Video Object Segmentation via Structural Feature Reconfiguration

Zhenyu Chen[1], Ping Hu[2], Lu Zhang[1], Huchuan Lu[1], You He[3]*, Shuo Wang[4], Xiaoxing Zhang[4], Maodi Hu[4], and Tao Li[4]

[1] Dalian University of Technology, China
dlutczy@mail.dlut.edu.cn luzhangdut@gmail.com lhchuan@dlut.edu.cn
[2] Boston University, USA
[3] Naval Aeronautical University, China
pinghu@bu.edu youhe_nau@163.com
[4] Meituan, China
{wangshuo28, zhangxiaoxing, humaodi, litao19}@meituan.com

Abstract. Recent memory-based methods have made significant progress for semi-supervised video object segmentation, by explicitly modeling the semantic correspondences between the target frame and the historical ones. However, the indiscriminate acceptance of historical frames into the memory bank and the lack of fine-grained extraction for target objects may incur high latency and information redundancy in these approaches. In this paper, we circumvent the challenges by developing a Structural Feature Reconfiguration Network (SFRNet) . The proposed SFRNet consists of two core sub-modules, which are Global-temporal Attention Module (GAM) and Local-spatial Attention Module (LAM). In GAM, we exploit self-attention-based encoders to capture the target objects' temporal context from historical frames. The LAM then reconfigures features with the current frame's spatial structural prior, which reinforces the objectness of foreground objects and suppresses the interference from background regions. By doing so, our model reduces the reliance on the large memory bank containing redundant historical frames, while instead effectively segmenting video objects with spatio-temporal context aggregated from a small set of key frames. We conduct extensive experiments with benchmark datasets, and the results demonstrate our method's favorable performance against the state-of-the-art approaches. The code will be available at https://github.com/zy5037/SFRNet.

Keywords: Video Object Segmentation · Structural Feature Reconfiguration · Global-temporal Attention · Local-spatial Attention.

1 Introduction

Video Object Segmentation (VOS) aims to segment out the interested objects along the video sequence. It has received great attention recently because of its

[*] Corresponding author

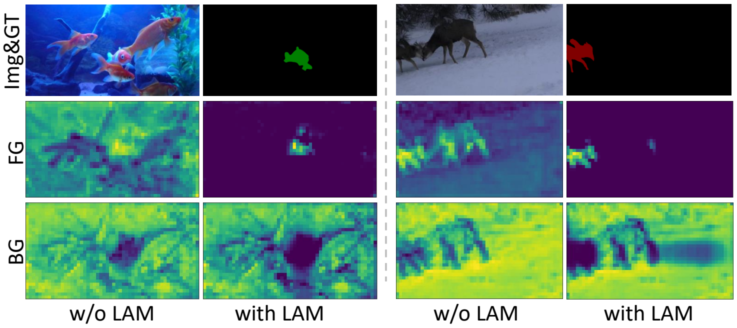

Fig. 1. Visualization of key feature channels. We observe that a target frame is structured as foreground (FG) and background (BG) regions by different feature channels. Our Local-spatial Attention Module (LAM) exploits the current frame's spatial structure to better extract discriminative structural feature representations, hence alleviating the demand for heavy historical memory.

benefits for applications like video surveillance, video editing, and multimedia analysis. In this paper, we focus on addressing semi-supervised video object segmentation, where the target objects are manually annotated in the first frame.

Thanks to the recent advances in deep learning techniques, state-of-the-art methods in VOS have achieved significant progress. Early methods typically propagate object masks over time via motion cues like optical flow [12,21–23,42,59] or adopting an online learning strategy [5,20,24,27,45,49] to finetune on the first frame with annotations. However, the motion-based mask propagation may accumulate errors and online finetuning suffers from very low efficiency. Recently, matching-based methods have emerged as a promising solution for this task [8,30,44,58,60]. Among these approaches, Space-Time Memory (STM) network [33] achieves great success, by extracting the spatio-temporal context from a memory bank, which is typically large and redundant to ensure effectiveness. In order to optimize the memory efficiency, several follow-ups of STM [11,37,38,46,50] have been proposed with improved encoding [38,46] and matching [11,37,50] process. Though significant progress toward this direction has been made, maintaining memory with both efficiency and effectiveness is still very challenging, due to the difficulties in balancing between the memory capacity and quality.

In this work, we circumvent the challenge by proposing a Structural Feature Reconfiguration Network (SFRNet), which alleviates the reliance on large memory banks by exploiting the spatial structural composition of testing frame. Given a video frame, we aim to discriminate foreground objects and background regions as different pixel sets that are spatially structured by their underlying semantic coherence. Therefore, the segmentation of video objects should not only benefit from pixel-level space-time correlations for referred object, but also a video frame's own spatial structural compositions as illustrated in Fig. 1. Based on this, we design SFRNet with two core components including a Global-temporal Attention Module (GAM) and a Local-spatial Attention Module (LAM).

The GAM extracts pixel-level spatio-temporal context from historical key frames with a Transformer-based architecture. And the LAM is adopted to further enhance the feature representations with spatial composition priors of the testing frame. To explicitly extract and represent a target frame's spatial composition, we draw inspiration from image subspace composition [9,40], where the deep model is trained to encode visual components in images as discriminative low-rank tensors [4,41,62]. Specifically, in LAM we explicitly construct low-rank feature maps by first collecting the contextual feature basis along the spatial dimensions of the feature maps. These basis are then aggregated via Kronecker Product to form a set of low-rank tensors, and finally combined with the input feature maps to represent different semantic components of the images. With end-to-end optimization, LAM is able to separate and encode visual information at object/region-level as illustrated in Fig. 1, hence achieving robustness for the quality of the temporal context aggregation in GAM. By combining GAM and LAM, our framework avoids heavy overheads caused by maintaining a large amount of memory, while achieving effectiveness in encoding and extracting visual objects in videos. Extensive experiments are performed to analyze the proposed method, and the results on multiple datasets [34,35,53] show that our proposed SFRNet can effectively segment video objects.

In summary, we have the following contributions:

- We propose a Local-spatial Attention Module that characterizes deep features with the structural composition prior to better extract video objects from the background.
- We develop a Structural Feature Reconfiguration Network (SFRNet), which utilizes space-time context aggregation and spatial structural composition of target objects, to relieve the dependency on heavily accumulated historical frames, and achieves effective video object segmentation.
- We conduct extensive experiments to demonstrate the effectiveness of the proposed method. Our SFRNet achieves the favorable performance against the state-of-the-art approaches on multiple datasets including *DAVIS16*, *DAVIS17*, and *YouTube-VOS*.

2 Related Work

Video Object segmentation. Learning video object segmentation via deep neural networks receives growing attention recently. To improve the model's generalization, early methods [5, 24, 27, 45, 49] usually rely on an online learning scheme that finetunes the deep model with the annotated first frame during testing. Despite the improvement in accuracy, the online finetuning process is quite time-consuming and hard to be applied in real-world applications. In recent years, with the advances in dense prediction tasks, STM [33] and CFBI [58] propose to build robust space-time correspondence modules and show great breakthroughs in performance against previous online approaches. They thus become the new baselines in VOS for further promotion on the accuracy [15,16,51] or efficiency [11,46]. The matching-based methods [15,44] like CFBI [58] usually build a multi-context feature matching mechanism between the query frame and key frames (typically the first frame and the recent frame) to encode the long-range similarity in semantics and the short-range similarity in appearance.

On the contrary, memory-based methods [11,19,28,46,51] like STM [33] aim to learn the pixel-wise space-time correspondence between the current frame and the historical ones. STM [33] introduces a memory mechanism, which resorts the non-local module to model the correspondence between query frame and memory frames. Through non-local module, the long-range dependencies among different frames can be established. However, simply including all the previous frames into the memory bank without selection would lead to memory explosion and a heavy computational burden. This motivates some recent attempts to improve the memory encoding strategy in STM. For example, SwiftNet [46] and AFB-URR [28] propose to further filter the redundant pixels and merge the similar ones in memory storage to alleviate the memory growth issue. RDE-VOS [25] instead limits the memory to a constant size to improve the model efficiency. RPCMVOS [54] suppress error propagation through a correction mechanism to avoid error accumulation. STCN [11] proposes to simplify the non-local calculation by replacing the original cosine distance with the L2 distance.

HMMN [38] builds a pyramid memory network where the multi-level features are incorporated to capture robust spatio-temporal correspondence. In this paper, we propose an effective memory encoding framework, in which the robust space-time coherence can be built on the key frames.

Transformer in Videos. Transformer [43] was originally proposed as a sequence-to-sequence model for machine translation and has become the mainstream baseline for natural language processing. Recently, it has been successfully applied to many computer vision tasks [6, 14, 29, 48, 61] and shown convincing performance w.r.t convolutional neural networks. Inspired by this, many attempts are made to explore the effectiveness of Transformer in video tasks. For example, TimeSformer [2] and STARK [55] extends the original Transformer to establish the spatio-temporal self-attention in video sequences. TransT [7] introduces Transformer to enhance the intra-correlation and inter-correlation, respectively. Recently, some Transformer-based models are proposed to tackle the VOS task. JOINT [32] incorporates the Transformer with an update mechanism for integrating transductive and inductive information into a unified framework. In this paper, we propose to build the Global-temporal Attention Module and Local-spatial Attention Module based on Transformer, for enhancing the structural representation of the target objects to achieve robust segmentation.

3 Method

We propose a Structural Feature Reconfiguration Network (SFRNet) for effective and efficient video object segmentation by enhancing the spatial structural representation of the referred objects. Given the ground-truth mask at the first frame, the proposed network aims to predict the masks in the following video sequence.

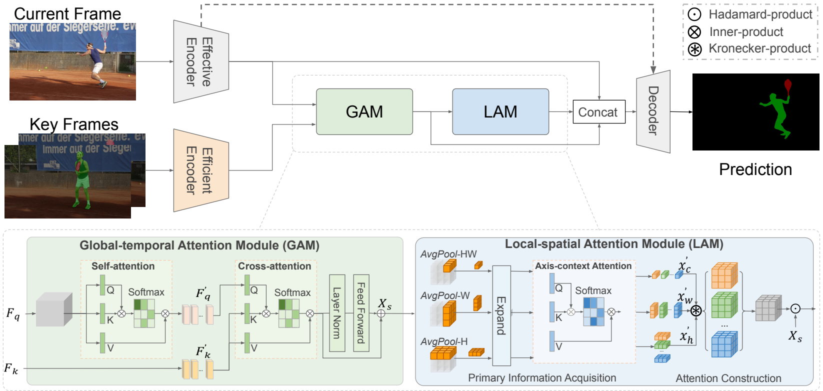

Fig. 2. An overview of our framework. Our network consists of an Effective Encoder and an Efficient Encoder for the feature extraction of query frame and key frames, respectively. The Global-temporal Attention Module (GAM) is used to construct the space-time correspondence at pixel level. The Local-spatial Attention Module (LAM) is used to enhance the structural composition of the target instances. Finally, the output of the two modules together with the skip-connections from effective encoder are fed to the decoder for mask generation.

The framework is shown in Fig. 2. We use two separate encoders to capture the embeddings for the current input I_t and the key frames I_k, respectively. To fix the memory storage issue, we use the first frame and a recent frame to form the key frames, *i.e.*, $I_k = \{I_1, I_r\}$. Following [11,46], we implement an Effective Encoder based on ResNet50 [17] to extract the features with rich semantics and spatial details for the current frame. Meanwhile, an Efficient Encoder is built on ResNet18 [17] to swiftly incorporate the memory embeddings for key frames. To capture historical information of the target objects, the concatenation of RGB images and predicted masks are fed to the efficient encoder as in [33].

With the extracted features of the current frame $F_q \in \mathbb{R}^{H\times W\times C}$ and key frames $F_k \in \mathbb{R}^{2\times H\times W\times C}$, SFRNet achieves memory encoding and video object segmentation via a Global-temporal Attention Module (GAM) and a Local-spatial Attention Module (LAM). The GAM is proposed to extract information for target object from the key-frame set, and a Transformer based attention formulation is designed to capture the globally long-term correlation for pixels across frames. To improve the robustness of captured features, we build LAM to further enhance with the referred objects' spatial structural composition, which strengthens the objectness of foreground objects in the current frame and suppresses the interference from background noise. The structural enhanced features, together with GAM output and skip-connections, are fed to the decoder model for mask generation.

We adopt a commonly used decoder architecture as [11,33,38], which stacks several refinement modules to incorporate the skip-connections from Effective Encoder.

3.1 Global-temporal Attention Module

To implement GAM, we leverage the advanced Transformer [43], which shows superior capability to model the global and long-range context in dense prediction tasks [6,61]. We start by briefly reviewing the multi-head attention module in Transformer, which is the core unit of the proposed GAM. Given the spatially flattened d_m-channel feature map $X \in \mathbb{R}^{N \times d_m}$, the attention formulation of each head is as follows:

$$h_i = Softmax(\frac{Q_i K_i^{\top}}{\sqrt{d_k}}) V_i \ . \tag{1}$$

where $Q_i \in \mathbb{R}^{N \times d_n}$, $K_i \in \mathbb{R}^{N \times d_n}$, and $V_i \in \mathbb{R}^{N \times d_n}$ are the *Query*, *Key*, and *Value* vectors respectively converted from the input vector X via several linear layers, d_k is a scaling factor equal to the output channel number d_n. The attention formulation in Eq. 1 can be performed in the form of multiple heads to produce richer representations.

$$MultiHead(Q, K, V) = Concat(h_1, ..., h_L) \cdot W^o. \tag{2}$$

where $Concat(\cdot)$ denotes the concatenation along the channel dimension, L is the number of heads, and $W^o \in \mathbb{R}^{(L \cdot d_n) \times d_m}$ is a linear layer that fuses the output vectors of the multi-head attention.

With the above-mentioned multi-head attention module, we build the Global-temporal Attention Module as in the left part of Fig. 2. Given the feature maps F_q of the current frame (also known as query frame), we first calculate the self-attention to exploit the spatial correlation. Specifically, we flatten the feature map $F_q \in \mathbb{R}^{H \times W \times C}$ over spatial dimensions and get a set of vectors $F'_q \in \mathbb{R}^{N \times C}$, with $N = H \times W$. The flattened vectors are then transferred to *Query*, *Key*, and *Value* vectors to formulate the multi-head attention as described in Eq. 1 and 2.

With the aggregated feature from the current frame, we then construct cross-context attention to capture target objects' temporal context from key frames. The cross-context attention is also based on multi-head attention, yet implemented on the enhanced current feature F'_q and the key frames' features F_k.

To better extract spatio-temporal context from key frames, we construct multiple stages of attention to aggregate information in an aggressive way. In detail, the flattened F'_q as *Query* and flattened F'_k are taken as *Key* and *Value* in cross-context attention of the first layer. In subsequent stages, the outputs of previous layer are sent to the self-attention for feature aggregation. The cross-context attention of subsequent stages takes the aggregated feature as *Query* and transform F'_k as *Key* and *Value* and finally generates the outputs of GAM, which are denoted as X_s.

3.2 Local-spatial Attention Module

Many efforts have been made to improve the efficiency of memory encoding and matching [11,28,46]. Yet it is still very challenging to balance between the efficiency and capacity, *i.e.*, a large memory contains redundant information incurring heavy computational cost, and a small memory bank benefits computational efficiency yet may lack sufficient information for extracting details of video objects. We argue that this challenge can be circumvented by exploiting the target objects structural composition in spatial domain of the current testing frame, which helps to reduce model's reliance on the quality of historical information, and thus avoiding large-size memory while keeping effectiveness in representations. We propose to learn to model and exploit the spatial structure of visual components in the the current frame. Inspired by the tensor low-rank decomposition of deep features [9,40], we recompose the feature maps via a set of low-rank components, which condense the key structural components for the objects and background regions. To this end, we build a Local-spatial Attention Module to characterize the output of cross-context features by GAM with region-level structure prior. The framework of LAM is shown in the right part of Fig. 2, which consists of two sub-modules: Primary Information Acquisition and Attention Construct Module.

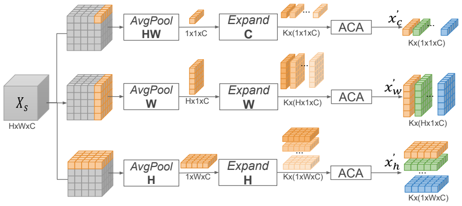

Fig. 3. Illustration of the Primary Information Acquisition (PIA) module. After pooling and expanding, we obtain the structural features in different spatial dimensions. "ACA" denotes the Axis-context Attention as described in Sec. 3.2.

Primary Information Acquisition (PIA). The goal of PIA is to learn to extract the discriminative basis vectors for the structural subspaces of the current frame. The architecture of PIA is shown in Fig. 3. Taking the cross-context feature $X_s \in \mathbb{R}^{H \times W \times C}$ from GAM as input, the PIA module first generates a set of spatial basis, which are later utilized to generate low-rank discriminative components. Specifically, we implement the average pooling on X_s along the *Height*, *Width*, and the full spatial dimensions to extract the contextual information of the target object and the current frame,

$$\begin{aligned} x_h &= AvgPool_H(X_s), \\ x_w &= AvgPool_W(X_s), \\ x_c &= AvgPool_{HW}(X_s) \ . \end{aligned} \tag{3}$$

where $AvgPool_H(\cdot)$, $AvgPool_W(\cdot)$ and $AvgPool_{HW}(\cdot)$ indicate the average pooling operation along height, width and spatial dimension, respectively. With this formulation, the cross-context feature can be compressed into three groups of basis feature vectors, which are expressed as $x_c \in \mathbb{R}^{1 \times 1 \times C}$, $x_h \in \mathbb{R}^{H \times 1 \times C}$ and $x_w \in \mathbb{R}^{1 \times W \times C}$. To further enhance the representation ability of the basis vectors, we feed them into a 1×1 convolutional layer and expand it by K times to obtain K groups of basis for each spatial dimension, $x'_c \in \mathbb{R}^{K \times 1 \times 1 \times C}$, $x'_h \in \mathbb{R}^{K \times H \times 1 \times C}$

and $x'_w \in \mathbb{R}^{K\times 1\times W\times C}$.

Axis-context Attention (ACA). After obtaining the semantic basis from PIA, we further build an ACA to enhance intra-correlation among the extracted basis. We implement this Axis-context Attention based on the multi-head self-attention. Taken $x'_h \in \mathbb{R}^{K\times H\times 1\times C}$ as example, we flatten it to shape $KH \times C$ to produce *Query*, *Key* and *Value* features. Then, self-attention as defined in Eq. 1 and 2 is applied. Following the multi-head attention layer, we further add a norm layer and a feed forward network to enhance the fitting ability of the subspace characteristics. Note that the ACA is also applied to x'_w and x'_c in a similar manner.

Attention Construct Module (ACM). With PIA and ACA, we can obtain the feature basis for different spatial dimensions, which can be utilized to represent the structural composition of the target objects in the testing frame. To transform the structural prior to the original input (*i.e.,* X_s from Global-temporal Attention Module), we further propose an Attention Construct Module. The architecture of ACM is shown in Fig. 4. Taking the feature basis of each group as input, the ACM first performs the low-rank structural component reconstruction by

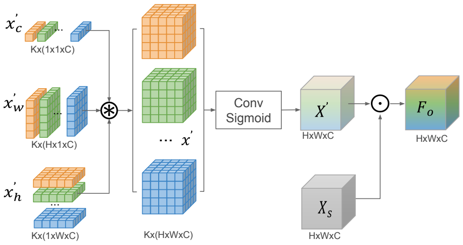

Fig. 4. Illustration of Attention Construct Module (ACM). $X^{'}$ indicates the reconstructed attention multiplied by X_s for generating final output F_o.

$$x'_i = x'_{c_i} \odot x'_{h_i} \odot x'_{w_i}. \tag{4}$$

where $x'_i \in \mathbb{R}^{H\times W\times C}$ indicates the combined feature of each group and $\odot$ is Kronecker Production. By this means, the obtained feature x'_i is reconstructed to the orginal shape. Then, the reconstructed structural components are combined via weighted summation,

$$x' = \sum_{k=1}^{K} \alpha_k \cdot x'_k \ . \tag{5}$$

where $\alpha = \{\alpha\}_{k=1}^{K}$ are learnable weights. A Sigmoid function is then applied to convert the x' into a 3D attention matrix X'. Finally, the attention matrix X' is applied to the cross-context feature X_s to construct the output feature F_o of the local-spatial attention module, which will be fed to the decoder for mask generation. As shown in Fig. 1, compared with the baseline model, the generated structural feature shows great capability in capturing the discriminative embeddings of the target object and suppressing the inferential instances from the background.

4 Experiments

4.1 Dataset and Evaluation Metric

We evaluate the proposed SFRNet on three benchmark datasets for video object segmentation including DAVIS2016 [34], DAVIS2017 [35], and YouTube-VOS [53]. DAVIS2016 contains 50 high-quality videos with per-frame fine-grained annotations. In this dataset, the multiple instances of the video sequence are grouped as one object for segmentation. DAVIS2017 is an extension version of DAVIS2016, and consists of 60 sequences for training and 30 sequences for testing. In DAVIS-2017, instance-level video object segmentation is evaluated in each frame. YouTube-VOS is a large scale VOS dataset, which contains 3471 video sequences for training and 474/507 videos for validation in the 2018/2019 version of dataset. Compared with the DAVIS benchmark, videos in YouTube-VOS are more challenging with large variations in object motion, deformation and cluttered background.

To evaluate the proposed model, we adopt three metrics including mean region similarity ($\mathcal{J}$), mean contour accuracy ($\mathcal{F}$) and their average ($\mathcal{J}\&\mathcal{F}$).

4.2 Implementation Details

Parameter Settings. In our model, the number of low-rank structural component K is set to 16 in LAM, the head numbers L in Transformer is set to 4, and the number of attention layers in GAM is set to 3. We apply fixed *Sine* spatial positional embedding in the self-attention and the cross-attention.

Training Details. Following previous methods [11,28,33,37,38,46,50], we conduct a two-stage training process. At first, we pretrain the model using static image datasets [13,26,39,47,56], by constructing synthetic video data through affinity transformation and image augmentation operations. The learning rate is

Table 1. Comparison results on DAVIS2017 validation set.

Method	FRTM [36]	PReMVOS [31]	LWL [3]	STM [33]	CFBI [58]	CoVOS [52]	GraphMem [30]	KMN [37]	JOINT [32]	RDE [25]	HMMN [38]	STCN [11]	**SFRNet**
$\mathcal{J}\&\mathcal{F}\uparrow$	76.7	77.8	81.6	81.8	81.9	82.4	82.8	82.8	83.5	84.2	84.7	85.4	**85.9**
$\mathcal{J}\uparrow$	73.9	73.9	79.1	79.2	79.1	79.7	80.2	80.0	80.8	80.8	81.9	82.2	**82.7**
$\mathcal{F}\uparrow$	79.6	81.7	84.1	84.3	84.6	85.1	85.2	85.6	86.2	87.5	87.5	88.6	**89.1**

Table 2. Comparison results on DAVIS2016 validation set.

Method	OSMN [57]	FEELVOS [44]	FRTM [36]	CINN [1]	CoVOS [52]	STM [33]	CFBI [58]	KMN [37]	HMMN [38]	RDE [25]	**SFRNet (ours)**
$\mathcal{J}\&\mathcal{F}\uparrow$	73.5	81.7	83.5	84.2	89.1	89.3	89.4	90.5	90.8	91.1	**91.3**
$\mathcal{J}\uparrow$	74.0	81.1	83.6	83.4	88.5	88.7	88.3	89.5	89.6	89.7	**90.5**
$\mathcal{F}\uparrow$	72.9	82.2	83.4	85.0	89.6	89.9	90.5	91.5	92.0	**92.5**	92.1
FPS	9.3	2.9	17.7	-	39.6	8.4	7.2	-	11.6	40.0	20.4

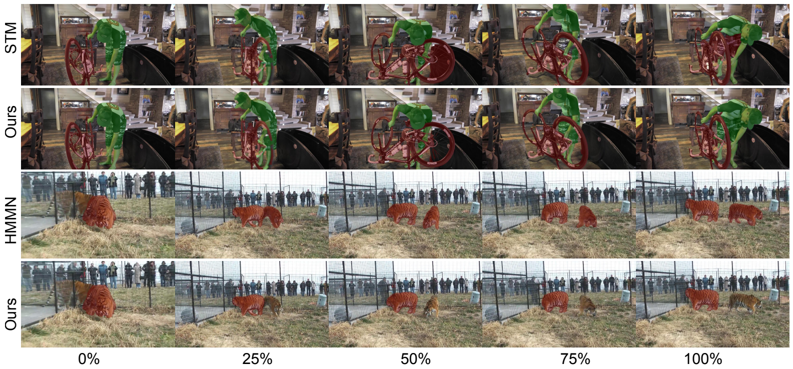

Fig. 5. Qualitative examples on DAVIS 2017 valid, and YouTube-VOS 2019 valid sets. The two examples show the comparisons of our method with STM and HMMN. Our method enables robust mask prediction in different scenarios.

set to 1e-5 in the first 150k iterations and decreases by 1/10 for the next 150k iterations. At the second stage, the pretrained model is further finetuned on YouTube-VOS and DAVIS2017. The initial learning rate is 1e-5 and decays by 1/10 at 125k iterations. The model converges after 150k iterations. The whole training process is conducted with 4 NVIDIA RTX 2080Ti GPUs. We use Adam optimizer with a batch size of 16 in pretraining and 8 at the finetuning stage. All the images and video frames are resized to 384×384 during training. The bootstrapped cross-entropy loss is applied for model optimization.

Testing Details. During testing, we use the first frame annotation and a recent prediction to form the key frames. The input images are kept at their original resolution for prediction. To avoid frequently running the efficient encoder on key frames, which decreases models' efficiency, we update the key frame features at fixed frequencies.

We compare the results on DAVIS2017 under different update frequencies in Tab. 7 and found that setting updating frequency to be 3 works best for our method.

4.3 Comparison to state-of-the-art

DAVIS Datasets. We first compare the performance on DAVIS2017 for multi-instance video object segmentation. The results on the validation split of DAVIS-2017 are shown in Tab. 1. As we can see, our SFRNet achieves superior performance against previous online learning based methods and memory-based methods. To further verify the generalization of our model, we conduct comparison experiments on the test-dev of DAVIS2017 in supplementary material, our model again achieves better performance against previous state-of-the-art approaches. We also evaluate the performance for single-object on DAVIS2016

Table 3. Comparison results on YouTube-VOS 2018 validation set.

Method	PReMVOS [31]	FRTM [36]	CoVOS [52]	STM [33]	AFB-URR [28]	CFBI [58]	KMN [37]	LWL [3]	LCM [18]	HMMN [38]	STCN [11]	JOINT [32]	**SFRNet**
$\mathcal{G} \uparrow$	66.9	72.1	79.0	79.4	79.6	81.4	81.4	81.5	82.0	82.6	83.0	83.1	**83.6**
$\mathcal{J}_\mathcal{S} \uparrow$	71.4	72.3	79.4	79.7	78.8	81.1	81.4	80.4	82.2	82.1	81.9	81.5	**82.4**
$\mathcal{F}_\mathcal{S} \uparrow$	75.9	76.2	83.6	84.2	83.1	85.8	85.6	84.9	86.7	87.0	86.5	85.9	**87.2**
$\mathcal{J}_\mathcal{U} \uparrow$	56.5	65.9	72.6	72.8	74.1	75.3	75.3	76.4	75.7	76.8	77.9	**78.7**	78.1
$\mathcal{F}_\mathcal{U} \uparrow$	63.7	74.1	80.4	80.9	82.6	83.4	83.3	84.4	83.4	84.6	85.7	86.5	**86.7**

Table 4. Comparison results on YouTube-VOS 2019 validation set.

Method	CFBI [58]	SST [15]	RDE [25]	MiVOS [10]	HMMN [38]	STCN [11]	JOINT [32]	**SFRNet**
$\mathcal{G} \uparrow$	81.0	81.8	81.9	82.4	82.5	82.7	82.8	**83.3**
$\mathcal{J}_\mathcal{S} \uparrow$	80.6	80.9	81.1	80.6	81.7	81.1	80.8	**82.0**
$\mathcal{F}_\mathcal{S} \uparrow$	85.1	-	85.5	84.7	86.1	85.4	84.8	**86.5**
$\mathcal{J}_\mathcal{U} \uparrow$	75.2	76.6	76.2	78.2	77.3	78.2	**79.0**	78.2
$\mathcal{F}_\mathcal{U} \uparrow$	83.0	-	84.8	85.9	85.0	85.9	**86.6**	**86.6**

in Tab. 2. As we can see, our method achieves favorable performance in terms of both accuracy and speed.

YouTube-VOS. Tab. 3 and Tab. 4 compare the proposed SFRNet with other state-of-the-art methods on the 2018/2019 validation sets of YouTube-VOS. Our method surpasses these existing top competitors with better overall accuracy on this benchmark. Among the existing approaches, JOINT [32] achieves the best $\mathcal{J}_\mathcal{U}$ accuracy for unseen objects. This is because JOINT adopts an online learning strategy, which achieve better performance for unseen categories of objects but sacrifices the efficiency. Without testing-time finetuning on the first frame, SFRNet achieves competitive accuracy for unseen objects compared to JOINT [32], and outperforms all the other methods, which demonstrates the superior generalization ability of our model.

Qualitative Results. Fig. 5 lists the visual comparison between our SFRNet with STM [33] and HMMN [38]. As we can see, the proposed SFRNet is able to effectively capture the structural representation of the target objects for high-quality mask generation. Besides, the proposed Global-temporal Attention Module and Local-spatial Attention Module cooperate to propose discriminative space-time correlation, which shows great effect in distinguishing the target object from interfered instances.

4.4 Method Analysis

In the following, we provide detailed analysis to demonstrate the effectiveness of the designs and modules in our method. For experiments in this part, we use the training splits of DAVIS2017 and YouTube-VOS for model training and report results on DAVIS2017 validation set.

Table 5. The effectiveness analysis of Global-temporal Attention and Local-spatial Attention Module on DAVIS2017 validation set.

	STM	GAM	LAM	TREnc	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
M1	✓				81.2	77.8	84.6	16.4
M2	✓		✓		82.5	79.4	85.7	17.8
M3		✓			82.9	79.5	86.3	9.1
M4		✓	✓		83.5	80.3	86.7	13.1
M5	✓			✓	81.7	78.7	84.8	15.4

Table 6. Performance comparison between key-frame-based model and memory-based model.

	Architecture	MEM	KF	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
M1	STM	✓		81.2	77.8	84.6	16.4
M2	STM		✓	80.8	77.3	84.3	21.5
M3	STM+LAM		✓	82.5	79.4	85.7	17.8
M4	GAM	✓		82.9	79.5	86.3	9.1
M5	GAM		✓	82.0	78.5	85.5	13.8
M6	GAM+LAM		✓	83.5	80.3	86.7	13.1

Table 7. Impact of key-frames update frequency on final accuracy.

DAVIS2017	Every 1	Every 2	Every 3
$\mathcal{J}\&\mathcal{F}$	84.0	85.3	85.9
DAVIS2017	Every 4	Every 7	Every 10
$\mathcal{J}\&\mathcal{F}$	85.7	83.2	83.0

Table 8. Impact of the number for structural components.

	$K = 4$	$K = 8$	$K = 16$	$K = 20$
$\mathcal{J}\&\mathcal{F}$	83.3	82.9	83.5	83.2
$\mathcal{J}$	80.1	79.8	80.3	80.3
$\mathcal{F}$	86.5	86.1	86.7	86.1

Effectiveness of Global-temporal Attention Module. We verify the effectiveness of GAM by comparing its performance with a baseline model based on STM [33]. Compared with GAM, the STM baseline only contains a single head self-attention architecture. Based on the results of M1 and M3 in Tab. 5, we can observe that GAM achieves 1.7% improvement on $\mathcal{J}\&\mathcal{F}$ against the STM baseline. Besides, when combined with LAM, the GAM still outperforms non-local module by 1.0% on $\mathcal{J}\&\mathcal{F}$ (see M2 and M4 in Tab. 5). These results demonstrate that the multi-layer based GAM achieves more robust results than the STM baseline on the VOS task.

Effectiveness of Local-spatial Attention Module. At first, we show the impact of LAM in Tab. 5. As shown, the application LAM (M2 and M4) helps to improve $\mathcal{J}\&\mathcal{F}$ by 1.3% and 0.6% for both the non-local baseline (M1) and GAM baseline (M3), respectively. To demonstrate the design of LAM, we compare it with a plain Transformer Encoder. As indicated by M1 and M5 in Tab. 5, adding a Transformer Encoder upon non-local module improves $\mathcal{J}\&\mathcal{F}$ by 0.5%. Yet, replacing the Transformer Encoder with LAM (M2 in Tab. 5) can further improve $\mathcal{J}\&\mathcal{F}$ by 0.8% and the speed from 15.4 FPS to 17.8 FPS. This demonstrates the effectiveness and efficiency of the LAM module.

Efficiency with Key-frames. In this part, we analyze the efficiency of SFRNet. As shown in Tab. 2, SFRNet achieves state-of-the-art accuracy at the speed of 20.4 FPS, which is faster than most of the previous methods. As introduced in the previous section, in SFRNet we adopt only two key frames for recording the historical information, which is in contrast to traditional memory-based methods that always add new frames into the memory bank. As shown in Tab. 6, for both STM-only and GAM-only baselines, adopting two-key-frame strategy

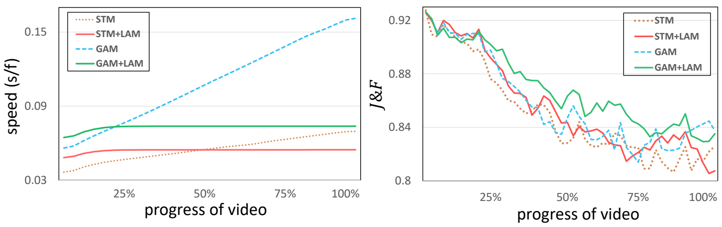

Fig. 6. Accuracy and time analysis for different methods on DAVIS2017 validation set.

Table 9. Impact of key-frames update frequency on final accuracy.

Memory Frames	2	4	6
$\mathcal{J}\&\mathcal{F}$	85.9	86.1	86.3
GFLOPS	91.6	138.3	185.0

Table 10. Impact of different LAM attention on the output of GAM.

	sigmoid	add	multi
$\mathcal{J}\&\mathcal{F}$	83.5	82.3	83.2

(M2 and M5) achieves faster speeds yet lower accuracy. This shows that a large memory bank is critical for effectiveness while bringing in more computational costs, which poses the challenge of balancing between capacity and efficiency in previous memory-based designs. In contrast, as indicated by M3 and M6, our proposed LAM with key frame based strategy keeps fast speeds while achieving the best accuracy. In Fig. 6, we compare efficiency and accuracy over time under different model designs. As we can see, for both STM-based and GAM-based model, the proposed LAM leads to better accuracy as well as faster and more stable speed over time. This again demonstrates the LAM's effectiveness to leverage local spatial composition of testing frames, which improves models' ability to effectively extract information from a small set of historical frames.

Update of Key-frames. In Tab. 7, we investigate the impact of updating key-frames frequency. We notice that setting frequency every 3 frames performs best for both datasets, and either a lower or higher frequency doesn't perform better. **Larger memory bank and LAM attention.** As shown in Tab. 9, increasing the memory frames can still improve accuracy, yet decrease computational efficiency. With only 2 memory frames, the proposed SFRNet can already achieve state-of-the-art accuracy in the experiments showed above. This demonstrates that GAM and LAM work complementarily to boost the effectiveness and alleviate the dependence on large memory banks. In Tab. 10, we compare the different LAM attention operation on the output feature of GAM and the sigmoid operation performs the best.

Analysis of Feature Basis in LAM. We analyze the feature basis extracted in LAM for constructing the low-rank structural component. As discussed in

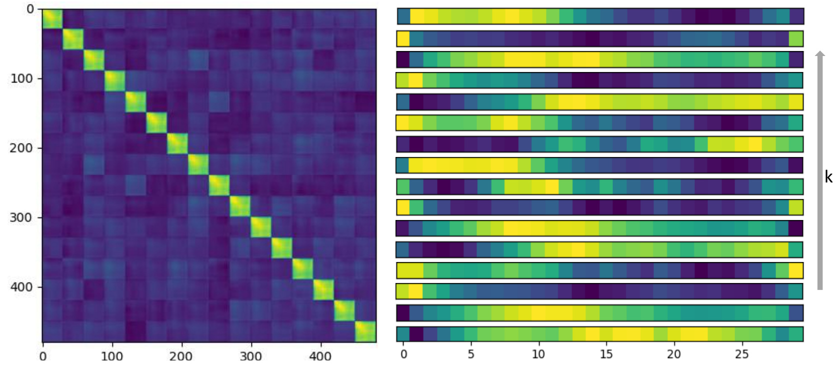

Fig. 7. Visualization of feature basis in LAM. The left side shows the affinity between the different feature basis. The right side visualizes a group of feature basis. Examples here are based on the features basis for "Height" dimension.

Sec. 3.2, we have K discriminative structural components produced by K groups of feature basis collected from different spatial dimensions. In Fig. 7, we visualize correlations between all $K \cdot H$ feature basis for the "Height" dimension. Note that these basis are organized by each of K groups. As we can see, the basis of the same group has a high correlation, and the basis of different groups presents a low correlation. This shows the orthogonality between different basis groups, which demonstrates that the constructed structural components are representing different aspects of structural composition for the target object as well as the scene. One group of the feature basis are also visualized in Fig. 7. In Tab. 8, we also vary the number of structural components in LAM and find that our model achieves the highest accuracy when $K = 16$.

5 Conclusion

We present SFRNet as a novel and effective framework for semi-supervised video object segmentation. In SFRNet, we first introduce a Global-temporal Attention Module (GAM) based on self-attention modules to capture the target objects' temporal context across frames. Then, the Local-spatial Attention Module (LAM) is proposed to further reconfigure features with a testing frame's spatial structural prior, so as to reinforce the objectness of foreground objects and suppress the interference from background regions. GAM and LAM work complementarily to extract target objects from video frames. Extensive experiments are conducted to analyze the effectiveness of SFRNet. The results demonstrate that our method achieves state-of-the-art results on multiple VOS benchmarks.

Acknowledgment This work is supported by National Postdoctoral Program for Innovative Talent (BX2021051), and National Natural Science Foundation of China under grant No. U1903215, 61725202, 61829102, 62206039.

References

1. Bao, L., Wu, B., Liu, W.: Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In: CVPR. pp. 5977–5986 (2018) 9
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095 (2021) 4
3. 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: ECCV (2020) 9, 11
4. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV (2010) 3
5. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR (2017) 2, 3
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 4, 6
7. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: CVPR (2021) 4
8. Chen, Y., Pont-Tuset, J., Montes, A., Van Gool, L.: Blazingly fast video object segmentation with pixel-wise metric learning. In: CVPR (2018) 2
9. Cheng, B., Liu, G., Wang, J., Huang, Z., Yan, S.: Multi-task low-rank affinity pursuit for image segmentation. In: ICCV (2011) 3, 7
10. Cheng, H.K., Tai, Y.W., Tang, C.K.: Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021) 11
11. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. arXiv preprint arXiv:2106.05210 (2021) 2, 3, 4, 5, 6, 9, 11
12. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: ICCV (2017) 2
13. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE transactions on pattern analysis and machine intelligence **37**(3), 569–582 (2014) 9
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
15. Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: SSTVOS: sparse spatiotemporal transformers for video object segmentation. In: CVPR (2021) 3, 11
16. Ge, W., Lu, X., Shen, J.: Video object segmentation using global and instance embedding learning. In: CVPR (2021) 3
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 5
18. Hu, L., Zhang, P., Zhang, B., Pan, P., Xu, Y., Jin, R.: Learning position and target consistency for memory-based video object segmentation. In: CVPR (2021) 11
19. Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F.: Temporally distributed networks for fast video semantic segmentation. In: CVPR (2020) 4
20. Hu, P., Liu, J., Wang, G., Ablavsky, V., Saenko, K., Sclaroff, S.: Dipnet: Dynamic identity propagation network for video object segmentation. In: WACV (2020) 2

21. Hu, P., Wang, G., Kong, X., Kuen, J., Tan, Y.P.: Motion-guided cascaded refinement network for video object segmentation. In: CVPR (2018) 2
22. Hu, P., Wang, G., Kong, X., Kuen, J., Tan, Y.P.: Motion-guided cascaded refinement network for video object segmentation. IEEE Trans. on PAMI (2019) 2
23. Hu, Y.T., Huang, J.B., Schwing, A.G.: Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In: ECCV (2018) 2
24. Khoreva, A., Benenson, R., Ilg, E., Brox, T., Schiele, B.: Lucid data dreaming for object tracking. CoRR **abs/1703.09554** (2017), `http://arxiv.org/abs/1703.09554` 2, 3
25. Li, M., Hu, L., Xiong, Z., Zhang, B., Pan, P., Liu, D.: Recurrent dynamic embedding for video object segmentation. In: CVPR (2022) 4, 9, 11
26. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: Fss-1000: A 1000-class dataset for few-shot segmentation. In: CVPR (2020) 9
27. Li, X., Change Loy, C.: Video object segmentation with joint re-identification and attention-aware mask propagation. In: ECCV (2018) 2, 3
28. 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) 4, 6, 9, 11
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021) 4
30. Lu, X., Wang, W., Danelljan, M., Zhou, T., Shen, J., Van Gool, L.: Video object segmentation with episodic graph memory networks. In: ECCV (2020) 2, 9
31. Luiten, J., Voigtlaender, P., Leibe, B.: Premvos: Proposal-generation, refinement and merging for video object segmentation. In: ACCV (2018) 9, 11
32. Mao, Y., Wang, N., Zhou, W., Li, H.: Joint inductive and transductive learning for video object segmentation. In: ICCV (2021) 4, 9, 11
33. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV (2019) 2, 3, 4, 5, 9, 11, 12
34. 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: CVPR (2016) 3, 9
35. 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) 3, 9
36. Robinson, A., Lawin, F.J., Danelljan, M., Khan, F.S., Felsberg, M.: Learning fast and robust target models for video object segmentation. In: CVPR (2020) 9, 11
37. Seong, H., Hyun, J., Kim, E.: Kernelized memory network for video object segmentation. In: ECCV (2020) 2, 9, 11
38. Seong, H., Oh, S.W., Lee, J.Y., Lee, S., Lee, S., Kim, E.: Hierarchical memory matching network for video object segmentation. In: ICCV (2021) 2, 4, 5, 9, 11
39. Shi, J., Yan, Q., Xu, L., Jia, J.: Hierarchical image saliency detection on extended cssd. IEEE transactions on pattern analysis and machine intelligence **38**(4), 717–729 (2015) 9
40. Tang, C., Yuan, L., Tan, P.: Lsm: Learning subspace minimization for low-level vision. In: CVPR (2020) 3, 7
41. Tao, L., Porikli, F., Vidal, R.: Sparse dictionaries for semantic segmentation. In: ECCV (2014) 3
42. Tsai, Y.H., Yang, M.H., Black, M.J.: Video segmentation via object flow. In: CVPR (2016) 2

43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017) 4, 6
44. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.: FEELVOS: fast end-to-end embedding learning for video object segmentation. In: CVPR (2019) 2, 3, 9
45. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. In: BMVC (2017) 2, 3
46. Wang, H., Jiang, X., Ren, H., Hu, Y., Bai, S.: Swiftnet: Real-time video object segmentation. In: CVPR (2021) 2, 3, 4, 5, 6, 9
47. Wang, L., Lu, H., Wang, Y., Feng, M., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (2017) 9
48. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122 (2021) 4
49. Xiao, H., Feng, J., Lin, G., Liu, Y., Zhang, M.: Monet: Deep motion exploitation for video object segmentation. In: CVPR (2018) 2, 3
50. Xie, H., Yao, H., Zhou, S., Zhang, S., Sun, W.: Efficient regional memory network for video object segmentation. In: CVPR (2021) 2, 9
51. Xie, H., Yao, H., Zhou, S., Zhang, S., Sun, W.: Efficient regional memory network for video object segmentation. In: CVPR (2021) 3, 4
52. Xu, K., Yao, A.: Accelerating video object segmentation with compressed video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 9, 11
53. 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) 3, 9
54. Xu, X., Wang, J., Li, X., Lu, Y.: Reliable propagation-correction modulation for video object segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022) 4
55. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking (2021) 4
56. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR (2013) 9
57. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.K.: Efficient video object segmentation via network modulation. In: CVPR (2018) 9
58. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by foreground-background integration. In: ECCV (2020) 2, 3, 9, 11
59. Zhang, L., Lin, Z., Zhang, J., Lu, H., He, Y.: Fast video object segmentation via dynamic targeting network. In: ICCV (2019) 2
60. Zhang, L., Zhang, J., Lin, Z., Měch, R., Lu, H., He, Y.: Unsupervised video object segmentation with joint hotspot tracking. In: ECCV (2020) 2
61. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR (2021) 4, 6
62. Zohrizadeh, F., Kheirandishfard, M., Kamangar, F.: Image segmentation using sparse subset selection. In: WACV (2018) 3