Local-Global Interaction and Progressive Aggregation for Video Salient Object Detection

Dingyao Min[1], Chao Zhang[2,3,⋆], Yukang Lu[1], Keren Fu[1,⋆], and Qijun Zhao[1]

[1] College of Computer Science, Sichuan University, Chengdu 610065, China
mindingyao@qq.com, luyukang6@163.com, {fkrsuper, qjzhao}@scu.edu.cn
[2] Intelligent Policing Key Laboratory of Sichuan Province, Luzhou 646000, China
[3] Sichuan Police College, Luzhou 646000, China
galoiszhang@gmail.com

Abstract. Video salient object detection (VSOD) aims at locating and segmenting visually distinctive objects in a video sequence. There still exist two problems that are not well handled in VSOD. First, facing unequal and unreliable spatio-temporal information in complex scenes, existing methods only exploit local information from different hierarchies for interaction and neglect the role of global saliency information. Second, they pay little attention to the refinement of the modality-specific features by ignoring fused high-level features. To alleviate the above issues, in this paper, we propose a novel framework named IANet, which contains local-global interaction (LGI) modules and progressive aggregation (PA) modules. LGI locally captures complementary representation to enhance RGB and OF (optical flow) features mutually, and meanwhile globally learns confidence weights of the corresponding saliency branch for elaborate interaction. In addition, PA evolves and aggregates RGB features, OF features and up-sampled features from the higher level, and can refine saliency-related features progressively. The sophisticated designs of interaction and aggregation phases effectively boost the performance. Experimental results on six benchmark datasets demonstrate the superiority of our IANet over nine cutting-edge VSOD models.

Keywords: Video salient object detection · saliency detection · local-global interaction · progressive aggregation.

1 Introduction

Video salient object detection (VSOD) aims to model the mechanism of human visual attention and locate visually distinctive objects in a video sequence, which often serves as an important pre-processing step for many downstream vision tasks, such as video compression [10], video captioning [17], and person re-identification [31].

Compared with image salient object detection (ISOD) which only exploits spatial information in a static image, there exists temporal information in a

⋆ Corresponding authors: Chao Zhang and Keren Fu

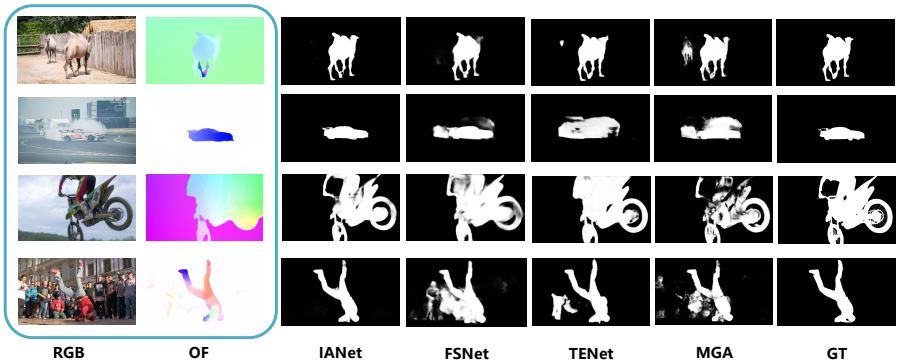

Fig. 1. Examples of challenging cases in video salient object detection. Compared to other state-of-the-art (SOTA) methods, *e.g.*, FSNet[11], TENet[21], and MGA[14], our model shows better detection performance under, *e.g.*, foreground-background visual similarity, environmental interference, and fast object motion. GT means ground truth.

video sequence. Therefore, VSOD needs to capture motion cues. However, more available information also brings more challenges. First, as shown in Fig. 1, there are many complex scenarios in real-world scenarios. In the 1^{st} and 2^{nd} rows, existing methods [14, 21, 11] have difficulty in identifying salient objects from RGB images due to foreground-background visual similarity and environmental interference. The 3^{rd} row shows the salient object is clear in the RGB image but motion cues are misleading in the OF image due to fast motion. The 4^{th} row shows both RGB and OF images hold complex cluttered background. In contrast, our proposed method can accurately segment salient objects by exploiting both local and global information in various scenarios.

Second, the early deep model [29] first applied the fully convolutional network (FCN) to extract spatial and temporal features. After that, optical flow and long-short term memory (LSTM) are used to extract more robust temporal features. Nevertheless, they still adopted some simple strategies to fuse features without any refinement. [3, 11] firstly concatenated RGB and OF features, and then adopted a traditional UNet-like decoder to get the final prediction. [22, 6] extracted temporal features by LSTM and used addition or concatenation operation to integrate features. These operations directly fuse multi-modal and multi-level features without considering level-specific and modality-specific characteristic, and therefore are insufficient to mine discriminative saliency-related information from multi-modal and multi-level features.

In this work, we propose a novel framework, *i.e.*, IANet, which mainly considers addressing the aforementioned problem through more adequate investigation on interaction of cross-modal complementary information as well as aggregation of multi-modal and multi-level features. First, we propose a local-global interaction (LGI) module, which contains a local mutual enhancement (LME) module and a global weight generator (GWG). LGI can fully exploit complementary information between RGB and OF features in both local and global scopes. Specially, in the local scope, LME is designed to cross-enhance RGB and OF features hierarchically. Meanwhile, GWG can learn dynamic channel-wisely weight vec-

tors in the global scope, which represent confidence weights of the corresponding saliency branch. Liu *et al.* [23] also proposed a way to generate the corresponding weights by using single-modal and multi-level features, which ignored the role of multi-modal and multi-level global information. Different from previous methods [14, 23, 3, 11], we fully exploit the correlation of global features.

Second, a progressive aggregation (PA) module is employed to aggregate multi-modal and multi-level features in a progressive refinement way. The fact that is neglected by previous methods [23, 3, 11] is that features from the higher level provide discriminative semantic and saliency-related information, which can refine modality-specific features to focus on saliency-related regions and suppress background distractors. Therefore, we first evolve and aggregate RGB/OF features and up-sampled features from the higher level by element-wise multiplication and maximization, respectively. To fuse the features obtained above, we utilize the channel attention mechanism [9] to adaptively select informative channel features, aiming to control information flows from multi-modal and multi-level features.

Overall, our main contributions are two-fold:

- We propose a *local-global* interaction (LGI) module to excavate complementary RGB and OF features to enhance each other *locally* and automatically learn the confidence weights of the corresponding saliency branch *globally*.
- We design a progressive aggregation (PA) module, which can first evolve and aggregate modality-specific (RGB/OF) features and up-sampled features from the higher level, and then fuse the above features by emphasizing meaningful features along channel dimensions with the attention mechanism. As such, our model can make full use of the higher level features to refine the modality-specific features to further boost performance.

2 Related Work

Traditional video salient object detection methods [26, 30, 1] mainly rely on hand-crafted features and heuristic models. Due to the limitation of hand-crafted features and the low efficiency of heuristic models, these methods cannot handle complex scenarios and are hard to apply in practice.

With the development of deep learning, convolutional neural networks (CNNs) were employed into VSOD. Wang *et al.* [29] concatenated the current frame and output of the previous frame as input of a dynamic FCN to explore intra-frame temporal information. Song *et al.* [22] and Fan *et al.* [6] modeled spatio-temporal information with ConvLSTM. Gu *et al.* [7] proposed a pyramid constrained self-attention module to capture temporal information in video segments directly. Optical flow-based methods [14, 21, 3, 11] used a two-stream fashion to extract RGB and OF features, respectively. Li *et al.* [14] adopted a one-way motion-guided mode, using motion information to enhance spatial information. Ren *et al.* [21] designed a novel Triple Excitation Network, which reinforces the training from three aspects, namely RGB, OF, and online excitations. Chen *et al.* [3] proposed a framework which adaptively captures available information from spatial and temporal cues. Recently, Ji *et al.* [11] achieved bidirectional message transfer by a full-duplex strategy.

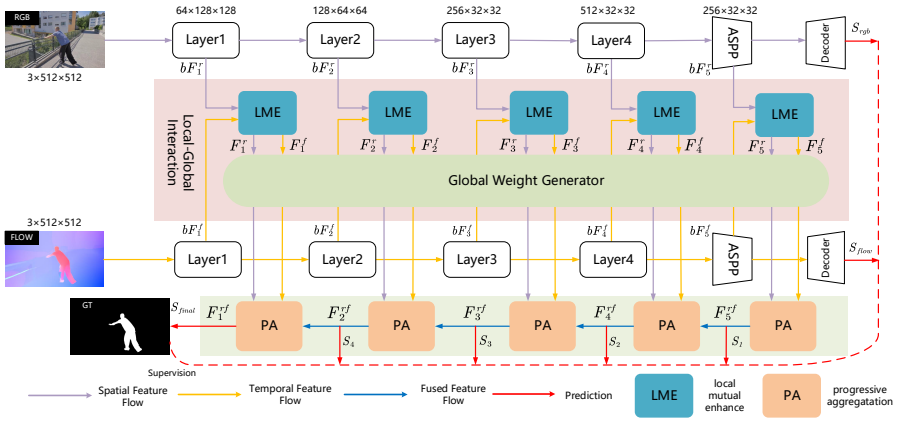

Fig. 2. Overall architecture of the proposed IANet.

3 The Proposed Method

3.1 Overview of Network Architecture

The overall framework of the proposed method is shown in Fig. 2. We propose a local-global interaction and progressive aggregation network (IANet) for video salient object detection. The network consists of symmetric ResNet-34 [8] backbones, local-global interaction (LGI) modules, and progressive aggregation (PA) modules. Given a pair of RGB and optical flow images, the symmetric backbones extract RGB and OF features at five different layers. After the top-down features extraction from the backbones, the multi-level RGB and OF features are fed to LGI that is composed of local mutual enhancement (LME) module and global weight generator (GWG). Finally, RGB features, OF features and up-sampled features from the higher level are independently forwarded to PA module. Details are described below.

3.2 Symmetric RGB and Flow Streams

Initially, we employ two symmetric ResNet-34 backbones as encoders to extract features from RGB and OF images. Let the feature outputs of the last four RGB/OF ResNet hierarchies be denoted as bF_i^m $(m \in \{r, f\}, i = 1, \ldots, 4)$. The symbols 'r' and 'f' mean RGB-related features and OF-related features, respectively. Following [14], we apply an atrous spatial pyramid pooling (ASPP [2]) module after the fourth layer. The feature after ASPP module is input into the decoder to generate S_{rgb}/S_{flow}. The decoder used in this paper contains three convolutional layers to realize channel reduction. Finally, we obtain five-level RGB and OF features, namely bF_i^m $(m \in \{r, f\}, i = 1, \ldots, 5)$.

3.3 Local-Global Interaction (LGI)

LGI is structured as two parts: LME module and GWG module. In realistic scenarios, both RGB and OF information are sometimes unreliable and unequal. Existing bi-modal methods [14, 3, 11] only focus on local interaction strategies and overlook the role of global information. To exploit complementary information between RGB and OF features and simultaneously balance the importance

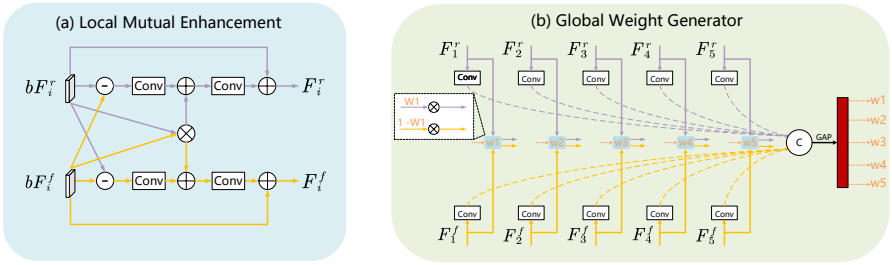

Fig. 3. Two parts of local-global interaction (LGI) module, namely LME and GWG. Here, "C" denotes feature concatenation, and $\ominus$, $\oplus$ and $\otimes$ denote element-wise subtraction, element-wise addition, and multiplication, respectively.

of RGB and OF features to the final saliency map, the obtained RGB and OF features are first locally enriched by LME. Different from previous methods [3, 11], after the local interaction, we further leverage the encoded global saliency information to dynamically scale all five-level features with adaptive weights.

Local mutual enhancement (LME). The structure of LME is shown in Fig. 3(a). For each saliency branch, we firstly extract complementary parts by the mutual subtraction operation, as follows:

$$\begin{aligned} C_i^r &= \mathcal{F}_{conv}(bF_i^f \ominus bF_i^r), \\ C_i^f &= \mathcal{F}_{conv}(bF_i^r \ominus bF_i^f), \end{aligned} \tag{1}$$

where $\mathcal{F}_{conv}$ denotes convolution operation. In addition, the common parts are usually able to locate salient objects accurately. So we extracts the common parts between bF_i^r and bF_i^f by element-wise multiplication, then combines them with C_i^r and C_i^f respectively by element-wise addition. The re-calibrated RGB and OF features can be written as:

$$\begin{aligned} F_i^r &= \mathcal{F}_{conv}(C_i^r + bF_i^r \otimes bF_i^f) + bF_i^r, \\ F_i^f &= \mathcal{F}_{conv}(C_i^f + bF_i^r \otimes bF_i^f) + bF_i^f, \end{aligned} \tag{2}$$

where $\otimes$ is element-wise multiplication with broadcasting.

LME takes a selective enhancement strategy, where redundant information will be suppressed to avoid contamination between features by element-wise multiplication and important features will complement each other by element-wise subtraction. The visual examples are presented in Fig. 4. As shown, OF features help RGB features focus on saliency-related regions, and in turn RGB features help OF features eliminate part of background noise to some extent.

Global weight generator (GWG). The structure of GWG is shown in Fig. 3(b). We first squeeze all feature maps into 64 channels. The compressed features are uniformly upsampled to the same scale. We concatenate RGB and OF features to get F_r and F_f with 320 channels, respectively. Specifically, global average pooling (GAP) is employed to F_r and F_f. Next, we concatenate the two tensors into one tensor, where $[\cdot, \cdot]$ means the concatenation operation:

$$w = [GAP(F_r), GAP(F_f)]. \tag{3}$$

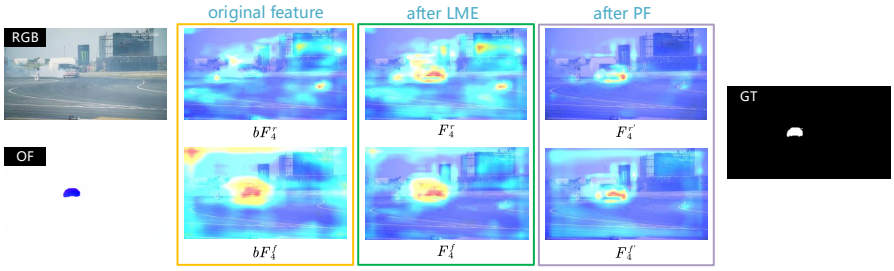

Fig. 4. Visualization of the features before and after LME and pre-fusion.

The obtained tensor $w \in \mathbb{R}^{640\times1\times1}$ contains global information, which reflects the contribution of F_i^r/F_i^f $(i = 1, \ldots, 5)$ to final performance of the model. We further use w to learn adaptive weights for RGB and OF branches respectively. For each layer $i \in \{1, \ldots, 5\}$ of the RGB branch, we can obtain its corresponding adaptive weights by feeding w to two consecutive fully connected layers:

$$w_i = \sigma(FC_i(ReLU(FC(w))), \tag{4}$$

where FC and FC_i represent two fully connected layers (the latter is i-th layer specific), $ReLU$ is non-linear activation, and σ is a sigmoid function that scales weights to interval $(0, 1)$. The derived $w_i \in \mathbb{R}^{c_i\times1\times1}$ has the same number of channels as the i-th layer feature, which can channel-wisely scale F_i^r / F_i^f as:

$$F_i^r \leftarrow w_i \otimes F_i^r, F_i^f \leftarrow (1 - w_i) \otimes F_i^f. \tag{5}$$

The obtained F_i^r and F_i^f are the feature maps dynamically scaled with adaptive weights w_i and $1 - w_i$, respectively. Benefiting from GWG, we can obtain features that are more valuable for final saliency prediction. Specifically, by learning adaptively, the model will leverage more RGB features if the RGB images contribute more to final prediction, and vice versa.

3.4 Progressive Aggregation (PA)

In this section, we describe the structure of PA (see Fig. 5). Given the features $\{F_i^r, F_i^f, F_{i+1}^{rf}\}$ $(i = 1, \ldots, 4)$ with the same size, we propose a PA module to progressively evolve and aggregate the above features. F_i^r and F_i^f denote RGB and OF features of the current layer, which contain unclear semantic information and redundant details. F_{i+1}^{rf} denotes fused features from the higher level by using up-sampling operation and two convolutional layers, which usually provide rich semantic information. Previous methods [23, 3, 11] first fuse RGB and OF features and then combine them with the higher-level features in a UNet-like way. This manner neglects the fact that the higher-level features can refine the modality-specific features. Therefore, before the final fusion phase, we perform pre-fusion (PF) on F_i^r and F_{i+1}^{rf}, F_i^f and F_{i+1}^{rf} by element-wise multiplication and maximization, respectively:

$$\begin{aligned} F_i^{r'} &= [(F_i^r \otimes F_{i+1}^{rf}), Max(F_i^r, F_{i+1}^{rf})], \\ F_i^{f'} &= [(F_i^f \otimes F_{i+1}^{rf}), Max(F_i^f, F_{i+1}^{rf})]. \end{aligned} \tag{6}$$

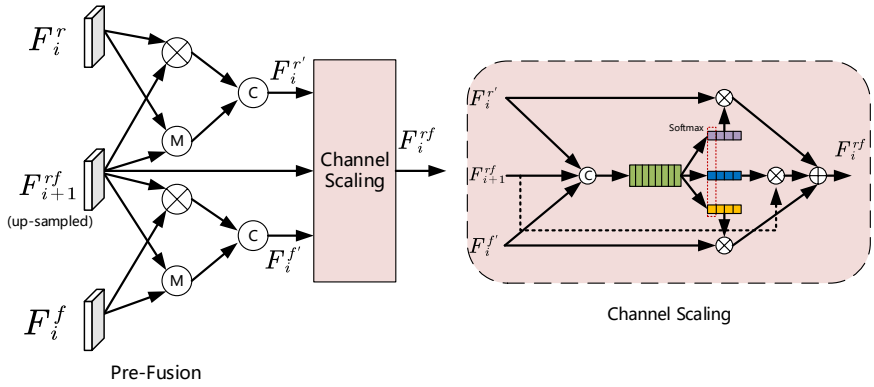

Fig. 5. The scheme of progressive aggregation (PA) module, where "C" denotes feature concatenation, and $\oplus$, $\otimes$ and "M" denote element-wise addition, multiplication, and maximization, respectively.

As shown in Fig. 4, the features after pre-fusion have clearer salient regions, while the cluttered backgrounds are concurrently reduced, simultaneously.

After that, to fuse the features obtained above, we utilize the channel attention mechanism [9] to adaptively select informative channel features, namely channel scaling (CS). We first use channel attention to control the information flows from $\{F_i^{r'}, F_i^{f'}, F_{i+1}^{rf}\}$:

$$w_c = FC(GAP([F_i^{r'}, F_i^{f'}, F_{i+1}^{rf}])). \tag{7}$$

Then, we employ another three FC layers to generate the weight vectors with the same number of channels as $\{F_i^{r'}, F_i^{f'}, F_{i+1}^{rf}\}$, respectively. Further, we utilize channel-wise softmax operation to generate final adaptive weights $\{w_1^c, w_2^c, w_3^c\} \in \mathbb{R}^{c\times 1\times 1}$. At last, we use $\{w_1^c, w_2^c, w_3^c\}$ to adjust $\{F_i^{r'}, F_i^{f'}, F_{i+1}^{rf}\}$ as below:

$$F_i^{rf} = w_1^c \otimes F_i^{r\prime} + w_2^c \otimes F_i^{f\prime} + w_3^c \otimes F_{i+1}^{rf}, \tag{8}$$

where F_i^{rf} is the fused features that will be passed to the next PA module. Notably, since F_5^r and F_5^f are already the highest level features, we directly fuse them by the above channel attention mechanism to generate F_5^{rf}.

3.5 Loss Function

Given four side outputs in the network, the overall loss can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{final} + \mathcal{L}_{rgb} + \mathcal{L}_{flow} + \sum_{i=1}^{4} \lambda_i \mathcal{L}_i, \tag{9}$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 balance the effect of each loss function on the training process. In our experiments, $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are set to $0.4, 0.4, 0.8$ and 0.8, respectively. The side output of each stage corresponds to $\mathcal{L}_i (i = 1, \ldots, 4)$.

Table 1. Quantitative comparisons with SOTA methods on five public VSOD datasets in term of three evaluation metrics: S_α, $F_\beta^{\max}$ and M. The best results are highlighted in **bold**.

	Metric	SCNN [24]	FGR [13]	PDBM [22]	SSAV [6]	MGA [14]	PCSA [7]	TENet [21]	CAG [3]	FSNet [11]	IANet
	Year	2018	2018	2018	2019	2019	2020	2020	2021	2021	2022
DAVIS	$S_\alpha \uparrow$	0.761	0.838	0.882	0.893	0.913	0.868	0.916	0.906	0.920	**0.927**
	$F_\beta^{\max} \uparrow$	0.679	0.783	0.855	0.861	0.893	0.880	0.904	0.898	0.907	**0.918**
	$M \downarrow$	0.077	0.043	0.028	0.028	0.022	0.022	0.019	0.018	0.020	**0.016**
FBMS	$S_\alpha \uparrow$	0.794	0.809	0.851	0.879	0.912	0.868	**0.915**	0.870	0.890	0.892
	$F_\beta^{\max} \uparrow$	0.762	0.767	0.821	0.865	**0.909**	0.837	0.897	0.858	0.888	0.888
	$M \downarrow$	0.095	0.088	0.064	0.040	**0.026**	0.040	0.026	0.039	0.041	0.033
SegV2	$S_\alpha \uparrow$	-	0.664	0.864	0.851	**0.895**	0.866	0.868	0.865	0.870	0.891
	$F_\beta^{\max} \uparrow$	-	0.748	0.808	0.798	0.840	0.811	0.810	0.826	0.806	**0.857**
	$M \downarrow$	-	0.169	0.024	0.023	0.024	0.024	0.027	0.027	0.025	**0.013**
ViSal	$S_\alpha \uparrow$	0.847	0.861	0.907	0.942	0.945	0.946	0.946	-	0.928	**0.947**
	$F_\beta^{\max} \uparrow$	0.831	0.848	0.888	0.938	0.942	0.941	0.948	-	0.913	**0.951**
	$M \downarrow$	0.071	0.045	0.032	0.021	0.016	0.017	0.014	-	0.022	**0.011**
DAVSOD	$S_\alpha \uparrow$	0.680	0.701	-	0.755	0.757	0.741	0.780	0.762	0.773	**0.784**
	$F_\beta^{\max} \uparrow$	0.494	0.589	-	0.659	0.662	0.656	0.664	0.670	0.685	**0.710**
	$M \downarrow$	0.127	0.095	-	0.084	0.079	0.086	0.074	0.072	0.072	**0.067**
VOS	$S_\alpha \uparrow$	0.704	0.715	0.817	0.819	0.807	0.828	-	-	0.703	**0.829**
	$F_\beta^{\max} \uparrow$	0.609	0.669	0.742	0.742	0.743	0.747	-	-	0.659	**0.762**
	$M \downarrow$	0.109	0.097	0.078	0.074	0.069	0.065	-	-	0.108	**0.063**

$\mathcal{L}_{rgb}, \mathcal{L}_{flow}$ and $\mathcal{L}_{final}$ denote loss functions for S_{rgb}, S_{flow} and S_{final}, respectively. For each loss function, it is defined as:

$$\mathcal{L} = \mathcal{L}_{bce}(S, G) + \mathcal{L}_{iou}(S, G), \tag{10}$$

where $\mathcal{L}_{bce}$ and $\mathcal{L}_{iou}$ are binary cross-entropy loss and intersection over union loss [20], respectively. G and S denote the ground truth and predicted saliency map, respectively.

4 Experiments

4.1 Datasets and Metrics

To demonstrate the effectiveness of our method, we conduct experiments on six public VSOD benchmatk datasets, including DAVIS [19], FBMS [16], SegV2 [12], ViSal [28], DAVSOD [6], VOS [15]. For quantitative evaluation, we adopt three metrics, *i.e.* S-measure [4] ($S_\alpha, \alpha = 0.5$), max F-measure [5] ($F_\beta^{\max}, \beta^2 = 0.3$), MAE [18] ($M$). See the related papers for specific definitions.

4.2 Implementation Details

In all experiments, the input RGB and optical flow images are uniformly resized to 512×512. Following [3], optical flow is rendered by the state-of-the-art optical flow prediction method [25]. The optical flow corresponding to the current frame is generated from the current frame and the previous frame. Different from the previous methods [14, 7, 21], we train our model with a simple one-step end-to-end strategy on the training set of DUTS [27], DAVIS, DAVSOD, instead of

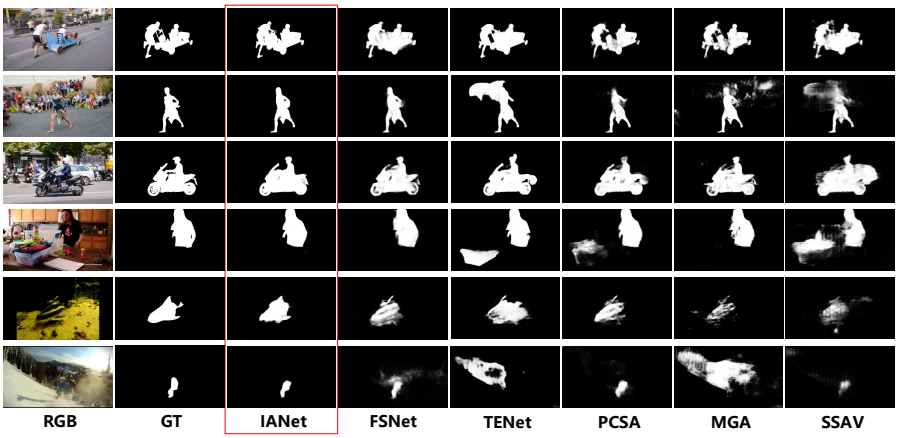

Fig. 6. Visual comparisons of the proposed method with SOTA models.

a multi-stage strategy. Since DUTS is a static image dataset, we fill the input optical flow with zeros. We use Adam optimizer and set batch size 8 in all experiments. The initial learning rate of the backbone and other parts is set to 5e-4 and 5e-3, respectively. The learning rate decays by 0.1 in the 15th epoch. The entire training process takes 20 epochs. The inference time of IANet is 34 FPS on Titan X (Pascal) GPU.

4.3 Comparisons to SOTAs

We compare our proposed IANet with 9 SOTA deep learning-based VSOD methods: SCNN [24], FGR [13], PDBM [22], SSAV [6], MGA [14], PCSA [7], TENet [21], CAG [3] and FSNet [11]. For fair comparison, we use the evaluation code from [6] to get results in our experimets.

Quantitative evaluation. As shown in Tab. 1, our proposed IANet achieves the best result overall. Specifically, our method exceeds other SOTA methods on DAVIS, SegV2, ViSal, DAVSOD, and VOS. The performance on FBMS is relatively worse since we did not use the train set of FBMS for training. Notably, on the most challenging dataset DAVSOD, our method achieves a significant improvement compared with the second-best TENet (0.710 *vs.* 0.664, 0.067 *vs.* 0.074 in terms of $F_\beta^{\max}$ and M respectively). Note we do not employ any post-processing compared with TENet. This shows that our method performs well in various scenarios.

Visual evaluation. Visual comparisons are shown in Fig. 6. Overall, for several complex scenarios, such as complex boundary (1^{st} row), cluttered fore-background (2^{nd}, 4^{th} and 6^{th} rows), multiple moving objects (3^{rd} row) and low contrast (5^{th} row), our method can better locate and segment salient objects with fine-grained details.

4.4 Ablation Study

To verify the impact of our key modules, we conduct experiments on DAVIS and DAVSOD datasets, which include the largest numbers of video sequences and the most complex/diverse scenarios. Quantitative results are shown in Tab. 2.

Local-global interaction (LGI) validation. We compare three variants: w/o LGI, w/o LME, w/o GWG. w/o LGI represents our full model without LGI

Fig. 7. Visual comparisons of different model variants.

Table 2. Ablation studies of different model variants from our IANet. The best results are highlighted in **bold**.

Modules	Models	DAVIS			DAVSOD		
		$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$M \downarrow$
	Full model	**0.927**	**0.918**	**0.015**	**0.784**	**0.710**	**0.067**
LGI	w/o LGI	0.919	0.908	0.016	0.767	0.690	0.073
	w/o LME	0.923	0.912	0.016	0.778	0.703	0.068
	w/o GWG	0.923	0.912	0.015	0.775	0.702	0.068
PA	w/o PA	0.917	0.906	0.017	0.765	0.691	0.074
	w/o PF	0.921	0.911	0.016	0.776	0.705	0.069
	w/o CS	0.921	0.914	0.017	0.778	0.701	0.071
GLI	global-local	0.925	0.916	0.015	0.775	0.694	0.069

module. *w/o* LME and *w/o* GWG correspond to removing LME and GWG. In Fig. 7, the prediction results of three variants contain noise and holes. Compared with the full model, the results indicate that the proposed LGI module is essential for improving the performance as presented in Tab. 2.

Progressive aggregation (PA) validation. We compare three variants: *w/o* PA, *w/o* PF, *w/o* CS. *w/o* PA denotes the fusion of features using the U-Net like approach. *w/o* PF and *w/o* CS denote removal of two sub-modules respectively. Observing Fig. 7 and Tab. 2, our proposed progressive aggregation strategy improves the model performance. The visual results reflect that three baselines generate foreground noise and incomplete salient object. The results show that our proposed progressive aggregation strategy improves the model performance.

About interaction order. Besides, we also compare a reverse interaction pattern for the LGI module, namely global-local interaction (GLI), which conducts GWG before LME. From Tab. 2, the results show a certain degree of degradation. This proves that it is more reasonable to perform the local interaction before the global interaction. GLI may introduce too much global noise by performing the global interaction first.

In summary, the ablation studies demonstrate the effectiveness and advantages of the proposed modules. In addition, the results of the ablation experiments also demonstrate that careful feature interaction and well-designed feature aggregation could effectively improve the performance of VSOD.

5 Conclusion

In this paper, we propose a novel video SOD network equipped with local-global interaction and progressive aggregation. The former adopts a new interaction

strategy to achieve full interaction of multi-modal features, whereas the latter progressively refines saliency-related features. The experimental results demonstrate that both modules boost the performance of the final model. The elaborate designs of the interaction and aggregation phases enable our approach to achieve overall state-of-the-art performance on six benchmark datasets.

Acknowledgements. This work was supported by the NSFC (62176169, 621761 70, 61971005), SCU-Luzhou Municipal Peoples Government Strategic Cooperation Project (2020CDLZ-10), and Intelligent Policing Key Laboratory of Sichuan Province (ZNJW2022KFMS001, ZNJW2022ZZMS001).

References

1. Chen, C., Li, S., Wang, Y., Qin, H., Hao, A.: Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Transactions on Image Processing **26**(7), 3156–3170 (2017)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs **40**(4), 834–848 (2017)
3. Chen, P., Lai, J., Wang, G., Zhou, H.: Confidence-guided adaptive gate and dual differential enhancement for video salient object detection. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
4. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557 (2017)
5. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018)
6. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8554–8564 (2019)
7. Gu, Y., Wang, L., Wang, Z., Liu, Y., Cheng, M.M., Lu, S.P.: Pyramid constrained self-attention network for fast video salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
10. Itti, L.: Automatic foveation for video compression using a neurobiological model of visual attention. IEEE TIP **13**(10), 1304–1318 (2004)
11. Ji, G.P., Fu, K., Wu, Z., Fan, D.P., Shen, J., Shao, L.: Full-duplex strategy for video object segmentation. In: ICCV (2021)
12. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2192–2199 (2013)
13. Li, G., Xie, Y., Wei, T., Wang, K., Lin, L.: Flow guided recurrent neural encoder for video salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3243–3252 (2018)

14. Li, H., Chen, G., Li, G., Yizhou, Y.: Motion guided attention for video salient object detection. In: Proceedings of International Conference on Computer Vision (2019)
15. Li, J., Xia, C., Chen, X.: A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. IEEE Transactions on Image Processing **27**(1), 349–364 (2017)
16. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE transactions on pattern analysis and machine intelligence **36**(6), 1187–1200 (2013)
17. Pan, Y., Yao, T., Li, H., Mei, T.: Video captioning with transferred semantic attributes. In: CVPR. pp. 6504–6512 (2017)
18. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 733–740. IEEE (2012)
19. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
20. Rahman, M.A., Wang, Y.: Optimizing intersection-over-union in deep neural networks for image segmentation. In: ISVC. pp. 234–244 (2016)
21. Ren, S., Han, C., Yang, X., Han, G., He, S.: Tenet: Triple excitation network for video salient object detection. In: European Conference on Computer Vision (ECCV) (2020)
22. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 715–731 (2018)
23. Su, Y., Wang, W., Liu, J., Jing, P., Yang, X.: Ds-net: Dynamic spatiotemporal network for video salient object detection. arXiv preprint arXiv:2012.04886 (2020)
24. Tang, Y., Zou, W., Jin, Z., Chen, Y., Hua, Y., Li, X.: Weakly supervised salient object detection with spatiotemporal cascade neural networks. IEEE Transactions on Circuits and Systems for Video Technology **29**(7), 1973–1984 (2018)
25. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. pp. 402–419. Springer (2020)
26. Tu, W.C., He, S., Yang, Q., Chien, S.Y.: Real-time salient object detection with a minimum spanning tree. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2334–2342 (2016)
27. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 136–145 (2017)
28. Wang, W., Shen, J., Shao, L.: Consistent video saliency using local gradient flow optimization and global refinement. IEEE Transactions on Image Processing **24**(11), 4185–4196 (2015)
29. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing **27**(1), 38–49 (2017)
30. Xi, T., Zhao, W., Wang, H., Lin, W.: Salient object detection with spatiotemporal background priors for video. IEEE Transactions on Image Processing **26**(7), 3425–3436 (2016)
31. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR. pp. 3586–3593 (2013)