Distance-Guided Mask Propagation Model for Efficient Video Object Segmentation

Jiajia Liu
School of Electronic and Information Engineering
South China University of Technology
Guangzhou, China
eeliujiajia@mail.scut.edu.cn

Hongning Dai
Faculty of Information Technology
Macau University of Science and Technology
Macau, China
hndai@ieee.org

Bo Li*
School of Electronic and Information Engineering
South China University of Technology
Guangzhou, China
leebo@scut.edu.cn

Gaozhong Tang
School of Electronic and Information Engineering
South China University of Technology
Guangzhou, China
eegztang@mail.scut.edu.cn

***Abstract*—Video object segmentation (VOS) is a significant yet challenging task in computer vision. In VOS, two challenging problems, including occlusions and distractions, are needed to be handled especially in multi-object videos. However, most existing methods have difficulty in efficiently tackling these two factors. To this end, a new semi-supervised VOS model, called *Distance-Guided Mask Propagation Model* (DGMPM), is proposed in this paper. Specifically, a novel *embedding distance module*, which is utilized to generate a soft cue for handling occlusions, is implemented by calculating distance difference between target features and the centers of foreground/background features. This non-parametric module that is based on global contrast between the target and reference features to detect object regions even if occlusions still exist, is less sensitive to the feature scale. The prior knowledge of the previous frame is applied as spatial guidance in the decoder to reduce the effect of distractions. In addition, spatial attention blocks are designed to strengthen the network to focus on the target object and rectify the prediction results. Extensive experiments demonstrate that the proposed DGMPM achieves competitive performance on accuracy and runtime in comparison with state-of-the-art methods.**

***Index Terms*—Video Object Segmentation, Spatial Guidance, Attention Mechanism, Fully Convolutional Neural Networks**

I. Introduction

Video object segmentation (VOS) aims to segment the specified object regions from the background throughout a video sequence. It is a fundamental task in computer vision, which can serve as an essential step for object tracking [1] and action detection [2] so as to support other video applications, such as video processing and editing [3], [4]. The task of VOS can be mainly divided into two categories: semi-supervised VOS [5], [6] and unsupervised VOS [7]. In this work, we mainly focus on the semi-supervised VOS, in which only the first frame is annotated in a test video. Semi-supervised VOS is a challenging task due to the frequent occurrence in videos of the occlusions, background distractions, appearances changes, etc. In particular, the occlusions and background distractions are critically needed to be handled in multi-object videos. Some examples of video frames with occlusions and distractions are shown in Fig. 1.

Fig. 1. Examples of video frames with occlusions and distractions. (a) and (c) exhibit the example of single-object and multi-object segmentation, while yellow square boxes are shown in (b). It can be seen that objects are more likely to be occluded in multi-object videos, not only by background but also by neighboring objects. (d) displays another frame in multi-object videos, and exists distractions from other similar objects, for example, multiple similar fish appear in the same frame.

As other tasks in computer vision, many methods based on deep learning [1], [5], [6], [8]–[21] have been developed to solve the semi-supervised VOS. In existing deep learning-based VOS methods, there exist two advanced frameworks including *mask propagation-based framework* and *matching-based framework*. The former one regards the VOS as a guided instance segmentation task that takes the previous frame's mask as the spatial guidance for segmenting the specific object [6], [11]–[16], [19], [20]. The latter one resolves the semi-supervised VOS by applying the deep embedding learning and pixel-level matching between the first frame and target frame [17], [18]. These two frameworks are described below.

1) Mask Propagation-based framework: A glimpse of the mask propagation-based framework is shown in Fig. 2(a). Due to temporal coherence, the mask propagated from the previous frame can provide rough spatial guidance for the segmentation of the current frame. With the guidance of the previous frame's mask, networks are more likely to detect the target object regions, which is favorable when encountering distractions from other objects or background. However, the input image and previous frame's mask are not aligned on account of the inaccurate mask and discrepancies between sequential frames

* Bo LI (leebo@scut.edu.cn) is the corresponding author.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

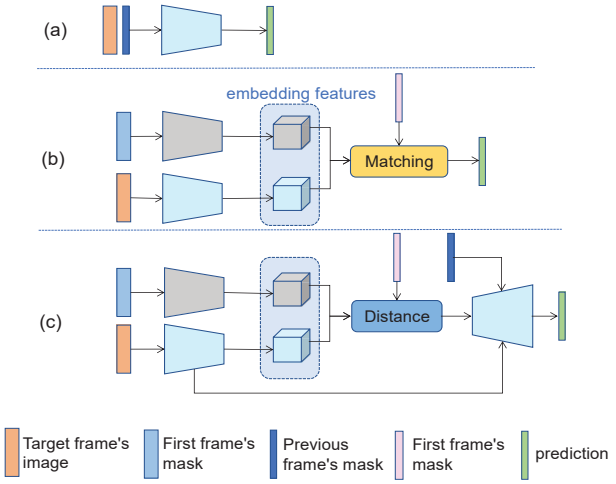

Fig. 2. Comparison with existing frameworks. (a) mask propagation-based framework, (b) matching-based framework and (c) the proposed DGMPM.

(e.g. position variations). Consequently, these unreliable masks may result in drifts or even disappearances when the target object is occluded. Many existing methods related to mask propagation [6], [13]–[15] adopt online learning to overcome occlusions. Online learning learns about the appearance of target objects by iterative optimization on the annotated first frame. It is clear that the high time consumption in online learning inevitably hinders the practical application of the VOS.

2) Matching-based framework: A draft of the matching-based framework is shown in Fig. 2(b). The main idea of matching-based framework maps the reference and target frame to metric space via a Siamese network, and classify every pixel in the target frame by pixel-wise matching between two frames. The matching-based framework can deal with position variations and occlusions, since it can perform global pixel-wise matching between the target frame and accurately annotated frame. Unfortunately, it may lead to mismatching thereby being difficult to distinguish similar objects, due to the fact that the embedding vectors extracted by deep networks may be similar if objects have the same semantics and appearances. Moreover, owing to the high computational complexity of pixel-wise matching and segmentation network [22], the efficiency of matching-based methods can be improved further.

To address occlusions and distractions efficiently, in this paper, a new semi-supervised VOS model is proposed, called *Distance-Guided Mask Propagation Model* (DGMPM). The intention of DGMPM is inspired by both the mask propagation-based framework and the matching-based framework. The sketch of the proposed DGMPM is depicted in Fig. 2(c). Different from recently-developed matching-based methods that adopt pixel-wise matching, DGMPM employs a novel *embedding distance module*. The strategy of the embedding distance module is to calculate the distance difference between target features and the centers of foreground/background features. This embedding distance module has less sensitivity to the feature scale, and meanwhile provides a soft cue for the latter decoder to restrain the effect of occlusions. Inspired by the mask propagation-based framework, DGMPM also uses the previous frame's mask as the spatial guidance to cope with distractions. Different from [6], [15], the previous frame's mask in DGMPM is downsampled and then merged into the decoder to mitigate the effect of the unreliable mask. Additionally, spatial attention blocks are designed in the decoder of DGMPM to extract the high-level semantic features for guiding the screening of low-level details stage-by-stage, which strengthens the network to focus on the target regions and rectify the prediction mask. The proposed DGMPM works efficiently at test time without online learning and time-consuming forward propagation process. Experiments on multiple benchmark datasets show that the proposed DGMPM has competitive accuracy and runtime in comparison with state-of-the-art methods. The contributions of this work are mainly threefold:

- A new Distance-Guided Mask Propagation Model is proposed for efficient semi-supervised VOS with competitive accuracy and runtime.
- By calculating the distance difference between target features and the centers of foreground/background features, a novel embedding distance module is realized to generate soft cue to mitigate the effect of occlusions.
- Spatial attention blocks are introduced to focus on the target regions and rectify prediction results.

II. Related Work

A. Semi-supervised Video Object Segmentation

In this paper, we focus on the semi-supervised VOS aiming to segment the specified object regions given from the first annotated frame. In recently-developed works of semi-supervised VOS, mask propagation-based and matching-based methods are relevant to this work.

Mask propagation-based methods [6], [11]–[16], [19], [20], [23] regard the video object segmentation as a guided instance segmentation. MSK [6] guides the network towards the target object by feeding in the mask of the previous frame and adopting online learning. Many works in DAVIS competition, such as [14], [15], also propagate the mask from the previous frame and achieve outstanding performance. The above-mentioned methods rely on online learning, which is a time-consuming training strategy that also used in other CNNs-based methods [5], [8], [10]. To promote speed performance, many mask propagation-based methods without online learning [11], [12], [16], [19], [20] are developed recently. OSMN [16] extracts information from the annotated first frame and produce a list of parameters, which manipulate layer-wise feature in the segmentation network. RVOS [11] incorporates recurrence on spatial and temporal domains, and feeds the mask of previous frame to the recurrent network for semi-supervised VOS. These offline methods have a good trade-off between speed and accuracy, but stills, have suboptimal performance due to the unreliable prediction mask, when it exists occlusions problems. In this work, mask propagation

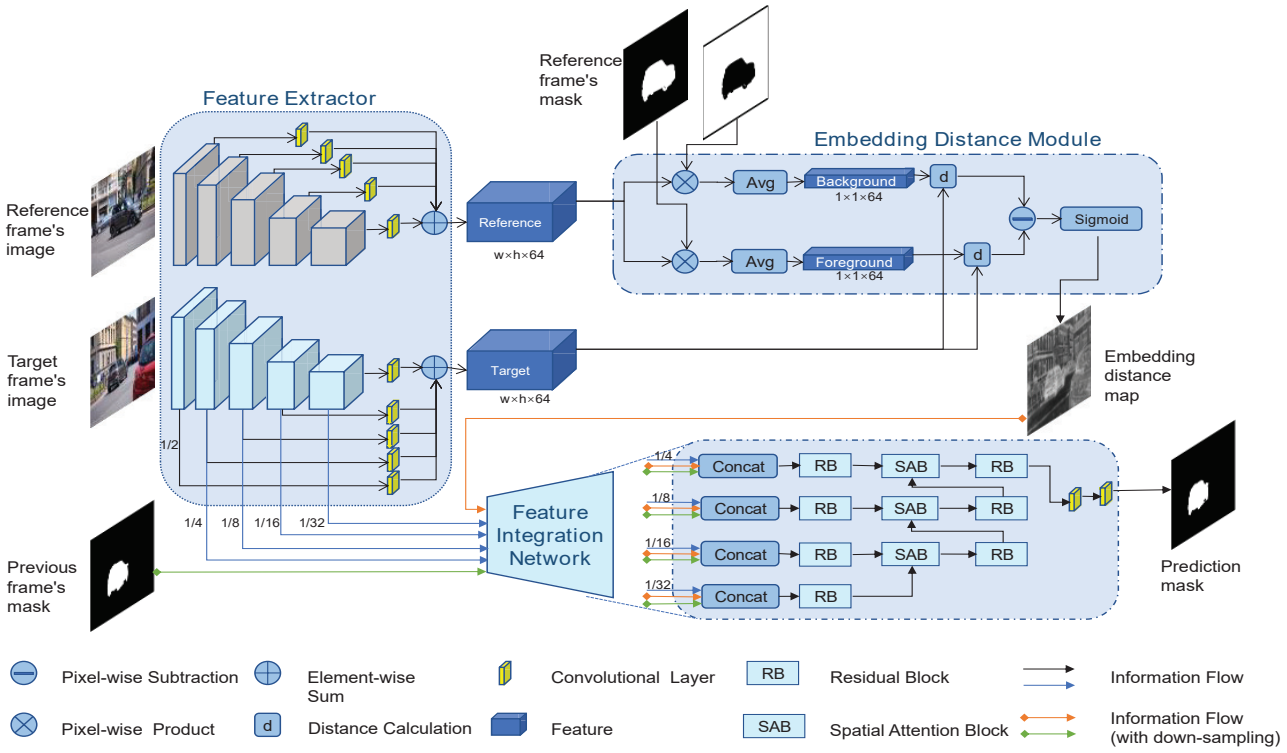

Fig. 3. The architecture of the proposed DGMPM. Feature extractor extracts the target feature and reference one with the shape of $w \times h \times 64$ (w, h is the width and height of the input frame). Embedding distance module generates embedding distance maps with the shape of $w \times h \times 1$. The final convolutional layers are applied to resize the feature to the output shape of $w \times h \times 2$. Best viewed in color.

strategy is also applied to reduce the effect of distractions. However, we design a module with global contrast between reference frame and target frame to tackle occlusions rather than relying on time-consuming online learning.

Matching-based methods [17], [18], [21] calculate pixel-level matching between the features of the reference and target frame in videos. PLM [21] consists of encoding and decoding models and conducts pixel-level object matching to detect the target object. PML [17] employs the k nearest neighbors to classify the target pixels in learned feature space. Due to the point-to-point correspondence strategy, these two methods are subject to distractions. VideoMatch [18] adopts soft matching upon the averaged top K similarity score maps to produce smooth predictions, and applies an outlier removal process to reduce the effect of distractions. These methods can tackle occlusions due to global pixel-wise matching. Owing to the high computational complexity of pixel-wise matching and segmentation network [22], [24], [25], the efficiency of these methods can be improved further. In this paper, we design a new embedding distance module that efficiently handles occlusions without time-consuming forward propagation process.

B. Spatial attention

Recently, attention mechanism is widely used in natural language processing and computer vision. In video captioning, Chen et al. [26] adopt channel attention and spatial attention to provide what and where the attention is. In semi-supervised VOS, Xu et al. [10] also design an attention module to merge features and focus on object regions. Li et al. [15] obtain the attention map by applying a convolutional layer and a softmax layer to the feature, which aims to reduce the distractions from the background and become more robust. Inspired by Yu et al. [27], in this paper, spatial attention blocks are employed to utilize high-level features to guide the screening of low-level details, which strengthens the network to focus on the target regions and rectify prediction results.

III. Proposed Method

In this work, an efficient *Distance-Guided Mask Propagation Model* (DGMPM) is introduced for semi-supervised VOS. The architecture is shown in Fig. 3. It is composed of three components, including feature extractor, embedding distance module and feature integration network. In the following sections, every component will be discussed in detail.

A. Feature extractor

Feature extractor is developed to map the input frames to metric space and provides multi-level features. Rather than utilizing semantic segmentation network [22], [24], [25] as previous works [6], [18] have done, for the sake of the efficiency, we devise a light-weight one for our method. Feature extractor is a pair of Siamese network that contains a reference branch and a target branch. In each branch, an encoder and five convolutional layers map the input frame into a 64 channels feature with the same resolution as the input. The encoder

is a trained ResNet [28] in which the last pooling layer and fully connected layer are removed. Convolutional layers are utilized to extract multi-level features and unify their numbers of channels. These convolutional layers with kernel size of 3×3 extract features from the side outputs of *conv1, res2, res3, res4, res5* in ResNet. Then the outputs of these convolutional layers are up-sampled by bilinear interpolation and merged by element-wise addition. In this way, features with multi-level representation are generated by feature extractor, while each pixel in the input frame corresponds to an embedding vector in the feature.

B. Embedding distance module

In semi-supervised VOS, the first frame's mask in the test video is provided to determine the target objects. It provides an important cue of visual appearance. To exploit the annotated first frame efficiently, a novel embedding distance module is implemented in this work. The module produces embedding distance maps with the global view of the reference feature, which can provide spatial guidance for the latter decoder and is helpful to tackle occlusions.

The details of the embedding distance module are shown in Fig. 3. Let $\boldsymbol{H} \in \mathbb{R}^{w\times h\times c}$ denote the reference feature, and $\boldsymbol{T} \in \mathbb{R}^{w\times h\times c}$ denotes the target feature. Embedding vectors $\boldsymbol{t}_n \in \mathbb{R}^{1\times1\times c}$ and $\boldsymbol{h}_n \in \mathbb{R}^{1\times1\times c}$ are obtained from $\boldsymbol{T}$ and $\boldsymbol{H}$ respectively, while the index n denotes the position of the pixel. The label of a pixel at position n in reference frame is $m_n \in \boldsymbol{M}$, where $\boldsymbol{M} \in \{0,1\}^{w\times h}$ represents the reference frame's mask. Then, the foreground feature and background feature at position n ($\boldsymbol{h}_{n,f}$ and $\boldsymbol{h}_{n,b}$) are individually defined as

$$\boldsymbol{h}_{n,f} = m_n \boldsymbol{h}_n, \tag{1}$$

$$\boldsymbol{h}_{n,b} = |1 - m_n|\, \boldsymbol{h}_n. \tag{2}$$

Let $h^c_{n,f}$ denotes the c-th channel of $\boldsymbol{h}_{n,f}$, and $h^c_{n,b}$ denotes the c-th channel of $\boldsymbol{h}_{n,b}$. Then, the foreground anchor and background anchor are defined by

$$a^c_f = \frac{\sum_n h^c_{n,f}}{\sum_n m_n} \tag{3}$$

$$a^c_b = \frac{\sum_n h^c_{n,b}}{\sum_n |1 - m_n|} \tag{4}$$

where a^c_f represents the c-th channel of the foreground anchor $\boldsymbol{a}_f \in \mathbb{R}^{1\times1\times c}$, a^c_b denotes the c-th channel of the background anchor $\boldsymbol{a}_b \in \mathbb{R}^{1\times1\times c}$. In this way, the foreground and background anchor are yielded by calculating the average of foreground and background vectors, which can be cast as the cluster centers of the foreground and background. The embedding distance map is denoted as $\boldsymbol{E} \in \mathbb{R}^{w\times h\times1}$ and $e_n \in \boldsymbol{E}$ in position n. The embedding distance map is calculated by

$$e_n = Sigmoid\left(Dis\left(\boldsymbol{t}_n, \boldsymbol{a}_b\right) - Dis\left(\boldsymbol{t}_n, \boldsymbol{a}_f\right)\right) \tag{5}$$

where $Dis(\cdot,\cdot)$ represents Euclidean distance, and the sigmoid activation function normalizes the range of e_n to [0,1]. In this way, the pixel whose embedding vector is near to the foreground anchor and far from the background anchor will be highlighted.

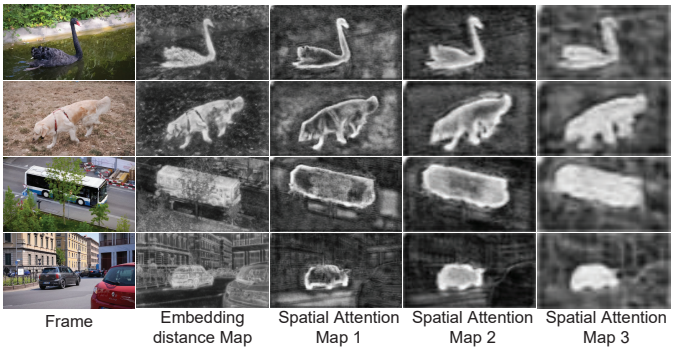

Fig. 4. Some examples of embedding distance maps and spatial attention maps. The spatial attention maps generated by the multiple spatial attention blocks in Fig. 3, are shown in columns 3-5 from shallow to deep layers respectively. Spatial attention maps have been resized for a better view.

The embedding distance module has two advantages: a) Embedding distance module is based on the global contrast between target and reference features. The anchors are generated by feature averaging with a global view of reference frame. Each pixel in the embedding distance map is determined by the distance difference between the embedding vector and the anchors. Thus, the embedding distance module is based on global contrast between target and reference features, which is advantageous to detect and highlight the object regions in the target frame even though the target objects are partially occluded or reappears after occlusions. b) The embedding distance module is non-parametric and simple. It conducts the distance calculation between the target feature and two anchors in metric space. Therefore, this module has low computation costs and is less sensitive to the feature scale.

Several examples of embedding distance maps are visualized in the second column of Fig. 4. It can be observed that the foreground regions are highlighted in these feature maps. In practice, the embedding distance maps generated by embedding distance module provide strong guidance for the latter decoder to handle occlusions. The effectiveness of embedding distance module will be discussed in Section IV-D.

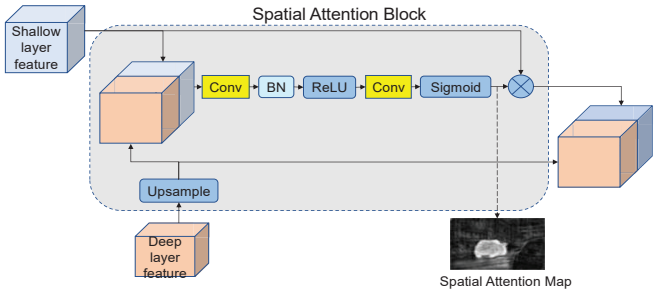

Fig. 5. The structure of spatial attention block. It inputs the shallow layer and deep layer feature, and outputs a feature that has the same shape as the shallow layer feature. '→' denotes information flow and '⇢' signifies the attention map, which is the output of the Sigmoid layer.

C. Feature integration network

There are a variety of distractions from other objects or background in videos, which have similar semantics or appearances with the target object. The regions of these objects or background may be highlighted in the embedding distance map. As illustrated in the final row of Fig. 4, not only the regions of the target car but also the regions of another red car are highlighted simultaneously. Hence, it is incapable of distinguishing distractors only under the guidance of the embedding distance map. In contrast to embedding distance map, the previous frame's mask can provide spatial guidance, which leads the network to segment the target regions and helps to reduce the effect of distractions. Therefore, we devise a feature integration network to merge the multi-level features with the guidance of both the previous frame's mask and the embedding distance map. This decoder network takes advantage of two guiding maps and extracts discriminative features for the specific object to generate final segmentation.

1) The architecture of feature integration network: The structure of the feature integration network is inspired by a semantic segmentation network DFN [27]. The details are shown in Fig. 3. There are two main differences between the feature integration network and the smooth network of DFN. Firstly, spatial attention blocks are applied in our network rather than the channel attention. The channel attention blocks are introduced in DFN to decrease intra-class inconsistency for semantic segmentation. However, VOS is an instance segmentation task, in which there exist multiple instances with the same semantics. Thus, channel attention is not perfectly appropriate for VOS. In this work, spatial attention blocks are developed to focus on the target regions and rectify prediction results. Secondly, two guiding maps, which are the embedding distance map and the previous frame's mask, are concatenated with side outputs of ResNet to extract multi-level features for the specific object. These two feature maps are down-sampled to mitigate the effect of inaccurate guidance. In addition to the two main differences, residual blocks [28] with output channels of 128 are employed to extract features for the specific object and merge multi-level features. An up-sampling layer and two convolutional layers are utilized to expand the resolution and decrease the channels of the outputs.

2) Spatial attention block: In the feature integration network, spatial attention blocks are developed to change the weights of multi-layer features in space. It utilizes high-level semantic features to guide the screening of shallow details stage-by-stage. The structure of a spatial attention block is shown in Fig. 5. The features in shallow and deep layers are defined as $\boldsymbol{F}_l$ and $\boldsymbol{F}_d$ respectively. The spatial attention map $\boldsymbol{\alpha}$ is generated by

$$\boldsymbol{\alpha} = \mathit{Sigmoid}\left(f(\mathit{Cat}\left(\mathit{Up}\left(\boldsymbol{F}_d\right), \boldsymbol{F}_l\right))\right) \tag{6}$$

where $\mathit{Up}(\cdot)$ is the bilinear interpolation, $\mathit{Cat}(\cdot)$ is a concatenation operation, and the sigmoid activation function transforms the feature into a probability map. The $f(\cdot)$ is formulated as

$$f(\boldsymbol{x}) = W_2 * \sigma\left(W_1 * \boldsymbol{x}\right) \tag{7}$$

where $\sigma(\cdot)$ indicates the ReLU activation function, * denotes the convolution operation, and W_1, W_2 denote the convolutions with the weights of $3 \times 3 \times 32$ and $3 \times 3 \times 1$. In this way, a spatial attention map is extracted from the deep layer feature $\boldsymbol{F}_d$ and shallow layer feature $\boldsymbol{F}_l$. Then, the spatial attention map is regarded as additional spatial weights for $\boldsymbol{F}_l$, and the output of spatial attention block $\boldsymbol{F}_o$ is generated by

$$\boldsymbol{F}_o = \mathit{Cat}\left(\mathit{Up}\left(\boldsymbol{F}_d\right), \boldsymbol{\alpha} \otimes \boldsymbol{F}_l\right) \tag{8}$$

where $\otimes$ denotes the element-wise product.

The spatial attention block applies an attention map on the feature, which represents the spatial feature selection. With this design, the network is strengthened to focus on the target regions. As shown in Fig. 4, object regions are highlighted in these attention maps. Moreover, note that in Fig. 4, the spatial attention block in shallow layer pays attention to the object boundary, while the block in deep layer focuses on the overall regions. It demonstrates that the spatial attention blocks can fully explore the feature of each layer to rectify the segmentation. The effectiveness of the spatial attention blocks will be discussed in Section IV-D.

IV. Experiments

A. Datasets and Evaluation Metrics

To validate the effectiveness of the proposed model, we validate on three widely-used benchmark sets in VOS, including DAVIS2017 [29], DAVIS [30] and SegTracks v2 [31]. DAVIS2017 is a multi-object dataset that has 60 sequences and 4219 annotated frames in the training set, and 30 sequences and 2023 annotated frames in the valuation set. DAVIS, a subset of DAVIS2017, is a single-object dataset with 30 training videos and 20 validation videos. SegTracks v2 consists of 14 videos with 976 annotated frames.

The evaluation metrics used in this paper for DAVIS2017 and DAVIS are defined in [30], including the region similarity J and contour accuracy F. The region similarity J is the mean intersection over union between the prediction mask and the ground truth. The contour accuracy F is calculated by the F-measure between the contour points of the ground truth and prediction mask. The average of the mean of J and F ($J\&F$) is also used to present the overall performance. In terms of the SegTrack v2 dataset, the mean intersection-over-union (mIOU) is adopted for performance evaluation.

B. Implement details

1) Training: Considering that the main training set DAVIS2017 [29] only has a limited amount of sequences and object classes, the proposed DGMPM is trained with three stages to avoid the over-fitting problem. In the first stage, we pre-train on MS COCO [32], a large semantic segmentation dataset, to avoid the over-fitting problem. To simulate training samples from this static image dataset, by following the same practice in [6], the previous frame's masks are synthesized by using multiple augmentation schemes, including affine transformation, random scaling and random shifting, whereas the reference frames are generated by employing

TABLE I
THE QUANTITATIVE EVALUATIONS ON THE VALIDATION SET OF DAVIS2017. THE OL DENOTES ONLINE LEARNING. $J\&F$ IS THE AVERAGE OF THE MEAN OF J AND F.

Methods	OL	DAVIS2017		
		J	F	$J\&F$
OnAVOS [8]	√	**61.6**	**69.1**	**65.4**
STCNN [10]	√	58.7	64.6	61.7
OSVOS [5]	√	55.1	62.1	58.6
MSK [6]	√	51.2	57.3	54.3
VideoMatch [18]		56.5	**68.2**	62.4
RVOS [11]		57.6	63.6	60.6
FAVOS [12]		54.6	61.8	58.2
SiamMask [1]		54.3	58.5	56.4
OSMN [16]		52.5	57.1	54.8
DGMPM		**61.6**	67.8	**64.7**

lucid data dreaming [14]. In the second stage, we further fine-tune the model on YouTube VOS [33], a large benchmark containing 3471 videos, 65 categories and 5945 objects, consequently improving the generalization performance. In the third stage, we fine-tune our DGMPM on the main training set DAVIS2017 [29]. During the second and third stages, we randomly select two frames as the target and reference frames, and the mask of a frame near the target frame as the previous frame's mask (the maximum interval is 3 in our experiment).

The proposed DGMPM is trained using Adam optimizer to minimize the cross-entropy loss. The learning rates are set to 10^{-5}, 5×10^{-6}, 10^{-6} in three stages respectively. All inputs are scaled into 800×400 pixels, and then apply data augmentation containing random scaling, random flipping, random rotation and normalization. Erosion is also applied to the previous frame's mask to imitate prediction masks with coarse boundary. All experiments in this section are trained on a single NVIDIA Titan XP.

2) Inference: We set the annotated first frame as the reference and predict the masks of subsequent frames in a video. For multi-object videos, we predict the masks frame-by-frame and run each object independently in the embedding distance module and feature integration network. The label of the largest prediction is assigned to the final result.

C. Performance comparison and analysis

The performance of DGMPM is evaluated on the validation sets of DAVIS2017 [29], DAVIS [30] and SegTrack v2 [31], and compared with multiple recently-developed CNNs-based offline methods, including VideoMatch [18], SiamMask [1], RVOS [11], FAVOS [12], OSMN [16], PML [17], CTN [19], VPN [20] and PLM [21]. The performance resulted from multiple online learning methods are also shown including OSVOS [5], STCNN [10], MSK [6], OnAVOS [8] and MoNet [9]. The performance comparison among various methods on DAVIS2017 [29], DAVIS [30] and SegTrack v2 [31] are shown on Table I, Table II and Table III respectively. Note that the performance of previous works are available by corresponding published papers and the DAVIS challenge website.

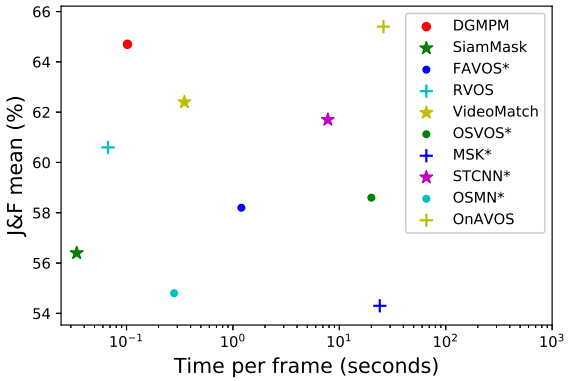

Fig. 6. Accuracy and runtime comparison among various methods on the DAVIS2017 dataset. '$*$' indicates that the runtime is extrapolated from DAVIS assuming linear scaling in the number of objects.

TABLE II
THE QUANTITATIVE EVALUATIONS ON THE VALIDATION SET OF DAVIS. THE SPEED SHOWS THE AVERAGE TEST TIME PER FRAME.

Methods	OL	Speed(s)	DAVIS		
			J	F	$J\&F$
OnAVOS [8]	√	13	**86.1**	**84.9**	**85.5**
MoNet [9]	√	14.1	84.7	84.8	84.8
STCNN [10]	√	**3.9**	83.8	83.8	83.8
OSVOS [5]	√	10	79.8	80.6	80.2
MSK [6]	√	12	79.7	75.4	77.6
SFL [13]	√	7.9	74.8	74.5	74.7
VideoMatch [18]		0.32	**81.0**	**80.8**	**80.9**
PML [17]		0.28	75.5	79.3	77.4
FAVOS [12]		0.6	77.9	76.0	77.0
OSMN [16]		0.14	74.0	72.9	73.5
CTN [19]		30	73.5	69.3	71.4
VPN [20]		0.3	70.0	62.0	66.0
PLM [21]		0.3	70.0	62.0	66.0
DGMPM		**0.071**	79.0	78.7	78.9

TABLE III
THE QUANTITATIVE EVALUATIONS ON THE SEGTRACK V2. THE METRIC MIOU IS THE MEAN INTERSECTION OVER UNION BETWEEN THE PREDICTION AND GROUND TRUTH.

Methods	OSVOS [5]	MSK [6]	MoNet [9]	DGMPM
mIOU	65.4	70.3	72.4	**73.5**

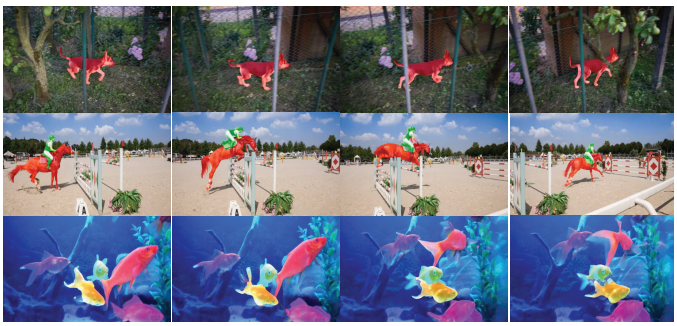

Fig. 7. The qualitative results of DGMPM on the DAVIS2017 validation set.

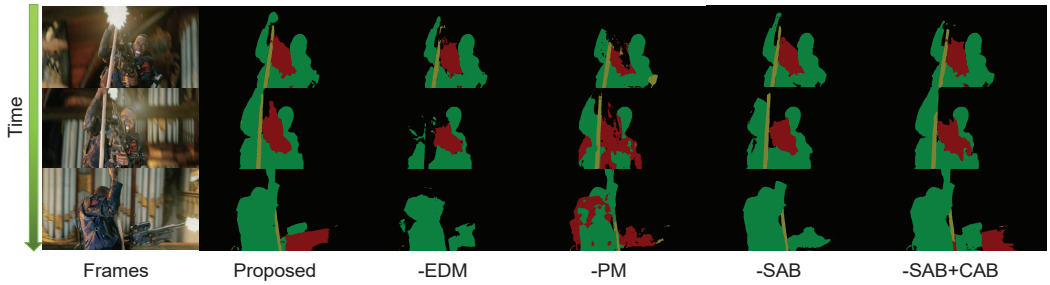

Fig. 8. Visual results of the proposed and ablated models. -EDM, -PM and -SAB individually denote embedding distance module, previous frame's mask and spatial attention blocks are removed from DGMPM, whereas -SAB+CAB denotes the spatial attention blocks are replaced by the channel attention blocks.

The performance of DGMPM and multiple online and offline methods on the DAVIS2017 validation set are shown in Table I. One can see that DGMPM method achieves the higher J, F and $J\&F$ among most of the online and offline methods under comparison, and has competitive J and $J\&F$ in comparison with the OnAVOS [8] and comparable F to VideoMatch [18]. Moreover, from the accuracy and runtime comparison shown in Fig. 6, it can be observed that the runtime of DGMPM is significantly lower than that of all the online learning methods. This is due to the efficient inference without online learning and the low computational cost in the forward propagation process in our proposed DGMPM. Besides, the qualitative results of DGMPM on DAVIS2017 are shown in Fig. 7. It can be seen that DGMPM is capable of efficiently dealing with the problems of occlusions and distractions, and achieve desirable performance results, especially in the case of the multi-object VOS task.

Table II shows the performance of DGMPM and multiple online and offline learning methods on the DAVIS dataset. Likewise, one can see that the performance of J, F and $J\&F$ of DGMPM is better or competitive in comparison with multiple online and offline methods. Although the performance results of DGMPM are lower than several online methods, DGMPM under comparison achieves a faster speed(0.071 seconds per frame). The proposed DGMPM performs a bit lower than VideoMatch [18] that adopts additional online updating and outlier removal to achieve higher performance. Online updating mechanism can be complementary to our method and need further exploration in our future work.

To evaluate the generalization performance, a cross-dataset evaluation is performed among the proposed DGMPM and three methods including OSVOS [5], MSK [6], and MoNet [9]. Specifically, we evaluate it on a SegTrack V2 dataset, a completely unseen dataset. The corresponding results are shown in Table III, it can be seen that the best result (i.e., 73.5% mIOU) yielded by DGMPM is consistently higher than that of OSVOS [5], MSK [6], and MoNet [9]. This study shows that DGMPM has better generalization performance.

D. Ablation study

To illustrate the effectiveness of the individual module in our proposed DGMPM, the ablation experiments are conducted on the DAVIS2017 validation set. The models of ablation experiments are trained on the MS COCO [32] and DAVIS2017 [29], and the corresponding results are shown in Table IV.

TABLE IV
ABLATION STUDY. -EDM, -PM AND -SAB INDIVIDUALLY DENOTE EMBEDDING DISTANCE MODULE, PREVIOUS FRAME'S MASK AND SPATIAL ATTENTION BLOCKS ARE REMOVED FROM DGMPM, WHEREAS -SAB+CAB DENOTES THE SPATIAL ATTENTION BLOCKS IN DGMPM ARE REPLACED BY THE CHANNEL ATTENTION BLOCKS. THE FINAL LINE DENOTES THE $J\&F$ PERFORMANCE CHANGES IN THE ABLATION STUDY.

Metric	Proposed	-EDM	-PM	-SAB	-SAB+CAB
J	59.1	50.6	44.2	55.0	57.2
F	65.1	53.2	51.5	60.3	62.7
$J\&F$	62.1	51.9	47.9	57.7	60.0
△	-	**-10.2**	**-14.2**	**-4.4**	**-2.1**

From the results, note that the performance degradation of DGMPM is incurred after disabling the *embedding distance module* (-EDM), *previous frame's mask* (-PM). The $J\&F$ decrement of 10.2 and 14.2 is incurred after disabling the embedding distance module and previous frame's mask respectively. This indicates that the embedding distance map generated by the embedding distance module provides a strong cue for DGMPM to handle the occlusions, and the temporal information from the previous frame's mask can mitigate the effect of distractions. Moreover, the ablation experiments with respect to *spatial attention blocks* are also conducted in this work. Besides studying the performance after disabling the spatial attention blocks (-SAB) in DGMPM, we also explore the performance after replacing the spatial attention blocks by channel attention blocks (-SAB+CAB). As shown in Table IV, the $J\&F$ decrement of 4.4 and 2.1 are incurred in the model -SAB and -SAB+CAB respectively. This result demonstrates that spatial attention blocks can strengthen the network to focus on the target regions and rectify the prediction results.

In addition to the quantitative results in Table IV, Fig. 8 also visualize the results among DGMPM and above-mentioned ablation schemes. As shown in Fig. 8, similar conclusions can be demonstrated from the qualitative results. It shows that DGMPM embedded with embedding distance module and the previous frame's mask can effectively mitigate the effect of occlusions and distractions, while spatial attention blocks in DGMPM provide better attention on target regions and helps for more accurate predictions.

V. Conclusion

In this paper, a new *Distance-Guided Mask Propagation Model* (DGMPM) is proposed for semi-supervised video object segmentation to efficiently tackle occlusions and distractions. An embedding distance module with global contrast between the target and reference features, has less sensitive to the feature scale and is helpful to mitigate the effect of occlusions. The prior knowledge of the previous frame provides spatial guidance to reduce the effect of distractions. In addition, spatial attention blocks strengthen the network to focus on target regions and rectify prediction results. Extensive experiments demonstrate that DGMPM is capable of efficiently dealing with occlusions and distractions, as well as achieves competitive accuracy and runtime in comparison with state-of-the-art methods.

VI. Acknowledgment

This research was supported by National Key R&D Program of China (No. 2017YFC0806000), by National Natural Science Foundation of China (No. 11627802, 51678249), State Key Lab of Subtropical Building Science, South China University of Technology (2018ZB33), and the State Scholarship Fund of China Scholarship Council (201806155022).

References

[1] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, "Fast online object tracking and segmentation: A unifying approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1328–1338.

[2] R. Hou, C. Chen, and M. Shah, "An end-to-end 3d convolutional neural network for action detection and segmentation in videos," *arXiv preprint arXiv:1712.01111*, 2017.

[3] W. Wang, X. Lu, D. Crandall, J. Shen, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," *IEEE International Conference on Computer Vision*, 2019.

[4] W. Wang, J. Shen, and F. Porikli, "Selective video object cutout," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5645–5655, 2017.

[5] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 221–230.

[6] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2663–2672.

[7] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli, "Unsupervised online video object segmentation with motion property understanding," *IEEE Transactions on Image Processing*, vol. 29, pp. 237–249, 2019.

[8] P. Voigtlaender and B. Leibe, "Online adaptation of convolutional neural networks for video object segmentation," *arXiv preprint arXiv:1706.09364*, 2017.

[9] H. Xiao, J. Feng, G. Lin, Y. Liu, and M. Zhang, "Monet: Deep motion exploitation for video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1140–1148.

[10] K. Xu, L. Wen, G. Li, L. Bo, and Q. Huang, "Spatiotemporal cnn for video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1379–1388.

[11] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro-i Nieto, "Rvos: End-to-end recurrent network for video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5277–5286.

[12] J. Cheng, Y.-H. Tsai, W.-C. Hung, S. Wang, and M.-H. Yang, "Fast and accurate online video object segmentation via tracking parts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7415–7424.

[13] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," in *IEEE International Conference on Computer Vision*, 2017, pp. 686–695.

[14] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele, "Lucid data dreaming for multiple object tracking," *arXiv preprint arXiv:1703.09554*, 2017.

[15] X. Li and C. Change Loy, "Video object segmentation with joint re-identification and attention-aware mask propagation," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 90–105.

[16] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos, "Efficient video object segmentation via network modulation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6499–6507.

[17] Y. Chen, J. Pont-Tuset, A. Montes, and L. Van Gool, "Blazingly fast video object segmentation with pixel-wise metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1189–1198.

[18] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, "Videomatch: Matching based video object segmentation," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 54–70.

[19] W.-D. Jang and C.-S. Kim, "Online video object segmentation via convolutional trident network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5849–5858.

[20] V. Jampani, R. Gadde, and P. V. Gehler, "Video propagation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 451–461.

[21] J. Shin Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. So Kweon, "Pixel-level matching for video object segmentation using convolutional neural networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 2167–2176.

[22] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[23] J. Luiten, P. Voigtlaender, and B. 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*, vol. 1, no. 2, 2018, p. 6.

[24] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[26] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5659–5667.

[27] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1857–1866.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[29] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675*, 2017.

[30] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 724–732.

[31] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *IEEE International Conference on Computer Vision*, 2013, pp. 2192–2199.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[33] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "Youtube-vos: Sequence-to-sequence video object segmentation," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 585–601.