Hierarchical Feature Alignment Network for Unsupervised Video Object Segmentation

Gensheng Pei[1] , Fumin Shen[2(✉)] , Yazhou Yao[1] , Guo-Sen Xie[1(✉)] ,
Zhenmin Tang[1] , and Jinhui Tang[1]

[1] Nanjing University of Science and Technology, Nanjing, China
yazhou.yao@njust.edu.cn
[2] University of Electronic Science and Technology of China, Chengdu, China
https://github.com/NUST-Machine-Intelligence-Laboratory/HFAN

Abstract. Optical flow is an easily conceived and precious cue for advancing unsupervised video object segmentation (UVOS). Most of the previous methods directly extract and fuse the motion and appearance features for segmenting target objects in the UVOS setting. However, optical flow is intrinsically an instantaneous velocity of all pixels among consecutive frames, thus making the motion features not aligned well with the primary objects among the corresponding frames. To solve the above challenge, we propose a concise, practical, and efficient architecture for appearance and motion feature alignment, dubbed hierarchical feature alignment network (HFAN). Specifically, the key merits in HFAN are the sequential **F**eature **A**lign**M**ent (FAM) module and the **F**eature **A**dapta**T**ion (FAT) module, which are leveraged for processing the appearance and motion features hierarchically. FAM is capable of aligning both appearance and motion features with the primary object semantic representations, respectively. Further, FAT is explicitly designed for the adaptive fusion of appearance and motion features to achieve a desirable trade-off between cross-modal features. Extensive experiments demonstrate the effectiveness of the proposed HFAN, which reaches a new state-of-the-art performance on DAVIS-16, achieving 88.7 $\mathcal{J}\&\mathcal{F}$ Mean, *i.e.*, a relative improvement of 3.5% over the best published result.

Keywords: Video object segmentation · Feature alignment

1 Introduction

Video object segmentation (VOS) aims to segment objects for each frame in a video sequence. Compared to semi-supervised VOS (SVOS), in which annotations are provided for the first frame at test time, unsupervised VOS (UVOS) is particularly challenging as it involves no prior knowledge and human interposing. This work focuses on the UVOS task, which has motivated numerous downstream segmentation topics [5,78,26,3,71].

UVOS approaches can be grouped into three main subcategories: motion-based, appearance-based, and motion-appearance-based, depending on the utilizations of different feature types. By merely using motion information [45,59],
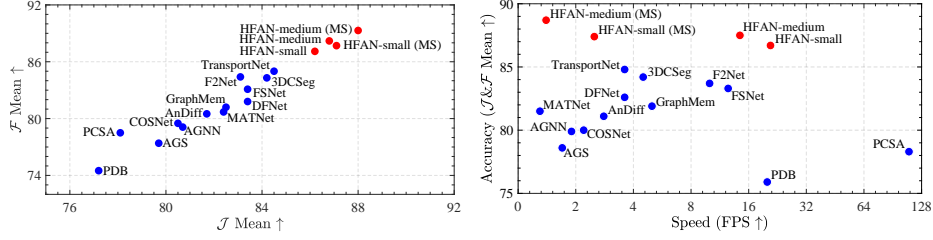
Fig. 1. Performance $\mathcal{J}\&\mathcal{F}$ Mean versus inference speed FPS (frames per second) on DAVIS-16 [47]. Existing and proposed methods are marked with ● and ●, respectively.

the UVOS is transformed into a moving object segmentation (MOS) task. The main drawback of MOS is the risk of losing targets when the object is moving slowly or is stationary. Further, appearance-based methods [63,38,34,77] usually describe the target in detail using mature image segmentation techniques. However, the lack of prior knowledge on the primary objects in the unsupervised solution, almost always, can lead to mis-segmentation cases. By contrast, motion-appearance schemes [1,69,75,50,79] can mitigate the deficiencies of the above two types of methods. Appearance features compensate for the shortage of motion descriptions on semantic representations, and motion cues enable the high-quality candidate regions to be selected for appearance features.

As the leading motion-appearance scenario, optical flow guided UVOS methods have significantly advanced the performances of segmentation. Abandoning appearance modeling by converting VOS into a foreground motion prediction based purely on optical flow information does not handle static foreground objects well. However, two intrinsic drawbacks exist in these approaches. **First**, optical flow describes the velocity vector of per-pixel motion in a video, capturing the motion information between consecutive frames. As such, the positions of the primary objects in a frame and/or its corresponding optical flow are usually not well preserved. Most of the existing methods solely fuse the features of a frame and its corresponding optical flow without considering the respective alignments between the primary objects and the appearance/motion features. This inevitably leads to the loss of boundary information of the primary objects. **Second**, when facing occlusions, motion blur, fast-moving objects, and even stationary objects in UVOS, the inferior optical flow estimation directly affects the segmentation results of the final models. Especially, if optical flow estimation fails, the motion features of the primary objects in the video are invalidated accordingly. In this case, the unselective fusion of appearance and motion features is probably harmful to segmentation accuracy.

To tackle the above challenges, inspired by the current trend of optical flow guided UVOS [59,1,22,69,75,50,79], we propose a hierarchical feature alignment network (HFAN). It aligns the object positions with the motion/appearance features and adapts the aligned features for mitigating the cross-modal mismatch. Specifically, we construct a **F**eature **A**lign**M**ent (FAM) module to implement

object-level alignment with appearance/motion features in the multi-level feature encoding stage. Considering that the spatial locations of appearance and ground-truth target regions are seamlessly matched, we generate the coarse segmentation probability mask by appearance features. Next, FAM leverages the same object regions (*i.e.*, coarse masks) to represent object-level alignment features for appearance and motion. Moreover, we build the **F**eature **A**dapta**T**ion (FAT) module to combine appearance and motion features after the alignment step. FAT aims to ensure the robustness of fused features by constructing an adaptive weight between appearance and motion features. Notably, the adaptive fusion of appearance and motion features could effectively relieve the harms of optical flow estimation failure and motion blur on segmentation results.

We assess the effectiveness and reliability of the proposed model on three widely-used datasets. On DAVIS-16 [47], our HFAN-small and HFAN-medium achieve 86.7 and 87.5 $\mathcal{J}\&\mathcal{F}$ Mean, respectively, at 20.8 and 14.4 FPS. These are new state-of-the-art (SOTA) results in terms of accuracy and speed, as shown in Fig. 1. On YouTube-Objects [49], the proposed HFAN-small represents a relative improvement of 2.0% over the reported best result. Furthermore, the proposed method achieves an equivalent performance to SVOS models on Long-Videos [32]. Meanwhile, HFAN also reaches the best reported result on DAVIS-16 for the video salient object detection (VSOD), which aims to detect salient regions in videos. In summary, HFAN provides an efficient solution and a new perspective on optical flow guided UVOS.

2   Related Work

2.1   Video Object Segmentation

Current video object segmentation is broadly classified as unsupervised VOS [63,65,70,58] and semi-supervised VOS [39,44,52,8] tasks. The main difference is whether they provide accurate pixel-level annotations for the first frame of the segmented video at inference. As research on VOS has progressed, interactive VOS methods [43,15,5] that utilize user interaction (*e.g.*, scribbles or clicks) as input to iteratively optimize segmentation results have yielded good performance. Referring VOS setting [18,51,24] arises from considering a different type of interaction, language-guided video expressions, *i.e.*, target objects referred by a given language descriptions. However, the expensive nature of high-quality annotated video data motivates the need for an elegant and unrestricted VOS setting. This paper focuses on UVOS, which does not use any human intervention during testing. Depending on whether the current methods use deep features or not, we further divide UVOS into two subcategories: *traditional* and *deep*.

The computer vision community has extensively studied the task of automatic binary video segmentation over nearly three decades. Early *traditional* models were typically based on specific heuristics related to the foreground (*i.e.*, target proposals [48], motion boundaries [45], salient objects [64]). They required hand-crafted low-level features (*e.g.*, SIFT, edges). Later, several methods (*e.g.*, point trajectories [42], background subtraction [12] and over-segmentation [10])

were proposed to segment and track all targets with different motions and appearances in the video. More recently, with the renaissance of artificial neural networks, *deep* models (*e.g.*, CNN [77,70], RNN [55,65,1], GNN [63,36]) have enabled UVOS to evolve rapidly. A quintessential example of an attempt to apply deep learning techniques in this field is LSMO [59], which learns a multilayer perceptron to detect moving objects. The computational burden is reduced by many subsequent approaches based on fully convolutional networks, such as two-stream structures [58,21,28,54], CNN-based encoder-decoder architectures [79,6,78], and Siamese network [37,34]. As the field of optical flow estimation [19,57,56] has flourished, more and more optical flow based UVOS methods [22,69,75,50] have gained tremendous performance improvements. The major difference from the optical flow-based approaches described above is that we reconsider the mismatch between frames and optical flow. Our HFAN performs hierarchical feature alignment and adaptation of motion-appearance features to achieve accurate feature representation of the primary objects in a video.

2.2   Feature Alignment

Feature alignment is widely used in various fields, *e.g.*, object detection [4,13,33], image segmentation [16,31,74,17], and person re-identification [73,40,62,72]. For object detection, feature alignment mainly involves the misalignment between anchor boxes and convolutional features, in addition to multiple anchors for the same point in the feature map. Existing image segmentation models generally adopt the feature pyramid networks (FPN) [33] to obtain different resolution feature maps to improve performance. However, this increases the loss of boundary information during downsampling and unaligned feature maps with different resolutions for upsampling. An effective way [16,31] is to align the features from the coarsest resolution to the finest resolution to match positions between feature maps. Aligning and adapting motion and appearance features of multi-level representations from the same encoder is implemented by our HFAN. Thus, it is guaranteed that hierarchical feature maps between the two modalities align their respective features based on the same primary objects.

3   The Proposed Method

Our HFAN consists of two modules: *feature alignment* (FAM, §3.2) and *feature adaptation* (FAT, §3.3). FAM aligns the hierarchical features of appearance and motion feature maps with the primary objects. FAT fuses these two aligned feature maps at the pixel-level with a learnable adaptive weight.

3.1   Task Definitions

Given an input video $\mathcal{I}$ with N frames, we can select each frame $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, and calculate the relative optical flow $\mathbf{O} \in \mathbb{R}^{H \times W \times 3}$ (visualized as an RGB image) by [57]. In the i-th stage of the multi-level feature representation ($i \in$
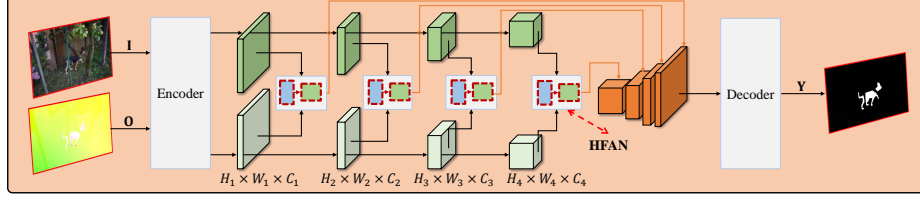
Fig. 2. The pipeline of HFAN. Frame **I** and optical flow **O** are used as inputs to extract hierarchical appearance and motion features, respectively, through an encoder with HFAN. And the excellent segmentation mask **Y** is obtained by the decoder.

$\{1, 2, 3, 4\}$), the appearance and motion features are denoted as $\mathbf{I}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ and $\mathbf{O}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, respectively. $H_i \times W_i$ indicates the feature resolution, where the value is set to $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$. The proposed HFAN aims to generate object-level aligned, high-quality adapted features,

$$\mathbf{U}_i = \mathcal{F}_{\text{HFAN}}(\mathbf{I}_i, \mathbf{O}_i) \in \mathbb{R}^{H_i \times W_i \times C_i}. \tag{1}$$

Here, $\mathcal{F}_{\text{HFAN}}(\cdot, \cdot)$ contains two main modules, which are:

$$\hat{\mathbf{I}}_i, \hat{\mathbf{O}}_i = \mathcal{F}_{\text{FAM}}(\mathbf{I}_i, \mathbf{O}_i), \mathbf{U}_i = \mathcal{F}_{\text{FAT}}(\hat{\mathbf{I}}_i, \hat{\mathbf{O}}_i), \tag{2}$$

where $\mathcal{F}_{\text{FAM}}$ conducts feature alignment for $\mathbf{I}_i$ and $\mathbf{O}_i$, and $\mathcal{F}_{\text{FAT}}$ fuses aligned feature maps from $\mathcal{F}_{\text{FAM}}$ by performing a multi-modal adaptive feature fusion. The overall architecture of the proposed method is shown in Fig. 2.

We adopt a lightweight MiT [66] backbone (ResNet [14] is also studied, see §4.2 for details.) and employ a decoder to yield the primary object binary mask $\mathbf{Y} \in \{0, 1\}^{H \times W}$ of the frame $\mathbf{I}$. Next, we illustrate in detail the two main modules of our proposed HFAN model, along with the training and inference phases.

3.2 Feature Alignment Module

Optical flow methods produce a dense motion vector field by generating a vector for each pixel, which is important auxiliary information for studying video analysis and representation. Former works [79,69,75,50] using optical flow guidance have defaulted one video frame and its optical flow to aligned images. However, this hypothesis then holds approximately only if the motion between two consecutive frames is small. In addition, this solidification tends to result in poor accuracy along moving object boundaries. An intuitive concept is that although appearance features and motion features are unaligned, the bond between them is that they share the primary objects. Motivated by this, we design a feature alignment module specifically for the frame and optical flow to alleviate these issues. Firstly, FAM predicts the coarse segmentation probability of $\mathbf{I}_i$ to obtain

$$\mathbf{P}_i = \mathcal{F}_{\text{CS}}(\mathbf{I}_i) \in \mathbb{R}^{H_i \times W_i \times N_{cls}}, \tag{3}$$

where $\mathcal{F}_{\mathrm{CS}}(\cdot)$ represents the coarse segmentation probability mask implemented by a convolution block $\mathtt{Conv}_{1\times1}(C_i, N_{cls}) \rightarrow \mathtt{BN} \rightarrow \mathtt{ReLU}$ on the appearance feature map $\mathbf{I}_i$, and N_{cls} denotes the number of categories. Here, $\mathtt{BN}$ indicates the batch normalization [20] and $\mathtt{ReLU}$ is the rectified linear unit [41]. This paper focuses on single foreground and background, so N_{cls} is set to 2.

The regions contained in the coarse probability mask $\mathbf{P}_i$ obtained by the original frame $\mathbf{I}_i$ are consistent with the primary object's areas to be segmented. Therefore, we design the feature alignment module, which only aligns the appearance and motion features separately for the mask regions. This way has the merit of reducing the computational cost while weakening the negative impact of the optical flow background noise on the segmentation. Subsequently, $\mathbf{P}_i$ obtained by Eq. (3) is a contextual representation of primary object regions co-built with the original appearance feature map. We design the category-specific semantic (CSS) module to represent the category semantic, formulated as

$$
\begin{aligned}
\mathbf{I}_i^{'} &= \mathtt{permute}(\mathtt{view}(\mathbf{I}_i)), \mathbf{P}_i^{'} = \mathtt{softmax}(\mathtt{view}(\mathbf{P}_i)), \\
\mathbf{M}_i &= \mathcal{F}_{\mathrm{CSS}}(\mathbf{I}_i, \mathbf{P}_i) = \mathtt{matmul}(\mathbf{I}_i^{'}, \mathbf{P}_i^{'}) \in \mathbb{R}^{C_i \times N_{cls} \times 1},
\end{aligned}
\tag{4}
$$

where $\mathtt{view}$, $\mathtt{permute}$ and $\mathtt{matmul}$ indicate the reshaping, permuting tensor dimension and tensor product. $\mathtt{softmax}$ is also known as normalized exponential function. The prominent role of $\mathbf{M}_i$ is summarized in two points: **1)** the spatial compression of appearance features within a specific region $\mathbf{P}_i$; **2)** the construction of category-specific information shared by appearance and motion features. The interchange on feature and semantic levels performed by $\mathcal{F}_{\mathrm{CSS}}$ makes it possible to seek common contexts for appearance-motion features.

Immediately afterward, the primary object context (POC) module is devised to perform object-level contextual alignment of appearance and motion features with the same $\mathbf{M}_i$. Inspired by self-attention [60], $\mathcal{F}_{\mathrm{SA}}$ is achieved by

$$
\begin{aligned}
\mathcal{F}_{\mathrm{SA}} &= \mathtt{softmax}(\alpha \mathbf{Q}\mathbf{K}^{\mathrm{T}})\mathbf{V}, \mathbf{Q} \in \{\mathbf{I}_i, \mathbf{O}_i\} \\
\mathbf{Q}, \mathbf{K}, \mathbf{V} &= \mathcal{F}_{\mathrm{Query}}(\mathbf{Q}), \mathcal{F}_{\mathrm{Key}}(\mathbf{M}_i), \mathcal{F}_{\mathrm{Value}}(\mathbf{M}_i),
\end{aligned}
\tag{5}
$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ denote the query, key and value obtained by three transformation operations of $\mathcal{F}_{\mathrm{Query}}$, $\mathcal{F}_{\mathrm{Key}}$ and $\mathcal{F}_{\mathrm{Value}}$, respectively. They are formed by $\mathtt{Conv}_{1\times1}(C_i, C_i/r) \rightarrow \mathtt{BN} \rightarrow \mathtt{ReLU}$. $\alpha = \frac{1}{\sqrt{C_i}}$ is a scaling factor. r is set to $C_i/16$ for channel reduction ratio and $\mathtt{concat}$ indicates the concatenation operation.

Our proposed POC module helps refine the target boundaries and alleviate the primary object shifts between the frame and optical flow. The POC module can be represented as follows:

$$
\begin{aligned}
\hat{\mathbf{I}}_i &= \mathcal{F}_{\mathrm{POC}}(\mathbf{I}_i, \mathbf{M}_i) \in \mathbb{R}^{H_i \times W_i \times C_i}, \\
\hat{\mathbf{O}}_i &= \mathcal{F}_{\mathrm{POC}}(\mathbf{O}_i, \mathbf{M}_i) \in \mathbb{R}^{H_i \times W_i \times C_i},
\end{aligned}
\tag{6}
$$

where $\hat{\mathbf{I}}_i$ and $\hat{\mathbf{O}}_i$ are appearance-aligned and motion-aligned feature maps, respectively.
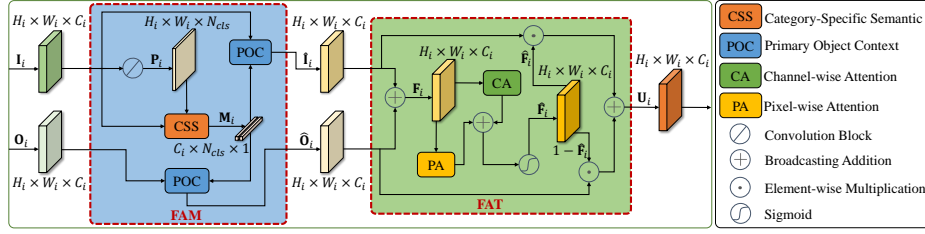
Fig. 3. Illustration of the proposed FAM and FAT modules. *Feature Alignment* and *Feature Adaptation* are applied on each hierarchical feature map to resolve the position and modal mismatches between optical flow and video frames. $\mathbf{U}_i$ denotes the alignment and adaptation features of stage $i \in \{1, 2, 3, 4\}$.

Compared to previous methods, FAM does not interact directly with appearance and motion features but employs CSS and POC modules to achieve contextual alignment of different modal features. As shown in Fig. 3-FAM, when $\mathbf{I}_i$ and $\mathbf{O}_i$ go through FAM, their respective features represent the shared primary object region $\mathbf{P}_i$, guided by $\mathbf{M}_i$. It ensures the feature independence of appearance $\hat{\mathbf{I}}_i$ and motion $\hat{\mathbf{O}}_i$ before the feature adaptation fusion phase.

3.3 Feature Adaptation Module

After the appearance and motion features are expressed based on the same object-contextual region, the aligned feature maps $\hat{\mathbf{I}}_i$ and $\hat{\mathbf{O}}_i$ have more boundary information and less background noise. However, when the optical flow estimation fails due to slow motion or stationary target objects, retaining all the optical flow features would cause a tremendous loss in segmentation performance. To this end, we require adaptive operations between cross-modal features for them. In this work, we propose the feature adaptation (FAT) module.

Specifically, we aggregate appearance-aligned and motion-aligned features and accordingly get the fused feature map $\mathbf{F}_i$, which embraces all the information of $\hat{\mathbf{I}}_i$ and $\hat{\mathbf{O}}_i$. The formula is directly expressed as $\mathbf{F}_i = \hat{\mathbf{I}}_i + \hat{\mathbf{O}}_i$. Here, $\mathbf{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ is treated as a semantic feature map after the superposition of appearance and motion contexts, equivalent to performing a skip connection operation [14,33] on different modal features of the same resolutions. Inspired by [30], channel-level and pixel-level semantic representations are obtained by

$$
\begin{aligned}
\mathbf{F}_i^{\mathrm{CA}} &= \mathcal{F}_{\mathrm{CA}}(\mathbf{F}_i) \in \mathbb{R}^{H_i \times W_i \times C_i}, \\
\mathbf{F}_i^{\mathrm{PA}} &= \mathcal{F}_{\mathrm{PA}}(\mathbf{F}_i) \in \mathbb{R}^{1 \times 1 \times C_i},
\end{aligned}
\tag{7}
$$

where $\mathcal{F}_{\mathrm{CA}}(\cdot)$ and $\mathcal{F}_{\mathrm{PA}}(\cdot)$ indicate channel-wise and pixel-wise attention operations are performed on $\mathbf{F}_i$.

Instead of using $\mathbf{F}_i$ directly as the fused appearance and motion features as in existing approaches [50,79,21], we propose to adapt these features. Specifically, we transform $\mathbf{F}_i$ into basis weight with feature adaptation to ensure stable

feature representation capability even under low-quality motion information conditions (*e.g.*, occlusion and slow motion). The formula is expressed as

$$\hat{\mathbf{F}}_i = \texttt{sigmoid}(\mathbf{F}_i^{\text{CA}} + \mathbf{F}_i^{\text{PA}}) \in \mathbb{R}^{H_i \times W_i \times C_i},$$
$$\mathbf{U}_i = \hat{\mathbf{I}}_i \odot \hat{\mathbf{F}}_i + \hat{\mathbf{O}}_i \odot (1 - \hat{\mathbf{F}}_i) \in \mathbb{R}^{H_i \times W_i \times C_i}, \tag{8}$$

where $\odot$ denotes the element-wise multiplication. At this point, details of FAT are introduced, and the workflow is illustrated in Fig. 3-FAT. Further observation of Eq. (8) shows that when $(1 - \hat{\mathbf{F}}_i)$ approaches 0, all information of $\mathbf{U}_i$ is provided by appearance features, while when $\hat{\mathbf{F}}_i$ reaches 0, all information of $\mathbf{U}_i$ is supplied by motion features. Meanwhile, $\hat{\mathbf{F}}_i$ is learnable, so it realizes the feature self-adaptation of the frame and optical flow.

3.4   Training and Inference

The multi-level features $\mathbf{U}_i$ ($i \in \{1, 2, 3, 4\}$) obtained through HFAN are fed to the decoder $\mathcal{F}_{\text{DEC}}$, and the predicted segmentation mask $\mathbf{Q}$ is acquired

$$\mathbf{Q} = \mathcal{F}_{\text{DEC}}(\mathbf{U}_{i|i=1,2,3,4}) \in \mathbb{R}^{H \times W \times N_{cls}}, \tag{9}$$

where $\mathcal{F}_{\text{DEC}}(\cdot)$ utilizes a lightweight All-MLP decoder provided by [66] to ensure consistency with the encoder network MiT.

Our model is trained to minimize the loss function $\mathcal{L}$ as follows

$$\mathcal{L} = \frac{1}{H \times W} \sum_{p,q} \mathcal{L}_{\text{CE}}(\mathbf{Q}_{[\cdot,p,q]}, \mathbf{G}_{[p,q]}), \tag{10}$$

where $\mathcal{L}_{\text{CE}}$ is the Cross Entropy Loss. $\mathbf{G}$ stands for the ground-truth mask. $\sum_{p,q}$ denotes the sum over all positions on the frame $\mathbf{I}$. In the inference stage, $\mathbf{Q}$ from the decoder is passed directly through the $\texttt{argmax}$ function to infer the final binary mask $\mathbf{Y}$. Prediction segmentation of video $\mathcal{I}$ without applying any post-processing techniques can be phrased as

$$\mathbf{Y} = \texttt{argmax}(\mathbf{Q}) \in \{0, 1\}^{H \times W}. \tag{11}$$

4   Experiments

4.1   Experimental Setup

Datasets. We evaluate HFAN on three publicly available datasets with UVOS: DAVIS-16 [47], YouTube-Objects [49] and Long-Videos [32]. DAVIS-16 [47] contains a total of 50 videos, including 30 videos for training and 20 videos for validation. YouTube-Objects [49] includes 126 web videos divided into 10 categories with a total of more than 20,000 frames. Long-Videos [32] consists of three videos, each of which contains about **2500** frames per video sequence.

Implementation Details. We utilize PyTorch [46] and MMSegmentation code-base [7] to implement our model and train on two NVIDIA V100 with a mini-batch size of 8 per GPU. To achieve a better trade-off between accuracy and speed, we choose lightweight MiT-b1 and middleweight MiT-b2 as the backbones rather than the better but larger MiT-b3 to MiT-b5 [66]. Following [79,61,38], we pre-train our network on YouTube-VOS [67] and fine-tune on the training set of DAVIS-16 [47]. During training, we augment data online by random horizontal flipping, random resize with ratio $0.5 \sim 2.0$, and random cropping to 512×512. We use AdamW optimizer to pre-train for 160K iterations and fine-tune for 4K iterations. The learning rates of pre-training and fine-tuning are set to $6e - 5$ with a *poly* schedule and $1e - 5$ with a *fixed* schedule, respectively. For obtaining an elegant end-to-end model, we do not employ training tricks like auxiliary head loss and online hard example mining [53]. Moreover, no post-processing techniques (*e.g.*, the widely used CRF [25]) are used in the inference phase. All inference processes for the experiments are performed using a single V100. We report UVOS performance using two standard evaluation metrics recommended by [47], *i.e.*, region similarity $\mathcal{J}$ and boundary accuracy $\mathcal{F}$.

4.2 Ablation Studies

To quantify the effect of each fundamental component in HFAN, we perform an exhaustive ablation study on the DAVIS-16 val-set [47]. For the fairness of the ablation results, we do not perform any post-processing techniques.

Impact of Data Input. To analyze the effect of appearance and motion features on performance, we first conduct an ablation study on the data input in Table 1. We adopt the video frame and corresponding optical flow as data inputs. A simple additive feature fusion approach is employed as the baseline. Compared to using a single input type, the baseline improves performance by providing richer appearance and motion cues. The ablation results illustrate that optical flow, which is deemed as the temporal consistency between video frames, requires the coaction of appearance features to achieve the desired effect.

Efficacy of Crucial Modules. When comparing our baseline with FAM, FAT, and HFAN, the results in Table 2 reveal that HFAN is a superior aggregate of FAM and FAT. Specifically, FAM improves 2.2% and 2.3% in terms of $\mathcal{J}$ Mean and $\mathcal{F}$ Mean, respectively. FAT increases by 2.0% on $\mathcal{J}$ Mean and 2.8% $\mathcal{F}$ Mean. The best performance gains achieved by HFAN, which is implemented by combining FAM and FAT modules, further demonstrates the effectiveness of

Table 1. Ablation study for data input. All ablated versions utilize hierarchical architecture MiT-b1 as the backbone.

Input	$\mathcal{J}$ Mean ↑	$\Delta\mathcal{J}$	$\mathcal{F}$ Mean ↑	$\Delta\mathcal{F}$
Image frame only	79.1	-3.9	79.8	-3.5
Optical flow only	77.9	-5.1	76.5	-6.8
Baseline	83.0	-	83.3	-

Table 2. Ablation study for module. HFAN indicates a full model with integrated FAM and FAT modules.

Variant	$\mathcal{J}$ Mean ↑	$\Delta\mathcal{J}$	$\mathcal{F}$ Mean ↑	$\Delta\mathcal{F}$	FPS ↑
Baseline	83.0	-	83.3	-	22.0
Baseline + FAM	85.2	+2.2	85.6	+2.3	21.0
Baseline + FAT	85.0	+2.0	86.1	+2.8	21.4
Baseline + HFAN	86.2	+3.2	87.1	+3.8	20.8

Table 3. Ablation study on different backbones. Transformer-like and CNN-like versions are considered in the experimental ablation. For the test setup, SS / MS denotes single / multi-scale test.

Backbone	Test Setup	$\mathcal{J}$ Mean ↑	$\mathcal{F}$ Mean ↑	FPS ↑
MiT-b0	SS	81.5	80.8	24.0
	MS	83.4	82.3	3.4
MiT-b1	SS	86.2	87.1	20.8
	MS	87.1	87.7	2.5
MiT-b2	SS	86.8	88.2	14.4
	MS	88.0	89.3	1.4
MiT-b3	SS	86.8	88.8	10.6
	MS	88.2	90.0	1.0
Swin-Tiny	SS	86.0	87.3	12.8
	MS	87.2	87.9	1.1
ResNet-101	SS	86.6	87.3	12.4
	MS	87.3	87.9	1.3

Table 4. Ablation study on different input sizes and optical flow with MS test.

Method Size	RAFT		PWCNet	
	$\mathcal{J}$ Mean ↑	$\mathcal{F}$ Mean ↑	$\mathcal{J}$ Mean ↑	$\mathcal{F}$ Mean ↑
384×384	86.2	86.6	84.5	84.7
448×448	86.9	87.5	85.3	85.7
480×480	86.9	87.6	85.5	85.9
512×512	87.1	87.7	85.7	86.0

Table 5. Ablation study on Transformer-like and CNN-like network architectures. Ablated results are obtained in same setups (RAFT, 512×512, and MS test).

Method Backbone	MATNet + CRF		Ours	
	$\mathcal{J}$ Mean ↑	$\mathcal{F}$ Mean ↑	$\mathcal{J}$ Mean ↑	$\mathcal{F}$ Mean ↑
MiT-b1	83.8	82.6	87.1	87.7
MiT-b2	84.7	83.8	88.0	89.3
ResNet-101	84.0	82.9	87.3	87.9

proposed approach. For aligning features of co-foreground objects in different modal images, HFAN achieves a simple way to correct shift differences between video frames and their corresponding optical flow features. In addition, HFAN achieves adaptive selection in the feature fusion phase by learning feature adaptation weights. Fig. 4 visualizes the ablated versions for Table 2. It can be found that FAM aligns the image and optical flow features to yield smoother and more refined object boundaries. Meanwhile, FAT enhances the image and optical flow features by adaptive transformation. Our HFAN inherits advantages of FAM and FAT, obtaining more finesse in the target region and removing a larger amount of noise in the non-target region.

Efficacy of Backbone. We investigate the effect of different backbone networks on accuracy and speed. The results of MiT-b0 to MiT-b3 [66] are shown in Table 3 (Note that we do not run experiments using MiT-b4 and MiT-b5 due to GPU memory limitations.). We find that the performance increases when enlarging the size of backbone networks. However, a larger network leads to a lower model efficiency and real-time speed. In addition, other types of backbone networks (*e.g.*, Swin Transformer [35] and ResNet [14]) also achieve competitive results. This adequately demonstrates the generality of the proposed method. Given the trade-off between the model size and performance, we choose MiT-b1 and MiT-b2 as the small and medium backbone networks for HFAN, respectively.
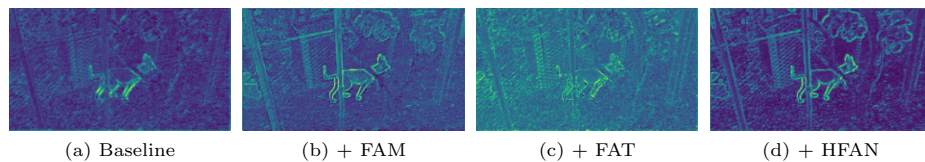

(a) Baseline	(b) + FAM	(c) + FAT	(d) + HFAN

Fig. 4. Illustration of the first stage feature maps $\mathbf{U}_1$ from four ablated models.

Effect of Image Size and Optical Flow. Low-resolution image inputs generally degrade the performance of the segmentation model, while the use of different optical flow estimation methods also affects the final segmentation results. To study the effects of image size and optical flow estimation methods on the proposed method, we explore four different image size inputs and two well-known optical flow estimation methods. The ablation results are shown in Table 4, and we can find that **1)** the proposed method still has good performance under the low-resolution condition; **2)** RAFT [57] has better results than PWC-Net [56] at the same resolution. The comprehensive analysis suggests that our method is not sensitive to the image resolution, while the optical flow estimation of different quality has a more obvious impact on the segmentation results.

Impact of Network Architecture. We further explore the impact of different network architectures on video segmentation methods. Table 5 shows the ablation results of Transformer-like (MiT [66]) and CNN-like (ResNet [14]) networks, and the analysis reveals that **1)** the performance ranking order of both methods (MATNet [79] and ours) is MiT-b2 > ResNet-101 > MiT-b1, and **2)** the proposed method outperforms MATNet (Note that results of MATNet are obtained by the CRF post-processing technique, while our results are not.) above 4.1% in terms of $\mathcal{J}\&\mathcal{F}$ Mean for the same network architecture. The above ablation results show that the large Transformer-like MiT-b2 [66] benefits from better visual perception and achieves better segmentation performance compared to CNN-like ResNet-101 [14].

4.3   Quantitative Results for UVOS

DAVIS-16. We compare the proposed model HFAN with SOTA methods on the public benchmark DAVIS-16 [47]. Table 6 shows the quantitative results. Our method outperforms all existing SOTA models by a significant margin on DAVIS-16. Specifically, our HFAN-small scores 86.7% $\mathcal{J}\&\mathcal{F}$ Mean and reaches 20.8 FPS in real-time speed. In contrast to RTNet [50], which employs both forward and backward optical flow and uses post-processing, HFAN-medium achieves 88.7% $\mathcal{J}\&\mathcal{F}$ Mean using only forward optical flow without any post-processing techniques. Compared with previous methods [79,22,69,75,50] using optical flow, our method exhibits significant superiority in terms of inference speed and segmentation accuracy. The main reason is that the FAM and FAT modules in HFAN perform feature alignment and adaptation for unaligned cross-modal features, allowing the decoder to utilize a more accurate feature representation. Quantitative results of different metrics demonstrate that our method achieves a nice trade-off between accuracy and speed in the UVOS task.

YouTube-Objects. To explore the universality of our proposed method to other video datasets, we perform validation experiments on the YouTube-Objects [49] test set without further fine-tuning its training set. The quantitative results of 10 categories in this dataset are shown in Table 7. Our method HFAN-small dose not reach SOTA across all categories but has better stability than other comparative methods. The proposed method is 2.0% higher than the second-best GraphMem [36] in terms of *average* $\mathcal{J}$ Mean. For 10 different object categories,

Table 6. Evaluation on DAVIS-16 [47]. 'small' and 'medium' indicate that the backbone networks of HFAN are MiT-b1 and MiT-b2, respectively. '†' means that the optical flow is used. 'PP' denotes post-processing. The three best scores are marked in red, blue and green for each metric, respectively. The inference speed (FPS) of each model contains all the necessary aspects for its generation of final results.

Method	Publication	PP	$\mathcal{J}$ Mean ↑	Recall ↑	Decay ↓	$\mathcal{F}$ Mean ↑	Recall ↑	Decay ↓	$\mathcal{J}\&\mathcal{F}$ Mean ↑	FPS ↑
PDB [55]	ECCV 2018	✓	77.2	90.1	0.9	74.5	84.4	-0.2	75.9	20.0
UOVOS† [80]	TIP 2019	✓	73.9	88.5	0.6	68.0	80.6	0.7	71.0	-
LSMO† [59]	IJCV 2019	✓	78.2	89.1	4.1	75.9	84.7	3.5	77.1	-
MotAdapt† [54]	ICRA 2019	✓	77.2	87.8	5.0	77.4	84.4	3.3	77.3	-
AGS [65]	CVPR 2019	✓	79.7	91.1	1.9	77.4	85.8	1.6	78.6	1.7
AGNN [63]	ICCV 2019	✓	80.7	94.0	0.0	79.1	90.5	0.0	79.9	1.9
COSNet [37]	CVPR 2019	✓	80.5	93.1	4.4	79.5	89.5	5.0	80.0	2.2
AnDiff [70]	ICCV 2019		81.7	90.9	2.2	80.5	85.1	0.6	81.1	2.8
PCSA [11]	AAAI 2020		78.1	90.0	4.4	78.5	88.1	4.1	78.3	110
EPO+† [1]	WACV 2020	✓	80.6	95.2	2.2	75.5	87.9	2.4	78.1	-
MATNet† [79]	AAAI 2020	✓	82.4	94.5	3.8	80.7	90.2	4.5	81.5	1.3
GraphMem [36]	ECCV 2020	✓	82.5	94.3	4.2	81.2	90.3	5.6	81.9	5.0
DFNet [77]	ECCV 2020	✓	83.4	94.4	4.2	81.8	89.0	3.7	82.6	3.6
3DCSeg [38]	BMVC 2020		84.2	95.8	7.4	84.3	92.4	5.5	84.2	4.5
F2Net [34]	AAAI 2021		83.1	95.7	0.0	84.4	92.3	0.8	83.7	10.0
FSNet† [22]	ICCV 2021	✓	83.4	94.5	3.2	83.1	90.2	2.6	83.3	12.5
AMC-Net† [69]	ICCV 2021	✓	84.5	96.4	2.8	84.6	93.8	2.5	84.6	-
TransportNet† [75]	ICCV 2021		84.5	-	-	85.0	-	-	84.8	3.6
RTNet† [50]	CVPR 2021	✓	85.6	96.1	-	84.7	93.8	-	85.2	-
Ours-small†(SS/MS)			86.2/87.1	96.7/96.8	4.6/4.8	87.1/87.7	95.5/95.3	2.3/2.5	86.7/87.4	20.8/2.5
Ours-medium†(SS/MS)	-		86.8/88.0	96.1/96.2	4.3/4.5	88.2/89.3	95.3/95.4	1.1/2.0	87.5/88.7	14.4/1.4

Table 7. Evaluation on YouTube-Objects [49]. The three best scores are marked in red, blue, and green for each object category over $\mathcal{J}$ Mean ↑.

Method	MOTAdapt [54]	LSMO [59]	LVO [58]	FSEG [21]	PDB [55]	SFL [6]	AGS [65]	COSNet [37]	AGNN [63]	MATNet [79]	AMCNet [69]	GraphMem [50]	RTNet [36]	Ours-small
Airplane	77.2	60.5	86.2	81.7	78.0	65.6	87.7	81.1	81.1	72.9	78.9	86.1	84.1	84.7
Bird	42.2	59.3	81.0	63.8	80.0	65.4	76.7	75.7	75.9	77.5	80.9	75.7	80.2	80.0
Boat	49.3	62.1	68.5	72.3	58.9	59.9	72.2	71.3	70.7	66.9	67.4	68.6	70.1	72.0
Car	68.6	72.3	69.3	74.9	76.5	64.0	78.6	77.6	78.1	79.0	82.0	82.4	79.5	76.1
Cat	46.3	66.3	58.8	68.4	63.0	58.9	69.2	66.5	67.9	73.7	69.0	65.9	71.8	76.0
Cow	64.2	67.9	68.5	68.0	64.1	51.2	64.6	69.8	69.7	67.4	69.6	70.5	70.1	71.2
Dog	66.1	70.0	61.7	69.4	70.1	54.1	73.3	76.8	77.4	75.9	75.8	77.1	71.3	76.9
Horse	64.8	65.4	53.9	60.4	67.6	64.8	64.4	67.4	67.3	63.2	63.0	72.2	65.1	71.0
Motorbike	44.6	55.5	60.8	62.7	58.4	52.6	62.1	67.7	68.3	62.6	63.4	63.8	64.6	64.3
Train	42.3	38.0	66.3	62.2	35.3	34.0	48.2	46.8	47.8	51.0	57.8	47.8	53.3	61.4
Average	58.1	64.3	67.5	68.4	65.5	57.1	69.7	70.5	70.8	69.0	71.1	71.4	71.0	73.4

the proposed method achieves its balanced performance over various challenging (*e.g.*, motion blur, occlusion, scale variation) video sequences. This is made possible by the sensible interaction of the proposed modules (FAM and FAT) for appearance and motion information.

Long-Videos. DAVIS [47] (**60+** frames per video sequence in average) only contains short-term video clips, while real-world videos tend to have more frames. To verify the performance of our HFAN in long-term video object segmentation, we evaluate it on the Long-Videos [32] val-set (approximate **2500** frames per video sequence). Table 8 shows the results under two types of supervision, SVOS and UVOS. By further observation, we can find that the proposed HFAN-medium has obtained the best result, achieving 81.7% over $\mathcal{J}\&\mathcal{F}$ Mean under the UVOS setting. Compared with the second-best method AGNN [63], our small model obtains an improvement of 7.0% on $\mathcal{J}\&\mathcal{F}$ Mean. Meanwhile, HFAN-medium

Table 8. Evaluation on Long-Videos [32]. The best results of SVOS and UVOS methods are marked in <u>underline</u> and **bold**, respectively.

Method	Supervision	$\mathcal{J}$			$\mathcal{F}$			$\mathcal{J}\&\mathcal{F}$ Mean ↑
		Mean ↑	Recall ↑	Decay ↓	Mean ↑	Recall ↑	Decay ↓	
RVOS [61]		10.2	6.7	13.0	14.3	11.7	<u>10.1</u>	12.2
A-GAME [23]	SVOS	50.0	58.3	39.6	50.7	58.3	45.2	50.3
STM [44]		79.1	88.3	11.6	79.5	90.0	15.4	79.3
AFB-URR [32]		<u>82.7</u>	<u>91.7</u>	<u>11.5</u>	<u>83.8</u>	<u>91.7</u>	13.9	<u>83.3</u>
3DCSeg [38]		34.2	38.6	11.6	33.1	28.1	15.6	33.7
MATNet [79]		66.4	73.7	10.9	69.3	77.2	10.6	67.9
AGNN [63]	UVOS	68.3	77.2	13.0	68.6	77.2	16.6	68.5
Ours-small		74.9	82.5	14.8	76.1	86.0	16.0	75.5
Ours-medium		**80.2**	**91.2**	**9.4**	**83.2**	**96.5**	**7.1**	**81.7**

Table 9. Evaluation on DAVIS [47] for VSOD. The best scores are marked in **bold**.

Method	FGRN [27]	LTSI [2]	RCR [68]	MBN [29]	SSAV [9]	PCSA [11]	DCFNet [76]	FSNet [22]	Ours-small	Ours-medium
S_α ↑	0.838	0.876	0.886	0.887	0.893	0.902	0.914	0.920	0.934	**0.938**
E_ξ^{max} ↑	0.917	0.957	0.947	0.966	0.948	0.961	-	0.970	**0.983**	0.983
F_β^{max} ↑	0.783	0.850	0.848	0.862	0.861	0.880	0.900	0.907	0.929	**0.935**
MAE ↓	0.043	0.034	0.027	0.031	0.028	0.022	0.016	0.020	0.009	**0.008**

achieves appealing results compared to SVOS methods. The results show that the temporal consistency provided by optical flow is also effective for long-term video object segmentation.

4.4 Quantitative Results for VSOD

The additional task VSOD, like UVOS, does not require first-frame annotation. To verify the performance of the proposed method on the VSOD setting, we perform a quantitative comparison with eight SOTA models on DAVIS [47].
Metrics. We employ four widely-used evaluation metrics including structure-measure S_α ($\alpha = 0.5$), max enhanced-alignment measure E_ξ^{max}, max F-measure F_β^{max} ($\beta^2 = 0.3$), and mean absolute error (MAE).
Results. As shown in Table 9, our HFAN outperforms all SOTA models. In particular, compared with DCFNet [76], S_α and F_β^{max} are improved by ∼2% and ∼3%, respectively. Compared to FSNet [22], HFAN achieves >1.3% perfor-mance gains on S_α, E_ξ^{max} and F_β^{max}, and reduces MAE by a factor of two. This significantly proves the adaptability of our method to similar tasks.

4.5 Qualitative Results

Fig. 5 shows qualitative results of our HFAN model. We select five videos from DAVIS-16 [47], YouTube-Objects [49] and Long-Videos [32] test sets. These videos consist of several challenging frame sequences (*e.g.*, fast motion, scale variation, interacting objects and occlusion). As shown in the top two rows,
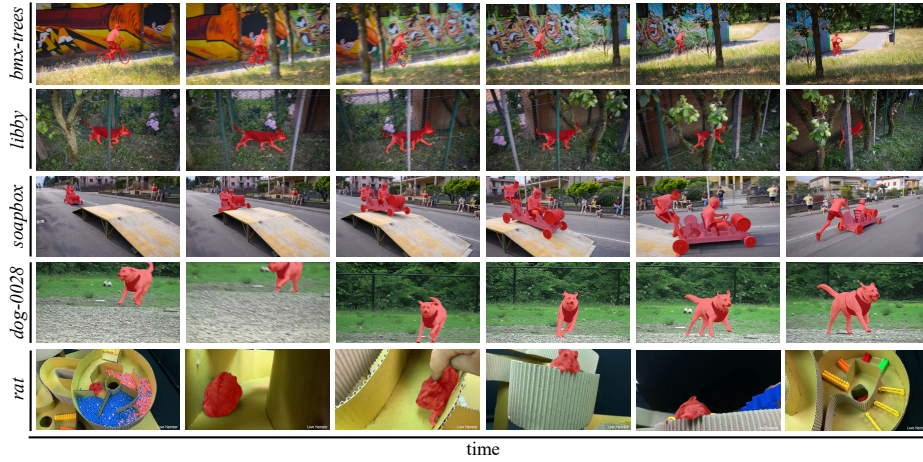
Fig. 5. Qualitative results on three challenging video clips over time. From top to bottom: *bmx-trees*, *libby* and *soapbox* from DAVIS-16 [47], *dog-0028* from YouTube-Objects [49], and *rat* from Long-Videos [32].

our method yields desirable results for dynamic, similar, and complex backgrounds. Moreover, our proposed model has an accurate prediction for the occlusion boundary. In the third and fourth rows, satisfactory segmentation results are acquired in the presence of large-scale variation and object interaction cases.

5   Conclusion

We present a hierarchical feature alignment network, termed as HFAN, for addressing the contextual mismatch between appearance and motion features in the UVOS task. Firstly, to address the mismatch of primary object positions between video frames and their corresponding optical flows, our proposed FAM module relies on sharing primary objects in images across modalities to align appearance and motion features. Subsequently, for tackling the modal mismatch problem between aligned feature maps, the FAT module is designed to construct a feature adaptation weight to automatically enhance cross-modal features. With the alignment and adaptation of appearance and motion features achieved by FAM and FAT, HFAN could achieve a more accurate object segmentation. Experimental results show that the proposed method achieves SOTA performance in the unsupervised video object segmentation task.

References

1. Akhter, I., Ali, M., Faisal, M., Hartley, R.: Epo-net: Exploiting geometric constraints on dense trajectories for motion saliency. In: WACV (2020)
2. Chen, C., Wang, G., Peng, C., Zhang, X., Qin, H.: Improved robust video saliency detection based on long-term spatial-temporal information. TIP (2019)
3. Chen, T., Yao, Y., Zhang, L., Wang, Q., Xie, G., Shen, F.: Saliency guided inter- and intra-class relation constraints for weakly supervised semantic segmentation. TMM (2022)
4. Chen, Y., Han, C., Wang, N., Zhang, Z.: Revisiting feature alignment for one-stage object detection. arXiv preprint arXiv:1908.01570 (2019)
5. Cheng, H.K., Tai, Y.W., Tang, C.K.: Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In: CVPR (2021)
6. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: ICCV (2017)
7. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation` (2020)
8. Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: Sstvos: Sparse spatiotemporal transformers for video object segmentation. In: CVPR (2021)
9. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: CVPR (2019)
10. Giordano, D., Murabito, F., Palazzo, S., Spampinato, C.: Superpixel-based video object segmentation using perceptual organization and location prior. In: CVPR (2015)
11. Gu, Y., Wang, L., Wang, Z., Liu, Y., Cheng, M.M., Lu, S.P.: Pyramid constrained self-attention network for fast video salient object detection. In: AAAI (2020)
12. Han, B., Davis, L.S.: Density-based multifeature background subtraction with support vector machine. TPAMI (2011)
13. Han, J., Ding, J., Li, J., Xia, G.S.: Align deep features for oriented object detection. TGRS (2021)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. Heo, Y., Koh, Y.J., Kim, C.S.: Guided interactive video object segmentation using reliability-based attention maps. In: CVPR (2021)
16. Huang, S., Lu, Z., Cheng, R., He, C.: Fapn: Feature-aligned pyramid network for dense image prediction. In: ICCV (2021)
17. Huang, Z., Wei, Y., Wang, X., Shi, H., Liu, W., Huang, T.S.: Alignseg: Feature-aligned segmentation networks. TPAMI (2021)
18. Hui, T., Huang, S., Liu, S., Ding, Z., Li, G., Wang, W., Han, J., Wang, F.: Collaborative spatial-temporal modeling for language-queried video actor segmentation. In: CVPR (2021)
19. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR (2017)
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
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: CVPR (2017)

22. Ji, G.P., Fu, K., Wu, Z., Fan, D.P., Shen, J., Shao, L.: Full-duplex strategy for video object segmentation. In: ICCV (2021)
23. Johnander, J., Danelljan, M., Brissman, E., Khan, F.S., Felsberg, M.: A generative appearance model for end-to-end video object segmentation. In: CVPR (2019)
24. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: ACCV (2018)
25. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NeurIPS (2011)
26. Lao, D., Zhu, P., Wonka, P., Sundaramoorthi, G.: Flow-guided video inpainting with scene templates. In: ICCV (2021)
27. Li, G., Xie, Y., Wei, T., Wang, K., Lin, L.: Flow guided recurrent neural encoder for video salient object detection. In: CVPR (2018)
28. Li, H., Chen, G., Li, G., Yu, Y.: Motion guided attention for video salient object detection. In: ICCV (2019)
29. Li, S., Seybold, B., Vorobyov, A., Lei, X., Kuo, C.C.J.: Unsupervised video object segmentation with motion-based bilateral networks. In: ECCV (2018)
30. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: CVPR (2019)
31. Li, X., You, A., Zhu, Z., Zhao, H., Yang, M., Yang, K., Tan, S., Tong, Y.: Semantic flow for fast and accurate scene parsing. In: ECCV (2020)
32. Liang, Y., Li, X., Jafari, N., Chen, Q.: Video object segmentation with adaptive feature bank and uncertain-region refinement. In: NeurIPS (2020)
33. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
34. Liu, D., Yu, D., Wang, C., Zhou, P.: F2net: Learning to focus on the foreground for unsupervised video object segmentation. In: AAAI (2021)
35. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
36. Lu, X., Wang, W., Danelljan, M., Zhou, T., Shen, J., Van Gool, L.: Video object segmentation with episodic graph memory networks. In: ECCV (2020)
37. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: CVPR (2019)
38. 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: BMVC (2020)
39. Mao, Y., Wang, N., Zhou, W., Li, H.: Joint inductive and transductive learning for video object segmentation. In: ICCV (2021)
40. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: ICCV (2019)
41. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
42. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. TPAMI (2013)
43. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Fast user-guided video object segmentation by interaction-and-propagation networks. In: CVPR (2019)
44. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV (2019)
45. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV (2013)
46. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z.,

Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)

47. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)

48. Perazzi, F., Wang, O., Gross, M., Sorkine-Hornung, A.: Fully connected object proposals for video segmentation. In: ICCV (2015)

49. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: CVPR (2012)

50. Ren, S., Liu, W., Liu, Y., Chen, H., Han, G., He, S.: Reciprocal transformations for unsupervised video object segmentation. In: CVPR (2021)

51. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV (2020)

52. Seong, H., Oh, S.W., Lee, J.Y., Lee, S., Lee, S., Kim, E.: Hierarchical memory matching network for video object segmentation. In: ICCV (2021)

53. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: CVPR (2016)

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: ICRA (2019)

55. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper con-vlstm for video salient object detection. In: ECCV (2018)

56. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: CVPR (2018)

57. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020)

58. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: ICCV (2017)

59. Tokmakov, P., Schmid, C., Alahari, K.: Learning to segment moving objects. IJCV (2019)

60. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)

61. Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F., Giro-i Nieto, X.: Rvos: End-to-end recurrent network for video object segmentation. In: CVPR (2019)

62. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: ICCV (2019)

63. Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L.: Zero-shot video object segmentation via attentive graph neural networks. In: ICCV (2019)

64. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: CVPR (2015)

65. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: CVPR (2019)

66. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS (2021)

67. 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: ECCV (2018)

68. Yan, P., Li, G., Xie, Y., Li, Z., Wang, C., Chen, T., Lin, L.: Semi-supervised video salient object detection using pseudo-labels. In: ICCV (2019)
69. Yang, S., Zhang, L., Qi, J., Lu, H., Wang, S., Zhang, X.: Learning motion-appearance co-attention for zero-shot video object segmentation. In: ICCV (2021)
70. Yang, Z., Wang, Q., Bertinetto, L., Hu, W., Bai, S., Torr, P.H.: Anchor diffusion for unsupervised video object segmentation. In: ICCV (2019)
71. Yao, Y., Chen, T., Xie, G.S., Zhang, C., Shen, F., Wu, Q., Tang, Z., Zhang, J.: Non-salient region object mining for weakly supervised semantic segmentation. In: CVPR (2021)
72. Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., Tang, Z.: Jo-src: A contrastive approach for combating noisy labels. In: CVPR (2021)
73. Yao, Y., Zhang, J., Shen, F., Hua, X., Xu, J., Tang, Z.: Exploiting web images for dataset construction: A domain robust approach. TMM (2017)
74. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: ECCV (2020)
75. Zhang, K., Zhao, Z., Liu, D., Liu, Q., Liu, B.: Deep transport network for unsupervised video object segmentation. In: ICCV (2021)
76. Zhang, M., Liu, J., Wang, Y., Piao, Y., Yao, S., Ji, W., Li, J., Lu, H., Luo, Z.: Dynamic context-sensitive filtering network for video salient object detection. In: ICCV (2021)
77. Zhen, M., Li, S., Zhou, L., Shang, J., Feng, H., Fang, T., Quan, L.: Learning discriminative feature with crf for unsupervised video object segmentation. In: ECCV (2020)
78. Zhou, T., Li, J., Li, X., Shao, L.: Target-aware object discovery and association for unsupervised video multi-object segmentation. In: CVPR (2021)
79. Zhou, T., Wang, S., Zhou, Y., Yao, Y., Li, J., Shao, L.: Motion-attentive transition for zero-shot video object segmentation. In: AAAI (2020)
80. Zhuo, T., Cheng, Z., Zhang, P., Wong, Y., Kankanhalli, M.: Unsupervised online video object segmentation with motion property understanding. TIP (2019)