Learning Motion-Appearance Co-Attention for Zero-Shot Video Object Segmentation

Shu Yang[1], Lu Zhang[1], Jinqing Qi[1], Huchuan Lu[1], Shuo Wang[2], Xiaoxing Zhang[2]
[1]Dalian University of Technology, China
[2]Meituan, China
{yangshua, luzhang_dut}@mail.dlut.edu.cn, {jinqing, lhchuan}@dlut.edu.cn,
{wangshuo28, zhangxiaoxing}@meituan.com

Abstract

How to make the appearance and motion information interact effectively to accommodate complex scenarios is a fundamental issue in flow-based zero-shot video object segmentation. In this paper, we propose an Attentive Multi-Modality Collaboration Network (AMC-Net) to utilize appearance and motion information uniformly. Specifically, AMC-Net fuses robust information from multi-modality features and promotes their collaboration in two stages. First, we propose a Multi-Modality Co-Attention Gate (MCG) on the bilateral encoder branches, in which a gate function is used to formulate co-attention scores for balancing the contributions of multi-modality features and suppressing the redundant and misleading information. Then, we propose a Motion Correction Module (MCM) with a visual-motion attention mechanism, which is constructed to emphasize the features of foreground objects by incorporating the spatio-temporal correspondence between appearance and motion cues. Extensive experiments on three public challenging benchmark datasets verify that our proposed network performs favorably against existing state-of-the-art methods via training with fewer data. The code is released at `https://github.com/isyangshu/AMC-Net`.

1. Introduction

Zero-shot Video Object Segmentation (ZVOS) aims to automatically separate the primary object(s) from the background in a video sequence without any human interaction. Since ZVOS does not require manual intervention, it has significant value in a wide range of applications, such as video compression [11], visual tracking [45], and person re-identification [51]. How to distinguish the target object(s) from complex and diverse background without any prior knowledge is an open challenge in ZVOS.

To address this issue, several methods [38, 21, 1, 19, 53] design various multi-modality interaction schemes to leverage external object motion information, which hypothesizes that the object moving across the video sequence is highly related to the primary object. Despite the impressive performance, there remain some issues in existing interaction approaches. Early methods [38, 21, 1] directly execute feature alignment such as concatenation or addition to produce object masks (see Figure 1 (a)). Due to the redundant and invalid information in flow maps and video frames, the direct feature alignment of multi-modality features would limit the accuracy of segmentation (see the third column in Figure 2). Several methods [53, 19] propose to build a motion-based attention mechanism to enhance the feature learning of object appearance (see Figure 1 (b)). These methods learn to

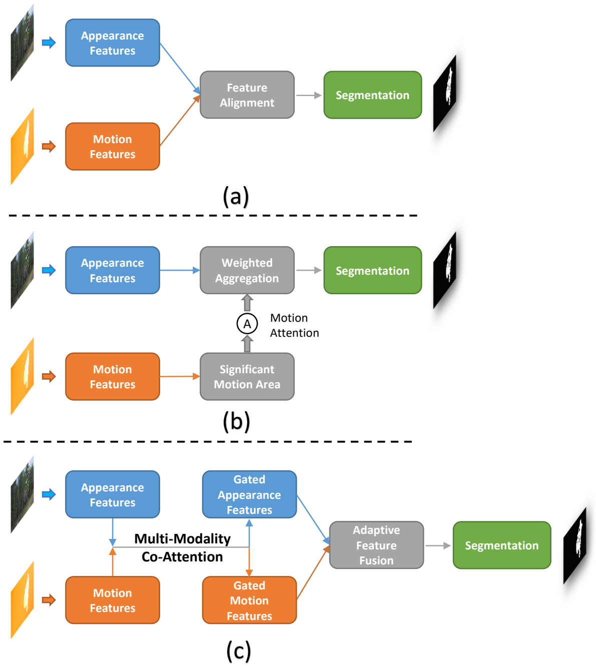

Figure 1. Illustration of various multi-modality interaction approaches for fusing appearance and motion information.

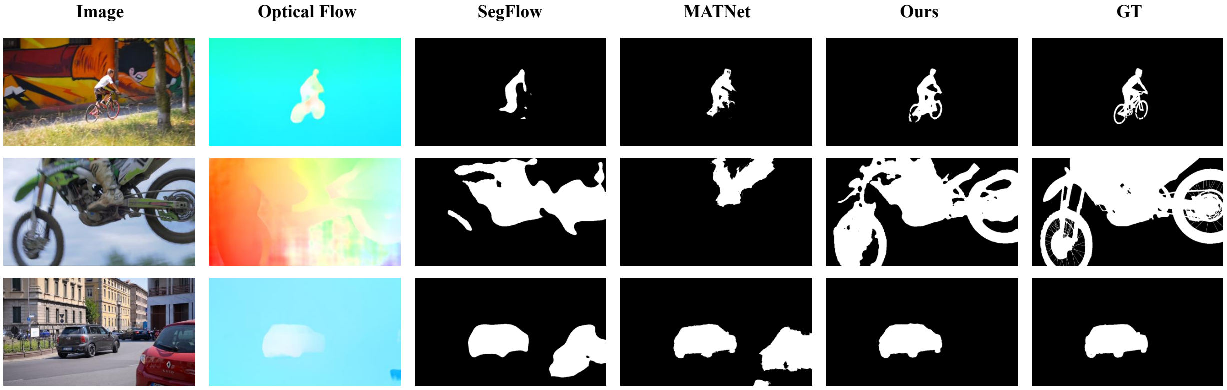

Figure 2. Visual comparison of different flow-based methods. We show the video frames, optical flow maps, SegFlow [1] predictions, MATNet [53] predictions, and our predictions in the ***bmx-trees***, ***motocross-jump***, and ***car-roundabout*** video sequences. The optical flow maps are predicted by PWCNet [35].

enhance the appearance features of the significant motion areas, which makes them rely on the quality of optical flow. However, when complex motion conditions (e.g., deformation, motion blur, fast motion, and clutters) occur in the video sequence, optical flow might fail to capture the object location and influence the accuracy of object segmentation (see the fourth column in Figure 2).

Motivated by the above observations, in this paper, we propose an Attentive Multi-Modality Collaboration Network (AMC-Net) for zero-shot video object segmentation, which builds a novel co-attention mechanism for effective multi-modality interaction. AMC-Net adaptively fuses robust spatio-temporal representations from multi-modality features and promotes their collaboration in two stages to thoroughly combine the merits of appearance and motion features (see Figure 1 (c)). In the first stage, we propose a Multi-Modality Co-Attention Gate (MCG), which is used to unify the appearance and motion information into valid spatio-temporal feature representations. Considering the disparity of the contributions of different modality features, we utilize a gate function to predict the co-attention scores, which are used to balance the contributions of multi-modality features and suppress the redundant and misleading information. In the second stage, we propose a Motion Correction Module (MCM) to perform adaptive feature fusion, in which a visual-motion attention mechanism is constructed to emphasize the features of foreground objects by incorporating the spatio-temporal correspondence between appearance and motion cues. Specifically, different from the single-directional attention guidance from motion to appearance, we model the attention based on visual saliency and motion saliency to facilitate the feature learning of foreground objects.

To investigate the effectiveness of our proposed model, we conduct comprehensive experiments including overall comparison and ablation studies on three benchmark datasets [31, 33, 29]. The results show that our proposed method can achieve superior performance against the state-of-the-arts by using only DAVIS-16 [31] for training.

Our contributions can be summarized as follows:

- We propose an Attentive Multi-Modality Collaboration Network for zero-shot video object segmentation, which promotes deep collaboration of appearance and motion information to generate accurate object segmentation.
- We propose a Multi-Modality Co-Attention Gate to unify the multi-modality information. A gate function is used to produce co-attention scores to adaptively balance the contributions of appearance and motion information.
- Our proposed method performs favorably against the state-of-the-art methods on three public challenging benchmark datasets (DAVIS-16 [31], Youtube-Objects [33], and FBMS [29]).

2. Related work

According to whether the manual intervention is required during testing, video object segmentation (VOS) can be broadly categorized into *zero-shot* (ZVOS) and *one-shot* (OVOS). In this paper, we focus on an object-level ZVOS setting (i.e., do not discriminate different instances), which extracts primary object(s) without manual annotation.

Zero-shot video object segmentation aims to automatically generate masks of the primary objects in the video sequence without any human interaction, which is also called unsupervised video object segmentation (UVOS) [14, 18, 32, 49, 21]. Early non-learning methods based on handcrafted features leverage low-level cues, such as visual saliency information [42, 9, 5, 39], object proposals [18, 27, 49, 14, 6], or optical flow [16, 21, 13, 20, 30], which are used as reliable prior knowledge to guide object segmentation. Later, inspired by the success of deep learning on segmentation tasks, more research efforts focus on

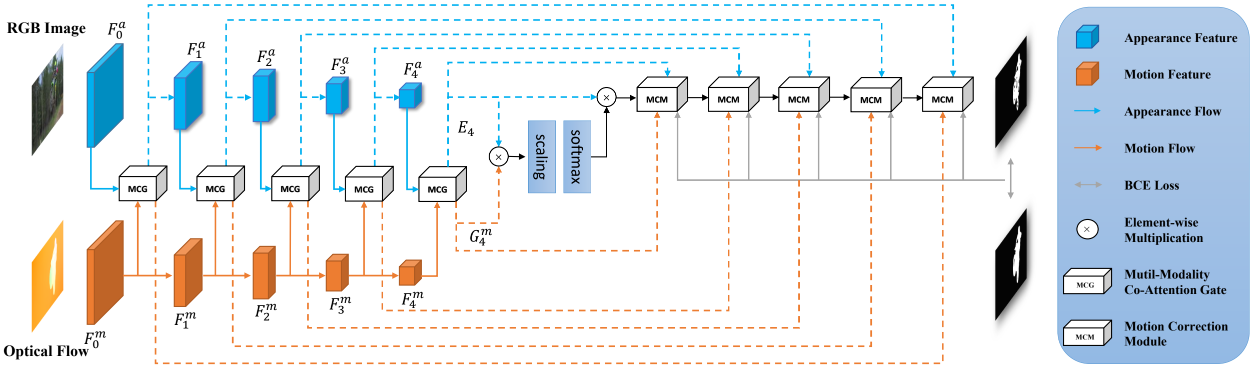

Figure 3. Overview of Attentive Multi-Modality Collaboration Network (AMC-Net). Given an input frame and the corresponding optical flow map, we first use bilateral encoder branches to extract multi-modality features. Then we adopt multi-level MCG on multiple encoder side outputs to filter information and unify robust information into spatio-temporal feature representations. Finally, we stack multiple MCMs in a coarse-to-fine manner to produce final predictions.

fully convolutional networks based ZVOS models. For example, several methods [34, 38, 43] use variants of recurrent neural networks to store previously computed segmentation information implicitly. Inspired by the non-local operation, COSNet [26], AGNN [41] and AnDiff [48] model the long-term correlations between frames to explore global information and gain a more comprehensive understanding of the video contents. WCS [50] encodes the pixel-wise correspondence between frames and takes object hotspots as guidance to enhance the influence of salient regions. 3DC-Seg [28] leverages 3D convolution to jointly learn spatial and temporal features. Recent method [25] proposes episodic graph memory to store cross-frame correlations and learn to update the segmentation model.

An alternative choice is to extract effective guidance about motion information from optical flow. LMP [37] discards the modeling of appearance features and purely relies on optical flow to predict foreground motion, which results in incorrect results for static foreground objects. To address this issue, several flow-based approaches [38, 1, 12, 19, 53] generally adopt dual-branch full convolution networks with feature fusion schemes, e.g., concatenation or attention mechanism, to aggregate appearance and motion information. Without considering the disparity of the contributions of multi-modality features, recent flow-based methods [19, 53] only consider the one-directional attention guidance from motion to appearance and enhance the appearance features of significant motion areas. When motion or appearance information is not significant from the background, these methods can not achieve satisfactory results which are shown in Figure 2. In this work, we consider the importance of deep collaboration between appearance and motion in learning richer spatio-temporal feature representations. We propose an Attentive Multi-Modality Collaboration Network (AMC-Net) to fuse robust information from multi-modality features and promote their collaboration for accurate zero-shot video object segmentation.

3. Method

3.1. Architecture Overview

In this work, we propose an Attentive Multi-Modality Collaboration Network (AMC-Net) for accurate zero-shot video object segmentation, which constructs a two-stage multi-modality integration system. The framework of our AMC-Net is shown in Figure 3. Specifically, given an input frame $I \in \mathbb{R}^{H \times W \times 3}$ and the corresponding flow map $O \in \mathbb{R}^{H \times W \times 3}$, we employ parallel encoder branches to capture the multi-level appearance and motion features, which are represented as $\{F_i^a\}_{i=0}^4$ and $\{F_i^m\}_{i=0}^4$. In the first stage, with the appearance feature F_i^a and motion feature F_i^m as input, a Multi-Modality Co-Attention Gate (MCG) is proposed to suppress the interference from redundant and invalid information and obtain effective spatio-temporal feature representations. We implement MCG on multi-level encoder side outputs to integrate the multi-modality features and propagate more valuable feature representations to the decoder. In the second stage, we use Motion Correction Module (MCM) to further emphasize the features of foreground objects by constructing a visual-motion attention mechanism. We stack multiple MCMs in a coarse-to-fine manner to facilitate the feature learning of foreground objects and generate the final segmentation results.

3.2. Multi-Modality Co-Attention Gate

Recent flow-based methods [19, 53] explore the appearance features of significant motion areas and utilize single-directional attention guidance from motion to appearance, which makes them rely on the dominant motion in the scene and ignore inherent noises in optical flow maps or images. In this paper, we propose a Multi-Modality Co-Attention Gate (MCG) for attentive motion-appearance interaction. We identify effective information from motion and appearance, and integrate cross-modality features into unified spatio-temporal feature representations.

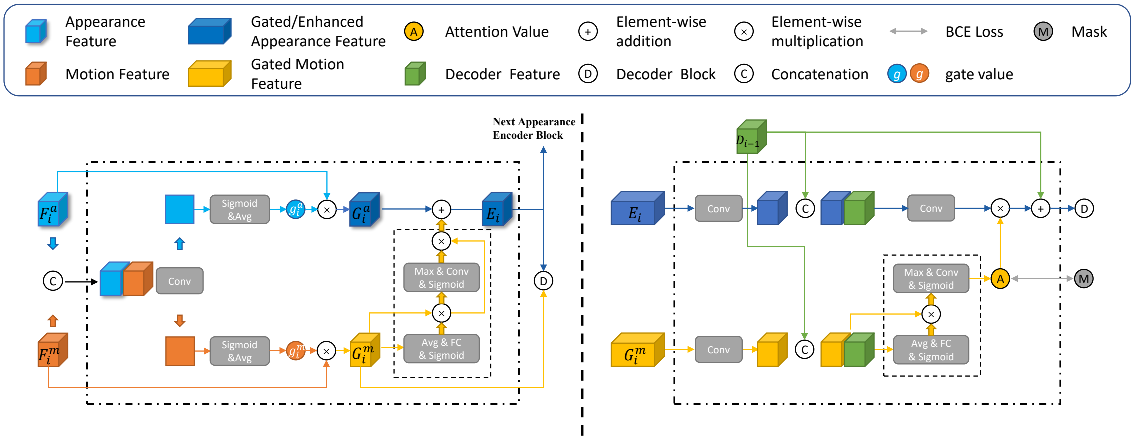

Figure 4. Frameworks of Multi-Modality Co-Attention Gate (*left*) and Motion Correction Module (*right*).

The framework of MCG is shown in Figure 4. Given an appearance feature F_i^a and a motion feature F_i^m at level i, we first combine them using cross-channel concatenation and convolution operation for implicit interaction, which aligns the motion feature with the appearance feature and helps to capture the relative relationship between multi-modality features. Then we obtain the fused feature $H \in \mathbb{R}^{h\times w\times 2}$ and spilt the channel-wise feature maps to two sub-branches. We perform the sigmoid function and global average pooling on each channel to obtain a pair of co-attention scores g_i^a and g_i^m, which reflect the importance of each modality feature on the final result. Specifically, we integrate the features from appearance and motion to formulate gate function, which plays the role of modality-wise attention and models the overall distribution of the contributions of multi-modality features in the network from global perspective. The entire gate function can be formulated as

$$g_i = Avg(\sigma(Conv(Cat(F_i^a, F_i^m)))), \tag{1}$$

where g_i contains a pair of co-attention scores g_i^a and g_i^m. $Avg(\cdot)$ is the global average pooling. σ denotes sigmoid function scaling the weight value into (0, 1). $Conv(\cdot)$ refers to the convolution layer with output channel 2 and $Cat(\cdot)$ is the concatenation operation among channel axis. In Figure 5, we give some visual examples of the co-attention scores in various videos. The tagged values show the disparity of the contributions of images and corresponding optical flow maps. A higher co-attention score indicates that the corresponding modality feature contains effective information for accurate segmentation. On the contrary, the modality feature with a lower score may contain noise that affects performance.

We apply the co-attention scores to the corresponding features to generate gated appearance feature G_i^a and gated motion feature G_i^m,

$$G_i^a = F_i^a * g_i^a, G_i^m = F_i^m * g_i^m, \tag{2}$$

which can be viewed as suppressing redundant and misleading information and enhancing more effective information. By assigning reliability scores for appearance and motion features, the network would not rely too much on either of the two information. It learns to adaptively exploit the merits of appearance and motion information to obtain satisfactory results. Considering the uncertainty of the predicted flow maps, we use channel-wise and spatial-wise attentions on motion features to emphasize the motion areas. Instead of treating all channels equally, we first build the inter-channel relationship of G_i^m,

$$M' = \sigma(MLP_{fc}(Avg(G_i^m))) * G_i^m, \tag{3}$$

where MLP_{fc} denotes fully connected layers. By doing this, we strengthen the responses of multi-level attributes, including texture, boundaries, color and semantics. Then we utilize the inter-spatial relationship of M' to emphasize the spatial locations of salient motion areas,

$$M'' = \sigma(MLP_{conv}(Max(M')) * M', \tag{4}$$

where MLP_{conv} denotes convolution layers and $Max(\cdot)$ is the global max pooling. We emphasize the features closely associated with motion-salient objects, which can take advantage of more effective salient motion information and temporal features. With the implicit interaction of appearance and motion information, we can aggregate features complementarily to obtain unified spatio-temporal feature representations. We devise the aggregation operation as an element-wise addition of the two modalities,

$$E_i = G_i^a + M''. \tag{5}$$

E_i is fed to the next stage of the appearance branch and exploited in the decoder for mask generation. In particular, G_i^m is applied to the decoder instead of the next stage

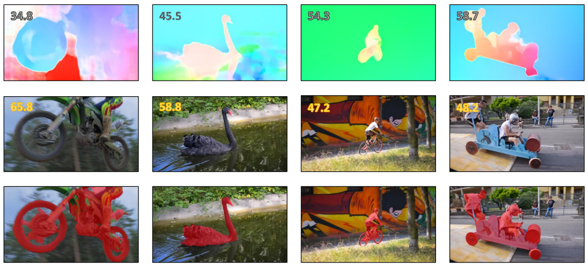

Figure 5. Visual samples of the co-attention scores. We show various video frames and the corresponding flow maps, along with the tagged scores to characterize the disparity of the contributions in *motocross-jump*, *blackswan*, *bmx-trees* and *soapbox*.

of the motion branch, ensuring the effectiveness of motion features during the long-term spread.

In order to capture the global features of primary objects, we compute the similarity between temporal and spatial information on the last encoder block. Specifically, we calculate a similarity matrix P to establish dense correspondences between each pair of pixels from gated motion feature G_4^m and enhanced appearance feature E_4. Similarly to [44], we combine the similarity matrix P with E_4 to yield the attention-enhanced feature map $\hat{E}_4$,

$$\hat{E}_4 = softmax(\frac{1}{\sqrt{c}} G_4^m \odot E_4^T) \odot E_4, \tag{6}$$

where $\odot$ indicates matrix multiplication. Following [23, 40], we scale the dot product by $\frac{1}{\sqrt{c}}$, where c is the channel size of G_4^m and E_4.

3.3. Motion Correction Module

With the adaptively multi-modality interaction by MCG, we obtain effective spatial-temporal correspondence encoder features, which may be well-aligned with the primary foreground objects in frames. Considering the complex scenes in video sequences, we unify the visual and motion significant areas to correct the intermediate features. We employ Motion Correction Module (MCM) to leverage appearance and motion features to supplement effective details and suppress the activation of non-salient regions. Besides, we construct visual-motion attention to emphasize the features of foreground objects by combining visual and motion cues. The details of MCM are shown in Figure 4.

Taking the gated motion feature G_i^m, enhanced appearance feature E_i from MCG and the decoder feature D_{i-1} from the previous decoder block as input, MCM aims to promote deep collaboration further to generate accurate results. When the previous decoding step is not available, we use $\hat{E}_4$ as the decoder feature to provide global information. One branch of the parallel branch aims to fuse visual saliency and motion saliency to extract visual-motion attention. This operations could be formulated as

$$F = Cat(Up(D_{i-1}), G_i^m), \tag{7}$$

where $Up(\cdot)$ is the upsampling operation with stride 2. Then we exploit the inter-channel relationship of fused feature F to improve the channel response of the significant information (visual or motion),

$$F' = \sigma(MLP_{fc}(Avg(F))) * F. \tag{8}$$

We select and strengthen the responses with higher significance. Then we utilize the inter-spatial relationship to calculate the comprehensive responses (visual and motion) of each pixel to generate attention maps,

$$A = \sigma(MLP_{conv}(Max(F'))), \tag{9}$$

where A refers to the visual-motion attention, which can be viewed as salient areas to suppress the activation of background and facilitate the feature learning of foreground objects. We argue these attention maps should be consistent with the final mask and equally weigh the cross-entropy loss for both the intermediate attention maps and the final prediction.

The other branch is designed as a residual structure, which is used to enhance the details of D_{i-1} by encoder features E_i with higher resolution. Finally, we combine A with the fused feature to correct feature and add D_{i-1} to obtain D_i. The entire process can be formulated as

$$D_i = Conv(Cat(Up(D_{i-1}), E_i)) \cdot A + D_{i-1}, \tag{10}$$

where $\cdot$ refers to element-wise multiplication.

3.4. Training and Inference

Implementation details. Following the most state-of-the-art flow-based methods, we take the ResNet-101 [8] pre-trained on ImageNet [3] as backbone. Given the current frame I_t and next frame I_{t+1}, we adopt PWCNet [35] to formulate optical flow maps O_t. Unlike the training strategy of other methods, we only use DAVIS-16 for training without applying any image datasets [47, 2, 22].

Training. Given images and optical flow maps as input, we train the model for 100 epochs with a mini-batch size as 4. We resize the images and flow maps to 384 × 384 for the balance between speed and performance. We adopt stochastic gradient descent (SGD) to train our AMC-Net, where the momentum, weight decay, and initial learning rate are set as 0.9, 0.0005 and 0.001. We use the “poly” policy [24] with the power of 0.9 to adjust the learning rate during training. We conduct data augmentation with random horizontally flipping and random rotation to avoid over-fitting and make the learned model more robust.

Inference. We use the same resolution 384 × 384 for each testing video sequence without any data augmentation and human interaction. Following the common protocol in ZVOS, we employ the fully-connected CRF [17] to obtain the final binary segmentation results.

Method	$\mathcal{J}$ mean ↑	$\mathcal{F}$ mean ↑	$\mathcal{J}\&\mathcal{F}$ ↑
Baseline	77.4	77.4	77.4
+ Multi-Modality Co-Attention Gate			
w/o gate	78.5(+1.1)	78.9(+1.5)	78.7(+1.3)
w/ gate	80.7(+3.3)	81.3(+3.9)	81.0(+3.6)
+ Motion Correction Module			
single MCM	81.1(+0.4)	81.7(+0.4)	81.4(+0.4)
3 stacked MCMs	81.9(+1.2)	83.9(+2.6)	82.9(+1.9)
5 stacked MCMs	83.0(+2.3)	84.3(+3.0)	83.7(+2.7)
+ Fully-Connected CRF			
AMC_{pwc} w/ crf	84.2(+1.2)	84.5(+0.2)	84.4(+0.7)
+ Better Quality Optical Flow			
AMC_{raft} w/ crf	84.5(+0.3)	84.6(+0.1)	84.6(+0.2)

Table 1. Ablation analysis of our proposed AMC-Net on DAVIS-16, measured by the $\mathcal{J}$ mean, $\mathcal{F}$ mean and $\mathcal{J}\&\mathcal{F}$. Red indicates the performance improvement compared to the previous setting.

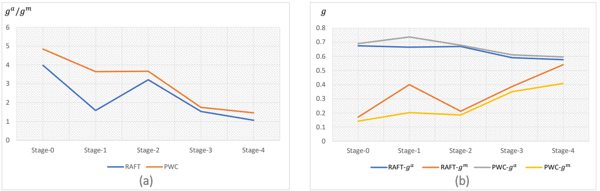

Figure 6. Distribution of the co-attention scores at each stage. With the convolution stage as the horizontal axis, we statistically demonstrate the curves about g^a/g^m in (a), evaluating the relative importance between multi-modality features. Meanwhile, we demonstrate the curves of co-attention scores g_i^a and g_i^m in (b).

Runtime. For each test image of size 384 × 384 × 3, the forward inference of our AMC-Net takes about 0.057s with a single Nvidia 1080Ti GPU.

4. Experiments

4.1. Dataset and Metrics

To evaluate the performance of our proposed method AMC-Net, we conduct comparison experiments on three public challenging benchmark datasets, including DAVIS-16 [31], Youtube-Objects [33] and FBMS [29].

DAVIS-16 consists of total 50 high-quality video sequences with 3455 densely annotated frames. We use 30 video sequences for training and the remaining 20 for testing. Each frame contains pixel-precise annotation of only one foreground object.

Youtube-Objects contains 126 video sequences of 10 object categories. The ground-truth in Youtube-Objects is sparsely labeled in every ten frames.

FBMS is comprised of 59 video sequences, with only 720 frames sparsely annotated. Certain video sequences are annotated with multiple target foreground objects.

Stage	w/o	0	0-1	0-2	0-3	0-4
$\mathcal{J}\&\mathcal{F}$	81.1	82.7	82.8	83.2	83.5	83.7

Table 2. Effectiveness of the Multi-Modality Co-Attention Gate (MCG) at multiple encoder side on DAVIS-16.

Supervision	0	1	2	3	4	5
$\mathcal{J}\&\mathcal{F}$	81.5	82.6	82.7	82.7	83.5	83.7

Table 3. Performance comparisons with different numbers of supervision for visual-motion attention on DAVIS-16.

Model	MATNet [53]	AMC-Net	AMC-Net	AMC-Net
Flow Model	PWCNet [35]	FlowNetS [4]	FlowNet2CS [10]	PWCNet [35]
$\mathcal{J}\&\mathcal{F}$	81.6	80.5	83.1	84.4

Table 4. Performance comparisons with different quality of optical flow maps on DAVIS-16.

4.2. Ablation Study

In this section, we conduct ablation studies of AMC-Net on the DAVIS-16. We add each component into the network in turn to verify the effectiveness. In Table 1, we report detailed results in terms of $\mathcal{J}$ mean, $\mathcal{F}$ mean and their average $\mathcal{J}\&\mathcal{F}$. To analyze the contribution of each component, we implement a simple baseline by employing bilateral encoder branches with performing a concatenation to integrate the multi-modality cues (as shown in Figure 1 (a)).

Effectiveness of Multi-Modality Co-Attention Gate. By adding MCG into the baseline, the model ("w/ gate") significantly outperforms the baseline by 3.6% on $\mathcal{J}\&\mathcal{F}$. To investigate the effect of our proposed gate function, we implement a variant ("w/o gate") without the gate function. This variant encounters a huge performance degradation (2.2% in $\mathcal{J}$ mean and 2.4% in $\mathcal{F}$ mean), which demonstrates the effectiveness of the gate function. Furthermore, we add MCG into the bilateral encoder branches in turn from the first stage and use simple feature aggregation at other stages. As shown in Table 2, MCG propagates more valuable information while minimizing interference.

Effectiveness of Motion Correction Module. To explore the impact of the number of stacked MCMs on performance, we first deploy a single MCM on the last decoder block, which has an improvement of 0.4% on $\mathcal{J}\&\mathcal{F}$. The variant ("3 stacked MCMs") with MCMs staked at last three decoder blocks achieves gains of 1.2% and 2.6% on $\mathcal{J}$ mean and $\mathcal{F}$ mean, respectively. When we use MCMs on all decoder blocks, the variant ("5 stacked MCMs") achieves the highest performance with a $\mathcal{J}$ mean score of 83.0. In addition, we explore the impact of the visual-motion attention mechanism on performance with five stacked MCMs. In Table 3, we report the results of adding supervision for visual-motion attention in turn from the first decoder block.

Effect of the quality of optical flow maps. To verify the effect of the optical flow quality on our full model, we first implement a variant (AMC_{raft}) by using the op-

Model	Operations				Metrics $\mathcal{J}$			Metrics $\mathcal{F}$			$\mathcal{J}$ & $\mathcal{F}$
	S	D	Y	PP	Mean ↑	Recall ↑	Decay ↓	Mean ↑	Recall ↑	Decay ↓	
SFL [1]		✓		✓	67.4	81.4	6.2	66.7	77.1	5.1	67.1
LMP [37]	✓	✓		✓	70.0	85.0	1.3	65.9	79.2	2.5	68.0
PDB [34]	✓	✓		✓	77.2	90.1	0.9	74.5	84.4	-0.2	75.9
ARP [14]		✓			76.2	89.1	7.0	70.6	83.5	7.9	73.4
AGS [43]	✓	✓		✓	79.7	91.1	1.9	77.4	85.8	1.6	78.6
COSNet [26]	✓	✓		✓	80.5	93.1	4.4	79.5	89.5	5.0	80.0
AGNN [41]	✓	✓		✓	80.7	94.0	**0.0**	79.1	90.5	**0.0**	79.9
ANDiff [48]	✓	✓		✓	81.7	90.9	2.2	80.5	85.1	0.6	81.1
MATNet [53]		✓	✓*	✓	82.4	94.5	5.5	80.7	90.2	4.5	81.6
EGMN [25]	✓	✓		✓	82.5	94.3	4.2	81.2	90.3	5.6	81.9
WCS [50]	✓	✓			82.2	-	-	80.7	-	-	81.5
DFNet [52]	✓	✓		✓	83.4	-	-	81.8	-	-	82.6
3DC-Seg [28]		✓			-	-	-	-	-	-	82.2
3DC-Seg* [28]	✓	✓	✓		84.3	95.7	7.4	**84.7**	92.6	5.2	84.5
Ours		✓		✓	84.2	96.0	3.5	84.5	**94.4**	2.2	84.4
Ours-$raft$		✓		✓	**84.5**	**96.4**	2.8	84.6	93.8	2.5	**84.6**

Table 5. Overall comparison with the state-of-the-arts on DAVIS-16 validation dataset. Use "✓" in the table to indicate whether the method uses the static segmentation datasets (S), DAVIS-16 (D), Youtube-VOS (Y) or Post-processing (PP). The "✓*" in the Y-column indicates that MATNet uses a subset of 12K frames selected from the training set of Youtube-VOS.

Model	Aeroplane	Bird	Boat	Car	Cat	Cow	Dog	Horse	Motorbike	Train	Avg ↑
ARP [14]	73.6	56.1	57.8	33.9	30.5	41.8	36.8	44.3	48.9	39.2	46.2
FST [30]	70.9	70.6	42.5	65.2	52.1	44.5	65.3	53.5	44.2	29.6	53.8
SFL [1]	65.6	65.4	59.9	64.0	58.9	51.1	54.1	64.8	52.6	34.0	57.0
PDB [34]	78.0	80.0	58.9	76.5	63.0	64.1	70.1	67.6	58.3	35.2	65.4
FSEG [12]	81.7	63.8	72.3	74.9	68.4	68.0	69.4	60.4	62.7	62.2	68.4
MATNet [53]	72.9	77.5	66.9	79.0	73.7	67.4	75.9	63.2	62.6	51.0	69.0
AGS [43]	87.7	76.7	72.2	78.6	69.2	64.6	73.3	64.4	62.1	48.2	69.7
COSNet [26]	81.1	75.7	71.3	77.6	66.5	69.8	76.8	67.4	67.7	46.8	70.5
AGNN [41]	81.1	75.9	70.7	78.1	67.9	69.7	77.4	67.3	68.3	47.8	70.8
WCS [50]	81.8	81.2	67.6	79.5	65.8	66.2	73.4	69.5	69.3	49.7	70.9
EGMN [25]	86.1	75.7	68.6	82.4	65.9	70.5	77.1	72.2	63.8	47.8	**71.4**
Ours	78.9	80.9	67.4	82.0	69.0	69.6	75.8	63.0	63.4	57.8	71.1

Table 6. Overall comparison with the state-of-the-arts on Youtube-Objects dataset. We report the per-category performance and the average result of the 10 categories with $\mathcal{J}$ mean.

tical flow maps generated by a more accurate network RAFT [36]. Compared to AMC_{pwc}, AMC_{raft} has an improvement of 0.3%, 0.1% in terms of the $\mathcal{J}$ mean and $\mathcal{F}$ mean, respectively. Furthermore, we implement several variants to use the optical flow maps in slightly inferior quality calculated by FlowNetS [4] and FlowNet2CS [10]. As shown in Table 4, all of the network variants produce competitive results, which demonstrates that our proposed model can benefit from optical flow of different quality and adaptively fuse multi-modality features.

Effect of optical flow quality on co-attention scores. In order to qualitatively measure the effect of optical flow quality on co-attention scores, we calculate the average scores of appearance and motion features from AMC_{pwc} and AMC_{raft}. As shown in Figure 6 (a), we compute g^a/g^m to evaluate the relative importance between appearance and motion features. For both AMC_{pwc} and AMC_{raft}, the high-level ratios g^a/g^m are significantly smaller than the low-level ones. The relative importance of optical flow maps gradually increases with the promotion of levels. More importantly, compared to AMC_{pwc}, AMC_{raft} gets lower ratios at low-level. In Figure 6 (b), we statistically demonstrate the curves of co-attention scores g_i^a and g_i^m. It can be seen that the high-level motion features contribute more effective guidance than low-level ones. This trend is just the opposite of the appearance features. Since AMC_{raft} uses flow maps with more precise image boundary details, its gate values corresponding to low-level motion features significantly increase. It can be seen that AMC-Net can adaptively and fully combine the merits of appearance and motion features.

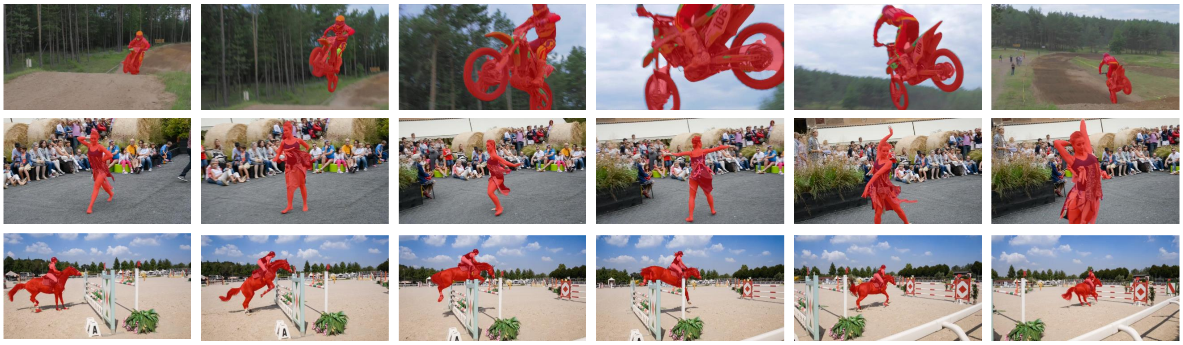

Figure 7. Qualitative results on three video sequences from DAVIS-16. From top to bottom: *motocross-jump*, *dance-twirl*, *horsejump-high*.

Method	NLC [5]	FST [30]	SFL [1]	ARP [14]
$\mathcal{J}$ mean ↑	44.5	55.5	56.0	59.8
Method	MSTP [9]	FSEG [12]	IET [20]	OBN [21]
$\mathcal{J}$ mean ↑	60.8	68.4	71.9	73.9
Method	PDB [34]	COSNet [26]	MATNet [53]	Ours
$\mathcal{J}$ mean ↑	74.0	75.6	76.1	**76.5**

Table 7. Overall comparison with the state-of-the-arts on FBMS. We use the mean region similarity ($\mathcal{J}$ mean) to measure the segmentation performance.

4.3. Quantitative and Qualitative Results

Evaluation on DAVIS-16. We compare the performance of our proposed method with the state-of-the-art methods. In Table 5, we list comparison results with ZVOS methods. We list the datasets widely used in existing methods, including DAVIS-16 [31], Youtube-VOS [46], and static segmentation datasets (DUT [47], MSRA10k [2], COCO [22], etc). Besides, we provide the indicator of post-processing (PP) in the existing methods. Compared with the existing ZVOS methods, our model outperforms the best performance 3DC-Seg [28] by 0.1% on $\mathcal{J}$&$\mathcal{F}$. In addition, compared to our proposed method, 3DC-Seg uses pre-trained weights from much larger video datasets IG-65M [7] and Kinetics [15] (65.8M video clips), and further fine-tunes jointly on three datasets [22, 46, 31] to yield the best score.

Evaluation on Youtube-objects. Table 6 illustrates the results of all compared methods for different categories. Our proposed method is comparable with recent ZVOS methods AGS [43], AGNN [41] and WCS [50] across all categories. Since memory mechanism can better handle static objects in video sequences, EGMN [25] achieves a higher $\mathcal{J}$&$\mathcal{F}$ score (0.3%) on YTB-Objects.

Evaluation on FBMS. Multiple target objects labeled in the video sequence share a similar appearance but have different motion patterns (moving or non-moving), which weakens the role of optical flow in the corresponding scene. In order to deal with the above problem, we design a variant by adding a separate FCN-like decoder that uses the gated appearance features from MCG. We train the entire network variant on DAVIS-16 and combine two decoder branches to produce the final results. Table 7 shows that our proposed method performs better than the state-of-the-art methods.

Qualitative results. As shown in Figure 7, we illustrate the segmentation results of our proposed method on three video sequences. These three video sequences contain some tough challenges: *Deformation*, *Scale-Variation*, etc. The qualitative results show that our method can cope well with the tough challenges posed by the tricky motion conditions in the video sequences and generate precise segmentation masks with well-defined details.

5. Conclusion

In this paper, we propose an Attentive Multi-Modality Collaboration Network for ZVOS, which adopts a novel mechanism to achieve deep collaboration between appearance and motion. AMC-Net adaptively fuses robust information from multi-modality features and promotes their collaboration in two stages. We first adopt multi-level MCG to balance the contributions of multi-modality features at each stage and suppress redundant and misleading information, propagating valid spatio-temporal feature representations while minimizing interference. Then we adopt five stacked MCMs with a visual-motion attention mechanism to emphasize the features of foreground objects utilizing the spatio-temporal correspondence between appearance and motion cues. The experimental results on three benchmarks demonstrate that AMC-Net learns from fewer data and outperforms existing competitors. We yield a neat yet effective framework with a novel strategy for the interaction of motion and appearance information, which will generalize well to ZVOS in complex scenes.

Acknowledgments

The paper is supported by the National Natural Science Foundation of China under Grant No.U1903215, No.61829102, No.61725202, and the Dalian Innovation leader's support Plan under Grant No.2018RD07. This research is also supported by Meituan.

References

[1] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, pages 686–695, 2017.

[2] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015.

[5] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, page 8, 2014.

[6] Huazhu Fu, Dong Xu, Bao Zhang, and Stephen Lin. Object-based multiple foreground video co-segmentation. In *CVPR*, pages 3166–3173, 2014.

[7] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, pages 12046–12055, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *ECCV*, pages 786–802, 2018.

[10] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017.

[11] Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE transactions on image processing*, 13(10):1304–1318, 2004.

[12] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, pages 2117–2126, 2017.

[13] Won-Dong Jang, Chulwoo Lee, and Chang-Su Kim. Primary object segmentation in videos via alternate convex optimization of foreground and background distributions. In *CVPR*, pages 696–704, 2016.

[14] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *CVPR*, pages 3442–3450, 2017.

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[16] Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *ICCV*, pages 3271–3279, 2015.

[17] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[18] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *ICCV*, pages 1995–2002, 2011.

[19] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, pages 7274–7283, 2019.

[20] Siyang Li, Bryan Seybold, Alexey Vorobyov, Alireza Fathi, Qin Huang, and C-C Jay Kuo. Instance embedding transfer to unsupervised video object segmentation. In *CVPR*, pages 6526–6535, 2018.

[21] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*, pages 207–223, 2018.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[23] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, pages 1449–1457, 2015.

[24] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[25] Xinkai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *ECCV*, 2020.

[26] 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 *CVPR*, pages 3623–3632, 2019.

[27] Tianyang Ma and Longin Jan Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, pages 670–677, 2012.

[28] 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. In *BMVC*, 2020.

[29] 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.

[30] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, pages 1777–1784, 2013.

[31] 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 *ICCV*, pages 724–732, 2016.

[32] Federico Perazzi, Oliver Wang, Markus Gross, and Alexander Sorkine-Hornung. Fully connected object proposals for video segmentation. In *ICCV*, pages 3227–3234, 2015.

[33] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, pages 3282–3289, 2012.

[34] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, pages 715–731, 2018.

[35] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018.

[36] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020.

[37] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, pages 3386–3394, 2017.

[38] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *ICCV*, pages 4481–4490, 2017.

[39] Yi-Hsuan Tsai, Guangyu Zhong, and Ming-Hsuan Yang. Semantic co-segmentation in videos. In *ECCV*, pages 760–775, 2016.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[41] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, pages 9236–9245, 2019.

[42] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, pages 3395–3402, 2015.

[43] 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 *CVPR*, pages 3064–3074, 2019.

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[45] Hefeng Wu, Guanbin Li, and Xiaonan Luo. Weighted attentional blocks for probabilistic object tracking. *The Visual Computer*, 30(2):229–243, 2014.

[46] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018.

[47] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013.

[48] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, pages 931–940, 2019.

[49] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, pages 628–635, 2013.

[50] Lu Zhang, Jianming Zhang, Zhe Lin, Radomír Měch, Huchuan Lu, and You He. Unsupervised video object segmentation with joint hotspot tracking. In *ECCV*, 2020.

[51] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013.

[52] 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. *ECCV*, 2020.

[53] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Transactions on Image Processing*, 29:8326–8338, 2020.