

# Tackling Background Distraction in Video Object Segmentation

Suhwan Cho<sup>1</sup> Heansung Lee<sup>1</sup> Minhyeok Lee<sup>1</sup> Chaewon Park<sup>1</sup>  
Sungjun Jang<sup>1</sup> Minjung Kim<sup>1</sup> Sangyoun Lee<sup>1,2</sup>

<sup>1</sup> Yonsei University

<sup>2</sup> Korea Institute of Science and Technology (KIST)

**Abstract.** Semi-supervised video object segmentation (VOS) aims to densely track certain designated objects in videos. One of the main challenges in this task is the existence of background distractors that appear similar to the target objects. We propose three novel strategies to suppress such distractors: 1) a spatio-temporally diversified template construction scheme to obtain generalized properties of the target objects; 2) a learnable distance-scoring function to exclude spatially-distant distractors by exploiting the temporal consistency between two consecutive frames; 3) swap-and-attach augmentation to force each object to have unique features by providing training samples containing entangled objects. On all public benchmark datasets, our model achieves a comparable performance to contemporary state-of-the-art approaches, even with real-time performance. Qualitative results also demonstrate the superiority of our approach over existing methods. We believe our approach will be widely used for future VOS research. Code and models are available at <https://github.com/suhwan-cho/TBD>.

**Keywords:** Video object segmentation, Metric learning, Temporal consistency, Video data augmentation

## 1 Introduction

Video object segmentation (VOS) aims to keep tracking certain designated objects within an entire given video at the pixel level. Depending on the type of guidance provided with regard to the target objects, VOS can be divided into semi-supervised VOS, unsupervised VOS, referring VOS, and other sub-categories. We focus on the semi-supervised setting, in which densely-annotated objects are provided in the initial frame of a video. Given its diverse applicability, semi-supervised VOS has recently attracted extensive attention in many vision fields such as autonomous driving, video editing, and video surveillance.

In semi-supervised VOS, one of the main challenges is the existence of background distractors that have a similar appearance to the target objects. As comparing visual properties is a fundamental technique to detect and track the designated objects, visual distractions can severely lower the reliability of a system. We propose three novel strategies to suppress the negative influence

of background distractions in VOS: 1) a spatio-temporally diversified template construction scheme to prepare various object properties for reliable and stable prediction; 2) a learnable distance-scoring function to consider the temporal consistency of a video; 3) swap-and-attach data augmentation to provide hard training samples showing severe occlusions.

As VOS is a pixel-level classification task that requires fine-grained information, most feature matching-based approaches including VideoMatch [12] and RANet [36] adopt pixel-level fine matching. While this approach is popular owing to its superior ability to capture details, it also has the shortcoming of easily causing noise. Given that each element in its template has small receptive fields, it is very susceptible to background distractors. To address this issue, we propose a novel way to construct a coarse matching template by compressing the fine matching template considering the probability of each pixel location. Unlike the fine matching template, each element in the coarse matching template covers large receptive fields that are dynamically defined based on past predictions. By employing fine matching and coarse matching simultaneously, more stable predictions can be obtained thanks to improved spatial diversity. In addition, to construct various templates exploiting multiple temporal properties in a video, we operate multiple templates that are independently built and updated based on their own strategies. The proposed spatio-temporally diversified template construction enables the model to capture local details while simultaneously obtaining a clear distinction between foreground and background as well as learning the ability to utilize various time-specific properties.

While visual information is important, exploiting temporal consistency between adjacent frames is also important for a robust VOS. To leverage the temporal consistency of a video, FEELVOS [34] and CFBI [41] apply a square window when transferring the reference frame information to the query frame. RMNet [38] reduces the search area based on the coarsely predicted segmentation mask of the query frame. These hard windowing methods, which totally exclude the non-candidates, are effective for capturing the locality of a video, but as the size of a window is a discrete value, a human-tuning process is required for setting the hyper-parameters, which makes the solution less elegant and incomplete. To overcome this issue, we propose a learnable distance-scoring function that takes the distance between two pixels of consecutive frames as its input and simply outputs a spatial distance score from 0 to 1. When transferring the information from a reference frame pixel to a query frame pixel, it outputs a low score for distant pixels and vice versa.

We also propose a novel data augmentation technique for VOS, termed swap-and-attach augmentation. By simply swapping the objects between multiple sequences and attaching the swapped objects on the frames, realistic video data with severe occlusions can be simulated. As this creates training snippets containing multiple entangled objects, the model can learn strong feature representations to force each object to have unique features. Furthermore, it generalizes the model by dramatically increasing the amount of VOS training data, which is scarce as compared to other vision tasks.

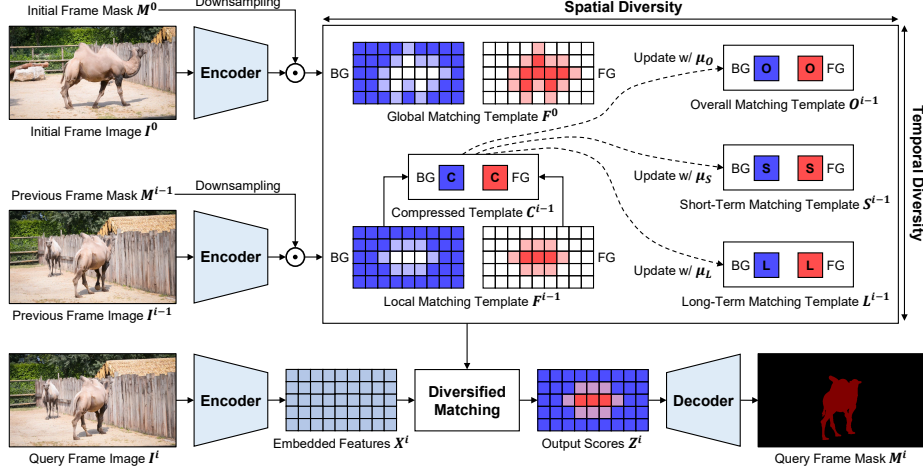
We validate the effectiveness of our method by evaluating it on three public VOS benchmark datasets, i.e., DAVIS 2016 [28], DAVIS 2017 [29], and YouTube-VOS 2018 [40] datasets. On all datasets, our method achieves a competitive performance compared to the current state-of-the-art methods while maintaining a real-time inference speed. In particular, in complex scenarios that have background distractors, our approach is proven to be much more effective than state-of-the-art solutions. Note that all our experiments are implemented on a single GeForce RTX 2080 Ti GPU, which makes our method much more accessible than other state-of-the-art methods, which require powerful hardware resources. We believe our proposed three novel strategies will be widely used for future VOS research.

Our main contributions can be summarized as follows:

- We introduce a spatio-temporally diversified template construction mechanism to prepare diverse features for stable and robust feature matching.
- We propose an end-to-end learnable distance-scoring function to fully exploit the temporal consistency between two consecutive frames in a video.
- We present swap-and-attach augmentation to learn a strong VOS model by ensuring the diversity of training data and simulating occluded samples.
- Our proposed approach achieves a competitive performance on the DAVIS and YouTube-VOS datasets, while maintaining a real-time performance.

## 2 Related Work

**Feature similarity matching.** Contemporary VOS methods track and segment the designated objects based on the similarity of the embedded features. Given that semi-supervised VOS is a pixel-level classification task, pixel-level fine matching is adopted in most cases. PML [3] learns a pixel-wise embedding using the nearest neighbor classifier. VideoMatch [12] produces foreground and background similarity maps using a soft matching layer. Extending these works, FEELVOS [34] and CFBI [41] exploit global matching and local matching to obtain long-term and short-term appearance information. RANet [36] maximizes the use of similarity maps from feature similarity matching by ranking and selecting conformable feature maps as per their importance. To fully exploit the information extracted from all previous frames, STM [27] adopts a memory network by defining the past frames with object masks as external memory and the current frame as the query. The query and the memory are densely matched in feature space, covering all the space-time pixel locations. Extended from STM, KMN [31] proposes memory-to-query matching to reduce background distractions when transferring the memorized information by applying a 2D Gaussian kernel. EGMN [25] exploits an episodic memory network where frames are stored as nodes so that cross-frame correlations can be captured effectively. These memory-based methods achieve admirable performance but face a serious issue: the size of memory increases over time, and therefore, they are not efficient for long videos. To address this, GC [19] stores and updates a key-value mapping instead of storing all the key and value features of



**Fig. 1.** Architecture of our proposed method. The templates for feature similarity matching are independently constructed and updated as per their own objectives. The spatial diversity and temporal diversity of the templates are represented horizontally and vertically, respectively. Several components, e.g., skip connections and mask propagation, are omitted for better clarification of the overall architecture.

the past frames. AFB-URR [21] introduces an adaptive feature bank update scheme to efficiently discard obsolete features and absorb new features based on a pre-defined policy.

**Leveraging temporal consistency.** As a video shares some common properties between the consecutive frames, there are several VOS methods that leverage this temporal consistency of a video. FEELVOS and CFBI apply a fixed-sized square window for a local feature similarity matching. When transferring information from the previous adjacent frame to the query frame, each query frame pixel can only refer to the region within a square window centered on the spatial location of that pixel. Similarly, RMNet [38] also reduces the search area by applying a square window. An optical flow model first coarsely predicts the segmentation mask of the query frame by warping the previous frame segmentation mask. Then, a square window is generated and applied based on the coarsely predicted query frame segmentation mask. These hard windowing methods are effective at leveraging the temporal consistency of a video, but as the size of a window is a discrete value, it needs a human-tuning process and is also not end-to-end learnable.

### 3 Approach

In our framework, frames in a video sequence are segmented using the ground truth segmentation mask given in the initial frame. Mask prediction is performed based on feature similarity between the embedded features. To prepare various

template features for feature matching, we employ a spatio-temporally diversified template construction mechanism. The results of feature matching are then fed into the decoder with low-level features extracted from the encoder and a downsampled previous adjacent frame mask. The overview of our framework is illustrated in Figure 1.

### 3.1 Deriving Spatial Diversity

Following previous approaches [6, 12, 34, 36, 41], we employ pixel-level fine matching to capture the fine-grained information that is essential to generate detailed segmentations. Given that  $I^i$  is an input image at frame  $i$ , the embedded features for that frame are defined as  $X^i \in [-1, 1]^{C \times H \times W}$  after channel L2 normalization. The provided or predicted segmentation mask and its downsampled version are denoted as  $M^i \in [0, 1]^{2 \times H_0 \times W_0}$  and  $m^i \in [0, 1]^{2 \times H \times W}$ , respectively. The first channel indicates the probabilities of the background, and the second channel indicates those of a foreground object. Using  $X^i$  and  $m^i$ , we define the pixel-level fine matching template  $F^i \in [-1, 1]^{C \times H \times W}$  as

$$\begin{aligned} F_{BG}^i &= X^i \odot m_0^i \\ F_{FG}^i &= X^i \odot m_1^i, \end{aligned} \quad (1)$$

where  $\odot$  denotes the Hadamard product and  $m_0^i$  and  $m_1^i$  denote the first and second channels of  $m^i$ , respectively.

Fine matching is effective at capturing the details that are inherent at the pixel-level, but thereafter, is susceptible to background distractors that have a similar appearance to a foreground object. In order to obtain more general features for an object and its background, i.e., to obtain spatial diversity, we propose coarse matching to complement the fine matching. Unlike the fine matching template whose receptive fields are small and local, the receptive fields of the coarse matching template are dynamically defined covering a large range. Therefore, coarse matching is able to obtain a clear distinction between foreground and background because outlier features are less likely to negatively affect the feature similarity matching scores.

To design the coarse matching template, we first build a compressed template  $C^i \in [-1, 1]^{C \times 1 \times 1}$  that indicates the normalized mean features of  $X^i$  weighted by the probabilities of each class in every spatial location. By using pre-computed  $F^i$ , it can be summarized as

$$\begin{aligned} C_{BG}^i &= \mathcal{N}(\mathcal{S}(X^i \odot m_0^i)) = \mathcal{N}(\mathcal{S}(F_{BG}^i)) \\ C_{FG}^i &= \mathcal{N}(\mathcal{S}(X^i \odot m_1^i)) = \mathcal{N}(\mathcal{S}(F_{FG}^i)), \end{aligned} \quad (2)$$

where  $\mathcal{S}$  and  $\mathcal{N}$  indicate channel-wise summation and channel L2 normalization, respectively. The compressed template is used to build various coarse matching templates according to respective strategies that are pre-defined based on certain objectives to exploit various temporal characteristics.

### 3.2 Deriving Temporal Diversity

As much as spatial diversity is effective to enrich the semantic information of the embedded features, utilizing temporal diversity is also important considering the nature of a video. When constructing the fine matching templates, we use the initial frame to exploit accurate information and the previous adjacent frame to leverage the most related information, similar to FEELVOS [34], CFBI [41], and BMVOS [6]. Those matching templates are defined as the global matching template and local matching template, respectively.

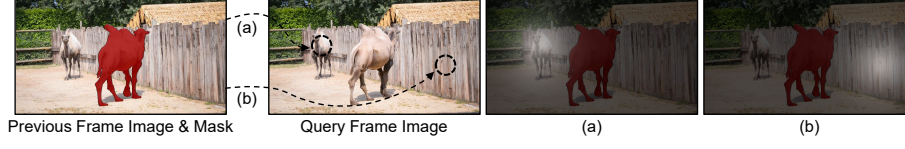
As opposed to the fine matching template whose elements contain their unique positional information, the features in the compressed template do not contain any positional information because it only has a single spatial location. Therefore, it is able to integrate multiple compressed templates of historic frames based on specific goals. Using this, we build temporally diversified coarse matching templates to fully exploit various time-specific properties of a video. We define the coarse matching templates as follows: 1) an overall matching template that memorizes the general features along an entire video; 2) a short-term matching template that is built by paying more attention to temporally-near frames; 3) a long-term matching template that is built by paying more attention to temporally-distant frames. Assuming frame  $i$  is the current frame being processed, the overall matching template  $O^i \in [-1, 1]^{C \times 1 \times 1}$ , short-term matching template  $S^i \in [-1, 1]^{C \times 1 \times 1}$ , and long-term matching template  $L^i \in [-1, 1]^{C \times 1 \times 1}$  are defined as follows.

$$\begin{aligned}
\mu_{O,BG} &= \Sigma_{k=0}^{i-1} \mathcal{S}(m_0^k) / \Sigma_{k=0}^i \mathcal{S}(m_0^k) \\
\mu_{O,FG} &= \Sigma_{k=0}^{i-1} \mathcal{S}(m_1^k) / \Sigma_{k=0}^i \mathcal{S}(m_1^k) \\
O_{BG}^i &= \mathcal{N}(\mu_{O,BG} O_{BG}^{i-1} + (1 - \mu_{O,BG}) C_{BG}^i) \\
O_{FG}^i &= \mathcal{N}(\mu_{O,FG} O_{FG}^{i-1} + (1 - \mu_{O,FG}) C_{FG}^i)
\end{aligned} \tag{3}$$

$$\begin{aligned}
S_{BG}^i &= \mathcal{N}(\mu_{S,BG} S_{BG}^{i-1} + (1 - \mu_{S,BG}) C_{BG}^i) \\
S_{FG}^i &= \mathcal{N}(\mu_{S,FG} S_{FG}^{i-1} + (1 - \mu_{S,FG}) C_{FG}^i)
\end{aligned} \tag{4}$$

$$\begin{aligned}
L_{BG}^i &= \mathcal{N}(\mu_{L,BG} L_{BG}^{i-1} + (1 - \mu_{L,BG}) C_{BG}^i) \\
L_{FG}^i &= \mathcal{N}(\mu_{L,FG} L_{FG}^{i-1} + (1 - \mu_{L,FG}) C_{FG}^i)
\end{aligned} \tag{5}$$

Considering that the compressed template is a channel normalized vector, taking the average over two vectors will generate a vector with a scale lower than 1, if two vectors have different directions. Therefore, the scale of the coarse matching templates will keep decreasing over time, which will lead to a scale-vanishing problem that engenders low confidence in the coarse matching. To prevent this, we apply channel re-normalization after every template update step to maintain the scale of the coarse matching templates at 1. The inertia values for short-term matching and long-term matching, i.e.,  $\mu_S$  and  $\mu_L$ , are



**Fig. 2.** Visualized information transfer flow for local matching when spatial distance scoring is applied. The information of each reference frame pixel is weighted by a distance score that indicates how close two pixels are spatially.

learnable parameters and trained during the network training stage. To press the values to be trained for respective objectives, we set the initial values to  $\frac{1}{1+\exp(1)}$  and  $\frac{1}{1+\exp(-1)}$ , respectively. At the initial frame, all  $\mu$  values are set to 0 because there is no previous frame.

### 3.3 Diversified Feature Similarity Matching

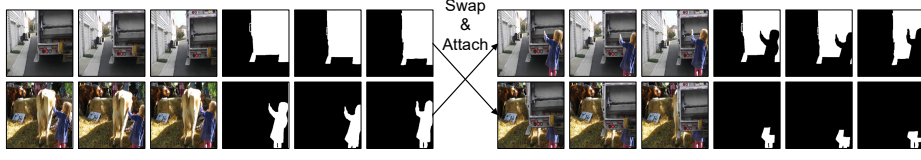
Assuming frame  $i$  is the query frame, the spatio-temporally diversified templates, i.e., global matching template  $F^0$ , local matching template  $F^{i-1}$ , overall matching template  $O^{i-1}$ , short-term matching template  $S^{i-1}$ , and long-term matching template  $L^{i-1}$ , are compared to the query frame embedded features  $X^i$ . Given that a feature similarity matching template has a size of  $C \times M \times N$ , taking a matrix inner product along the channel dimension with  $X^i$  will output a similarity map with the size of  $MN \times HW$ . By taking query-wise maximum operation and reshaping, a score map with the size of  $H \times W$  can be obtained for each template and class. As there are five different matching templates consisting of foreground and background, the output scores of feature similarity matching are defined as  $Z^i \in [-1, 1]^{10 \times H \times W}$ . These scores are fed into a decoder as an input, to provide visually derived information that can specify the target object.

### 3.4 Spatial Distance Scoring

As described before, the feature similarity matching operates in a non-local manner, which excludes the importance of temporal consistency of a video. However, adjacent frames in a video sequence are strongly dependent on each other, and therefore, are predominantly local. To exploit this locality of VOS, FEELVOS [34] and CFBI [41] restrict the search area by applying a square window for every spatial location when transferring the previous adjacent frame information to the query frame. Similarly, RMNet [38] reduces the search area under a hard square that is generated by warping the previous frame mask using a pre-trained optical flow model. These hard windowing methods are effective at capturing the temporal consistency of a video, but as the size of a window is a discrete value, it needs a human-tuning process, which makes the solution less elegant and complete.

To obtain temporal consistency automatically without any human-tuning process, we propose an end-to-end learnable distance-scoring function that takes





**Fig. 3.** Example video sequences before and after applying swap-and-attach data augmentation method.

a spatial distance between two pixels as its input and simply outputs a distance score from 0 to 1. When performing local matching with the previous frame, we first calculate a distance matrix that comprises of the spatial distance of every previous adjacent frame pixel location  $p$  and query frame pixel location  $q$  based on Euclidean distance as

$$d = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}, \quad (6)$$

where  $p_x$ ,  $p_y$ ,  $q_x$ , and  $q_y$  are  $x$ ,  $y$  coordinates of  $p$  and  $q$ . Then, by independently applying the distance-scoring function to every single element in the matrix, spatial distance score  $D$  can be obtained as

$$D = f(d; w) = \text{Sigmoid}(w_2 * \max(0, w_1 * d)) , \quad (7)$$

where  $f$  is the distance-scoring function and  $w$  indicates learnable parameters of  $f$ . By multiplying the spatial distance scores to feature similarity scores obtained from the local matching, the spatial distance as well as the feature distance can be reflected in the similarity matching scores. Figure 2 shows the information transfer flow when applying spatial distance scoring. As information transfer is adjusted by spatial distance scores between two pixels, temporal consistency between adjacent frames can be effectively considered.

### 3.5 Swap-and-Attach Augmentation

As annotating video data in a pixel level is extremely laborious, VOS data is not sufficient compared to other computer vision tasks. To increase the amount of VOS training data, we propose a swap-and-attach data augmentation method. In Figure 3, we show some example sequences when applying swap-and-attach augmentation. During the network training stage, each video sequence of a single batch exchanges its information with another video sequence of a different batch. The objects from multiple frames are jointly swapped between different sequences and attached on respective frames. As this process does not harm the temporal consistency of each video sequence, the generated video sequences are also realistic and have temporal connectivity between the consecutive frames. The swap-and-attach augmentation has two main advantages. First, as the amount of training data is dramatically increased, the model can be much more generalized. Second, as the objects may be severely occluded by the attached objects, the model can learn strong feature representations to force each object to have unique features.



### 3.6 Implementation Details

**Encoder.** We take the DenseNet-121 [13] architecture as our encoder. To initialize the model with rich feature representations, we use the ImageNet [17] pre-trained version. As the highest-level features extracted from original DenseNet-121 have low resolution, we skip the last block to use a higher feature resolution.

**Decoder.** Taking processed features from the encoder and cues from feature matching module, the decoder generates the final segmentation. To better exploit the locality of VOS,  $m^{i-1}$  is fed into the decoder when predicting frame  $i$ . As  $Z^i$  and  $m^{i-1}$  have a small spatial size compared to the required output size, low-level features extracted from the encoder are also fed into the middle of the decoder via skip connections. For fast decoding, a channel-reducing deconvolutional layer [43] is added after every skip connection as in CRVOS [5]. To obtain rich feature representations, we adopt CBAM [37] after every deconvolutional layer.

### 3.7 Network Training

Following most state-of-the-art methods [10, 21, 27, 31, 38], we simulate training samples using static image dataset, COCO [23], to obtain sufficient training data. After pre-training the network using static images, either the DAVIS 2017 [29] training set or YouTube-VOS 2018 [40] training set is used for the main training depending on the testing dataset. From a video sequence, we randomly crop  $384 \times 384$  patches from 10 consecutive frames jointly. If the first frame does not contain any object, the cropping is repeated until it contains at least one object. After that, a single designated object is defined as a foreground object and other objects are considered as background. Similar to CFBI [41], if the number of foreground pixels is not enough, that sequence is not used for network training to prevent the model from being biased to background attributes. We use cross-entropy loss and Adam optimizer [16] with a learning rate of  $1e-4$  without learning rate decay. During the network training, the backbone encoder is frozen and all batch normalization layers [14] are disabled. The swap-and-attach augmentation is applied with a probability of 20%.

## 4 Experiments

In this section, we first describe the datasets and evaluation metrics used in this study in Section 4.1. Each component of our proposed approach is validated in Section 4.2. Quantitative and qualitative comparison with state-of-the-art methods can be found in Section 4.3 and Section 4.4, respectively. Our method is abbreviated as TBD.

### 4.1 Experimental Setup

**Datasets.** The task of semi-supervised VOS is associated with two popular benchmark datasets, i.e., DAVIS [28, 29] and YouTube-VOS [40] datasets. DAVIS

**Table 1.** Ablation study on the proposed components. G, L, O, and T indicate the use of global, local, overall, and short-term & long-term matching templates, respectively. DS, HW, and Size denote spatial distance scoring, hard windowing, and window size for the hard windowing (the distance between the center point and side of a window). Aug indicates the use of swap-and-attach augmentation. The models are tested on the DAVIS 2017 validation set.

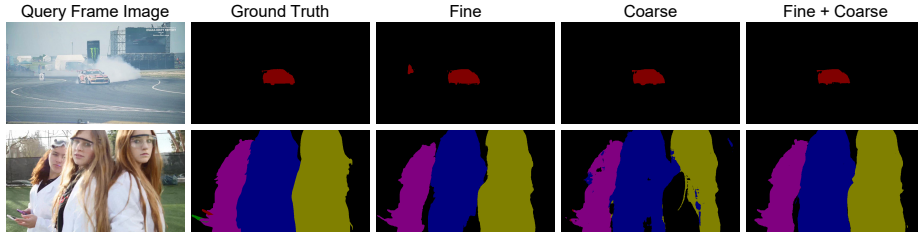
Version	G	L	O	T	DS	HW	Size	Aug	$\mathcal{G}_M$	$\mathcal{J}_M$	$\mathcal{F}_M$
I	✓	✓			✓		-	✓	77.2	75.3	79.1
II			✓	✓	✓		-	✓	72.7	70.2	75.2
III	✓	✓	✓		✓		-	✓	79.1	76.7	81.6
IV	✓	✓		✓	✓		-	✓	78.7	76.7	80.7
V	✓	✓	✓	✓			-	✓	78.8	76.7	80.9
VI	✓	✓	✓	✓		✓	1	✓	79.2	76.9	81.5
VII	✓	✓	✓	✓		✓	2	✓	79.3	76.9	81.8
VIII	✓	✓	✓	✓		✓	4	✓	79.9	77.4	82.5
IX	✓	✓	✓	✓		✓	8	✓	78.5	76.3	80.7
X	✓	✓	✓	✓	✓		-		78.6	76.1	81.2
XI	✓	✓	✓	✓	✓		-	✓	80.0	77.6	82.3

2016 consists of 30 videos in the training set and 20 videos in the validation set. DAVIS 2017 comprises 60 videos in the training set, 30 videos in the validation set, and 30 videos in the test-dev set. Each video is 24 fps and all frames are annotated. YouTube-VOS 2018 is the largest dataset for VOS, containing 3,471 videos in the training set and 474 videos in the validation set. All videos are 30 fps, annotated every five frames.

**Evaluation metrics.** We use the standard evaluation metrics for VOS. The predicted segmentation masks with hard labels are evaluated using region accuracy and contour accuracy. Region accuracy  $\mathcal{J}$  can be obtained by computing the number of pixels in the intersection between the predicted segmentation mask and the ground truth segmentation mask and then dividing it by the size of their union. Contour accuracy  $\mathcal{F}$  can be obtained using the same process but is evaluated only on the object boundaries. The overall accuracy  $\mathcal{G}$  is the average of the  $\mathcal{J}$  and  $\mathcal{F}$  measures.

## 4.2 Analysis

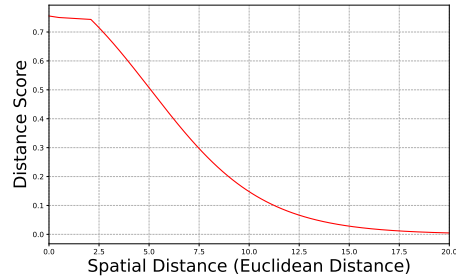
**Spatio-temporal diversity of template.** To capture the local keypoints while perceiving the global tendencies, we employ a fine matching template and coarse matching template simultaneously. In addition, we build each template using multiple sub-templates to exploit various temporal properties. In Table 1, various model versions with different template construction schemes are compared. As can be seen, a joint use of fine matching and coarse matching significantly improves the performance compared to when each one is used individually. The use of diverse temporal properties is also quantitatively validated to be effective. In



**Fig. 4.** Qualitative comparison of various model versions with different spatial template construction schemes. Fine and coarse indicate the use of fine matching and coarse matching, respectively.

Figure 4, we compare the results of various model versions with different spatial diversities of a feature matching template. As can be seen, fine matching is effective for capturing local details (second sequence), but thereafter, is susceptible to visual distractions as it misses the global tendencies (first sequence). By contrast, coarse matching can capture the general features for an object (first sequence), but is not able to capture the details of the objects (second sequence). By using fine and coarse matching methods simultaneously, their respective advantages can be effectively leveraged.

**Spatial distance scoring.** In order to capture the temporal consistency between consecutive frames in a video without a human-tuning process, we present a novel spatial distance scoring method by learning a distance-scoring function that takes a spatial distance between two pixels as its input and outputs a naive distance score. According to Table 1, our proposed spatial distance scoring method is more effective than hard windowing method used in FEELVOS [34] and CFBI [41], even without the need for manual tuning. We also visualize a fully-trained distance-scoring function to confirm that it really works. As can be seen in Figure 5, the fully-trained distance-scoring function outputs a high distance score when two pixels are spatially near, and vice versa, which accords with the expected behavior.



**Fig. 5.** Visualized input and output of a fully-trained distance-scoring function. Spatial distance indicates the Euclidean distance between two pixels after feature embedding.

**Swap-and-attach augmentation.** Our proposed swap-and-attach augmentation is also demonstrated to be effective through a quantitative evaluation in Table 1. Simply swapping and attaching the objects of different batches brings 1.4% improvement on a  $\mathcal{G}$  score, as it increases the amount of training data and provides hard training samples that contain some severe occlusions.

**Table 2.** Quantitative evaluation on the DAVIS datasets. OL denotes online learning. (+S) denotes the use of static image datasets during the network training.

Method	OL	fps	2016 val			2017 val			2017 test-dev		
			$\mathcal{G}_M$	$\mathcal{J}_M$	$\mathcal{F}_M$	$\mathcal{G}_M$	$\mathcal{J}_M$	$\mathcal{F}_M$	$\mathcal{G}_M$	$\mathcal{J}_M$	$\mathcal{F}_M$
STCNN (+S) [39]		0.26	83.8	83.8	83.8	61.7	58.7	64.6	-	-	-
FEELVOS (+S) [34]		2.22	81.7	80.3	83.1	69.1	65.9	72.3	54.4	51.2	57.5
DMM-Net (+S) [44]		-	-	-	-	70.7	68.1	73.3	-	-	-
AGSS-VOS (+S) [22]		10.0	-	-	-	67.4	64.9	69.9	57.2	54.8	59.7
RANet (+S) [36]		30.3	85.5	85.5	85.4	65.7	63.2	68.2	55.3	53.4	57.2
DTN (+S) [45]		14.3	83.6	83.7	83.5	67.4	64.2	70.6	-	-	-
STM (+S) [27]		6.25	86.5	84.8	<u>88.1</u>	71.6	69.2	74.0	-	-	-
DIPNet (+S) [11]		0.92	86.1	85.8	86.4	68.5	65.3	71.6	55.2	-	-
LWL (+S) [1]	✓	14.0	-	-	-	74.3	72.2	76.3	-	-	-
CFBI (+S) [41]		5.56	86.1	85.3	86.9	74.9	72.1	77.7	-	-	-
GC (+S) [19]		25.0	86.6	<u>87.6</u>	85.7	71.4	69.3	73.5	-	-	-
KMN (+S) [31]		8.33	<u>87.6</u>	<u>87.1</u>	<u>88.1</u>	76.0	74.2	77.8	-	-	-
AFB-URR (+S) [21]		4.00	-	-	-	74.6	73.0	76.1	-	-	-
STG-Net (+S) [24]		-	85.7	85.4	86.0	74.7	71.5	77.9	<u>63.1</u>	<u>59.7</u>	<u>66.5</u>
RMNet (+S) [38]		11.9	81.5	80.6	82.3	75.0	72.8	77.2	-	-	-
LCM (+S) [10]		8.47	-	-	-	75.2	73.1	77.2	-	-	-
SSTVOS (+S) [8]		-	-	-	-	78.4	75.4	81.4	-	-	-
HMMN (+S) [32]		10.0	<b>89.4</b>	<b>88.2</b>	<b>90.6</b>	<b>80.4</b>	<b>77.7</b>	<b>83.1</b>	-	-	-
<b>TBD (+S)</b>		<b>50.1</b>	86.8	87.5	86.2	<u>80.0</u>	<u>77.6</u>	<u>82.3</u>	<b>69.4</b>	<b>66.6</b>	<b>72.2</b>
AGSS-VOS [22]		10.0	-	-	-	66.6	63.4	69.8	54.3	51.5	57.1
RANet [36]		30.3	-	73.2	-	-	-	-	-	-	-
STM [27]		6.25	-	-	-	43.0	38.1	47.9	-	-	-
FRTM [30]	✓	21.9	81.7	-	-	68.8	66.4	71.2	-	-	-
TVOS [46]		37.0	-	-	-	72.3	69.9	74.7	<u>63.1</u>	58.8	<u>67.4</u>
JOINT [26]	✓	4.00	-	-	-	<b>78.6</b>	<b>76.0</b>	<b>81.2</b>	-	-	-
BMVOS [6]		<u>45.9</u>	<u>82.2</u>	<u>82.9</u>	<u>81.4</u>	72.7	70.7	74.7	62.7	<u>60.7</u>	64.7
<b>TBD</b>		<b>50.1</b>	<b>84.3</b>	<b>85.2</b>	<b>83.4</b>	<u>75.2</u>	<u>73.2</u>	<u>77.2</u>	<b>66.0</b>	<b>63.1</b>	<b>68.9</b>

### 4.3 Quantitative Results

**DAVIS.** We present a quantitative comparison of our proposed method with state-of-the-art methods on the DAVIS 2016 [28] and DAVIS 2017 [29] datasets in Table 2. For all methods, inference speed is calculated on the DAVIS 2016 validation set with the 480p resolution. TBD shows competitive performance on all datasets, even maintaining a real-time inference speed. Compared to the real-time methods faster than 24 fps (DAVIS default setting), it outperforms all existing methods by a significant margin.

**YouTube-VOS.** In Table 3, our method is compared to state-of-the-art methods on the YouTube-VOS 2018 validation set. Gen. indicates the generalization gap proposed in MAST [18]. In order to capture small objects, we use the 720p resolution if a target object contains less than 1,000 pixels; if not, we use the

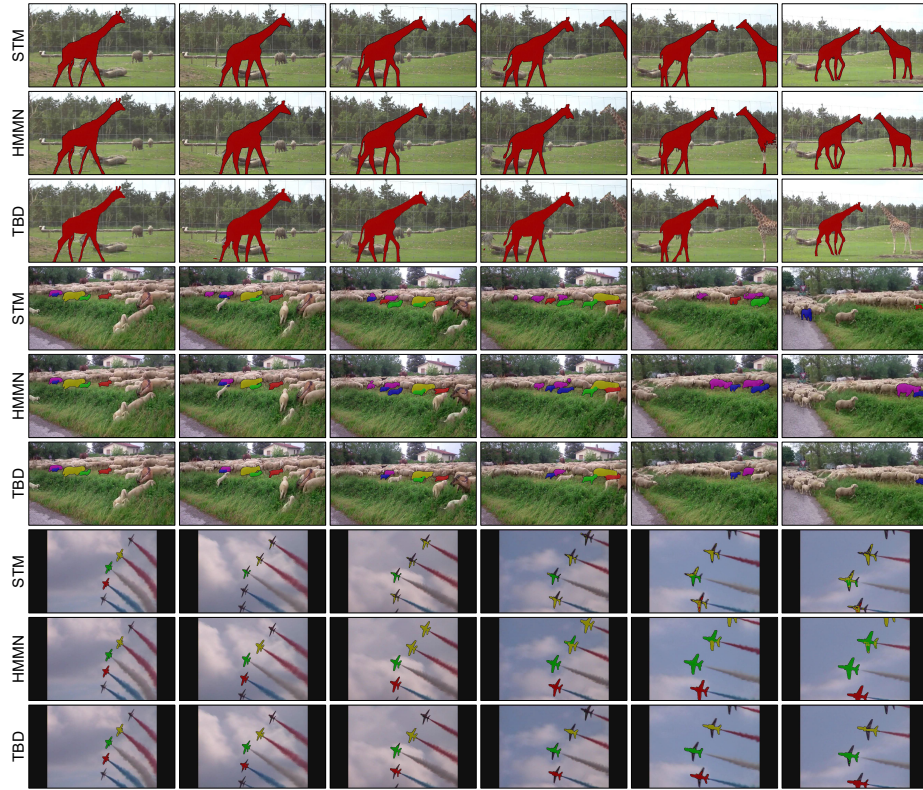
**Table 3.** Quantitative evaluation on the YouTube-VOS 2018 validation set. OL denotes online learning and (+S) denotes the use of static image datasets during the network training. Gen. denotes the generalization gap between seen and unseen categories.

Method	OL	fps	$\mathcal{G}_M$	$\mathcal{J}_S$	$\mathcal{J}_U$	$\mathcal{F}_S$	$\mathcal{F}_U$	Gen.
DMM-Net (+S) [44]		12.0	51.7	58.3	41.6	60.7	46.3	15.6
STM (+S) [27]		-	79.4	79.7	72.8	84.2	80.9	5.10
SAT (+S) [2]		39.0	63.6	67.1	55.3	70.2	61.7	10.2
LWL (+S) [1]	✓	-	81.5	80.4	76.4	84.9	84.4	2.25
EGMN (+S) [25]		-	80.2	80.7	74.0	85.1	80.9	5.45
CFBI (+S) [41]		-	81.4	81.1	75.3	85.8	83.4	4.10
GC (+S) [19]		-	73.2	72.6	68.9	75.6	75.7	<u>1.80</u>
KMN (+S) [31]		-	81.4	81.4	75.3	85.6	83.3	4.20
AFB-URR (+S) [21]		-	79.6	78.8	74.1	83.1	82.6	2.60
STG-Net (+S) [24]		6.00	73.0	72.7	69.1	75.2	74.9	1.95
RMNet (+S) [38]		-	81.5	82.1	75.7	85.7	82.4	4.85
LCM (+S) [10]		-	82.0	<u>82.2</u>	75.7	86.7	83.4	4.90
GIEL (+S) [9]		-	80.6	80.7	75.0	85.0	81.9	4.40
SwiftNet (+S) [35]		-	77.8	77.8	72.3	81.8	79.5	3.90
SSTVOS (+S) [8]		-	81.7	81.2	76.0	-	-	-
JOINT (+S) [26]	✓	-	<u>83.1</u>	81.5	<b>78.7</b>	85.9	<b>86.5</b>	<b>1.10</b>
HMMN (+S) [32]		-	82.6	82.1	76.8	<u>87.0</u>	84.6	3.85
AOT-T (+S) [42]		<u>41.0</u>	80.2	80.1	74.0	84.5	82.2	4.20
SwinB-AOT-L (+S) [42]		9.30	<b>84.5</b>	<b>84.3</b>	<u>77.9</u>	<b>89.3</b>	<u>86.4</u>	4.65
STCN (+S) [4]		-	83.0	81.9	<u>77.9</u>	86.5	85.7	2.40
TBD (+S)		30.4	80.5	79.4	75.5	83.8	83.2	2.25
RVOS [33]		22.7	56.8	63.6	45.5	67.2	51.0	17.2
A-GAME [15]		-	66.1	67.8	60.8	-	-	-
AGSS-VOS [22]		12.5	71.3	71.3	65.5	76.2	73.1	4.45
CapsuleVOS [7]		13.5	62.3	67.3	53.7	68.1	59.9	10.9
STM [27]		-	68.2	-	-	-	-	-
FRTM [30]	✓	-	72.1	72.3	65.9	76.2	74.1	4.25
TVOS [46]		37.0	67.8	67.1	63.0	69.4	71.6	<b>0.95</b>
STM-cycle [20]		<b>43.0</b>	69.9	71.7	61.4	75.8	70.4	7.85
BMVOS [6]		28.0	<u>73.9</u>	<u>73.5</u>	<u>68.5</u>	<u>77.4</u>	<u>76.0</u>	3.20
TBD		30.4	<b>77.8</b>	<b>77.4</b>	<b>72.7</b>	<b>81.2</b>	<b>79.9</b>	<u>3.00</u>

480p resolution. The experimental results indicate TBD is comparable to other competitive methods, while showing a fast inference speed of 30.4 fps.

#### 4.4 Qualitative Results

In Figure 6, we qualitatively compare our method to state-of-the-art STM [27] and HMMN [32]. As can be seen, TBD predicts more stable outputs compared to other methods, thanks to its diverse feature similarity matching templates and spatial distance scoring. Exceptional temporal consistency and robustness against occlusion can also be observed.



**Fig. 6.** Qualitative evaluation of TBD obtained by comparing it to other state-of-the-art methods.

## 5 Conclusion

We introduce a spatio-temporally diversified template construction scheme for robust and stable VOS. In addition, a spatial distance scoring method with a learnable distance-scoring function and a novel swap-and-attach augmentation method are proposed. On public benchmark datasets, our approach achieves a comparable performance to the current state-of-the-art solutions while maintaining a real-time inference speed. We believe our approach will be widely used for future VOS research and various fields in computer vision.

**Acknowledgement.** This research was supported by R&D program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of KOREA (NRF) funded by Ministry of Science and ICT (NRF-2018M3E3A1057289) and the KIST Institutional Program (Project No.2E31051-21-203).

## References

1. Bhat, G., Lawin, F.J., Danelljan, M., Robinson, A., Felsberg, M., Van Gool, L., Timofte, R.: Learning what to learn for video object segmentation. In: European Conference on Computer Vision. pp. 777–794. Springer (2020)
2. Chen, X., Li, Z., Yuan, Y., Yu, G., Shen, J., Qi, D.: State-aware tracker for real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9384–9393 (2020)
3. Chen, Y., Pont-Tuset, J., Montes, A., Van Gool, L.: Blazingly fast video object segmentation with pixel-wise metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1189–1198 (2018)
4. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems* **34** (2021)
5. Cho, S., Cho, M., Chung, T.y., Lee, H., Lee, S.: Crvos: Clue refining network for video object segmentation. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 2301–2305. IEEE (2020)
6. Cho, S., Lee, H., Kim, M., Jang, S., Lee, S.: Pixel-level bijective matching for video object segmentation. *arXiv preprint arXiv:2110.01644* (2021)
7. Duarte, K., Rawat, Y.S., Shah, M.: Capsulevos: Semi-supervised video object segmentation using capsule routing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8480–8489 (2019)
8. Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: Sstvos: Sparse spatiotemporal transformers for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5912–5921 (2021)
9. Ge, W., Lu, X., Shen, J.: Video object segmentation using global and instance embedding learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16836–16845 (2021)
10. Hu, L., Zhang, P., Zhang, B., Pan, P., Xu, Y., Jin, R.: Learning position and target consistency for memory-based video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4144–4154 (2021)
11. Hu, P., Liu, J., Wang, G., Ablavsky, V., Saenko, K., Sclaroff, S.: Dipnet: Dynamic identity propagation network for video object segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1904–1913 (2020)
12. Hu, Y.T., Huang, J.B., Schwing, A.G.: Videomatch: Matching based video object segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 54–70 (2018)
13. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
14. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015)
15. Johnander, J., Danelljan, M., Brissman, E., Khan, F.S., Felsberg, M.: A generative appearance model for end-to-end video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8953–8962 (2019)



16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60**(6), 84–90 (2017)
18. Lai, Z., Lu, E., Xie, W.: Mast: A memory-augmented self-supervised tracker. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6479–6488 (2020)
19. Li, Y., Shen, Z., Shan, Y.: Fast video object segmentation using the global context module. In: *European Conference on Computer Vision*. pp. 735–750. Springer (2020)
20. Li, Y., Xu, N., Peng, J., See, J., Lin, W.: Delving into the cyclic mechanism in semi-supervised video object segmentation. arXiv preprint arXiv:2010.12176 (2020)
21. Liang, Y., Li, X., Jafari, N., Chen, Q.: Video object segmentation with adaptive feature bank and uncertain-region refinement. arXiv preprint arXiv:2010.07958 (2020)
22. Lin, H., Qi, X., Jia, J.: Agss-vos: Attention guided single-shot video object segmentation. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 3949–3957 (2019)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *European conference on computer vision*. pp. 740–755. Springer (2014)
24. Liu, D., Xu, S., Liu, X.Y., Xu, Z., Wei, W., Zhou, P.: Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, pp. 2100–2108 (2021)
25. Lu, X., Wang, W., Danelljan, M., Zhou, T., Shen, J., Van Gool, L.: Video object segmentation with episodic graph memory networks. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III* 16. pp. 661–679. Springer (2020)
26. Mao, Y., Wang, N., Zhou, W., Li, H.: Joint inductive and transductive learning for video object segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9670–9679 (2021)
27. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 9226–9235 (2019)
28. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 724–732 (2016)
29. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)
30. Robinson, A., Lawin, F.J., Danelljan, M., Khan, F.S., Felsberg, M.: Learning fast and robust target models for video object segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7406–7415 (2020)
31. Seong, H., Hyun, J., Kim, E.: Kernelized memory network for video object segmentation. In: *European Conference on Computer Vision*. pp. 629–645. Springer (2020)
32. Seong, H., Oh, S.W., Lee, J.Y., Lee, S., Lee, S., Kim, E.: Hierarchical memory matching network for video object segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 12889–12898 (2021)

33. Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F., Giro-i Nieto, X.: Rvos: End-to-end recurrent network for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5277–5286 (2019)
34. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.C.: Feelvos: Fast end-to-end embedding learning for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9481–9490 (2019)
35. Wang, H., Jiang, X., Ren, H., Hu, Y., Bai, S.: Swiftnet: Real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1296–1305 (2021)
36. Wang, Z., Xu, J., Liu, L., Zhu, F., Shao, L.: Ranet: Ranking attention network for fast video object segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 3978–3987 (2019)
37. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
38. Xie, H., Yao, H., Zhou, S., Zhang, S., Sun, W.: Efficient regional memory network for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1286–1295 (2021)
39. Xu, K., Wen, L., Li, G., Bo, L., Huang, Q.: Spatiotemporal cnn for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1379–1388 (2019)
40. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
41. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by foreground-background integration. arXiv preprint arXiv:2003.08333 (2020)
42. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems* **34** (2021)
43. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: 2011 International Conference on Computer Vision. pp. 2018–2025. IEEE (2011)
44. Zeng, X., Liao, R., Gu, L., Xiong, Y., Fidler, S., Urtasun, R.: Dmm-net: Differentiable mask-matching network for video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3929–3938 (2019)
45. Zhang, L., Lin, Z., Zhang, J., Lu, H., He, Y.: Fast video object segmentation via dynamic targeting network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5582–5591 (2019)
46. Zhang, Y., Wu, Z., Peng, H., Lin, S.: A transductive approach for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6949–6958 (2020)