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

Shu Yang\textsuperscript{1}, Lu Zhang\textsuperscript{1}, Jinqing Qi\textsuperscript{1}, Huchuan Lu\textsuperscript{1}, Shuo Wang\textsuperscript{2}, Xiaoxing Zhang\textsuperscript{2}
\textsuperscript{1}Dalian University of Technology, China
\textsuperscript{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 \url{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
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 hand-crafted 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
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 $\{F_a^i\}_{i=0}^4$ and $\{F_m^i\}_{i=0}^4$. In the first stage, with the appearance feature $F^a_i$ and motion feature $F^m_i$ 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 produce final predictions.

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.
The framework of MCG is shown in Figure 4. Given an appearance feature \( F_a^i \) and a motion feature \( F_m^i \) 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 \( g_a^i \) and \( g_m^i \), 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

\[
g_a^i = \text{Avg}(\sigma(\text{Conv}(\text{Cat}(F_a^i, F_m^i))))\tag{1}
\]

where \( g_a^i \) contains a pair of co-attention scores \( g_a^i \) and \( g_m^i \). \( \text{Avg}(\cdot) \) is the global average pooling. \( \sigma \) denotes sigmoid function scaling the weight value into \((0, 1)\). \( \text{Conv}(\cdot) \) refers to the convolution layer with output channel 2 and \( \text{Cat}(\cdot) \) 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 \( G_a^i \) and gated motion feature \( G_m^i \).

\[
G_a^i = F_a^i \ast g_a^i, \quad G_m^i = F_m^i \ast g_m^i\tag{2}
\]

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 \( G_m^i \),

\[
M' = \sigma(\text{MLP}_{fc}(\text{Avg}(G_m^i))) \ast G_m^i,\tag{3}
\]

where \( \text{MLP}_{fc} \) 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,

\[
M'' = \sigma(\text{MLP}_{conv}(\text{Max}(M'))) \ast M',\tag{4}
\]

where \( \text{MLP}_{conv} \) denotes convolution layers and \( \text{Max}(\cdot) \) is the global max pooling. We emphasize the features closely associated with motion-salient objects and 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,

\[
E_i = G_a^i + M''.\tag{5}
\]

\( E_i \) is fed to the next stage of the appearance branch and exploited in the decoder for mask generation. In particular, \( G_a^i \) is applied to the decoder instead of the next stage.
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 $G^m_i$ and enhanced appearance feature $E_i$. Similarly to [44], we combine the similarity matrix $P$ with $E_i$ to yield the attention-enhanced feature map $\hat{E}_4$,

$$\hat{E}_4 = \text{softmax}(\frac{1}{\sqrt{c}} G^m_i \odot E_i^T) \odot E_i,$$

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 $G^m_i$ and $E_i$.

### 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 $G^m_4$, 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

$$F = \text{Cat}(Up(D_{i-1}), G^m_i),$$

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),

$$F' = \sigma(\text{MLP}_f(\text{Avg}(F))) \ast F.$$  \hspace{1cm} (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,

$$A = \sigma(\text{MLP}_{\text{conv}}(_{\text{Max}}(F'))),$$  \hspace{1cm} (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

$$D_i = \text{Conv}(_{\text{Cat}}(Up(D_{i-1}), E_i)) \ast A + D_{i-1},$$

where $\ast$ 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 $384 \times 384$ 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 $384 \times 384$ 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.
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 $J$ mean, $F$ mean and their average $J \& F$. 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 $J \& F$. 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 $J$ mean and 2.4% in $F$ 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 $J \& F$. The variant (“3 stacked MCMs”) with MCMs staked at last three decoder blocks achieves gains of 1.2% and 2.6% on $J$ mean and $F$ mean, respectively. When we use MCMs on all decoder blocks, the variant (“5 stacked MCMs”) achieves the highest performance with a $J$ 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-


<table>
<thead>
<tr>
<th>Model</th>
<th>Operations</th>
<th>Metrics $J$</th>
<th>Metrics $F$</th>
<th>$J$ &amp; $F$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S D Y PP</td>
<td>Mean ↑ Recall ↑ Decay ↓</td>
<td>Mean ↑ Recall ↑ Decay ↓</td>
<td></td>
</tr>
<tr>
<td>SF [14]</td>
<td>✓ ✓ ✓ ✓</td>
<td>67.4 69.1 0.3 69.1</td>
<td>51.1 49.7 1.9 51.1</td>
<td></td>
</tr>
<tr>
<td>LMP [37]</td>
<td>✓ ✓ ✓ ✓</td>
<td>70.0 85.0 1.3 85.0</td>
<td>65.9 79.2 2.5 65.9</td>
<td></td>
</tr>
<tr>
<td>PDB [34]</td>
<td>✓ ✓ ✓ ✓</td>
<td>77.2 90.1 0.9 90.1</td>
<td>74.5 84.4 -0.2 74.5</td>
<td></td>
</tr>
<tr>
<td>ARP [14]</td>
<td>✓ ✓ ✓ ✓</td>
<td>76.2 89.1 7.0 89.1</td>
<td>70.6 83.5 7.9 70.6</td>
<td></td>
</tr>
<tr>
<td>AGS [43]</td>
<td>✓ ✓ ✓ ✓</td>
<td>79.7 91.1 1.9 91.1</td>
<td>77.4 85.8 1.6 77.4</td>
<td></td>
</tr>
<tr>
<td>COSNet [26]</td>
<td>✓ ✓ ✓ ✓</td>
<td>80.5 93.1 4.4 93.1</td>
<td>79.5 89.5 5.0 79.5</td>
<td></td>
</tr>
<tr>
<td>AGNN [41]</td>
<td>✓ ✓ ✓ ✓</td>
<td>80.7 94.0 0.0 94.0</td>
<td>79.1 90.5 0.0 79.1</td>
<td></td>
</tr>
<tr>
<td>ANDiff [48]</td>
<td>✓ ✓ ✓ ✓</td>
<td>81.7 90.9 2.2 90.9</td>
<td>80.5 85.1 0.6 80.5</td>
<td></td>
</tr>
<tr>
<td>MATNet [53]</td>
<td>✓ ✓ ✓ ✓</td>
<td>82.4 94.5 5.5 94.5</td>
<td>80.7 90.2 4.5 80.7</td>
<td></td>
</tr>
<tr>
<td>WCS [50]</td>
<td>✓ ✓ ✓ ✓</td>
<td>82.2 - - -</td>
<td>80.7 - - -</td>
<td></td>
</tr>
<tr>
<td>DFNet [52]</td>
<td>✓ ✓ ✓ ✓</td>
<td>83.4 - - -</td>
<td>81.8 - - -</td>
<td></td>
</tr>
<tr>
<td>3DC-Seg [28]</td>
<td>✓ ✓ ✓ ✓</td>
<td>84.3 95.7 7.4 95.7</td>
<td>84.7 92.6 5.2 84.7</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>✓ ✓ ✓ ✓</td>
<td>84.2 96.0 3.5 96.0</td>
<td>84.5 94.4 2.2 94.4</td>
<td></td>
</tr>
<tr>
<td>Ours-raft</td>
<td>✓ ✓ ✓ ✓</td>
<td>84.5 96.4 2.8 96.4</td>
<td>84.6 93.8 2.5 93.8</td>
<td></td>
</tr>
</tbody>
</table>

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.

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 $J$ mean and $F$ 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 $g_a^i$ and $g_m^i$. 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 DA VIS-16. From top to bottom: motocross-jump, dance-twirl, horsejump-high.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( J \text{ mean} \uparrow )</td>
<td>44.5</td>
<td>55.5</td>
<td>56.0</td>
<td>59.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( J \text{ mean} \uparrow )</td>
<td>60.8</td>
<td>68.4</td>
<td>71.9</td>
<td>73.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>PDB [34]</th>
<th>COSNet [26]</th>
<th>MATNet [53]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>( J \text{ mean} \uparrow )</td>
<td>74.0</td>
<td>75.6</td>
<td>76.1</td>
<td>76.5</td>
</tr>
</tbody>
</table>

Table 7. Overall comparison with the state-of-the-arts on FBMS.

We use the mean region similarity (\( J \)) to measure the segmentation performance.

4.3. Quantitative and Qualitative Results

Evaluation on DA VIS-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 DA VIS-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 \( J \& F \). 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] (65.8M video clips), and further fine-tunes jointly on three datasets [22, 46, 31] to yield the best score.

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 on DA VIS-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


