E3SN: Efficient End-to-End Siamese Network for Video Object Segmentation

Meng Lan , **Yipeng Zhang** , **Qinning Xu** and **Lefei Zhang** *

National Engineering Research Center for Multimedia Software,
Institute of Artificial Intelligence and School of Computer Science, Wuhan University, Wuhan, China
{menglan, zyp91, qinning.xu, zhanglefei}whu.edu.cn

Abstract

In the semi-supervised video object segmentation (VOS) field, SiamMask has achieved competitive accuracy and the fastest running speed. However, the two-stage training procedure requires additional manual intervention, and using only single-level features does not maximize the rich hierarchical feature information. This paper proposes an efficient end-to-end Siamese network for VOS. In particular, a supervised sampling strategy is designed to optimize the training procedure. Such an optimization facilitates the training of the entire model in an end-to-end manner. Moreover, a multilevel feature aggregation module is developed to enhance feature representability and improve segmentation accuracy. Experimental results on DAVIS2016 and DAVIS2017 datasets show that the proposed approach outperforms the SiamMask in accuracy with similar FPS. Moreover, this approach also achieves good accuracy-speed trade-off compared with that of other state-of-the-art VOS algorithms.

1 Introduction

Semi-supervised video object segmentation (VOS) [Perazzi *et al.*, 2016], in which the ground truth information of the object is given in the first frame, aims to find the pixel-level position of specified object(s) in a short video. This challenging task is fundamental for high-level video analysis tasks such as video understanding, and importantly useful in video editing [Wang *et al.*, 2017]. VOS in the following part refers to semi-supervised VOS.

Early VOS methods [Caelles *et al.*, 2017; Luiten *et al.*, 2018] are mainly based on online learning (OL), which needs fine-tuning on the first frame of the test video. Fine-tuning generally leads to low frames-per-second (FPS) speed. Subsequently, matching- and propagation-based methods [Chen *et al.*, 2018b; Oh *et al.*, 2018; Yang *et al.*, 2018] are proposed for fast VOS by disregarding fine-tuning. Matching-based approaches employ a metric learning strategy to calculate similarity maps between the features of the first frame and those of the current frame. By contrast, propagation-based methods mainly exploit consistency between video frames and propagate the mask of the first or the previous frame as supervision to the current one. However, both methods suffer from suboptimal accuracy due to mismatching or drifting problems. Moreover, although these offline methods are faster than the OL-based methods, their speed remains unsatisfactory.

Providing a strong pixel-level prior mask as the guidance in the first frame for VOS is difficult in practical applications. Thus, finding a relatively weak prior knowledge in testing process is crucial in facilitating a friendly human-computer interaction process [Han *et al.*, 2014; Song *et al.*, 2020]. An example of a weak prior knowledge is a boundary box, which merely indicates the location of the target.

SiamMask [Wang *et al.*, 2019] which has the fastest speed at present with competitive accuracy, recently achieved good accuracy and speed trade-off from the perspective of tracking. This method extends the SiamRPN tracker [Li *et al.*, 2018] by adding a branch as postprocessing to produce the segmentation mask of the specified object. However, as illustrated in Fig. 1 (a), the high-accuracy SiamMask with mask refinement module has to be fine-tuned on a base model to avoid excessive memory requirements. This condition results in a two-stage training manner. Additional manual intervention is also required in this process. That is, the base model must be evaluated to find the optimal model parameters, and different hyper-parameters must be set for the refined model. Thus, additional manual interventions are inconvenient in practice. Moreover, the mask refinement module only takes the single-level features as input, thereby lacking rich feature representation.

In this paper, A supervised sampling strategy is proposed to address the aforementioned flaws and reduce the training samples in the mask refinement pathway, which significantly reduces the memory requirements (Fig. 1 (b)). An end-to-end framework is then constructed for VOS. Furthermore, a multilevel feature aggregation module is presented to enrich the feature representation and aggregate multilevel upsampled depth-wise cross-correlation feature maps [Li *et al.*, 2019; Zhang *et al.*, 2016]. The segmentation performance of the end-to-end model is boosted through the aforementioned techniques.

In the testing process, the proposed E3SN can only require a single bounding box initialization rather than a pixel-

*Corresponding author

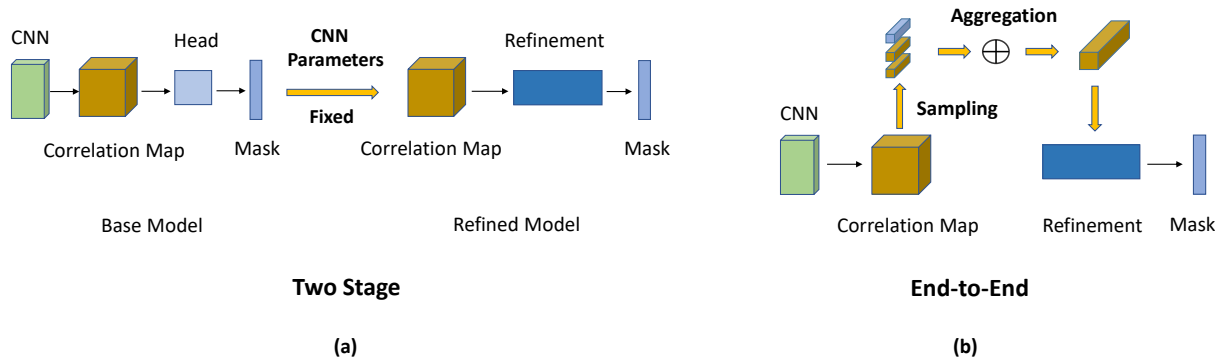

Figure 1: (a): Two-stage SiamMask; (b): End-to-end E3SN. The proposed sampling strategy alleviates the training gap of SiamMask.

level object mask to specify the object and then operate online without fine-tuning. Both conditions indicate the convenience and remarkable generalization capability of E3SN in practical use. The experimental results on DAVIS2016 and DAVIS2017 demonstrate that the proposed E3SN outperforms the SiamMask and achieves competitive performance compared with some matching- and propagation-based VOS algorithms. However, E3SN runs considerably fast.

Overall, the main contributions of this work are listed below in threefold:

- A supervised sampling strategy is proposed to alleviate the training gap of SiamMask and a fast end-to-end Siamese network is further developed for the semi-supervised VOS task.
- A multilevel feature aggregation module is developed to improve the segmentation performance. Multilevel correlation maps computed from hierarchical backbone features are fused to enrich features representation before feeding into the mask refinement pathway.
- The experiments on the two benchmark datasets demonstrate that the proposed E3SN outperforms the SiamMask in accuracy with similar FPS. Moreover, E3SN achieves good accuracy-speed trade-off with bounding box initialization compared with that of other state-of-the-art VOS algorithms.

2 Related Work

Online learning based methods. Early VOS approaches commonly involve online learning, which comprises a fine-tuning procedure using the first-frame ground truth. OSVOS [Caelles *et al.*, 2017] adopts a pre-trained CNN for foreground-background segmentation and fine-tunes it on the first frame of the test video to extract a specific object. OSVOS-S [Maninis *et al.*, 2019] enhances the performance with semantic information from an instance segmentation network. PReMVOS [Luiten *et al.*, 2018] incorporates four different techniques, such as optical flow, re-identification, integration with a blending algorithm, and extensive fine-tuning, to outperform its rivals in the 2018 DAVIS Challenge.

All the aforementioned approaches achieve impressive performance and prove online learning as an effective technique for VOS. However, satisfying the requirement of sufficient speed in real-time tasks is difficult due to the high computational cost.

Propagation based methods. Propagation-based methods resort to previous frames to capture the temporal coherence of sequential frames. MaskTrack [Perazzi *et al.*, 2017] propagates the segmentation mask of one frame to the next using optical flow. OSMN [Yang *et al.*, 2018] proposes a modulator network trained to manipulate the intermediate layers of the segmentation network. RGMP [Oh *et al.*, 2018] introduces a deep Siamese encoder-decoder network. The two encoder streams respectively encode the video frame with the estimated segmentation mask of the previous frame and the ground-truth segmentation mask of the first frame beforeintegration of the features by a global convolution block. Although the mask propagation strategy is effective for VOS, drifting remains a major problem with the presence of fast motions between sequential frames.

Matching based methods. Matching-based methods solve the VOS task from the perspective of metric learning which segments the current frame based on the pixel-wise matching distance between the features of the current and reference frames. PML [Chen *et al.*, 2018b] first establishes an embedding model with a triplet loss and represents each pixel as an embedding vector. A kNN classifier is then deployed to segment the frame. However, the point-to-point matching strategy often introduces noise. VideoMatch [Hu *et al.*, 2018] uses a soft matching layer to match extracted features and generate smooth predictions. However, the mismatching problem is due to the direct derivation of the final segmentation result from the matching of embedding space. FEELVOS [Voigtlaender *et al.*, 2019] proposes global and local pixel-level matching mechanisms to utilize rich information from the first and previous frames and only calculates extreme value maps to reduce computation. This approach results in information loss in the segmentation process. Matching-based methods achieve competitive accuracy in comparison to OL-based methods, while the problems of mismatching and relatively low running speed still exist.

SiamMask. SiamMask [Wang *et al.*, 2019] recently addresses the VOS problem from a new aspect involving the location of the target followed by segmentation. This approach extends a tracking model [Li *et al.*, 2018] by adding a mask generation branch and runs an order of magnitude faster than

that of top VOS methods with competitive accuracy. However, predicting a mask for each element of the correlation map introduces substantial redundancy, resulting in a separate two-stage training process. Moreover, the mask refinement module considering only single-level features as input ignores the multilevel feature information in the backbone to obtain strong representational features for the final mask refinement pathway. Thus, a supervised sampling strategy is proposed in the current study to reduce the redundancy and realize multilevel feature aggregation to further boost the performance. The proposed strategy reaches high accuracy with fast running speed and achieves a good speed-accuracy tradeoff.

3 Methodology

The framework of E3SN is illustrated in Fig. 2. E3SN is a fully-convolutional Siamese network with three output branches for classification, localization and segmentation. The proposed supervised sampling strategy reduces the substantial memory requirements and the multilevel feature aggregation module generates fused features with rich representation for the mask refinement pathway to further improve the segmentation accuracy. The adopted tracking model SiamRPN is first introduced in the following part, and then the proposed framework is described.

3.1 Fully-convolutional Siamese network for visual tracking

SiamMask extends from the popular tracking system SiamRPN [Li *et al.*, 2018] which compares an exemplar frame z against a larger search frame x to obtain a dense response map. z and x are crops of different sizes centered on the target object of different frames in the same video. The two inputs are processed by the same CNN f_θ, producing two feature maps that are cross-correlated:

$$g_\theta(z, x) = f_\theta(z) \star f_\theta(x) \tag{1}$$

Each spatial element of the response map (left-hand side of Eq. 1) encodes the similarity between the exemplar z and the corresponding candidate window in x. The maximum value of the response map corresponds to the target location in the search area x. Moreover, the region proposal network [Ren *et al.*, 2015] branch further improve the location performance.

3.2 E3SN

The proposed model based on SiamRPN is constructed through extension with an extra branch and loss to build a fast end-to-end framework for VOS. E3SN adopts a two-step procedure for inference: locating the target region and then predicting a binary mask for the target.

Resnet50 [He *et al.*, 2016] is employed as the shared backbone network of E3SN. Meanwhile, the stride of *conv4* is reduced, and then dilated convolutions [Chen *et al.*, 2018a] are introduced. Thus, the backbone output has high resolution with a total stride of 8 pixels but a larger receptive field. An extra 1×1 convolution layer is used to reduce the channel dimension before the depth-wise cross-correlation operation. Binary masks for the selected candidate windows of the cross-correlation map are then predicted by a learnable refinement pathway h_ϕ as shown in Fig. 2. Let m_n denotes the predicted mask corresponding to the n-th candidate window as follows:,

$$m_n = h_\phi\left(g_\theta^n(z, x)\right) \tag{2}$$

The mask prediction is a function of the search frame x and the target object in z. Therefore, z can be used as a reference to guide the segmentation process. This approach means that different reference frames will produce different segmentation masks for x.

Mask refinement pathway. A common mask refinement strategy is proposed to predict an accurate mask for each selected sample (candidate window in correlation map). This strateg upsamples the features layer by layer with the addition of the high-resolution features from the backbone network to enrich the spatial information. Fig. 3 illustrates a detailed partial refinement process. Specifically, the unfold operation denotes that we slide a fixed size window on the large feature map to capture the regional feature similar to a convolution operation without any computation and then select the feature that corresponds to the sample location in correlation map.

Loss function. A multitask loss based on SiamRPN is derived to optimize the proposed model during the training process because E3SN is derived from the tracking model.

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{cls} + \lambda_2 \cdot \mathcal{L}_{reg} + \lambda_3 \cdot \mathcal{L}_{mask} \tag{3}$$

where $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = 32$ are set similar to that in SiamMask. The classification and bounding box losses are respectively the cross-entropy and the smooth L_1 losses similar to those defined in SiamRPN. For the mask branch, each sample is labeled with a binary label $y_n \in \{\pm 1\}$ and allocated with a pixel-wise ground truth mask m_n of size $w \times h$, in which $m_n^{i,j} \in \{\pm 1\}$ is the label of pixel (i, j). p_n is the predicted mask. The binary logistic regression loss is adopted as the loss function for mask prediction over all samples:

$$\mathcal{L}_{mask} = \sum_n \left(\frac{1 + y_n}{2wh} \sum_{ij} \log\left(1 + e^{-m_n^{ij} p_n^{ij}}\right) \right) \tag{4}$$

3.3 Supervised Sampling Strategy

SiamMask [Wang *et al.*, 2019] proposes the second version of SiamMask to generate high-quality object masks. This version replaces the simple mask head with an elaborate refinement module comprising upsampling layers and skip connections. The new refinement module improves performance. However, the refinement module has to be fine-tuned on the first simple model due to the training strategy that predicts a mask for each candidate window. This fine-tuning is conducted to avoid substantial memory cost, which results in an inconvenient two-stage framework with different hyperparameters.

The supervised sampling strategy is proposed to build an end-to-end framework. According to this strategy, a sample distribution matrix is established in the data loading process. The matrix is of the same size as the subsequent correlation maps, and all the elements are initialized zero. Then,

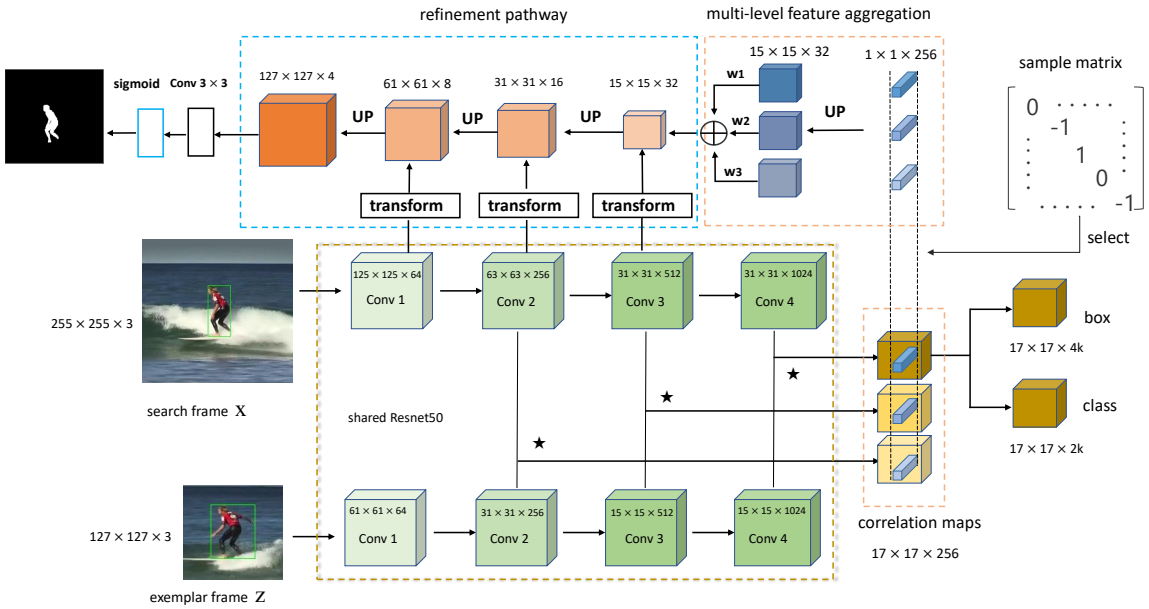

Figure 2: Architecture of the proposed E3SN. Multiple correlation maps are calculated on the basis of multilevel features extracted by the shared Siamese network. A fixed number of positive and negative samples are selected from the correlation maps, and these samples are aggregated for the final mask refinement module.

the value of each element in the matrix is changed in accordance with the IOU between the corresponding window in the search image and the target. The IOU is set positive if it is larger than a high threshold and negative if it becomes lower than the low threshold. Before the mask refinement pathway, candidate windows of correlation maps based on the sample distribution matrix are selected as positive and negative samples with a ratio of 3:1 rather than all candidate windows in SiamMask. Only the selected samples will be upsampled to predict the masks in the refinement pathway. Four samples are selected for each image.

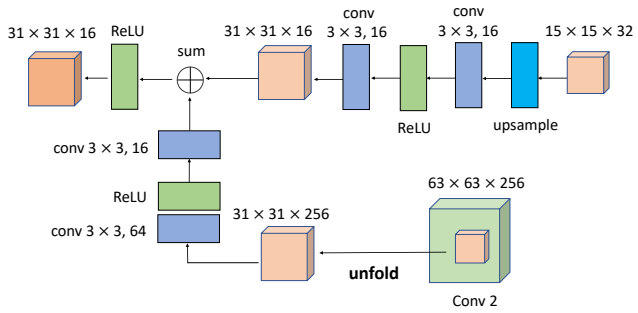

Figure 3: Detailed partial refinement process.

3.4 Multilevel Feature Aggregation Module

ResNet50, which produces hierarchical features in different layers, is employed as the backbone. Features of earlier layers contain additional low-level spatial information, such as color, shape, and texture, which are crucial for localization. However, these layers lack semantic information. Features of top layers have rich semantic information suitable to solve some tricky scenes, including motion blur and substantial deformation in the VOS task. Aggregating these hierarchical feature information for rich feature representation is possible to improve the segmentation quality.

In the multilevel feature aggregation module, hierarchical features extracted from the last three residual block of ResNet50 until $conv4$ are exploited. Each two extracted features from the same layer of x and z are cross-correlated, generating three depth-wise cross correlation maps referred as $g_2(z, x)$, $g_3(z, x)$, and $g_4(z, x)$. These correlation feature maps are upsampled to the same spatial resolution and then combined to one fused feature with learnable weights.

$$\mathcal{M}_{sum} = \sum_{i=2}^{4} \mathrm{w}_{i-1} \cdot \mathcal{M}_i \tag{5}$$

The weight parameters are optimized offline together with the network. The the refinement pathway then takes the fused feature as input to complete the mask prediction process.

4 Experiments

The implementation details and experimental setting are first described in this section. The proposed approach is then evaluated on two benchmarks compared with the state-of-the-art VOS methods. Furthermore, some qualitative segmentation results are presented to elicit intuitive feelings. Finally, the ablation study is performed to demonstrate the effectiveness of the proposed module and technique.

4.1 Implementation Details

Training. The exemplar is randomly chosen following the setting of SiamFC [Bertinetto *et al.*, 2016]. Image pairs are then searched from the same video, and cropping or padding is used to scale their size to 127×127 and 255×255. The ImageNet pre-trained model is loaded as the initial parameters of the backbone network, and SGD with an initial learning rate of 2×10^{-3} which logarithmically decreases to 2×10^{-4} in 20 epochs. Particularly, the learning rate of the mask branch is multiplied by 0.1. The two thresholds of sampling strategy are set to 0.6 and 0.3. The proposed model is trained on ImageNet-VID [Russakovsky *et al.*, 2015], COCO [Lin *et al.*, 2014], and YouTube-VOS [Xu *et al.*, 2018]. By contrast, only COCO and YouTube-VOS, which have mask labels, are useful for training the mask branch. NVIDIA TITAN RTX is the GPU used for training and evaluation.

Inference. E3SN can accept only bounding box of the target as initialization to determine the exemplar and operate online without any adaptation. During inference, the region of target is obtained in accordance with the maximum score in the classification branch. The mask of the selected region is then predicted with a sigmoid and threshold of 0.5. The predicted mask will eventually be mapped back to the corresponding position in the frame.

4.2 Experimental Setting

Datasets and evaluation metrics. The performance of E3SN is evaluated on DAVIS-2016 [Perazzi *et al.*, 2016] and DAVIS-2017 [Pont-Tuset *et al.*, 2017] validation sets for single- and multi-object segmentations, respectively. The DAVIS 2016 validation set comprises 20 videos, and each video sequence is annotated with a single pixel-wise object mask. The DAVIS 2017 validation set extends the DAVIS 2016 validation set to 30 videos with multiple object annotations.

The benchmark datasets only provide the pixel-level mask initialization in the first frame, whereas the proposed method merely requires a bounding box prior to the exemplar acquisition. Therefore, the strong mask prior to a bounding box initialization is weakened using the axis-aligned bounding rectangle strategy during the test phase.

For the evaluation metrics, the official performance criteria are adopted for both datasets: the Jaccard index ($\mathcal{J}$) to denote the mean intersection-over-union (mIoU) between the predicted and the ground-truth masks and the F-measure ($\mathcal{F}$) to represent contour accuracy, including $\mathcal{J}$ Mean, $\mathcal{J}$ Recall, $\mathcal{F}$ Mean, and $\mathcal{F}$ Recall, wherein a high value indicates good performance.

Comparison methods. The proposed model is compared with the following two kinds of state-of-the-art methods: online methods, such as OSVOS, OnAVOS, PReMVOS and MaskTrack; offline methods, including FAVOS [Cheng *et al.*, 2018], RGMP, FEELVOS, OSMN, VPN [Jampani *et al.*, 2017] and SiamMask, Offline methods have been the mainstream for fast inference speed with competitive accuracy.

4.3 Evaluation Results

Quantitative Result. (1) Evaluation on DAVIS2016. Table 1 shows the quantitative performance comparisons on DAVIS 2016. The table indicates the E3SN achieves a competitive performance of 73.0% at $\mathcal{J}_{Mean}$ and 69.3% at $\mathcal{F}_{Mean}$ with a faster speed of 45 FPS compared with most of listed methods. Especially for SiamMask, our proposed E3SN achieves 1.3% higher at $\mathcal{J}_{Mean}$ and 1.5% at $\mathcal{F}_{Mean}$ with slight speed sacrifice. Furthermore, E3SN is almost 60 times faster in running speed compared with that of the high-accuracy offline method FAVOS.

Method	FT	M	$\mathcal{J}_{Mean}$	$\mathcal{J}_{Recall}$	$\mathcal{F}_{Mean}$	$\mathcal{F}_{Recall}$	FPS
OnAVOS	✓	✓	**86.1**	**96.1**	84.9	89.7	**0.08**
PReMVOS	✓	✓	84.9	96.1	**88.6**	**94.7**	0.03
MaskTrack	✓	✓	79.7	93.1	75.4	87.1	0.08
FAVOS	×	✓	**82.4**	**96.5**	79.5	89.4	0.8
RGMP	×	✓	81.5	91.7	82.0	**90.8**	8.0
FEELVOS	×	✓	81.1	90.5	**82.2**	86.6	2
OSMN	×	✓	74.0	87.6	72.9	84.0	8.0
VPN	×	✓	70.2	82.3	65.5	69.0	1.6
SiamMask	×	×	71.7	86.8	67.8	79.8	**55**
E3SN	×	×	**73.0**	**88.3**	**69.3**	**80.9**	46

Table 1: Results on DAVIS2016 validation set. FT denotes the fine-tuning requirement of the method and M denotes mask initialization (✓) or a bounding box (×).

Method	FT	M	$\mathcal{J}_{Mean}$	$\mathcal{J}_{Recall}$	$\mathcal{F}_{Mean}$	$\mathcal{F}_{Recall}$	FPS
OnAVOS	✓	✓	**61.6**	**67.4**	69.1	75.4	**0.1**
OSVOS	✓	✓	56.6	63.8	63.9	73.8	0.1
FAVOS	×	✓	54.6	61.1	61.8	72.3	0.8
OSMN	×	✓	52.5	62.8	57.1	66.1	8.0
SiamMask	×	×	54.3	62.8	58.5	67.5	55
E3SN	×	×	**56.1**	**63.6**	**59.8**	**68.3**	46

Table 2: Results on DAVIS 2017 validation set. FT denotes the fine-tuning requirement of the method and M denotes mask initialization (✓) or a bounding box (×).

(2) Evaluation on DAVIS2017. The evaluation on DAVIS2017 is challenging for the multiple object scenarios. The evaluation results presented in Table 2 indicate that E3SN achieves the $\mathcal{J}_{Mean}$ of 56.1% and $\mathcal{F}_{Mean}$ of 59.8%, which are 1.8% and 1.3% higher than those of SiamMask, respectively. Moreover, the proposed method is about six times faster than the fast offline method OSMN.

E3SN has a good speed-accuracy trade-off compared with that of state-of-the-art VOS methods. An end-to-end trainable framework is constructed particularly for SiamMask to improve its performance.

Qualitative Result. Fig. 4 presents some qualitative visual results on DAVIS2016 and DAVIS2017. ES3N is found to have improved segmentation performance, especially on the edge, demonstrating the effectiveness of the multilevel feature aggregation module.

4.4 Ablation study

The contribution of the individual component of E3SN to its functionality is discussed in this subsection.

Figure 4: Qualitative visual results. The first column is the initialization of the first frame, and the other columns are the segmentation results of the subsequent frames. The first four rows are results on DAVIS2016, and the last two rows are results on DAVIS2017.

Supervised Sampling Strategy. Some indicators of the two-stage SiamMask (without sampling strategy) and the end-to-end E3SN with only sampling strategy (E3SN only S) are compared using the same setting to prove the effectiveness of the proposed strategy. The results presented in Table 3 show that the accuracy of E3SN (with only sampling strategy) is slightly worse than that of SiamMask but has less total training time without additional manual intervention. This finding suggests that the proposed sampling strategy helps alleviate the training gap of two-stage SiamMask and the end-to-end framework is training fast and convenient.

Method	$\mathcal{J}_{Mean}$	$\mathcal{F}_{Mean}$	Training time
SiamMask	71.7	67.8	22
E3SN only S	70.8	66.3	20

Table 3: Performance comparison of the two-stage SiamMask and the end-to-end E3SN on DAVIS2016. The training time is measured in hours.

Method	$\mathcal{J}_{Mean}$	$\mathcal{F}_{Mean}$
E3SN only S	70.8	66.3
E3SN	73.0	69.3

Table 4: Ablation study of the feature aggregation module on DAVIS2016.

Multilevel Feature Aggregation. E3SN and E3SN with only sampling strategy (E3SN only S) are evaluated on DAVIS2016 to assess the contribution of the multilevel feature aggregation module. As listed in Table 4, E3SN improves the performance by 2.2% at $\mathcal{J}_{Mean}$ and 3.0% at $\mathcal{F}_{Mean}$ through the proposed module compared with E3SN only S.

5 Conclusion

An E3SN for semi-supervised VOS is proposed in this study. Different from the two-stage SiamMask, which generates a mask for each candidate window of the correlation map, the proposed E3SN employs the supervised sampling strategy to select useful samples for the mask refinement training. This approach leads to a significant reduction in memory requirements and facilitates the establishment of an end-to-end framework. Furthermore, a multilevel feature aggregation module is developed to fuse the hierarchical features to enrich the feature representation capability. This module helps improve the final mask accuracy. Additionally, the proposed model only requires a bounding box initialization instead of a carefully crafted pixel-level mask in practice due to the advantage of the tracking model. The experimental results on benchmark datasets DAVIS2016 and DAVIS2017 demonstrate that E3SN achieves a satisfactory accuracy-speed trade-off.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 61771349 and 61822113, and by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170.

References

[Bertinetto *et al.*, 2016] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. In *ECCV Workshops*, volume 9914, pages 850–865, 2016.

[Caelles *et al.*, 2017] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, pages 5320–5329, 2017.

[Chen *et al.*, 2018a] 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.

[Chen *et al.*, 2018b] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *CVPR*, pages 1189–1198, 2018.

[Cheng *et al.*, 2018] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *CVPR*, pages 7415–7424, 2018.

[Han *et al.*, 2014] Junwei Han, Dingwen Zhang, Gong Cheng, Lei Guo, and Jinchang Ren. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3325–3337, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hu *et al.*, 2018] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G. Schwing. Videomatch: Matching based video object segmentation. In *ECCV*, pages 56–73, 2018.

[Jampani *et al.*, 2017] Varun Jampani, Raghudeep Gadde, and Peter V. Gehler. Video propagation networks. In *CVPR*, pages 3154–3164, 2017.

[Li *et al.*, 2018] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018.

[Li *et al.*, 2019] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, pages 4282–4291, 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, and Serge J. Belongie et al. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.

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

[Maninis *et al.*, 2019] K.-. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. Video object segmentation without temporal information. *TPAMI*, 41(6):1515–1530, 2019.

[Oh *et al.*, 2018] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, pages 7376–7385, 2018.

[Perazzi *et al.*, 2016] 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*, pages 724–732, 2016.

[Perazzi *et al.*, 2017] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, pages 3491–3500, 2017.

[Pont-Tuset *et al.*, 2017] Jordi Pont-Tuset, Federico Perazzi, and Sergi Caelles et al. The 2017 DAVIS challenge on video object segmentation. abs/1704.00675, 2017.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, and Hao Su et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[Song *et al.*, 2020] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, page 107173, 2020.

[Voigtlaender *et al.*, 2019] 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.

[Wang *et al.*, 2017] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Selective video object cutout. *TIP*, 26(12):5645–5655, 2017.

[Wang *et al.*, 2019] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, pages 1328–1338, 2019.

[Xu *et al.*, 2018] Ning Xu, Linjie Yang, and Yuchen Fan et al. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 603–619, 2018.

[Yang *et al.*, 2018] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K. Katsaggelos. Efficient video object segmentation via network modulation. In *CVPR*, pages 6499–6507, 2018.

[Zhang *et al.*, 2016] Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.