The 2020 DAVIS Challenge on Video Object Segmentation - CVPR 2020 Workshops

Depth-aware Space-Time Memory Network for Video Object Segmentation

Haozhe Xie Yunmu Huang Anni Xu Jinpeng Lan Wenxiu Sun

SenseTime Research

{xiehaozhe, huangyunmu, xuanni, lanjinpeng, sunwenxiu}@sensetime.com

Abstract

In this paper, we propose Depth-aware Space-Time Memory (D-STM) Network for semi-supervised Video Object Segmentation (VOS). Space-Time Memory (STM) Network learns the feature embedding of the foreground objects and archives promising results in VOS. However, STM focus on the appearances of objects without explicitly considering the spatial location, which leads to poor segmentation results when objects having similar appearances. To solve this problem, we estimate the depth maps from a video sequence to alleviate the ambiguity of objects with similar appearances. Besides, an ASPP module is incorporated to increase the semantic receptive field on different scales. Together with the multi-scale ensemble, the proposed D-STM archives a $\mathcal{J}\&\mathcal{F}$ score of 76.9% in the 2020 DAVIS challenge on semi-supervised VOS.

1. Introduction

Video Object Segmentation (VOS) is the task of automatically generating accurate and consistent pixel masks for objects in a video sequence. It is a fundamental task with many potential applications, including video editing [13], autonomous driving [5], and 3D reconstruction [14]. In this paper, we focus on the semi-supervised VOS, which aims to segment particular object instances across the entire video sequence given the first frame ground truth annotations. It is a very challenging task due to the occlusions, drifts, and appearance change of the target objects over time.

Many VOS methods have been proposed in the past few years. Several methods [9, 11] rely on temporal continuity and propagate the segmentation mask from the first frame to the next. However, these methods have difficulties in handling occlusions and drifts due to the error accumulation. To solve this issue, STM [10] introduces the space-time memory that encodes the appearances of target objects to deal with appearance changes, occlusions, and drifts. However, STM pays little attention to the spatial location and only focuses on exploring feature matching for the target objects, which fails to distinguish objects with similar appearances.

To solve this problem, we take the spatial locations of objects into account. Intuitively, it is easy to distinguish similar objects with different spatial locations. To this aim, we present Depth-aware Space-Time Memory (D-STM) Network that introduces the depth maps to VOS (Figure 1). Specifically, we use [7] to estimate depth maps from video sequences and feed both the frames and depth maps into space-time memory networks. To further improve the segmentation accuracy, we introduce Lovász Loss [2] and replace the ResNet with ResNeSt [16]. Besides, the atrous spatial pyramid pooling (ASPP) [3] is employed to robustly segment objects at multiple scales. Experimental results on the DAVIS *test-dev* set indicate that the proposed D-STM outperforms STM and archives a $\mathcal{J}\&\mathcal{F}$ score of 76.9% in the 2020 DAVIS challenge on semi-supervised VOS.

2. Methods

2.1. Space-Time Memory Network

In this section, we briefly review the space-time memory (STM) network [10] that becomes the basic architecture of our method. STM is composed of four components, memory encoder, query encoder, space-time memory reader, and decoder. At time step t, the image I_t is regarded as a query image while the past frames $[I_0, \ldots, I_{t-1}]$ and masks $[M_0, \ldots, M_{t-1}]$ are memories. The key-value pair $\left\{k_t^Q, v_t^Q\right\}$ for time step t is encoded by the query encoder. Similarly, the memory encoder generates $\left[\left\{k_0^M, v_0^M\right\}, \ldots, \left\{k_{t-1}^M, v_{t-1}^M\right\}\right]$ by encoding the past frames and masks, where key is used for encoding semantics for matching robust to appearance variations and value stores detailed information for producing the mask estimation. The space-time memory reader produces features f_t for I_t by combining the keys and values from memories, which can be formulated as

$$f_t = \left[v_t^Q, \frac{1}{Z}\sum_{i=0}^{t-1} R(k_t^Q, k_i^M) v_i^M\right] \tag{1}$$

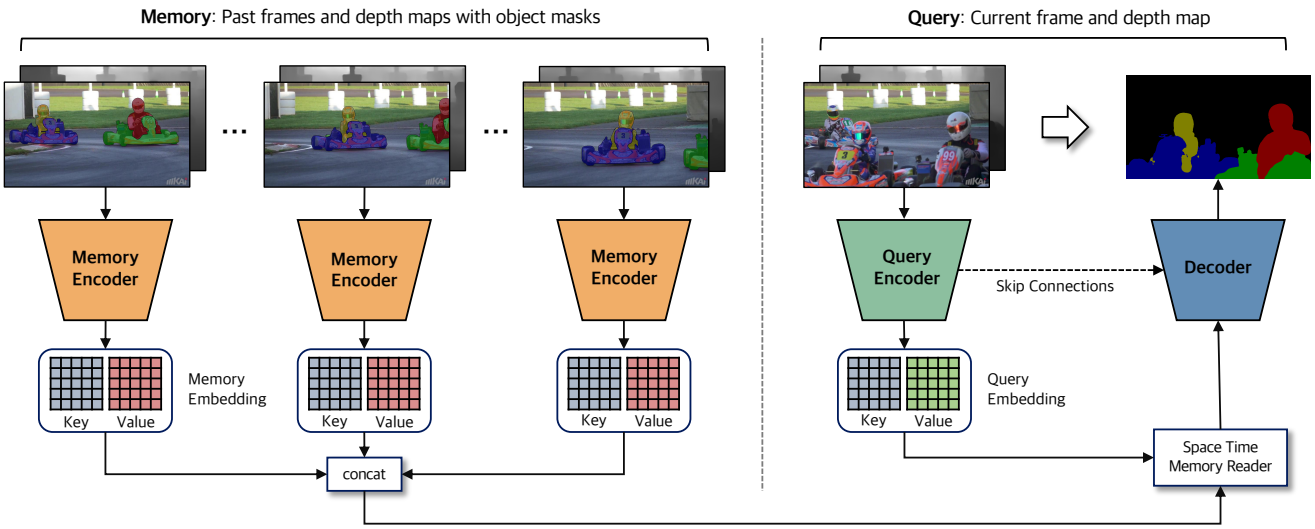

Figure 1. Overview of D-STM. D-STM consists of two encoders for the memory and the query frame, a space-time memory read block, and a decoder. The memory encoder takes an RGB frame, a depth map and the object mask. The query encoder takes the query image and depth map as input.

where $[\cdot]$ denotes the concatenation operation , Z represents the normalization factor, and $R(\cdot)$ is the correlation function to measure the similarity between the current frame and memories. Formally,

$$Z = \sum_{i=0}^{t-1} R(k_t^Q, k_i^M) \tag{2}$$

$$R(k_t^Q, k_i^M) = \exp(k_t^Q \cdot k_i^M) \tag{3}$$

The f_t is then fed into the decoder to obtain the final mask of the current frame.

2.2. Depth-aware Space-Time Memory Network

Depth Estimation. Depth maps provide spatial locations of all objects of a frame, which is helpful to alleviate the ambiguity of objects with similar appearances. Moreover, it provides an initial segmentation of an object. However, recovering the dense depth maps from a video is challenging because the moving objects violate the epipolar constraint used in 3D vision, and are often treated as noise or outliers in existing structure from motion (SfM) and multi-view stereo (MVS) methods. To solve this problem, we follow [7] to train a neural network on videos where people imitate mannequins, *i.e.,* freeze in elaborate, natural poses, while a hand-held camera tours the scene. Once trained, the network can handle natural videos with an arbitrary camera and human motion. In D-STM, the backbone network (*i.e.,* ResNeSt50) take the both RGB frame and depth maps as input, as shown in Figure 1.

ResNeSt50. In D-STM, ResNeSt50 is adopted as the backbone in both memory encoder and query encoder. As discussed in [16], ResNet is originally designed for image classification and may not be suitable for image segmentation due to the limited receptive-field size and lack of cross-channel interaction. In contrast, ResNeSt incorporates feature map split attention within the individual network blocks. More specifically, each of blocks divides the feature map into several groups (along the channel dimension) and finer-grained subgroups or splits, where the feature representation of each group is determined by a weighted combination of the representations of its splits (with weights chosen based on the global contextual information). Besides, ResNeSt requires no more computation than existing ResNet-variants, and is easy to be adopted as a backbone for other computer vision tasks.

ASPP. One of the major challenges in the segmentation is caused by the existence of objects at multiple scales. To solve this issue, we introduce ASPP [3] to probe an image with multiple filters that have complementary effective fields of view, thus capturing objects as well as useful image context at multiple scales. Consequently, D-STM archives better results in segmenting small objects.

Lovász Loss. Lovász Loss [2] is used to direct optimization of the mean intersection-over-union loss in neural networks in the context of VOS. Formally, the Lovász Loss can be defined as

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \overline{\Delta_{J_c}}(\hat{\mathbf{y}}, \mathbf{y}) \tag{4}$$

Let $\hat{\mathbf{y}}$ and $\mathbf{y}$ be the segmentation output and ground truth, respectively. The set of mispredicted pixels $\mathbf{M}_c$ for class c can be defined as

$$\mathbf{M}_c(\hat{\mathbf{y}}, \mathbf{y}) = \{\mathbf{y} = c, \hat{\mathbf{y}} \neq c\} \cup \{\mathbf{y} \neq c, \hat{\mathbf{y}} = c\} \tag{5}$$

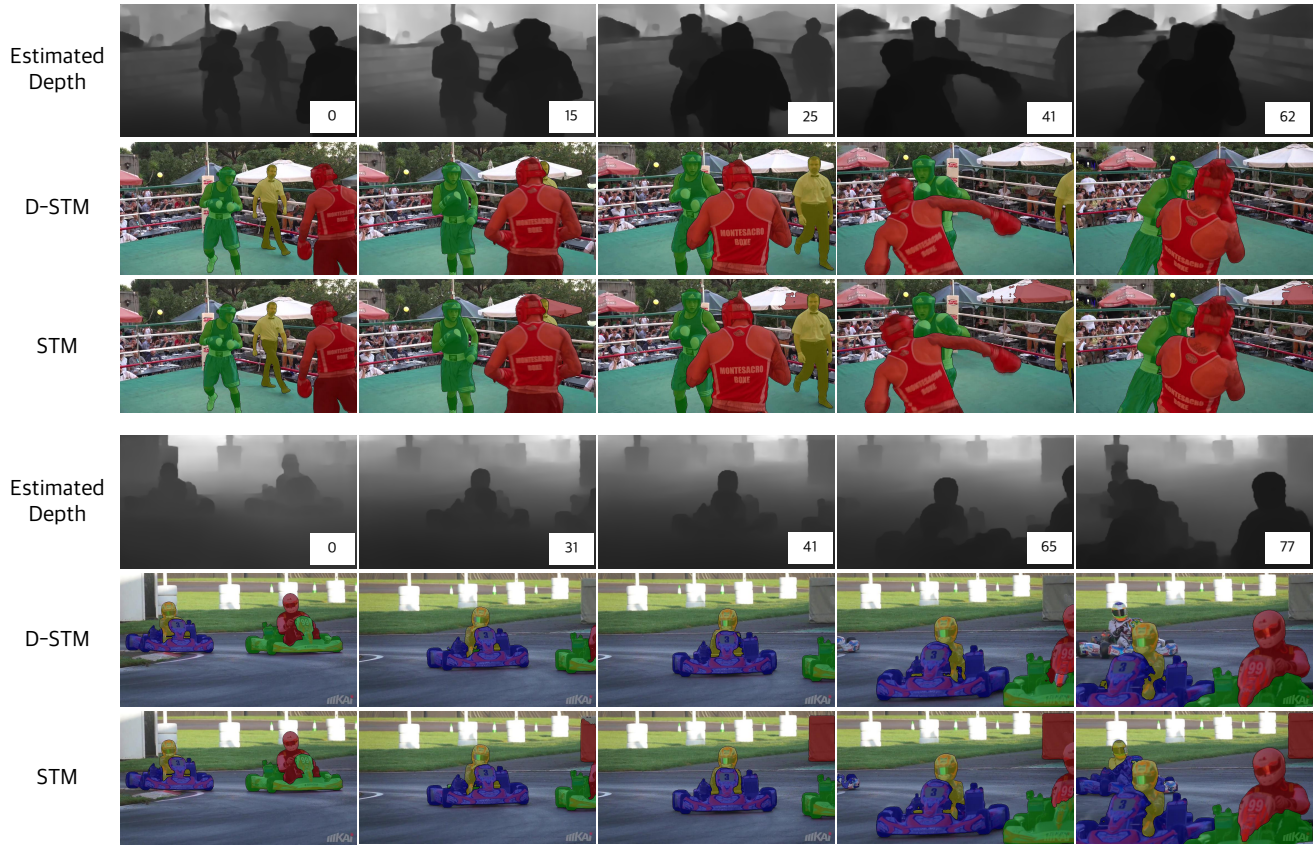

Figure 2. The depth estimation and video object segmentation results on the DAVIS *test-challenge* set. Frames are sampled at important moments (*e.g.,* before and after occlusions).

Therefore, the Jaccard loss Δ_{J_c} can be written as a function of the set of mispredictions

$$\Delta_{J_c} : \mathbf{M}_c \in \{0,1\}^p \mapsto \frac{|\mathbf{M}_c|}{|\{\mathbf{y}=\mathbf{c}\} \cup \mathbf{M}_c|} \tag{6}$$

To use Jaccard loss in a a continuous optimization, the Lovász extension of a set function $\Delta : \{0,1\}^p \rightarrow \mathbb{R}$ such that $Delta(\mathbf{0}) = 0$ is defined by

$$\overline{\Delta} : \mathbf{m} \in \mathbb{R}^p \mapsto \sum_{i=1}^{p} m_i g_i(\mathbf{m}) \tag{7}$$

with

$$g_i(\mathbf{m}) = \Delta(\{\pi_1, \ldots, \pi_i\}) - \Delta(\{\pi_1, \ldots, \pi_{i-1}\}) \tag{8}$$

where $\mathbf{m}$ is the a vector of pixel errors and π is a permutation ordering the components of $\mathbf{m}$ in decreasing order, *i.e.,* $x_{\pi_1} \geq x_{\pi_2} \geq x_{\pi_p}$.

3. Experiments

3.1. Implementation Details

We use a two-stage training strategy to train the D-STM following [10]. Specifically, D-STM is firstly pretrained on a simulation datasets generated from several image datasets, including MSRA10K [4], MSCOCO [8], and PascalVOC [6]. After pretraining, D-STM is fed with real videos for main training. Both videos from the DAVIS [12] and YouTube-VOS [15] datasets are used in this stage. Particularly, we sample 3 temporally ordered frames from the video during training. To learn the appearance change over a long time, we randomly skip frames during the sampling. The number of frames to skip is gradually increased from 0 to 25 during the training as in curriculum learning [1].

We randomly crop 384×384 patches from images for training. The batch size is set to 12 using four NVIDIA TITAN Xp GPUs. We adopt an Adam optimizer with a β_1 of 0.9 and β_2 of 0.999. The initial learning rate is set to 10^{-5} and decayed by 2 after 100 epochs. Both pretraining and main training is set to stop after 150 epochs.

3.2. Evaluation on the DAVIS Benchmark

We evaluate D-STM on the 2020 DAVIS challenge on semi-supervised VOS. The *test-challenge* set is composed of 92 objects in 30 videos. The quantitative results are reported in Table 1. Figure 2 shows estimated depth maps and qualitative examples of D-STM and STM, which indicates

Table 1. The semi-supervised video segmentation results on the DAVIS *test-challenge* set. Our results are highlighted in bold.

#	User	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\ \&\ \mathcal{F}$
1	pengzhang0x	0.815	0.867	0.841
2	captain	0.811	0.865	0.838
3	hongje	0.770	0.821	0.795
4	tmtriet	0.765	0.821	0.793
5	**hzxie**	**0.744**	**0.795**	**0.769**
6	vltanh	0.733	0.787	0.760
7	JingshanXu	0.722	0.777	0.750
8	littleboy	0.698	0.746	0.722
9	mingmingdiii	0.676	0.722	0.699
10	Mustansar	0.625	0.673	0.649

Table 2. The ablation studies for D-STM on the DAVIS *validation* set. Note that Lovász denotes the Lovász Loss and TTA represents the multi-scale ensemble during inference.

ResNeSt	Lovász	Depth	ASPP	TTA	$\mathcal{J}\&\mathcal{F}$
					0.783
✓					0.799
✓	✓				0.805
✓	✓	✓			0.813
✓	✓	✓	✓		0.820
✓	✓	✓	✓	✓	**0.835**

that D-STM performs better than STM in distinguishing objects with similar appearances.

3.3. Ablation Study

To demonstrate the effectiveness of the key components, we conduct ablation studies for D-STM. As shown in Table 2, test-time augmentation (TTA) brings the largest improvement in terms of $\mathcal{J}\&\mathcal{F}$ of 0.02. Removing either Depth, ResNeSt, ASPP, or Lovász loss causes considerable degeneration in segmentation accuracy.

4. Conclusion

In this paper, we propose Depth-aware Space-Time Memory (D-STM) network for semi-supervised video object segmentation. The depth maps estimated from video sequences alleviate the ambiguity of objects having similar appearances. Together with ResNeSt, ASPP, Lovász Loss, and the test-time augmentation, the proposed D-STM outperforms STM [10] on the DAVIS benchmark and archives a $\mathcal{J}\&\mathcal{F}$ score of 76.9% in the 2020 DAVIS challenge on semi-supervised video object segmentation.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009. 3

[2] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018. 1, 2

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018. 1, 2

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

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1

[6] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. 3

[7] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 1, 2

[8] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 3

[9] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV 2018*, 2018. 1

[10] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 1, 3, 4

[11] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017. 1

[12] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 3

[13] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, 2020. 1

[14] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2Vox: Context-aware 3D reconstruction from single and multi-view images. In *ICCV*, 2019. 1

[15] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. YouTube-VOS: A large-scale video object segmentation benchmark. *arXiv*, 1809.03327, 2018. 3

[16] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander J. Smola. Resnest: Split-attention networks. *arXiv*, 2004.08955, 2020. 1, 2