

# Full-Duplex Strategy for Video Object Segmentation

Ge-Peng Ji<sup>1,2</sup> Keren Fu<sup>3</sup> Zhe Wu<sup>4</sup> Deng-Ping Fan<sup>1,\*</sup> Jianbing Shen<sup>5</sup> Ling Shao<sup>1</sup>

<sup>1</sup> IIAI <sup>2</sup> School of CS, Wuhan University <sup>3</sup> College of CS, Sichuan University

<sup>4</sup> Peng Cheng Laboratory <sup>5</sup> Department of CIS, University of Macau

<http://dpfan.net/FSNet/>

## Abstract

Appearance and motion are two important sources of information in video object segmentation (VOS). Previous methods mainly focus on using simplex solutions, lowering the upper bound of feature collaboration among and across these two cues. In this paper, we study a novel framework, termed the *FSNet* (Full-duplex Strategy Network), which designs a relational cross-attention module (RCAM) to achieve the bidirectional message propagation across embedding subspaces. Furthermore, the bidirectional purification module (BPM) is introduced to update the inconsistent features between the spatial-temporal embeddings, effectively improving the model robustness. By considering the mutual restraint within the full-duplex strategy, our *FSNet* performs the cross-modal feature-passing (i.e., transmission and receiving) simultaneously before the fusion and decoding stage, making it robust to various challenging scenarios (e.g., motion blur, occlusion) in VOS. Extensive experiments on five popular benchmarks (i.e., DAVIS<sub>16</sub>, FBMS, MCL, SegTrack-V2, and DAVSOD<sub>19</sub>) show that our *FSNet* outperforms other state-of-the-arts for both the VOS and video salient object detection tasks.

## 1. Introduction

Video object segmentation (VOS) [12, 32, 101, 104] is a fundamental topic in computer vision for intelligent video analysis, whose purpose is to delineate pixel-level moving object<sup>1</sup> masks in each frame. It has been widely applied to robotic manipulation [1], autonomous cars [58], video editing [34], medicine [36], optical flow estimation [18], interactive segmentation [9, 29, 60], URVOS [75], and video captioning [65]. There are two settings for addressing this task (i.e., *semi-supervised* [95] and *unsupervised* [59] VOS), depending on whether or not the candidate object is given manually in the first frame. In this

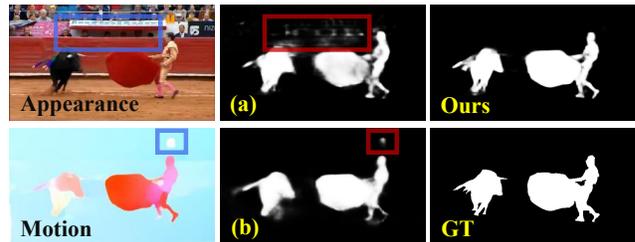


Figure 1: Visual comparison between the *simplex* (i.e., (a) appearance-refined motion and (b) motion-refined appearance) and our *full-duplex strategy*. In contrast, our *FSNet* offers a collaborative way to leverage the appearance and motion cues under the mutual restraint of full-duplex strategy, thus providing more accurate structure details and alleviating the short-term feature drifting issue [117].

work, we focus on the unsupervised setting, i.e., zero-shot VOS [126, 127]. For semi-supervised VOS, we refer readers to prior works [5, 8, 43, 53, 73, 76, 114, 116, 120, 122].

Recent years have witnessed promising progress of addressing video content understanding by exploiting appearance (e.g., color frame [119]) and motion (e.g., optical flow [33, 83] and pixel trajectory [78]) correlation between frames. However, short-term dependency estimation (i.e., one-step motion cues [33, 83]) produces unreliable results and suffers the common ordeals [30] (e.g., diffusion, noise, and deformation), while the capability of appearance-based modeling (e.g., recurrent neural network (RNN) [59, 85]) is severely hindered by blurred foregrounds or cluttered backgrounds [14]. Those conflicts are prone to accumulating inaccuracies with the propagation of spatial-temporal embeddings, which cause short-term feature drifting [117].

Earlier solutions address this issue using *direction-independent strategy* [16, 35, 38, 85, 108], which would be to encode the appearance and motion features individually and fuse them directly. However, this implicit strategy will cause feature conflicts, since motion and appearance are two distinctive modalities, extracted from separate branches. A reasonable idea is to integrate them in a guided manner, and

\*Corresponding author: Deng-Ping Fan ([dengpfan@gmail.com](mailto:dengpfan@gmail.com)). Work was done while Ge-Peng Ji was an intern mentored by Deng-Ping Fan.

<sup>1</sup>We use ‘foreground object’ & ‘target object’ interchangeably.

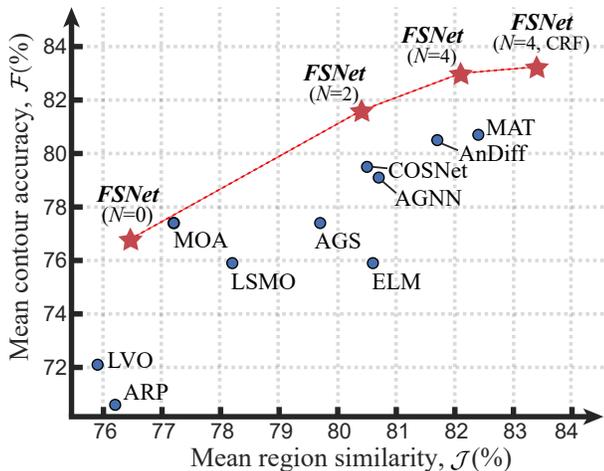


Figure 2: Mean contour accuracy ( $\mathcal{F}$ ) vs. mean region similarity ( $\mathcal{J}$ ) scores on DAVIS<sub>16</sub> [71]. Circles indicate UVOS methods. Four variants of our *FSNet* are shown in **bold-italic**, in which ‘ $N$ ’ indicates the number of BPM. Compared with the best unsupervised VOS model (MAT [127] with CRF [42] post-processing), the proposed method *FSNet* ( $N=4$ , CRF) achieves the new SOTA by a large margin.

thus, several recent approaches opt for the *simplex strategy* [30, 50, 54, 62, 68, 88, 127], which is either appearance-based or motion-guided. Although these two strategies have achieved remarkable advances, they both fail to infer the mutual restraint between the appearance and motion cues that both guide human visual attention allocation during dynamic observation, according to previous studies in cognitive psychology [40, 87, 105] and computer vision [35, 93].

For the same object, we argue that appearance and motion characteristics should be homogeneous to a certain degree. Intuitively, as shown in Fig. 1, the foreground region of appearance (top-left) and motion (bottom-left) maps intrinsically share the correlative patterns about perceptions, including semantic structure, movement posture. However, misguided knowledge in the individual modality, *e.g.*, static spectators at the bullring and dynamic watermark on TV (blue boxes), will produce inaccuracies during the feature propagation, and thus, it easily stains the result (red boxes).

To alleviate the above conflicts, it is important to introduce a new modality transmission scheme, instead of embedding them individually. Inspired by this, we introduce the idea of *full-duplex*<sup>2</sup> from the field of wireless communication. As shown in Fig. 4 (c) & Fig. 5 (c), this is a bidirectional-attention scheme across motion and appearance cues, which explicitly incorporates the appearance and motion patterns in a unified framework. As can be seen in the first row of Fig. 1, the proposed **Full-duplex Strategy**

<sup>2</sup>On the same channel, information can be transmitted and received simultaneously [4].

**Network** (*FSNet*) visually performs better than the one with simplex strategy. To understand what enables good learning strategies, we comprehensively delve into the simplex and full-duplex strategies of our framework and present the following contributions:

- We emphasize the importance of the full-duplex strategy for the spatial-temporal representations. Specifically, a bidirectional interaction module, termed the relational cross-attention module (RCAM), is used to extract discriminative features from the appearance and motion branches, which ensures the *mutual restraint* between each other.
- To further improve the model robustness, we introduce a bidirectional purification module (BPM), which is equipped with an interlaced decremental connection (IDC) to automatically update inconsistent features between the spatial-temporal embeddings.
- We demonstrate that our *FSNet* performs superior performance on five mainstream benchmarks, especially for *FSNet* ( $N=4$ , CRF) outperforms the SOTA UVOS model (*i.e.*, MAT [127]) on the DAVIS<sub>16</sub> [71] leaderboard by a margin of 2.4% in terms of  $\mathcal{F}$  score (see Fig. 2), with less training data (*i.e.*, Ours-13K vs. MAT-16K). This suggests that the mutual restraints within full-duplex strategy is promising for the spatial-temporal learning tasks.

## 2. Related Works

### 2.1. Unsupervised VOS

Although there are many works [7, 15, 37, 69, 89, 107] addressing the VOS task in a semi-supervised manner, *i.e.*, by supposing an object mask annotation is given in the first frame, other researchers have attempted to address the more challenging unsupervised VOS (UVOS) problem. Early UVOS models resort to low-level handcrafted features for heuristic segmentation inference, such as long sparse point trajectories [6, 23, 63, 79, 97], object proposals [47, 48, 57, 72], saliency priors [20, 92, 94], optical flow [88], or superpixels [24, 25, 109]. As such, these traditional models have limited generalizability, and thus low accuracy in highly dynamic and complex scenarios, due to their lack of semantic information and high-level content understanding. Recently, RNN-based models [3, 81, 85, 99, 114, 125] have become popular due to their better capability of capturing long-term dependencies, as well as their use of deep learning. In this case, UVOS is formulated as a recurrent modeling issue over time, where spatial features are jointly exploited with long-term temporal context.

How to combine motion cues with appearance features is a long-standing problem in this field. To this end, Tokmakov *et al.* [84] proposed to simply use the motion patterns required from the video. However, their method is

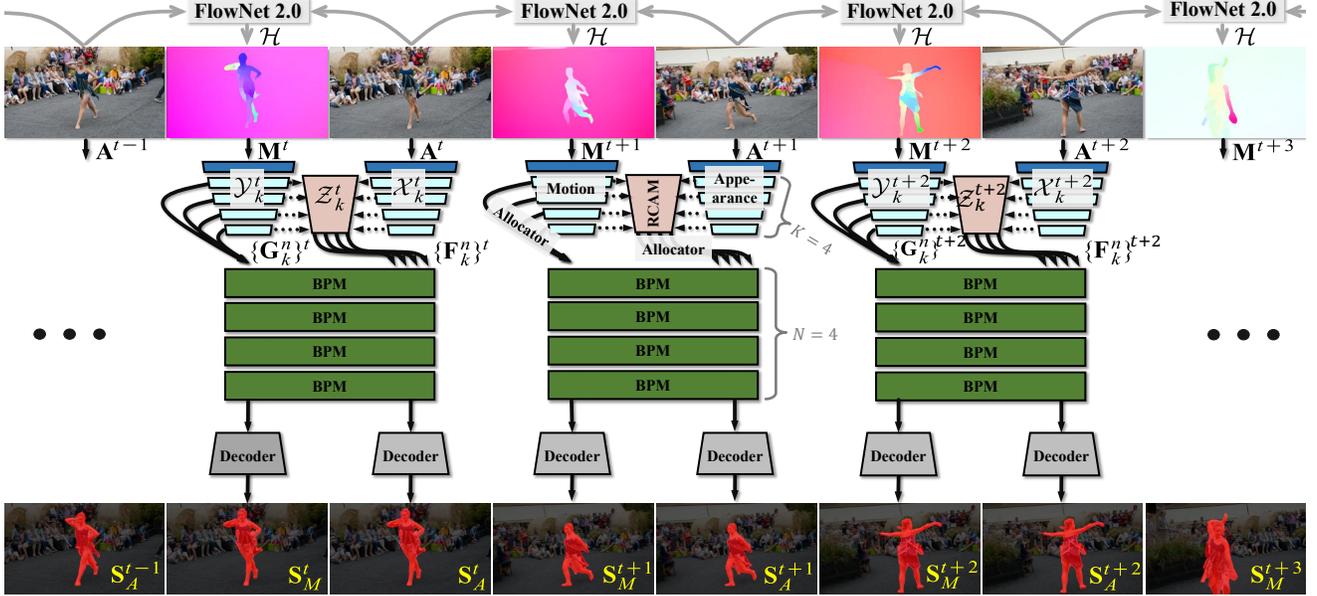


Figure 3: The pipeline of our *FSNet*. The Relational Cross-Attention Module (RCAM) abstracts more discriminative representations between the motion and appearance cues using the full-duplex strategy. Then four Bidirectional Purification Modules (BPM) are stacked to further re-calibrate inconsistencies between the motion and appearance features. Finally, we utilize the decoder to generate our prediction.

unable to accurately segment objects between two similar consecutive frames, since it relies heavily on the guidance of optical flow. To resolve this, several works [16, 80, 85] have integrated the spatial and temporal features from the parallel network, which can be viewed as plain feature fusion from the independent spatial and temporal branch with an implicit modeling strategy. Li *et al.* [51] proposed a multi-stage processing method to tackle UVOS, which first utilizes a fixed appearance-based network to generate objectness and then feeds this into the motion-based bilateral estimator to segment the objects.

## 2.2. Attention-based VOS

The attention-based VOS task is closely related to UVOS, since it aims at extracting attention-aware object(s) from a video clip. Traditional methods [31, 98, 111, 113, 128] first compute the single-frame saliency based on various hand-crafted static and motion features, and then conduct spatial-temporal optimization to preserve coherency across consecutive frames. Recent works [45, 61, 96, 121] aim to learn a highly-semantic representation and usually perform spatial-temporal detection in an end-to-end manner. Many schemes have been proposed to employ deep networks that consider temporal information, such as ConvLSTM [22, 49, 81], take optical-flows/adjacent-frames as input [50, 96, 110], 3D convolutional [45, 61], or directly exploit temporally concatenated deep features [46]. Besides, long-term influences are often taken into account and combined with deep learning. Li *et al.* [52] proposed a key-frame strategy to locate representative high-quality video frames of

salient objects and diffused their saliency to ill-detected non-key frames. Chen *et al.* [10] improved saliency detection by leveraging long-term spatial-temporal information, where high-quality “beyond-the-scope frames” are aligned with the current frames and both types of information are fed to deep neural networks for classification. Besides considering how to better leverage temporal information, other researchers have attempted to address different problems in VSOD, such as reducing the data labeling requirements [115], developing semi-supervised approaches [82], or investigating relative saliency [102]. Fan *et al.* [22] introduced a VSOD model equipped with a saliency shift-aware ConvLSTM, together with an attention-consistent VSOD dataset with high-quality annotations.

## 3. Methodology

### 3.1. Overview

Suppose that a video clip contains  $T$  consecutive frames  $\{\mathbf{A}^t\}_{t=1}^T$ . We first utilize optical flow field generator  $\mathcal{H}$ , *i.e.*, FlowNet 2.0 [33], to generate  $T - 1$  optical flow maps  $\{\mathbf{M}^t\}_{t=1}^{T-1}$ , which are all computed by two adjacent frames ( $\mathbf{M}^t = \mathcal{H}[\mathbf{A}^t, \mathbf{A}^{t+1}]$ ). To ensure the inputs match, we discard the last frame in the pipeline. Thus, the proposed pipeline takes both the appearance image  $\{\mathbf{A}^t\}_{t=1}^{T-1}$  and its paired motion map  $\{\mathbf{M}^t\}_{t=1}^{T-1}$  as the input. First,  $\mathbf{M}^t$  &  $\mathbf{A}^t$  pairs at frame  $t^3$  are fed to two independent ResNet-50 [28] branches (*i.e.*, motion and appearance blocks in Fig. 3). The appearance features  $\{\mathcal{X}_k\}_{k=1}^K$  and motion features  $\{\mathcal{Y}_k\}_{k=1}^K$

<sup>3</sup>Here, we omit the superscript “ $t$ ” for the convenient expression.

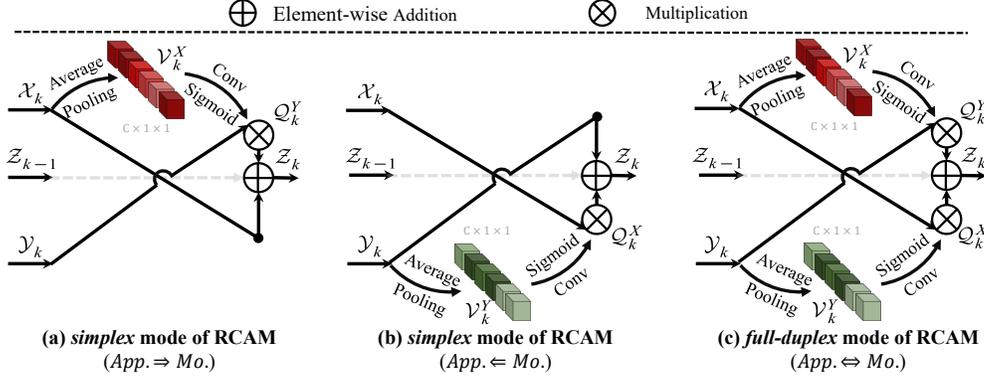


Figure 4: Illustration of our Relational Cross-Attention Module (RCAM) with a *simplex* (a & b) and *full-duplex* (c) strategy.

extracted from  $K$  layers are then sent to the Relational Cross-Attention Modules (RCAMs), which allows the network to embed spatial-temporal cross-modal features. Next, we employ the Bidirectional Purification Modules (BPMs) with  $N$  cascaded units. BPMs focus on distilling representative carriers from fused features  $\{\mathbf{F}_k^n\}_{n=1}^N$  and motion-based features  $\{\mathbf{G}_k^n\}_{n=1}^N$ . Finally, the predictions (*i.e.*,  $\mathbf{S}_M^t$  and  $\mathbf{S}_A^t$ ) at frame  $t$  are generated from two decoder blocks.

### 3.2. Relational Cross-Attention Module

As discussed in § 1, a single-modality (*i.e.*, motion or appearance) guided stimulation may cause the model to make incorrect decisions. To alleviate this, we design a cross-attention module (RCAM) via the channel-wise attention mechanism, which focuses on distilling out effective squeezed cues from two modalities and then modulating each other. As shown in Fig. 4 (c), the two inputs of RCAM are appearance features  $\{\mathcal{X}_k\}_{k=1}^K$  and motion features  $\{\mathcal{Y}_k\}_{k=1}^K$ , which are obtained from the two different branches of the standard ResNet-50 [28]. Specifically, for each  $k$ -level, we first perform global average pooling (GAP) to generate channel-wise vectors  $\mathcal{V}_k^X$  and  $\mathcal{V}_k^Y$  from each  $\mathcal{X}_k$  and  $\mathcal{Y}_k$ . Next, two  $1 \times 1$  conv layers, *i.e.*,  $\phi(x; \mathbf{W}_\phi)$  and  $\theta(x; \mathbf{W}_\theta)$ , with learnable parameters  $\mathbf{W}_\phi$  and  $\mathbf{W}_\theta$ , generate two discriminated global descriptors. The sigmoid function  $\sigma[x] = e^x / (e^x + 1)$ ,  $x \in \mathbb{R}$  is then applied to convert the final descriptors into the interval  $[0, 1]$ , *i.e.*, into the valid attention vector for channel weighting. Then, we perform outer product  $\otimes$  between  $\mathcal{X}_k$  and  $\sigma[\theta(\mathcal{V}_k^Y; \mathbf{W}_\theta)]$  to generate a candidate feature  $Q_k^X$ , and vice versa, as follows:

$$Q_k^X = \mathcal{X}_k \otimes \sigma[\theta(\mathcal{V}_k^Y; \mathbf{W}_\theta)], \quad (1)$$

$$Q_k^Y = \mathcal{Y}_k \otimes \sigma[\phi(\mathcal{V}_k^X; \mathbf{W}_\phi)]. \quad (2)$$

Then, we combine  $Q_k^X$ ,  $Q_k^Y$ , and lower-level fused feature  $Z_{k-1}$  for in-depth feature extraction. With an element-wise addition operation  $\oplus$ , conducted in the corresponding  $k$ -th level block  $\mathcal{B}_k[x]$  in the ResNet-50, we finally obtain

the fused features  $Z_k$  that contain comprehensive spatial-temporal correlations:

$$Z_k = \mathcal{B}_k [Q_k^X \oplus Q_k^Y \oplus Z_{k-1}], \quad (3)$$

where  $k \in \{1 : K\}$  denotes different feature hierarchies in the backbone. Note that  $Z_0$  denotes the zero tensor. In our implementation, we use the top four feature pyramids, *i.e.*,  $K = 4$ , which is suggested by [103, 123].

### 3.3. Bidirectional Purification Module

In addition to the RCAM described above, which integrates common cross-modality features, we further introduce the bidirectional purification module (BPM) to improve the model robustness. Following the standard in action recognition [77] and saliency detection [106], our bidirectional purification phase is composed of  $N$  BPMs connected in a cascaded manner. As shown in Fig. 3, we first employ the feature allocator  $\psi_{\{F,G\}}(x; \mathbf{W}_\psi^{\{F,G\}})$  to unify the feature representations from the previous stage:

$$\mathbf{F}_k^n = \psi_F(Z_k; \mathbf{W}_\psi^F), \quad \mathbf{G}_k^n = \psi_G(\mathcal{Y}_k; \mathbf{W}_\psi^G), \quad (4)$$

where  $k \in \{1 : K\}$  and  $n \in \{1 : N\}$  denote different feature hierarchies and number of BPM, respectively. To be specific,  $\psi_{\{F,G\}}(x; \mathbf{W}_\psi^{\{F,G\}})$  is composed of two  $3 \times 3$  conv, each with 32 filters to reduce the feature channels. Note that the allocator is conducive to reduce the computational burden as well as facilitate various element-wise operations.

Here, we consider a *full-duplex scheme* (see Fig. 5 (c)) that contains two *simplex* strategies (see Fig. 5 (a & b)) in the BPM. On one hand, the motion features  $\mathbf{G}_k^n$  contain temporal cues and can be used to enrich the fused features  $\mathbf{F}_k^n$  by the concatenation operation. On the other, the distractors in the motion features  $\mathbf{G}_k^n$  can be suppressed by multiplying the fused features  $\mathbf{F}_k^n$ . Besides, to acquire robust feature representation, we introduce an efficient cross-modal fusion strategy in this scheme, which broadcasts high-level, semantically strong features to low-level,

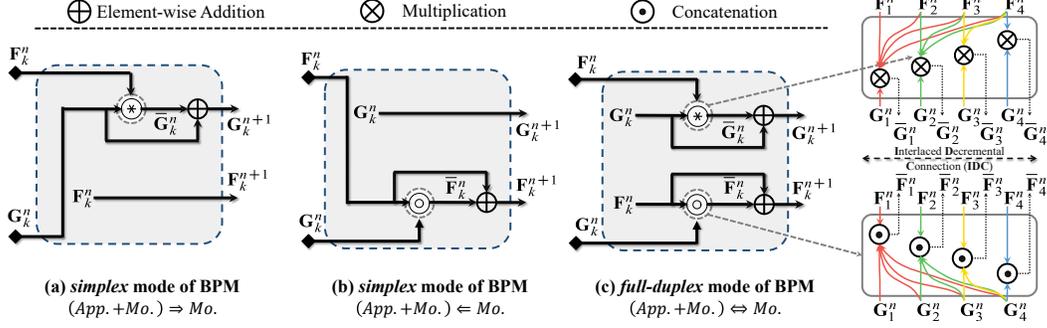


Figure 5: Illustration of our Bidirectional Purification Module (BPM) with a *simplex* and *full-duplex* strategy.

semantically weak features via interlaced decremental connection (IDC) with a top-down pathway [55]. Specifically, as the first part, the spatial-temporal feature combination branch (see Fig. 5 (b)) is formulated as:

$$\mathbf{F}_k^{n+1} = \mathbf{F}_k^n \oplus \bigcup_{i=k}^K [\mathbf{F}_k^n, \mathcal{P}(\mathbf{G}_i^n)], \quad (5)$$

where  $\mathcal{P}$  is an up-sampling operation followed by a  $1 \times 1$  convolutional layer (conv) to reshape the candidate guidance to a consistent size with  $\mathbf{F}_k^n$ . Symbols  $\oplus$  and  $\bigcup$  respectively denote element-wise addition and concatenation operations with an IDC strategy<sup>4</sup>, followed by a  $1 \times 1$  conv with 32 filters. For the other part, we formulate the temporal feature re-calibration branch (see Fig. 5 (a)) as:

$$\mathbf{G}_k^{n+1} = \mathbf{G}_k^n \oplus \bigcap_{j=k}^K [\mathbf{G}_k^n, \mathcal{P}(\mathbf{F}_j^n)], \quad (6)$$

where  $\bigcap$  denotes element-wise multiplication with an IDC strategy, followed by a  $1 \times 1$  conv with 32 filters.

### 3.4. Decoder

After feature aggregation and re-calibration with multi-pyramidal interaction, the last BPM unit produces two groups of discriminative features (*i.e.*,  $\mathbf{F}_k^N$  &  $\mathbf{G}_k^N$ ) with a consistent channel number of 32. We integrate pyramid pooling module (PPM) [124] into each skip connection of the U-Net [74] as our decoder, and only adopt the top four layers in our implementation ( $K = 4$ ). Since the features are fused from high to low level, global information is well retained at different scales of the designed decoder:

$$\hat{\mathbf{F}}_k^N = \mathcal{C}[\mathbf{F}_k^N \odot \mathcal{UP}(\hat{\mathbf{F}}_{k+1}^N)], \quad (7)$$

$$\hat{\mathbf{G}}_k^N = \mathcal{C}[\mathbf{G}_k^N \odot \mathcal{UP}(\hat{\mathbf{G}}_{k+1}^N)]. \quad (8)$$

Here,  $\mathcal{UP}$  indicates the upsampling operation after the pyramid pooling layer, while  $\odot$  is a concatenation operation between two features. Then, a conv  $\mathcal{C}$  is used for reducing the

<sup>4</sup>For instance,  $\bar{\mathbf{G}}_2^n = \bigcup_{i=2}^{K=4} [\mathbf{F}_2^n, \mathcal{P}(\mathbf{G}_i^n)] = \mathbf{F}_2^n \odot \mathcal{P}(\mathbf{G}_2^n) \odot \mathcal{P}(\mathbf{G}_3^n) \odot \mathcal{P}(\mathbf{G}_4^n)$  when  $k = 2$  and  $K = 4$ .

channels from 64 to 32. Lastly, we use a  $1 \times 1$  conv with a single filter after the upstream output (*i.e.*,  $\hat{\mathbf{F}}_1^N$  &  $\hat{\mathbf{G}}_1^N$ ), followed by a sigmoid activation function to generate the predictions (*i.e.*,  $\mathbf{S}_A^t$  &  $\mathbf{S}_M^t$ ) at frame  $t$ .

### 3.5. Training

Given a group of predictions  $\mathbf{S}^t \in \{\mathbf{S}_A^t, \mathbf{S}_M^t\}$  and the corresponding ground-truths  $\mathbf{G}^t$  at frame  $t$ , we employ the standard binary *cross-entropy* loss  $\mathcal{L}_{bce}$  to measure the dissimilarity between output and target, which computes:

$$\begin{aligned} \mathcal{L}_{bce}(\mathbf{S}^t, \mathbf{G}^t) = & - \sum_{(x,y)} [\mathbf{G}^t(x,y) \log(\mathbf{S}^t(x,y)) \\ & + (1 - \mathbf{G}^t(x,y)) \log(1 - \mathbf{S}^t(x,y))], \end{aligned} \quad (9)$$

where  $(x, y)$  indicates a coordinate in the frame. The overall loss function is then formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{bce}(\mathbf{S}_A^t, \mathbf{G}^t) + \mathcal{L}_{bce}(\mathbf{S}_M^t, \mathbf{G}^t). \quad (10)$$

For final prediction, we use  $\mathbf{S}_A^t$  since our experiments show that it better combines appearance and motion cues.

### 3.6. Implementation Details

**Training Settings.** We implement our model in PyTorch [67], accelerated by an NVIDIA RTX TITAN. All the inputs are uniformly resized to  $352 \times 352$ . To enhance the stability and generalizability of the learning algorithm, we employ the multi-scale ( $\{0.75, 1, 1.25\}$ ) training strategy [27] in the training phase. Based on experiments in Tab. 4,  $N=4$  (the number of BPM) achieves the best performance. We utilize the stochastic gradient descent (SGD) algorithm to optimize the entire network, with a momentum of 0.9, learning rate of  $2e^{-3}$ , and weight decay of  $5e^{-4}$ .

**Testing Settings and Runtime.** Given a frame along with its motion map, we resize them to  $352 \times 352$  and feed them into the corresponding branch. Similar to [56, 99, 127], We employ the conditional random field (CRF) [42] post-processing technique for a fair comparison. The inference time of our method is 0.08s per frame, regardless of flow generation and CRF post-processing.

Table 1: Video object segmentation (VOS) performance of our *FSNet*, compared with 14 SOTA unsupervised and seven semi-supervised models on DAVIS<sub>16</sub> [71] validation set. ‘w/ Flow’: the optical flow algorithm is used. ‘w/ CRF’: conditional random field [42] is used for post-processing. The best scores are marked in **bold**.

Metric	Unsupervised														Semi-supervised								
	<i>FSNet</i> (Ours)	MAT [127]	AGNN [91]	AnDiff [117]	COSNet [56]	AGS [99]	EpO+ [19]	MOA [80]	LSMO [86]	ARP [41]	LVO [85]	LMP [84]	SFL [16]	ELM [44]	FST [66]	CFBI [118]	AGA [37]	RGM [107]	FEEL [89]	FA [15]	OS [7]	MSK [69]	
w/ Flow	✓	✓																					
w/ CRF	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓			✓		✓	
Mean- $\mathcal{J}$ ↑	<b>83.4</b>	82.1	82.4	80.7	81.7	80.5	79.7	80.6	77.2	78.2	76.2	75.9	70.0	67.4	61.8	55.8	<b>85.3</b>	81.5	81.5	81.1	82.4	79.8	79.7
Mean- $\mathcal{F}$ ↑	<b>83.1</b>	83.0	80.7	79.1	80.5	79.5	77.4	75.5	77.4	75.9	70.6	72.1	65.9	66.7	61.2	51.1	<b>86.9</b>	82.2	82.0	82.2	79.5	80.6	75.4

Table 2: Video salient object detection (VSOD) performance of our *FSNet*, compared with 13 SOTA models on several VSOD datasets. ‘†’ denotes that we generate non-binary saliency maps without CRF [42] for fair comparison. ‘N/A’ means the results are not available.

Model	DAVIS <sub>16</sub> [71]				MCL [39]				FBMS [64]				DAVSOD <sub>19</sub> -Easy35 [22]				
	$S_\alpha$ ↑	$E_\xi^{max}$ ↑	$F_\beta^{max}$ ↑	$\mathcal{M}$ ↓	$S_\alpha$ ↑	$E_\xi^{max}$ ↑	$F_\beta^{max}$ ↑	$\mathcal{M}$ ↓	$S_\alpha$ ↑	$E_\xi^{max}$ ↑	$F_\beta^{max}$ ↑	$\mathcal{M}$ ↓	$S_\alpha$ ↑	$E_\xi^{max}$ ↑	$F_\beta^{max}$ ↑	$\mathcal{M}$ ↓	
2018	MBN [51]	0.887	0.966	0.862	0.031	0.755	0.858	0.698	0.119	0.857	0.892	0.816	0.047	0.646	0.694	0.506	0.109
	FGRN [49]	0.838	0.917	0.783	0.043	0.709	0.817	0.625	0.044	0.809	0.863	0.767	0.088	0.701	0.765	0.589	0.095
	SCNN [82]	0.761	0.843	0.679	0.077	0.730	0.828	0.628	0.054	0.794	0.865	0.762	0.095	0.680	0.745	0.541	0.127
	DLVS [96]	0.802	0.895	0.721	0.055	0.682	0.810	0.551	0.060	0.794	0.861	0.759	0.091	0.664	0.737	0.541	0.129
	SCOM [13]	0.814	0.874	0.746	0.055	0.569	0.704	0.422	0.204	0.794	0.873	0.797	0.079	0.603	0.669	0.473	0.219
2019~2020	RSE [113]	0.748	0.878	0.698	0.063	0.682	0.657	0.576	0.073	0.670	0.790	0.652	0.128	0.577	0.663	0.417	0.146
	SRP [17]	0.662	0.843	0.660	0.070	0.689	0.812	0.646	0.058	0.648	0.773	0.671	0.134	0.575	0.655	0.453	0.146
	MESO [112]	0.718	0.853	0.660	0.070	0.477	0.730	0.144	0.102	0.635	0.767	0.618	0.134	0.549	0.673	0.360	0.159
	LTISI [10]	0.876	0.957	0.850	0.034	0.768	0.872	0.667	0.044	0.805	0.871	0.799	0.087	0.695	0.769	0.585	0.106
	SPD [52]	0.783	0.892	0.763	0.061	0.685	0.794	0.601	0.069	0.691	0.804	0.686	0.125	0.626	0.685	0.500	0.138
	SSAV [22]	0.893	0.948	0.861	0.028	0.819	0.889	0.773	0.026	0.879	0.926	0.865	0.040	0.755	0.806	0.659	0.084
	RCR [115]	0.886	0.947	0.848	0.027	0.820	0.895	0.742	0.028	0.872	0.905	0.859	0.053	0.741	0.803	0.653	0.087
	PCSA [26]	0.902	0.961	0.880	0.022	N/A	N/A	N/A	N/A	0.868	0.920	0.837	<b>0.040</b>	0.741	0.793	0.656	0.086
	<i>FSNet</i> <sup>†</sup> (Ours)	<b>0.920</b>	<b>0.970</b>	<b>0.907</b>	<b>0.020</b>	<b>0.864</b>	<b>0.924</b>	<b>0.821</b>	<b>0.023</b>	<b>0.890</b>	<b>0.935</b>	<b>0.888</b>	0.041	<b>0.773</b>	<b>0.825</b>	<b>0.685</b>	<b>0.072</b>

## 4. Experiments

### 4.1. UVOS and VSOD

**Datasets.** We evaluate the proposed model on four widely used VOS datasets. DAVIS<sub>16</sub> [71] is the most popular of these, and consists of 50 (30 training and 20 validation) high-quality and densely annotated video sequences. MCL [39] contains 9 videos and is mainly used as testing data. FBMS [64] includes 59 natural videos, in which 29 sequences are used as the training set and 30 are for testing. SegTrack-V2 [48] is one of the earliest VOS dataset, and consists of 13 clips. In addition, DAVSOD<sub>19</sub> [22] was specifically designed for the VSOD task. It is the most challenging visual attention consistent VSOD dataset with high-quality annotations and diverse attributes.

**Metrics.** We adopt six standard metrics including: mean region similarity ( $\mathcal{J}$ ) [71], mean contour accuracy ( $\mathcal{F}$ ) [71], structure-measure ( $S_\alpha$ ,  $\alpha=0.5$ ) [11], maximum enhanced-alignment measure ( $E_\xi^{max}$ ) [21], maximum F-measure ( $F_\beta^{max}$ ,  $\beta^2=0.3$ ) [2], and MAE ( $\mathcal{M}$ ) [70].

**Training.** Following a similar multi-task training setup as [50], we divide our training procedure into three steps: (i) We first use a well-known static saliency dataset DUTS [90] to train the spatial branch to avoid over-fitting, like in [22, 81, 96], (ii) We then train the temporal branch on the generated optical flow maps, and (iii) We finally load the weights

pretrained on two sub-tasks into the spatial and temporal branches, and thus, the whole network is end-to-end trained on the training set of DAVIS<sub>16</sub> (30 clips) and FBMS (29 clips). Last step takes about 4 hours and converges after 20 epochs with a batch size of 8.

**Testing.** We follow the standard benchmarks [22, 71] to test our model on the validation set (20 sequences) of DAVIS<sub>16</sub>, the test set of FBMS (30 clips), the test set (Easy35 split) of DAVSOD<sub>19</sub> (35 clips), the whole of MCL (9 clips), and the whole of SegTrack-V2 (13 clips).

**Evaluation on DAVIS<sub>16</sub>.** As shown in Tab. 1, we compare our *FSNet* with 14 SOTA UVOS models on the DAVIS<sub>16</sub> public leaderboard. We also compare it with seven recent semi-supervised approaches as reference. For fair comparison, we use a threshold of 0.5 to generate the final binary maps, as recommended by [117]. Our *FSNet* outperforms the best model (AAAI’20-MAT [127]) by a margin of 2.4% in  $\mathcal{F}$  and 1.0% in  $\mathcal{J}$ , achieving the new SOTA performance. Notably, the proposed UVOS model also outperforms the semi-supervised model (e.g., AGA [37]), even though it utilizes the first GT mask as the reference of object location.

We also compare *FSNet* against 13 SOTA VSOD models. We obtain the non-binary saliency maps<sup>5</sup> from the standard benchmark [22]. This can be seen from Tab. 2, our method consistently outperforms all other models since

<sup>5</sup>Note that all compared maps in VSOD, including ours, are non-binary.

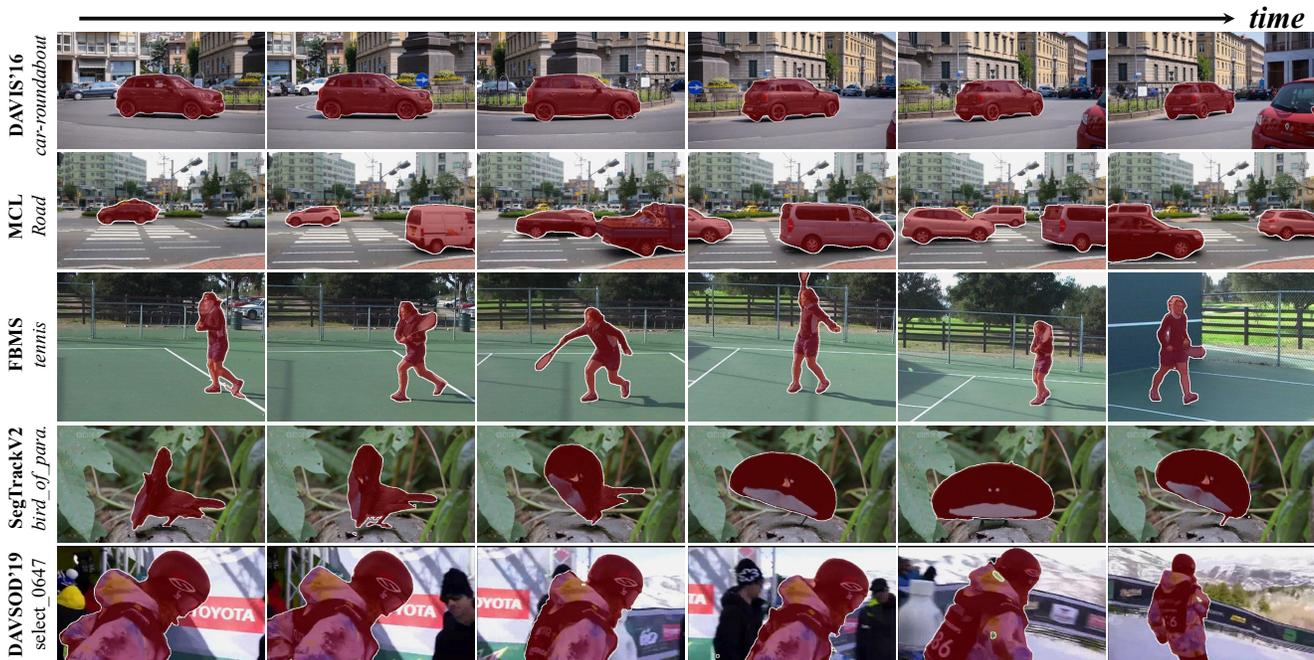


Figure 6: Qualitative results on five datasets, including DAVIS<sub>16</sub> [71], MCL [39], FBMS [64], SegTrack-V2 [48], and DAVSOD<sub>19</sub> [22].

2018, on all metrics. In particular, for the  $S_\alpha$  and  $F_\beta^{max}$  metrics, our method improves the performance by  $\sim 2.0\%$  compared with the best AAAI’20-PCAS [26] model.

**Evaluation on MCL.** This dataset has fuzzy object boundaries in the low-resolution frames, due to fast object movements. Therefore, the overall performance is lower than on DAVIS<sub>16</sub>. As shown in Tab. 2, our method still stands out in these extreme circumstances, with a 3.0~8.0% increase in all metrics compared with ICCV’19-RCR [115] and CVPR’19-SSAV [22].

**Evaluation on FBMS.** This is one of the most popular VOS datasets with diverse attributes, such as interacting objects and dynamic backgrounds, and no per-frame annotation. As shown in Tab. 2, our model achieves competitive performance in terms of  $\mathcal{M}$ . Further, compared to the previous best-performing SSAV [22], it obtains improvements in other metrics, including  $S_\alpha$  (0.890 vs. SSAV=0.879) and  $F_\xi^{max}$  (0.935 vs. SSAV=0.926), making it more suitable to the human visual system (HVS) as mentioned in [11, 21].

**Evaluation on SegTrack-V2.** This is the earliest VOS dataset from the traditional era. Thus, only a limited number of deep UVOS models have been tested on it. We only compare our *FSNet* against the top-3 models: AAAI’20-PCAS [26] ( $S_\alpha=0.866$ ), ICCV’19-RCR [115] ( $S_\alpha=0.842$ ), and CVPR’19-SSAV [22] ( $S_\alpha=0.850$ ). Our method achieves the best performance ( $S_\alpha=0.870$ ).

**Evaluation on DAVSOD<sub>19</sub>.** Most of the video sequences in DAVSOD<sub>19</sub> are similar to those in the challenging DAVIS<sub>16</sub> dataset. It also contains a large amount of sin-

gle (salient) objects. We find that *FSNet* outperforms all the reported algorithms. Compared with the current best solution (*i.e.*, AAAI’20-PCAS), our model achieves large improvements of 3.2% in terms of  $S_\alpha$ .

**Qualitative Results.** Some qualitative results on the five datasets are shown in Fig. 6, validating that our method achieves high-quality UVOS and VSOD results. As can be seen in the 1<sup>st</sup> row, the red car in the bottom-right corner moves slowly, so it does not get noticed. However, as our full-duplex strategy model considers both appearance and motion bidirectionally, it can automatically predict the smaller car in the center of the video. Overall, for these challenging situations, *e.g.*, dynamic background (1<sup>st</sup> & 5<sup>th</sup> rows), occlusion (2<sup>nd</sup> row), fast-motion (3<sup>rd</sup> row), and deformation (4<sup>th</sup> row), our model is able to infer the real target object(s) with fine-grained details. From this point of view, we demonstrate that *FSNet* is a general framework for both UVOS and VSOD tasks.

## 4.2. Ablation Study

### 4.2.1 Stimulus Selection

We explore the influence of different stimuli (appearance only vs. motion only) in our framework. We use only video frames or motion maps (using [33]) to train the ResNet-50 [28] backbone together with the proposed decoder block (see § 3.4). As shown in Tab. 3, *Mo.* performs slightly better than *App.* in terms of  $S_\alpha$  on DAVIS<sub>16</sub>, which suggests that the “optical flow” setting can learn more visual cues than “video frames”. Nevertheless, *App.* outperforms *Mo.*

Table 3: Ablation studies (§ 4.2.1, § 4.2.2, & § 4.2.3) for our components on DAVIS<sub>16</sub> and MCL. We set  $N = 4$  for BPM.

	Component Settings				DAVIS <sub>16</sub>		MCL	
	Appearance	Motion	RCAM	BPM	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
<i>App.</i>	✓				0.834	0.047	0.754	0.038
<i>Mo.</i>		✓			0.858	0.039	0.763	0.053
Vanilla	✓	✓			0.871	0.035	0.776	0.046
Rel.	✓	✓	✓		0.900	0.025	0.833	0.031
Bi-Purf.	✓	✓		✓	0.904	0.026	0.855	0.023
<i>FSNet</i>	✓	✓	✓	✓	<b>0.920</b>	<b>0.020</b>	<b>0.864</b>	<b>0.023</b>

Table 4: Ablation study for the number ( $N$ ) of BPMs on DAVIS<sub>16</sub> [71] and MCL [39], with the focus on parameter and FLOPs of BPMs, and runtime of *FSNet*.

	Param.	FLOPs	Runtime	DAVIS <sub>16</sub>		MCL	
	(M)	(G)	(s/frame)	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
$N = 0$	0.000	0.000	0.03	0.900	0.025	0.833	0.031
$N = 2$	0.507	1.582	0.05	0.911	0.026	0.843	0.028
$N = 4$	1.015	3.163	0.08	<b>0.920</b>	<b>0.020</b>	<b>0.864</b>	<b>0.023</b>
$N = 6$	1.522	4.745	0.10	0.918	0.023	0.863	0.023
$N = 8$	2.030	6.327	0.13	0.920	0.023	0.864	0.023

in  $\mathcal{M}$  metric on MCL. This motivates us to explore how to effectively use appearance and motion cues simultaneously.

#### 4.2.2 Effectiveness of RCAM

To validate the effectiveness of our RCAM (Rel.), we replace our fusion strategy with the vanilla fusion (Vanilla) using a concatenate operation followed by a convolutional layer to fuse two modalities. As expected (Tab. 3), the proposed Rel. performs consistently better than the vanilla fusion strategy on both DAVIS<sub>16</sub> and MCL. We would like to point out that our RCAM has two important properties: (i) it enables mutual correction and attention, and (ii) it can alleviate error propagation within a network to an extent due to the mutual correction and bidirectional interaction.

#### 4.2.3 Effectiveness of BPM

To illustrate the effectiveness of the BPM (with  $N = 4$ ), we derive two different models: Rel. and *FSNet*, referring to the framework *without* or *with* BPM. We observe that the model with BPM gains 2.0~3.0% than the one without BPM, according to the statistics in Tab. 3. We attribute this improvement to BPM’s introduction of an interlaced decremental connection, which enables it to effectively fuse the different signals. Similarly, we remove the RCAM and derive another pair of settings (Vanilla & Bi-Purf.) to test the robustness of our BPM. The results show that even using the bidirectional vanilla fusion strategy (Bi-Purf.) can still enhance the stability and generalization of the model. This benefits from the purification forward process and re-calibration backward process in the whole network.

#### 4.2.4 Number of Cascaded BPMs

Intuitively, more cascaded BPMs should lead to better boost performance. This is investigated and the evaluation results

Table 5: Ablation study for the *simplex* and *full-duplex* strategies on DAVIS<sub>16</sub> [71] and MCL [39]. We set  $N = 4$  for BPM.

	Direction Setting		DAVIS <sub>16</sub>		MCL	
	RCAM	BPM	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
simplex	<i>App.</i> $\Rightarrow$ <i>Mo.</i>	$(App. + Mo.) \Rightarrow Mo.$	0.896	0.026	0.816	0.038
	<i>App.</i> $\Rightarrow$ <i>Mo.</i>	$(App. + Mo.) \Leftarrow Mo.$	0.902	0.025	0.832	0.031
	<i>App.</i> $\Leftarrow$ <i>Mo.</i>	$(App. + Mo.) \Rightarrow Mo.$	0.891	0.029	0.806	0.039
	<i>App.</i> $\Leftarrow$ <i>Mo.</i>	$(App. + Mo.) \Leftarrow Mo.$	0.897	0.028	0.840	0.028
full-dup.	<i>App.</i> $\Leftrightarrow$ <i>Mo.</i>	$(App. + Mo.) \Leftrightarrow Mo.$	<b>0.920</b>	<b>0.020</b>	<b>0.864</b>	<b>0.023</b>

are shown in Tab. 4, where  $N = \{0, 2, 4, 6, 8\}$ . Note that  $N = 0$  means that **NO** BPM is used. Clearly, as can be seen from Fig. 2 and Tab. 4, more BPMs leads to better results, but the performance reaches saturation after  $N = 4$ . Further, too many BPMs (*i.e.*,  $N > 4$ ) will cause high model-complexity and may increase the risk of over-fitting. As a trade-off, we use  $N = 4$  throughout our experiments.

#### 4.2.5 Effectiveness of Full-Duplex Strategy

To investigate the effectiveness of the RCAM and BPM modules with the full-duplex strategy, we study two unidirectional (simplex, see Fig. 4 & Fig. 5) variants of our model. In Tab. 5, the symbols  $\Rightarrow$ ,  $\Leftarrow$ , and  $\Leftrightarrow$  indicate the feature transmission directions in the designed RCAM or BPM. Specifically, *App.*  $\Leftarrow$  *Mo.* indicates that the attention vector in the optical flow branch weights the features in the appearance branch, and vice versa.  $(App. + Mo.) \Leftarrow Mo.$  indicates that motion cues are used to guide the fused features extracted from both appearance and motion. The comparison results show that our elaborately designed modules (RCAM and BPM) jointly cooperate in a full-duplex fashion and outperform all simplex (*unidirectional*) settings.

## 5. Conclusion

We explore a simple yet efficient full-duplex strategy network (*FSNet*) that fully leverages the complementarity of appearance and motion cues to address the video object segmentation problem. This architecture consists of a relational cross-attention module (RCAM) and an efficient bidirectional purification module (BPM). The former is used to abstract features from a dual-modality, while the latter is utilized to re-calibrate inaccurate features step-by-step. In the BPM, the interlaced decremental connection is critical for broadcasting high-level coarse features to low-level fine-grained features. We thoroughly validate each module of our *FSNet*, providing several interesting findings. Finally, *FSNet* acts as a unified solution significantly advancing the SOTA of both VOS and VSOD. How to learn short-/long-term in an efficient Transformer-like [100, 129] scheme under the complicated scenarios seems to be interesting work.

**Acknowledgments.** This work was supported by the NSFC (No. 62176169, 61703077) and China Postdoctoral Science Foundation Funded Project (No. 2020M682829).

## References

- [1] Alexey Abramov, Karl Pauwels, Jeremie Papon, Florentin Wörgötter, and Babette Dellen. Depth-supported real-time video segmentation with the kinect. In *IEEE WACV*, pages 457–464, 2012. [1](#)
- [2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009. [6](#)
- [3] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In *ICLR*, 2016. [2](#)
- [4] Dinesh Bharadia, Emily McMilin, and Sachin Katti. Full duplex radios. In *ACM SIGCOMM*, pages 375–386, 2013. [2](#)
- [5] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *ECCV*, 2020. [1](#)
- [6] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, pages 282–295, 2010. [2](#)
- [7] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *IEEE CVPR*, pages 221–230, 2017. [2](#), [6](#)
- [8] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019. [1](#)
- [9] Bowen Chen, Huan Ling, Xiaohui Zeng, Gao Jun, Ziyue Xu, and Sanja Fidler. Scribblebox: Interactive annotation framework for video object segmentation. In *ECCV*, 2020. [1](#)
- [10] Chenglizhao Chen, Guotao Wang, Chong Peng, Xiaowei Zhang, and Hong Qin. Improved robust video saliency detection based on long-term spatial-temporal information. *IEEE TIP*, 29:1090–1100, 2019. [3](#), [6](#)
- [11] Ming-Ming Chen and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 2021. [6](#), [7](#)
- [12] Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. In *IEEE CVPR*, pages 9384–9393, 2020. [1](#)
- [13] Yuhuan Chen, Wenbin Zou, Yi Tang, Xia Li, Chen Xu, and Nikos Komodakis. Scom: Spatiotemporal constrained optimization for salient object detection. *IEEE TIP*, 27(7):3345–3357, 2018. [6](#)
- [14] Zixuan Chen, Chunchao Guo, Jianhuang Lai, and Xiaohua Xie. Motion-appearance interactive encoding for object segmentation in unconstrained videos. *IEEE TCSVT*, 2019. [1](#)
- [15] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *IEEE CVPR*, pages 7415–7424, 2018. [2](#), [6](#)
- [16] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *IEEE ICCV*, pages 686–695, 2017. [1](#), [3](#), [6](#)
- [17] Runmin Cong, Jianjun Lei, Huazhu Fu, Fatih Porikli, Qingming Huang, and Chunping Hou. Video saliency detection via sparsity-based reconstruction and propagation. *IEEE TIP*, 28(10):4819–4831, 2019. [6](#)
- [18] Mingyu Ding, Zhe Wang, Bolei Zhou, Jianping Shi, Zhiwu Lu, and Ping Luo. Every Frame Counts: Joint Learning of Video Segmentation and Optical Flow. In *AAAI*, pages 10713–10720, 2020. [1](#)
- [19] Muhammad Faisal, Ijaz Akhter, Mohsen Ali, and Richard Hartley. Exploiting geometric constraints on dense trajectories for motion saliency. In *IEEE WACV*, 2020. [6](#)
- [20] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014. [2](#)
- [21] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, 6, 2021. [6](#), [7](#)
- [22] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *IEEE CVPR*, pages 8554–8564, 2019. [3](#), [6](#), [7](#)
- [23] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *IEEE CVPR*, pages 1846–1853, 2012. [2](#)
- [24] Fabio Galasso, Roberto Cipolla, and Bernt Schiele. Video segmentation with superpixels. In *ACCV*, pages 760–774, 2012. [2](#)
- [25] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph-based video segmentation. In *IEEE CVPR*, pages 2141–2148, 2010. [2](#)
- [26] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Mingming Cheng, and Shaoping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, 2020. [6](#), [7](#)
- [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE TPAMI*, 37(9):1904–1916, 2015. [5](#)
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. [3](#), [4](#), [7](#)
- [29] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Interactive video object segmentation using global and local transfer modules. In *ECCV*, 2020. [1](#)
- [30] Ping Hu, Gang Wang, Xiangfei Kong, Jason Kuen, and Yap-Peng Tan. Motion-guided cascaded refinement network for video object segmentation. *IEEE TPAMI*, pages 1400–1409, 2020. [1](#), [2](#)
- [31] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *ECCV*, pages 786–802, 2018. [3](#)
- [32] Xuhua Huang, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Fast video object segmentation with temporal aggre-

- gation network and dynamic template matching. In *IEEE CVPR*, pages 8879–8889, 2020. 1
- [33] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE CVPR*, pages 2462–2470, 2017. 1, 3, 7
- [34] Suyog Dutt Jain and Kristen Grauman. Click carving: Segmenting objects in video with point clicks. In *IJCV*, 2016. 1
- [35] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *IEEE CVPR*, pages 2117–2126, 2017. 1, 2
- [36] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Debesh Jha, Huazhu Fu, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *MICCAI*, 2021. 1
- [37] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation. In *IEEE CVPR*, pages 8953–8962, 2019. 2, 6
- [38] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for object tracking. In *IEEE CVPRW*, 2017. 1
- [39] Hansang Kim, Youngbae Kim, Jae-Young Sim, and Chang-Su Kim. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE TIP*, 24(8):2552–2564, 2015. 6, 7, 8
- [40] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141, 1987. 2
- [41] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *IEEE CVPR*, pages 7417–7425, 2017. 6
- [42] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, pages 109–117, 2011. 2, 5, 6
- [43] Meng Lan, Yipeng Zhang, Qinning Xu, and Lefei Zhang. E3SN: Efficient End-to-End Siamese Network for Video Object Segmentation. In *IJCAI*, pages 701–707, 2020. 1
- [44] Dong Lao and Ganesh Sundaramoorthi. Extending layered models to 3d motion. In *ECCV*, pages 435–451, 2018. 6
- [45] Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *BMVC*, volume 1, page 3, 2017. 3
- [46] Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE TIP*, 27(10):5002–5015, 2018. 3
- [47] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *IEEE ICCV*, pages 1995–2002, 2011. 2
- [48] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *IEEE ICCV*, pages 2192–2199, 2013. 2, 6, 7
- [49] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *IEEE CVPR*, pages 3243–3252, 2018. 3, 6
- [50] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *IEEE ICCV*, pages 7274–7283, 2019. 2, 3, 6
- [51] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*, pages 207–223, 2018. 3, 6
- [52] Y. Li, S. Li, C. Chen, A. Hao, and H. Qin. Accurate and robust video saliency detection via self-paced diffusion. *IEEE TMM*, 22(5):1153–1167, 2020. 3, 6
- [53] Yu Li, Zhuoran Shen, and Ying Shan. Fast video object segmentation using the global context module. In *ECCV*, 2020. 1
- [54] Fanqing Lin, Yao Chou, and Tony Martinez. Flow adaptive video object segmentation. *Image and Vision Computing*, 94:103864, 2020. 2
- [55] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE CVPR*, pages 2117–2125, 2017. 5
- [56] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *IEEE CVPR*, pages 3623–3632, 2019. 5, 6
- [57] Tianyang Ma and Longin Jan Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *IEEE CVPR*, pages 670–677, 2012. 2
- [58] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *IJRR*, 36(1):3–15, 2017. 1
- [59] Sabarinath Mahadevan, Ali Athar, Aljosa Osep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. Making a case for 3d convolutions for object segmentation in videos. In *BMVC*, 2020. 1
- [60] Jiayu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *IEEE CVPR*, pages 10366–10375, 2020. 1
- [61] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *IEEE ICCV*, pages 2394–2403, 2019. 3
- [62] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *IEEE CVPR*, pages 6819–6828, 2018. 2
- [63] Peter Ochs and Thomas Brox. Higher order motion models and spectral clustering. In *IEEE CVPR*, pages 614–621, 2012. 2
- [64] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE TPAMI*, 36(6):1187–1200, 2013. 6, 7
- [65] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *IEEE CVPR*, pages 6504–6512, 2017. 1
- [66] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *IEEE ICCV*, pages 1777–1784, 2013. 6

- [67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 5
- [68] Qinmu Peng and Yiu-Ming Cheung. Automatic video object segmentation based on visual and motion saliency. *IEEE TMM*, 21(12):3083–3094, 2019. 2
- [69] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *IEEE CVPR*, pages 2663–2672, 2017. 2, 6
- [70] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE CVPR*, pages 733–740, 2012. 6
- [71] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE CVPR*, pages 724–732, 2016. 2, 6, 7, 8
- [72] Federico Perazzi, Oliver Wang, Markus Gross, and Alexander Sorkine-Hornung. Fully connected object proposals for video segmentation. In *IEEE ICCV*, pages 3227–3234, 2015. 2
- [73] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shabbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *IEEE CVPR*, 2020. 1
- [74] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 5
- [75] Seonguk Seo, Joon-Young Lee, and Bohyung Han. UR-VOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In *ECCV*, 2020. 1
- [76] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, 2020. 1
- [77] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J Black. On the integration of optical flow and action recognition. In *GCPR*, pages 281–297, 2018. 4
- [78] Jianbo Shi et al. Good features to track. In *IEEE CVPR*, pages 593–600, 1994. 1
- [79] Jianbo Shi and Jitendra Malik. Motion segmentation and tracking using normalized cuts. In *IEEE ICCV*, pages 1154–1160, 1998. 2
- [80] Mennatullah Siam, Chen Jiang, Steven Lu, Laura Petrich, Mahmoud Gamal, Mohamed Elhoseiny, and Martin Jagersand. Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In *IEEE ICRA*, pages 50–56, 2019. 3, 6
- [81] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, pages 715–731, 2018. 2, 3, 6
- [82] Yi Tang, Wenbin Zou, Zhi Jin, Yuhuan Chen, Yang Hua, and Xia Li. Weakly supervised salient object detection with spatiotemporal cascade neural networks. *IEEE TCSVT*, 29(7):1973–1984, 2018. 3, 6
- [83] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 1
- [84] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *IEEE CVPR*, pages 3386–3394, 2017. 2, 6
- [85] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *IEEE ICCV*, pages 4481–4490, 2017. 1, 2, 3, 6
- [86] Pavel Tokmakov, Cordelia Schmid, and Karteek Alahari. Learning to segment moving objects. *IJCV*, 127(3):282–301, 2019. 6
- [87] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 2
- [88] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *IEEE CVPR*, pages 3899–3908, 2016. 2
- [89] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *IEEE CVPR*, pages 9481–9490, 2019. 2, 6
- [90] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *IEEE CVPR*, pages 136–145, 2017. 6
- [91] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J. Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *IEEE ICCV*, 2019. 6
- [92] Wenguan Wang, Jianbing Shen, Xuelong Li, and Fatih Porikli. Robust video object cosegmentation. *IEEE TIP*, 24(10):3137–3148, 2015. 2
- [93] Wenguan Wang, Jianbing Shen, Xiankai Lu, Steven CH Hoi, and Haibin Ling. Paying attention to video object pattern understanding. *IEEE TPAMI*, 2020. 2
- [94] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *IEEE CVPR*, pages 3395–3402, 2015. 2
- [95] Wenguan Wang, Jianbing Shen, Fatih Porikli, and Ruigang Yang. Semi-supervised video object segmentation with super-trajectories. *IEEE TPAMI*, 41(4):985–998, 2018. 1
- [96] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 27(1):38–49, 2017. 3, 6
- [97] Wenguan Wang, Jianbing Shen, Jianwen Xie, and Fatih Porikli. Super-trajectory for video segmentation. In *IEEE ICCV*, pages 1671–1679, 2017. 2
- [98] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE TPAMI*, 40(1):20–33, 2017. 3
- [99] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through

- visual attention. In *IEEE CVPR*, pages 3064–3074, 2019. [2](#), [5](#), [6](#)
- [100] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE ICCV*, 2021. [8](#)
- [101] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE CVPR*, 2021. [1](#)
- [102] Zheng Wang, Xinyu Yan, Yahong Han, and Meijun Sun. Ranking video salient object detection. In *ACM MM*, pages 873–881, 2019. [3](#)
- [103] Jun Wei, Shuhui Wang, and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020. [4](#)
- [104] Peisong Wen, Ruolin Yang, Qianqian Xu, Chen Qian, Qingming Huang, Runmin Cong, and Jianlou Si. DMVOS: Discriminative matching for real-time video object segmentation. In *ACM MM*, 2020. [1](#)
- [105] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: an alternative to the feature integration model for visual search. *J EXP PSYCHOL HUMAN*, 15(3):419, 1989. [2](#)
- [106] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *IEEE ICCV*, pages 7264–7273, 2019. [4](#)
- [107] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *IEEE CVPR*, pages 7376–7385, 2018. [2](#), [6](#)
- [108] Huaxin Xiao, Bingyi Kang, Yu Liu, Maojun Zhang, and Jiashi Feng. Online meta adaptation for fast video object segmentation. *IEEE TPAMI*, 42(5):1205–1217, 2019. [1](#)
- [109] Chenliang Xu, Caiming Xiong, and Jason J Corso. Streaming hierarchical video segmentation. In *ECCV*, pages 626–639, 2012. [2](#)
- [110] M. Xu, P. Fu, B. Liu, and J. Li. Multi-stream attention-aware graph convolution network for video salient object detection. *IEEE TIP*, 30:4183–4197, 2021. [3](#)
- [111] M. Xu, P. Fu, B. Liu, H. Yin, and J. Li. A novel dynamic graph evolution network for salient object detection. *Applied Intelligence*, 2021. [3](#)
- [112] Mingzhu Xu, Bing Liu, Ping Fu, Junbao Li, and Yu Hen Hu. Video saliency detection via graph clustering with motion energy and spatiotemporal objectness. *IEEE TMM*, 21(11):2790–2805, 2019. [6](#)
- [113] Mingzhu Xu, Bing Liu, Ping Fu, Junbao Li, Yu Hen Hu, and Shou Feng. Video salient object detection via robust seeds extraction and multi-graphs manifold propagation. *IEEE TCSVT*, 2019. [3](#), [6](#)
- [114] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018. [1](#), [2](#)
- [115] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *IEEE ICCV*, pages 7284–7293, 2019. [3](#), [6](#), [7](#)
- [116] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *IEEE ICCV*, pages 5188–5197, 2019. [1](#)
- [117] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *IEEE ICCV*, pages 931–940, 2019. [1](#), [6](#)
- [118] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020. [6](#)
- [119] Lih Zelnik-Manor and Michal Irani. Event-based analysis of video. In *IEEE CVPR*, volume 2, pages II–II, 2001. [1](#)
- [120] Kaihua Zhang, Long Wang, Dong Liu, Bo Liu, Qingshan Liu, and Zhu Li. Dual temporal memory network for efficient video object segmentation. In *ACM MM*, 2020. [1](#)
- [121] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *IEEE ICCV*, 2021. [3](#)
- [122] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *IEEE CVPR*, pages 6949–6958, 2020. [1](#)
- [123] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *ECCV*, pages 269–284, 2018. [4](#)
- [124] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE CVPR*, pages 2881–2890, 2017. [5](#)
- [125] Jia Zheng, Weixin Luo, and Zhixin Piao. Cascaded convlts using semantically-coherent data synthesis for video object segmentation. *IEEE Access*, 7:132120–132129, 2019. [2](#)
- [126] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE TIP*, 29:8326–8338, 2020. [1](#)
- [127] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, 2020. [1](#), [2](#), [5](#), [6](#)
- [128] Xiaofei Zhou, Zhi Liu, Chen Gong, and Wei Liu. Improving video saliency detection via localized estimation and spatiotemporal refinement. *IEEE TMM*, 20(11):2993–3007, 2018. [3](#)
- [129] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *IEEE CVPR*, pages 12647–12657, 2021. [8](#)