Dual Prototype Attention for Unsupervised Video Object Segmentation

Suhwan Cho[1,*] Minhyeok Lee[1,*] Seunghoon Lee[1] Dogyoon Lee[1]
Heeseung Choi[2] Ig-Jae Kim[2] Sangyoun Lee[1]

[1] Yonsei University
[2] Korea Institute of Science and Technology (KIST)

Abstract

Unsupervised video object segmentation (VOS) aims to detect and segment the most salient object in videos. The primary techniques used in unsupervised VOS are 1) the collaboration of appearance and motion information; and 2) temporal fusion between different frames. This paper proposes two novel prototype-based attention mechanisms, inter-modality attention (IMA) and inter-frame attention (IFA), to incorporate these techniques via dense propagation across different modalities and frames. IMA densely integrates context information from different modalities based on a mutual refinement. IFA injects global context of a video to the query frame, enabling a full utilization of useful properties from multiple frames. Experimental results on public benchmark datasets demonstrate that our proposed approach outperforms all existing methods by a substantial margin. The proposed two components are also thoroughly validated via ablative study. Code and models are available at https://github.com/Hydragon516/DPA.

1. Introduction

Video object segmentation (VOS) is a fundamental task in computer vision. Given a video sequence as input, the objective is to segment objects for the entire frames. It can be divided into several categories depending on how the objects to be detected are defined. In this study, we deal with the unsupervised setting, i.e., detecting and segmenting the most salient object in a video sequence without any external guidance such as target mask or reference text.

In unsupervised VOS, collaboration of different modalities and different frames is widely adopted. As salient object usually shows distinctive movements compared to the background, existing approaches including MATNet [34], FSNet [4], and HFAN [14] leverage motion cues in addition to appearance cues. For each video frame, an RGB image and an optical flow map generated by pre-trained optical flow estimation models are used as input to exploit appearance and motion information simultaneously. In order to fuse the multi-modal cues, they mainly focus on designing a reciprocity framework that blends the embedded features obtained from each modality. On the other hand, there are also studies, such as COSNet [11], AGNN [25], and AD-Net [29], that focus on exploiting temporal coherence of a video. They transfer information of the initial frame to each query frame or iteratively refines each frame via temporally connecting different frames.

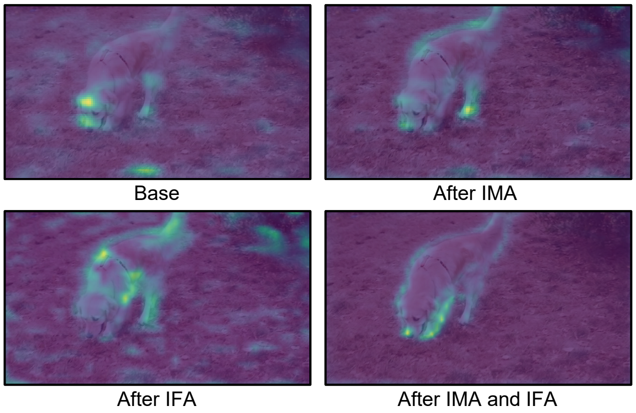

Figure 1. Visualized feature maps after applying IMA and IFA.

However, existing modality fusion methods and temporal aggregation methods have significant limitations. First, conventional multi-modality solutions are not carefully designed to be robust against various situations. As they fuse multi-modal cues via a direct summation, concatenation, or modulating channel weights, respective cues can act as noise if their quality is not reliable. Second, existing temporal fusion methods do not fully consider global context of a video or require high computational cost. They only consider the initial frame as an anchor frame that provides external guidance or perform iterative refinement over all frames, which severely degrades their efficiency.

[*] These authors contribute equally to this work.

In this paper, we propose two novel modules to overcome the aforementioned limitations. First, we introduce an inter-modality attention (IMA) to refine cues of respective modalities by densely integrating context information of both modalities. For each modality, useful cues are first extracted and refined to provide valuable supervision to each other. Then, instead of a naive fusion, the features of each modality is adaptively allocated to other modality based on mutual feature propagation. Second, we introduce inter-frame attention (IFA) to leverage global context of a video without requiring heavy computational cost. Once a video sequence is given as input, a designated number of frames are first sampled from the entire video sequence and features from those frames are stored in an external memory bank. When predicting each frame, the stored features are adaptively propagated to the query frames to provide overall properties of a video. Finally, we extend the proposed two modules by incorporating a prototype framework. Through converting pixel-level information to prototype-level information, more reliable and comprehensive cues can be leveraged, as each prototype is constructed as having spatial structure knowledge of the scenes.

We evaluate our proposed approach on three popular benchmark datasets, DAVIS 2016 [15] validation set, FBMS [13] test set, and YouTube-Objects [16] dataset. On all of them, our method surpasses all existing methods by a substantial margin. Extensive experiments are also conducted to demonstrate the effectiveness of each proposed component.

Our main contributions can be summarized as follows:

- We propose dual attention modules, IMA and IFA, to effectively leverage multi-modality fusion and temporal aggregation for unsupervised VOS.
- We incorporate a prototype framework into the proposed attention mechanisms to further improve their efficacy by refining the source information.
- On all public benchmark datasets, our approach sets a new state-of-the-art performance, while avoiding high computational complexity.

2. Related Work

Multi-modality fusion. In unsupervised VOS, two-stream architectures that jointly leverage appearance cues and motion cues have been attracting extensive attention. MATNet [34] designs a two-stream encoder that utilizes an RGB image and an optical flow map to enhance spatio-temporal object representation. RTNet [17] proposes a reciprocal transformation network to identify and segment primary objects in videos. FSNet [4] proposes a full-duplex strategy to effectively fuse RGB images and optical flow maps; specifically, a bidirectional interaction module is used to ensure the mutual restraint between appearance and motion cues. AMC-Net [28] proposes a co-attention gate that modulates the impacts of appearance and motion cues. Based on the learned weights, appearance and motion information can be leveraged adaptively. TransportNet [31] establishes the correspondence between appearance and motion cues while suppressing the distractions via optimal structural matching. HFAN [14] proposes a hierarchical feature alignment network that aligns the object positions using the appearance and motion features. The cross-modal mismatch can be mitigated by adapting the aligned features. PMN [7] stores prototypes of appearance cues as well as and motion cues to fully leverage multiple modalities. TMO [3] optionally employs motion stream on top of appearance stream for robust learning of the motion encoder. However, the performance of these methods can be further improved as the modality fusion is performed based on a simple summation or concatenation.

Temporal aggregation. Unlike existing two-stream methods, some studies focus on fully exploiting the temporal coherence of a video. COSNet [11] employs the co-attention layers to capture global correlations and scene context by propagating semantic information in the reference frames to the query frame. AGNN [25] builds fully connected graphs to represent frames as nodes and relations between those frames as edges. Rich relations between arbitrary frames can be obtained through parametric message passing. ADNet [29] and F2Net [9] regard the initial frame of a video as a reference frame, and leverages the reference frame information for query frame prediction. IMP [8] iteratively propagates the segmentation mask of an easy reference frame to other frames by using a pre-trained semi-supervised VOS algorithm. These methods can capture temporal coherence in a video, but still suffer from certain problems such as the global context of a video not being completely leveraged and requiring heavy computational complexity owing to the iterative inferring process.

3. Approach

3.1. Problem Formulation

The goal of an unsupervised VOS algorithm is to identify the most salient object for all frames of a video. Following common protocol in the VOS community, we collaboratively use RGB images and optical flow maps as the input of our network. The network output is binary segmentation masks that have the same resolution as the input information. RGB images, optical flow maps, and output segmentation masks are denoted as $I := \{I^0, I^1, ..., I^{L-1}\}$, $F := \{F^0, F^1, ..., F^{L-1}\}$, and $O := \{O^0, O^1, ..., O^{L-1}\}$, respectively, where L is the number of total frames.

3.2. Network Architecture

Following existing two-stream approaches for unsupervised VOS, such as MATNet [34], FSNet [4], HFAN [14],

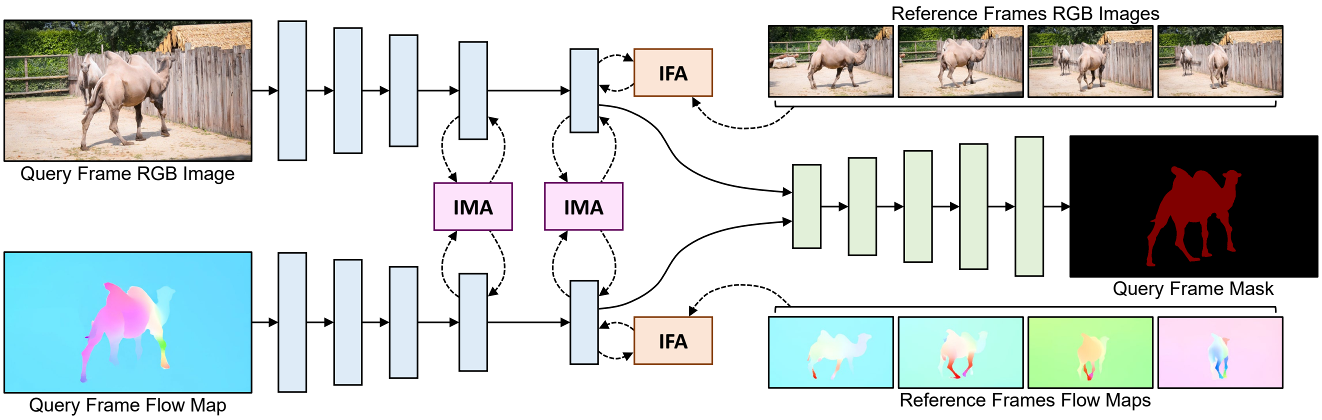

Figure 2. Architecture of our proposed network. Based on a two-stream encoder-decoder architecture, IMA and IFA modules are employed. For simplicity, skip connections between encoding blocks and decoding blocks are omitted in the illustration.

PMN [7], and TMO [3], our network is designed based on a simple two-stream encoder-decoder architecture. As both RGB image and optical flow map are given as input, two separate encoders are adopted. The features obtained from those encoders are decoded using a decoder that outputs a binary segmentation mask. In the middle of the encoding and decoding processes, the proposed IMA and IFA are adopted for mutual modality fusion and temporal cue aggregation, respectively. The visualized pipeline of our network can be found at Figure 2.

3.3. Inter-Modality Attention (IMA)

Existing two-stream approaches, including the aforementioned methods, focus on fusing the multi-modal cues, i.e., appearance cues and motion cues. However, the fusion has been implemented using a simple summation or concatenation, resulting in unstable cue generation, particularly in challenging scenarios. To enhance multi-modality fusion for unsupervised VOS, we propose IMA to densely and thoroughly exchange information between appearance and motion cues based on prototype attention mechanism. The proposed IMA consists of three parts: prototype generation, self-correlation calculation, and mutual feature refinement. In Figure 3, we visualize the architecture of IMA.

Prototype generation. Inspired by OCR [30], we first generate prototypes based on learnable object regions. For each input feature map $X \in \mathbb{R}^{C\times HW}$, soft object region $S \in [0,1]^{C\times HW}$ is calculated by applying a simple channel-wise softmax operation as

$$S = Softmax(X)\ . \tag{1}$$

Considering the backbone encoder is learned with large-scale ImageNet [6], each channel in S already contains the clustering ability that helps spatially separate the input feature map into semantic parts. Then, using X and S, prototypes $P1 \in \mathbb{R}^{C\times C'}$ are obtained as

$$P1 = X \otimes S^T\ , \tag{2}$$

where $\otimes$ indicates matrix multiplication. In $P1$, C' prototypes with a channel size of C are contained. Note that C' is equal to C, but is adopted for better clarification.

Self-correlation calculation. In order to incorporate the information of the constructed prototypes $P1$ into the input features X, we first calculate self-correlation map $\Psi 1 \in [-1,1]^{C'\times HW}$ for each modality as

$$\Psi 1 = \mathcal{N}(P1)^T \otimes \mathcal{N}(X)\ , \tag{3}$$

where $\mathcal{N}$ indicates channel L2 normalization. The generated $\Psi 1$ represents the cosine similarities between each prototype and each input feature. Here, we embed key and value features of each modality from $\Psi 1$ instead of directly extracting them from input features X. To validate the incorporation of a prototype framework helps better fusion of two modalities, we conduct an ablation study regarding the key-value extraction (normal embedding vs. prototype-based self-correlation calculation) in Section 4.3.

Mutual feature refinement. After generating the self-correlation maps for each modality, the information of each modality can now be effectively transferred to each other using a cross attention mechanism. From $\Psi 1$, the key features $K \in \mathbb{R}^{C'\times HW}$ and value features $V \in \mathbb{R}^{C'\times HW}$ are first calculated as

$$\begin{aligned} K &= \sigma_K(\Psi 1) \\ V &= \sigma_V(\Psi 1)\ , \end{aligned} \tag{4}$$

where σ indicates a pixel-wise (HW-wise) fully connected layer. Next, the correspondence map $\Phi 1 \in \mathbb{R}^{C'\times C'}$ of the

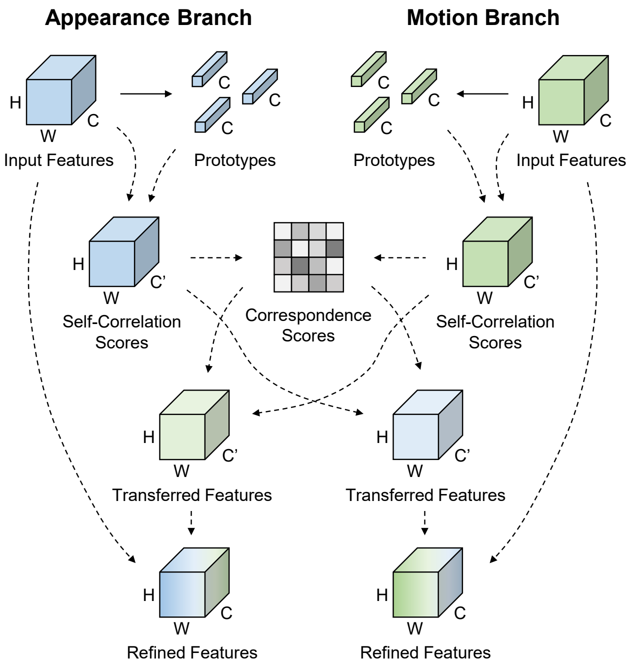

Figure 3. Visualized pipeline of IMA.

key features from each modality are computed as

$$\Phi 1 = K_A \otimes K_M^T \ . \tag{5}$$

As each channel in K contains certain properties of prototypes, relations between every constructed prototype are contained in $\Phi 1$. Based on $\Phi 1$, V from one modality are propagated to the other modality as

$$\begin{aligned} T_A &= Softmax(\Phi 1) \otimes V_M \\ T_M &= Softmax(\Phi 1^T) \otimes V_A \ , \end{aligned} \tag{6}$$

where T indicates the transferred features. Finally, the input features are concatenated with the transferred features, and the refined features $X' \in \mathbb{R}^{C \times HW}$ are obtained as

$$\begin{aligned} X'_A &= Conv(X_A \oplus T_A) \\ X'_M &= Conv(X_M \oplus T_M) \ , \end{aligned} \tag{7}$$

where $\oplus$ indicates channel concatenation and $Conv$ is a convolutional layer for feature refinement. Through this mutual refinement process, each modality can appropriately reflect semantic information from another modality.

3.4. Inter-Frame Attention (IFA)

As much as fusing the features of multiple modalities is important, exploiting temporal coherence of a video is also an effective strategy for unsupervised VOS. Existing approaches, such as COSNet [11], AD-Net [29], F2Net [9], and IMP [8], design their network architectures to utilize this temporal coherence. However, they are either time-consuming owing to their iterative workflows or not fully leveraging the global context of a video. To overcome these limitations, we propose IFA to efficiently leverage the temporal coherence of a video. The proposed IFA consists of three parts: reference frame sampling, prototype generation, and temporal propagation. In Figure 4, we visualize the architecture of IFA.

Reference frame sampling. The objective of IFA is to collect and store the global context of a video and propagate it to each query frame. To efficiently obtain the global context, we sample the frames in a video and store only the sampled frames rather than storing all frames. As the sampling method, we adopt uniform sampling strategy, i.e., frames are sampled while keeping the intervals identical. For example, if we want to sample N frames from L-length video, the sampled frames are defined as $\{I^k\}_{i=0}^{N-1}$ where

$$k = \lfloor \frac{i * (L-1)}{N-1} \rfloor \ . \tag{8}$$

Prototype generation. Before transferring semantic context of the reference frames to the query frame, we first transform the input features $Y \in \mathbb{R}^{D \times HW}$ to prototypes $P2 \in \mathbb{R}^{D \times D'}$ similar to *prototype generation* in Section 3.3. Y can be directly used instead of $P2$ for attention embedding in later step, but we use $P2$ to obtain better feature representations. The effects of this feature-prototype conversion process is reported in Section 4.3.

Temporal context propagation. After constructing prototypes, we extract key features $K \in \mathbb{R}^{D \times ND'}$ and value features $V \in \mathbb{R}^{D \times ND'}$ from the reference frames' prototypes, and query features $Q \in \mathbb{R}^{D \times D'}$ from the query frame's prototypes. Here, all embedding processes are implemented for each frame separately. The correspondence map $\Phi 2 \in \mathbb{R}^{D' \times ND'}$ can be obtained as

$$\Phi 2 = Q^T \otimes K \ . \tag{9}$$

Based on $\Phi 2$, the context of the reference frames is adaptively read and stored in read features $R \in \mathbb{R}^{D \times D'}$ as

$$R = (Softmax(\Phi 2) \otimes V^T)^T \ . \tag{10}$$

The generated R has D' prototypes with feature size of D, which contain information of the sampled frames. As it does not have spatial information related to the query frame, it cannot be directly leveraged for the feature fusion process. Therefore, we calculate the correlation scores $\Psi 2 \in [-1, 1]^{D' \times HW}$ between the query frame's Y and R to force the temporally transferred information to have the same spatial size as the input features, as follows:

$$\Psi 2 = \mathcal{N}(R)^T \otimes \mathcal{N}(Y) \ . \tag{11}$$

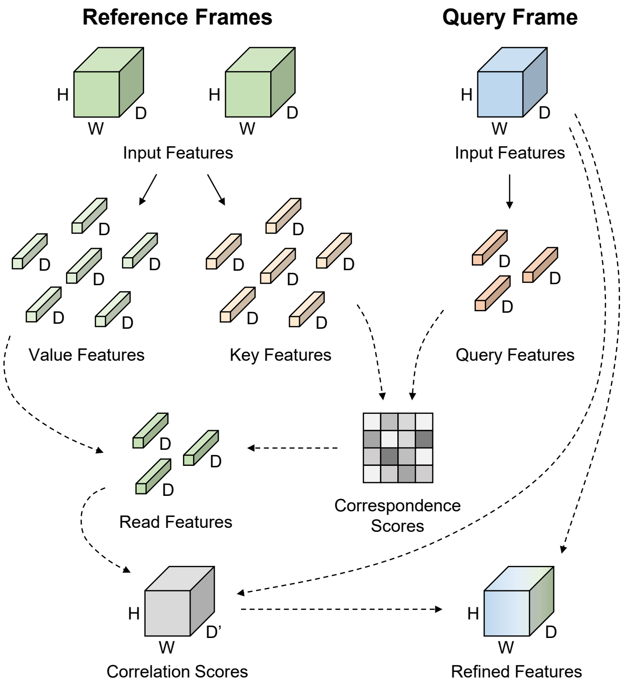

Figure 4. Visualized pipeline of IFA.

As Y and R have the same spatial size, feature fusion between them can now be easily achieved. The refined features $Y' \in \mathbb{R}^{D \times HW}$ for the query frame is defined as

$$Y' = Conv(Y \oplus \Psi 2)\ . \tag{12}$$

By employing the proposed IFA, semantic context of the reference frames is propagated to the query frame for reliable cue generation. In particular, when reliable information cannot be obtained from a single frame owing to difficulties such as occlusions, the use of IFA can effectively lead to stable functioning of a system.

3.5. Implementation Details

Optical flow map. Following a common protocol for two-stream approaches in unsupervised VOS, we generate optical flow maps from a pre-trained optical flow estimation model. The generated two-channel motion flow maps are converted to three-channel RGB flow maps and then saved in advance. As an optical flow estimation model, we adopt RAFT [23] pre-trained on the Sintel [1] dataset.

Network design. We adopt VGG-16 [20] as our backbone encoder for the appearance branch and motion branch. The encoded features are refined using IMA and IFA, where IMA is adopted for the fourth and fifth encoding blocks and IFA is adopted at the fifth encoding block. Note that IMA and IFA are separately adopted in the fifth encoding block, that is, they are employed in a parallel manner and then fused later. After refining the encoded features using IMA and IFA, an ASPP module [2] is applied to obtain stronger feature representations. The decoder takes the features from the ASPP module as its input, and gradually refines those features using lower-level features from the encoder.

Two-stage network training. Following previous methods, such as F2Net [9], RTNet [17], FSNet [4], and PMN [7], we train our network using multiple steps. As the first step, a salient object detection dataset DUTS [24] is adopted to pre-train the model on large-scale data. Both DUTS training set and test set are used as our training dataset. As it is an image-level dataset, only RGB images are available. Therefore, we only train the appearance branch and copy the learned parameters to the motion branch after the pre-training is done. Then, the entire model is trained on the DAVIS 2016 [15] training set and YouTube-VOS 2018 [27] training set with both appearance branch and motion branch. If a video sequence contains multiple objects, we regard them as a single object to obtain binary ground truth masks. Training snippets are randomly sampled from the DAVIS 2016 training set and YouTube-VOS 2018 training set with the same probabilities. The length of each snippet is fixed at four frames.

Training details. For network optimization, we use cross-entropy loss and the Adam optimizer [5]. The learning rate is decayed from 1e-4 to 1e-5 using the cosine annealing scheduler [10], and the batch size is set to 16. For network training, two GeForce RTX 3090 GPUs are used.

4. Experiments

In Section 4.1 and Section 4.2, the datasets and metrics used in this study are first introduced. Each proposed component is analyzed in Section 4.3. Quantitative and qualitative comparison can be found at Section 4.4 and Section 4.5, respectively. Our method is abbreviated as DPA.

4.1. Datasets

In this study, we use three datasets for network training: DUTS [24] dataset, DAVIS 2016 [15] training set, and YouTube-VOS 2018 [27] training set; and three datasets for network testing: DAVIS 2016 validation set, FBMS [13] test set, and YouTube-Objects [16] dataset.

4.2. Evaluation Metrics

We employ three evaluation metrics in this study: region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and their average $\mathcal{G}$. $\mathcal{J}$ and $\mathcal{F}$ can be calculated as follows:

$$\mathcal{J} = \left| \frac{M_{gt} \cap M_{pred}}{M_{gt} \cup M_{pred}} \right|\ , \tag{13}$$

$$\mathcal{F} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\ . \tag{14}$$

Table 1. Ablation study on the proposed components. P and N indicate the use of prototype embedding and the number of reference frames used in IFA, respectively.

Version	IMA	IFA	N	$\mathcal{G}_\mathcal{M}$	$\mathcal{J}_\mathcal{M}$	$\mathcal{F}_\mathcal{M}$
I	×	×	-	83.4	83.2	83.5
II	w/ P	×	-	85.9	85.4	86.3
III	×	w/ P	4	85.4	85.0	85.8
IV	w/ P	w/ P	4	86.9	86.3	87.4
V	w/ P	w/ P	1	86.3	85.8	86.9
VI	w/ P	w/ P	2	86.5	86.0	87.1
VII	w/ P	w/ P	3	86.8	86.2	87.5
VIII	w/ P	w/ P	5	86.9	86.3	87.5
IX	w/o P	w/ P	4	86.1	85.5	86.6
X	w/ P	w/o P	4	86.3	85.7	87.1
XI	w/o P	w/o P	4	85.3	84.6	86.0

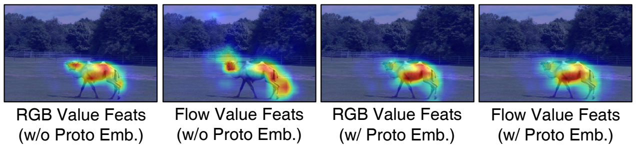
RGB Value Feats (w/o Proto Emb.)
Flow Value Feats (w/o Proto Emb.)
RGB Value Feats (w/ Proto Emb.)
Flow Value Feats (w/ Proto Emb.)

Figure 5. Visualized activation maps of different IMA versions.

Table 2. Cost analysis of the proposed components.

Version	IMA	IFA	Param #	Time (s)	$\mathcal{G}_\mathcal{M}$
I			39.5M	0.0164	83.4
II	✓		41.5M	0.0183	85.9
III		✓	40.5M	0.0322	85.4
IV	✓	✓	43.5M	0.0414	86.9

For the DAVIS 2016 [15] validation set, $\mathcal{G}$, $\mathcal{J}$, and $\mathcal{F}$ are used for network evaluation, while only $\mathcal{J}$ is used for the FBMS [13] test set and YouTube-Objects [16] dataset.

4.3. Analysis

To verify the effectiveness of the proposed components, we perform an ablation study on them, the results of which are present in Table 1 and Table 2. Note that the models are trained and tested with 352×352-resolution videos and evaluated on the DAVIS 2016 validation set.

Use of IMA and IFA. To compare the model performance with and without IMA and IFA, we compare model I, II, III, and IV in Table 1. As presented in the table, IMA and IFA both bring significant performance improvements to the baseline model. When IMA is employed alone, 2.5% improvements on $\mathcal{G}_\mathcal{M}$ are obtained, which implies that identifying inter-relationships between RGB images and optical flow maps via cross attention is effective for blending two distinct streams. IFA also brings meaningful improvements, $\mathcal{G}_\mathcal{M}$ score of 2.0%, backing up the need for global observation for specifying the primary objects. If IMA and IFA are used together, outstanding performance is obtained, $\mathcal{G}_\mathcal{M}$ of 86.9%. This indicates that IMA and IFA can constructively compensate each other as they focus on separate problems lying in two-stream unsupervised VOS.

Qualitative effects of adopting IMA and IFA can also be found at Figure 1. In the figure, feature maps of various embedding stages are visualized. We compare four feature maps, i.e., feature maps before IMA and IFA, feature maps after IMA, feature maps after IFA, and feature maps after IMA and IFA. It can be seen that IMA and IFA are both effective for capturing and specifying the salient objects (clearer edges and higher confidence for object regions). The joint use of IMA and IFA is even better than using IMA or IFA alone, validating the compatibility.

Number of reference frames. As described, IFA can take an arbitrary number of frames as reference frames given that it is based on an attention mechanism. To determine the optimal number of frames, we compare models with different number of reference frames. As shown in model V, VI, VII, and VIII in Table 1, employing more reference frames generally leads to a higher segmentation performance. This proves that as the amount of information from a video increases, the model becomes more generalized and robust against challenges such as occlusion. However, if the number of reference frames is larger than three, the performance gain is not satisfactory considering additional computations. In other words, semantic cues obtained from three representative frames are generally sufficient to encompass the global properties of a video.

Prototype embedding. In IMA and IFA, we employ a prototype framework before the attention mechanism as a preprocessing step. In Table 1, we quantitatively compare the model versions with and without this protocol. For both IMA and IFA, it brings meaningful improvements, demonstrating its effectiveness as a feature refinement tool. In Figure 5, we also visually compare the value feature maps with input features and self-correlation scores as the embedding source. As discussed in TMO [3], when there is a difference in quality between two domains, the information from the less reliable domain can act as noise. Considering this, applying prototype embedding significantly increases the stability of the network, as the features from both domains focus on the same regions. The qualitative results align with improvements in quantitative performance.

Cost analysis. In Table 2, we compare the number of parameters and inference time of different model versions. As shown in the table, IMA and IFA only introduce a small increase in the number of parameters compared to the baseline model. While there is some inference time slowdown due to the incremental computational cost, the performance gain is substantial considering the trade-offs.

Table 3. Quantitative evaluation on the DAVIS 2016 validation set and FBMS test set. OF and PP indicate the use of optical flow estimation models and post-processing techniques, respectively. * denotes speed calculated on our hardware.

						DAVIS 2016			FBMS
Method	Publication	Resolution	OF	PP	fps	$\mathcal{G}_\mathcal{M}$	$\mathcal{J}_\mathcal{M}$	$\mathcal{F}_\mathcal{M}$	$\mathcal{J}_\mathcal{M}$
PDB [21]	ECCV'18	473×473		✓	20.0	75.9	77.2	74.5	74.0
MOTAdapt [19]	ICRA'19	-		✓	-	77.3	77.2	77.4	-
AGS [26]	CVPR'19	473×473		✓	10.0	78.6	79.7	77.4	-
COSNet [11]	CVPR'19	473×473		✓	-	80.0	80.5	79.4	75.6
AD-Net [29]	ICCV'19	480×854		✓	4.00	81.1	81.7	80.5	-
AGNN [25]	ICCV'19	473×473		✓	3.57	79.9	80.7	79.1	-
MATNet [34]	AAAI'20	473×473	✓	✓	20.0	81.6	82.4	80.7	76.1
WCS-Net [32]	ECCV'20	320×320			33.3	81.5	82.2	80.7	-
DFNet [33]	ECCV'20	-		✓	3.57	82.6	83.4	81.8	-
3DC-Seg [12]	BMVC'20	480×854		✓	4.55	84.5	84.3	84.7	-
F2Net [9]	AAAI'21	473×473			10.0	83.7	83.1	84.4	77.5
RTNet [17]	CVPR'21	384×672	✓	✓	-	85.2	85.6	84.7	-
FSNet [4]	ICCV'21	352×352	✓	✓	12.5	83.3	83.4	83.1	-
TransportNet [31]	ICCV'21	512×512	✓		12.5	84.8	84.5	85.0	78.7
AMC-Net [28]	ICCV'21	384×384	✓	✓	17.5	84.6	84.5	84.6	76.5
D^{2}Conv3D [18]	WACV'22	480×854			-	86.0	85.5	86.5	-
IMP [8]	AAAI'22	-			1.79	85.6	84.5	86.7	77.5
HFAN [14]	ECCV'22	512×512	✓		11.0*	$\underline{87.0}$	$\underline{86.6}$	87.3	-
PMN [7]	WACV'23	352×352	✓		$\underline{41.3}$*	85.9	85.4	86.4	77.7
TMO [3]	WACV'23	384×384	✓		**43.2***	86.1	85.6	86.6	79.9
OAST [22]	ICCV'23	384×640	✓		-	85.9	85.4	86.3	$\underline{81.9}$
DPA		352×352	✓		24.2*	86.9	86.3	$\underline{87.4}$	81.2
DPA		512×512	✓		19.5*	**87.6**	**86.8**	**88.4**	**83.4**

Table 4. Quantitative evaluation on the YouTube-Objects dataset. Performance is reported using the $\mathcal{J}$ mean.

Method	Aeroplane	Bird	Boat	Car	Cat	Cow	Dog	Horse	Motorbike	Train	Mean
AGS [26]	**87.7**	76.7	**72.2**	78.6	69.2	64.6	73.3	64.4	62.1	48.2	69.7
COSNet [11]	81.1	75.7	$\underline{71.3}$	77.6	66.5	69.8	76.8	$\underline{67.4}$	67.7	46.8	70.5
AGNN [25]	71.1	75.9	70.7	78.1	67.9	69.7	$\underline{77.4}$	67.3	$\underline{68.3}$	47.8	70.8
MATNet [34]	72.9	77.5	66.9	79.0	$\underline{73.7}$	67.4	75.9	63.2	62.6	51.0	69.0
WCS-Net [32]	81.8	$\underline{81.1}$	67.7	79.2	64.7	65.8	73.4	**68.6**	**69.7**	49.2	70.5
RTNet [17]	84.1	80.2	70.1	$\underline{79.5}$	71.8	$\underline{70.1}$	71.3	65.1	64.6	$\underline{53.3}$	71.0
AMC-Net [28]	78.9	80.9	67.4	**82.0**	69.0	69.6	75.8	63.0	63.4	**57.8**	71.1
TMO [3]	85.7	80.0	70.1	78.0	73.6	**70.3**	76.8	66.2	58.6	47.0	$\underline{71.5}$
DPA	$\underline{87.5}$	**85.6**	70.1	77.7	**81.2**	69.0	**81.8**	61.9	62.1	51.3	**73.7**

4.4. Quantitative Results

In Table 3 and Table 4, we present quantitative comparison between our proposed method and existing state-of-the-art methods on the DAVIS 2016 [15] validation set, FBMS [13] test set, and YouTube-Objects [16] dataset. Note that all methods are evaluated using CNN-based backbone networks for a fair comparison. Our model is tested on a single GeForce RTX 2080 Ti GPU.

DAVIS 2016. On the DAVIS 2016 validation set, the methods using the estimated optical flow maps generally exhibit notable performance. Specifically, recent methods achieve high performance based on a two-stream architecture without the need for post-processing steps. Our proposed DPA outperforms all other methods at the same input resolution. With a resolution of 352×352, it achieves a $\mathcal{G}_\mathcal{M}$ score of 86.9%, whereas with a higher resolution of 512×512, a $\mathcal{G}_\mathcal{M}$ score of 87.6% is obtained. While demonstrating such satisfactory performance, the inference speed is also comparable to other fast methods.

FBMS. Unlike the DAVIS 2016 validation set, the FBMS test set also contains multi-object scenarios as well as the single-object scenarios. DPA outperforms all other existing approaches by a significant margin with a $\mathcal{J}_\mathcal{M}$ score of

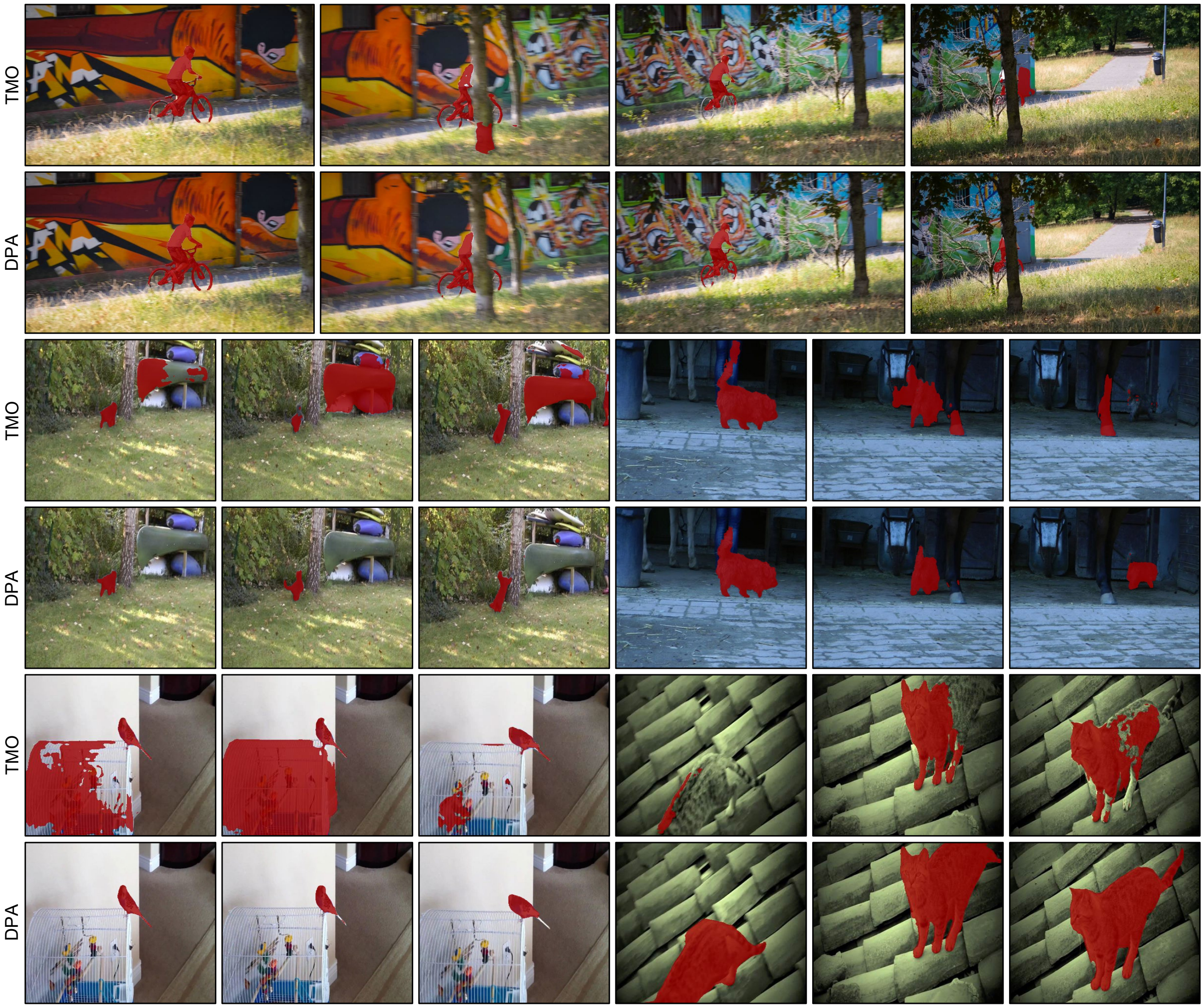

Figure 6. Qualitative comparison between state-of-the-art TMO and the proposed DPA.

1.5%, which demonstrates robustness of DPA even against videos containing multiple objects.

YouTube-Objects. Compared to the DAVIS 2016 validation set and FBMS test set, the YouTube-Objects dataset presents a significantly more challenging environment, as salient objects are often less distinctive. DPA surpasses all other methods on the YouTube-Objects dataset, showcasing its effectiveness even in challenging scenarios.

4.5. Qualitative Results

In Figure 6, we qualitatively compare our proposed DPA to the state-of-the-art TMO [3]. While TMO is often distracted by background elements, DPA consistently captures the primary objects. Even when the target object is occluded by obstacles, DPA maintains stable tracking of the object.

5. Conclusion

In unsupervised VOS, multi-modality fusion and temporal aggregation are recognized as essential components. However, they have limitations, such as a lack of thorough information exchange or substantial time consumption. To address these limitations and further enhance performance, we propose two novel attention modules, IMA and IFA. Additionally, each module is further enhanced by incorporating a prototype framework. By leveraging IMA and IFA in a collaborative manner, we achieve outstanding performance on all public benchmark datasets.

Acknowledgement. This work was supported in part by the Korea Institute of Science and Technology (KIST) Institutional Program under Project 2E33001.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625. Springer-Verlag, 2012.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[3] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Chaewon Park, Donghyeong Kim, and Sangyoun Lee. Treating motion as option to reduce motion dependency in unsupervised video object segmentation. *arXiv preprint arXiv:2209.03138*, 2022.

[4] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4922–4933, 2021.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[7] Minhyeok Lee, Suhwan Cho, Seunghoon Lee, Chaewon Park, and Sangyoun Lee. Unsupervised video object segmentation via prototype memory network. *arXiv preprint arXiv:2209.03712*, 2022.

[8] Youngjo Lee, Hongje Seong, and Euntai Kim. Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1245–1253, 2022.

[9] Daizong Liu, Dongdong Yu, Changhu Wang, and Pan Zhou. F2net: Learning to focus on the foreground for unsupervised video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2109–2117, 2021.

[10] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[11] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3623–3632, 2019.

[12] Sabarinath Mahadevan, Ali Athar, Aljoša Ošep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. Making a case for 3d convolutions for object segmentation in videos. *arXiv preprint arXiv:2008.11516*, 2020.

[13] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6): 1187–1200, 2013.

[14] Gensheng Pei, Fumin Shen, Yazhou Yao, Guo-Sen Xie, Zhenmin Tang, and Jinhui Tang. Hierarchical feature alignment network for unsupervised video object segmentation. In *European Conference on Computer Vision*, pages 596–613. Springer, 2022.

[15] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[16] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *2012 IEEE Conference on computer vision and pattern recognition*, pages 3282–3289. IEEE, 2012.

[17] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15455–15464, 2021.

[18] Christian Schmidt, Ali Athar, Sabarinath Mahadevan, and Bastian Leibe. D2conv3d: Dynamic dilated convolutions for object segmentation in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1200–1209, 2022.

[19] Mennatullah Siam, Chen Jiang, Steven Lu, Laura Petrich, Mahmoud Gamal, Mohamed Elhoseiny, and Martin Jagersand. Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 50–56. IEEE, 2019.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 715–731, 2018.

[22] Tiankang Su, Huihui Song, Dong Liu, Bo Liu, and Qingshan Liu. Unsupervised video object segmentation with online adversarial self-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 688–698, 2023.

[23] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[24] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.

[25] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9236–9245, 2019.

[26] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through

visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3074, 2019.

[27] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[28] Shu Yang, Lu Zhang, Jinqing Qi, Huchuan Lu, Shuo Wang, and Xiaoxing Zhang. Learning motion-appearance co-attention for zero-shot video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1564–1573, 2021.

[29] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 931–940, 2019.

[30] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pages 173–190. Springer, 2020.

[31] Kaihua Zhang, Zicheng Zhao, Dong Liu, Qingshan Liu, and Bo Liu. Deep transport network for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8781–8790, 2021.

[32] Lu Zhang, Jianming Zhang, Zhe Lin, Radomír Měch, Huchuan Lu, and You He. Unsupervised video object segmentation with joint hotspot tracking. In *European Conference on Computer Vision*, pages 490–506. Springer, 2020.

[33] Mingmin Zhen, Shiwei Li, Lei Zhou, Jiaxiang Shang, Haoan Feng, Tian Fang, and Long Quan. Learning discriminative feature with crf for unsupervised video object segmentation. In *European Conference on Computer Vision*, pages 445–462. Springer, 2020.

[34] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13066–13073, 2020.