Computational Visual Media
https://doi.org/10.1007/s41095-021-0262-4 Vol. 9, No. 1, March 2023, 155–175

Research Article

Full-duplex strategy for video object segmentation

Ge-Peng Ji[1], Deng-Ping Fan[2] (✉), Keren Fu[3], Zhe Wu[4], Jianbing Shen[5], and Ling Shao[6]

Abstract Previous video object segmentation approaches mainly focus on simplex solutions linking appearance and motion, limiting effective feature collaboration between these two cues. In this work, we study a novel and efficient full-duplex strategy network (*FSNet*) to address this issue, by considering a better mutual restraint scheme linking motion and appearance allowing exploitation of cross-modal features from the fusion and decoding stage. Specifically, we introduce a relational cross-attention module (RCAM) to achieve bidirectional message propagation across embedding sub-spaces. To improve the model's robustness and update inconsistent features from the spatiotemporal embeddings, we adopt a bidirectional purification module after the RCAM. Extensive experiments on five popular benchmarks show that our *FSNet* is robust to various challenging scenarios (e.g., motion blur and occlusion), and compares well to leading methods both for video object segmentation and video salient object detection. The project is publicly available at https://github.com/GewelsJI/FSNet.

Keywords video object segmentation (VOS); video salient object detection (V-SOD); visual attention

1 Introduction

Over the past three years, social platforms have accumulated a large number of short videos. Analyzing these videos efficiently and intelligently has become a challenging issue today. Video object segmentation (VOS) [1, 2] is a fundamental technique in addressing this issue; its purpose is to delineate pixel-level moving object (*foreground object* or *target object*) masks in each frame. Besides video analysis, many other applications have also benefited from VOS, such as robotic manipulation [3], autonomous cars [4], video editing [5], action segmentation [6], optical flow estimation [7], medical diagnosis [8], interactive segmentation [9], referring VOS [10], and video captioning [11].

Recently, we have witnessed rapid development in video object understanding which exploits the relationships between frames' appearances [12, 13] and is motion-aware [14, 15]. Unfortunately, short-term dependency prediction [14, 15] generates unreliable estimates and suffers from common problems [16] (e.g., noise, deformation, and diffusion). In addition, the capability of appearance-based modelling, e.g., using recurrent neural networks (RNNs) [17, 18], is severely hindered by blurred foregrounds or cluttered backgrounds [19]. Such issues are prone to lead to accumulating inaccuracies and the propagation of spatiotemporal embeddings, which cause the problem of short-term feature drift [20].

As Fig. 1(a) shows, the *direction-independent strategy* [17, 21–24] is the earliest solution; it encodes appearance and motion features separately and fuses them directly. However, this intuitive approach implicitly causes feature conflicts since the motion and appearance features are derived from two distinctive modalities, extracted from separate branches. An alternative approach is to integrate them in a guided manner. As illustrated in Fig. 1(b), several recent methods opt for a *simplex strategy*

1 School of Computer Science, Wuhan University, Wuhan, China. E-mail: gepengai.ji@gmail.com.
2 Computer Vision Lab, ETH Zürich, ETF C113.2, Sternwartstrasse 7, 8092 Zürich, Switzerland. E-mail: dengpingfan@mail.nankai.edu.cn (✉).
3 College of Computer Science, Sichuan University, Chengdu, China. E-mail: fkrsuper@scu.edu.cn.
4 Peng Cheng Laboratory, Shenzhen, China. E-mail: wuzh02@pcl.ac.cn.
5 School of Computer Science, Beijing Institute of Technology, Beijing, China. E-mail: shenjianbingcg@gmail.com.
6 Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. E-mail: ling.shao@ieee.org.
Manuscript received: 2021-09-01; accepted: 2021-10-16

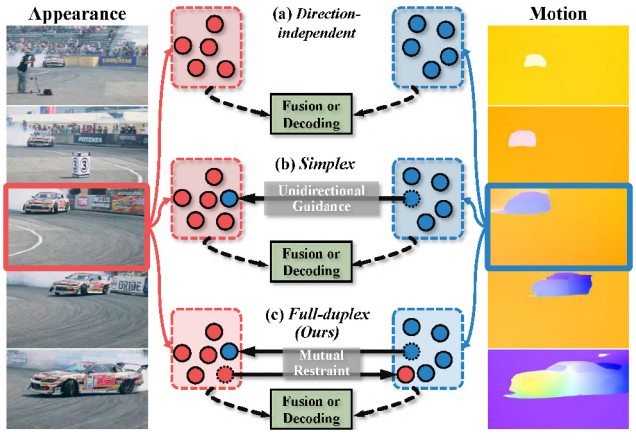

Fig. 1 Three strategies for embedding appearance and motion patterns before the fusion and decoding stage. (a) *Direction-independent* strategy [21] without information transmission. (b) *Simplex* strategy [25] with only unidirectional information transmission, with motion guiding appearance or vice versa. (c) Our *full-duplex* strategy with simultaneous bidirectional information transmission. This paper mainly focuses on discussing directional modelling (b, c) in the deep learning era.

[16, 25–30], which can be either appearance-based or motion-guided. Although these two strategies have achieved promising results, they both fail to consider *mutual restraints* between appearance and motion features that guide human visual attention allocation during dynamic observation, according to previous studies in cognitive psychology [31, 32] and computer vision [21, 33].

Intuitively, appearance and motion characteristics should be homogeneous to a certain degree for the same object within a short time. As Fig. 2 shows, the foreground region of appearance and motion intrinsically share correlated patterns of perception, including semantic structure and movement trends. Nevertheless, misguided knowledge in each individual modality, e.g., a static shadow under the chassis and small car in the background, produces inaccuracies

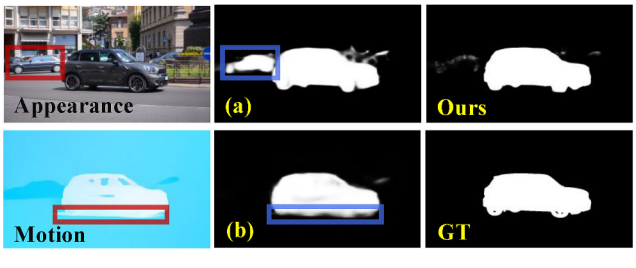

Fig. 2 Results using the *simplex* strategy (i.e., (a) appearance-refined motion and (b) motion-refined appearance) and our *full-duplex* strategy, which offers a collaborative way of using appearance and motion cues under *mutual restraints*. It thus provides more accurate structural details and alleviates the short-term feature drift issue [20].

during feature propagation. This easily taints the results (see blue boxes).

To address these challenges, we introduce a novel modality transmission strategy (*full-duplex* [34]) between spatial and temporal information, instead of embedding them individually. The proposed strategy is a bidirectional attention scheme across motion and appearance cues, which explicitly incorporates appearance and motion patterns in a unified framework, as depicted in Fig. 1(c). As seen in Fig. 2, our method visually performs better than ones with a simplex strategy in Figs. 1(a) and 1(b).

In fully investigating simplex and full-duplex strategies for our framework, we present the following contributions:

- a unified framework full-duplex strategy network (*FSNet*) for robust video object segmentation, which makes full use of spatiotemporal representations,
- a bidirectional interaction module, dubbed the relational cross-attention module (RCAM), to extract discriminative features from appearance and motion branches, which ensures mutual restraints between them, and to improve model robustness, a bidirectional purification module (BPM), which is equipped with an interlaced decremental connection to automatically update inconsistent features between the spatiotemporal embeddings, and finally
- a demonstration that our *FSNet* achieves favourable performance on five mainstream benchmarks; in particular our *FSNet* (N=4, CRF) outperforms the SOTA U-VOS model (i.e., MAT [25]) on the DAVIS$_{16}$ [35] leader board by a margin of 2.4% in terms of mean-$\mathcal{F}$ score, with less training data (13k for ours versus 16k for MAT).

As an extension of our ICCV-2021 paper [36], additions include:

- improved presentation, in particular Fig. 1, Fig. 2, and Fig. 7), and discussions (see Section 4.5).
- an investigation of the self-purification mode of BPM under our *FSNet* (see Fig. 9 and Section 4.5.4), the relation between RCAM and BPM (see Section 4.5.5), and training effectiveness with less data (see Section 4.5.3). The results further demonstrate the validity and rationality of our current design under various conditions.

- extra details of the backbone (see Section 3.6.1), evaluation metrics (see Section 4.2), prediction selection (see Section 4.5.1), and post-processing techniques (see Section 4.5.2).
- further results using different thresholds (PR curve in Fig. 6). Additional test results ($DAVSOD_{19}$-Normal25, $DAVSOD_{19}$-Difficult20) on a recent challenging dataset confirm that our framework is superior to existing SOTA models (see Table 3).

2 Related work

Depending on whether or not the first frame of ground truth is given, the VOS task can be divided into two scenarios, i.e., *semi-supervised* or *few-shot* and *unsupervised* or *zero-shot*. Typical semi-supervised VOS models include Refs. [37–40]. This paper studies the unsupervised setting [25, 41], leaving the semi-supervised setting as future work.

2.1 Unsupervised VOS

Although there are many works addressing the VOS task in a semi-supervised manner, supposing an object mask annotation is given in the first frame, other researchers have attempted to address the more challenging unsupervised VOS (U-VOS) problem. Early U-VOS models resort to low-level handcrafted features for heuristic segmentation inference, such as long sparse point trajectories [42, 43], object proposals [44, 45], saliency priors [46, 47], optical flow [26], or superpixels [48, 49]. 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 [50–52] have become popular due to their better ability to capture long-term dependencies and their use of deep learning. In this case, U-VOS is formulated as a recurrent modelling 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. [53] proposed to simply use motion patterns acquired from the video. However, their method cannot accurately segment objects between two similar consecutive frames since it relies heavily on the guidance of optical flow. To resolve this, several works [17, 23, 54] have integrated spatial and temporal features from parallel networks, which can be viewed as plain feature fusion from independent spatial and temporal branches with an implicit modelling strategy. Li et al. [55] proposed a multi-stage processing method to tackle U-VOS, which first utilizes a fixed appearance-based network to generate objectness and then feeds this into a motion-based bilateral estimator to segment the objects.

2.2 Attention-based VOS

The attention-based VOS task is closely related to U-VOS since it extracts attention attracting object(s) from a video clip. Traditional methods [56–59] first compute single-frame saliency based on various handcrafted static and motion features, and then conduct spatiotemporal optimization to preserve coherence across consecutive frames. Recent works [60–62] aim to learn a highly semantic representation and usually perform spatiotemporal detection end-to-end. Many schemes have been proposed that employ deep networks that consider temporal information, such as ConvLSTM [18, 50, 63], take optical-flows and adjacent-frames as input [29, 60], use 3D convolutional information [61, 62], or directly exploit temporally concatenated deep features [64]. Furthermore, long-term influences are often taken into account and combined with deep learning. Li et al. [65] proposed a key-frame strategy to locate representative high-quality video frames with salient objects [66, 67] and diffused their saliency to ill-detected non-key frames. Chen et al. [68] improved saliency detection by leveraging long-term spatiotemporal information, where high-quality *beyond-the-scope frames* are aligned with the current frames. Both types of information are fed to deep neural networks for classification. Besides considering how to better make use of temporal information, other researchers have attempted to address different problems in video salient object detection (V-SOD), such as reducing the data labelling requirements [69], developing semi-supervised approaches [70], or investigating relative saliency [71]. Fan et al. [18] recently introduced a V-SOD model equipped with a saliency shift-aware ConvLSTM, together with an attention-consistent V-SOD dataset with high-quality annotations. Zhao et al. [72] built a large-scale dataset with scribble annotation for weakly

supervised video salient object detection. They proposed an appearance–motion fusion module to aggregate the spatiotemporal features attentively.

3 Methodology

3.1 Overview

Suppose that a video clip contains T consecutive frames $\{\boldsymbol{A}^t\}_{t=1}^{T}$. We first utilize an optical flow field generator $\mathcal{H}$, i.e., FlowNet 2.0 [14], to generate $T-1$ optical flow maps $\{\boldsymbol{M}^t\}_{t=1}^{T-1}$, each of which is computed from two consecutive frames ($\boldsymbol{M}^t = \mathcal{H}[\boldsymbol{A}^t, \boldsymbol{A}^{t+1}]$). To ensure the inputs match, we discard the last frame in the pipeline. Thus, the proposed pipeline takes both appearance images $\{\boldsymbol{A}^t\}_{t=1}^{T-1}$ and their paired motion maps $\{\boldsymbol{M}^t\}_{t=1}^{T-1}$ as the input. First, pairs $\boldsymbol{M}^t$ and $\boldsymbol{A}^t$ pairs at frame t are fed to two independent ResNet-50 [73] branches (i.e., motion and appearance blocks in Fig. 3). We now omit the superscript t to simplify the notation. The appearance features $\{\mathcal{X}_k\}_{k=1}^{K}$ and motion features $\{\mathcal{Y}_k\}_{k=1}^{K}$ extracted from K layers are then sent to the relational cross-attention modules (RCAMs), which allows the network to embed spatiotemporal cross-modal features. Next, we employ N cascaded bidirectional purification modules (BPMs). The BPMs focus on distilling representative carriers from fused features $\{\boldsymbol{F}_k^n\}_{n=1}^{N}$ and motion-based features $\{\boldsymbol{G}_k^n\}_{n=1}^{N}$. Finally, the predictions (i.e., $\boldsymbol{S}_M^t$ and $\boldsymbol{S}_A^t$) at frame t are generated from two decoder blocks.

3.2 Relational cross-attention module

As discussed in Section 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) based on the channel-wise attention mechanism, which focuses on extracting cues from the two modalities and then using them to modulate 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 a standard ResNet-50 [73]. Specifically, for each level k, 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 1×1 conv layers, i.e., $\phi(x;\boldsymbol{W}_\phi)$ and $\theta(x;\boldsymbol{W}_\theta)$, with learnable parameters $\boldsymbol{W}_\phi$ and $\boldsymbol{W}_\theta$ respectively, generate two discriminative global descriptors. The sigmoid function $\sigma[x] = \mathrm{e}^x/(\mathrm{e}^x+1)$, $x \in \mathbb{R}$, is then applied to convert the final descriptors into the interval [0, 1], i.e., into a valid attention vector for channel weighting. Then, we perform an outer product $\otimes$ between $\mathcal{X}_k$ and $\sigma\left[\theta(\mathcal{V}_k^Y;\boldsymbol{W}_\theta)\right]$ to generate a candidate feature $\mathcal{Q}_k^X$, and vice versa, as Eqs. (1) and (2):

$$\mathcal{Q}_k^X = \mathcal{X}_k \otimes \sigma\left[\theta(\mathcal{V}_k^Y;\boldsymbol{W}_\theta)\right] \tag{1}$$

$$\mathcal{Q}_k^Y = \mathcal{Y}_k \otimes \sigma\left[\phi(\mathcal{V}_k^X;\boldsymbol{W}_\phi)\right] \tag{2}$$

Next, we combine $\mathcal{Q}_k^X$, $\mathcal{Q}_k^Y$, and lower-level fused feature $\mathcal{Z}_{k-1}$ for in-depth feature extraction. With element-wise addition $\oplus$, conducted in the corresponding k-th level block $\mathcal{B}_k[x]$ in ResNet-50, we finally obtain the fused features $\mathcal{Z}_k$ that contain comprehensive spatiotemporal correlations:

$$\mathcal{Z}_k = \mathcal{B}_k\left[\mathcal{Q}_k^X \oplus \mathcal{Q}_k^Y \oplus \mathcal{Z}_{k-1}\right] \tag{3}$$

where $k \in \{1 : K\}$ denotes different feature hierarchies in the backbone. Note that $\mathcal{Z}_0$ denotes the zero tensor. In our implementation, we use the top four feature pyramid levels, i.e., $K = 4$, as suggested by Refs. [74, 75].

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 model robustness. Following the standard in action recognition [76] and saliency detection [77], our bidirectional purification phase comprises N cascaded BPMs. As shown in Fig. 3, we first employ the feature allocator $\psi_{\{F,G\}}(x;\boldsymbol{W}_\psi^{\{F,G\}})$ to unify the feature representations from the previous stage:

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

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

Here, we consider a *bidirectional attention* scheme (see Fig. 5(c)) that contains two *simplex* strategies (see Figs. 5(a) and 5(b)) in the BPM. On the one hand, the motion features $\boldsymbol{G}_k^n$ contain temporal cues and can be used to enrich the fused features $\boldsymbol{F}_k^n$ by concatenation. On the other hand, we would

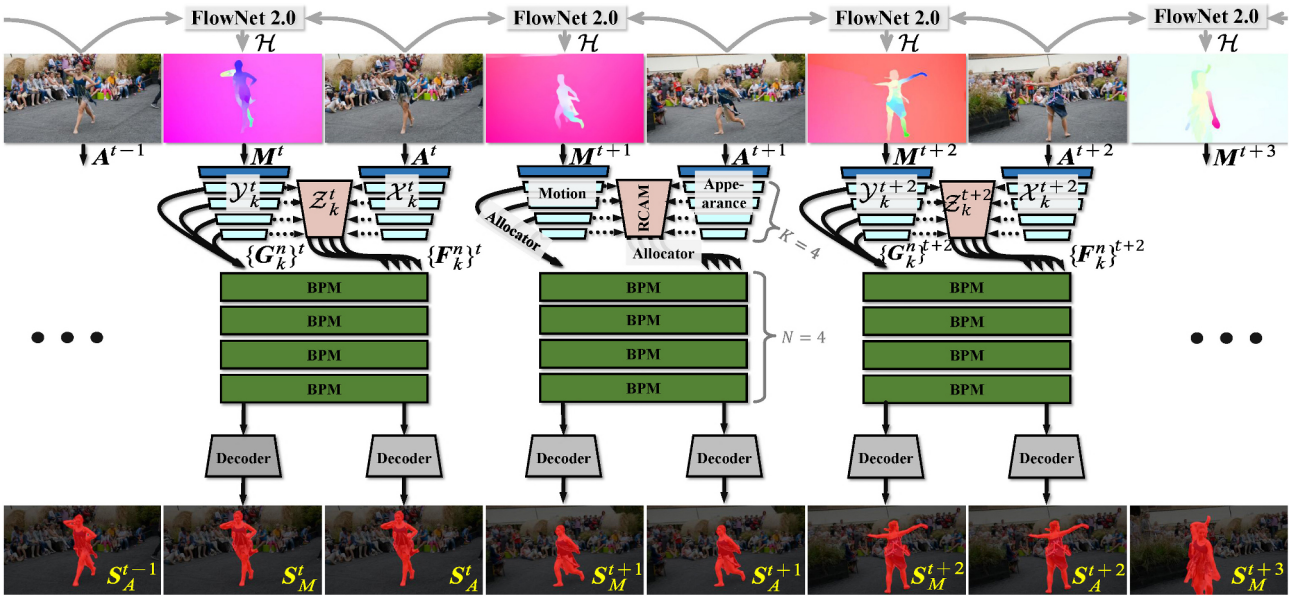

Fig. 3 Architecture of our *FSNet* for video object segmentation. The relational cross-attention module (RCAM) abstracts more discriminative representations linking motion and appearance cues using the full-duplex strategy. Then four bidirectional purification modules (BPM) are stacked to further resolve inconsistencies between the motion and appearance features. Finally, we utilize a decoder to generate our prediction.

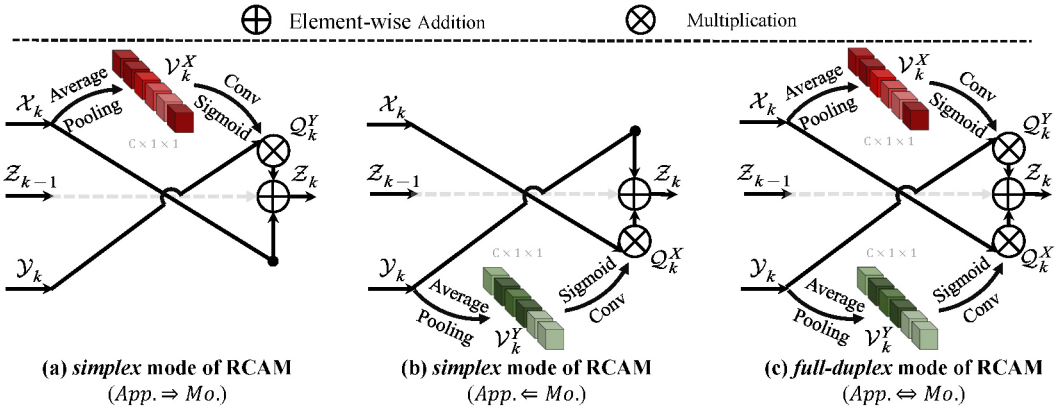

Fig. 4 Relational cross-attention module (RCAM) with *simplex* (a, b) and *full-duplex* (c) strategy.

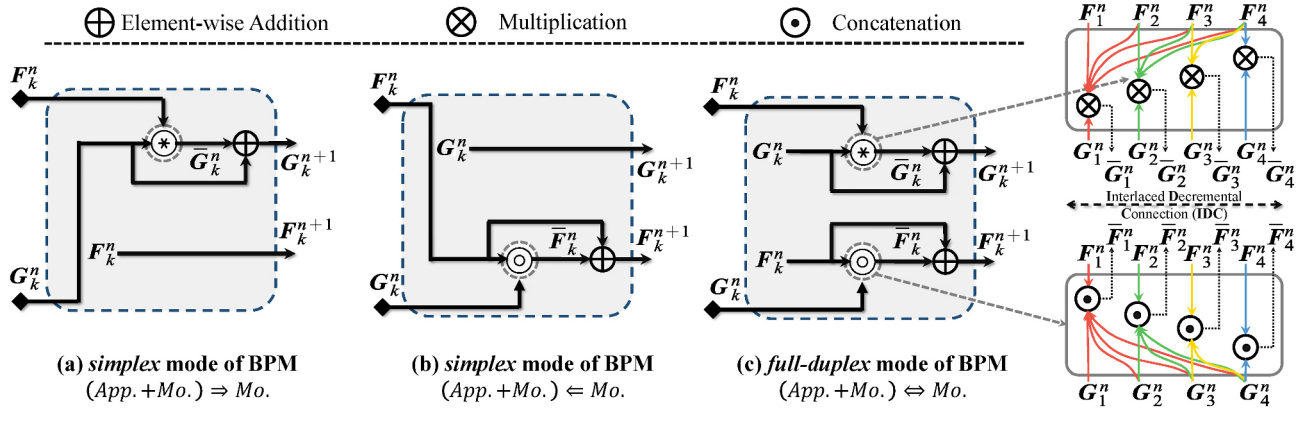

Fig. 5 Bidirectional purification module (BPM) with *simplex* and *full-duplex* strategy.

compress the distractors in the motion feature $\boldsymbol{G}_k^n$ by multiplying the fused features $\boldsymbol{F}_k^n$. Besides providing a robust feature representation, we introduce an efficient cross-modal fusion strategy in this scheme,

which broadcasts high-level, semantically strong features to low-level, semantically weak features by interlaced decremental connection (IDC) with a top–down pathway [78]. Specifically, as the first part, the spatiotemporal feature combination branch (see Fig. 5(b)) is formulated as

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

where $\mathcal{P}$ is an up-sampling operation followed by a 1×1 convolutional layer (conv) to reshape the candidate guidance to a consistent size with $\boldsymbol{F}_k^n$. Symbols $\oplus$ and $\bigcup$ respectively denote element-wise addition and concatenation operations with an IDC strategy, followed by a 1×1 conv with 32 filters. For instance, $\bar{\boldsymbol{G}}_2^n = \bigcup_{i=2}^{K=4}[\boldsymbol{F}_2^n, \mathcal{P}(\boldsymbol{G}_i^n)] = \boldsymbol{F}_2^n \odot \mathcal{P}(\boldsymbol{G}_2^n) \odot \mathcal{P}(\boldsymbol{G}_3^n) \odot \mathcal{P}(\boldsymbol{G}_4^n)$ when $k = 2$ and $K = 4$. For the other part, we formulate the temporal feature re-calibration branch (see Fig. 5(a)) as

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

where $\bigcap$ denotes element-wise multiplication with an IDC strategy, followed by a 1×1 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, $\boldsymbol{F}_k^N$ and $\boldsymbol{G}_k^N$, with a consistent number 32 of channels. We integrate a pyramid pooling module (PPM) [79] into each skip connection of the U-Net [80] 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:

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

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

Here, $\mathcal{UP}$ indicates the upsampling operation after the pyramid pooling layer, while $\odot$ concatenates two features. Then, a conv $\mathcal{C}$ is used to reduce the number of channels from 64 to 32. Lastly, we use a 1×1 conv with a single filter after the upstream output (i.e., $\hat{\boldsymbol{F}}_1^N$ and $\hat{\boldsymbol{G}}_1^N$), followed by a sigmoid activation function to generate the predictions $\boldsymbol{S}_A^t$ and $\boldsymbol{S}_M^t$ at frame t.

3.5 Learning objective

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

$$\mathcal{L}_{\text{bce}}(\boldsymbol{S}^t, \boldsymbol{G}^t) = -\sum_{(x,y)} [\boldsymbol{G}^t(x,y)\log(\boldsymbol{S}^t(x,y)) + (1-\boldsymbol{G}^t(x,y))\log(1-\boldsymbol{S}^t(x,y))] \tag{9}$$

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

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{bce}}(\boldsymbol{S}_A^t, \boldsymbol{G}^t) + \mathcal{L}_{\text{bce}}(\boldsymbol{S}_M^t, \boldsymbol{G}^t) \tag{10}$$

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

3.6 Implementation details

3.6.1 Backbone details

Without any modification, three standard ResNet-50 [73] backbones, removing the top-three layers (average pooling, fully-connected, and softmax layers), are adopted for the appearance branch, the motion branch, and the merging branch. Each ResNet-50 backbone results in $K = 4$ hierarchies following previous work [74]. After removing the top fully connected layers, the feature hierarchies ($\{\mathcal{X}_k, \mathcal{Y}_k, \mathcal{Z}_k\}, k \in \{2:5\}$) from shallow to deep are extracted from the conv2_3 ($k = 2$), conv3_4 ($k = 3$), conv4_6 ($k = 4$), and conv5_3 ($k = 5$) layers of the ResNet-50, respectively.

We also tried a two-branch setting, removing the merging branch and letting $\mathcal{Z}_k = \mathcal{Q}_k^X \oplus \mathcal{Q}_k^Y \oplus \mathcal{Z}_{k-1}$ instead of $\mathcal{Z}_k = \mathcal{B}_k[\mathcal{Q}_k^X \oplus \mathcal{Q}_x^Y \oplus \mathcal{Z}_{k-1}]$ in Eq. (3). Unfortunately, this leads to a 2.5% drop in performance for S_α on the DAVIS$_{16}$ [35] dataset. This is because the third merging branch can sequentially enhance and promote the spatiotemporal features from RCAMs, leading to better segmentation accuracy.

3.6.2 Training settings

We implemented our model in PyTorch [81], accelerated by an NVIDIA RTX TITAN GPU. All inputs were uniformly resized to 352×352. To enhance the stability and generalizability of our learning algorithm, we employed a multi-scale (i.e., $\{0.75, 1, 1.25\}$) training strategy [82] in the training phase. As can be seen from the experimental results in Table 5, the variant with N=4 (the number of BPMs) achieves the best performance. We utilized stochastic gradient descent (SGD) to optimize the entire network, with a momentum of 0.9, a learning

rate of $2e^{-3}$, and a weight decay of $5e^{-4}$. The learning rate decreased by 10% per 20 epochs.

3.6.3 Testing settings and runtime

Given a frame along with its motion map, we resized them to 352×352 and fed them into the corresponding branch. Following Refs. [25, 51, 83], we employed a conditional random field (CRF) [84] technique. The inference time of our method is 0.08 s/frame, ignoring flow generation and CRF post-processing.

4 Experiments

4.1 Experimental protocols

4.1.1 Datasets

We evaluated the proposed model on four widely used VOS datasets. DAVIS_{16} [35] is the most popular of these, and consists of 50 (30 training and 20 validation) high-quality and densely annotated video sequences. MCL [85] contains 9 videos and is mainly used as testing data. FBMS [86] includes 59 natural videos, in which 29 sequences are used for training and 30 for testing. SegTrack-V2 [44] is one of the earliest VOS datasets and consists of 13 clips. In addition, DAVSOD_{19} [18] was specifically designed for the V-SOD task. It is the most challenging visual attention consistent V-SOD dataset with high-quality annotation and diverse attributes.

4.1.2 Training

Following a similar multi-task training setup as Ref. [29], we divided our training procedure into three steps:

- We first adopted a well-known static saliency dataset DUTS [87] to train the spatial branch to avoid over-fitting, as in Refs. [18, 50, 60]. This step lasts for 50 epochs with a batch size of 8 under the same training settings as in Section 3.6.2.
- We then train the temporal branch on the generated optical flow maps. This step lasts for 50 epochs with a batch size of 8 under the same training settings as in Section 3.6.2.
- We finally loaded the weights pre-trained on the two sub-tasks into the spatial and temporal branches, and then, the whole network was end-to-end trained on the DAVIS_{16} (30 clips) and FBMS (29 clips) training sets. This last step took about 4 hours and converges after 20 epochs with a mini-batch size of 8 under same the training settings as in Section 3.6.2.

4.1.3 Testing

We used standard benchmarks [18, 35] to test our model on the validation set of DAVIS_{16} (20 clips), the test set of FBMS (30 clips), the test set (Easy35 split) of DAVSOD_{19} (35 clips), the whole of MCL (9 clips), and the whole of SegTrack-V2 (13 clips).

4.2 Evaluation metrics

We define a predicted map at frame t as $\boldsymbol{S}_A^t$ and its corresponding ground-truth mask as $\boldsymbol{G}^t$. The evaluation metrics used are given as follows.

4.2.1 Metrics for the U-VOS task

Following Ref. [20], we utilized two standard metrics to evaluate the U-VOS models. Note that all predicted maps are binary in in this task.

1. **Mean Region Similarity**. This metric, also called Jaccard similarity coefficient, is defined as the intersection-over-union of the predicted map and the ground-truth mask:
$$\mathcal{J} = \frac{|\boldsymbol{S}_A^t \cap \boldsymbol{G}^t|}{|\boldsymbol{S}_A^t \cup \boldsymbol{G}^t|} \tag{11}$$
where $|\cdot|$ is the number of pixels in an area. In all of our experiments, we also report the mean value, Mean-$\mathcal{J}$, following Ref. [20].
2. **Mean Contour Accuracy**. The contour accuracy metric we used is also called the contour $\mathcal{F}$-measure. We compute the contour-based precision and recall between the contour points of $c(\boldsymbol{S}_A^t)$ and $c(\boldsymbol{G}^t)$, where $c(\cdot)$ contains the extracted contour points of a mask. $\mathcal{F}$ is defined as
$$\mathcal{F} = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{12}$$
where $\text{Precision}_c = |c(\boldsymbol{S}_A^t) \cap c(\boldsymbol{G}^t)|/|c(\boldsymbol{S}_A^t)|$ and $\text{Recall}_c = |c(\boldsymbol{S}_A^t) \cap c(\boldsymbol{G}^t))|/|c(\boldsymbol{G}^t)|$. Following Ref. [20], we also report the mean value, Mean-$\mathcal{F}$ in all of our experiments.

4.2.2 Metrics for the V-SOD task

Unlike the U-VOS task, the predicted map can be non-binary in V-SOD benchmarking: see Section 4.3.1.

1. **Mean Absolute Error (MAE)**. This is a typical pixel-wise measure, defined as
$$\mathcal{M} = \frac{1}{WH} \sum_x^W \sum_y^H |\boldsymbol{S}_A^t(x,y) - \boldsymbol{G}^t(x,y)| \tag{13}$$
where W and H are the width and height of ground-truth $\boldsymbol{G}^t$, and (x, y) are the coordinates of a pixel in $\boldsymbol{G}^t$.

2. **Precision–Recall (PR) Curve**. Precision and recall [88–90] are defined as

$$\text{Precision} = \frac{|\boldsymbol{S}_A^t(T) \cap \boldsymbol{G}^t|}{|\boldsymbol{S}_A^t(T)|} \tag{14}$$

$$\text{Recall} = \frac{|\boldsymbol{S}_A^t(T) \cap \boldsymbol{G}^t|}{|\boldsymbol{G}^t|} \tag{15}$$

where $\boldsymbol{S}_A^t(T)$ is the binary mask obtained by directly thresholding the predicted map $\boldsymbol{S}_A^t$ with threshold $T \in [0, 255]$, and $|\cdot|$ is the total area of the mask inside the map. By varying T, a precision–recall curve can be obtained.

3. **Maximum F-measure**. This is defined as

$$F_\beta = \frac{(1+\beta^2)\text{Precision} \times \text{Recall}}{\beta^2\,\text{Precision} + \text{Recall}} \tag{16}$$

where β^2 is set to 0.3 to emphasise precision over recall, as recommended in Ref. [90]. We convert the non-binary predicted map into binary masks with threshold values from 0 to 255. In this paper, we report the maximum (i.e., $F_\beta^{\max}$) of a series of F-measure values calculated from the precision–recall curve by iterating over all thresholds.

4. **Maximum Enhanced-Alignment Measure**. As a recently proposed metric, E_ξ [88] is used to evaluate both local and global similarity between two binary maps. Its formulation is as Eq. (17):

$$E_\xi = \frac{1}{W \times H} \sum_x^W \sum_y^H \phi\left[\boldsymbol{S}_A^t(x,y), \boldsymbol{G}^t(x,y)\right] \tag{17}$$

where ϕ is the enhanced-alignment matrix. As for $F_\beta^{\max}$, we report the maximum E_ξ value computed over all the thresholds in all of our comparisons.

5. **Structure Measure**. Fan et al. [91] proposed a metric to measure the structural similarity between a non-binary saliency map and a ground-truth mask:

$$S_\alpha = (1-\alpha) S_{\text{o}}(\boldsymbol{S}_A^t, \boldsymbol{G}^t) + \alpha S_{\text{r}}(\boldsymbol{S}_A^t, \boldsymbol{G}^t) \tag{18}$$

where α balances the object-aware similarity S_{o} and region-aware similarity S_{r}. We use the default setting ($\alpha = 0.5$) suggested in Ref. [91].

4.3 U-VOS and V-SOD tasks

4.3.1 Evaluation on DAVIS$_{16}$ dataset

Table 1 compares results from our *FSNet* with 14 state of the art (SOTA) U-VOS models on the DAVIS$_{16}$ public leaderboard. We also compare it to 7 recent semi-supervised approaches as reference. We use a threshold of 0.5 to generate the final binary maps to ensure a fair comparison, as recommended by Ref. [20]. Our *FSNet* outperforms the best other model (AAAI'20-MAT [25]) by a margin of 2.4% in Mean-$\mathcal{F}$ and 1.0% in Mean-$\mathcal{J}$. Notably, the proposed U-VOS model also outperforms the semi-supervised models (e.g., AGA [99]), even though they utilize an initial ground-truth mask to locate objects.

We also compare *FSNet* against 13 SOTA V-SOD models. All compared maps in the V-SOD task, including ours, are non-binary. The non-binary saliency maps are obtained from the standard benchmark [18]. As can be seen from Table 2, our method consistently outperforms all other models since 2018 on all metrics. In particular, for the S_α and $F_\beta^{\max}$ metrics, our method improves the results by about 2% compared to the best AAAI'20-PCAS [108] model.

4.3.2 Evaluation on MCL dataset

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$_{16}$. As shown in Table 2, our method still stands out in these extreme circumstances, with a 3%–8% increase in all metrics compared to ICCV'19-RCR [69] and CVPR'19-SSAV [18].

Table 1 Video object segmentation (VOS) results of our *FSNet*, compared to 14 state-of-the-art unsupervised models and 7 semi-supervised models on the DAVIS$_{16}$ [35] validation set. *w*/Flow: the optical flow algorithm was used. *w*/CRF: a conditional random field [84] was used for post-processing. Best scores are marked in **bold**

Metric	Unsupervised																	Semi-supervised						
	FSNet (ours)		MAT [25]	AGNN [92]	AnDiff [20]	COS [83]	AGS [51]	EpO+ [93]	MOA [54]	LSMO [94]	ARP [95]	LVO [17]	LMP [53]	SFL [23]	ELM [96]	FST [97]	CFBI [98]	AGA [99]	RGM [100]	FEEL [101]	FA [102]	OS [103]	MSK [104]
w/Flow	✓	✓	✓					✓	✓	✓	✓	✓	✓	✓	✓	✓							✓
w/CRF	✓		✓	✓		✓	✓	✓	✓	✓		✓	✓						✓		✓		✓
Mean-$\mathcal{J}$	**83.4**	82.3	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	**85.3**	81.5	81.5	81.1	82.4	79.8	79.7
Mean-$\mathcal{F}$	83.1	**83.3**	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	**86.9**	82.2	82.0	82.2	79.5	80.6	75.4

Table 2 Video salient object detection (V-SOD) results for our *FSNet*, compared to 13 state-of-the-art models on three popular V-SOD datasets, including DAVIS_{16} [35], MCL [85], and FBMS [86]. † indicates that we generate non-binary saliency maps without CRF [84] to enable a fair comparison. N/A means results are not available

		DAVIS_{16} [35]				MCL [85]				FBMS [86]			
	Model	$S_\alpha \uparrow$	$E_\xi^{\max} \uparrow$	$F_\beta^{\max} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\max} \uparrow$	$F_\beta^{\max} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\max} \uparrow$	$F_\beta^{\max} \uparrow$	$\mathcal{M} \downarrow$
2018	MBN [55]	0.887	0.966	0.862	0.031	0.755	0.858	0.698	0.119	0.857	0.892	0.816	0.047
	FGRN [63]	0.838	0.917	0.783	0.043	0.709	0.817	0.625	0.044	0.809	0.863	0.767	0.088
	SCNN [70]	0.761	0.843	0.679	0.077	0.730	0.828	0.628	0.054	0.794	0.865	0.762	0.095
	DLVS [60]	0.802	0.895	0.721	0.055	0.682	0.810	0.551	0.060	0.794	0.861	0.759	0.091
	SCOM [105]	0.814	0.874	0.746	0.055	0.569	0.704	0.422	0.204	0.794	0.873	0.797	0.079
2019–2020	RSE [58]	0.748	0.878	0.698	0.063	0.682	0.657	0.576	0.073	0.670	0.790	0.652	0.128
	SRP [106]	0.662	0.843	0.660	0.070	0.689	0.812	0.646	0.058	0.648	0.773	0.671	0.134
	MESO [107]	0.718	0.853	0.660	0.070	0.477	0.730	0.144	0.102	0.635	0.767	0.618	0.134
	LTSI [68]	0.876	0.957	0.850	0.034	0.768	0.872	0.667	0.044	0.805	0.871	0.799	0.087
	SPD [65]	0.783	0.892	0.763	0.061	0.685	0.794	0.601	0.069	0.691	0.804	0.686	0.125
	SSAV [18]	0.893	0.948	0.861	0.028	0.819	0.889	0.773	0.026	0.879	0.926	0.865	0.040
	RCR [69]	0.886	0.947	0.848	0.027	0.820	0.895	0.742	0.028	0.872	0.905	0.859	0.053
	PCSA [108]	0.902	0.961	0.880	0.022	N/A	N/A	N/A	N/A	0.868	0.920	0.837	**0.040**
	FSNet†	**0.920**	**0.970**	**0.907**	**0.020**	**0.864**	**0.924**	**0.821**	**0.023**	**0.890**	**0.935**	**0.888**	0.041

4.3.3 Evaluation on FBMS dataset

This is one of the most popular VOS datasets with diverse attributes, such as interacting objects, dynamic backgrounds, and no per-frame annotation. As shown in Table 2, our model achieves competitive performance in terms of $\mathcal{M}$. Further, compared to the previous best-performing SSAV [18], it obtains improvements in other metrics, including $\mathcal{S}_\alpha$ (0.890 vs. SSAV=0.879) and $E_\xi^{\max}$ (0.935 vs. SSAV=0.926), making it more suited to the human visual system (HVS) as mentioned in Refs. [91, 109].

4.3.4 Evaluation on SegTrack-V2 dataset

This is the earliest VOS dataset from the traditional era. Thus, only a limited number of deep U-VOS models have been tested on it. We only compareour *FSNet* against the top-3 models: AAAI'20-PCAS [108] (S_α=0.866), ICCV'19-RCR [69] (S_α=0.842), and CVPR'19-SSAV [18] (S_α=0.850). Our method achieves the best results (S_α=0.870).

4.3.5 Evaluation on $DAVSOD_{19}$ dataset

Recently published DAVSOD_{19} [18] is the most challenging visual attention consistent V-SOD dataset with high-quality annotation and diverse attributes. It contains diverse challenging scenarios: the video sequences contain shifts in attention. DAVSOD_{19} is divided into three subsets, according to difficulty: DAVSOD_{19}-Easy35 (35 clips), DAVSOD_{19}-Normal25 (25 clips), and DAVSOD_{19}-Difficult20 (20 clips). Note that, in the saliency field, non-binary maps are required for evaluation; thus, we only report the results of *FSNet* without CRF post-processing when benchmarking the V-SOD task. We adopt the four metrics: S_α, $E_\xi^{\max}$, $F_\beta^{\max}$, and $\mathcal{M}$. To show the robustness of *FSNet*, in Table 3, we also make the first effort to benchmark all 11 SOTA models since 2018, at the three difficulty levels:

- **Easy35 subset**. Most of these video sequences are similar to those in the DAVIS_{16} dataset, with a large number having single video objects. We see that *FSNet* outperforms all reported algorithms across all metrics. As shown in Table 3, compared to the recent method PCSA, our model achieves a large improvements of 3.2% in S_α.
- **Normal25 subset**. This subset includes multiple moving salient objects. Thus, it is more difficult than traditional V-SOD datasets due to the attention shift phenomenon [18]. As hoped, *FSNet* still provides the best results, with significant improvementa, e.g., 6.4% for the $F_\beta^{\max}$ metric.
- **Difficult20 subset**. This is the most challenging of existing V-SOD datasets since it contains a large number of attention shifting sequences in cluttered scenarios. Unsurprisingly, the quality of results shown in Table 3 decrease dramatically for all the compared models (e.g., $F_\beta^{\max} \leqslant 0.5$). Even though our framework is not specifically designed for the V-SOD task, we still easily obtain the best performance in two metrics (e.g., $\mathcal{S}_\alpha$ and

Table 3 Benchmarking results of 13 state-of-the-art V-SOD models on three subsets of $DAVSOD_{19}$ [18]. † denotes that we generate non-binary saliency maps without CRF [84] to enable a fair comparison. N/A means results are not available

		$DAVSOD_{19}$-Easy35				$DAVSOD_{19}$-Normal25				$DAVSOD_{19}$-Difficult20			
	Model	$S_\alpha \uparrow$	$E_\xi^{\max} \uparrow$	$F_\beta^{\max} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\max} \uparrow$	$F_\beta^{\max} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\max} \uparrow$	$F_\beta^{\max} \uparrow$	$\mathcal{M} \downarrow$
2018	MBN [55]	0.646	0.694	0.506	0.109	0.597	0.665	0.436	0.127	0.561	0.635	0.352	0.140
	FGRN [63]	0.701	0.765	0.589	0.095	0.638	0.700	0.468	0.126	0.608	0.698	0.390	0.131
	SCNN [70]	0.680	0.745	0.541	0.127	0.589	0.685	0.425	0.193	0.533	0.677	0.345	0.234
	DLVS [60]	0.664	0.737	0.541	0.129	0.599	0.670	0.416	0.147	0.571	0.687	0.336	0.128
	SCOM [105]	0.603	0.669	0.473	0.219	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2019–2020	RSE [58]	0.577	0.663	0.417	0.146	0.549	0.590	0.360	0.170	0.555	0.644	0.306	0.130
	SRP [106]	0.575	0.655	0.453	0.146	0.545	0.601	0.387	0.169	0.555	0.682	0.341	0.123
	MESO [107]	0.549	0.673	0.360	0.159	0.542	0.597	0.354	0.165	0.556	0.661	0.310	0.127
	LTSI [68]	0.695	0.769	0.585	0.106	0.658	0.723	0.499	0.128	0.618	0.718	0.406	0.112
	SPD [65]	0.626	0.685	0.500	0.138	0.596	0.633	0.443	0.171	0.574	0.688	0.345	0.137
	SSAV [18]	0.755	0.806	0.659	0.084	0.661	0.723	0.509	0.117	0.619	0.696	0.399	0.114
	RCR [69]	0.741	0.803	0.653	0.087	0.674	0.729	0.533	0.118	0.644	**0.768**	0.444	**0.094**
	PCSA [108]	0.741	0.793	0.656	0.086	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	FSNet†	**0.773**	**0.825**	**0.685**	**0.072**	**0.707**	**0.764**	**0.597**	**0.104**	**0.662**	0.752	**0.487**	0.099

$F_\beta^{\max}$). Unlike the best two models, which utilize additional training data (RCR leverages pseudo-labels, SSAV utilizes the validation set), our model does not use any additional training data and still outperforms the SSAV model by 8.8% ($F_\beta^{\max}$), and achieves comparable performance to the second-best RCR (ICCV'19) model. These results are also supported by recent conclusions that "human visual attention should be an underlying mechanism that drives U-VOS and V-SOD" (TPAMI'20 [33]).

4.3.6 PR curves

Figure 6 shows precision–recall curves for different models on six V-SOD datasets: $DAVIS_{16}$ [35], MCL [85], FBMS [86], and $DAVSOD_{19}$ [18] Easy35, Normal25, and Difficult20. Note that the higher and further to the right the PR curve, the more accurate the results. Even though existing SOTA methods

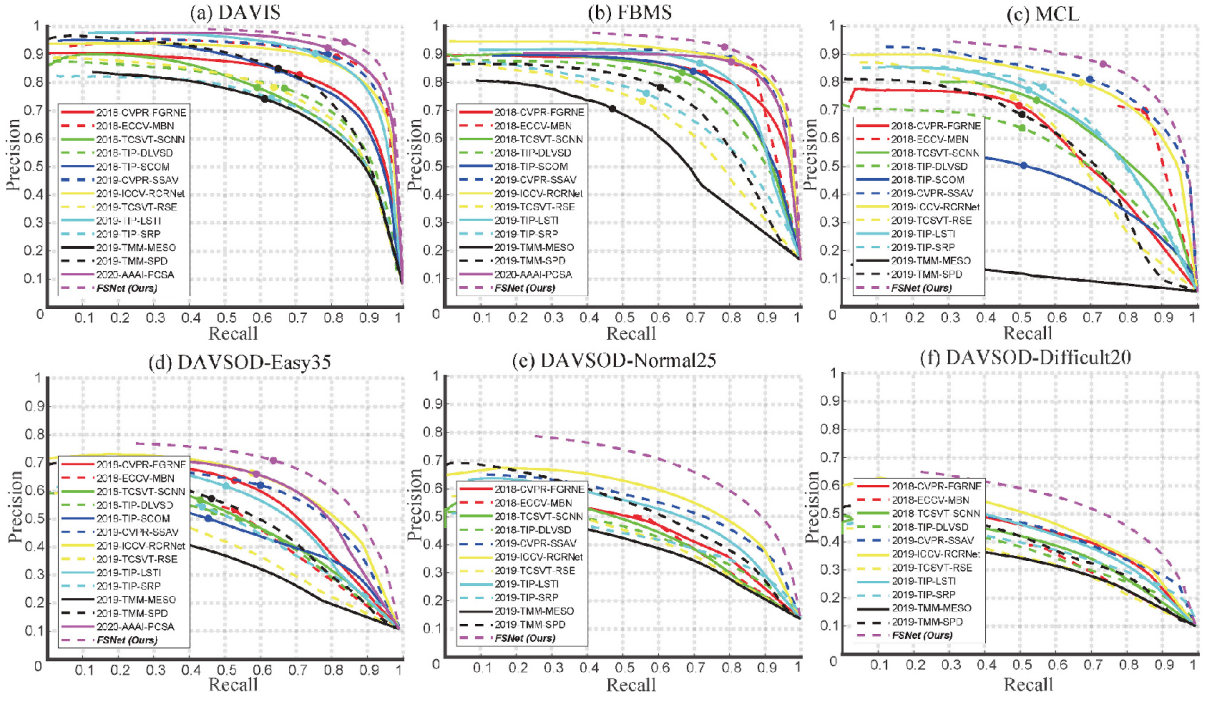

Fig. 6 Precision–recall curves of SOTA V-SOD methods and our proposed *FSNet* across six datasets.

have achieved significant progress in the V-SOD task on three typical benchmark datasets, we still obtain the best performance under all thresholds. While the overall performances on the three subsets of the recent and challenging DAVSOD_{19} [18] dataset are relatively poor, our *FSNet* again achieves more better results by large margins.

4.3.7 Qualitative results

Some qualitative results on five datasets are shown in Fig. 7, validating that our method achieves high-quality U-VOS and V-SOD results. As can be seen in the first row, the camel in the background did not move, so it does not get noticed: as our full-duplex strategy model considers both appearance and motion bidirectionally, it can automatically detect the dominant camel in the centre of the video instead of the camel behind. A similar outcome is also presented in the 5th row, where our method successfully detects dynamic skiers in the video clip rather than the static man in the background. Overall, in these challenging situations, e.g., dynamic backgrounds (1st and 5th rows), fast-motion (4th row), out-of-view (6th and 7th rows), occlusion (7th row), and deformation (8th row), our model is able to infer the real target object(s) with accurate details., demonstrating that *FSNet* is a general framework for both U-VOS and V-SOD tasks.

4.4 Ablation and other studies

We conduct ablation and related studies to analyse our *FSNet*, including stimulus selection, effectiveness of RCAM and BPM, number of cascaded BPMs, and effectiveness of the full-duplex strategy.

4.4.1 Stimulus selection

We now explore the influence of different stimuli (appearance only or motion only) in our framework.

Fig. 7 Qualitative results on five datasets, including DAVIS_{16} [35], MCL [85], FBMS [86], SegTrack-V2 [44], and DAVSOD_{19} [18].

We use only video frames or motion maps (using Ref. [14]) to train the ResNet-50 [73] backbone together with the proposed decoder block (see Section 3.4). As shown in Table 4, motion performs slightly better than appearance in terms of S_α on DAVIS$_{16}$, which suggests that *optical flow* can learn more visual cues than *video frames*. Nevertheless, appearance outperforms motion in terms of the $\mathcal{M}$ metric on MCL. This motivates us to explore how to use appearance and motion cues simultaneously effectively.

4.4.2 *Effectiveness of RCAM*

To validate the effectiveness of RCAM, we replaced our fusion strategy with simple fusion using a concatenate operation followed by a convolutional layer to fuse the two modalities. As expected (Table 4), RCAM performs consistently better than the simple fusion strategy on both DAVIS$_{16}$ and MCL datasets. We would like to point out that our RCAM has two important properties:

- It enables mutual correction and attention.
- It can alleviate error propagation within a network to an extent, due to the mutual correction and bidirectional interaction.

4.4.3 *Effectiveness of BPM*

To illustrate the effectiveness of BPM (with $N = 4$), we derive two different models: RCAM and *FSNet*, frameworks *without* and *with* BPM, respectively. We observe that the model with BPM gains 2%–3% over the one without BPM, according to the statistics in Table 4. We attribute this improvement to BPM's introduction of an interlaced decremental connection, enabling it to fuse the different signals effectively. Similarly, we removed the RCAM to give another pair of models (simple and BPM) to test the robustness of BPM. The results show that even using the bidirectional simple fusion strategy (BPM) by itself can still enhance the stability and generalization of the model. The whole network benefits from the purification forward process and re-calibration backward process.

Table 4 Ablation studies for our components on DAVIS$_{16}$ and MCL. $N = 4$ for BPM. A=Appearance. M=Motion. S=Simple. RCAM=RCAM only. BPM = BPM only

	Component setting				DAVIS$_{16}$		MCL	
	A	M	RCAM	BPM	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
A	✓				0.834	0.047	0.754	0.038
M		✓			0.858	0.039	0.763	0.053
S	✓	✓			0.871	0.035	0.776	0.046
RCAM	✓	✓	✓		0.900	0.025	0.833	0.031
BPM	✓	✓		✓	0.904	0.026	0.855	0.023
FSNet†	✓	✓	✓	✓	**0.920**	**0.020**	**0.864**	**0.023**

4.4.4 *Number of cascaded BPMs*

Naturally, more cascaded BPMs should lead to better boosting. This was investigated and results are shown in Table 5, for $N = \{0, 2, 4, 6, 8\}$. Note that $N = 0$ means that BPM is not used. The improvements can be seen in Fig. 8 (red star), which compares four variants of our *FSNet*, including N=0 (Mean-$\mathcal{J}$=76.4, Mean-$\mathcal{F}$=76.8), N=2 (Mean-$\mathcal{J}$=80.4, Mean-$\mathcal{F}$=81.4), N=4 (Mean-$\mathcal{J}$=82.3, Mean-$\mathcal{F}$=83.3), and N=4 with CRF (Mean-$\mathcal{J}$=83.4, Mean-$\mathcal{F}$=83.1). Using more BPMs leads to better results, which saturate after $N = 4$. Further, too many BPMs (i.e., $N > 4$) lead to high model-complexity and increase the risk of over-fitting. As a trade-off, we used $N = 4$ throughout our experiments.

Table 5 Effect of the number (N) of BPMs on results on DAVIS$_{16}$ [35] and MCL [85], focusing on number of parameters and FLOPs for BPMs, and runtime of *FSNet*

	Param. (M)	FLOPs (G)	Runtime (s/frame)	DAVIS$_{16}$		MCL	
				$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
$N = 0$	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	**0.920**	**0.020**	**0.864**	**0.023**
$N = 6$	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

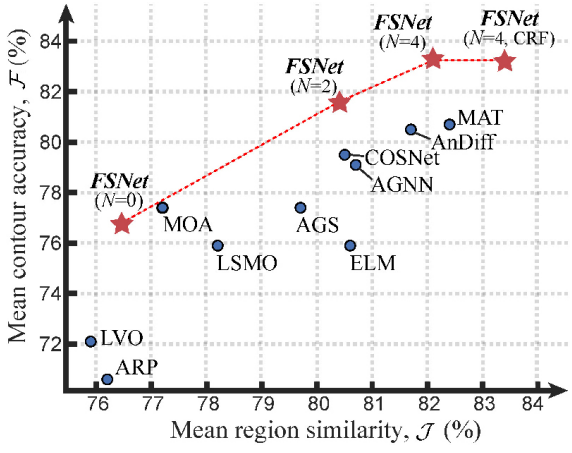

Fig. 8 Mean contour accuracy ($\mathcal{F}$) versus mean region similarity ($\mathcal{J}$) scores on the DAVIS$_{16}$ dataset [35]. Circles indicate U-VOS methods. Four variants of our FSNet are shown in ***bold-italic***. N indicates the number of bidirectional purification modules (BPM) and CRF means CRF [84] post-processing was used. Compared to the best unsupervised VOS model, MAT [25] also with CRF, the proposed method *FSNet* (N=4, CRF) achieves the new SOTA by a large margin.

4.4.5 *Effectiveness of full-duplex strategy*

To investigate the effectiveness of the RCAM and BPM modules with the full-duplex strategy, we studied two unidirectional (i.e., simplex strategies in variants of our model in Fig. 4 and Fig. 5). In Table 6, the symbols $\Rightarrow$, $\Leftarrow$, and $\Leftrightarrow$ indicate the feature transmission directions in the designed RCAM or BPM. Specifically, A $\Leftarrow$ M indicates that the attention vector in the optical flow branch weights the features in the appearance branch and vice versa. A + M $\Leftarrow$ M indicates that motion cues are used to guide the fused features extracted from both appearance and motion. The comparison shows that our elaborately designed modules (RCAM and BPM) jointly cooperate in a full-duplex fashion and outperform all simplex (*unidirectional*) settings.

Table 6 Ablation study for the *simplex* and *full-duplex* strategies on DAVIS$_{16}$ [35] and MCL [85]. $N = 4$ for BPM

	Direction setting		DAVIS$_{16}$		MCL	
	RCAM	BPM	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
simplex	A $\Rightarrow$ M	A + M $\Rightarrow$ M	0.896	0.026	0.816	0.038
	A $\Rightarrow$ M	A + M $\Leftarrow$ M	0.902	0.025	0.832	0.031
	A $\Leftarrow$ M	A + M $\Rightarrow$ M	0.891	0.029	0.806	0.039
	A $\Leftarrow$ M	A + M $\Leftarrow$ M	0.897	0.028	0.840	0.028
self-purif.	A $\Leftrightarrow$ M	A+M $\Leftrightarrow$ M	0.899	0.026	0.854	0.023
full-duplex	A $\Leftrightarrow$ M	A + M $\Leftrightarrow$ M	**0.920**	**0.020**	**0.864**	**0.023**

4.5 Further discussion

4.5.1 *Prediction selection*

Which is the final prediction, $\boldsymbol{S}_A^t$ or $\boldsymbol{S}_M^t$? As mentioned in Section 3.5, we choose $\boldsymbol{S}_A^t$ as our final segmentation result instead of $\boldsymbol{S}_M^t$. The major reasons for doing so can be summarized as

- we employ auxiliary supervision for the motion-based branch to learn more motion patterns inspired by Ref. [53], and
- more informative appearance and motion cues are contained in another branch at the phase of bidirectional purification.

As shown in Table 7, three experiments were conducted to verify our assumption: choosing $\boldsymbol{S}_M^t$, $(\boldsymbol{S}_A^t + \boldsymbol{S}_M^t)/2$, or $\boldsymbol{S}_A^t$ (as per our method) as the final result. All three choices achieve very similar results, but $\boldsymbol{S}_A^t$ performs slightly better than the other two. Besides, considering the reduction of unnecessary computational cost, we choose $\boldsymbol{S}_A^t$ as our final result in our other tests.

Table 7 Choice of final segmentation result for DAVIS$_{16}$ [35] and MCL [85] datasets

Used as result	DAVIS$_{16}$		MCL	
	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
$\boldsymbol{S}_M^t$	0.920	0.022	0.862	0.024
$(\boldsymbol{S}_A^t + \boldsymbol{S}_M^t)/2$	0.920	0.022	0.863	0.023
$\boldsymbol{S}_A^t$ (ours)	0.920	0.022	0.864	0.023

4.5.2 *Effectiveness of CRF*

From Fig. 8 we can see that *FSNet* without CRF, i.e., *FSNet* (N=4), still outperforms the best model AAAI'20-MAT in terms of Mean-$\mathcal{F}$ metric. This means that our initial method (i.e., *FSNet* without CRF) can distinguish hard samples around the object boundaries without post-processing techniques. When equipped with CRF post-processing [84], our *FSNet* (N=4, CRF) achieves the best performance in terms of both Mean-$\mathcal{J}$ and Mean-$\mathcal{F}$ metrics.

4.5.3 *Training effectiveness with less data*

As shown in Fig. 8, the proposed method, *FSNet* (N=4, CRF), surpasses the best U-VOS model MAT [25] (also with CRF), while our *FSNet* uses less labelled data in the training phase (13k versus 16k for MAT). We further observe that the recently proposed 3DC-Seg method [110], based on a 3D convolutional network, achieves the new state-of-the-art (Mean-$\mathcal{J}$=84.3, Mean-$\mathcal{F}$=84.7), but relies on a massive amount of labelled training samples for expert knowledge in the fine-tuning phase, including 158k images (from COCO [111], YouTube-VOS [112], and DAVIS$_{16}$ [35]). It requires about ten times more training data than the best MAT model [25] (16k images) in the fine-tuning phase. This demonstrates the efficient training process in our pipeline.

4.5.4 *Self-purification strategy in BPM*

We provide more details on the different variants mentioned in Section 4.4.5 including A+M $\Leftarrow$ M, A+M $\Rightarrow$ M, and A + M $\Leftrightarrow$ M in BPM. The implementations of A $\Leftarrow$ M and A $\Rightarrow$ M in RCAM are illustrated in Figs. 4(a) and 4(b), while the implementations of A+M $\Leftarrow$ M and A+M $\Rightarrow$ M in BPM are illustrated in Figs. 5(a) and 5(b). Note that all of these variants indicate unidirectional refinement, in contrast to our bi-directional schemes.

Last but not least, to validate that the gains of bi-directional schemes in practice do come from the bi-directional procedure and not more complex model structures, we implemented another variant using the

same structures but without any branch interactions before the decoding stage. This is done by exchanging the places of $\boldsymbol{G}_k^n$ and $\boldsymbol{F}_k^n$ as illustrated in Fig. 9(b), leading to a kind of *self-purification* strategy. Symbol $\nLeftrightarrow$ in Fig. 9(a) means that there is no interaction between the two branches, there is only interaction within a branch itself. The uni- and bidirectional strategies are compared in Table 6. The results show that our carefully designed modules RCAM and BPM jointly cooperate in a bidirectional manner and outperform all unidirectional settings. Furthermore, our bidirectional purification scheme (full duplex in Table 6) also achieves very notable improvements (2.1% and 1.0% gains in S_α on DAVIS_{16} [35] and MCL [85], respectively) against the self-purification variant (self-purif. in Table 6), which has a similarly complex structure, further validating the benefit of the bidirectional behavior claimed in this study.

4.5.5 Relation between RCAM and BPM

The two new modules, RCAM and BPM, focus on using appearance and motion features while ensuring information flow between them. They can work collaboratively under the mutual restraint of our full-duplex strategy, but they cannot be substituted for one another. This is because RCAM transmits the features at each level in a *point-to-point* manner (e.g., $\mathcal{X}_1 \rightarrow \mathcal{Y}_1$), and thus it fits in with the progressive feature extraction in the encoder. The BPM, on the other hand, broadcasts high-level features to low-level features via an interlaced decremental connection in a *set-to-point* manner (e.g., $\{\boldsymbol{F}_2^n, \boldsymbol{F}_3^n, \boldsymbol{F}_4^n\} \rightarrow \boldsymbol{G}_2^n$), which is more suitable for the multi-level feature interaction.

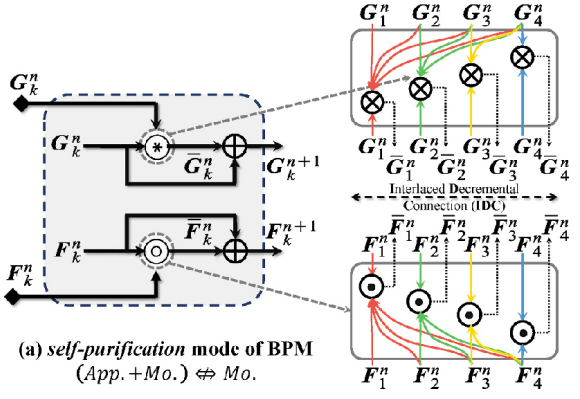

Fig. 9 Self-purification strategy (a) and the proposed bidirectional purification strategy (b) (the latter repeats Fig. 5(c) for convenience). $\oplus$, $\otimes$, and $\odot$ denote element-wise addition, multiplication, and concatenation, respectively.

5 Conclusions

In this paper, we presented a simple yet efficient framework, termed full-duplex strategy network (*FSNet*), that fully leverages the mutual constraints on appearance and motion cues to address the video object segmentation problem. It consists of two core modules: a relational cross-attention module (RCAM) in the encoding stage and an efficient bidirectional purification module (BPM) in the decoding stage. The former is used to abstract features from a dual-modality, while the latter is utilized to re-calibrate inconsistent features step-by-step. We thoroughly validated the functional modules of our architecture by extensive experiments, leading to several interesting findings. Finally, *FSNet* acts as a unified solution that significantly advances the state of the art for both U-VOS and V-SOD tasks. In future, we may extend our scheme to learn short-term and long-term information in an efficient Transformer-like framework [113, 114] to further boost accuracy.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62176169, 61703077, and 62102207).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Wang, Y. Q.; Xu, Z. L.; Wang, X. L.; Shen, C. H.; Cheng, B. S.; Shen, H.; Xia, H. End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8737–8746, 2021.

[2] Chen, X.; Li, Z. X.; Yuan, Y.; Yu, G.; Shen, J. X.; Qi, D. L. State-aware tracker for real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9381–9390, 2020.

[3] Abramov, A.; Pauwels, K.; Papon, J.; Wörgötter, F.; Dellen, B. Depth-supported real-time video segmentation with the Kinect. In: Proceedings of the IEEE Workshop on the Applications of Computer Vision, 457–464, 2012.

[4] Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research* Vol. 36, No. 1, 3–15, 2017.

[5] Jain, S.; Grauman, K. Click carving: Segmenting objects in video with point clicks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* Vol. 4, No. 1, 89–98, 2016.

[6] Wang, H.; Deng, C.; Ma, F.; Yang, Y. Context modulated dynamic networks for actor and action video segmentation with language queries. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 12152–12159, 2020.

[7] Ding, M. Y.; Wang, Z.; Zhou, B. L.; Shi, J. P.; Lu, Z. W.; Luo, P. Every frame counts: Joint learning of video segmentation and optical flow. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 10713–10720, 2020.

[8] Ji, G. P.; Chou, Y. C.; Fan, D. P.; Chen, G.; Fu, H.; Jha, D.; Shao, L. Progressively normalized self-attention network for video polyp segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Lecture Notes in Computer Science, Vol. 12901*. Springer Cham, 142–152, 2021.

[9] Chen, B.; Ling, H.; Zeng, X.; Gao, J.; Xu, Z.; Fidler, S. ScribbleBox: Interactive annotation framework for video object segmentation. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12358*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 293–310, 2020.

[10] Seo, S.; Lee, J. Y.; Han, B. URVOS: Unified referring video object segmentation network with a large-scale benchmark. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12360*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 208–223, 2020.

[11] Pan, Y. W.; Yao, T.; Li, H. Q.; Mei, T. Video captioning with transferred semantic attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 984–992, 2017.

[12] Lee, S. H.; Jang, W. D.; Kim, C. S. Contour-constrained superpixels for image and video processing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5863–5871, 2017.

[13] Reso, M.; Jachalsky, J.; Rosenhahn, B.; Ostermann, J. Temporally consistent superpixels. In: Proceedings of the IEEE International Conference on Computer Vision, 385–392, 2013.

[14] Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; Brox, T. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1647–1655, 2017.

[15] Teed, Z.; Deng, J. RAFT: Recurrent all-pairs field transforms for optical flow. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12347*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 402–419, 2020.

[16] Hu, P.; Wang, G.; Kong, X.; Kuen, J.; Tan, Y. Motion-guided cascaded refinement network for video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 42, No. 8, 1957–1967, 2020.

[17] Tokmakov, P.; Alahari, K.; Schmid, C. Learning video object segmentation with visual memory. In: Proceedings of the IEEE International Conference on Computer Vision, 4491–4500, 2017.

[18] Fan, D. P.; Wang, W. G.; Cheng, M. M.; Shen, J. B. Shifting more attention to video salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8546–8556, 2019.

[19] Chen, Z. X.; Guo, C. C.; Lai, J. H.; Xie, X. H. Motion-appearance interactive encoding for object segmentation in unconstrained videos. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 30, No. 6, 1613–1624, 2020.

[20] Yang, Z.; Wang, Q.; Bertinetto, L.; Bai, S.; Hu, W.; Torr, P. Anchor diffusion for unsupervised video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 931–940, 2019.

[21] Jain, S. D.; Xiong, B.; Grauman, K. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2117–2126, 2017.

[22] Khoreva, A.; Benenson, R.; Ilg, E.; Brox, T.; Schiele, B. Lucid data dreaming for object tracking. In: Proceedings of the 2017 DAVIS Challenge on Video Object Segmentation - CVPR 2017 Workshops, 2017.

[23] Cheng, J.; Tsai, Y.-H.; Wang, S.; Yang, M.-H. SegFlow: Joint learning for video object segmentation and optical flow. In: Proceedings of the IEEE International Conference on Computer Vision, 686–695, 2017.

[24] Xiao, H. X.; Kang, B. Y.; Liu, Y.; Zhang, M. J.; Feng, J. S. Online meta adaptation for fast video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 42, No. 5, 1205–1217, 2020.

[25] Zhou, T. F.; Wang, S. Z.; Zhou, Y.; Yao, Y. Z.; Li, J. W.; Shao, L. Motion-attentive transition for zero-shot video object segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 13066–13073, 2020.

[26] Tsai, Y.-H.; Yang, M.-H.; Black, M. J. Video segmentation via object flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3899–3908, 2016.

[27] Lin, F. Q.; Chou, Y.; Martinez, T. Flow adaptive video object segmentation. *Image and Vision Computing* Vol. 94, 103864, 2020.

[28] Nilsson, D.; Sminchisescu, C. Semantic video segmentation by gated recurrent flow propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6819–6828, 2018.

[29] Li, H.; Chen, G.; Li, G.; Yu, Y. Motion guided attention for video salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 7273–7282, 2019.

[30] Peng, Q. M.; Cheung, Y. M. Automatic video object segmentation based on visual and motion saliency. *IEEE Transactions on Multimedia* Vol. 21, No. 12, 3083–3094, 2019.

[31] Koch, C.; Ullman, S. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology* Vol. 4, No. 4, 219–227, 1985.

[32] Wolfe, J. M.; Cave, K. R.; Franzel, S. L. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology*: *Human Perception and Performance* Vol. 15, No. 3, 419–433, 1989.

[33] Wang, W. G.; Shen, J. B.; Lu, X. K.; Hoi, S. C. H.; Ling, H. B. Paying attention to video object pattern understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43, No. 7, 2413–2428, 2021.

[34] Bharadia, D.; McMilin, E.; Katti, S. Full duplex radios. *ACM SIGCOMM Computer Communication Review* Vol. 43, No. 4, 375–386, 2013.

[35] Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; van Gool, L.; Gross, M.; Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 724–732, 2016.

[36] Ji, G. P.; Fu, K. R.; Wu, Z.; Fan, D. P.; Shen, J. B.; Shao, L. Full-duplex strategy for video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4902–4913, 2021.

[37] Seong, H.; Hyun, J.; Kim, E. Kernelized memory network for video object segmentation. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12367*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 629–645, 2020.

[38] 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: Proceedings of the Computer Vision – ECCV 2020: 16th European Conference, 777–794, 2020.

[39] Hu, L.; Zhang, P.; Zhang, B.; Pan, P.; Xu, Y. H.; Jin, R. Learning position and target consistency for memory-based video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4142–4152, 2021.

[40] Duke, B.; Ahmed, A.; Wolf, C.; Aarabi, P.; Taylor, G. W. SSTVOS: Sparse spatiotemporal transformers for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5908–5917, 2021.

[41] Zhou, T.; Li, J.; Wang, S.; Tao, R.; Shen, J. MATNet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Transactions on Image Processing* Vol. 29, 8326–8338, 2020.

[42] Ochs, P.; Brox, T. Higher order motion models and spectral clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 614–621, 2012.

[43] Fragkiadaki, K.; Zhang, G.; Shi, J. B. Video segmentation by tracing discontinuities in a trajectory embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1846–1853, 2012.

[44] Li, F.; Kim, T.; Humayun, A.; Tsai, D.; Rehg, J. M. Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE International Conference on Computer Vision, 2192–2199, 2013.

[45] Perazzi, F.; Wang, O.; Gross, M.; Sorkine-Hornung, A. Fully connected object proposals for video segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, 3227–3234, 2015.

[46] Wang, W. G.; Shen, J. B.; Porikli, F. Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3395–3402, 2015.

[47] Wang, W. G.; Shen, J. B.; Li, X. L.; Porikli, F. Robust video object cosegmentation. *IEEE Transactions on Image Processing* Vol. 24, No. 10, 3137–3148, 2015.

[48] Galasso, F.; Cipolla, R.; Schiele, B. Video

segmentation with superpixels. In: *Computer Vision – ACCV 2012. Lecture Notes in Computer Science, Vol. 7724*. Lee, K. M.; Matsushita, Y.; Rehg, J. M.; Hu, Z. Eds. Springer Berlin Heidelberg, 760–774, 2013.

[49] Xu, C.; Xiong, C.; Corso, J. J. Streaming hierarchical video segmentation. In: *Computer Vision – ECCV 2012. Lecture Notes in Computer Science, Vol. 7577*. Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer Berlin Heidelberg, 626–639, 2012.

[50] Song, H.; Wang, W.; Zhao, S.; Shen, J.; Lam, K. M. Pyramid dilated deeper ConvLSTM for video salient object detection. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11215*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 744–760, 2018.

[51] Wang, W. G.; Song, H. M.; Zhao, S. Y.; Shen, J. B.; Zhao, S. Y.; Hoi, S. C. H.; Ling, H. Learning unsupervised video object segmentation through visual attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3059–3069, 2019.

[52] Zheng, J.; Luo, W. X.; Piao, Z. X. Cascaded ConvLSTMs using semantically-coherent data synthesis for video object segmentation. *IEEE Access* Vol. 7, 132120–132129, 2019.

[53] Tokmakov, P.; Alahari, K.; Schmid, C. Learning motion patterns in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 531–539, 2017.

[54] Siam, M.; Jiang, C.; Lu, S.; Petrich, L.; Gamal, M.; Elhoseiny, M.; Jagersand, M. Video object segmentation using teacher-student adaptation in a human robot interaction (HRI) setting. In: Proceedings of the International Conference on Robotics and Automation, 50–56, 2019.

[55] Li, S.; Seybold, B.; Vorobyov, A.; Lei, X.; Kuo, C. C. J. Unsupervised video object segmentation with motion-based bilateral networks. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11207*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 215–231, 2018.

[56] Wang, W.; Shen, J.; Yang, R.; Porikli, F. Saliency-aware video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 1, 20–33, 2018.

[57] Zhou, X. F.; Liu, Z.; Gong, C.; Liu, W. Improving video saliency detection via localized estimation and spatiotemporal refinement. *IEEE Transactions on Multimedia* Vol. 20, No. 11, 2993–3007, 2018.

[58] Xu, M. Z.; Liu, B.; Fu, P.; Li, J. B.; Hu, Y. H.; Feng, S. Video salient object detection via robust seeds extraction and multi-graphs manifold propagation. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 30, No. 7, 2191–2206, 2020.

[59] Hu, Y. T.; Huang, J. B.; Schwing, A. G. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11205*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 813–830, 2018.

[60] Wang, W. G.; Shen, J. B.; Shao, L. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing* Vol. 27, No. 1, 38–49, 2018.

[61] Le, T. N.; Sugimoto, A. Deeply supervised 3D recurrent FCN for salient object detection in videos. In: Proceedings of the British Machine Vision Conference, 38.1–38.13, 2017.

[62] Min, K.; Corso, J. TASED-net: Temporally-aggregating spatial encoder–decoder network for video saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2394–2403, 2019.

[63] Li, G. B.; Xie, Y.; Wei, T. H.; Wang, K. Z.; Lin, L. Flow guided recurrent neural encoder for video salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3243–3252, 2018.

[64] Le, T. N.; Sugimoto, A. Video salient object detection using spatiotemporal deep features. *IEEE Transactions on Image Processing* Vol. 27, No. 10, 5002–5015, 2018.

[65] Li, Y. X.; Li, S.; Chen, C.; Hao, A. M.; Qin, H. Accurate and robust video saliency detection via self-paced diffusion. *IEEE Transactions on Multimedia* Vol. 22, No. 5, 1153–1167, 2020.

[66] Borji, A.; Cheng, M. M.; Hou, Q. B.; Jiang, H. Z.; Li, J. Salient object detection: A survey. *Computational Visual Media* Vol. 5, No. 2, 117–150, 2019.

[67] Zhou, T.; Fan, D. P.; Cheng, M. M.; Shen, J. B.; Shao, L. RGB-D salient object detection: A survey. *Computational Visual Media* Vol. 7, No. 1, 37–69, 2021.

[68] Chen, C.; Wang, G. T.; Peng, C.; Zhang, X. W.; Qin, H. Improved robust video saliency detection based on long-term spatial-temporal information. *IEEE Transactions on Image Processing* Vol. 29, 1090–1100, 2020.

[69] Yan, P. X.; Li, G. B.; Xie, Y.; Li, Z.; Wang, C.; Chen, T. S.; Lin, L. Semi-supervised video salient object detection using pseudo-labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 7283–7292, 2019.

[70] Tang, Y.; Zou, W. B.; Jin, Z.; Chen, Y. H.; Hua, Y.; Li, X. Weakly supervised salient object detection with spatiotemporal cascade neural networks. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 29, No. 7, 1973–1984, 2019.

[71] Wang, Z.; Yan, X. Y.; Han, Y. H.; Sun, M. J. Ranking video salient object detection. In: Proceedings of the 27th ACM International Conference on Multimedia, 873–881, 2019.

[72] Zhao, W. B.; Zhang, J.; Li, L.; Barnes, N.; Liu, N.; Han, J. W. Weakly supervised video salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16821–16830, 2021.

[73] He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, 2016.

[74] Wei, J.; Wang, S. H.; Huang, Q. M. F^3Net: Fusion, feedback and focus for salient object detection. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 12321–12328, 2020.

[75] Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. ExFuse: Enhancing feature fusion for semantic segmentation. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11214*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 273–288, 2018.

[76] Sevilla-Lara, L.; Liao, Y.; Güney, F.; Jampani, V.; Geiger, A.; Black, M. J. On the integration of optical flow and action recognition. In: *Pattern Recognition. Lecture Notes in Computer Science, Vol. 11269*. Brox, T.; Bruhn, A.; Fritz, M. Eds. Springer Cham, 281–297, 2019.

[77] Wu, Z.; Su, L.; Huang, Q. Stacked cross refinement network for edge-aware salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 7263–7272, 2019.

[78] Lin, T. Y.; Dollár, P.; Girshick, R.; He, K. M.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 936–944, 2017.

[79] Zhao, H. S.; Shi, J. P.; Qi, X. J.; Wang, X. G.; Jia, J. Y. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6230–6239, 2017.

[80] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351*. Navab, N.; Hornegger, J.; Wells, W.; Frangi, A. Eds. Springer Cham, 234–241, 2015.

[81] Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. et al. PyTorch: An imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 8026–8037, 2019.

[82] He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 37, No. 9, 1904–1916, 2015.

[83] Lu, X. K.; Wang, W. G.; Ma, C.; Shen, J. B.; Shao, L.; Porikli, F. See more, know more: Unsupervised video object segmentation with co-attention Siamese networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3618–3627, 2019.

[84] Krähenbühl, P.; Koltun, V. Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, 109–117, 2011.

[85] Kim, H.; Kim, Y.; Sim, J. Y.; Kim, C. S. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE Transactions on Image Processing* Vol. 24, No. 8, 2552–2564, 2015.

[86] Ochs, P.; Malik, J.; Brox, T. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 36, No. 6, 1187–1200, 2014.

[87] Wang, L. J.; Lu, H. C.; Wang, Y. F.; Feng, M. Y.; Wang, D.; Yin, B. C.; Ruan, X. Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3796–3805, 2017.

[88] Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1597–1604, 2009.

[89] Cheng, M. M.; Mitra, N. J.; Huang, X. L.; Torr, P. H. S.; Hu, S. M. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 37, No. 3, 569–582, 2015.

[90] Borji, A.; Cheng, M. M.; Jiang, H. Z.; Li, J. Salient object detection: A benchmark. *IEEE Transactions on Image Processing* Vol. 24, No. 12, 5706–5722, 2015.

[91] Fan, D. P.; Cheng, M. M.; Liu, Y.; Li, T.; Borji, A. Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision, 4558–4567, 2017.

[92] Wang, W. G.; Lu, X. K.; Shen, J. B.; Crandall, D.; Shao, L. Zero-shot video object segmentation via attentive graph neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9235–9244, 2019.

[93] Faisal, M.; Akhter, I.; Ali, M.; Hartley, R. EpO-net: Exploiting geometric constraints on dense trajectories for motion saliency. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 1873–1882, 2020.

[94] Tokmakov, P.; Schmid, C.; Alahari, K. Learning to segment moving objects. *International Journal of Computer Vision volume* Vol. 127, No. 3, 282–301, 2019.

[95] Koh, Y. J.; Kim, C. S. Primary object segmentation in videos based on region augmentation and reduction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7417–7425, 2017.

[96] Lao, D.; Sundaramoorthi, G. Extending layered models to 3D motion. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11214*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 441–457, 2018.

[97] Papazoglou, A.; Ferrari, V. Fast object segmentation in unconstrained video. In: Proceedings of the IEEE International Conference on Computer Vision, 1777–1784, 2013.

[98] Yang, Z.; Wei, Y.; Yang, Y. Collaborative video object segmentation by foreground-background integration. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12350*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 332–348, 2020.

[99] Johnander, J.; Danelljan, M.; Brissman, E.; Khan, F. S.; Felsberg, M. A generative appearance model for end-to-end video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8945–8954, 2019.

[100] Oh, S. W.; Lee, J. Y.; Sunkavalli, K.; Kim, S. J. Fast video object segmentation by reference-guided mask propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7376–7385, 2018.

[101] Voigtlaender, P.; Chai, Y. N.; Schroff, F.; Adam, H.; Leibe, B.; Chen, L. C. FEELVOS: Fast end-to-end embedding learning for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9473–9482, 2019.

[102] Cheng, J. C.; Tsai, Y. H.; Hung, W. C.; Wang, S. J.; Yang, M. H. Fast and accurate online video object segmentation via tracking parts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7415–7424, 2018.

[103] Caelles, S.; Maninis, K. K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; van Gool, L. One-shot video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5320–5329, 2017.

[104] Perazzi, F.; Khoreva, A.; Benenson, R.; Schiele, B.; Sorkine-Hornung, A. Learning video object segmentation from static images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3491–3500, 2017.

[105] Chen, Y. H.; Zou, W. B.; Tang, Y.; Li, X.; Xu, C.; Komodakis, N. SCOM: Spatiotemporal constrained optimization for salient object detection. *IEEE Transactions on Image Processing* Vol. 27, No. 7, 3345–3357, 2018.

[106] Cong, R. M.; Lei, J. J.; Fu, H. Z.; Porikli, F.; Huang, Q. M.; Hou, C. P. Video saliency detection via sparsity-based reconstruction and propagation. *IEEE Transactions on Image Processing* Vol. 28, No. 10, 4819–4831, 2019.

[107] Xu, M. Z.; Liu, B.; Fu, P.; Li, J. B.; Hu, Y. H. Video saliency detection via graph clustering with motion energy and spatiotemporal objectness. *IEEE Transactions on Multimedia* Vol. 21, No. 11, 2790–2805, 2019.

[108] Gu, Y. C.; Wang, L. J.; Wang, Z. Q.; Liu, Y.; Cheng, M. M.; Lu, S. P. Pyramid constrained self-attention network for fast video salient object detection. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 10869–10876, 2020.

[109] Fan, D.-P.; Ji, G.-P.; Qin, X.; Cheng, M.-M. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis* Vol. 51, No. 9, 1475–1489, 2021. (in Chinese)

[110] Mahadevan, S.; Athar, A.; Ošep, A.; Hennen, S.; Leal-Taixé, L.; Leibe, B. Making a case for 3D convolutions for object segmentation in videos. In: Proceedings of the 31st British Machine Vision Conference, 2020.

[111] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. L. Microsoft COCO: Common objects in context. In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Vol. 8693*. Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 740–755, 2014.

[112] Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; Huang, T. YouTube-VOS: Sequence-to-sequence video object segmentation. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11209*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 603–619, 2018.

[113] Wang, W. H.; Xie, E. Z.; Li, X.; Fan, D. P.; Song, K. T.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 548–558, 2021.

[114] Zhuge, M. C.; Gao, D. H.; Fan, D. P.; Jin, L. B.; Chen, B.; Zhou, H. M.; Qiu, M.; Shao, L. Kaleido-BERT: Vision-language pre-training on fashion domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12642–12652, 2021.

Ge-Peng Ji received his master degree in communication and information systems from the School of Computer Science, Wuhan University, in 2021. He is currently a research intern at the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. His research interests lie in designing deep neural networks and applying deep learning to various fields of computer vision, such as camouflaged and salient object detection, video salient object detection, and medical image segmentation.

Deng-Ping Fan received his Ph.D. degree from Nankai University in 2019. He joined the Inception Institute of Artificial Intelligence (IIAI) in 2019. He has published about 30 top journal and conference papers in outlets such as IEEE TPAMI, IEEE TMI, IJCV, CVPR, ICCV, ECCV, etc. His research interests include computer vision, deep learning, and saliency detection. He served as a senior program committee member for IJCAI 2021.

Keren Fu received dual Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, and Chalmers University of Technology, Gothenburg, Sweden, under the joint supervision of Prof. Jie Yang and Prof. Irene Yu-Hua Gu. He is currently a research associate professor with the College of Computer Science, Sichuan University, China. His current research interests include visual computing, saliency analysis, and machine learning.

Zhe Wu received his Ph.D. degree in computer science from the School of Computer and Control Engineering, University of the Chinese Academy of Sciences, Beijing, in 2020. He is a post-doctoral researcher in the Peng Cheng Laboratory, Shenzhen, China. His current research interests include visual attention, computer vision, and traffic prediction.

Jianbing Shen is a full professor in the School of Computer Science, Beijing Institute of Technology. He has published about 100 journal and conference papers in outlets such as IEEE TPAMI, CVPR, and ICCV. He has received many honors, including a Fok Ying Tung Education Foundation from the Ministry of Education, and awards from the Program for Beijing Excellent Youth Talents from Beijing Municipal Education Commission, and the Program for New Century Excellent Talents from the Ministry of Education. His research interests include computer vision and deep learning. He is an Associate Editor of IEEE TNNLS and IEEE TIP.

Ling Shao is the CEO and Chief Scientist of the Inception Institute of Artificial Intelligence (IIAI). He was the initiator and the Founding Provost and Executive Vice President of the Mohamed bin Zayed University of Artificial Intelligence (the world's first AI University), United Arab Emirates. His research interests include computer vision, machine learning, and medical imaging. He is a fellow of the IEEE, the IAPR, the IET, and the BCS.