Motion-Guided Spatial Time Attention for Video Object Segmentation

Qiang Zhou[1]*, Zilong Huang[1], Lichao Huang[2], Yongchao Gong[2], Han Shen[2], Wenyu Liu[1], Xinggang Wang[1]
[1]School of EIC, Huazhong University of Science and Technology
[2]Horizon Robotics
{theodoruszq,hzl,liuwy,xgwang}@hust.edu.cn
{lichao.huang,yongchao.gong,han.shen}@horizon.ai

Abstract

In this paper, we propose a novel motion-guided attention module to implant the spatial and time consistency in the correlation map of the current frame with the historical frames. Unlike other mask propagation based methods, our method regards the previous mask as a strong prior instead of concatenating it to the current frame or feature for propagation. Additionally, to reduce the gap between training and testing phase, we propose an improved optimization strategy, named sequence learning, which feeds a video in chronological order into the end-to-end network instead of several random-sampling frames when training. Sequence learning helps our model be better aware of the concept of tracking and recognition of object. We evaluated the proposed algorithm on the second YouTube-VOS test-challenge set and achieved a $\mathcal{J}\&\mathcal{F}$ mean score of 81.7%, ranked the second place on the VOS track. In the challenge, our method only uses ResNet-50 as the backbone and our score is very slightly worse than the first place score, i.e.*, 0.1%, which implies that our VOS framework is the state-of-the-art one.*

1. Introduction

Semi-supervised Video Object Segmentation (VOS) is a fundamental task in computer vision for years and widely applied in video editing, autonomous driving, *etc*. Given a video and the first frame's annotation of single or multiple object(s), the algorithm has to provide the instance segmentation maps of the specified object(s) in the following frames. Challenges like large appearance changes, similar instance distractors, occlusion, fast motions *etc*. are frequently appeared in this task.

Recently embedding matching based methods make great advances in VOS. PML [1] assigns each pixel in current frame to foreground or background by measuring the distances between the pixel with all pixels in the reference pool. FEELVOS [9] extends it with global and local embedding matching, and implements in an end-to-end manner. STM [6] further brings the embedding matching with historical frames into the non-local operation, boosting the performance a lot. However, such embedding-based methods usually emphasize on the appearance matching. In other words, the predicted masks of historical frames are only used as a reference, but the temporal consistency is not fully utilized. As we known, the previous predicted mask is a strong spatial prior for VOS, but it is not fully used. Moreover, these methods are usually trained only with several cropped random-sampling frames for each iteration, *e.g*. all the mentioned methods (PML, FEELVOS, STM) sample three cropped frames when training, but the testing videos mostly contain dozens of frames. The difference between training and testing will bring a definite gap for algorithms.

To alleviate the problems above, we propose a novel motion-guided attention module, unifying the spatial prior and appearance matching concisely. This module takes the previous predicted mask and the current frame's feature to learn a one-channel spatial probability map, then does multiply operation with the correlation map of the current frame and the selected historical frames. This spatial probability map provides a coarse prior when predicting the current frame, so that it is able to suppress unrelated but similar pixels (False Positive) far from the interested objects in the current frame. Another contribution we made is sequence learning. For each iteration, we feed the network a full resolution video in chronological order. In this way, our model could be trained on dozens of continuous frames, which helps the network be aware of tracking concept instead of pure appearance matching.

2. Method

Our method is derived from the STM [6] framework, which classifies each pixel in the current frame a foreground or background by applying non-local matching with predicted historical frames. To be specific, we implant spa-

*The work was mainly done during an internship at Horizon Robotics.

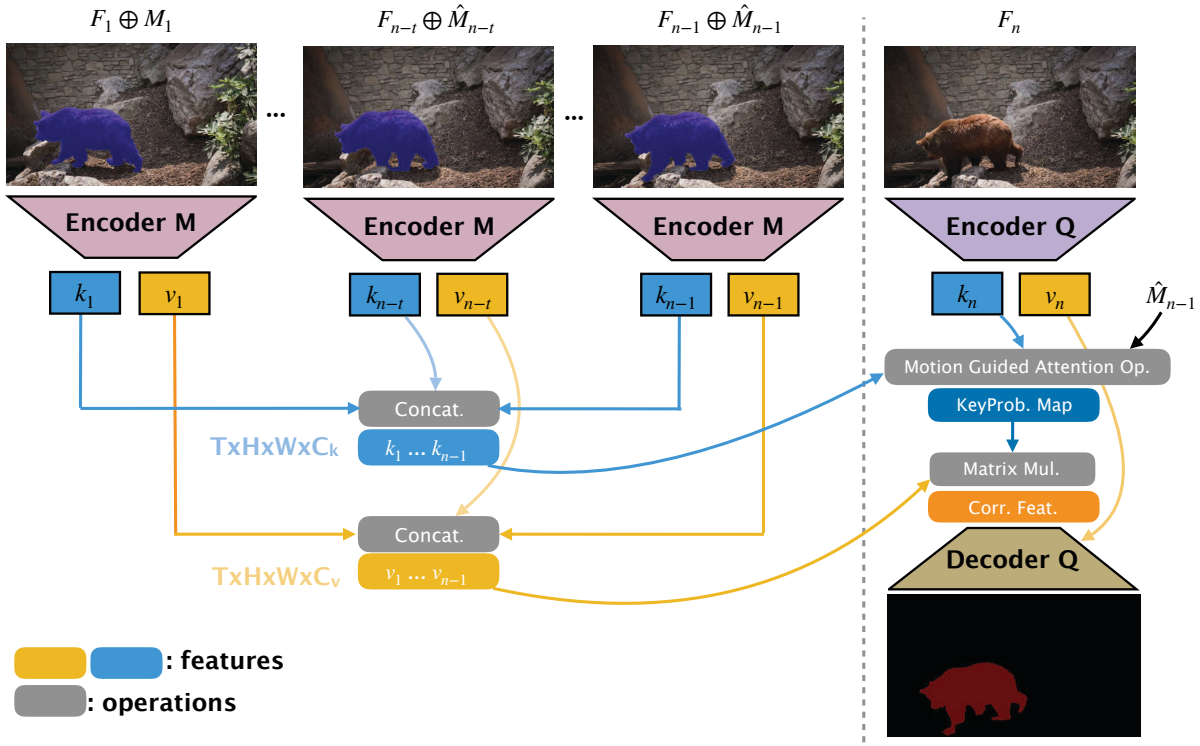


Figure 1. Overview of the proposed method. Given a certain video and the first frame's annotation (F_1 and M_1), we encode the concatenated input ($F_1 \oplus M_1$) to get the key and value embeddings (k_1 and v_1). Similarly we can obtain key and value embeddings for frames $t, ..., n-t, n-1$ before predicting frame n. For the current frame, we use another encoder to get embeddings. Then we use the previous $\hat{M}_{n-1}$ as the spatial prior, applying non-local operation between the current and all historical embeddings. The correlation feature with the current frame can be extracted from the historical value embeddings by matrix multiplication. The main differences between our method and STM are the motion-guided attention operation and optimization strategy.

tial prior into the non-local matching operator, which is described in Sec. 2.1. In Sec. 2.2, we propose an improved optimization strategy named sequence learning. The framework is shown in Fig 1.

2.1. Motion-Guided Attention Module

Our proposed motion-guided attention module aims to introduce spatial prior in the standard non-local operation when predicting the current frame, and the pipeline is depicted in Fig 2. It takes all saved historical key embeddings, current frame's key embedding and previous predicted mask as inputs, denoted as $\{k_1, ..., k_{n-1}\}$, k_n and $\hat{M}_{n-1}$ respectively.

Firstly we follow the non-local operation (Space-Time Memory Reader) proposed in STM in left part of Fig 2. The key embedding of the current frame is flatten into a vector in spatial dimension (from $H \times W \times C_k$ to $HW \times C_k$), also for the memorized key embeddings (from $T \times H \times W \times C_k$ to $THW \times C_k$, T means the number of frames). Then the correlation probability map with shape of $THW \times HW$ can be obtained by matrix multiplication and SoftMax operation. As the temporal consistency in video, the previous frame is always high correlated with the current frame in

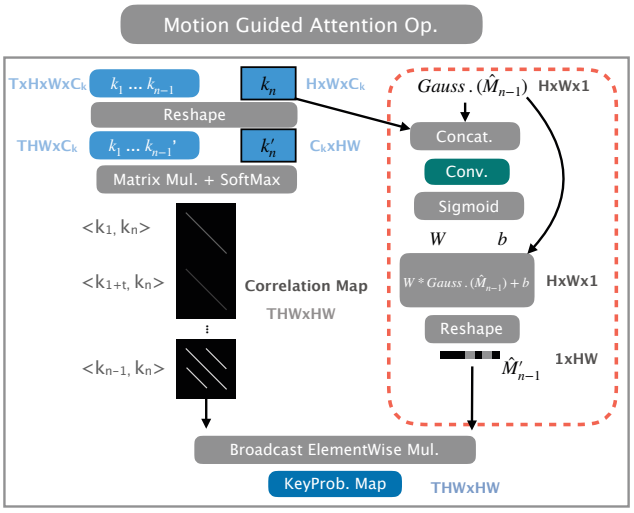
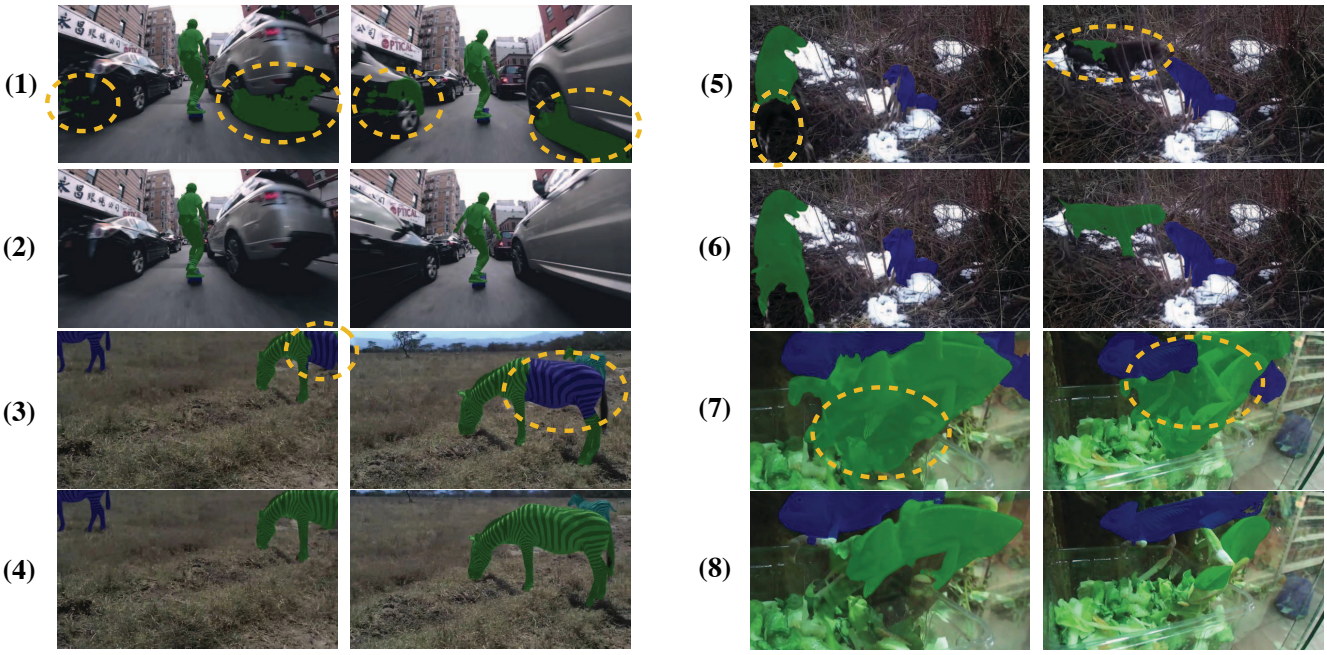


Figure 2. Motion-guided attention module diagram.

the diagonal direction. Secondly on the right half is our proposed motion-guided part. The previous predicted result $\hat{M}_{n-1}$ is a 0-1 mask with shape of $H \times W \times 1$. Specifically, we encode $\hat{M}_{n-1}$ with a two-dimensional Gaussian



Baseline (1), (3) **vs.** Baseline+Motion-Guided (2), (4)

Baseline (5), (7) **vs.** Baseline+Sequence Learning (6), (8)

Figure 3. Typical visualization comparison with the two proposed modules. The left half figure shows our motion-guided module suppresses the similar pixels and is better at recognizing a complete object. The right half shows sequence learning improves the completion of interested objects.

heatmap similar to [11]. If the mask is empty, *i.e*. the value of each pixel is 0, we simply reverse the value to 1. Then we concatenate the current key embedding k_n and the encoded mask $Gauss.(\hat{M}_{n-1})$ to learn W and b, which are parameters to control the weight of spatial prior. The shapes of W and b are both $H \times W \times 1$. After a linear combination $W * Gauss.(\hat{M}_{n-1}) + b$ and multiplication with the mentioned correlation map, we can suppress the spatial unrelated but appearance high relevant pixels effectively. We show some typical examples in Fig 3.

2.2. Sequence Learning

In the training phase, most embedding based methods sample several cropped random-sampling frames from a video, to learn the tracking concept. However, the testing videos contain mostly dozens of frames. To eliminate the training and testing gap, for each iteration we feed the network a full-resolution video in chronological order. To be specific, given the first full-resolution frame F_1 and its annotation M_1, the network outputs all frames' predictions frame by frame, *e.g*. $\hat{M}_2, \hat{M}_3, ...$ In some state we give one frame *e.g*. F_m to the network, and the network outputs the prediction $\hat{M}_m$. In the next, F_{m+1} is given to the network which outputs $\hat{M}_{m+1}$. The process is repeated till the end of the video. Because of the GPU RAM limitation, we back propagate the gradients only on each predicted frame individually. The negative impacts are that this strategy is GPU memory and computation unfriendly. In practice we use eight NVIDIA V100 GPUs with 16 GB GPU memory.

3. Experiments

We follow the training settings in STM, which uses ECSSD [8], MSRA10k [2], VOC0712 [3], COCO [5] as pretraining data, and YouTube-VOS [10] as main training data. The encoders use ResNet-50 [4] as backbone. The validation and test-challenge set (2019 version) contain 507 and 541 videos respectively. The evaluation metrics are the region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$ proposed in [7].

3.1. Results

As shown in Table 1, our proposed method achieves overall score of 81.7% on the second YouTube-VOS test-challenge set (Semi-supervised VOS Track) and ranks the second place. Our method has a better generalization ability than the first entry, both in $\mathcal{J}$ and $\mathcal{F}$, but is weaker in fitting the seen categories.

3.2. Ablation Study

To study the contribution of the two proposed components, we show some quantitative results in Table 2. The baseline is our re-implementation of STM, without any challenge tricks like online learning, multi-scale testing or hard mining.

Table 1. Ranking results in the YouTube-VOS 2019 test-challenge. In parentheses we place the ranking for each measure category. "seen" and "unseen" indicate whether the categories of tracking instances appeared in training set or not. We mark our results in blue.

Team	Overall	$\mathcal{J}_{seen}$	$\mathcal{J}_{unseen}$	$\mathcal{F}_{seen}$	$\mathcal{F}_{unseen}$
zszhou	**81.8 (1)**	**80.7 (1)**	77.3 (2)	**84.7** (1)	84.7 (2)
Ours	81.7 (2)	80.0 (2)	**77.9 (1)**	83.3 (2)	**85.5 (1)**
zxyang1996	80.4 (3)	79.4 (3)	75.9 (4)	83.3 (3)	83.1 (4)
swoh	80.2 (4)	78.8 (4)	75.9 (3)	82.5 (4)	83.5 (3)
youtube_test	79.1 (5)	77.9 (5)	74.7 (5)	81.5 (5)	82.2 (5)
Jono	71.4 (6)	70.3 (9)	68.0 (6)	73.6 (9)	74.0 (7)
andr345	71.0 (7)	69.9 (10)	66.7 (7)	73.2 (10)	74.0 (6)
hthieu	68.8 (8)	70.7 (7)	61.9 (8)	74.2 (8)	68.5 (8)
JLU_thunder	68.7 (9)	71.3 (6)	61.0 (9)	75.0 (6)	67.3 (9)
NotRaining	67.6 (10)	70.4 (8)	59.7 (10)	74.2 (7)	66.2 (10)

Table 2. Ablation study of the proposed components on YouTube-VOS 2019 validation set.

Baseline and components	Overall	Δ
Baseline (Our STM re-implementation)	70.7	-
+ Motion-Guided Attention Module	72.5	+1.8
+ Sequence Learning	72.2	+1.5
+ Motion-Guided. & Sequence Learning	73.8	+3.1

4. Conclusion

In this paper, we propose a motion-guided attention module and sequence learning for VOS. However, some hard cases are not exhaustively solved. Hybrid methods *e.g.* introducing proposals like PTSNet [12] in the embedding-based methods maybe an effective direction.

References

[1] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1189–1198, 2018.

[2] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015.

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Cconference on Computer Vision*, pages 740–755, 2014.

[6] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. *arXiv preprint arXiv:1904.00607*, 2019.

[7] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[8] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015.

[9] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019.

[10] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *European Conference on Computer Vision*, pages 603–619, 2018.

[11] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K. Katsaggelos. Efficient video object segmentation via network modulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6499–6507, 2018.

[12] Qiang Zhou, Zilong Huang, Lichao Huang, Shen Han, Yongchao Gong, Chang Huang, Wenyu Liu, and Xinggang Wang. Proposal, tracking and segmentation (pts): A cascaded network for video object segmentation. *arXiv preprint arXiv:1907.01203v2*, 2019.