Reciprocal Transformations for Unsupervised Video Object Segmentation

Sucheng Ren[1], Wenxi Liu[2], Yongtuo Liu[1], Haoxin Chen[1], Guoqiang Han[1], Shengfeng He[1]*
[1] School of Computer Science and Engineering, South China University of Technology
[2] College of Mathematics and Computer Science, Fuzhou University

Abstract

Unsupervised video object segmentation (UVOS) aims at segmenting the primary objects in videos without any human intervention. Due to the lack of prior knowledge about the primary objects, identifying them from videos is the major challenge of UVOS. Previous methods often regard the moving objects as primary ones and rely on optical flow to capture the motion cues in videos, but the flow information alone is insufficient to distinguish the primary objects from the background objects that move together. This is because, when the noisy motion features are combined with the appearance features, the localization of the primary objects is misguided. To address this problem, we propose a novel reciprocal transformation network to discover primary objects by correlating three key factors: the intra-frame contrast, the motion cues, and temporal coherence of recurring objects. Each corresponds to a representative type of primary object, and our reciprocal mechanism enables an organic coordination of them to effectively remove ambiguous distractions from videos. Additionally, to exclude the information of the moving background objects from motion features, our transformation module enables to reciprocally transform the appearance features to enhance the motion features, so as to focus on the moving objects with salient appearance while removing the co-moving outliers. Experiments on the public benchmarks demonstrate that our model significantly outperforms the state-of-the-art methods. Code is available at https://github.com/OliverRensu/RTNet.

1. Introduction

Video object segmentation (VOS) aims at localizing and segmenting objects in videos. As one of the fundamental tasks in computer vision, VOS has many applications, *e.g.*, object tracking [22, 30, 51] autonomous driving [5, 13], video surveillance [45]. In specific, the existing techniques of VOS can be roughly categorized into: *semi-supervised video object segmentation* [14, 32, 56] in which the segmentation mask of the primary object(s) is given at the first frame, and *unsupervised video object segmentation* (UVOS) [31, 47, 48, 60] that aims to extract the mask of the primary object(s) without any prior knowledge. In this paper, we focus on the task of UVOS.

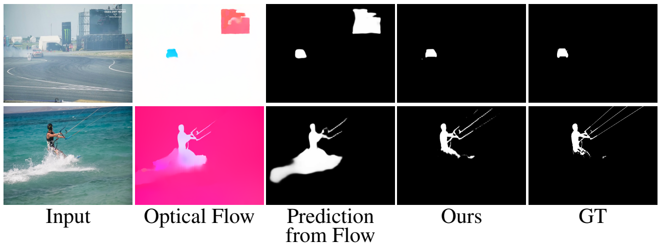

Figure 1: Segmenting primary objects based on optical flow is usually distracted by the co-moving outliers. We attack this problem by reciprocally transforming appearance features to motion features, and thus avoid misleading motion information from corrupting the localization of primary objects.

Due to the lack of prior knowledge, the UVOS models have to handle the major concern for identifying the sources of primary objects in the videos. We observe that there are three types of candidate primary objects: the salient objects in a single frame, the moving objects, and recurring objects in the video. In general, human attention will be drawn on the salient objects [18, 28, 52] within an image, thus these visually distinct objects may be the candidate primary objects. However, these methods may not be applicable for identifying primary objects in videos, as human attention will naturally shift to various patterns of dynamics or motions in video [9, 37]. Thus, the objects that are indistinct in a single frame but moving in video may be the primary objects according to motion cues, yet ignored by the image-based models. Besides, people also tend to memorize the objects that appear repeatedly in the video, therefore these objects may be treated as another type of primary ones.

Previous methods [24, 65] apply optical flow to capture motion information. However, the optical flow can hardly

*Corresponding author (hesfe@scut.edu.cn).

distinguish the dynamic background objects from the foreground objects. For instance, in Fig. 1, it is ambiguous to categorize the car and the billboard, or the human and the spray, into foreground and background by optical flow only. Therefore, directly mapping these motion cues to the appearance features [24, 65] may misguide UVOS models when localizing the primary objects.

To address the aforementioned limitation, we propose a unified framework, Reciprocal Transformation Network (RTNet), to identify primary objects beyond the distraction of co-moving outliers. Our idea is to mutually evolve and integrate the appearance and motion representations in the network such that all three types of candidate primary objects can be taken into consideration and produces a holistic decision. To this end, we propose a Reciprocal Transformation Module (RTM) within the network to enable in-domain and cross-domain feature interactions. In particular, the proposed reciprocal transformation scheme computes similarities for all the pairwise features including motion-motion, appearance-appearance, and appearance-motion pairs of features. The underlying information will be transformed to each other in order to replenish the appearance/motion object representation and remove the ambiguity from the inconsistent appearance or the inaccurate optical flow.

Applying the proposed RTM on different source features results in different types of primary object properties, *i.e.*, 1) self-similarities of appearance and motion features lead to intra-frame contrast; 2) appearance-motion similarity produces motion cues; and 3) cross-frame appearance-appearance and motion-motion feature similarities yield temporal coherence. Each corresponds to one of the three types of primary objects. Besides, instead of simply skip connecting the encoder and decoder as FCN [29] does, we propose a Spatial Temporal Attentive Fusion Module (STAFM) to leverage the appearance and motion features from the corresponding encoder stage, and segment spatio-temporally consistent primary objects. In experiments, we evaluate our approach against the state-of-the-art methods on public benchmarks DAVIS [34] and achieve the performance gain of 4% on region similarity $\mathcal{J}$ and 5% on boundary accuracy $\mathcal{F}$ over the second best method [65].

To sum up, the contributions of our paper are three-fold:

- We delve into three types of primary objects in videos, and present a novel reciprocal transformation network (RTNet), which is able to effectively exploit the intra-frame contrast, motion cues, and temporal coherence of recurring objects to identify and segment primary objects from the videos.

- To eliminate the co-ocurring moving outliers from the optical flow and extract the moving objects with salient appearance, we propose a new reciprocal transformation approach that mutually evolves the appearance features to the motion features.

- We propose a Spatial Temporal Attentive Fusion Module (STAFM) to selectively integrate the appearance and motion features.

- Our method significantly outperforms state-of-the-art methods in the public benchmark. Even if we use a much smaller backbone and less training data, our lightweight model can still achieve comparable performance against latest competitors.

2. Related Work

In this section, we will survey the works on video object segmentation as well as attention mechanism.

2.1. Video Object Segmentation

Prior methods on video object segmentation can be divided into two categories: *semi-supervised video object segmentation* and *unsupervised video object segmentation*.

Semi-supervised Video Object Segmentation. The semi-supervised VOS methods assume that the ground-truth mask of the target object(s) is provided at the first frame. These methods can be further categorized into two types, i.e., online-learning methods [1,4,7,26] and offline-learning methods [6,20]. Online-learning methods fine-tune the pretrained model based on the given ground-truth mask and predict the segmentation results at the cost of the inference time. On the contrary, offline-learning methods utilize the given mask as the guidance to update the pretrained model at the inference time. Despite the superior performance of semi-supervised VOS methods, the annotating ground-truth masks involves manual efforts and may introduce bias, which limits its application in real-world scenarios.

Unsupervised Video Object Segmentation. Compared with semi-supervised VOS methods, UVOS methods do not require any manual annotations. Early UVOS methods are mainly based on object proposal [21, 23], temporal trajectory [3, 11, 33] and saliency prior [19, 53, 54]. With the development of deep convolutional neural networks and the establishment of large-scale datasets [34], deep learning based methods are proposed for modeling the spatio-temporal information. To capture the motion cues, MP-Net [46] focuses on optical flow only, but it is difficult to segment the static objects, due to the insufficient appearance information. In addition, several methods apply two stream networks to capture and then fuse the appearance and motion features. Fragkiadaki *et al.* [10] use fully-connected layers to integrate optical flow and static boundaries to rank the segment proposals. MBN [25] includes a bilateral network for background estimation and integrates it with the appearance features into a graph. Besides, to capture long-term temporal information, several methods [43] process

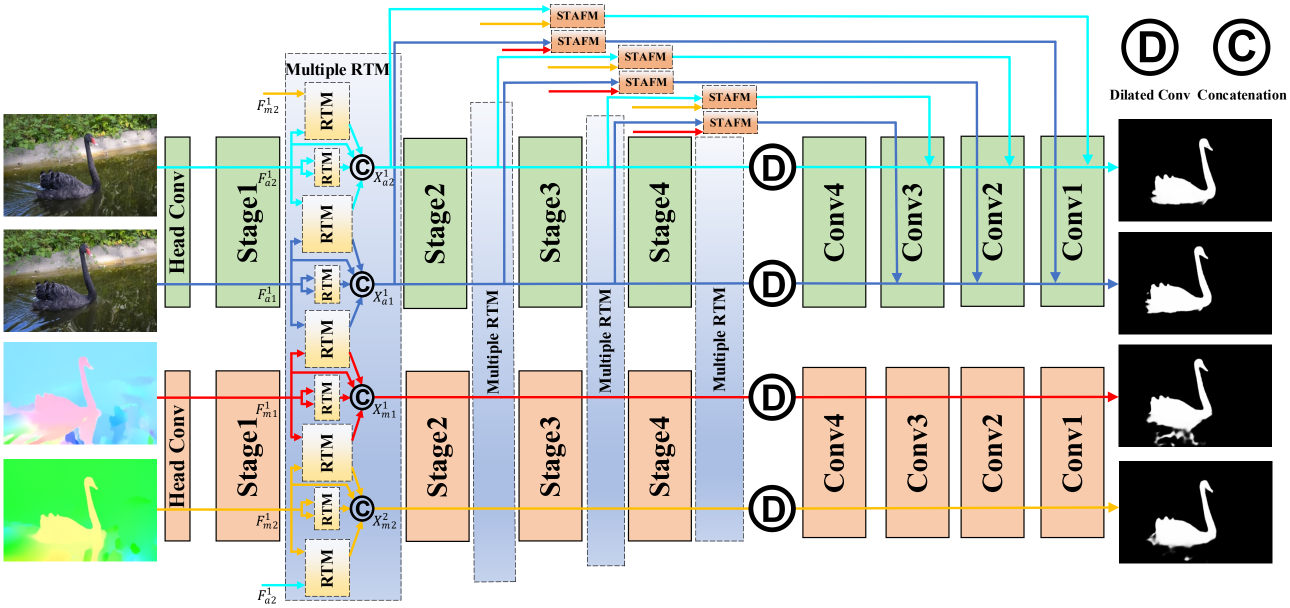

Figure 2: Illustration of Reciprocal Transformation Network which is composed of two network streams: the appearance stream and the motion stream. A pair of frames $\{I_{a_1}, I_{a_2}\}$ and their corresponding optical flows $\{I_{m_1}, I_{m_2}\}$ are fed into two streams, respectively. To correlate the features of the spatial and temporal domains, we propose the Reciprocal Transformation Modules (RTM) to transform the motion features, $F^i_{m_1}$ and $F^i_{m_2}$ $(i = \{1, 2\})$, and the appearance features, $F^i_{a_1}$ and $F^i_{a_2}$. Given pairwise features, RTM enables in-domain and cross-domain feature interactions for identifying salient objects in a single frame, moving objects, and recurring objects, respectively. In the end, the appearance features and motion features are fused by the Spatial Temporal Attentive Fusion Module (STAFM) and fed to the decoder of the appearance stream to generate the final mask.

video in RNN-based architecture. Song *et al.* [43] introduce a bi-directional ConvLSTM [41] for extracting features for multiple frames. However, RNN-based models suffer from the gradient vanishing problem and they are difficult to run in parallel. Furthermore, the attention-based methods are proposed for capturing long range dependency. COSNet [31] propose a co-attention layer to extract the discriminative foreground in a short video. ADNet [60] introduces an anchor frame to model the long-term dependency. MATNet [65] uses a motion-attentive transition to model motion information and spatio-temporal representation. Different from prior works, our reciprocal transformations leverage long range intra-frame contrast, temporal coherence, and motion-appearance similarity to enhance the appearance feature representation.

Attention Mechanism. Attention mechanism has been demonstrated effective and efficiency in many tasks due to its flexibility [27, 40, 49, 63]. The core idea of attention mechanisms is to highlight the task-specific discriminative regions in features. The attention scheme for videos has been explored in many aspects, including gating or pooling [14, 32], pose primitives [2, 15], graph representations [17, 57], recurrent memory models [31, 39], and self-attention [56]. In contrast to previous works, our proposed model measures the similarity for any pair of feature maps of frames and flows, and reciprocally transform the similarity to each other.

3. Proposed Method

3.1. System Overview

Given a pair of frames $\{I_{a_1}, I_{a_2}\}$ in a video and their corresponding optical flow $\{I_{m_1}, I_{m_2}\}$ computed by [44], we aims at segmenting the primary objects within I_{a_1} and I_{a_2}. Fig. 2 shows the pipeline of our method, which is composed of two main streams: the appearance stream with $\{I_{a_1}, I_{a_2}\}$ as inputs and the motion stream with $\{I_{m_1}, I_{m_2}\}$ as inputs. Each stream is an encoder-decoder architecture with skip connections [38]. In specific, we adopt ResNet [16] with dilated convolution [61] as backbone.

In each stage of the encoder, we introduce a Reciprocal Transformation Module (RTM), $\mathcal{F}_{RTM}$. As the main component of our framework, RTM is consisted of three submodules: reciprocal scaling, reciprocal transformation, and reciprocal gating. We will elaborate it in Sec. 3.2. Hence, we leverage RTM to enhance the pairwise appearance and

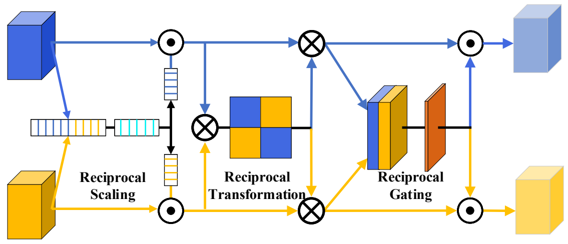

Figure 3: Our Reciprocal Transformation Module (RTM) includes reciprocal scaling, reciprocal transformation and reciprocal gating. $\odot$ and $\otimes$ indicate element-wise multiplication and matrix multiplication respectively.

motion features for identifying three types of candidate primary objects, i.e., salient objects within a frame, moving objects, and recurring objects in the video, which will be described in Sec. 3.3.

Last, in the decoder network of the motion stream, the skip connections bridge the low-level encoded features with the corresponding decoding layer, while the decoder of the appearance stream leverages the Spatial Temporal Attentive Fusion Module (STAFM) (Sec. 3.4) to fuse appearance and motion features to yield the final results.

3.2. Reciprocal Transformation Module

In this section, we will describe the structure of Reciprocal Transformation Module (RTM) (see Fig. 3) . Given two features (denoted as F_a and F_b which could be any pair from the appearance or motion features), we can mutually evolve and integrate the pairwise features via RTM. As described in the following subsections, our RTM includes three sub-modules: reciprocal scaling for adjusting the weights of different semantics, reciprocal transformation for measuring the similarity between feature maps, and reciprocal gating for balancing the transformed features.

3.2.1 Reciprocal Scaling

In deep neural networks, the channels of features represent different semantic meanings [12]. Thus, the values of channels can be accordingly scaled so as to bridge the attention of different primary objects for the features of different sources. In particular, as the motion features mainly focus on moving objects while the appearance features focus on salient objects, the reciprocal scaling enables to bridge their attention gap. Similarly, for the appearance features of two different frames, the reciprocal scaling may enable them to cast attention on different objects. To do so, we squeeze the combined feature maps to generate the representative value A_c for each channel and estimate the scaling factors w_1, w_2 that indicate the importance of semantics:

$$\begin{aligned} A_c &= \mathcal{F}_c(F_a; F_b), \\ w_1 &= \sigma(FC(\phi(FC(A_c; \theta_1); \theta_{21}))), \\ w_2 &= \sigma(FC(\phi(FC(A_c; \theta_1); \theta_{22}))), \end{aligned} \tag{1}$$

where $\mathcal{F}_c$ refers to the concatenation and squeezing (i.e., global average pooling) operation. $FC(\cdot; \theta)$ refers to the fully connected layer and ϕ is a ReLU activation function. Thus, we scale the channels of the features F_a and F_b according to w_1 and w_2:

$$\begin{aligned} F_a^{'} &= w_1 \odot F_a + F_a, \\ F_b^{'} &= w_2 \odot F_b + F_b, \end{aligned} \tag{2}$$

where $\odot$ is the element-wise multiplication.

3.2.2 Reciprocal Transformation

In this sub-module, we aim at measuring the similarity between feature maps, and transform both features to enhance the representation. Before explaining our reciprocal transformation, we introduce the principle of the vanilla self-transformation and nonreciprocal transformation.

Vanilla self-transformation. The local receptive fields of standard fully convolution network (FCN) based methods [29] confines the capability of networks in segmentation tasks that require rich context information. To capture long range dependency, Wang *et al.* [56] propose the self-attention scheme in which the feature maps are applied to measure the holistic similarity with itself to estimate its attention.

Nonreciprocal transformation. Nonreciprocal transformation is designed to compute the holistic similarity across different feature maps and manages to deliver the information from one feature to another. The nonreciprocal transformation is computed based on the similarity of the features from different sources. Given the feature maps for query $F_a^{'} \in \mathcal{R}^{c \times w \times h}$ and the target feature map $F_b^{'} \in \mathcal{R}^{c \times w \times h}$, it measures the holistic positional similarity matrix S using a non-local network structure:

$$S = Softmax(F_b^{'\mathrm{T}} W F_a^{'}), \tag{3}$$

where $W \in \mathcal{R}^{c \times c}$ is the similarity matrix. In practice, W contains a large amount of trainable parameters. To reduce the network complexity, we approximate W via two separate convolution operations:

$$\begin{aligned} \hat{F}_a^{'} &= Conv(F_a^{'}; \theta_a), \\ \hat{F}_b^{'} &= Conv(F_b^{'}, \theta_b), \\ S_{a \to b} &= Softmax(\hat{F}_b^{'\mathrm{T}} \times \hat{F}_a^{'}), \end{aligned} \tag{4}$$

where $Conv(F;\theta)$ is the convolution layer with the parameters θ on the feature map F and $\times$ refer to the matrix multiplication operator. Then, we derive the features according to the holistic positional similarity matrix S:

$$F^{''}_{a\to b} = Conv(F^{'}_a, \theta_{x_a}) \times S_{a\to b}, \tag{5}$$

where $F^{''}_{a\to b}$ is the enhanced features of $F^{'}_b$ from interacting with the feature $F^{'}_a$.

Reciprocal Transformation. Unlike nonreciprocal transformation, our reciprocal transformation use two features to mutually compensate each other, i.e.,:

$$\begin{aligned} F^{''}_{a\to b} &= Conv(F^{'}_a, \theta_{x_a}) \times S_{a\to b}, \\ F^{''}_{b\to a} &= Conv(F^{'}_b, \theta_{x_b}) \times S_{b\to a}. \end{aligned} \tag{6}$$

Using reciprocal transformation, the motion cues can be strengthened across the features of different sources. In specific, we can first transform the motion features to the appearance features to improve the segmentation ability for moving objects, and then the appearance features are reciprocally transformed to the motion features to eliminate the co-moving outliers.

3.2.3 Reciprocal Gating

Transformed features often have the different extents of importance. For instance, the appearance or motion noise like background variation or motion blur should be regarded as distractors with less importance. Therefore, we design a reciprocal gating mechanism to balance different transformed features:

$$\begin{aligned} G_a &= \sigma(Conv(F^{''}_{a\to b} \oplus F^{''}_{b\to a}; \theta_1)), \\ G_b &= \sigma(Conv(F^{''}_{a\to b} \oplus F^{''}_{b\to a}; \theta_2)), \end{aligned} \tag{7}$$

where $\oplus$ is the concatenation operation and σ is the sigmoid function. $G_a, G_b \in (0,1)$ are the reciprocal gates to balance the transformed features. Thus, we apply these gates to the original features:

$$\begin{aligned} X_{a\to b} &= G_b \odot F^{''}_{a\to b} + F^{''}_{a\to b}, \\ X_{b\to a} &= G_a \odot F^{''}_{b\to a} + F^{''}_{b\to a}. \end{aligned} \tag{8}$$

where $X_{a\to b}, X_{b\to a}$ are the final feature maps.

3.3. RTM-based Video Object Segmentation

Depending on the input features, RTM is enabling to interact and enhance the features of different sources for identifying the salient objects in a single frame, the moving objects, and the recurring objects, respectively. In general, on the i^{th} stage of the encoder network, we obtain the appearance features $F^i_{a_1}$ and $F^i_{a_2}$ from the input frames $\{I_{a_1}, I_{a_2}\}$, as well as the motion features $F^i_{m_1}$ and $F^i_{m_2}$ from the corresponding optical flows $\{I_{m_1}, I_{m_2}\}$.

Salient objects. To identify the salient objects in a single frame, RTM is utilized to obtain the intra-frame contrast by measuring the self-similarity of the appearance features or the motion features, which can be expressed as below.

$$\begin{aligned} X^i_{a_1\to a_1} &= \mathcal{F}_{RTM}(F^i_{a_1}, F^i_{a_1}; \theta_a), \\ X^i_{a_2\to a_2} &= \mathcal{F}_{RTM}(F^i_{a_2}, F^i_{a_2}; \theta_a), \\ X^i_{m_1\to m_1} &= \mathcal{F}_{RTM}(F^i_{m_1}, F^i_{m_1}; \theta_m), \\ X^i_{m_2\to m_2} &= \mathcal{F}_{RTM}(F^i_{m_2}, F^i_{m_2}; \theta_m), \end{aligned} \tag{9}$$

where $X^i_{a_1\to a_1}$ and $X^i_{a_2\to a_2}$ represent the self-similarity of the appearance features $F^i_{a_1}$ and $F^i_{a_2}$. $X^i_{m_1\to m_1}$ and $X^i_{m_2\to m_2}$ are the self-similarity of the motion features $F^i_{m_1}$ and $F^i_{m_2}$. θ_a, θ_m are the parameters of the corresponding RTMs.

Recurring objects. To identify the recurring objects, the spatio-temporal correlation between the input frames will be measured, so as to capture the long range dependency in two separate frames. Thus, we have:

$$\begin{aligned} X^i_{a_1\to a_2}, X^i_{a_2\to a_1} &= \mathcal{F}_{RTM}(F^i_{a_1}, F^i_{a_2}; \theta_{aa}), \\ X^i_{m_1\to m_2}, X^i_{m_2\to m_1} &= \mathcal{F}_{RTM}(F^i_{m_1}, F^i_{m_2}; \theta_{mm}), \end{aligned} \tag{10}$$

where $X^i_{a_1\to a_2}$ and $X^i_{a_2\to a_1}$ are the similarity between the appearance features of two frames, $F^i_{a_1}$ and $F^i_{a_2}$. $X^i_{m_1\to m_1}$ and $X^i_{m_2\to m_2}$ refer to the similarity between the motion features of two frames, $F^i_{m_1}$ and $F^i_{m_2}$.

Moving objects. We associate the motion features with the appearance features for identifying the moving objects, by computing the similarity of salient appearance and motion cues. Likewise, by associating the motion features with the appearance features, we can eliminate the co-moving outliers:

$$\begin{aligned} X^i_{m_1\to a_1}, X^i_{a_1\to m_1} &= \mathcal{F}_{RTM}(F^i_{a_1}, F^i_{m_1}; \theta_{am}), \\ X^i_{m_2\to a_2}, X^i_{a_2\to m_2} &= \mathcal{F}_{RTM}(F^i_{a_2}, F^i_{m_2}; \theta_{am}), \end{aligned} \tag{11}$$

where $X^i_{m_1\to a_1}$ and $X^i_{a_1\to m_1}$ are the simiarity between the motion and appearance features, $F^i_{m_1}$ and $F^i_{a_1}$. $X^i_{m_2\to a_2}$ and $X^i_{a_2\to m_2}$ refer to the similarity between the motion and appearance features, $F^i_{m_2}$ and $F^i_{a_2}$.

Hence, the final appearance features are the combination of the intra-frame contrast, temporal coherence, and motion cues, as below:

$$\begin{aligned} X^i_{a_1} &= F^i_{a_1} + X^i_{a_1\to a_1} + X^i_{a_2\to a_1} + X^i_{m_1\to a_1}, \\ X^i_{a_2} &= F^i_{a_2} + X^i_{a_2\to a_2} + X^i_{a_1\to a_2} + X^i_{m_2\to a_2}. \end{aligned} \tag{12}$$

Similarly, the motion features are defined as below.

$$\begin{aligned} X^i_{m_1} &= F^i_{m_1} + X^i_{m_1\to m_1} + X^i_{m_2\to m_1} + X^i_{a_1\to m_1}, \\ X^i_{m_2} &= F^i_{m_2} + X^i_{m_2\to m_2} + X^i_{m_1\to m_2} + X^i_{a_2\to m_2}. \end{aligned} \tag{13}$$

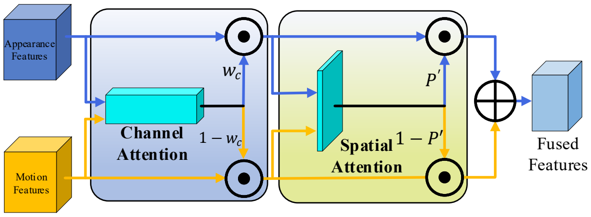

Figure 4: Our Spatial Temporal Attentive Fusion Module (STAFM) including channel attention and spatial attention. $\odot$ and $\oplus$ indicate element-wise addition and matrix multiplication respectively.

3.4. Spatial Temporal Attentive Fusion Module

Inspired by CBAM [59], we aim at selectively fusing appearance features and motion features from each network stage via channel attention and spatial attention. As illustrated in Fig. 4, with the appearance features $X_a \in \mathcal{R}^{c\times w\times h}$ and motion features $X_m \in \mathcal{R}^{c\times w\times h}$, we design a gating mechanism to adjust the weight of each channel for all the feature maps:

$$\begin{aligned} A_c &= \mathcal{F}_c(X_a; X_m), \\ w_c &= \sigma(FC(\phi(FC(A_c;\theta_1);\theta_2))), \\ w_c^a &= w_c,\ w_c^m = 1 - w_c, \end{aligned} \tag{14}$$

where $\mathcal{F}_c$ is the concatenation and squeezing operation. $FC(X;\theta)$ is a fully connected layer, ϕ is the ReLU activation function. The fully connected layers and these activation functions jointly serve as the excitation operation. $\{w_c \in (0,1)\,|w_c \in \mathcal{R}^c\}$ refers to the channel gate. Thus, we use the channel gate to enhance motion and appearance features as below:

$$\begin{aligned} X_a^{'} &= w_c^a \odot X_a, \\ X_m^{'} &= w_c^m \odot X_m, \end{aligned} \tag{15}$$

In addition, we exploit the spatial relationship across motion and appearance features to infer the spatial attention, which can guide the adaptive fusion of the motion and appearance features. To do so, we first calculate the spatial features for the appearance features $X_a^{'}$ and motion features $X_m^{'}$ enhanced by channel attention:

$$\begin{aligned} P_a &= MaxPool(X_a^{'}) \oplus AvgPool(X_a^{'}), \\ P_m &= MaxPool(X_m^{'}) \oplus\ AvgPool(X_m^{'}), \\ P^{'} &= \sigma(Conv(P_a \oplus P_m; \theta_p)), \\ P_a^{'} &= P^{'}, P_m^{'} = 1 - P^{'}, \end{aligned} \tag{16}$$

where P_a and P_m represent the spatial attention maps of $X_a^{'}$ and $X_m^{'}$. $MaxPool$ represents the max pooling layer and $AvgPool$ represents the average pooling layer. Then, we leverage the attention maps to fuse the appearance and motion features, as follows.

$$X_a^{''} = P_a^{'} \odot X_a^{'} + P_m^{'} \odot X_m^{'}, \tag{17}$$

where $X_a^{''}$ is the fused features which will be passed to the decoder of the appearance network stream.

3.5. Loss Function

We adopt the same loss function as BASNet [36] to jointly measure the prediction in the pixel level by cross entropy loss [8], in the patch level by SSIM loss [58], as well as in the region level by IoU loss [62]:

$$L(M,G) = l_{ce}(M,G) + l_{ssim}(M,G) + l_{iou}(M,G), \tag{18}$$

where M denotes the segmentation mask and G refers to the ground-truth.

Given an pair of image I_{a_1}, I_{a_2} and their corresponding optical flow I_{m_1}, I_{m_2}, we have four output segmentation masks from four stages of the decoder. Hence, the total loss is defined as below:

$$\mathcal{L} = \sum_{i=1}^{2}\sum_{j=1}^{4} L(M_{a_i}^j, G_{a_i}^j) + L(M_{m_i}^j, G_{m_i}^j). \tag{19}$$

4. Experiment

4.1. Implementation Details and Datasets

Implementation Details. In the following experiments, we test our model with ResNet-34 and ResNet-101 as backbone networks. We pre-train our appearance stream and the motion stream without the transformation and fusion modules on the saliency detection dataset DUTS [50] and video object segmentation dataset DAVIS-16 [34] dataset respectively. Then, we freeze the weights of the backbone networks and train the transformation modules and the feature fusion module on DAVIS-16 dataset. During above training period, the learning rate is set as 1e-3 and we fine-tune the whole model with the learning rate of 5e-4. We apply SGD as the optimizer with momentum of 5e-4. During training, we randomly sample frame pairs in the same video. We adopt the data argumentation strategies including the vertical/horizontal flip and multi-scale training. On the test time, to produce the segmentation masks of primary objects in the target frame, we randomly select another frame in the same video as reference. Following the practice of [31], we also apply CRF-based post-processing technique.

Datasets. We evaluate our method on two public dataset: DAVIS-16 [34] contains 50 high-quality videos of 480p and 720p with high quality dense pixel-level annotations and YouTube-Objects [35] contains 126 videos over 20,000 frames and 10 semantic categories. We adopt region similarity $\mathcal{J}$ and boundary accuracy $\mathcal{F}$ as evaluation metrics.

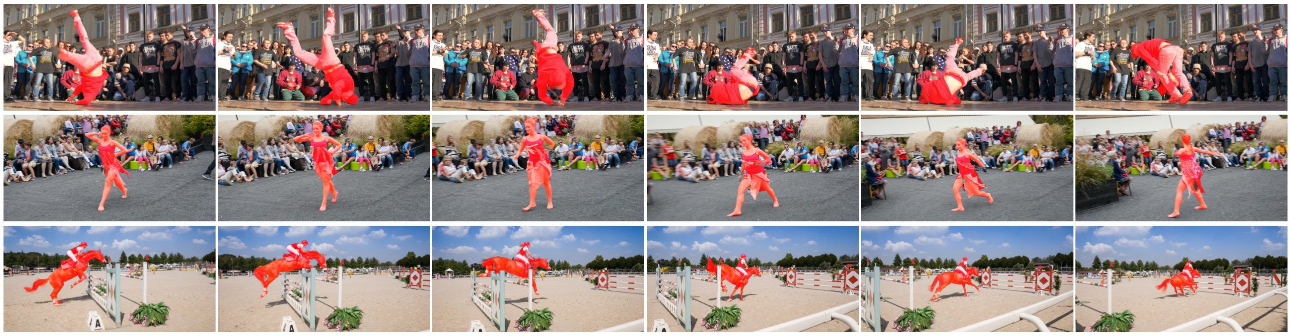

Figure 5: Qualitative results on DAVIS-16: *breakdance*, *horsejump-high* and *dance-twirl*

Method	Backbone	$\mathcal{J}$		$\mathcal{F}$	
		Mean	Recall	Mean	Recall
LMP	FLowNet	70.0	85.0	65.9	79.2
LVO	DeepLab	75.9	89.1	72.1	83.4
PDB	ResNet-50	77.2	90.1	74.5	84.4
LMSO	ResNet-101	78.2	89.1	75.9	84.7
MOTAdapt	ResNet-101	77.2	87.8	77.4	84.4
AGS	ResNet-101	79.7	91.1	77.4	85.8
COSNet	ResNet-101	80.5	93.1	79.5	89.5
ADNet	ResNet-101	81.7	90.9	80.5	85.1
MATNet	ResNet-101	82.4	94.5	80.7	90.2
DFNet	DeepLab	83.4	-	81.8	-
Ours-Light	ResNet-34	84.8	95.8	83.5	93.1
Ours	ResNet-101	85.6	96.1	84.7	93.8

Table 1: Quantitative results on DAVIS-16. Res-*m* indicates the number of layers in the backbone. The top three performers are marked in Red, Green, and Blue, respectively.

4.2. Comparison with State-of-the-arts

We compare with the previous methods: LMP [29] LVO [47], PDB [43], LSMO [48], MOTAdapt [42], AGS [55], COSNet [31], ADNet [60], MATNet [65], DFNet [64].

Evaluation on DAVIS-16. We evaluate our RTNet with state-of-the-art unsupervised video object segmentation methods. The quantitative results are reported in Table 1. LMP [29] tries to prediction primary objects based on the optical flow only. However, due to the lack of appearance features, this method has the worst performance in localizing and segmenting the primary objects. Some methods [31, 43, 60] rely on the appearance features without the guidance of optical flow also achieves comparable performance, because these methods extract the appearance features in sequences based on the spatial-temporal architectures like ConvLSTM. Among them, COSNet [31] and ADNet [60] leverage the attention mechanism and show strong ability to model global sequence information. AGS [55] takes extra visual attention annotations and is more powerful to locate primary objects. More methods [47, 48, 65] including us take both appearance featrues and motion features into consideration. With the reciprocal transformation, our method is able to identify all kinds of candidate primary objects. In particular, our lightweight model using ResNet-34 as backbone ("Ours-Light" in Table 1) outperforms the both motion and appearance based method MATNet by 2.91% on $\mathcal{J}$ and 3.47% on $\mathcal{F}$. MATNet uses ResNet-101 as backbone and around 12k video frames for training, while our lightweight version model adopts ResNet-34 as backbone and around 2k frames for training. Our full model using ResNet-101 ("Ours" in Table 1) shows even better performance. It achieves 0.94% on $\mathcal{J}$ over our lightweight weight model and 3.88% on $\mathcal{J}$ over the second best model.

Fig. 5 shows our qualitative results on DAVIS-16 which contains challenging scenarios like complex background and motion blur. Our RTNet precisely captures the location of primary objects and segments them with sharp boundaries, thanks to the transformed intra-frame contrast, moving cues and temporal coherence. The effectiveness can be specially observed in *breakdance*, where there are plenty of appearance-similar humans standing behind the dancing man. Besides, with our multi-stage reciprocal transformation, objects of different scales in *horsejump-high* can be accurately segmented.

Evaluation on YouTube-Objects. We report the performance of our RTNetet on YouTube-Objects dataset in Table 2. Our method achieve the best performance over all the comparison methods under the region similarity $\mathcal{J}$. For the slow moving objects (*i.e.Airplane* and *Boat*) with ambiguous background, it is difficult for the optical flow based model (i.e. MATNet) to capture the primary objects, while our model fuses temporal coherence and intra-frame contrast information and thus significantly outperform MATNet. For the moving objects with salient appearance (*i.e.*, *Bird* and *Cat*), motion based method including ours and MATNet outperform the appearance based method (i.e. COSNet).

Category	LVO	PDB	MATNet	AGS	COSNet	Ours
Airplane	86.2	78.0	72.9	87.7	81.1	84.1
Bird	81.0	80.0	77.5	76.7	75.7	80.2
Boat	68.5	58.9	66.9	72.2	71.3	70.1
Car	69.3	76.5	79.0	78.6	77.6	79.5
Cat	58.8	63.0	73.7	69.2	66.5	71.8
Cow	68.5	64.1	67.4	64.6	69.8	70.1
Dog	61.7	70.1	75.9	73.3	76.8	71.3
Horse	53.9	67.6	63.2	64.4	67.4	65.1
Motorbike	60.8	58.3	62.6	62.1	67.7	64.6
Train	66.3	35.2	51.0	48.2	46.8	53.3
Mean $\mathcal{J}$	67.5	65.4	69.0	69.7	70.5	**71.0**

Table 2: Quantitative results of each category on YouTube-Objects dataset over regional similarity (mean $\mathcal{J}$).

Model	Mean $\mathcal{J}$	$\Delta\mathcal{J}$	Mean $\mathcal{F}$	$\Delta\mathcal{F}$
Baseline	77.53	-	76.26	-
Baseline+S	78.19	0.66	77.30	1.04
Baseline+M	83.01	5.48	82.19	5.93
Baseline+R	79.57	2.04	79.22	2.96
Baseline+SM	83.51	5.98	81.97	5.71
Baseline+SR	80.11	2.58	79.95	3.69
Baseline+MR	83.43	5.90	82.44	6.18
Baseline+SMR	83.96	6.43	82.65	6.39

Table 3: Ablation study for three types of primary objects. S, M, R indicate intra-frame salient objects, moving objects, and recurring objects, respectively.

4.3. Ablation Study

Our ablation experiment is conducted based on our lightweight model without applying the CRF post-processing operation on DAVIS-16 dataset.

Primary Objects. We evaluate the effectiveness of our proposed model for transforming intra-frame saliency, moving cues, and temporal coherence. The results are reported in Table 3. We use the vanilla encoder-decoder architecture without any transformation or feature fusion module as *Baseline*. Then, we search three kinds of candidate primary objects and transform the intra-frame contrast features for local salient objects (denoted as *S*), the motion features for moving objects (denoted as *M*) and the temporal coherence for recurring objects (denoted as *R*). On the one hand, we find that identifying the three kinds of primary objects all contribute to the whole model according to performance gain while considering one more kind of primary objects. On the other hand, we find that the performance improvement is most obvious. Therefore, the moving objects play the most important roles in primary objects comparing with salient objects and the recurring objects.

Model	Mean $\mathcal{J}$	$\Delta\mathcal{J}$	Mean $\mathcal{F}$	$\Delta\mathcal{F}$
Baseline	77.53	-	76.26	-
Nonreciprocal	83.01	5.51	82.19	5.93
Reciprocal	83.74	6.21	82.57	6.31

Table 4: Ablation study of reciprocal mechanism.

Model	Mean $\mathcal{J}$	$\Delta\mathcal{J}$	Mean $\mathcal{F}$	$\Delta\mathcal{F}$
Ours w/o STAFM	84.15	-	82.89	-
Ours w/ STAFM	84.31	0.16	83.01	0.12

Table 5: Ablation study for STFAM.

Reciprocal Mechanism. We study the quality of the motion features transformed to the appearance to show the effectiveness of our reciprocal transformation between motion features and appearance features. We adopt the vanilla encoder-decoder architecture without transforming any motion features as *Baseline*. Then, we transform the motion features directly from the motion stream without salient appearance as *Nonreciprocal*. Furthermore, our reciprocal transformation for transforming between appearance and motion with moving primary objects as *Reciprocal*. The results are reported in Table 4.

STAFM. To evaluate the effectiveness of our STAFM, we compare the performance for our RTNet with STAFM (*Ours w/ STAFM*) and the model simply skip connect both motion and appearance features (*Ours w/o STAFM*). The results are reported in Table 5. The gain of our STAFM comes from the fusion of the crucial spatial and temporal features while removing the redundant features.

5. Conclusion

In these paper, we propose the reciprocal transformations to identify the three kind of primary objects: salient objects, recurring objects and moving objects by searching the intra-frame dependency, the correlation between motion and appearance features, and temporal coherence to the appearance features. Besides, to eliminate the moving background objects, the reciprocal scheme transform appearance features back to motion features to filter moving objects with distinct appearance. Finally, we propose a spatial temporal attentive feature fusion module to dynamically and selective fuse spatial and temporal features.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61972162, 62072110), the Natural Science Foundation of Fujian Province under Grant 2018J07005, and CCF-Tencent Open Research fund.

References

[1] Linchao Bao, Baoyuan Wu, and Wei Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *CVPR*, pages 5977–5986, 2018. 2

[2] Fabien Baradel, Christian Wolf, and Julien Mille. Human action recognition: Pose-based attention draws focus to hands. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 604–613, 2017. 3

[3] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, pages 282–295. Springer, 2010. 2

[4] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, pages 221–230, 2017. 2

[5] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *ICCV*, pages 2722–2730, 2015. 1

[6] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *CVPR*, pages 1189–1198, 2018. 2

[7] 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

[8] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005. 6

[9] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019. 1

[10] Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving objects in videos. In *CVPR*, June 2015. 2

[11] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, pages 1846–1853. IEEE, 2012. 2

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019. 4

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 1

[14] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *NeurIPS*, pages 34–45, 2017. 1, 3

[15] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*, pages 971–980, 2017. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[17] Roei Herzig, Elad Levi, Huijuan Xu, Eli Brosh, Amir Globerson, and Trevor Darrell. Classifying collisions with spatio-temporal action graph networks. *arXiv preprint arXiv:1812.01233*, 2, 2018. 3

[18] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 3203–3212, 2017. 1

[19] 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. 2

[20] Varun Jampani, Raghudeep Gadde, and Peter V Gehler. Video propagation networks. In *CVPR*, pages 451–461, 2017. 2

[21] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *CVPR*, pages 3442–3450, 2017. 2

[22] Changick Kim and Jenq-Neng Hwang. Fast and automatic video object segmentation and tracking for content-based applications. *IEEE TCSVT*, 12(2):122–129, 2002. 1

[23] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *ICCV*, pages 1995–2002. IEEE, 2011. 2

[24] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, pages 7274–7283, 2019. 1, 2

[25] 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. 2

[26] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, pages 90–105, 2018. 2

[27] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017. 3

[28] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018. 1

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2, 4, 7

[30] Xiankai Lu, Chao Ma, Bingbing Ni, Xiaokang Yang, Ian Reid, and Ming-Hsuan Yang. Deep regression tracking with shrinkage loss. In *ECCV*, pages 353–369, 2018. 1

[31] 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. 1, 3, 6, 7

[32] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 1, 3

[33] Peter Ochs and Thomas Brox. Object segmentation in video: a hierarchical variational approach for turning point trajecto-

ries into dense regions. In *ICCV*, pages 1583–1590. IEEE, 2011. 2

[34] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2, 6

[35] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, pages 3282–3289. IEEE, 2012. 6

[36] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, June 2019. 6

[37] Sucheng Ren, Chu Han, Xin Yang, Guoqiang Han, and Shengfeng He. Tenet: Triple excitation network for video salient object detection. 1

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[39] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015. 3

[40] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*, 2018. 3

[41] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *NeurIPS*, 28:802–810, 2015. 3

[42] 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 *ICRA*, pages 50–56. IEEE, 2019. 7

[43] 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. 2, 3, 7

[44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. *arXiv preprint arXiv:2003.12039*, 2020. 3

[45] Ying-Li Tian, Max Lu, and Arun Hampapur. Robust and efficient foreground analysis for real-time video surveillance. In *CVPR*, volume 1, pages 1182–1187. IEEE, 2005. 1

[46] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, pages 3386–3394, 2017. 2

[47] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *ICCV*, Oct 2017. 1, 7

[48] Pavel Tokmakov, Cordelia Schmid, and Karteek Alahari. Learning to segment moving objects. *IJCV*, 127(3):282–301, 2019. 1, 7

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3

[50] 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. 6

[51] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, pages 1328–1338, 2019. 1

[52] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *CVPR*, pages 3127–3135, 2018. 1

[53] Wenguan Wang, Jianbing Shen, Xuelong Li, and Fatih Porikli. Robust video object cosegmentation. *IEEE TIP*, 24(10):3137–3148, 2015. 2

[54] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, pages 3395–3402, 2015. 2

[55] 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. 7

[56] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 1, 3, 4

[57] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, pages 399–417, 2018. 3

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[59] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 6

[60] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *CVPR*, pages 931–940, 2019. 1, 3, 7

[61] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[62] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *ACM MM*, pages 516–520, 2016. 6

[63] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, pages 3085–3094, 2019. 3

[64] 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 *ECCV*, pages 445–462. Springer, 2020. 7

[65] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE TIP*, 29:8326–8338, 2020. 1, 2, 3, 7