Dynamic Context-Sensitive Filtering Network for Video Salient Object Detection

Miao Zhang[1, 2*] Jie Liu[1*] Yifei Wang [1*] Yongri Piao[1†] Shunyu Yao[1] Wei Ji[4]
Jingjing Li[4] Huchuan Lu[1,3] Zhongxuan Luo [1]
[1] Dalian University of Technology, China
[2] Key Lab for Ubiquitous Network and Service Software of Liaoning Province,
Dalian University of Technology, China
[3] Pengcheng Lab
[4] University of Alberta, Canada
{miaozhang, yrpiao, lhchuan, zxluo}@dlut.edu.cn
{1605721375, dilemma, ysyfeverfew}@mail.dlut.edu.cn {wji3, jingjin1}@ualberta.ca

Abstract

The ability to capture inter-frame dynamics has been critical to the development of video salient object detection (VSOD). While many works have achieved great success in this field, a deeper insight into its dynamic nature should be developed. In this work, we aim to answer the following questions: How can a model adjust itself to dynamic variations as well as perceive fine differences in the real-world environment; How are the temporal dynamics well introduced into spatial information over time? To this end, we propose a dynamic context-sensitive filtering network (DCFNet) equipped with a dynamic context-sensitive filtering module (DCFM) and an effective bidirectional dynamic fusion strategy. The proposed DCFM sheds new light on dynamic filter generation by extracting location-related affinities between consecutive frames. Our bidirectional dynamic fusion strategy encourages the interaction of spatial and temporal information in a dynamic manner. Experimental results demonstrate that our proposed method can achieve state-of-the-art performance on most VSOD datasets while ensuring a real-time speed of 28 fps. The source code is publicly available at https://github.com/OIPLab-DUT/DCFNet.

1. Introduction

Videos as one of the most engaging mediums strike a deep connection with humans. As a fundamental task in video processing, video salient object detection (VSOD) aims to explore this connection and segment most visually distinctive regions in videos. This task has drawn broad attention due to a wide range of applications in video object segmentation [19, 33, 48], visual tracking [52], video captioning [36], video compression [17, 14] and medical analysis [18, 22]. Compared with still-image based SOD tasks, VSOD does not only suffer from processing a huge amount of data but also is directly affected by temporal dynamics. The substantial differences make VSOD more challenging than still-image based SOD task.

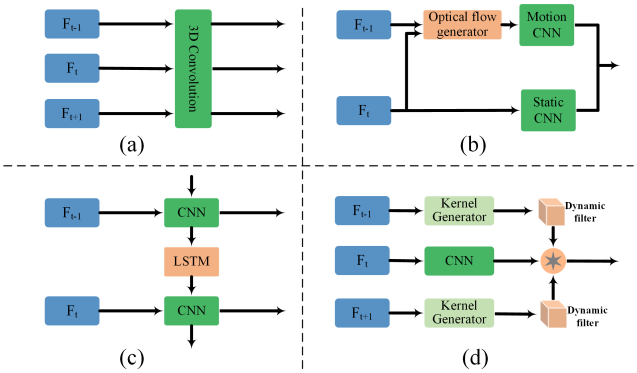


Figure 1. Architecture comparison of our dynamic filtering based method (d) with 3D Convolution (a), Optical Flow (b) and ConvLSTM (c) based methods.

Most existing VSOD methods, which can be classified into 3D convolution based [24, 25], ConvLSTM based [12, 44] and optical flow based [28, 43] methods as shown in Figure 1, employ fixed parameter layers during inference. Given that our world is constantly changing, performing convolution with dynamic parameters conditioned on inputs can better adapt to dynamic real-world environments [3, 34]. However, directly applying the dynamic fil-

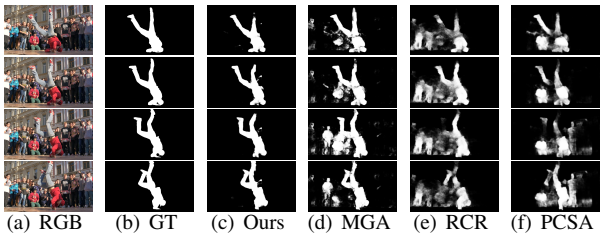
*Equal Contributions
†Corresponding Author

tering mechanism to the VSOD task may fail to comprehensively utilize inter-frame contextual information. Therefore, this may impair these methods to achieve high accuracy for saliency prediction.

Moreover, when any events that happen in the real world are condensed into seconds, the pixels in different frames can be temporally inconsistent over time. Such time taking the form of objects moving between consecutive frames makes VSOD very challenging. For instance, both moving foreground and background objects in a video clip enable some representative VSOD methods to be less effective, as illustrated in Figure 2. Given that spatial and temporal domains are entangled in video, sufficient spatiotemporal fusion is the cornerstone of VSOD. It further extends how the temporal dynamics are incorporated into spatial information over time.

In this paper, we strive to confront challenges towards accurate VSOD. The primary challenge towards this goal is to design a model capable of not only adapting to dynamic changes but also distinguishing fine differences in the real-world environment. The second challenge is to dynamically formulate the cross-domain complementarity, adaptively allowing more effective fusion. The key aspect in the success of our method is in its ability to better dynamically adjust itself to our constantly-changing world. Concretely, our contributions are fourfold:

- We propose a dynamic context-sensitive filtering module (DCFM). DCFM can estimate the location-related affinity weights to dynamically generate context-sensitive convolution kernels, thus promoting the model's adaptability to constantly changing scenes.
- We introduce a bidirectional dynamic fusion strategy to encourage the bidirectional interaction between spatial and temporal domains. As a result, the proposed strategy helps our network combine cross-domain features and ensures high stability for saliency detection in the challenging scenes.
- Furthermore, we conduct extensive experiments on 5 widely-used datasets and demonstrate that our method outperforms 12 state-of-the-art VSOD approaches in terms of 3 evaluation metrics. Especially, our approach reduces the *MAE* metric by 34.8% and 27.3% on SegV2 [26] and DAVIS [39] respectively, which are dominated by fast and moderate moving objects respectively, showing the adaptability of our model in different video scenarios.
- The proposed DCFM can be extended to improve the existing still-image SOD based models. Experiments demonstrate that compared with the original models, the new ones embedded with the DCFM achieve better performance on all the evaluation metrics.



(a) RGB (b) GT (c) Ours (d) MGA (e) RCR (f) PCSA

Figure 2. Sample prediction results of our methods compared to MGA [28], RCRNet [57] and PCSA [13]. Column 'RGB' shows raw images of a consecutive video sequence from DAVIS dataset. Column 'GT', 'Ours', 'MGA', 'RCR' and 'PCSA' denote ground truth, corresponding predictions from our methods, MGA, RCRNet and PCSA respectively.

2. Related Work

Video Salient Object Detection. Salient object detection (SOD) in color images [7, 62, 54], RGB-D images [4, 20, 21, 40], light-field images [30, 59, 60], and videos, have always been an active field of research. In this paper, we will mainly study the video-based SOD task. Existing video salient object detection methods can be generally classified into two categories: (1) traditional methods; (2) deep learning based methods. Traditional methods mainly rely on hand-crafted features and prior knowledge, such as color-contrast, background prior and morphology cues. Esa Rahtu *et al*. [42] combine a statistical framework and local feature contrast for reasoning saliency maps. Based on the superpixel representation of video frames, Liu *et al*. [32] extract motion and color histograms at both superpixel and frame level to generate predictions. Later, Xi *et al*. [56] compute the appearance and motion saliency maps using spatiotemporal background priors. Although traditional approaches are efficient for simple VSOD tasks, they may suffer from issues including sensitivity to complex scenes and limited capability for discriminative information.

Recently, deep learning based methods have shown promising prospects in VSOD. Gu *et al*. [13] design a pyramid constrained self-attention module for capturing temporal information directly. Yan *et al*. [57] extract spatiotemporal coherence by introducing a refinement network equipped with a non-locally recurrent module and propose a pseudo-label generator for auto labeling datasets. Fan *et al*. [12] present a baseline model with ConvLSTM and propose a densely annotated VSOD dataset. All these methods either adopt 3D convolution or ConvLSTM to model temporal coherence. Besides, to model motion cues explicitly, optical flow based methods are also proposed. Li *et al*. [28] propose a two-stream architecture to combine the appearance and optical flow information. Li *et al*. [27] design a flow guided recurrent encoder for enhancing temporal coherence by simultaneously utilizing an optical flow network and a feature extractor with ConvLSTM.

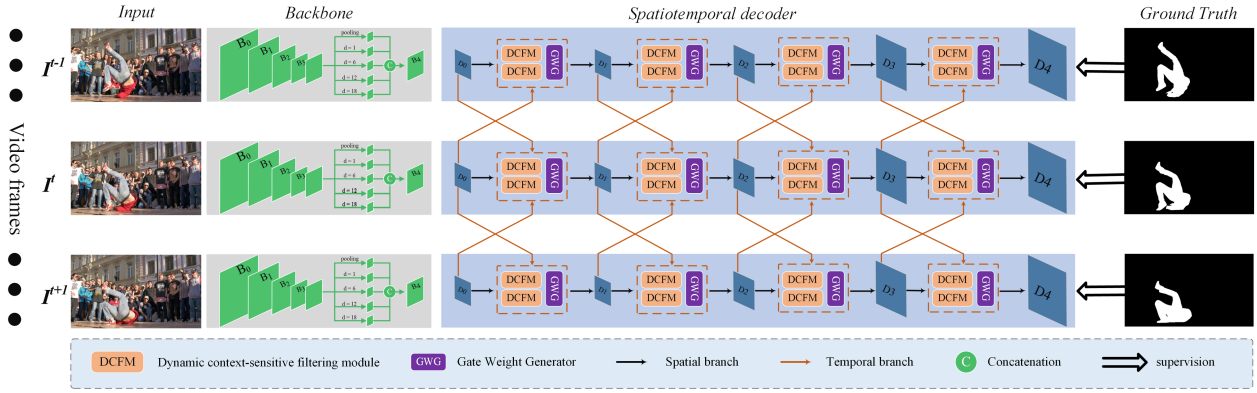


Figure 3. Overall architecture of our proposed DCFNet. D_0 to D_4 stand for feature maps with different spatial resolutions.

Dynamic Filtering Mechanism. Originally introduced by Brabandere *et al*. [3], the dynamic filtering mechanism provides adjustable convolution kernel parameters conditioned on different inputs, thus providing a powerful yet flexible way for feature utilization. Recently, several methods have explored dynamic filters in different fields. Wu *et al*. [53] propose a dynamic filtering strategy with large sampling field, enabling dynamic kernels to learn from diverse feature regions for image based tasks. He *et al*. [15] strive to adaptively capture multi-scale dynamic contents for predicting pixel-level semantic labels. In the field of RGB-D SOD, Pang *et al*. [37] integrate features of different modalities, and use their mixed features to generate dynamic filters. While the dynamic filtering mechanism has been adopted in video based tasks, such as human action recognition [10] and video deblurring [61], the significance of the dynamic filtering mechanism in video salient object detection has not been fully studied. Considering the huge amount of spatiotemporal data brought by the additional time dimension, simply applying the dynamic filtering mechanism to the video-based SOD task inevitably leads to inaccurate saliency predictions. Therefore, a suitable design tailored for the VSOD task is putting forward.

3. The Proposed Method

3.1. Architecture Overview

We first describe the overall architecture of the DCFNet shown in Figure 3. DCFNet follows the encoder-decoder architecture. It takes a video clip consisting of three consecutive frames I^{t-1}, I^t, and I^{t+1} as input, generating dense saliency prediction for I^t. In terms of the encoder, we utilize a ResNet-101 [16] as our backbone network for feature extraction. It generates four feature maps with different spatial resolution and channel number. Inspired by [5], last two layers are discarded to preserve spatial structure, then replaced with an atrous spatial pyramid pooling (ASPP) layer for extracting multi-scale contextual information. Outputs of first three blocks, denoted B_1^t, B_2^t, B_3^t, and output of ASPP B_4^t are served as inputs to the decoder.

Figure 4 shows the structure of the decoder. It first constructs an interleaved feature fusion layer, in which for each B_i^t, other three feature maps are resized to its spatial resolution and fused using point-wise addition to produce an enriched feature representation $\widetilde{B}_i^t$. After feature fusion, decoder is arranged into four stages. The input of each stage consists of two categories: spatial features from the feature fusion layer and temporal features from adjacent frames. Spatial and temporal features are progressively aggregated into spatiotemporal features by the decoder. It performs spatiotemporal fusion by leveraging DCFMs, and finally generates saliency prediction D_4 at stage four.

3.2. Dynamic Context-sensitive Filtering Module

Constant changes in the real world make fixed parameter networks less adaptive to dynamic video scenarios. A straightforward solution to this issue is to directly introduce the dynamic filtering mechanism into network design [3, 35]. However, direct application just simply stack several convolution layers to generate dynamic convolution kernels. This may prevent them from fully extracting contextual information of consecutive frames, due to limited size of receptive fields. In other words, the generated kernels cannot provide sufficient guidance for achieving high prediction accuracy. To solve this issue, we propose a dynamic context-sensitive filtering module (DCFM). DCFM estimates the location-related affinity weights by introducing matrix multiplication into the kernels' generation process. Therefore, dynamic convolution kernels can extract rich contextual features that are not restricted by the receptive fields, providing comprehensive guidance for finer saliency prediction.

Specifically, we arrange three context-sensitive filter-generating networks in a pyramid structure, generating three dynamic convolution kernels, then perform dilation convolution with different dilation rates respectively. Fi-

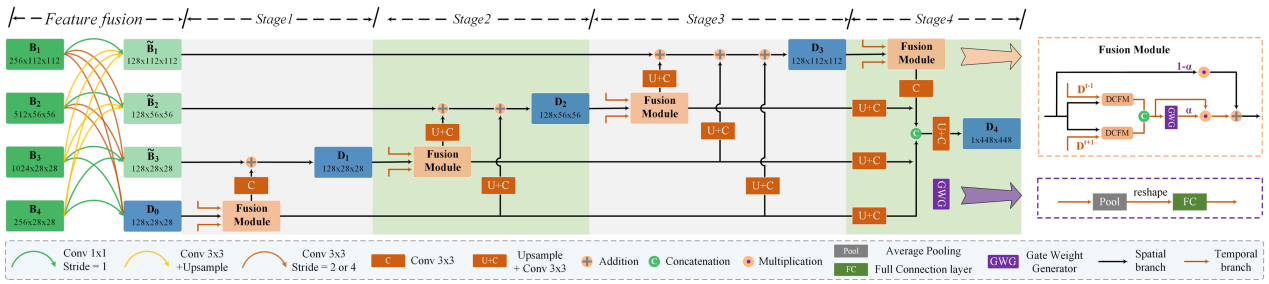


Figure 4. Detailed architecture of proposed decoder. D_0 to D_4 have the same meaning as in Figure 3.

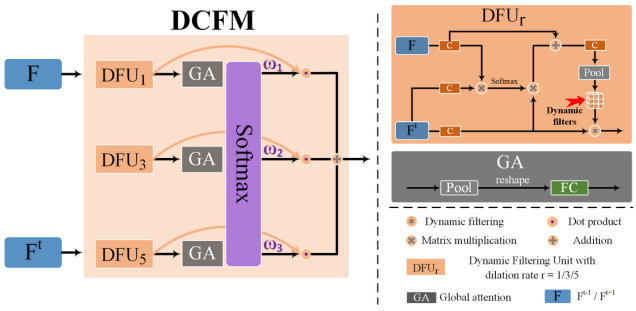


Figure 5. Detailed structure of the dynamic context-sensitive filtering module.

nally, the global attention mechanism is applied to combine three branches' outputs.

DCFM takes two feature maps F^{t-1}, F^t as input. Each filter-generating network then generates a dynamic kernel $\widetilde{K}_i^t$, where the subscript $i \in \{1, 2, 3\}$ is added to distinguish between different kernels. First, the network generates three feature maps as shown below

$$\begin{aligned} \Theta_i^t &= Conv_{1\times1}(F^{t-1}), \\ \Phi_i^t &= Conv_{1\times1}(F^t), \\ \Omega_i^t &= Conv_{1\times1}(F^t), \end{aligned} \tag{1}$$

where $Conv_{1\times1}$ denotes 1×1 convolution that reduce feature map's dimension to $\mathbb{R}^{H\times W\times \frac{C}{2}}$. It is noteworthy that three convolution operations do not share parameters. After that, all three feature maps are reshaped to $\mathbb{R}^{HW\times \frac{C}{2}}$, denoted as $\widetilde{\Theta}_i^t$, $\widetilde{\Phi}_i^t$ and $\widetilde{\Omega}_i^t$ respectively. Three maps then go through procedures expressed as below to generate the dynamic kernel $\widetilde{K}_i^t$

$$\begin{aligned} P_i^t &= Softmax((\widetilde{\Theta}_i^t \times \widetilde{\Phi}_i^t)^T), \\ \widetilde{P}_i^t &= P_i^t \times \widetilde{\Omega}_i^t + \widetilde{\Theta}_i^t, \\ K_i^t &= AvgPooling(Conv'_{1\times1}(\widetilde{P}_i^t)), \end{aligned} \tag{2}$$

where $Softmax$ denotes softmax operation, $Conv'_{1\times1}$ denotes 1×1 convolution which transforms $\mathbb{R}^{HW\times \frac{c}{2}}$ to $\mathbb{R}^{HW\times \frac{c^2}{4}}$, and $AvgPooling$ means average pooling layer with kernel size 3×3. '$\times$' denotes matrix multiplication, and the superscript T means matrix transpose. The affinity weight K_i^t contains location-correlated contextual information by adopting matrix multiplication, and finally it is reshaped to $\widetilde{K}_i^t \in \mathbb{R}^{3\times3\times\frac{c}{2}\times\frac{c}{2}}$ to be used as the dynamic convolution kernel. Instead of stacking several convolution layers, our kernel generation network condenses more contextual information into context-sensitive kernels $\widetilde{K}^t$. Then the feature map F^t is convolved with generated context-sensitive filters with dilation rate $d \in \{1, 3, 5\}$ to obtain scale-specific feature representation C_i^t, which can be defined as

$$C_i^t = DConv(Conv(F^t); \widetilde{K}^t, d), \tag{3}$$

where $DConv$ denotes the dilation convolution with dilation rate d. While above operations manage to capture inter-frame correlation, dynamic convolution is only performed at a single scale. As a result, it fails to exploit features at multiple scales. To break this constraint, we arrange three Dynamic Filtering Units (DFUs) in parallel and assign different d to them. Each DFU can capture features at a specific scale, providing three feature representations C_1^t, C_2^t and C_3^t with dilation rate 1, 3, 5 respectively, noting that here we add subscripts to distinguish three DFUs' outputs. When it comes to feature integration, most existing methods treat multi-scale features without distinction, which either perform element-wise summation or simple concatenation. Inspired by the visual attention mechanism, we perform integration through an attention-guided weighted summation:

$$\begin{aligned} w_i^t &= Fc_1(AvgPooling(C_i^t)), \\ \widetilde{w}_i^t &= \frac{e^{w_i^t}}{\sum_{j=1}^{3} e^{w_j^t}}, \end{aligned} \tag{4}$$

where $AvgPooling$ shrinks the input feature map from $\mathbb{R}^{h\times w\times c}$ to $\mathbb{R}^{1\times1\times c}$, and Fc_1 condenses it to a scalar value w_i. It is then normalized using softmax function to produce w_i'. The final output can be calculated by

$$O^t = \sum_{j=1}^{3} \widetilde{w}_j^t * C_j^t, \tag{5}$$

Table 1. Quantitative comparisons of S_λ, F_β and *MAE* on five widely-used VSOD datasets. The top four methods above the horizontal line are traditional methods (marked by superscript †), and the following methods are neural network based. The top three results are marked in **boldface**, *red*, *green* fonts respectively. In addition, ** denotes this model is trained on this dataset, thus cannot be used for evaluation.

		DAVIS			SegV2			ViSal			VOS			DAVSOD		
Method	**Years**	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$
MSTM†	CVPR'16	0.583	0.429	0.165	0.643	0.526	0.114	0.749	0.673	0.095	0.657	0.567	0.144	0.532	0.344	0.211
STBP†	TIP'16	0.677	0.544	0.096	0.735	0.640	0.061	0.629	0.622	0.163	0.576	0.526	0.163	0.568	0.410	0.160
SGSP†	TCSVT'16	0.692	0.655	0.138	0.681	0.673	0.124	0.706	0.677	0.165	0.557	0.426	0.236	0.577	0.426	0.236
SCOM†	TIP'18	0.832	0.783	0.048	0.815	0.764	0.030	0.762	0.831	0.122	0.712	0.690	0.162	0.599	0.464	0.220
SCNN	TCSVT'18	0.783	0.714	0.064	**	**	**	0.847	0.831	0.071	0.704	0.609	0.109	0.674	0.532	0.128
DLVS	TIP'18	0.794	0.708	0.061	**	**	**	0.881	0.852	0.048	0.760	0.675	0.099	0.657	0.521	0.129
FGRN	CVPR'18	0.838	0.783	0.043	**	**	**	0.861	0.848	0.045	0.715	0.669	0.097	0.693	0.573	0.098
PDB	CVPR'18	0.882	0.855	0.028	0.864	0.800	0.024	0.907	0.888	0.032	0.818	0.742	0.078	0.698	0.572	0.116
RCR	ICCV'19	0.886	0.848	0.027	0.842	0.781	0.035	0.922	0.906	0.026	**0.873**	**0.833**	**0.051**	0.741	0.653	0.087
SSAV	CVPR'19	0.893	0.861	0.028	0.851	0.801	0.023	0.943	0.939	0.020	0.819	0.742	0.073	0.724	0.603	0.092
MGA	ICCV'19	0.912	0.892	0.022	0.865	0.821	0.030	0.941	0.940	0.016	0.792	0.735	0.075	**0.751**	0.656	0.081
PCSA	AAAI'20	0.902	0.880	0.022	0.865	0.810	0.025	0.946	0.940	0.017	0.827	0.747	0.065	0.741	0.655	0.086
DCFNet	-	**0.914**	**0.900**	**0.016**	**0.883**	**0.839**	**0.015**	**0.952**	**0.953**	**0.010**	0.846	0.791	0.060	0.741	**0.660**	**0.074**

where O^t represents DCFM's final output.

3.3. Bidirectional Dynamic Fusion Strategy

Given that spatial and temporal dimensions are deeply connected in video scenes, exploring and modelling these cross-domain correlations is a critical and hard task for VSOD. Previous methods like 3D convolution ambiguously utilize high-dimensional convolution kernels, incorporating spatial and temporal features indiscriminately. This inevitably introduces negative features into saliency inference. To address this problem, we develop an insight into details of temporal and spatial characteristics by explicitly constructing spatiotemporal connections between features of input frames, and propose a bidirectional dynamic fusion strategy for better spatiotemporal feature fusion.

Given both spatial and temporal feature inputs, we strive to explicitly model the cross-domain interaction at multiple stages, and jointly fuse them in a progressive way. Instead of ambiguously fusing spatiotemporal features, we rearrange feature interactions in a progressive refinement architecture. We divide the whole process into four stages, using Fusion modules (FMs) in Figure 4 for feature fusion.

For stage i, its inputs consist of both spatial and temporal parts. Spatial inputs come from four enhanced spatial feature maps $\widetilde{B}_1^t$, $\widetilde{B}_2^t$, $\widetilde{B}_3^t$ and $\widetilde{B}_4^t$, as mentioned in the previous section. Temporal inputs D_i^{t-1} and D_i^{t+1} are aggregated inside FMs. The outputs of all previous stages' FMs together with spatial and temporal inputs of the current stage are fused hierarchically, generating D_i^t at the end of the ith stage, and D_4^t is used as the final prediction result. It is worth noting that $\widetilde{B}_4^t$ is an alias of D_0^t.

First, to comprehensively utilize the temporal coherence in both forward and backward directions, we arrange DCFMs in a bidirectional structure. In each FM, two DCFMs in parallel process the input of the previous and next frames separately, generating outputs O_+^t and O_-^t. It is worth noting that two DCFMs do not share parameters, *i.e.*, each direction's DCFM maintains its own set of parameters, depicted in Figure 4 using two separate blocks. The output of two DCFMs can be combined using

$$O_{full}^t = Conv(Cat(O_+^t, O_-^t)), \tag{6}$$

where $Conv$ denotes the convolution layer with kernel size 3, and O_{full}^t stands for each pair of DCFMs' output.

Second, to fuse D_i^t and temporal features D_i^{t-1}, D_i^{t+1} inside FM, a common approach is to combine the feature maps in a relatively undifferentiated manner (*e.g.*, direct point-wise summation). However, it may yield severe fusion ambiguity in the scenarios of the feature redundancy and noises among different domains. Instead of fusing spatiotemporal features indiscriminately, a finer approach is supposed to encourage the network to automatically balance cross-domain feature fusion proportion. Therefore, we formulate FM as

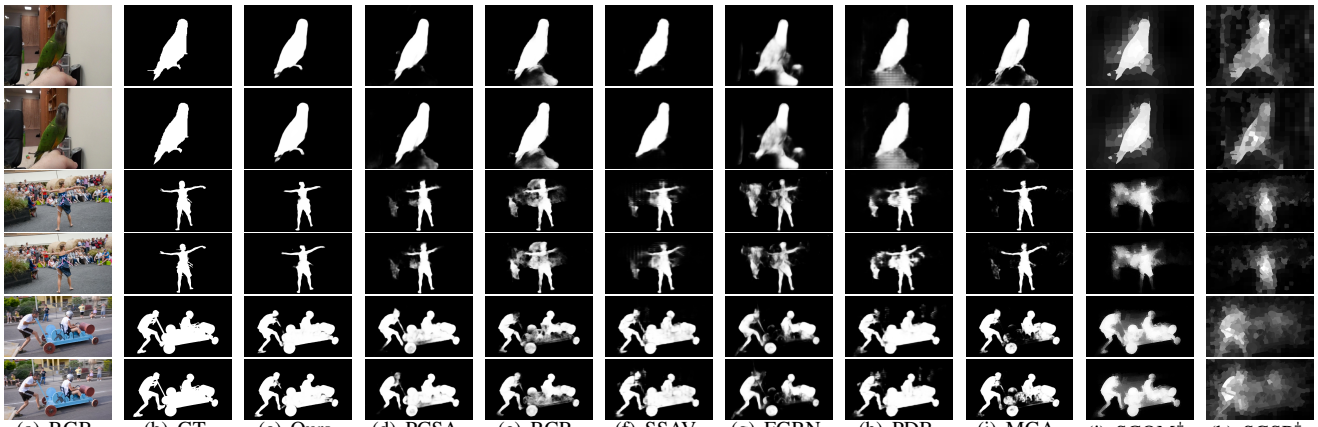
$$\begin{aligned} \widetilde{D}_i^t &= O_{full}^t, \\ F_i^t &= \alpha_i * \widetilde{D}_i^t + (1-\alpha_i) * D_i^t, \end{aligned} \tag{7}$$

The scalar value α_i is generated by a light-weight gate weight generation network (GWG) in Figure 4. GWG aims to generate a context-related scalar value α using

$$\alpha_i = Fc_2(AvgPooling(\widetilde{D}_i^t)), \tag{8}$$

where Fc_2 has the same meaning as Fc_1. Average Pooling shrinks the feature map down to a vector, then full connection layer reduces it to the scalar value α_i. GWG dynamically determines the fusing proportion, hence it enables model to be attentive on the desired features.

After F_i^t is generated, it absorbs both high-resolution spatial features $\widetilde{B}_{i-1}^t$ and all outputs of the pervious



(a) RGB (b) GT (c) Ours (d) PCSA (e) RCR (f) SSAV (g) FGRN (h) PDB (i) MGA (j) SCOM† (k) SGSP†

Figure 6. Qualitative comparisons of DCFNet with eight previous VSOD methods.

stages. We then merge them using element-wise addition to produce D_i^t. The feature maps of lower resolution are convolved and upsampled using bilinear interpolation. Through this bidirectional dynamic fusion strategy, progressive interaction of spatial and temporal domains allows more effective fusion.

4. Experiments

4.1. Experimental Setup

Datasets and Evaluation Metrics To evaluate the performance of our method, we conduct experiments on five widely-used VSOD datasets, *i.e.*, DAVIS [39], VOS [29], SegV2 [26], ViSal [49] and DAVSOD [12]. For fair comparisons, we split the above datasets as the same splitting way in [12, 13, 28], and evaluate our proposed method on the test datasets of all the five datasets. We adopt three widely-used metrics to evaluate our model performance, *i.e.*, max F-measure (F_β) [1], mean absolute error (MAE) [2] and structure-measure (S_λ) [11].

4.2. Implementation Details

Our method is implemented on the PyTorch toolbox [38] with a Nvidia GTX 2080Ti GPU. During training, we adopt the same loss with [41], which includes the binary cross entropy loss L_{bce} [8], IOU Loss L_{IoU} [58] and SSIM Loss L_{ssim} [51] to train our DCFNet. And the final loss L can be expressed as $L = L_{bce} + L_{IoU} + L_{ssim}$.

First, we initialize our backbone with a ResNet-101 [16] pre-trained on ImageNet [9]. Then, we remove the temporal modules, *i.e.*, DCFMs and GWGs in DCFNet and pre-train them on the same datasets used by PCSA [13] (include the training dataset of VOS [29]). We adopt Adam [23] as the optimizer with the initial learning rate of 1e-5 and batch sizes is set to 10 in the pre-training phase. The learning rate decays 0.1 times every 30 epochs. We resize the input frames to 448×448 and adopt the same data augmentation strategies in [57]. Next, we fine-tune the whole DCFNet with the video datasets utilized in pre-training phase. The number of input frames is set to 4 due to the limitation of GPU memory. The learning rate of backbone and temporal modules are set to 1e-6 and 1e-5, respectively. The proposed method takes approximately 0.036 seconds to generate a saliency map for a single frame, which reaches a real-time speed of 28 fps.

Table 2. Quantitative comparisons with different design options of DCFM. '1x', '3x' and '5x' denote dynamic filtering module with dilation rate 1, 3, 5 respectively. GA represents global attention and 'st' stands for 'stack'. DCFM-original refers to our proposed DCFNet. Top-1 results are marked in **boldface**.

Methods	Design Options					DAVIS		ViSal	
	1x	3x	5x	GA	st	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$
DCFNet-d1	√			√		0.893	0.017	0.948	0.012
DCFNet-d3		√		√		0.885	0.018	0.947	0.011
DCFNet-d5			√	√		0.888	0.017	0.950	0.012
DCFNet-noGA	√	√	√			0.895	**0.016**	0.951	0.011
DCFNet-st	√	√	√	√	√	0.889	0.017	0.949	0.011
Ours	√	√	√	√		**0.900**	**0.016**	**0.952**	**0.010**

4.3. Comparisons with State-of-the-arts

As shown in Table 1, we compare our methods with 12 video salient object detection methods including 4 traditional VSOD methods (remarked with †): MSTM† [46], STBP† [56], SGS P† [31], SCOM† [6] and 8 state-of-the-art CNNs based VSOD methods: SCNN [45], DLVS [50], FGRN [27], PDB [44], RCR [57], SSAV [12], MGA [28], PCSA [13]. To guarantee fair comparisons, we utilize the widely-used evaluation toolbox provided by [12].

Quantitative Evaluation. Table 1 shows the quantitative comparisons in terms of three metrics including F_β, MAE, S_λ on five widely-used VSOD datasets. It can be seen that our method significantly outperforms both traditional and CNNs based approaches across four datasets, and wins the second place on the VOS dataset, and the second place in

Table 3. Quantitative comparisons of DCFNet with different spatiotemporal fusion strategy. 'uni' denotes model with unidirectional DCFMs, 'S' denotes DCFNet with the absence of temporal information input, and 'T' means DCFNet without spatial input except D_0^t. Top-1 results are marked in **boldface**.

Methods	DAVIS			SegV2		
	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_\lambda \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
DCFNet-uni	0.908	0.895	0.017	0.873	0.825	0.017
DCFNet-S	0.903	0.887	0.018	0.882	0.831	0.018
DCFNet-T	0.889	0.867	0.022	0.852	0.798	0.026
DCFNet-add	0.900	0.881	0.018	0.868	0.811	0.019
DCFNet-original	**0.914**	**0.900**	**0.016**	**0.883**	**0.839**	**0.015**

S_λ on DAVSOD. In terms of the VOS dataset, RCR (the method for the best performance) utilizes additional pseudo label generation network to significantly enlarge the amount of VOS training dataset. By contrast, our method does not rely on any assistance of extra dataset augmentation strategy. In our opinion, since all the existing VSOD approaches are not trained with any subsets of ViSal, performance on it can be used to reflect the generalization of VSOD models. As can be seen in Table 1, on the ViSal dataset, our method outperforms the second-best model MGA [28] by approximately 37.5% in terms of *MAE*.

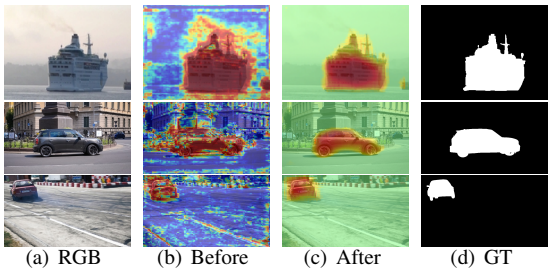
Qualitative Evaluation. Figure 6 shows visual comparisons to demonstrate the superiority of our proposed approach in an intuitive way. The visual results of the *Bird* sequence (Row 1-2) demonstrate that our method can segment salient objects with more accurate details. Compared with other methods negatively affected by distraction of moving crowd in the background, our method discriminates salient objects from clutter background and achieves more accurate predictions in the *dance-twirl* sequence (Row 3-4). Moreover, when facing objects that have both dynamic changes and subtle difference, our method is capable of generating accurate saliency predictions while maintaining fine details, as shown in the *soapbox* sequence (Row 5-6). More visualized results have been provided in supplementary materials.

4.4. Ablation Studies

In this section, we conduct extensive experiments to illustrate the effectiveness of each component of our method.

Effect of DCFM. To prove the effect of the dynamic context-sensitive filtering module (DCFM), in Figure 7, we visualize the feature maps of three typical video scenes before and after being processed by the DCFM. It can be seen from Column 'Before' and 'After' that the salient regions are emphasized after adopting DCFM, significantly improving detection accuracy.

In order to offer deeper insights into the proposed DCFM, we also perform four quantitative experiments to validate the effectiveness of each key components of DCFM. First, to prove the efficiency of generating location-related affinity weights, *i.e.*, performing matrix multiplication between consecutive frames, we replace each filter generation network with an ordinary multiple-layer convolution network (*i.e.*, two input feature maps are added element-wisely, then go through three convolution layers to produce dynamic kernels). We denote this method as 'DCFNet-st' in Table 2. Obviously, compared with our proposed model, 'DCFNet-st' degrades performance to some extent, demonstrating the advantage of our DCFM module in sufficiently utilizing inter-frame contextual information.



(a) RGB (b) Before (c) After (d) GT

Figure 7. Visualization for feature maps of several samples before and after adopting DCFM. Row 1: slow moving object in a monotonous background; Row 2: fast moving object in a complex background; Row 3: fast moving object from the near to the distant with rapid changes in scale.

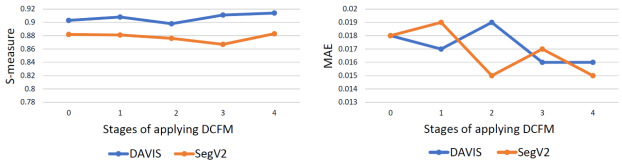


Figure 8. Performance comparison of DCFNet using DCFM at different stages.

Second, to validate the global attention module of DCFM in fusing dynamic filtering results at different scales, we replace the global attention module with a simple element-wise addition (denoted as 'DCFNet-noGA'). Results in Table 2 intuitively verifies the GA module's effect.

In addition, to further prove the effectiveness of the DCFM, we explore different design options for dilation rates. As shown in Table 2, our DCFM with multiple dilation rates (denoted as 'DCFNet-original') achieves significant improvements compared with other three dynamic filtering modules with single dilation rates (1, 3, 5 respectively, denoted as 'DCFNet-d1', 'DCFNet-d3' and 'DCFNet-d5'). Specifically, the *MAE* represents a 17.0%, 9.1% and 17.0% improvement respectively on ViSal.

Furthermore, we run ablation experiments to explore the optimal number of stages applying DCFM. We elaborately attach DCFM step by step to find the optimal setting. In Figure 8, the best S_λ and *MAE* are obtained synchronously when the number of stages containing DCFM is set to 4.

Effect of Bidirectional Dynamic Fusion Strategy. To highlight the importance of the bidirectional dynamic fusion strategy, first we validate the necessary of applying DCFM in a bidirectional way. We can see from Table 3

that compared with applying DCFM backward in a unidirectional way (denoted as 'DCFNet-uni'), the bidirectional DCFM (denoted as 'DCFNet-original') improves the performance by 5.9% and 11.8% on *MAE* towards DAVIS dataset and SegV2 dataset, respectively. Second, we conduct ablations with the absence of the temporal branch or the spatial branch (denoted as 'DCFNet-S' and 'DCFNet-T', respectively). The comparison results in Table 3 verify our claim, even the strategy that combines temporal and spatial information with simple element-wise addition shows impressive performance gains compared with that of 'DCFNet-S' and 'DCFNet-T'. More importantly, the superior performance of applying our bidirectional dynamic fusion strategy (denoted as 'DCFNet-original') powerfully demonstrates the effectiveness of explicit construction of spatiotemporal connections between features.

4.5. Application of DCFM

In order to further verify the effect and generalization of the proposed DCFM, we apply the DCFM in two top-ranking RGB models (CPD [55], GCPA [7]). We attach the DCFM to the first partial decoder of CPD and the aggregation module of GCPA, respectively. This enables the advanced RGB saliency models to capture temporal information and show impressive performance gains towards the VSOD task by simply appending the DCFM without the need to modify the original models.

For the training process, to prevent the dataset bias from affecting the fairness of comparison, we first fine-tune the original CPD and GCPA models with the same training settings adopted in our method's pre-training phase (denoted as '+F' as well as the orange bars in Figure 9). We then extend the models to be applicable for the VSOD task by training following our fine-tune settings (denoted as ' +DCFM' as well as gray bars in Figure 9). The quantitative comparisons in Figure 9 demonstrate that compared with two fine-tuned models ('CPD+F' and 'GCPA+F'), the performance of CPD embedded with DCFMs has been improved by 25.7% and 2.8% on *MAE* and F_β towards DAVIS dataset, respectively. Besides, some visual comparisons of challenging examples are illustrated in Figure 10. After applying DCFM, CPD generates more precise and consistent saliency predictions, especially in the case of complex backgrounds (Row 3 of Figure 10).

4.6. Limitations

While our model achieves robustness in many challenging scenes, some failure cases should not be ignored as they can help reveal opportunities for improvement. In complex multi-objects scenes, *e.g.*, a group of overlapped people, or one object occluded by another, our model may generate incomplete saliency predictions. Multiple objects or occlusions often bring difficulties to VSOD, in part because the single annotation in most VSOD datasets may fall short of

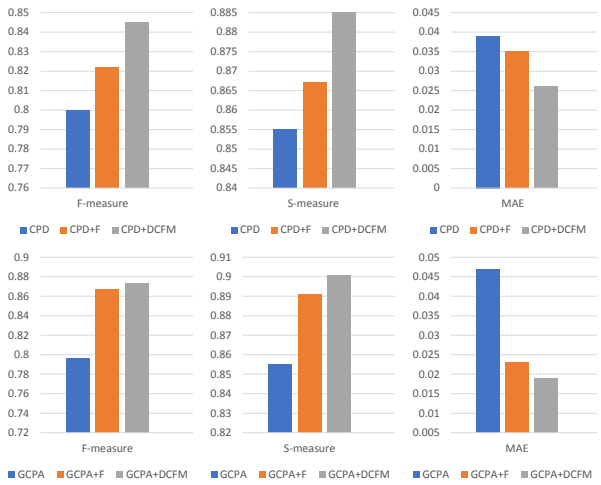
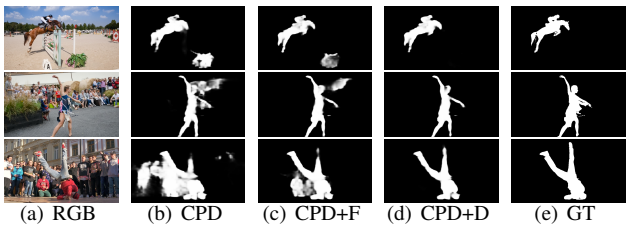


Figure 9. Application of the proposed DCFM in top-ranking RGB saliency models. Results are evaluated on the DAVIS dataset.



(a) RGB (b) CPD (c) CPD+F (d) CPD+D (e) GT

Figure 10. Visualization of application of the proposed DCFM in top-ranking RGB Saliency Models. 'CPD+D' stands for 'CPD+DCFM'.

describing the extreme situations. This might be improved by providing a dataset enriched with more diverse attribute annotations covering different object categories, scene categories, and challenging factors [47].

5. Conclusion

In this paper, we strive to face the challenge of accurate video salient object detection using dynamic filtering mechanism. We propose a dynamic context-sensitive filtering module (DCFM), which generates context-sensitive convolution kernels through estimating the location-related affinity weights, allowing for more adaptability to our constantly-changing world. To model interactions between entangled spatial and temporal information, we further propose a bidirectional dynamic fusion strategy to aggregate spatiotemporal information more sufficiently. Experimental results demonstrate that our proposed method can achieve state-of-the-art performance on most VSOD datasets while ensuring a real-time speed of 28 fps.

Acknowledgments. This work was supported by the Science and Technology Innovation Foundation of Dalian (#2019J12GX034), the National Natural Science Foundation of China (#61976035), and the Fundamental Research Funds for the Central Universities (#DUT20JC42).

References

[1] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 6

[2] Ali Borji, Dicky N. Sihite, and Laurent Itti. Salient object detection: a benchmark. In *ECCV*, pages 414–429, 2012. 6

[3] Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. *CoRR*, abs/1605.09673, 2016. 1, 3

[4] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *CVPR*, pages 3051–3060, 2018. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016. 3

[6] Yuhuan Chen, Wenbin Zou, Yi Tang, Xia Li, Chen Xu, and Nikos Komodakis. Scom: Spatiotemporal constrained optimization for salient object detection. *IEEE Transactions on Image Processing*, 27(7):3345–3357, 2018. 6

[7] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. *arXiv preprint arXiv:2003.00651*, 2020. 2, 8

[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] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. Dynamonet: Dynamic action and motion network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6192–6201, 2019. 3

[11] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4558–4567, 2017. 6

[12] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8554–8564, 2019. 1, 2, 6

[13] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):10869–10876, 2020. 2, 6

[14] Hadi Hadizadeh and Ivan V Bajić. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2013. 1

[15] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3562–3572, 2019. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR)*, pages 770–778, 2016. 3, 6

[17] Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *TIP*, 13(10):1304–1318, 2004. 1

[18] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Debesh Jha, Huazhu Fu, and Ling Shao. Progressively normalnized self-attention network for video polyp segmentation. In *MICCAI*, 2021. 1

[19] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *ICCV*, 2021. 1

[20] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, and Li Cheng. Calibrated rgb-d salient object detection. In *CVPR*, pages 9471–9481, June 2021. 2

[21] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In *ECCV*, 2020. 2

[22] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *CVPR*, pages 12341–12351, June 2021. 1

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *BMVC*, volume 1, page 3, 2017. 1

[25] Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE Transactions on Image Processing*, 27(10):5002–5015, 2018. 1

[26] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2199, 2013. 2, 6

[27] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3243–3252, 2018. 2, 6

[28] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7274–7283, 2019. 1, 2, 6, 7

[29] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Transactions on Image Processing*, 27(1):349–364, 2017. 6

[30] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. *PAMI*, 39(8):1605–1616, 2017. 2

[31] Zhi Liu, Junhao Li, Linwei Ye, Guangling Sun, and Liquan Shen. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation.

IEEE transactions on circuits and systems for video technology, 27(12):2527–2542, 2016. 6

[32] Zhi Liu, Xiang Zhang, Shuhua Luo, and Olivier Le Meur. Superpixel-based spatiotemporal saliency detection. *IEEE transactions on circuits and systems for video technology*, 24(9):1522–1540, 2014. 2

[33] Dwarikanath Mahapatra, Syed Omer Gilani, and Mukesh Kumar Saini. Coherency based spatio-temporal saliency detection for video object segmentation. *IEEE Journal of Selected Topics in Signal Processing*, 8(3):454–462, 2014. 1

[34] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018. 1

[35] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[36] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6504–6512, 2017. 1

[37] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. *arXiv preprint arXiv:2007.06227*, 2020. 3

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[39] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 2, 6

[40] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 2

[41] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019. 6

[42] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä. Segmenting salient objects from images and videos. In *European conference on computer vision*, pages 366–379. Springer, 2010. 2

[43] Sucheng Ren, Chu Han, Xin Yang, Guoqiang Han, and Shengfeng He. Tenet: Triple excitation network for video salient object detection. *arXiv preprint arXiv:2007.09943*, 2020. 1

[44] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 6

[45] Yi Tang, Wenbin Zou, Zhi Jin, Yuhuan Chen, Yang Hua, and Xia Li. Weakly supervised salient object detection with spatiotemporal cascade neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7):1973–1984, 2018. 6

[46] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2334–2342, 2016. 6

[47] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *TPAMI*, 2021. 8

[48] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3395–3402, 2015. 1

[49] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015. 6

[50] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017. 6

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

[52] Hefeng Wu, Guanbin Li, and Xiaonan Luo. Weighted attentional blocks for probabilistic object tracking. *Vis. Comput.*, 30(2):229–243, Feb. 2014. 1

[53] Jialin Wu, Dai Li, Yu Yang, Chandrajit Bajaj, and Xiangyang Ji. Dynamic filtering with large sampling field for convnets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 185–200, 2018. 3

[54] Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding. A mutual learning method for salient object detection with intertwined multi-supervision. In *CVPR*, pages 8150–8159, 2019. 2

[55] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2019. 8

[56] Tao Xi, Wei Zhao, Han Wang, and Weisi Lin. Salient object detection with spatiotemporal background priors for video. *IEEE Transactions on Image Processing*, 26(7):3425–3436, 2016. 2, 6

[57] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7284–7293, 2019. 2, 6

[58] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016. 6

[59] Miao Zhang, Wei Ji, Yongri Piao, Jingjing Li, Yu Zhang, Shuang Xu, and Huchuan Lu. LFNet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29:6276–6287, 2020. 2

[60] Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. In *NeurIPS*, pages 896–906, 2019. 2

[61] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2482–2491, 2019. 3

[62] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H. Li, and Ge Li. Pdnet: Prior-model guided depth-enhanced network for salient object detection. In *ICME*, 2019. 2