Video Object Segmentation with Weakly Temporal Information

Yikun Zhang[1], Rui Yao[1,*], Qingnan Jiang[1], Changbin Zhang[1], Shi Wang[1]
[1] School of Computer Science and Technology, CUMT
Xuzhou, 221116 - China
[e-mail: ruiyao@cumt.edu.cn]
*Corresponding author: Rui Yao

Received July 12, 2018; revised September 25, 2018; revised October 11, 2018; accepted October 22, 2018; published March 31 2019

Abstract

Video object segmentation is a significant task in computer vision, but its performance is not very satisfactory. A method of video object segmentation using weakly temporal information is presented in this paper. Motivated by the phenomenon in reality that the motion of the object is a continuous and smooth process and the appearance of the object does not change much between adjacent frames in the video sequences, we use a feed-forward architecture with motion estimation to predict the mask of the current frame. We extend an additional mask channel for the previous frame segmentation result. The mask of the previous frame is treated as the input of the expanded channel after processing, and then we extract the temporal feature of the object and fuse it with other feature maps to generate the final mask. In addition, we introduce multi-mask guidance to improve the stability of the model. Moreover, we enhance segmentation performance by further training with the masks already obtained. Experiments show that our method achieves competitive results on DAVIS-2016 on single object segmentation compared to some state-of-the-art algorithms.

Keywords: Video object segmentation, temporal feature, feed-forward architecture, further training

This work was supported by the Fundamental Research Funds for the Central Universities (No. 2017XKQY075).

http://doi.org/10.3837/tiis.2019.03.018 **ISSN : 1976-7277**

1. Introduction

Video object segmentation is a key task in computer vision with extensive applications, including video editing, video surveillance, video abstraction, video retrieval, motion analysis, and video semantics. The essence of video object segmentation is a pixel-level classification task, we assign a label for separating the foreground object and the background area to each pixel in the video frames.

Previous works on video object segmentation have been constrained by the lack of benchmark dataset. To address this problem, Perazzi et al. proposed a dataset DAVIS (Densely Annotated Video Segmentation) dedicated to this task in [1]. In addition, they also recommended three benchmark evaluation methods and opened the source code. Later, many DAVIS-based algorithms were proposed, the two representative architectures are the One-Shot network and MaskTrack network. Caelles et al. suggested a method based on VGG16 to segment each frame of video independently without temporal information in [2], it was the first attempt to use CNNs for the task of video object segmentation. OSVOS regards the video object segmentation task as the image segmentation task, and only using the first frame of the test set DAVIS during online training, so it is named "One-Shot". Motivated by the compatible results of the feed-forward architecture in DeepMask[3] and SharpMask[4], Perazzi et al. proposed a guidance segmentation framework and learned the idea of online fine-tune in object tracking to enhance the performance in [5]. In the offline training phase, the network is guided towards the foreground object by feeding the mask estimate of the previous frame. In online training, the network rapidly focuses on the specific target by online fine-tuning from object tracking. Therefore, this architecture is named "MaskTrack".

Fig. 1. The performance of OSVOS without temporal information is not satisfactory on the whole video. When the deformation of the target object is too large, the information in the first frame is not enough for video object segmentation.

Although the results of OSVOS are temporally coherent and stable, its temporal stability T is not satisfactory. Due to the segment method of One-Shot, OSVOS does not perform well in some situations, as discussed in detail below. We find that, with the deepening of the video sequence, the segmentation result is not satisfactory, especially when the object appearance of the following video frames has a visible difference from it in the first annotated frame. There are some typical examples that indications the shortcomings of OSVOS, as shown in **Fig. 1**. In addition, One-Shot convnet lacks the ability of further learning new information after online training. No matter how long the video is, the trained model can only use the knowledge learned from the first frame to complete the segmentation, which does not meet the practical application scenario of video object segmentation. We hope the system can improve continuously by learning new knowledge to achieve better robustness.

The video is a sequence of static images which transformed smoothly and slowly and played continuously, that is, the information carried by the neighbouring frames in the video sequence is very similar. Therefore, we consider introducing the segmentation result of the previous frame through the extended mask channel to guide the segmentation of the current frame. For the computer, as the video passes, the information contained in the first frame does not instruct well the network to process all frames. It is necessary to let the network learn new knowledge through further training, especially in practical applications.

To address the above problem, we propose a video object segmentation method utilizing weakly temporal information in this paper. The main improvements of our method are as follows: First, we introduce the temporal information into the video object segmentation through the feed-forward architecture. Secondly, the information in the previous frame mask is used multiple times to reduce the influence of random factors and enhance the stability of the model. Finally, we further train the model through online iteration to continuously update the network and further improve segmentation performance.

2. Related Work

Video object segmentation. Inspired by the satisfactory performance of OSVOS in video object segmentation, some improved algorithms based on it were proposed. In [6], Caelles et al. introduced instance-level semantic segmentation information into the architecture in order to enhance the performance of one-shot convnet. On the basis of OSVOS, Sharir et al. proposed a video object segmentation method, combining category-based object detection, category-independent object appearance segmentation and temporal object tracking in [7]. They obtained the segmentation mask and bounding box of the object through the One-Shot and Faster R-CNN networks, respectively. Then, the correct bounding box is filtered by the appearance-based filter and temporal filter. Finally, the high-precision bounding box is used to constrain the connection component of the segmented mask to enhance the segmentation performance. In [8], Amos et al. proposed a method to improve the result of OSVOS by online iterative. This method obtained several masks through the OSVOS network first. Then the refine masks were filtered out through the bounding box filter, which served as data for further training of the OSVOS model. As the appearance branch in OSVOS, the further trained model generated mask which fused with the output of contour capture branch to get the final mask. Moreover, they simplified the structure of OSVOS and improved the convergence speed of the network.

Bouwmans et al. firstly reviewed the application of the Robust PCA (RPCA) in image processing, video processing and 3D computer vision, and then pointed out the possible future

research directions of the method in [9]. The research of last seven years before 2013 done on video dynamic object segmentation was published in [10]. Recently, a number of methods for video object segmentation based on deep learning have been proposed. Yoon et al. proposed a network composed of encoding and decoding models which are suitable for pixel-level object matching in [11]. At the same time, they also proposed a feature compression technique that drastically reduced the memory requirements while maintaining the capability of feature representation. Moreover, this network was very robust and even had good performance on infrared data. [12] proposed a video object segmentation method based on super-trajectory representation. Combining two intuitive mechanisms for segmentation (reverse-tracking and object re-occurrence), this system was robust and performed well. An approach for video object segmentation utilizing frame-sequential label propagation was proposed in [13]. Chen et al. introduced TV-L1 to solve the problem of motion estimation while modelling the foreground object appearance in a range-adaptive way. Finally, a binary-level segmentation result was generated by blending the shape model and the appearance model via GraphCut. In [14], Li et al. proposed a method for unsupervised video object segmentation by transferring the knowledge encapsulated in image-based instance embedding. Instead of directly outputting the binary mask, they trained a network to generate embedding of the packaged instance information. As a result, this method adapted well to the changes of the foreground objects in the video. Khoreva et al. presented a method using language referring expressions to identify a target object for video object segmentation [15]. Given referring expression, they first localized the target object via the grounding model and enforced temporal consistency of bounding boxes across frames. Next, they applied a convnet-based pixel-wise segmentation model to recover detailed object masks. To address the rotational camera-motion, [16] suggested a method with multi-sprite backgrounds. Kumar et al. adopted a method using spatial-temporal filtering based on background subtraction to accomplish video object extraction and tracking task in complex environments [17]. Li et al. proposed an algorithm named Sub-Optimal Low-rank Decomposition (SOLD) in [18]-[19]. It performs efficient unsupervised video segmentation by suppressing the effects of data noises or corruptions. The method called Semantically-Guided Video Object Segmentation (SGV) is suggested in [20]. Caelles et al. introduced a semantic prior to guide the appearance model. Wang et al. introduced geodesic distance into saliency-aware video object segmentation to label the foreground objects more reliable [21].

Object tracking. Object tracking is one of the most critical tasks in computer vision and has many significant applications, including video surveillance, human-computer interaction, medical diagnosis and so on. Given the initial state (position and size) of a target object in the first frame of the video, its goal is to predict the state of the target in the subsequent frames. Existing object tracking algorithms can be classified into three categories: generating, discriminative, and deep learning based methods. The generating methods treat the tracking task as a template matching problem and use the tracker to find the most similar target region to the generated template [22]-[26]. While the discriminative method treats the object tracking as a classification task, which is also known as the tracking-by-detection method. What differs from the generative model is that tracking the maximum classification score between object and background is the goal of discriminative model. [27]-[31] are some attempts to handle the tracking problem with discriminative methods. In view of the outstanding performance of convolutional neural networks in the field of computer vision, recently, some tracking methods based on deep learning have emerged. [32]-[36] shows the state-of-the-art performance of some deep learning based target tracking algorithms.

Instance segmentation. Instance segmentation is a problem that detects and delineates each distinct object of interest that appears in the image. In recent research, the instance segmentation integrates the three tasks of object detection, image classification, and image segmentation, implementing these tasks through a framework. The latest representative work is the Mask R-CNN [37] which improved on Faster R-CNN [38]. It stems from the RCNN [40] framework proposed by Girshick R et al. for object detection in 2014. The input of RCNN is an image, and output the target's bounding box and category information. Given the input image, the output of the RCNN is the bounding box and category information of the target. Subsequently, Girshick R learnt the ideas of SPPNET [41] to improve the disadvantages of RCNN in a repetitive calculation and proposed FASTRCNN [39]. In the same year, Ren S et al. broke through the speed bottleneck of FAST-RCNN by resorting CNN to generate the regional hypothesis and proposed Faster-RCNN in [38]. Mask-RCNN added a segmentation task for each region of interest (RoI) and extended to three tasks. The performance of the model improved greatly by replacing the RoIPooling layer with the RoIAlign layer.

3. Method

In this section, we first briefly introduce the basic OSVOS network. Then we make a detailed description of the improved model we proposed, including the architecture and training details.

3.1 OSVOS Model

To reduce the impact of other factors, our improvement is based on the OSVOS model without boundary snapping branch. Our experiments are conducted on the Tensorflow code published by Caelles et al [2]. The OSVOS model implemented on TensorFlow is shown in **Fig. 2**. The OSVOS network is based on VGG16 and the fully-connected layer is removed. Skip paths from the last layer of each stage (before pooling) are suggested. The feature maps are recovered to the original image size by upscaling and then they are linearly fused into a single output.

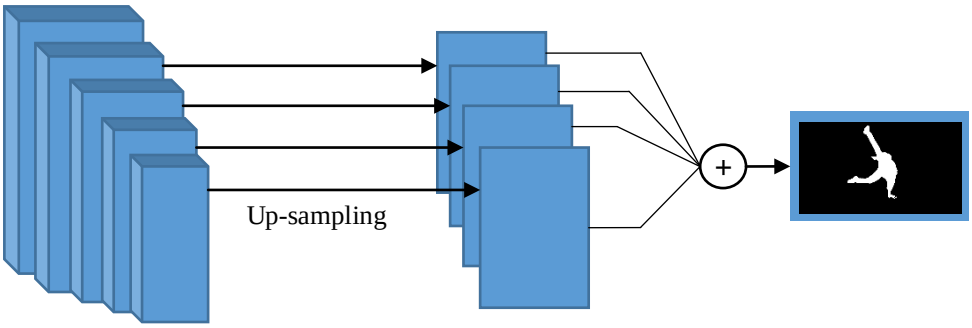

Fig. 2. The OSVOS model implemented on TensorFlow. The appearance network we adopted. It's based on VGG16 but the full connection layer has been removed and replaced with 1×1 convolution which more helpful for pixel-level classification.

OSVOS divides the video object segmentation into three phases. It starts with a basic CNN for image classification tasks pre-trained on ImageNet, that is, uses trained parameters to initialize the One-Shot network. Its results in terms of segmentation, although conform with

some image features, are not useful for video object segmentation. Then, the network called "parent network" is further trained on the train set of DAVIS with data augment. At this stages, the network has been able to separate the foreground object from the background area but not sensitive to the specific object. Finally, the network focuses on the specific object by fine-tuning with the first frame data of the test set in DAVIS.

3.2 Improved Model

To better illustrate the importance of temporal information for video object segmentation, our approach is implemented on the architecture of OSVOS (without temporal information). An extra branch used to extract timing features is added to the One-Shot network, hoping to get better results. After up-sampling, the feature map extracted by the newly added temporal branch is linearly fused with the original feature maps of OSVOS to generate the final mask of the current frame, and a loss function is assigned to it. The overview of our method is shown in **Fig. 3**.

3.2.1 Network Architecture

First, we extend the input channel from the original RGB to RGB+Mask. The extended mask channel is used to extract the temporal feature. We perform the affine transformation, the non-rigid deformation via thin-plate splines as well as the coarse on the previous frame mask to get the input of the mask channel. The transformation is to estimate the motion of the target object and to predict the position and shape of it in the current frame. Meanwhile, it also removesT some noise well, preventing errors from transmitting continuously in subsequent frames. The temporal feature is obtained by convolving the transformed previous frame, followed by an up-sampling operation to restore the original image size. Fusing it with the feature maps extracted from the OSVOS appearance branch to obtain the final refine mask of the current frame. In fact, the temporal branch is to make a prediction of the segmentation result of the current frame by transforming the mask of the previous frame, aiming to learn the transformation relationship between two adjacent frames of the target.

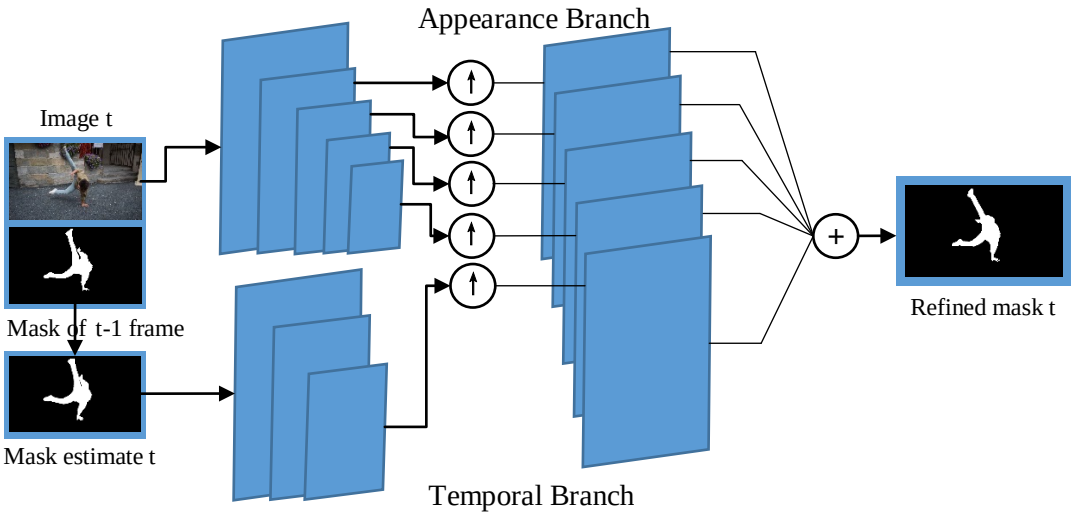

Fig. 3. Overview of our Network Architecture. (1) The original structure of OSVOS is preserved. (2) Before fusing features from each layer, we add the temporal branch to the framework. (3) The guidance information from the temporal branch is fused with the features extracted from the OSVOS model to generate the final mask.

Appearance branch. The One-Shot network has satisfactory performance in extracting the appearance characteristics of the target object. For appearance branch, we basically follow the structure of the OSVOS. The appearance network is based on VGG16 without full connection layer and it is fed with an RGB image (854×480×3).

Temporal branch. The extended mask channel serves as the input to the temporal branch. After inputting the mask of the previous frame, the affine, non-rigid deformation, and coarse operations are performed to estimate the position and shape of the target at the current frame, as well as removing some noises to prevent the error from expanding. After the convolution layer, the extracted feature map is up-sampled and recovered to the image size, follow by linearly fusing with the feature maps come from the appearance branch to generate the refine mask of the current frame. Because the mask is a binary image, a simple shallow model with 3 convolution layers is used when implementing temporal branching. For the convolutional layers of the extra mask channel, we use Gaussian initialization. Considering the computational complexity and the feasibility of the method, instead of using the strong timing information like optical flow to guide the segmentation, we employ the simple approach. However, the results of the experiment fully confirm the feasibility of our philosophy. A binary image (854×480×1) is fed into this branch.

The final output of our model is a refined mask (854×480×1) of the current frame and we apply a pixel-wise cross-entropy loss aimed at binary classification for it. In addition, we assign sigmoid as the activation function for the final layer as suggested in [2]. As for the activation functions of other layers, we adopt ReLU.

Although we have followed the architecture of OSVOS, our method do not separate each frame independently. The final mask of the current frame is generated with the motion of the previous frame as a guide. When testing, our architecture is a chained structure, as shown in **Fig. 4**.

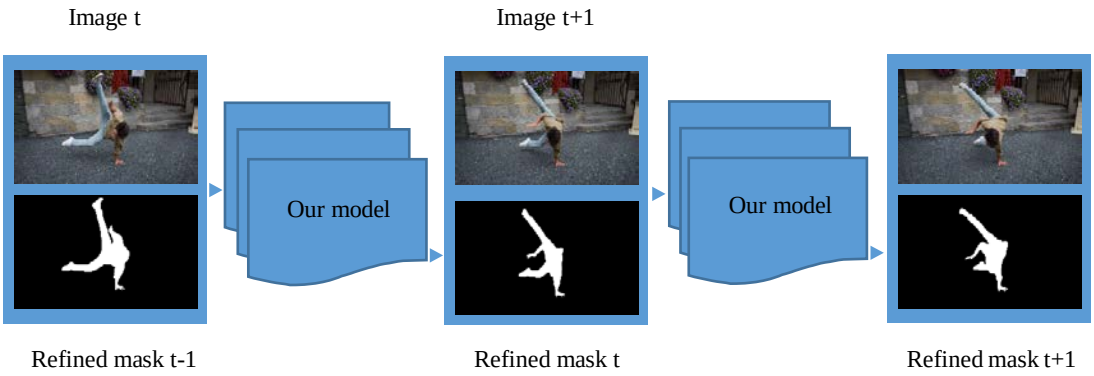

Fig. 4. In fact, the test network is a chained structure. The mask of t frame will be restricted by timing information from t-1 to t frame; and the temporal information from t to t+1 frame will affect the t+1 frame segmentation result.

3.2.2 Training Details

We adopt offline training and online training. When offline training, the network will learn the general appearance of foreground objects, but it is not sensitive to the specific goal. At the same time, how to use time information to guide the segmentation is significant for the network at this stage. For online training, given data of the first frame, the network rapidly focuses on the specific target.

Offline training. The architecture described in **Fig. 3** is iteratively trained 50,000 times on the DAVIS-2016 with data augment (scaling and mirroring), and Stochastic Gradient Descent with momentum of 0.9. While in the experiment we find that 50,000 times offline training is completely unnecessary because our network converges faster. For offline training, the processed (the affine, non-rigid and coarse deformation) ground-truth of the previous frame is used as the temporal branch input. For the affine transformation, non-rigid deformation and coarsening operations, we consider the suggestions proposed in [5]. We try to make some changes in the appeal parameters, but do not find any difference. Moreover, we find that the model only using the mask of the previous frame once has a certain degree of volatility, but this is not what we expect. Hoping to get a robust and stable system, we increase this proportion by deforming the mask of the previous frame five times to reduce the influence of random factors.

Online training/testing. In the online fine-tuning phase, we trained 500 times using the augmented data of the first frame in the video, allowing the network to focus on the specific object. At the same time, we discover that the original OSVOS network lacks the ability to further learn new information after online learning. For improving this problem, we add online iterations. We use the results of network segmentation to further train the model so that the network constantly learns new features to enhance segmentation performance. Noticing that the method of retraining will amplify the error when encountering bad segmentation results, so we only use the masks with satisfactory segmentation results of the first ten frames for online iteration, utilizing skip training to economize the train time. The number of online iteration is 300 times. As for testing, the segmentation result of the first frame is directly output by the original model. From the second frame, the refined mask of the current frame is segmented under the guidance of the previous frame.

4. Experiments

The experiment is implemented on the benchmark dataset DAVIS-2016 for video object segmentation. The DAVIS dataset focuses on the video object segmentation task and consists of 50 high-quality full-pixel video sequences, with totally 3455 frames, and each frame is annotated for pixel-level segmentation. The DAVIS dataset covers all challenging factors of video object segmentation, including Background Clutter (BC), Deformation (DEF), Motion Blur (MB), Fast Motion (FM), Low Resolution (LR), Occlusion (OCC), Out of View (OV), Scale Variation (SV), Appearance Change (AC), Edge Ambiguity (EA), Camera Shark (CS), Heterogeneus Object (HO), Interacting Objects (IO), Dynamic Background (DB) and Shape Complexity (SC).

Table 1. State-of-the-art comparison: Comparison of video object segmentation to the publicly available results on DAVIS-2016.

Measures		CVOS	CUT	BVS	JMP	FCP	NLC	OFL	MP-Net-F	OSVOS	VM	Ours
J	**Mean ↑**	48.2	55.2	60.0	60.7	63.1	64.1	68.0	70.0	74.2	75.9	**76.0**
	Recall ↑	54.0	57.5	66.9	69.3	77.8	73.1	75.6	85.0	84.8	89.1	**89.2**
	Decay ↓	10.5	2.3	28.9	37.2	3.1	8.6	26.4	**1.4**	16.5	-	14.8
F	**Mean ↑**	44.7	55.2	58.8	58.6	54.6	59.3	63.4	65.9	76.5	72.1	**78.7**
	Recall ↑	52.6	61.0	67.9	65.6	60.4	65.8	70.4	79.2	89.4	83.4	**92.5**
	Decay ↓	11.7	3.4	21.3	37.3	3.9	8.6	27.2	2.5	18.0	**1.3**	17.4
T	**Mean ↓**	24.4	26.3	34.7	**13.1**	28.5	35.6	22.2	56.3	42.6	25.5	38.8

We adopt the evaluation protocol provided by the benchmark [5]. Three metrics are used to evaluate our method: 1) region similarity (J) is adopted to measure pixel-level matching between segmented masks and the ground-truth; 2) as for the accuracy of contour, we use contour accuracy (F) to evaluate; 3) temporal stability (T) is introduced to punish unintended effects such as jitter and deformation.

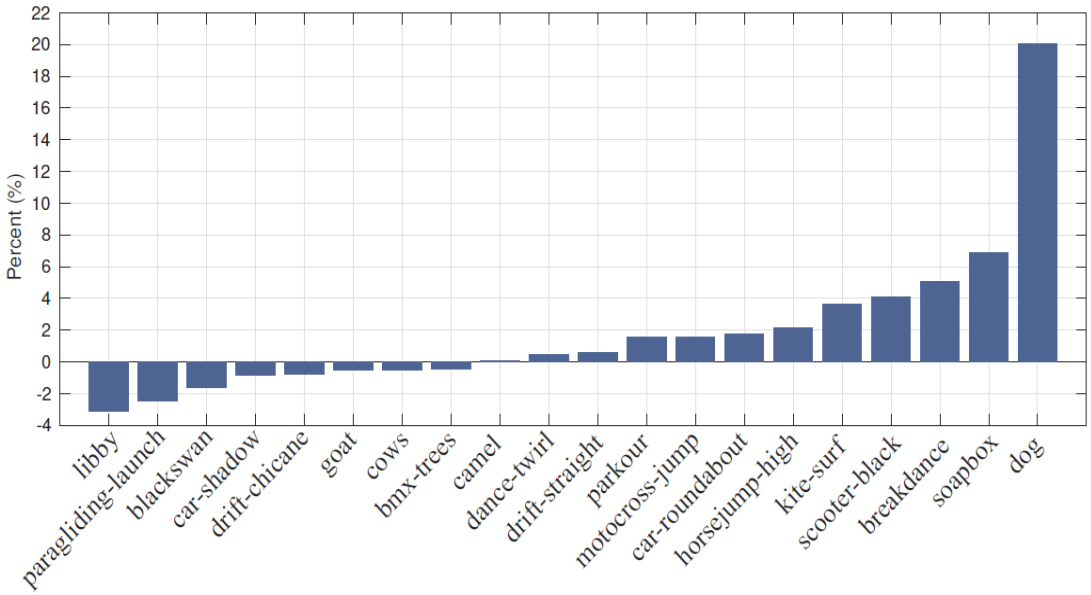

Fig. 5. The relative difference between our best performance (MmG+FT) and OSVOS on J Mean.

Table 2. The ablation study of our method on DAVIS-2016.
(SmG: Single-mask Guidance, MmG: Multi-mask Guidance, FT: Further Training)

SmG	MmG	FT	J			F			T
			Mean	**Recall**	**Decay**	**Mean**	**Recall**	**Decay**	**Mean**
Baseline(OSVOS)			74.2	84.8	16.5	76.5	89.4	18.0	42.6
√			75.0	85.8	15.5	77.8	90.9	17.4	36.6
	√		75.8	88.3	14.1	77.8	90.6	17.5	38.8
		√	75.1	88.0	**13.0**	77.1	91.7	**15.1**	**36.1**
√		√	75.2	87.2	15.4	77.9	93.2	18.0	36.8
	√	√	**76.0**	**89.2**	14.8	**78.7**	**92.5**	17.4	38.8

We evaluate our method with 10 state-of-the-art algorithms proposed for video object segmentation, including OSVOS[2], CVOS[47], CUT[48], BVS[44], JMP[49], FCP[45], NLC[46], OFL[43], MP-Net-F[42] and VM[12]. Among them, OSVOS, FCP, JMP, OFL, BVS, CUT and VM are semi-supervised methods, while MP-Net-F, CVOS and NLC are automated methods. Our algorithm consistently performs better than 10 recently proposed methods, as shown in **Table 1**. Before 2016, a key factor limiting various algorithms is the lack of large-scale datasets and benchmarks, including CVOS, CUT, NLC and JMP. After the datasets for video object segmentation is available, the performance for segmentation has been improved greatly and the regional similarity J of some methods has exceeded 0.7, such as

MP-Net-F, OSVOS and VM. MP-Net-F is an unsupervised method, which is superior to some semi-supervised methods because of the introduction of optical flow. In theory, our system can also boost performance with optical flow, but at the expense of huge computing resources. What we interested is a simple, economical and effective way to use temporal information, which is why our approach performs better than other methods.

Table 3. The total number of iterations until convergence.

	OSVOS	FT	SmG+FT	MmG+FT
J Mean	74.2	75.1	75.2	**76.0**
Number of Iterations	50k	50k	25k	**12k**

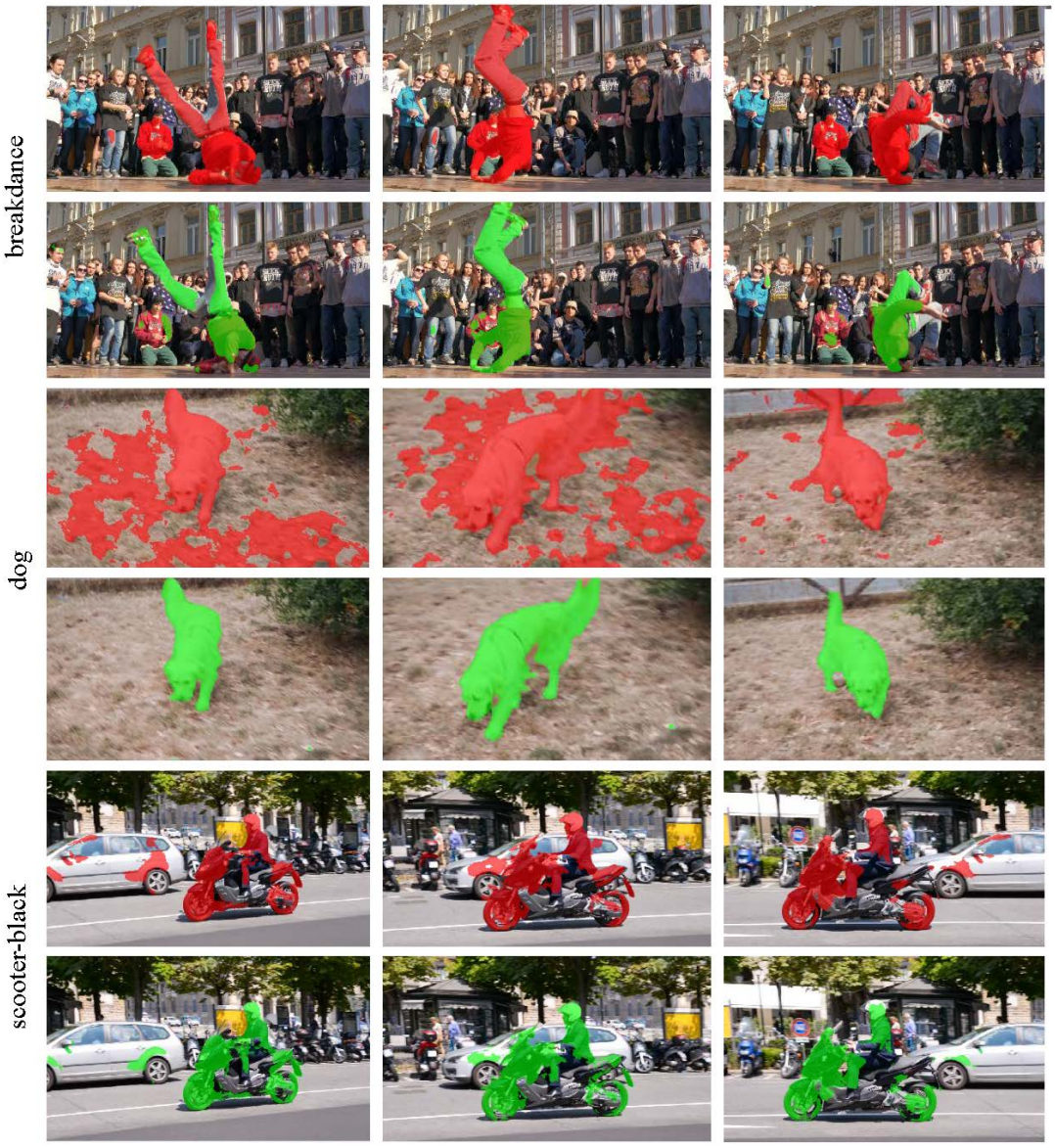

Fig. 6. Visualization results. The frames with red marks are the performance of OSVOS, and the green are ours.

Fig. 5 shows the relative difference for each sequence between our best performance (MmG+FT) and OSVOS. It reveals that the results of our method outperformed OSVOS on 12 test sequences among 20, while decreased on 8 sequences. Some visualization results (dog, breakdance and scooter-black) of improved sequences are shown in **Fig. 6**. The results show that the temporal information introduced is helpful for removing noise from similar backgrounds or other objects in segmentation. There is also a slight gain in the contour and connection components of the segmentation target object.

We also did ablation research on the three proposed improvements to more deeply explore the impact of various approaches, as shown in **Table 2**. Independently evaluating the three methods of SmG, MmG and FT, the results show they all have a boost relative to the baseline. Among them, the algorithm performance of the MmG promotes the most. FT achieves better performance because of the use of stronger temporal information (Multi-mask Guidance). Combining these three methods, MmG+FT obtains the highest performance improvement. Compared with SmG and FT, although SmG+FT has an improved performance, its best performance is lower than MmG+FT. In summary, the stronger temporal information helps to achieve better segmentation results, and further training can further improve segmentation performance.

Another advantage of our method is that it reduces the number of iterations of the model training. Using the mask of the previous frame as the guidance information can make the model converge toward the desired result more quickly. Our improved model can achieve better results with fewer iterations, as shown in **Table 3**. Compared with the baseline OSVOS, the number of iterations of MmG+FT drops to 12k, and the reduction of the training time by a factor of about 5.

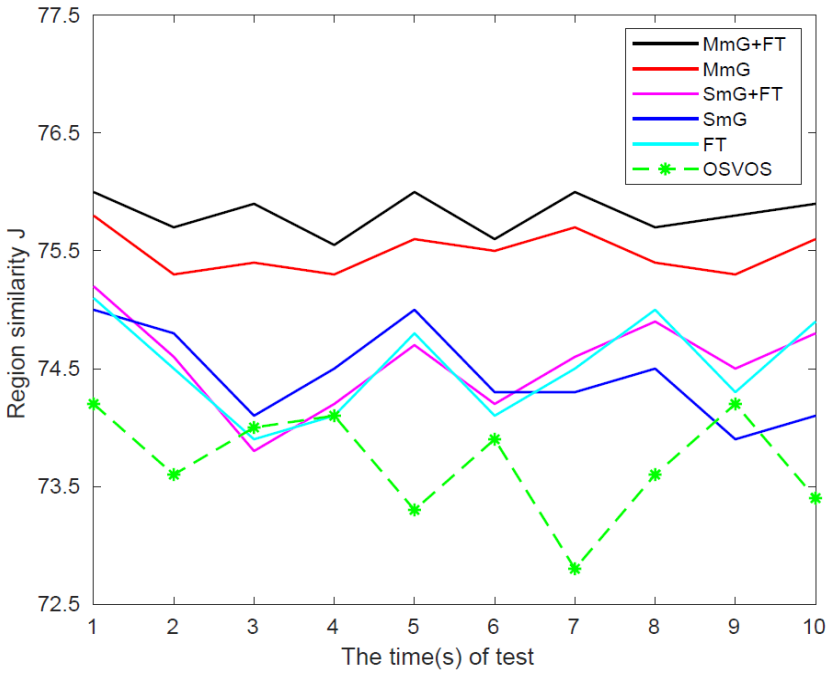

Fig. 7. Stability comparison of the proposed method. Green(OSVOS) is the baseline. Single-mask Guidance (SmG, purple and blue) and Further Training (FT, light blue) are helpful for segmentation but useless for the stability of our model. After introducing the Multi-mask Guidance (MmG, black and red), in addition to the improved segmentation results, our model also performs better in terms of stability.

Last but not least, we take stability into account when refining the model. **Fig. 7** shows ten test results for various methods including OSVOS, SmG, MmG, FT, Smg+FT, and Mmg+FT. Although the performance of FT, SmG, and SmG+FT has improved, it has similar amplitude fluctuations as OSVOS. In order to reduce the volatility caused by random factors, we propose an improved method of Multi-mask Guidance. By using the mask of the previous frame multiple times, the model achieves better stability (MmG, MmG+FT), and its fluctuation range is reduced from 1% (OSVOS) to 0.4% (MmG+FT).

5. Conclusion

Temporal information is especially significant for video object segmentation, but existing methods either treat segmentation as static image segmentation task without considering temporal information, or use optical flow with the cost of computing resources. To address this problem, we proposed a novel algorithm for video object segmentation exploiting weakly temporal features. Firstly, we added a temporal branch fed with the mask of the previous frame on an architecture without utilizing the interaction between adjacent frames, which transformed the independent segmentation of static images in OSVOS into a chained process. Second, we innovatively introduced the Multi-mask Guidance to improve the stability of the model by reducing random factors. Finally, we proposed to further train the model utilizing good results (not annotated data but the outputs of our model) in the testing process so that the network has the ability of learning new knowledge to enhance performance continuously. Although the temporal information used in our method is not strong, we still obtained competitive results on the DAVIS-2016 dataset compared to OSVOS and other the-state-of-art models.

We note that using weakly temporal information in this way is simpler and more economical than methods such as optical flow and it works. In addition, the idea of Further Training and Multi-mask Guidance has potential improvements for other systems. What's more, the methods in this paper can also be applied to other video tasks, such as detection and tracking. The position and appearance features of the previous frame can be extracted to guide the detection or tracking of the current frame. In future work, we will conduct experiments in related fields to verify the universality of our method, and visual tracking may be a good choice. Attempts on other architectures are also somethings we have to consider.

References

[1] F. Perazzi, J. Ponttuset, B. Mcwilliams, L. V. Gool, M. Gross, and A. Sorkinehornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 724-732, June 27-30, 2016. Article (CrossRef Link).

[2] S. Caelles, K. K. Maninis, J. Ponttuset, L. Lealtaixe, D. Cremers, and L. V. Gool, "One-shot video object segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5320-5329, July 21-26, 2017. Article (CrossRef Link).

[3] J. Gao, B. Wang, and Y. Qi, "DeepMask: Masking DNN models for robustness against adversarial samples," *arXiv:1702.06763 [cs.LG]*, February 2017. Article (CrossRef Link).

[4] Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert and Piotr Dollár, "Learning to refine object segments," in *Proc. of European Conference on Computer Vision*, pp. 75-91, September 17, 2016. Article (CrossRef Link).

[5] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkinehornung, "Learning video object segmentation from static images," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3491-3500, July 21-26, 2017. Article (CrossRef Link).
[6] K. K. Maninis, S. Caelles, Y. Chen, J. Ponttuset, L. Lealtaixe, D. Cremers, and L. Van Gool, "Video object segmentation without temporal information," *IEEE Transactions of Pattern Analysis & Machine Intelligence*, pp. 1-1, 2018. Article (CrossRef Link).
[7] Sharir, Gilad, E. Smolyansky, and I. Friedman, "Video object segmentation using tracked object proposals," *arXiv:1707.06545 [cs.CV]*, July 20, 2017. Article (CrossRef Link).
[8] Amos Newswanger and Chenliang Xu, "One-shot video object segmentation with iterative online fine-tuning," *CVPRW*, May 2017. Article (CrossRef Link).
[9] T. Bouwmans, S. Javed, H. Zhang, Z. Lin and R. Otazo, "On the applications of robust PCA in Image and video processing," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1427-1457, August 6, 2018. Article (CrossRef Link).
[10] RA. Graciela and CM. Mario, "New trends on dynamic object segmentation in video sequences: a survey," *DIEE&C*, vol. 11, no. 1, pp. 29-42, Dec. 2013. Article (CrossRef Link).
[11] J. S. Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. S. Kweon, "Pixel-level matching for video object segmentation using convolutional neural networks," *arXiv:1708.05137[cs.CV]*, August 17, 2017. Article (CrossRef Link).
[12] Tokmakov, Pavel, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," *arXiv:1704.05737 [cs.CV]*, July 12, 2017. Article (CrossRef Link).
[13] Y. Chen, C. Hao, W. Wu, and E. Wu, "Efficient frame-sequential label propagation for video object segmentation," *Multimedia Tools and Applications*, vol.77, no. 5, pp. 6117-6133, March 2018. Article (CrossRef Link).
[14] S. Li, B. Seybold, A. Vorobyov, A. Fathi, Q. Huang, and C. Kuo, "Instance embedding transfer to unsupervised video object segmentation," *arXiv:1801.00908 [cs.CV]*, February 2018. Article (CrossRef Link).
[15] Khoreva, Anna, A. Rohrbach, and B. Schiele, "Video Object Segmentation with Language Referring Expressions," *arXiv:1803.08006[cs.CV]*, Feb. 5, 2019. Article (CrossRef Link).
[16] D. Farin, P. de With, W. Effelsberg, "Video-object segmentation using multi-sprite background subtraction," in *Proc. of IEEE International Conference on Multimedia and Expo, ICME 2004*, pp. 343-346, June 27-30, 2004. Article (CrossRef Link).
[17] S. Kumar, J. Yadav, "Video object extraction and its tracking using background subtraction in complex environments," *Perspectives in Science*, vol. 8, pp. 317-322, September 2016. Article (CrossRef Link).
[18] C. Li, L. Lin, W. Zuo, W. Wang, and J. Tang, "SOLD: Sub-optimal low-rank decomposition for efficient video segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 5519-5527, June 7-12, 2015. Article (CrossRef Link).
[19] C. Li, L. Lin, W. Zuo, W. Wang, and J. Tang, "An approach to streaming video segmentation with sub-optimal low-rank decomposition," *IEEE Transactions on Image Processing*, vol.25, no.5, pp.1947-1960, May 2016. Article (CrossRef Link).
[20] S. Caelles, Y. Chen, J. Ponttuset, and L. Gool, "Semantically-guided video object segmentation," *arXiv:1704.01926v2[cs.CV]*, Jul. 17, 2018. Article (CrossRef Link).
[21] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20-33, Jan. 1, 2018. Article (CrossRef Link).
[22] K. Zhang, L. Zhang, M. H. and Yang, "Real-time compressive tracking," *European Conference on Computer Vision*, vol. 7574, pp. 864-877, October 2012. Article (CrossRef Link).
[23] J. Xing, J. Gao, B. Li, W. Hu, and S. Yan, "Robust object tracking with online multi-lifespan dictionary learning," in *Proc. of IEEE International Conference on Computer Vision*, pp. 665-672, Dec. 1-8, 2013. Article (CrossRef Link).
[24] D. A. Ross, Lim, R. S. Lin and M. H. Yang, "Incremental learning for robust visual tracking," *IEEE International Conference on Computer Vision*, vol. 77, no. 1-3, pp. 125-141, May 2008. Article (CrossRef Link).

[25] X. Mei, Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 33, no. 11, pp. 2259-2272, Nov. 2011. Article (CrossRef Link).
[26] B. Liu, J. Huang, L. Yang and C. Kulikowsk, "Robust tracking using local sparse appearance model and k-selection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 3619, pp. 1313-1320, June 20-25, 2011. Article (CrossRef Link).
[27] B. Babenko, M. H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 33, no. 8, pp. 1619-1632, Aug. 2011. Article (CrossRef Link).
[28] R. T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 10, pp. 1631-1643, Oct. 2005. Article (CrossRef Link).
[29] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M. M. Cheng, S. L. Hicks, and P. H. Torr, "Struck: Structured output tracking with kernels," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 38, no. 10, pp. 2096-2109, Oct. 1, 2016. Article (CrossRef Link).
[30] L. Zhang, and Van Der Maaten, "Preserving structure in model-free tracking," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 36, no. 4, pp. 756-769, April 2014. Article (CrossRef Link).
[31] J. Xing, J. Gao, B. Li, W. Hu, and S. Yan, "Robust object tracking with online multi-lifespan dictionary learning," in *Proc. of IEEE International Conference on Computer Vision*, pp. 665-672, Dec. 1-8, 2013. Article (CrossRef Link).
[32] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Stct: Sequentially training convolutional networks for visual tracking," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1373-1381, June 27-30, 2016. Article (CrossRef Link).
[33] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. of Computer Vision and Pattern Recognition*, pp. 3119-3127, Dec. 7-13, 2015. Article (CrossRef Link).
[34] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M. H. Yang, "Hedged deep tracking," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4303-4311, June 27-30, 2016. Article (CrossRef Link).
[35] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. of IEEE International Conference on Computer Vision*, pp. 3074-3082, Dec. 7-13, 2015. Article (CrossRef Link).
[36] S. Hong, T. You, S. Kwak, and B.Han, "Online tracking by learning discriminative saliency map with convolutional neural network," *arXiv:1502.06796 [cs.CV]*, February 24, 2015. Article (CrossRef Link).
[37] Hu, Yuan Ting, J. B. Huang, and A. G. Schwing, "Mask-RNN: Instance level video object segmentation," *arXiv:1803.11187[cs.CV]*, March 29, 2018. Article (CrossRef Link).
[38] S. Ren, K.He, R. Girshick, J. and Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp.1137-1149, June 1, 2017. Article (CrossRef Link).
[39] R. Girshick, "Fast r-cnn," *arXiv:1504.08083[cs.CV]*, September 27, 2015. Article (CrossRef Link).
[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, June 23-28, 2014. Article (CrossRef Link).
[41] Purkait, Pulak, C. Zhao, and C. Zach. "SPP-Net: Deep absolute pose regression with synthetic views," *arXiv:1712.03452[cs.CV]*, December 09, 2017. Article (CrossRef Link).
[42] Tokmakov, Pavel, K. Alahari, and C. Schmid. "Learning motion patterns in videos," *Computer Vision and Pattern Recognition*, pp. 531-539, April 10, 2017. Article (CrossRef Link).
[43] Tsai, Yi Hsuan, M. H. Yang, and M. J. Black. "Video segmentation via object flow," *Computer Vision and Pattern Recognition*, pp. 3899-3908, June 27-30, 2016. Article (CrossRef Link).

[44] N. Marki, F. Perazzi, O.Wang, and A. Sorkine, "Bilateral space video segmentation," in *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, pp. 743-751, June 27-30, 2016. Article (CrossRef Link).
[45] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung, "Fully connected object proposals for video segmentation," *in Proc. of 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3227-3234, Dec. 7-13, 2015. Article (CrossRef Link).
[46] Faktor Alon and Irani Michal, "Video segmentation by non-local consensus voting," *British Machine Vision Conference*, June 2014. Article (CrossRef Link).
[47] Taylor, Brian, V. Karasev, and S. Soattoc, "Causal video object segmentation from persistence of occlusions," in Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 4268-4276, June 7-12, 2015. Article (CrossRef Link).
[48] Keuper, Margret, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in Proc. of 2015 IEEE International Conference on Computer Vision, pp. 3271-3279, Dec. 7-13, 2015. Article (CrossRef Link).
[49] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen, "Jumpcut: Non-successive mask transfer and interpolation for video cutout," *Acm Transactions on Graphic*, vol. 34, no. 6, pp. 195, November 2015. Article (CrossRef Link).

Yikun Zhang is currently pursuing the M.S. degree from School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. His main research interests include semantic segmentation and deep learning.

Rui Yao is an associate professor in School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. His current research interests include computer vision and machine learning.

Qingnan Jiang is a junior student of China University of Mining and Technology in School of Computer Science and Technology. His main research interests include deep learning and image classification.

Changbin Zhang is a junior student of China University of Mining and Technology in School of Computer Science and Technology. His main research interests include deep learning and image classification.

Shi Wang is currently pursuing the M.S. degree from School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. His current interests include video object segmentation and object tracking.