

Integrating Long-Short Term Network for Efficient Video Object Segmentation

Jingjing Wang¹
18120412@bjtu.edu.cn

Zhu Teng¹
zteng@bjtu.edu.cn

Baopeng Zhang ✉¹
bpzhang@bjtu.edu.cn

Jianping Fan²
fanj1@lenovo.com

¹ School of Computer and Information
Technology
Beijing Jiaotong University
Beijing, China

² AI Lab
Lenovo Research
Beijing, China

Abstract

Real-world application of video object segmentation (VOS) is a very challenging problem, especially for multiple video object segmentation. The deep-learning-based approaches have recently dominated VOS by fine-tuning the networks at the first frame to seize the object dynamics, but they may result in impractical frame-rates and risk of over-fitting. To overcome this limitation, we develop an efficient and fully end-to-end model to achieve fast and accurate VOS, named Long-Short Term Network (LSTNet). It contains a long term network to encode absolute object variations and a short term network to capture relative object dynamics. The segmentation results of video objects can be directly acquired by an attentional gate operation based on these two networks. Our proposed model runs at a very high speed and can conveniently tackle multi-object segmentation without post-processing. Extensive experiments on widely used benchmarks including YouTube-VOS and DAVIS 2017 have demonstrated that our proposed model can achieve a competitive accuracy and speed in comparison to a number of state-of-the-art methods.

1 Introduction

Video object segmentation - segmenting target objects within the entire video sequence - is an important task in computer vision with various applications including autonomous driving, video editing, robotics *et al.* In this work, we focus on the task of Video Object Segmentation (VOS), where the ground-truth masks of target objects on the first frame are provided. In order to segment the target objects in the subsequent video frames, utilizing the features of target objects in the first frame can memorize the appearances of target objects and distinguish them from the background. Consequently, a number of methods rely on fine-tuning the CNNs [11, 13, 15, 20, 22, 30, 33] online to exploit the first-frame segmentation and learn the appearances of the target objects. Unfortunately, such VOS methods usually require a very long executing time, which is not suitable for most practical applications.

To alleviate the dependence on the online fine-tuning process, some template [4, 12, 34, 35, 38, 42, 46] and propagation-based methods [17, 21, 32, 35, 36, 38, 41] have emerged. Since the ground-truth mask of the first frame is given, the first frame can be used as a template to extract the object features. Subsequent frames can be judged by matching with the first frame [4, 12, 34, 35, 38, 42], and then set each pixel by a score. Finally, a score threshold is employed to determine whether the pixel belongs to the target objects or not. Although CNNs can extract target features to match the pixels, constant appearance change cannot be captured due to the limited adaptation. The intuition behind the propagation-based methods is that the changes between adjacent two frames are usually very tiny. Thus, in the propagation-based methods, the result of the previous frame is employed as an estimated or guided mask of the next frame [13, 38]. But when occlusions or fast motions occur, drifting problems might arise and errors from previous frames are easily accumulated thereafter.

Both the segmentation information on the first frame and the information from historical frames are important to conduct video object segmentation. How to achieve better utilization of this information is the key to reach a better speed-accuracy trade-off for the VOS task. In this paper, we develop a fast network to achieve a fine-grained VOS performance, named Long-Short Term Network (LSTNet). Our framework is formed by three branches, the *Long-Term Network (LTN)*, the *Short-Term Network (STN)* and the *Attentional Gate Network (AGN)* as presented in Fig. 1.

The LTN describes the absolute change of the target objects by capturing the target relationship of targets in the current frame with respect to those in the first frame. As the target objects in the first frame are segmented accurately, the absolute deformation of the target objects can be achieved. The STN leverages the temporal correlations between the adjacent frames to obtain the relative variations of target objects. The AGN effectively suppresses the unrelated features and enhances the target features for high-resolution segmentation. It is fed by the long and short term maps, and refines the features to estimate an accurate segmentation. In summary, the contributions of this work are three-fold:

- A fast end-to-end deep network is designed for video object segmentation. Both the absolute object changes and the relative object variations are captured to facilitate the model to segment multiple video objects accurately.
- A novel Long-Short Term Network (LSTNet) is designed to encode the evolutionary process for the objects. The long-term network exploits the object relationship between the current frame and the first frame, and the short-term network explores the immediate object variation. An attentional gate network is finally constructed to predict the object masks in the current frame.
- Extensive experiments are executed on two widely used segmentation benchmarks including YouTube-VOS and DAVIS 2017, and our results have demonstrated that our approach can exhibit high efficiency and accuracy against the state-of-the-arts.

2 Related Work

Fine-tuning based on the first frame methods: Many state-of-the-art approaches train a segmentation network offline and execute online learning by fine-tuning the segmentation networks [11, 13, 19, 21, 22, 30, 33] at the test time in order to memorize the appearance of the target object on the given object mask. For example, OSVOS [11] handles the online learning without temporal modeling but uses the first annotated frame to fine-tune a trained

network and segments the objects in other frames individually. LucidTracker [19] employs data augmentations on the first frame so that more target features can be learned during the online fine-tuning stage. OnAVOS [3] adjusts the fixed parameters online and utilizes online updates to adapt to changes in appearance. Other approaches further incorporate optical flow [15, 16] as an additional cue. For instance, PReMVOS [24, 25] integrates techniques from instance segmentation, optical flow, refinement, and re-identification together with extensive fine-tuning, and achieves a satisfactory performance. Although these algorithms achieve an impressive segmentation result, there is still a big gap for the processing speed to satisfy the requirements for practical applications due to the heavy computation burden of fine-tuning on the first frame.

Propagation-based methods: To leverage the temporal consistency between two adjacent frames and segment all the frames independently, many propagation-based methods [7, 21, 32, 35, 36, 38, 40] often transfer the mask of the previous frame to the current frame to adapt the changes of object appearances. A typical method MaskTrack [30] uses the output of the last frame as a guide for the interesting region of the next frame. Optical flow [15, 16] is also used to guide the propagation process in many methods [8, 13, 17]. Others utilize GANs [10] to capture propagation coherence effectively. For instance, GANs [2] is designed to learn spatio-temporal object models on a limited space-time window. And an adversarial fashion like GANs [40] is utilized to capture dynamic appearance and motion cues of video sequences to guide object segmentation. Recently, many approaches [32, 40] also adopt RNN [45] to focus on propagation information in videos. RVOS [32] proposes a recurrent network for multiple object video object segmentation to learn temporality. S2S [40] relies on Conv-LSTM [39] to build a memory module for recursively long-term prediction. However, these methods are vulnerable to temporal discontinuities like occlusions and rapid motion, and can suffer from drifting once the propagation becomes unreliable.

Template-based methods: The template-based methods [8, 14, 32, 35, 38, 42, 46] address the VOS task as a pixel-level object matching problem with the annotation of the object mask on the first frame. For example, VideoMatch [14] performs a soft segmentation upon the averaged similarity score maps of matching features to generate smooth predictions. RGMP [38] proposes a deep Siamese [9] encoder-decoder network that is designed to take advantage of the template. Other approaches propose a way of ranking to match the similarity with the template. Take RANet [37] as an example. It adopts a novel ranking attention module, which automatically ranks and selects these maps for fine-grained VOS performance. Furthermore, a pixel-wise embedding metric learning based approach predicts each pixel by nearest neighbor matching in pixel space to the template frame. FEELVOS [32] uses a semantic pixel-wise embedding together with a global and a local matching mechanism for more stable pixel-level matching. DDEAL [42] proposes a directional deep embedding and appearance learning method for fine-tuning-free fast VOS. These methods obtain good performance. However, due to the lack of temporal information, they still suffer from the mismatching problem. In this work, we explore both the strategies of matching for pixel-level object segmentation and temporal propagation, to handle the mismatching and drifting problem.

3 Long-Short Term Network

We design a model to learn the absolute variations and relative dynamics of targets with respect to the first frame and previous frames, respectively. An attentional gate network is composed to decode this information into the segmentation mask. Our elaborately designed net-

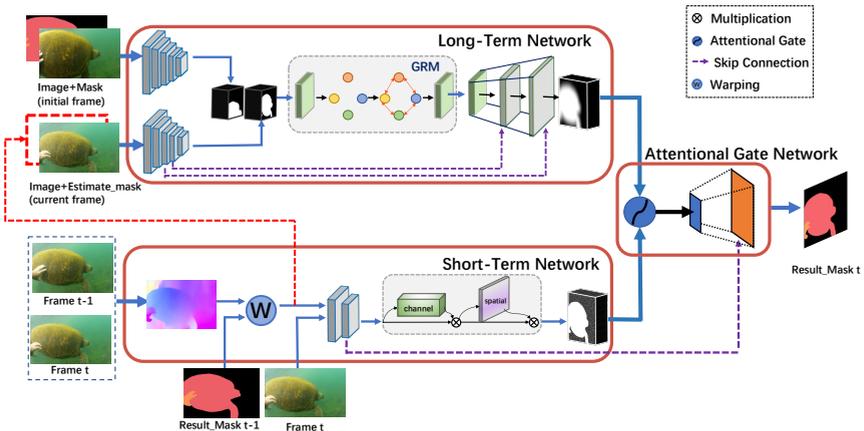


Figure 1: The architecture of LSTNet. In order to segment the target object from the current frame, the extracted long-term features and the short-term features are fed into the attentional gate model to produce the final segmentation.

work tackles multi-object segmentation in one forward pass without post-processing, which avoids repeating multiple times for multi-object. In the following, the pipeline of our proposed framework is first presented and then we describe three components and their network architectures in detail.

3.1 Pipeline

The pipeline of the proposed method is illustrated in Figure 1 where three seamless branches compose the segmentation model including a long-term network, a short-term network, and an attentional gate network. In the long-term network, the relationship information between the target of the current frame and the annotated first frame is extracted to get the absolute variations of the targets, which reasons the global information of specific positions and deformations of these targets. In the short-term network, the relative changes of targets against the previous frame are exploited, which facilitates to estimate the local and subtle information of the target. The attentional gate network focuses on the overall structures and refines the feature maps to decode the final result. With the above efforts, a single forward pass can adapt the proposed segmentation model to the appearances of specific objects fast and accurately by exploiting long-term and short-term information without online fine-tuning.

To give a more precise definition, assume there are K objects in total, let G_1 be the provided ground-truth mask of the first frame, and denote the coarse mask of frame t as \hat{M}_t . We introduce $\mathcal{I} = \{I_1, \dots, I_t, \dots\}$ to denote an input video with N frames, and our LSTNet is forwarded to generate the segmentation results that are represented by $\mathcal{M} = \{M_0, \dots, M_t, \dots\}$.

3.2 Long-Term Network

The task of long-term Network is to exploit long-term information of objects in a deep feature space. As the segmented objects are indicated on the first frame, it possesses the most accurate and richest information throughout the video sequence. In this part, the background and foreground features on the first frame are extracted, and features from the first frame and the current frame are concatenated on the channel dimension to extract the relation fea-

tures of intended objects. The absolute variations of the objects are perceived to predict the deformation of these objects.

A Siamese structure is employed to capture the differences of objects on the current frame and the first frame. Specifically, Resnet-50 [10] is employed as the shared feature extractor due to its well-balanced capacity and efficiency. In order to better distinguish the foreground from the background, the input of our Siamese structure is expanded from RGB to RGB+mask channel (4 channels). The extra mask channel is targeted to provide an estimate of the visible area, approximate location and shape of the objects in the current frame. In the first frame, the input image I_1 and the provided mask G_1 are fed into one branch of our Siamese structure (as shown in Figure 1). The current frame I_t and an estimated coarse mask \hat{M}_t which contains multiple targets that are generated by the short term network (see more details in Sec. 3.3) are delivered into the other branch of our Siamese structure. The Siamese structure outputs a hierarchical of features $F_t = \{F_{t,1}, F_{t,2}, \dots, F_{t,5}\}$ with different spatial information, where $F_{t,i}$ suggests the feature obtained from the i^{th} extractor level on the t^{th} frame of the video sequence.

After extracting the features from the first frame and the current frame, we concatenate $F_{1,5}$ and $F_{t,5}$, and feed them into the global relation module (GRM). GRM globally aggregates the concatenated feature over the coordinate space and then projects it into the interaction space where the relation is inferred. The feature is back-projected into the original coordinate space to calculate similar regions and judge the relation among them (see more details in Sec. 4.1). This module not only allows us to extract global features but also to collect the relationship information between the current and the first frame. It examines similar areas and determines the relationship between different areas. In order to match with the short term network, the feature maps are further upsampled. We additionally design skip connections with the upsampling map that are built by residual blocks on the hierarchical features $F_{t,3}$ and $F_{t,4}$ of the current frame to obtain more accurate target information.

3.3 Short-Term Network

The short-term Network is to capture the relative dynamics of objects and is guided by a coarse mask \hat{M}_t which contains multiple targets estimated based on optical flow and the segmentation result of the immediately previous frame M_{t-1} . Motion estimations based on optical flow reveal the pixel correspondence between frames and enable the propagation of foreground/background labels from one frame to the next. Since the guidance map provides the coarse information on the location and shape of target objects, the network concentrates full attention on the estimation of dominant objects in the given region with the provided coarse shape. In the case of multiple objects, the overall mask M_{t-1} can be separated into a series of masks of objects $\{M_{t-1,1}, M_{t-1,2}, \dots, M_{t-1,k}\}$.

Given a pair of two consecutive frames (I_{t-1}, I_t) , optical flow is acquired through a light-weight network, which incorporates building blocks from LiteFlowNet [15]. Both this optical flow and the object masks $\{M_{t-1,1}, M_{t-1,2}, \dots, M_{t-1,k}\}$ from the previous segmentation result (as described in Figure 1) are conveyed to the warping module to predict the masks $\{\hat{M}_{t,1}, \hat{M}_{t,2}, \dots, \hat{M}_{t,k}\}$ on the current frame according to flow map with the bi-linear operation. The estimated coarse mask \hat{M}_t is estimated by Eq.(1) and Eq.(2), where O is the optical flow operation and W denotes the warp operation.

$$\hat{M}_{t,i} = W(O(I_{t-1}, I_t), M_{t-1,i}) \quad (1)$$

$$\hat{M}_t = \max_{1 \leq i \leq k} \hat{M}_{t,i} \quad (2)$$

The predicted mask reflects the specific location and detailed information of the target, and is not only an input for the next module of the short-term network, but also provides as the estimated mask of the current frame in the long-term network.

We expand the estimated mask \hat{M}_t into masks of multiple objects $\{\hat{M}_{t,1}, \hat{M}_{t,2}, \dots, \hat{M}_{t,k}\}$, and feed them together with the current frame I_t into the layers that are consisted by *Conv1* and *Conv2* of Resnet-50. The output feature is denoted by $f_i = \{f_{i,1}, f_{i,2}, \dots, f_{i,k}\}$, where $f_{i,k}$ indicates the visual-feature of the k^{th} object. We further construct a light-weight channel & spatial encoding module to enhance the information of objects, and the results of this short term network are represented by $A_t = \{A_{t,1}, A_{t,2}, \dots, A_{t,k}\}$. The short-term network described above captures the relative dynamics of objects, which provides us the specific local and detailed information of objects.

3.4 Attentional Gate Network

In order to get the final segmentation result, the long-term and the short-term features should be further decoded, refined and upsampled to the shape aligned with the input. To facilitate this, an attentional gate module is established where the similarity and difference of long-term and short-term feature maps are perceived. It automatically captures the semantic information, globally organizes the location and relationship on a series of spatial levels, suppresses the background regions on the current frame, and highlights the prominent features of the foreground. The network starts with the gate operation [28] on the long-term features L_t with 3 dimensions and the short-term features $A_t = \{A_{t,1}, A_{t,2}, \dots, A_{t,k}\}$ to generate the gate features $U_t = \{U_{t,1}, U_{t,2}, \dots, U_{t,k}\}$. The gate operation is formulated in Eq. (3)-Eq. (5), where $*$ is a 1×1 convolution, b_{A_i} and b_{ψ_i} are bias terms, and \otimes denotes for an element-wise multiplication.

$$X_{t,i} = Relu(W_L^T * L_t + W_{A_i}^T * A_{t,i} + b_{A_i}) \quad (3)$$

$$G_{t,i} = Sigmoid(W_{\psi_i}^T * X_{t,i} + b_{\psi_i}) \quad (4)$$

$$U_{t,i} = A_{t,i} \otimes G_{t,i} \quad (5)$$

A refined prediction module is further constructed by skip-connections to reconstruct accurate masks. These masks are upsampled to the same size as those in the current frame, which are denoted by $R_t = \{R_{t,1}, R_{t,2}, \dots, R_{t,k}\}$.

So far, the segmentation result for multiple objects is obtained. Based on this, the background pixels can be extracted as

$$R_{t,0} = 1 - \max_{1 \leq i \leq k} R_{t,i} \quad (6)$$

In order to obtain a more accurate object boundary, we normalize the segmentation probability map of K objects to predict the result of each pixel. The segmented results of K object $R_t = \{R_{t,1}, R_{t,2}, \dots, R_{t,K}\}$ can be further refined and normalized by Eq. (7) where $\hat{R}_{t,c}$ is the refined result.

$$\hat{R}_{t,c} = \frac{R_{t,c}}{1 - R_{t,c}} \quad (7)$$

$$\hat{R}_{t,c} = \frac{R_{t,c}}{\sum_{i=0}^k \frac{R_{t,i}}{1 - R_{t,i}}}$$

The pixel in $\hat{R}_{t,c}$ belongs to the target whose probability is the largest among K probabilities, and if the maximum value is less than 0.5, the pixel is set as background. The above probabilistic normalization strategy also enables us to directly derive the object segmentation result $S_t^c \in \{0, 1\}$.

YouTube-VOS one-shot					
Method	\mathcal{G} overall(%)	\mathcal{J} seen(%)	\mathcal{J} unseen(%)	\mathcal{F} seen(%)	\mathcal{F} unseen(%)
LSTNet	71.8	70.9	66.8	74.9	74.8
LTN_free	68.7	69.5	62.6	73.1	69.8
STN_free	67.5	68.4	60.8	72.4	68.5
AGN_free	69.7	70.0	63.6	73.8	71.4

Table 1: Ablation study on three key components on the YouTube-VOS dataset. Models are trained on the training set and evaluated on the validation set.

4 Experiments

We evaluate our approach against numerous state-of-art VOS methods on two public datasets, namely the Youtube-VOS [42] and DAVIS 2017 [50]. In the following, we first introduce the implementation details of our model and then conduct ablative studies on the three key components of our LSTNet. Afterward, the performance of the proposed method is evaluated against state-of-the-art approaches on the two benchmarks.

Metrics: For a more comprehensive evaluation, we employ three metrics including the mean region similarity (\mathcal{J} Mean), mean contour accuracy (\mathcal{F} Mean) and their average (\mathcal{G} Mean) [49]. Note that \mathcal{F} and \mathcal{J} are separately calculated for seen and unseen classes in YouTube-VOS [42]. Besides, the run time is also measured for efficiency evaluation.

4.1 Implementation Details

In the long-term network, we rebuild the Resnet-50 [10], which is pre-trained on the ImageNet [8], by removing the last global pooling and fully-connected layers and adding attention [42] module in the block *Conv3*, as the feature extractor. In the global relation module, we firstly construct a block including a conv layer with a 3×3 kernel and stride 1, batch normalization layer and ReLU layer. The channels of feature maps are reduced from 4096 to 1024. Then, five convolutions [4] are employed to extract global relationships, two of which are used for dimension reduction and expansion, one of which is employed to generate bi-projection coordinates and latent interaction spaces, and the other two convolutions are utilized for global reasoning based on the graph in the interaction space. The relationship map outputted by this module is with 1024 channels. Finally, a 3×3 convolution with stride 1, followed by batch normalization and ReLU layer, is applied to aggregate global relationship information. This convolution reduces the channels of feature maps from 1024 to 256. In the short-term network, the channel module utilizes both max-pooling outputs and average-pooling outputs with a shared network and the spatial module employs similar two outputs that are pooled along the channel axis and forwards them to a convolution layer [42].

The loss function we use in the proposed LSTNet is IOU (Intersection over Union) between the predicted segmentation and final segmentation [42]. The loss is summed and minimized by Adam with a learning rate of $1e^{-4}$, and LSTNet is trained for 50 epochs using Youtube-VOS [42]. Our method is implemented using Python with PyTorch. All experiments are conducted on one NVIDIA Titan X Pascal GPU card.

4.2 Ablation Study

We perform an ablative analysis on the key components of LSTNet including LTN, STN and AGN on the YouTube-VOS official validation set [42]. The results are reported in Table 1.

YouTube-VOS one-shot							
Method	FT	\mathcal{G} overall(%)	\mathcal{J} seen(%)	\mathcal{J} unseen(%)	\mathcal{F} seen(%)	\mathcal{F} unseen(%)	FPS
S2S [14]	✓	64.4	71.0	55.5	70.0	61.2	0.11
MSK [81]	✓	53.1	59.9	45.0	59.5	47.9	0.08
OSVOS [10]	✓	58.8	59.8	54.2	60.5	60.7	0.10
OnAVOS [63]	✓	55.2	60.1	46.6	62.7	51.4	0.08
PRReMVOS [42, 43]	✓	66.9	71.4	75.9	56.5	63.7	0.03
STM [24]		79.4	79.7	72.8	84.2	80.9	<6.25*
STM-synth [24]		68.2	-	-	-	-	<6.25*
AGSS [25]		71.3	71.3	65.5	75.2	73.1	12.5
OSMN [15]		51.2	60.0	40.6	60.1	44.0	4.16
DMM-Net [16]		51.7	58.3	41.6	60.7	46.3	12
RGMP [33]		53.8	59.5	45.2	-	-	-
A-GAME [13]		66.0	66.9	61.2	-	-	-
DDEAL [22]		70.5	72.5	75.8	63.4	70.4	-
S2S (w/o OL) [14]		51.7	66.7	48.2	65.5	50.3	6.25
CapsuleVOS [1]		62.3	67.3	68.1	53.7	59.9	13.5
LSTNet		71.8	70.9	66.8	74.9	74.8	12.8

Table 2: Quantitative results on the Youtube-VOS validation set. Models marked with FT indicates that fine-tuning on the first frame is required. Bold font indicates the best result. Note that the performance data are directly copied from the corresponded published papers. The asterisk in FPS (*) indicates a speed for a single object and others are for multi-object.

Effectiveness of LTN: To demonstrate the effectiveness of LTN, we construct a model (named LTN_free) based on the LSTNet but only remove the global relation module in the long-term network. The extracted features of the current frame and the first frame are directly concatenated, and then retain the conv layer before and after the global relation module to match the channels. The segmentation results of LTN_free is reported in the second row of Table 1. Compared with the overall LSTNet, the LTN_free model drops the overall performance by a percentage of 3.1 evaluated by the average metric \mathcal{G} , which demonstrates the effectiveness of the long-term network.

Effectiveness of STN: To validate the effectiveness of STN, a model named STN_free is composed by disabling the STN from LSTNet and directly conveying the estimated mask of multiple objects $\{\hat{M}_{t,1}, \hat{M}_{t,2}, \dots, \hat{M}_{t,k}\}$ and the long-term features L_t as the input of AGN. As shown on the third row in Table 1, a decline of 4.3% is observed in average performance by removing the STN, which demonstrates the influence of the proposed short-term network for video object segmentation.

Effectiveness of AGN: In order to investigate the impacts of AGN, we establish a model AGN_free by removing the gate network from LSTNet. The LTN and STN information are directly operated by element-wise multiplication. The comparison results are presented in the fourth row of Table 1, where it can be observed that the overall score is decreased by 2.1% due to the elimination of AGN. This demonstrates that the AGN is able to learn crucial information for the video object segmentation.

4.3 State-of-the-Art Comparison

We compare the performance of our method with state-of-the-art approaches on Youtube-VOS [14] and the DAVIS 2017 dataset [81].

Youtube-VOS: The validation set of YouTube-VOS [14] comprises 474 videos labeled with one or multiple objects. We submitted our results to the official platform for a fair evaluation. The comparison results are reported in Table 2 where all models are trained

Method	FT	DAVIS _{17-val}			DAVIS _{17-testdev}			FPS
		\mathcal{G} (%)	\mathcal{J} (%)	\mathcal{F} (%)	\mathcal{G} (%)	\mathcal{J} (%)	\mathcal{F} (%)	
OSVOS [10]	✓	56.6	60.3	52.9	50.9	47.0	54.8	0.05
OSVOS-S [27]	✓	68.0	64.7	71.3	57.5	52.9	62.1	-
OnAVOS [33]	✓	63.6	61.0	66.1	56.5	53.4	59.6	0.04
PReMVOS [24, 25]	✓	77.8	73.9	81.7	71.6	67.5	75.7	0.03
OSMN [43]		54.8	52.5	57.1	41.3	37.7	44.9	3.57
VideoMatch [42]		62.4	56.5	68.3	-	-	-	2.86
FAVOS [6]		58.2	54.6	61.8	43.6	42.9	44.3	0.83
RANet [56]		65.7	63.2	68.2	55.3	53.4	57.2	-
RGMP [52]		66.7	64.8	68.6	52.8	51.3	54.3	3.57
STCNN [40]		61.7	58.7	64.6	-	-	-	0.25
LSTNet		67.5	64.6	70.4	56.2	53.4	58.9	8.33

Table 3: Quantitative results on the DAVIS 2017 validation set and test-dev set [61]. FPS is evaluated on the val set. Models marked with FT indicates that fine-tuning on the first frame is required. Bold font indicates the best result.

on YouTube-VOS [47]. Note that DMM-Net [46], A-GAME [18], STM [27], AGSS [23], CapsuleVOS [9] and DDEAL [44] are the latest segmentation models. The proposed method obtains a percentage of 71.8 measured by \mathcal{G} at a speed of 0.078 seconds (12.8FPS, Frames Per Second), which obtains the second best in the overall performance. STM [27] achieves the best performance at a \mathcal{G} -score of 79.4 by training the model on a simulation dataset and YouTube-VOS dataset (VOS) [47]. But it claims a \mathcal{G} -score of 68.2 if trained only on VOS. In contrast, our method achieves a \mathcal{G} -score of 71.8 training only on VOS, which boosts the \mathcal{G} -score by 3.6 on even ground. Furthermore, our model is several times faster than STM as it reports a speed of 6.25FPS for a single object and ours is 12.8FPS for multi-object. In other words, our method ranks number one considering both the overall performance and the efficiency, and significantly outperforms state-of-the-arts without invoking any online fine-tuning. The third best performer is AGSS [23], which reports a \mathcal{G} -score of 71.3 and is slightly lower than ours. In particular, for the unseen evaluation, ours leads AGSS by 1.5 (70.8 VS 69.3). The performer with the fastest running speed is CapsuleVOS [9] at about 13.5 FPS, which is only slightly faster than ours (13.5 FPS VS 12.8 FPS), but the average performance \mathcal{G} is dropped by 9.5% compared with the proposed method. Furthermore, in contrast to the state-of-the-art online methods such as PReMVOS [24, 25] and S2S [40], the proposed method still achieves a better performance.

DAVIS 2017: We also conduct the comparative experiments on DAVIS 2017 dataset [30] to verify the effectiveness of our method on multi-object segmentation. Table 3 shows the quantitative comparisons against a number of latest state-of-arts by two metrics on both the validation set and test-dev set. The methods can be divided into two categories according to the requirement of fine-tuning. It can be seen that the proposed method achieves the best performance and the fastest speed compared to the methods that do not require fine-tuning. When compared with the methods with online fine-tuning, our method occupies the third position and the best performer is PReMVOS [24, 25]. However, these online fine-tuning methods such as PReMVOS only runs at a speed of 0.03 FPS or so, which is 417x slower than ours. In other words, the proposed method achieves the best trade-off between the accuracy and the running speed.

Finally, parts of our segmentation results are visualized in Figure 2 where different time steps of each video sequence are uniformly sampled, and some qualitative comparisons

against several state-of-the-art methods are displayed in Figure 3. It is clear that our model consistently achieves a very good segmentation quality throughout the video sequence even under challenging situations, such as deformations, fast motions, large scale variations, and cluttered backgrounds.

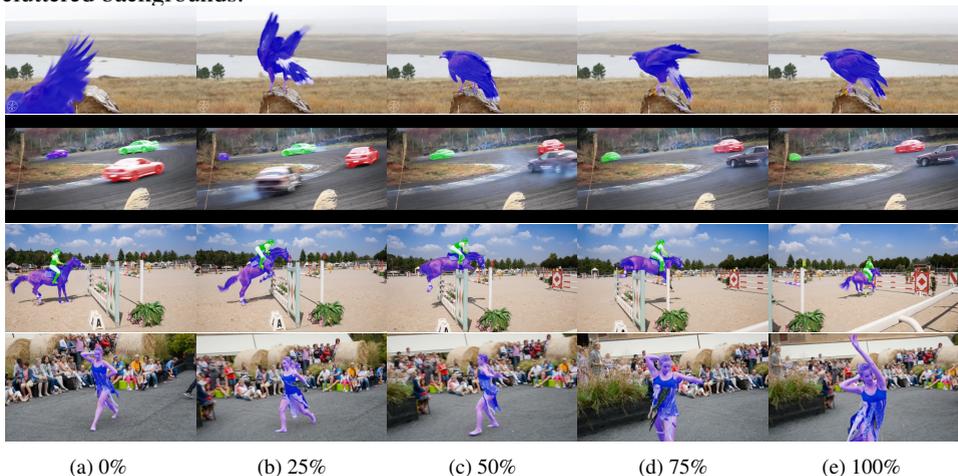


Figure 2: Visualization of our results on YouTube-VOS and DAVIS 2017

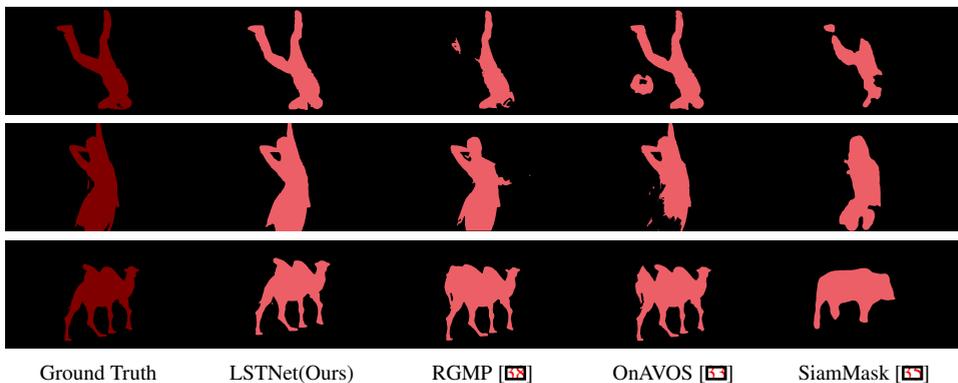


Figure 3: Qualitative comparison with three state-of-the-art approaches

5 Conclusions

A fast and end-to-end long-short term network is developed in this paper to conduct video object segmentation. The long-term network encodes the absolute object variations by exploiting the object relationship between the current and first frames, and the short-term network targets at relative object dynamics by taking advantage of optical flow and a channel & spatial network. This information is further decoded and refined by the attentional gate network to obtain the final segmentation results of multiple video objects. Extensive experiments are executed, which have demonstrated that our proposed method can achieve a superior performance and fast running speed (about 12.8 FPS) on both the YouTube-VOS dataset and DAVIS 2017 against numerous state-of-the-art methods.

Acknowledgments: This work was supported by the Natural Science Foundation of China (61972027). The Titan X Pascal used for this research was donated by the NVIDIA Corporation. The corresponding author is Baopeng Zhang.

References

- [1] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017.
- [2] Sergi Caelles, Albert Pumarola, Francesc Moreno-Noguer, Alberto Sanfeliu, and Luc Van Gool. Fast video object segmentation with spatio-temporal gans. *arXiv preprint arXiv:1903.12161*, 2019.
- [3] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1189–1198, 2018.
- [4] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.
- [5] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, pages 686–695, 2017.
- [6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7415–7424, 2018.
- [7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [9] Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Capsulevos: Semi-supervised video object segmentation using capsule routing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8480–8489, 2019.
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

- [13] Ping Hu, Gang Wang, Xiangfei Kong, Jason Kuen, and Yap-Peng Tan. Motion-guided cascaded refinement network for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1400–1409, 2018.
- [14] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 54–70, 2018.
- [15] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018.
- [16] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.
- [17] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2126. IEEE, 2017.
- [18] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8953–8962, 2019.
- [19] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for object tracking. In *The DAVIS Challenge on Video Object Segmentation*, 2017.
- [20] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for video object segmentation. *International Journal of Computer Vision*, 127(9):1175–1197, 2019.
- [21] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 90–105, 2018.
- [22] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 577–585, 2018.
- [23] Huaijia Lin, Xiaojuan Qi, and Jiaya Jia. Agss-vos: Attention guided single-shot video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3949–3957, 2019.
- [24] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, pages 565–580. Springer, 2018.

- [25] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for the davis challenge on video object segmentation 2018. In *The 2018 DAVIS Challenge on Video Object Segmentation-CVPR Workshops*, volume 1, page 6, 2018.
- [26] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1515–1530, 2018.
- [27] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9226–9235, 2019.
- [28] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [29] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.
- [30] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2663–2672, 2017.
- [31] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.
- [32] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5277–5286, 2019.
- [33] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017.
- [34] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019.
- [35] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1328–1338, 2019.
- [36] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3978–3987, 2019.

- [37] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.
- [38] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7376–7385, 2018.
- [39] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.
- [40] Kai Xu, Longyin Wen, Guorong Li, Liefeng Bo, and Qingming Huang. Spatiotemporal cnn for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1379–1388, 2019.
- [41] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.
- [42] 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.
- [43] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6499–6507, 2018.
- [44] Yingjie Yin, De Xu, Xingang Wang, and Lei Zhang. Directional deep embedding and appearance learning for fast video object segmentation. *arXiv preprint arXiv:2002.06736*, 2020.
- [45] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.
- [46] Xiaohui Zeng, Renjie Liao, Li Gu, Yuwen Xiong, Sanja Fidler, and Raquel Urtasun. Dmm-net: Differentiable mask-matching network for video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3929–3938, 2019.