Going Deeper into Embedding Learning for Video Object Segmentation

Zongxin Yang Peike Li Qianyu Feng Yunchao Wei Yi Yang
ReLER, Centre for Artificial Intelligence, University of Technology Sydney
{zongxin.yang,peike.li,qianyu.feng}@student.uts.edu.au {yunchao.wei,yi.yang}@uts.edu.au

Abstract

*In this paper, we investigate the principles of consistent training, between given reference and predicted sequence, for better embedding learning of semi-supervised video object segmentation. To accurately segment the target objects given the mask at the first frame, we realize that the expected feature embeddings of any consecutive frames should satisfy the following properties: **1)** global consistency in terms of both foreground object(s) and background; **2)** robust local consistency under a various object moving rate; **3)** environment consistency between the training and inference process; **4)** receptive consistency between the receptive fields of network and the variable scales of objects; **5)** sampling consistency between foreground and background pixels to avoid training bias. With the principles in mind, we carefully design a simple pipeline to lift both accuracy and efficiency for video object segmentation effectively. With the ResNet-101 as the backbone, our single model achieves a $\mathcal{J}\&\mathcal{F}$ score of 81.0% on the validation set of Youtube-VOS benchmark without any bells and whistles. By applying multi-scale & flip augmentation at the testing stage, the accuracy can be further boosted to 82.4%. Code will be made available.*

1. Introduction

Semi-supervised Video Object Segmentation (VOS) targets on segmenting a particular object instance across the entire video sequence based on the object mask given at the first frame. The VOS is a fundamental task in computer vision with many potential applications, including interactive video editing, augmented reality, and self-driving cars. A recent work, FEELVOS [10], uses a semantic pixel-wise embedding together with a global (between reference and current frames) and a local (between previous and current frames) matching mechanism to transfer information from the first frame and form the previous frame of the video to the current frame. In contrast to some previous works (*e.g.* PReMVOS [7]), which are complicated and heavily rely on fine-tuning on the first frame, FEELVOS enables the network can be learned in an end-to-end manner for the multiple object segmentation task. Although the matching mechanism proposed in [7] looks simple, it is actually beneficial.

Even though significant progress has been made in the research fields of VOS, the current state-of-the-art works are still with some significant problems or defects. **First**, previous works always focus on keeping consistency on only foreground objects. However, an excellent segmented background is equally important as the foreground. **Second**, the local matching between previous and current frames is usually limited in a fixed extent of neighboring pixels in the previous works. However, the offset of objects between two adjacent frames in real videos is often variable in terms of different moving rate or frame rate. For instance, the frame rate of the Youtube-VOS is 6 fps, which is much slower than of DAVIS benchmark (*i.e.* 24 fps) [8], leading to a larger variance of appearance for the objects across two adjacent frames. **Third**, in the training process of previous works, the mask of previous frames is always from ground truth data, which is not consistent with the situation at the inference stage, *i.e.* the mask of previous frames is generated by network itself. **Moreover**, the receptive fields of guidance information are usually not robust to different scales of objects for most previous works. For example, the matching mechanism in FEELVOS works on pixel-level (with a stride of 4), which is not sufficient and robust to match those objects with large scale. **Finally**, there is an apparent imbalance between foreground pixels and background pixels during the training process. The region of foreground objects is always much smaller than the background, which makes the networks easier to over-fit on background attributes.

To relief or overcome the above-mentioned issues, we propose an Embedding Learning with Consistency Perception (ELCP) approach for fast and robust semi-supervised VOS. **First**, apart from only matching on foreground objects, we make consistency in terms of both foreground objects and background by additionally leveraging background pixels to match background globally and locally. **Second**, we apply a multi-scale mechanism for local matching and let the network to learn how to choose the best local scale adaptively. In addition, we concatenate the embedding

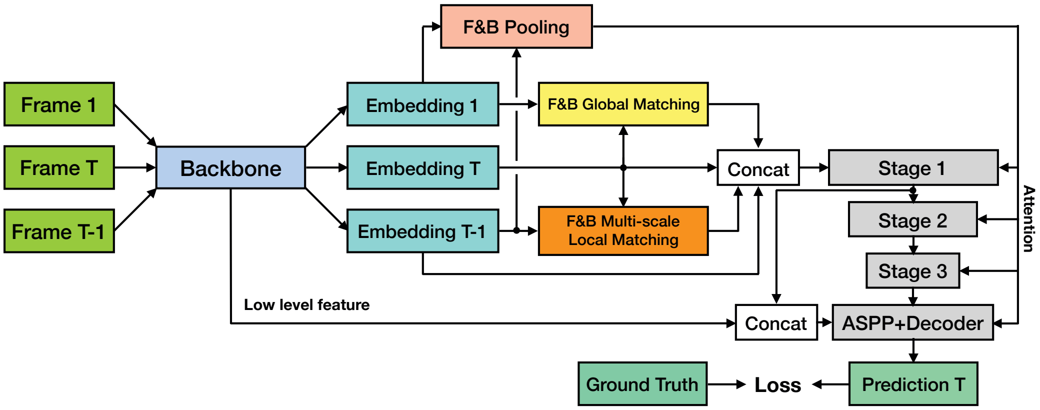

Figure 1. An **overview** of the structure of ELCP. "F&B" means "Foreground and Background".

features from the previous frame to the current one. Such an operation has been proven to be useful in learning the offset of content between adjacent frames in [3]. **Third**, we design a sequential training method for VOS to compel the network to keep the integrity of instances (or objects) during sequential prediction, which is closer to the environment at the inference stage. **Moreover**, we redesign a deep segmentation module for making larger receptive fields, which is helpful to relief the local ambiguities [9]. Beyond the pixel-level matching, we design an instance-level light-weight attention mechanism to guide the segmentation further. **Finally**, we design a balanced random-crop augmentation method, which crops a sequence of frames together with the same window and restricts the crop region to contain enough foreground information. All these proposed strategies can significantly improve the quality of the learned embeddings for conducting VOS while keeping the network simple yet effective simultaneously.

We conduct extensive experiments on the validation set of YouTube-VOS [11] 2019 to evaluate the effectiveness of the proposed ELCP approach. Using ResNet-101 [5] & Deeplab-V3+ [2] as backbone, our single model achieves $\mathbf{81.0}\%$ (w/o multi-scale & flip) and $\mathbf{82.4}\%$ (w/ multi-scale & flip) $\mathcal{J}\&\mathcal{F}$ on the validation set. We hope our ELCP will serve as a solid baseline to help ease future research in video object segmentation. We will make our code publicly available soon.

2. Approach

Compared to previous works, our proposed ELCP has advantages in two aspects. First, on the aspect of **model architecture**, we improve the robustness of local matching under various object moving rates. Furthermore, we augment the guidance information by matching on both foreground object(s) and background with both pixel-level matching and instance-level attention. Moreover, we make larger receptive fields to perceive individual instance of various scales. Second, on the aspect of **training method**, we design a balanced random-crop augmentation method to generate more appropriate foreground/background ratio to avoid training bias. Also, we design a novel training method for VOS to compel the network to keep the integrity of foreground instances (or objects) during sequence prediction. We show an overview of our model architecture in Fig. 1.

2.1. Model Architecture

2.1.1 Foreground and Background Matching

Similar to FEELVOS, we use a global and local matching mechanism to guide the segmentation in the current frame. The difference from FEELVOS is that we additionally incorporate background information. Concretely, the foreground matching is the same as the matching method proposed in FEELVOS [10]. Let $\mathcal{P}_t$ denote the set of all pixels (with a stride of 4) at time t and $\mathcal{P}_{t,o} \subseteq \mathcal{P}_t$ is the set of pixels at time t which belong to foreground object o. The global foreground matching between pixels p and q is,

$$G_{t,o}(p) = \min_{q \in \mathcal{P}_{1,o}} d(p,q), \tag{1}$$

where $d(p,q)$ is the distance defined in FEELVOS [10]. And the local foreground matching is,

$$\hat{G}_{t,o}(p) = \begin{cases} \min_{q \in \mathcal{P}_{t-1,o}} d(p,q) & \text{if } \mathcal{P}_{t-1,o} \neq \emptyset \\ 1 & \text{otherwise} \end{cases}. \tag{2}$$

Similarly, let $\mathcal{P}^{'}_{t,o} = \mathcal{P}_t \backslash \mathcal{P}_{t,o}$ denote the set of background pixels of object o at time t. The global background matching is,

$$G^{'}_{t,o}(p) = \min_{q \in \mathcal{P}^{'}_{1,o}} d(p,q). \tag{3}$$

And the local background matching is,

$$\hat{G}^{'}_{t,o}(p) = \begin{cases} \min_{q \in \mathcal{P}^{'}_{t-1,o}} d(p,q) & \text{if } \mathcal{P}^{'}_{t-1,o} \neq \emptyset \\ 1 & \text{otherwise} \end{cases}. \tag{4}$$

2.1.2 Foreground and Background Attention

In addition to the pixel-level matching, we design an instance-level light-weight attention mechanism to guide the segmentation further. Inspired by SE-Net [6], we first separately embed the foreground and background pixels of the reference and previous frames (*i.e.* $\mathcal{P}_{1,o}$, $\mathcal{P}_{t-1,o}$, $\mathcal{P}^{'}_{1,o}$, and $\mathcal{P}^{'}_{t-1,o}$) using average pooling. After generating the embedding vector, we use fully connected layers with non-linear activation to learn guidance information and adjust the channel scale of feature maps in our segmentation output module.

2.1.3 Multi-scale Local Matching

In the FEELVOS, the local matching is limited in a fixed extent of neighboring pixels, but the offset of objects across two adjacent frames in real videos is variable. In our ELCP, we apply the local matching mechanism on different scales and let the network to learn how to select the best local scale, which makes our framework more robust to various object moving rates. Benefiting from the effective engineering design, the increase of computational resource of our multi-scale matching is negligible.

In addition to the multi-scale local matching, we concatenate the embedding feature from the previous frame to the current frame. This simple concatenation method has proven to be useful in learning the offset of content (optical flow) between consecutive frames in [3].

2.1.4 Deep Segmentation Module

To relief the problem of local ambiguities, we design a deep segmentation module for making larger receptive fields. Inspired by ResNets [5] and Deeplabs [1, 2], which both have shown significant representational power in image segmentation tasks, our deep segmentation module contains three stages of Res-Blocks. The number of Res-Blocks in Stage 1, 2, and 3 are separately 2, 2, 3. At the beginning of Stage 2 and stage 3, the feature maps will be downsampled by a Res-Block with a stride of 2. After these three stages, similar to Deeplabs [1, 2], we employ the Atrous Spatial Pyramid Pooling (ASPP) module to increase the receptive fields further and make the network more robust to different scale of objects. Besides, we further use one decoder module to refine the boundary of prediction by utilizing the information from low-level layers of the backbone and our deep segmentation module.

2.2. Training Method

2.2.1 Balanced Random Crop

There is an apparent imbalance between foreground pixels and background ones on YouTube-VOS [11]. However, previous works for VOS did not focus on this problem, which makes the models easier to over-fit on background attributes and decrease the generalization ability. In order to relieve this problem, we design a specific balanced random-crop augmentation method, which crops a sequence of frames (*i.e.* reference frame, previous frame, and current frame) together with the same crop window and restricts the crop region to contain enough foreground information. The restriction method is simple but effective. Specifically, our balanced random-crop method will decide on whether the randomly cropped frame contains enough pixels from foreground objects or not. If not, the method will continually take the cropping operation until one expected frame is obtained.

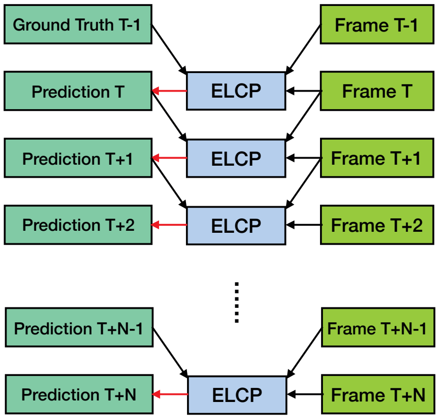

Figure 2. An illustration of the sequential training method. For brevity's sake, we omit the reference frame (*i.e.* Frame 1 and Ground Truth 1) used for the global matching in all the steps.

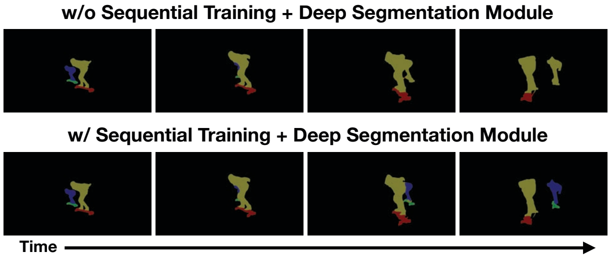
w/o Sequential Training + Deep Segmentation Module
w/ Sequential Training + Deep Segmentation Module
Time

Figure 3. The proposed sequential training and deep segmentation module effectively relief the problem of local ambiguities.

Moreover, our method will adjust the sequential frames to keep consistency with the reference frame. By doing this, we can successfully avoid false foreground object labels.

2.2.2 Sequential Training

To relief the problem of local ambiguities and over-fitting further, we design a sequential training method, which trains the network using a sequence of consecutive frames in each iteration. An illustration is shown in Fig. 2. During the sequential training process, we use the prediction in the last step to guide the segmentation in the next step. This method compels the network to learn how to keep the consistency of objects during sequence prediction at the inference stage, which is more similar to real video object segmentation environment.

Approach	score	boost
FEELVOS (after adjusting hyper-parameters for YouTube-VOS)	75.1%	-
+ Foreground and Background Matching	76.2%	1.1%
+ Foreground and Background Attention	77.1%	0.9%
+ Balanced Random Crop	78.4%	1.3%
+ Deep Segmentation Module	79.5%	1.1%
+ Multi-scale Local Matching	80.2%	0.7%
+ Sequential Training	81.0%	0.8%
+ Multi-scale & Flip in Testing	82.4%	1.4%

Table 1. The ablation study experiments on the validation set of YouTube-VOS 2019.

3. Experiments

We evaluate our method on the YouTube-VOS 2019 dataset, which contains 3471 videos in the training set, 507 videos in the validation set (26 unseen categories in training), and 541 videos in the test set (29 unseen categories). The evaluation metric is the $\mathcal{J}$ score, calculated as the average IoU between the prediction and the ground truth mask, and the $\mathcal{F}$ score, calculated as an average boundary similarity measure between the boundary of the prediction and the ground truth, and their average value over the seen and unseen categories.

3.1. Experimental Results

On the validation set, our ELCP (single model) achieves a $\mathcal{J}\&\mathcal{F}$ mean score over both the seen and unseen categories of 81.0% without any bells and whistles. By applying multi-scale & flip augmentation during the evaluation, the score can be further boosted to 82.4%. **Notably, the ELCP trained with short training schedule achieves 80.3 on the validation set and 80.4 on the test set (w/o multi-scale & flip), which ranks 3rd in Track 1 (VOS) of the 2nd Large-scale Video Object Segmentation Challenge. Moreover, our ELCP can be further applied to conduct the challenging video instance segmentation task [4].**

Moreover, the speed (w/o multi-scale & flip) of the proposed ELCP is about 3 fps in evaluation on YouTube-VOS (720P videos) using single Tesla V100 GPU, which is much faster than previous works with first-frame fine-tuning (*e.g.* [7]).

3.2. Ablation Study

We also study the contribution of all the components and methods in our framework. As shown in Table 1, all the components, and methods we proposed show significant improvement in performance. Fig. 3 gives a result comparison to demonstrate the effectiveness of Sequential Training and Deep Segmentation Module further. By applying all the proposed methods together, we extremely boost the performance of baseline from 75.1% to 82.4% on the validation set of Youtube-VOS 2019.

4. Conclusion

In this paper, we proposed a novel framework for semi-supervised video object segmentation by going deeper into embedding learning. The proposed approach, including both architectures and training methods for more robust embedding learning, significantly improves the quality of VOS but keeps the network simple and fast.

References

[1] 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, 2017.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.

[3] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015.

[4] Qianyu Feng, Zongxin Yang, Peike Li, Yunchao Wei, and Yi Yang. Dual embedding learning for video instance segmentation. In *IEEE International Conference on Computer Vision Workshop*, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[7] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, pages 565–580, 2018.

[8] 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 *CVPR*, pages 724–732, 2016.

[9] Antonio Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, 2003.

[10] 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*, pages 9481–9490, 2019.

[11] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.