 show results from OSVOS and our model. Fig 4 choose a simple dataset and Fig 5 choose some images from difficult datasets. In Fig 4, we can see that on a few simple datasets, our model is excellent at boxing out salient objects as OSVOS. Yet in terms of some difficult datasets, whether the integrity of the target or the accuracy of the edge, or the false rate, our model is better than OSVOS. It shows that our introduced model is useful. Because multi-scale fusion and dilated convolution will make full use of the information and transfer them to upper layer with skip layers. Then, all layer information will be fully used to get the final significant results. So, the integrity of the target and the accuracy of the edge will be improved.

Table 1 compares our model to OSVOS and other state of the art. We can see that in terms of region similarity $J$, our model is 0.7 point above OSVOS. In terms of contour accuracy $F$, our model is 1.0 point above OSVOS. So, it can prove that our model is better than OSVOS. And when compare our model to the rest of the state of the art, in terms of region similarity $J$, OSVOS is 8.2 points above the second best and 16.2 above the third best. In terms of contour accuracy $F$, OSVOS is 14.6 points above them. The results strongly demonstrate that the modules we introduce are valid in terms of completeness and accuracy.
4. CONCLUSIONS
OSVOS is a deep learning approach which doesn’t need a large amount of training data to solve a specific problem such as segmenting an object in a video. And it behaves best in DAVIS 2016 competition. In this paper, we propose a novel network architecture with introducing of dilated convolution, multi-scale fusion and skip layer to improve the performance of OSVOS. These operations can expand the receptive field, get better features and make enough use of information from all layers. The training was carried out on Davis 2016 and the results of different algorithms were compared. Objective and subjective evaluation results demonstrated the effectiveness of our model.

Acknowledgments.
This work is supported by National Nature Science Foundation of China (61471394); the Natural Science Foundation of Jiangsu Province for Excellent Young Scholars (BK20180080).

References
[1] Grundmann, M. , Kwatra, V. , Han, M. , & Essa, I. . (2010). Efficient hierarchical graph-based video segmentation. In:2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco.
[2] Ramakanth, S. A. , & Babu, R. V. . (2014). SeamSeg: Video Object Segmentation Using Patch Seams. In:2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus.
[3] Khoreva, A. , Perazzi, F. , Benenson, R. , Schiele, B. , & Sorkine-Hornung, A. . (2017). Learning video object segmentation from static images. In:2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu pp.3491-3500.
[4] Jampani, V. , Gadde, R. , & Gehler, P. V. . (2017). Video Propagation Networks. In:2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu pp.3154-3164.
[5] Maninis K K, Caelles S, Chen Y, et al. (2018). Video Object Segmentation Without Temporal Information. IEEE Transactions on Pattern Analysis & Machine Intelligence, J. Sci. Commun., 99:1-1.
[6] Perazzi, Wang, F. , Gross, O. , Sorkine-Hornung, M. , & Alexander. (2015). Fully Connected Object Proposals For Video Segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston.
[7] Fan, Q. , Zhong, F. , Lischinski, D. , Cohen-Or, D. , & Chen, B. . (2015). Jumpcut: non-successive mask transfer and interpolation for video cutout. ACM Transactions on Graphics J. Sci. Commun., 34:1-10.
[8] Chang, J. , Wei, D. , & Fisher, J. W. . (2013). A Video Representation Using Temporal Superpixels. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland.
[9] Papazoglou, A. , & Ferrari, V. . (2013). Fast Object Segmentation in Unconstrained Video. In: 2013 International Conference on Computer Vision. Sydney.
[10] Wang, W. , Jianbing Shen*, & Porikli, F. . (2015). Saliency-Aware Geodesic Video Object Segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston.

| Table1 the result of our mode comparing to the state of the art |
|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| MEASURE           | Semi-Supervised   | Unsupervised      |                   |                   |                   |
|                   | OURS | OSVOS | OP   | BVS  | PCP | JMP | HVS  | SEA  | TSP   | PST | NLC | MSG | KEY | CVOS | TRC | SAL |
| MEAN | 74.2 | 75.5 | 68.0 | 60.0 | 58.4 | 57.0 | 54.6 | 50.4 | 31.9  | 55.8 | 55.1 | 53.3 | 49.8 | 48.2 | 47.3 | 39.3 |
| RECALL↑ | 89.6 | 89.4 | 75.6 | 66.9 | 71.5 | 62.6 | 61.4 | 53.1 | 30.0  | 64.9 | 55.8 | 61.6 | 59.1 | 54.0 | 49.3 | 30.0 |
| DECAY↓ | 14.0 | 16.6 | 26.4 | 28.9 | -2.0 | 39.4 | 23.6 | 36.4 | 38.1  | 0.0  | 12.6 | 2.4  | 14.1 | 10.5 | 8.3  | 6.9  |
| MEAN | 78.0 | 77.0 | 63.4 | 58.8 | 49.2 | 53.1 | 52.9 | 48.0 | 29.7  | 51.1 | 52.3 | 50.8 | 42.7 | 44.7 | 44.4 | 34.4 |
| RECALL↑ | 92.9 | 92.4 | 70.4 | 67.9 | 49.5 | 54.2 | 61.0 | 46.3 | 23.0  | 51.6 | 51.9 | 60.0 | 37.5 | 52.6 | 43.6 | 45.4 |
| DECAY↓ | 14.1 | 18.2 | 27.2 | 21.3 | -1.1 | 38.4 | 22.7 | 34.5 | 35.7  | 2.9  | 11.4 | 5.1  | 10.6 | 11.7 | 12.9 | 4.3  |
| F MEAN | 44.0 | 41.0 | 21.7 | 34.5 | 29.6 | 15.3 | 35.0 | 14.9 | 42.2  | 34.3 | 41.4 | 29.1 | 23.2 | 24.4 | 37.6 | 64.1 |
[11] Lee, Y. J., Kim, J., & Grauman, K. (2011). Key-segments for video object segmentation. In: 2011 IEEE International Conference on Computer Vision, Barcelona. pp. 6-13.
[12] Brox, T., & Malik, J. (2010). Object Segmentation by Long Term Analysis of Point Trajectories. In: 11th European Conference on Computer Vision, Heraklion. pp. 5-11.
[13] Fragkiadaki, K., Zhang, G., & Shi, J. (2012). Video segmentation by tracing discontinuities in a trajectory embedding. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence.
[14] Taylor, B., Karasev, V., & Soatto, S. (2015). Causal video object segmentation from persistence of occlusions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston.
[15] Perazzi, F., Pont-Tuset, J., MeWiliams, B., Gool, L. V., & Sorkine-Hornung, A. (2016). A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas.