Global-Guided Weighted Enhancement for Salient Object Detection

Jizhe Yu[1(✉)], Yu Liu[1], Hongkui Wei[2], Kaiping Xu[1], Yifei Cao[1], and Jiangquan Li[1]

[1] DUT School of Software Technology and DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, China
yujizhe@mail.dlut.edu.cn

[2] State Key Laboratory of Intelligent Manufacturing System Technology, Beijing Institute of Electronic System Engineering, Beijing, China

Abstract. Salient Object Detection (SOD) benefits from the guidance of global context to further enhance performance. However, most works focus on treating the top-layer features through simple compression and nonlinear processing as the global context, which inevitably lacks the integrity of the object. Moreover, directly integrating multi-level features with global context is ineffective for solving semantic dilution. Although the global context is considered to enhance the relationship among salient regions to reduce feature redundancy, equating high-level features with global context often results in suboptimal performance. To address these issues, we redefine the role of global context within the network and propose a new method called Global-Guided Weighted Enhancement Network (GWENet). We first design a Deep Semantic Feature Extractor (DSFE) to enlarge the receptive field of network, laying the foundation for global context extraction. Secondly, we construct a Global Perception Module (GPM) for global context modeling through pixel-level correspondence, which employs a global sliding weighted technique to provide the network with rich semantics and acts on each layer to enhance SOD performance by Global Guidance Flows (GGFs). Lastly, to effectively merge multi-level features with the global context, we introduce a Comprehensive Feature Enhancement Module (CFEM) that integrates all features within the module through 3D convolution, producing more robust feature maps. Extensive experiments on five challenging benchmark datasets demonstrate that GWENet achieves state-of-the-art results.

Keywords: Salient object detection · Global context guidance · Sliding weighted enhancement · Comprehensive feature fusion · 3D convolution

1 Introduction

Salient Object Detection (SOD) mimics the human visual perception system to capture the most attractive parts of an image, and is widely applied in the preprocessing stages of visual tasks such as image editing, AR, VR, and autonomous

M. Wand et al. (Eds.): ICANN 2024, LNCS 15017, pp. 137–152, 2024.
https://doi.org/10.1007/978-3-031-72335-3_10

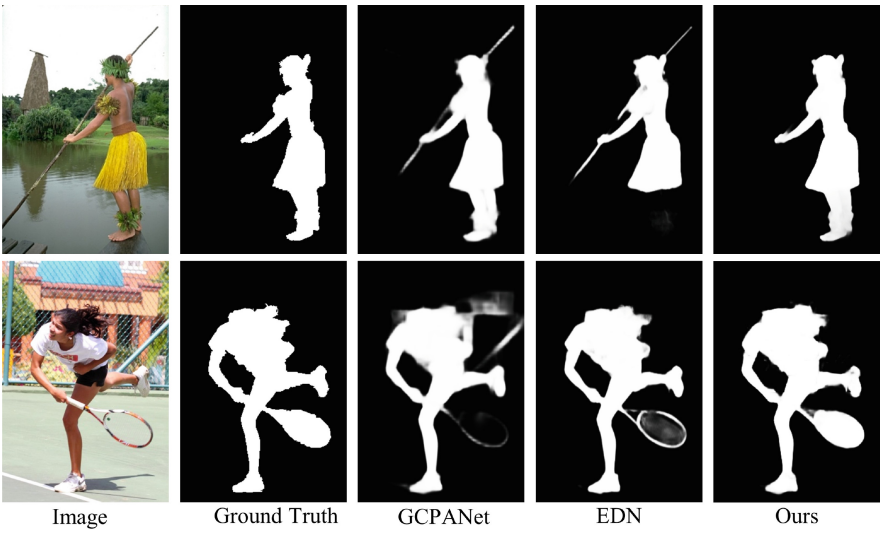

Fig. 1. The prediction results are more accurate than the predictions of other state-of-the-art networks in the field, such as GCPANet [9] and EDN [7].

driving. The popularity of SOD owes to the exceptional feature extraction capabilities of Convolutional Neural Networks (CNNs) in computer vision, marking a significant shift from traditional handcrafted feature extraction [1] to advanced feature representation based on encoder-decoder architectures [2]. Many CNN-based models [3,4] have significantly improved SOD performance through the collaborative work of high-level and low-level features. High-level features have a lower resolution but are rich in semantic information, making them ideal for generating coarse saliency maps. In contrast, low-level features offer larger spatial scales and finer details, crucial for reconstructing object structures. Unfortunately, the dilution of high-level features in the top-down transmission process and the large amount of noise in low-level features have prompted many studies to solve this issue by introducing global features. Some studies [5–7] utilize attention mechanisms to focus the global context on critical areas within high-level features. Other studies [8–11] integrate global context into various stages of the decoder, aiming to enhance the coherence of object prediction. Although these two strategies significantly improve, there remains room for enhancing the prediction integrity for irregular-scale objects in complex scenes, as shown in Fig. 1. Therefore, to enhance the semantics and minimize errors, we identify two main issues that need to be addressed: (1) Most previous works generate global context through simple spatial compression and nonlinear activation of deep features in the encoder, overlooking the fundamental differences between high-level features and global context within the network; (2) While global context guides the fusion of multi-level features by scene understanding of the overall image, simply combining semantic information, detail information, and global context is

suboptimal. This way fails to consider the interference of complex backgrounds and utilize the potential of global guidance in restoring object integrity.

To redefine the role of global context within networks, we propose a novel network called the Global-guided Weighted Enhancement Network (GWENet), comprising three key components and one enhancement technique. To address the first issue, we design a Global Perception Module (GPM) that focuses on learning the correspondence of each pixel to extract global context, which is then applied in the decoding stage through Global Guidance Flows (GGFs). According to EDN [7], further downsampling of the existing CNN backbone network can extract semantics and locate objects more effectively. For this purpose, we design a Deep Semantic Feature Extractor (DSFE) positioned before GPM, which lays the foundation for capturing global context by mining the correlations among feature channels. To enhance learning object integrity from a global view, we construct a Comprehensive Feature Enhancement Module (CFEM) to gradually aggregate multi-level features from the top-down. CFEM comprises two submodules: an Adaptive Feature Interaction Fusion Module (AFIFM) and a Scale Diversity Integration Module (SDIM), which enable the model to enrich feature diversity while extracting valuable complementary information. To address the second issue, the currently common practice involves element-wise addition, multiplication, or concatenation of global context with multi-level features, which is intuitive but not optimal. Therefore, we propose a more stable and efficient enhancement technique, namely a learnable weighted operator. We utilize the global attention guidance generated by GPM to perform sliding weighting on the appearance details and body region maps, which imparts low-level features with semantics and effectively prevents the dilution of high-level features. Furthermore, at the end of the feature aggregation stage in the CFEM module, we employ 3D convolution to capture more accurate and prosperous inter-feature correspondence, thereby enhancing the comprehensive feature fusion. As the visualization results in Fig. 1 show, our GWENet primarily employs global sliding weighted technique, complemented by comprehensive feature fusion, to maximize shape integrity and minimize background interference, as displayed with objects like tennis rackets and bamboo poles.

To sum up, our contributions are as follows:

- We explore the global context from a new perspective to restore object integrity learning, and propose a global sliding weighted enhancement technique to effectively address issues such as dilution and noise instead of previous improved high-level features acting as the global context to locate salient objects, which is expected to provide a new idea for SOD.
- We propose a novel Global-guided Weighted Enhancement Network for accurate salient object detection, which introduces three key components: the Global Perception Module (GPM), the Deep Semantic Feature Extractor (DSFE), and the Comprehensive Feature Enhancement Module (CFEM), where CFEM comprises the Adaptive Feature Interaction Fusion Module (AFIFM) and the Scale Diversity Integration Module (SDIM).

– Compared with the state-of-the-art methods on five challenging datasets, the proposed GWENet achieves the best performance in quantitative and qualitative evaluations.

2 Related Work

2.1 Methods Based on Global Context Guidance

In recent years, global context learning plays a vital role in enhancing the performance of SOD. Wei et al. [10] introduce the Side-out Aggregation Module to enhance the receptive field of the entire network, enabling it to capture more comprehensive information while avoiding the omission of crucial information. Zhao et al. [6] add a global average pooling layer at the end of the encoder to obtain global context, resulting in complete segmentation outcomes. Wu et al. [7] propose an extreme downsampling block to effectively capture global context, thereby achieving accurate salient object localization. To better address the semantic dilution problem of high-level features, Chen et al. [9] design a global context flow module to generate global context information for different decoding stages. Liu et al. [11] utilize the existing semantic segmentation module PPM [4] to capture global guiding information, compensating for the gradual dilution from the top-down.

2.2 Methods Based on Multi-level Feature Fusion

Most multi-level feature fusion methods adopt the principle of feature complementarity [12], that is, combining global structures with local detail information to aggregate multi-scale information. Wu et al. [8] develop a cascaded feedback decoder to fuse multi-level features through multiple iterations, narrowing the feature differences between different layers. Zhou et al. [3] design a two-stream feature decoder for details and structures to capture complementary information. Pang et al. [13] propose a mutual learning aggregation strategy, fusing only adjacent layer features to enhance the representational capability of different resolution features. To obtain richer scale information, Ma et al. [14] introduce atrous convolution in the feature fusion module, aimed at enhancing valuable information and suppressing noise. Zhuge et al. [15] believe that the rich receptive field of convolution kernels can further help the network capture features of different scales, hence a diversity aggregation module is designed to extract feature diversity.

3 Method

3.1 Overview of GCWNet

As shown in Fig. 2, our proposed GWENet employs a U-shaped network architecture based on the encoder-decoder. The encoder utilizes the VGG16 network as

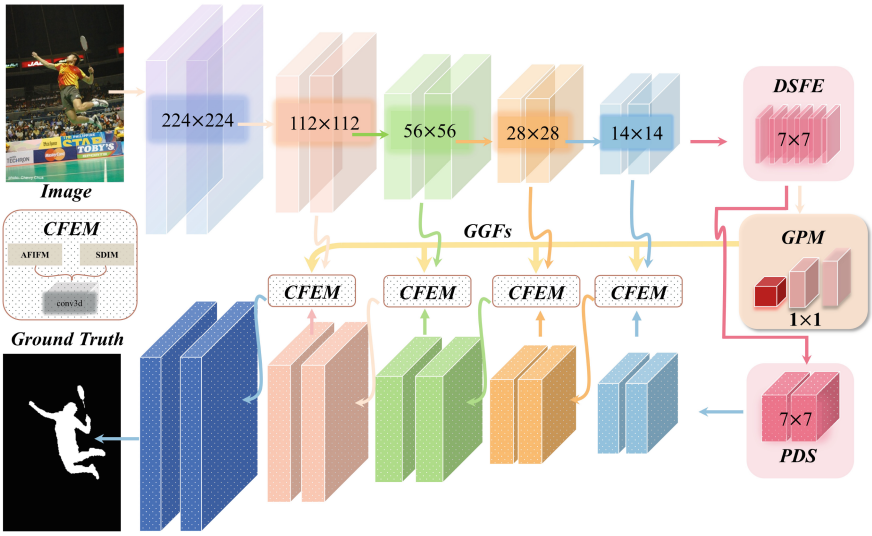

Fig. 2. Illustration of the overall network architecture of GWENet.

its backbone to extract initial features. Following prior studies [4,10], we remove the last pooling and fully connected layers for end-to-end saliency prediction. The network input is images with a resolution of $[224 \times 224]$ pixels. Given that VGG comprises four pooling layers, the subsequent output scales are $[112 \times 112]$, $[56 \times 56]$, $[28 \times 28]$, and $[14 \times 14]$, respectively. For simplicity, we denote these five stages as a set $\mathrm{V} = \{V_{E1}, V_{E2}, V_{E3}, V_{E4}, V_{E5}\}$.

Next, we further downsample the network through the Deep Semantic Feature Extractor (DSFE) to fully extract high-level features rich in semantic and localization information. Thereafter, the high-level features are passed to the Global Perception Module (GPM) to obtain a global view of the image by learning the relationship among pixels and act on the comprehensive feature enhancement module (CFEM) at each decoding stage through the Global Gudance Flow (GFFs). Then, we enhance multi-level feature aggregation with the Adaptive Feature Interaction Fusion Module (AFIFM) and capture multi-scale information using the Scale Diversity Integration Module (SDIM) under the guidance of the global context. Finally, we use 3D convolution to integrate features within the CFEM and output a robust prediction map.

3.2 Deep Semantic Feature Extractor

Wu et al. [7] indicates that further downsampling the network can capture a broader field of view to enhance high-level features, which play a crucial role in scene understanding and object localization [5,10,13]. To this end, we design the Deep Semantic Feature Extractor (DSFE) at the end of the encoder to fully extract high-level semantics. First, we adopt max pooling to downsample the

feature maps to 7×7, obtaining feature denoted as $F_{down} \in R^{N \times C \times H \times W}$($C$, H, W are the channel number, height, and width):

$$F_{down} = Conv^{3\times3}\left(MaxPool\left(V_{E5}\right)\right) \tag{1}$$

where $MaxPool(\cdot)$ means 2 times max pooling downsampling. $Conv^{3\times3}(\cdot)$ represents a 3×3 convolution followed by batch normalization and ReLU layers.

Thereafter, inspired by the self-attention [17], three 1×1 convolution layers are deployed to get three feature maps, namely $F_Q \in R^{N\times C\times H\times W}$, $F_K \in R^{N\times C\times H\times W}$, and $F_V \in R^{N\times \mathrm{C}\times H\times W}$. After reshaping F_Q, F_V and F_K to $R^{N\times HW\times C}$. Then we reshape the correlation strength map among pixels $F_{\mathrm{S}} \in R^{N\times HW\times C}$ to $R^{N\times C\times H\times W}$, and the above process is computed as:

$$F_S = Softmax\left(F_Q F_K^T\right) F_V + F_{down} \tag{2}$$

where T means transpose, and $Softmax(\cdot)$ represents the softmax layer for feature normalization.

Inspired by the Squeeze-and-Excitation [18], three 3×3 convolution layers are deployed to enhance global dependencies among channels. This process is calculated by

$$F_{D6} = Conv_3^{3\times3}\left(Conv_2^{3\times3}\left(Conv_1^{3\times3}\left(F_S\right)\right)\right) + F_S \tag{3}$$

where $Conv_1^{3\times3}(\cdot) \in R^{C\to C/2}$, $Conv_2^{3\times3}(\cdot) \in R^{C/2\to C/2}$, and $Conv_3^{3\times3}(\cdot) \in R^{C/2\to \mathrm{C}}$. $F_{D6} \in R^{N\times C\times H\times W}$ represents the deep high-level semantics that are the final output of DSFE.

Finally, we conduct max pooling to highlight salient regions and average pooling to suppress background on the F_{D6}, which captures 'pure' deep semantics F_{PDS} ("PDS" in Fig. 2) applied for top-down.

3.3 Global Perception Module

In the above, we have discussed that existing SOD methods often generate global features via simple compression and nonlinear activation of high-level features but overlook the fundamental differences between the two. Therefore, we design the GPM to capture the global context based on DSFE, enhancing object integrity learning.

We first model F_{D6} in pixel-level corresponding relationship to determine salient region. A location with a high correlation suggests a higher likelihood of being a salient region. Specifically, We pass F_{D6} through two 1×1 convolutional layers followed by matrix multiplication to obtain the affinity matrix map $A_{D6} \in R^{N\times HW\times HW}$:

$$A_{D6} = Conv_1^{1\times1}\left(F_{D6}\right) \otimes Conv_2^{1\times1}\left(F_{D6}\right)^T \tag{4}$$

where $Conv^{1\times1}(\cdot)$ is the 1×1 convolution, $\otimes$ represents matrix multiplication.

Next, To further get the global correlation of each pixel about spatial location, we apply max pooling on the affinity matrix A_{D6} by row to obtain an affinity matrix $A_{D6}^{\max} \in R^{N\times HW\times 1}$, then normalize and reshape it into affinity matrix $A'_{D6} \in R^{N\times 1\times HW}$:

$$A'_{D6} = Softmax\left(Maxpool\left(A_{D6}\right)\right) \tag{5}$$

Fig. 3. The effectiveness of GPM. (a) Input images. (b) Ground Truth. (c) Results of our method w/o GPM. (d) Results of our method. We can see that without the GPM, the proposed method will suffer from semantic dilution, interference from non-salient objects, and inaccurate localization of objects.

where $MaxPool(\cdot)$ is the max pooling along the row. Similarly, we repeat Eq. (5) for A'_{D6} to obtain the global affinity vector $A''_{D6} \in R^{N\times 1\times 1}$.

Thereafter, We transform the affinity vector A''_{D6} into an affinity weight map $A^{w}_{D6} \in R^{N\times\ 1\times 1\times 1}$ by unsqueeze operation, then we combine the affinity weight map A^{w}_{D6} and deep high-level semantic F_{D6} in an element-wise multiplication manner, thereby getting the correlation feature map $C^{w}_{D6} \in R^{N\times\ \mathrm{C}\times H\times W}$ with a high position relationship.

$$C^{w}_{D6} = A^{w}_{D6} \odot \rho\left(F_{D6}\right) \tag{6}$$

where $\odot$ is element-wise multiplication, $\rho(\cdot)$ represents L2 normalized function.

Finally, We sum C^{w}_{D6} along the width and height dimensions to get $R^{N\times C}$, and then convert to $R^{N\times C\times 1\times 1}$ by two unsqueeze operations. After that, we can get the global context $A^{w}_{G} \in R^{N\times C\times 1\times 1}$ in final feature aggregation.

By applying GPM, we learn a global correspondence for each pixel to learn a global understanding of the entire image. Figure 3 shows some examples. We can see that the proposed model without GPM is easily disturbed by background noise, e.g., the tennis racket and spider legs in Fig. 3(c). In contrast, the proposed GPM effectively prevents semantic dilution and accurately locates salient regions.

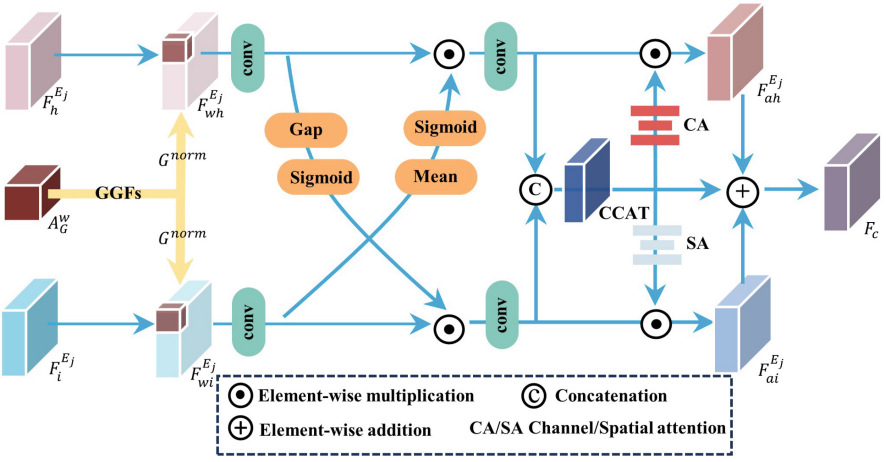

Fig. 4. The architecture of the Adaptive Feature Interactive Fusion Module (AFIFM).

3.4 Comprehensive Feature Enhancement Module

Recent work [8,11] has shown that fusing features from different levels can effectively preserve detail information and capture semantics. For this purpose, we propose the CFEM to better merge multi-level features. We first pass the initial and deep features to the AFIFM, which address the resolution differences between semantics and details under the guidance of global context. Thereafter, features are passed to the SDIM for extracting multi-scale and diverse features. Finally, 3D convolution is used to integrate complementary and multi-scale features.

Adaptive Feature Interactive Fusion Module. As shown in Fig. 4, we construct the AFIFM to address feature misalignment and noise interference in the feature fusion process. Suppose that $F_h^{Ej} \in R^{N\times C\times H\times W}$ represents the deep feature, $F_i^{Ej} \in R^{N\times C\times H\times 1}$ (j=2,3,4,5) represents the initial feature (j denotes the decoding layer). We first apply 1×1 convolutional layer for $A_G^w \in R^{N\times C\times 1\times 1}$ followed by L2 normalization to obtain global attention

guidance map $G^{norm} \in R^{N\times C\times 1\times 1}$. Then, the proposed global sliding weighted technique is used to enhance F_h^{Ej} and F_i^{Ej}, making them fully considering the corresponding relationship between each pixel and global attention map G^{norm} to enhance the recognition of salient regions and suppress irrelevant background. We can get weight features $F_{wh}^{Ej} \in R^{N\times C\times H\times W}$ and $F_{wi}^{Ej} \in R^{N\times C\times H\times W}$ in a residual connection manner, which can be represented as:

$$\begin{cases} F_{wh}^{Ej} = Conv^{G^{\text{norm}}}\left(F_h^{Ej}\right) \odot F_h^{Ej} + F_h^{Ej} \\ F_{wi}^{Ej} = Conv^{G^{\text{norm}}}\left(F_i^{Ej}\right) \odot F_i^{Ej} + F_i^{Ej} \end{cases} \tag{7}$$

where $Conv^{G^{norm}}(\cdot)$ denotes the convolution with G^{norm} as the convolution kernel. It is worth noting that the global sliding weighted technique is not used without the GPM module. With the enhancement of the sliding weighted technique, the model can recognize non-salient objects (billboard in the 3rd column) is illustrated in Fig. 3. Furthermore, unlike [6,7,22], we considered the distinct contributions to different stages, that is, transmitting the global context to each decoding stage through GGFs.

Next, inspired by PFSNet [14], we utilize dynamic weights to fuse semantic and detail features. We first use F_{wh}^{Ej} to generate the channel weights $F_{ch}^{Ej} = \sigma\left(\mathcal{G}\left(Conv_1^{1\times 1}\left(F_{wh}^{Ej}\right)\right)\right)$ for F_{wi}^{Ej}, and F_{wi}^{Ej} to generate the spatial weights $F_{si}^{Ej} = \sigma\left(\mathcal{M}\left(Conv_1^{1\times 1}\left(F_{wh}^{Ej}\right)\right)\right)$ for F_{wh}^{Ej}. Both work on opposite branches, which can help us accumulate more salient features at each level. The specific process can be described as follows:

$$CCAT = Concat\left(\left(F_{wi}^{Ej} \odot F_{ch}^{Ej}\right), \left(F_{wh}^{Ej} \odot F_{si}^{Ej}\right)\right) \tag{8}$$

where $CCAT \in R^{N\times C\times H\times W}$ represnts comprehensive feature map, $\sigma(\cdot)$ represents the Sigmoid activation function. $\mathcal{G}(\cdot)$ is global average pooling operation, and $\mathcal{M}(\cdot)$ means that the channel dimensions are averaged. *Concat* is concatenation operation.

Finally, we use channel and spatial attention [9] to further refine the two branch features, which can be expressed as follows.

$$F_c = CCAT + \left(F_{ai}^{Ej} \odot CA(CCAT)\right) + \left(F_{ah}^{Ej} \odot SA(CCAT)\right) \tag{9}$$

where CA denotes channel attention, and SA is spatial attention. $F_{ai}^{Ej} = F_{wi}^{Ej} \odot F_{ch}^{Ej}$, $F_{ah}^{Ej} = F_{wh}^{Ej} \odot F_{si}^{Ej}$. The final result $F_c \in R^{N\times C\times H\times W}$ is obtained by fusing all the information through the residual connection.

Scale Diversity Integrated Module. As shown in Fig. 5, we construct SDIM for efficient multi-scale learning. Inspired by the UNet architecture [16], we design SDIM with a U-shaped structure, where the resolution of the deepest feature is

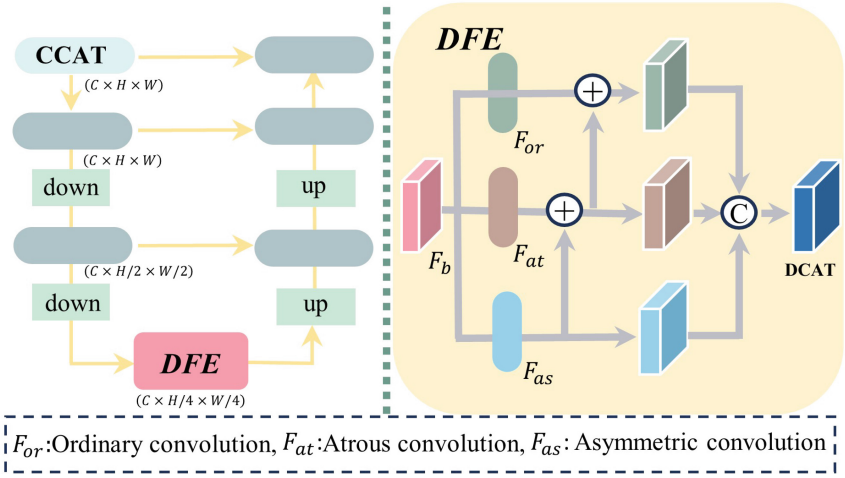

Fig. 5. Illustration of Scale Diversity Integrated Module(SDIM) for enriching feature space.

1/4 of the input size. This is mainly because features of adjacent layers are similar, and integrating features over a wide range can introduce noise [13,14]. Specifically, we first take the $CCAT$ as input of the SDIM to extract and aggregate multi-scale context information $U(CCAT)$. $U(\cdot)$ represents UNet architecture. At the last, we finish with a residual connection:

$$F_u = U(CCAT) + CCAT \tag{10}$$

Due to the challenge of predicting irregular-scale objects in complex scenes, we combine convolutional kernels of different shapes at the bottleneck layer of SDIM to capture features of various object sizes. The DFE is shown in Fig. 5. We employ ordinary convolution (or), atrous convolution [19] (at), and asymmetric convolution [7] (as) to enrich the diversity of the feature space. Furthermore, according to [7,11,21,22], inter-branch interaction helps enhance the multi-scale representation ability. Therefore, we achieve information communication and sharing between the three convolution branches.

$$\begin{cases} F_{\text{orc}} &= Conv_{or}^{3\times3}\left(F_b\right) \\ F_{\text{asc}} &= Conv_{as}^{3\times3}\left(F_b\right) + F_{\text{orc}} \\ F_{\text{atc}} &= Conv_{at}^{3\times3}\left(F_b\right) + F_{\text{asc}} \end{cases} \tag{11}$$

Finally, we concatenate multi-scale features, like

$$DCAT = Conv^{1\times1}\left(Concat\left(F_{\text{orc}}, F_{\text{asc}}, F_{\text{atc}}\right)\right) \tag{12}$$

In this way, SDIM effectively learns irregular-scale features by integrating features of different scales. Notably, SDIM adopts average pooling for down-

sampling and bilinear interpolation for upsampling, thereby achieving efficient transmission.

Final Aggregation. Due to the high efficiency of 3D convolution in processing video sequences [20], we employ 3D convolution to integrate comprehensive features $CCAT$ with multi-scale features $DCAT$, aiming to enhance model performance and reduce redundancy. This process can be represented as

$$F_A = \delta\left(Conv3D^{2\times3\times3}(Concat(CCAT, DCAT))\right) \tag{13}$$

where $Conv3D^{2\times3\times3}$ represents a 3D convolution with a kernel size of $2 \times 3 \times 3$, δ is RELU activation function. $DCAT$ and $CCAT \in R^{N\times\ \mathrm{C}\times1\times H\times W}$. As you know, our network can further enhance the integrity of salient objects.

3.5 Supervision Strategy

Inspired by F3Net [8], we adopt a hybrid loss scheme, utilizing BCE and IoU to train our model, where BCE is used to maintain smooth gradients of the loss function, and IoU is employed to draw more attention to object structures. BCE loss is defined as:

$$\mathcal{L}_{bce} = -\sum_{x=1}^{H}\sum_{y=1}^{W}[G(x,y)\log(P(x,y)) + (1-G(x,y))\log(1-P(x,y))] \tag{14}$$

where $G\,(x,y)$ and $P\,(x,y)$ are the ground truth label and the predicted saliency label at the location (x,y), respectively. H and W are the height and width of the images, respectively. Meanwhile, L_{iou} is defined as:

$$\mathcal{L}_{iou} = 1 - \frac{\sum_{x=1}^{H}\sum_{y=1}^{W}P(x,y)G(x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W}[P(x,y)+G(x,y)-P(x,y)G(x,y)]} \tag{15}$$

4 Experiments

4.1 Experimental Settings

Implementation Details. We use ImageNet to pre-train the backbone network and then use the DUTS-TR to fine-turn the proposed GWENet. Input images are resized to $[352 \times 352]$, $[320 \times 320]$, $[288 \times 288]$, $[256 \times 256]$, and $[224 \times 224]$ for data augmentation. Adam optimizer [7,16] is used to train our network and its hyper parameters are set to default (initial learning rate lr=1e-4, betas=(0.9, 0.999), eps=1e-8, weight decay=0). The warm-up learning rate strategy is also adopted. The batch size is set to 8 (VGG16) and 32 (ResNet50). We run all experiments on the publicly available Pytorch 1.10.0 platform. The network is trained for 50 epochs. Inference for a testing image takes around 30 fps on a single GPU. The code can be available at https://github.com/Gi-gigi/GWENet.

Testing Datasets and Evaluation Criteria. We evaluate all the models on five popular datasets: ECSSD, PASCAL-S, DUT-OMRON, DUTS, HKU-IS. We adopt the Mean Absolute Error (M), the mean E-measure (E_ξ^m), the weighted F-measure (F_β^ω), and the S-measure (S_m) to assess SOD models [15]. We plot Precision-Recall (PR) curves and F-measure curves to show overall performance.

4.2 Comparison with the State-of-the-Arts

We compare the proposed GWENet with twelve recent state-of-the-art models, including CPD [2], EGNet [12], ITSD [3], GateNet [5], MINet [13], F^{3}Net [8], U^{2}Net [16], GCPANet [9], PFSNet [14], PA-KRN [10], ICON [15], EDN [7], and CTD-L [22]. For a fair comparison, the saliency maps are either provided by the authors or obtained by running their released codes under the default parameters.

Table 1. Comparison of GWENet with state-of-the-art SOD methods. The best performance in each column is highlighted in bold.

Summary		ECSSD				PASCAL-S				DUTS-TE				HKU-IS				OMRON			
Method	Params	S_m	E_ξ^m	F_β^ω	M	S_m	E_ξ^m	F_β^ω	M	S_m	E_ξ^m	F_β^ω	M	S_m	E_ξ^m	F_β^ω	M	S_m	E_ξ^m	F_β^ω	M
VGG16-Based Methods																					
CPD	29.23	.91	.938	.895	.04	.845	.882	.796	.072	.867	.902	.8	.043	.904	.94	.879	.033	.818	.845	.715	.057
EGNet	108.07	.919	.936	.892	.041	.848	.877	.788	.077	.878	.898	.797	.044	.91	.938	.875	.035	.836	.853	.728	.057
ITSD	17.08	.914	.937	.897	.04	.856	.891	.811	.068	.877	.905	.814	.042	.906	.938	.881	.035	.829	.853	.734	.063
GateNet	100.02	.917	.932	.886	.041	.857	.886	.797	.068	.87	.893	.786	.045	.91	.934	.872	.036	.821	.84	.703	.061
MINet	47.56	.919	.943	.905	.036	.854	.893	.808	.064	.875	.907	.813	.039	.912	.944	.889	.031	.822	.846	.718	.057
ICON	19.17	.919	.946	.905	.036	.861	.902	.82	.064	.878	.915	.822	.043	.915	.95	.895	.032	.833	**.865**	.743	.065
EDN	21.83	**.928**	**.951**	**.915**	.034	.86	.896	.815	.066	.883	.912	.822	.041	.921	.95	.9	.029	.838	.863	.746	.057
Ours-V	18.47	**.928**	.950	**.915**	**.031**	**.87**	**.905**	**.833**	**.059**	**.895**	**.926**	**.848**	**.035**	**.922**	**.951**	**.906**	**.027**	**.84**	**.865**	**.756**	**.056**
ResNet50-Based Methods																					
CPD	47.85	.918	.942	.898	.037	.848	.882	.794	.071	.869	.898	.795	.043	.905	.938	.875	.034	.825	.847	.719	.056
EGNet	111.69	.925	.943	.903	.037	.852	.881	.795	.074	.887	.907	.815	.039	.918	.944	.887	.031	.841	.857	.738	.053
ITSD	26.47	.925	.947	.91	.034	.859	.894	.812	.066	.885	.913	.823	.041	.917	.947	.894	.031	.84	.865	.75	.061
GateNet	128.63	.92	.936	.894	.04	.858	.886	.797	.067	.885	.906	.809	.04	.915	.937	.88	.033	.838	.855	.729	.055
MINet	126.38	.925	.95	.911	.033	.856	.896	.809	.064	.884	.917	.825	.037	.919	.952	.897	.029	.833	.86	.738	.056
F^{3}Net	25.54	.924	.948	.912	.033	.861	.898	.816	061	.888	.92	.835	.035	.917	.952	.9	.028	.838	.864	.747	.053
U^{2}Net	46.21	.928	.924	.91	.033	.844	.841	.797	.074	.861	.886	.804	.044	.916	.948	.89	.031	.847	.871	.757	.054
GCPANet	67.06	.927	.920	.903	.036	.858	.846	.808	.063	.89	.89	.82	.038	.92	.949	.889	.031	.839	.860	.734	.057
PFSNet	31.18	.929	.927	**.919**	.031	.853	.855	.818	.062	.892	.902	.842	.036	.924	.956	.909	.026	.842	.874	.756	.055
PAKRN	141.06	.927	.924	.918	.032	.851	.857	.816	066	.9	.916	.86	**.033**	.923	.955	.909	.027	**.853**	**.885**	.779	**.05**
ICON	33.09	.929	**.954**	.918	.032	.861	.899	.818	.064	.888	.924	.836	.037	.92	.953	.902	.029	.844	.876	.761	.057
EDN	42.85	.927	.951	.918	.033	.865	.902	.827	.062	.892	.925	.844	.035	.924	.955	.908	.027	.849	.877	.77	**.05**
CTD-L	26.48	.921	.925	.913	.032	.868	.870	.825	.059	.891	.914	.849	.034	.922	.954	.905	**.026**	.845	.878	.776	.052
Ours-R	27.73	**.931**	**.954**	**.919**	**.03**	**.872**	**.907**	**.835**	**.058**	**.901**	**.931**	**.863**	**.033**	**.926**	**.957**	**.91**	**.026**	.851	.882	**.781**	**.05**

Quantitative Comparison. Table 1 reports the quantitative results on five benchmark datasets using the backbone networks VGG16 and ResNet50 in terms of S-measure, E-measure, weighted F-measure, and MAE. Obviously, the

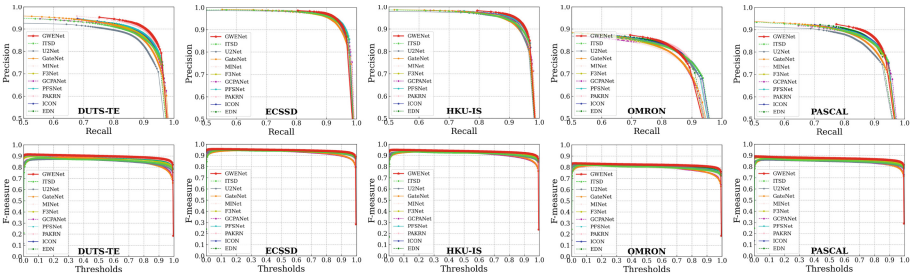

Fig. 6. Illustration of PR curves (1st row), F-measure curves (2nd row) on five datasets.

proposed GWENet outperforms other methods in both performance and efficiency. Although GWENet reach a competitive or comparable level on individual metrics, its overall performance emerged as the leader. In terms of the MAE metric, our GWENet achieves the lowest scores across all datasets, which demonstrates that the GPM assists in enhancing the model to locate salient objects. Compared with EDN [7] and GCPANet [9], the global sliding weighted technique can enhance the capability to prevent semantic dilution and suppress background interference. PR and F-Measure curves are shown in Fig. 6, respectively. GWENet performs best overall on PR and F-Measure curves, which further demonstrates the effectiveness of the proposed method based on the global context guidance.

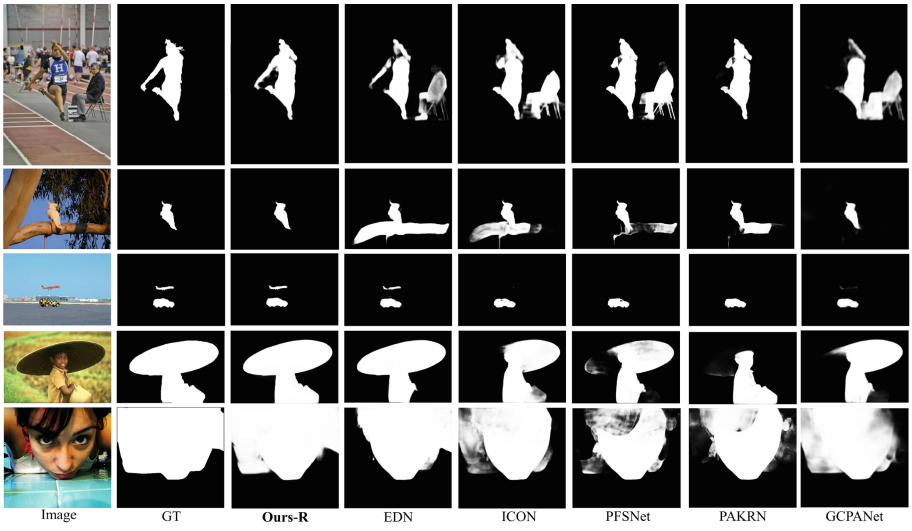

Fig. 7. Qualitative comparison of our method with five SOTA methods: GCPANet [9], PA-KRN [10], PFSNet [14], ICON [15], and EDN [7]. The proposed GWENet produces more accurate localization and complete objects with fewer background noises for various complex scenes.

Qualitative Evaluation. The qualitative comparison is shown in Fig. 7. Our GWENet generates more accurate and complete saliency maps than other methods for diverse challenging cases, e.g., Regular-scale objects in cluttered backgrounds (1st row), Small-scale objects and multi-object scenes (2nd and 3rd row), Large-scale objects (4th and 5th row). Besides, our model can highlight salient regions more clearly and suppress background noise. All visualization results demonstrate the accuracy and robustness of the proposed method.

4.3 Ablation Study

In this part, we conduct the ablation study to verify the effectiveness of the key components and technique proposed in our model. All studies are conducted on the ECSSD and PASCAL-S datasets, and VGG-16 is adopted as the backbone.

Table 2. Ablation study with different components combinations on ECSSD and PASCAL-S dataset.

Summary		ECSSD				PASCAL-S			
ID	Methods	S_m	E_ξ^m	F_β^ω	M	S_m	E_ξ^m	F_β^ω	M
1	Ours	**.928**	**.95**	**.915**	**.031**	**.87**	**.905**	**.833**	**.059**
2	w/o GPM	.921	.943	.903	.036	.865	.899	.819	.065
3	w/o AFIFM	.926	.944	.909	.035	.864	.897	.82	.064
4	w/o SDIM	.924	.948	.912	.034	.862	.902	.822	.063
5	CFEM-add	.925	.943	.907	.034	.864	.896	.82	.063
6	CFEM-multi	.926	.945	.908	.034	.869	.903	.828	.061
7	CFEM-2D	.925	.944	.909	.033	.868	.901	.826	.062

Effectiveness of Different Components. Table 2 shows that removing GPM (ID:2) significantly declines network performance on two datasets, with a 16% and 10% decrease in MAE, respectively. The effectiveness of GPM is exhaustively demonstrated in Fig. 3. We observe further performance degradation after separately removing AFIFM (ID:3) and SDIM (ID:4). Although the decline is not as significant as in model (ID:2), their indispensability is evident, highlighting the contribution of feature fusion and multi-scale information to performance improvement. As anticipated, integrating all components into the proposed model (ID:1) achieves the best performance.

Effectiveness of 3D Convolution from CFEM. In this part, we verify the effectiveness of 3D convolution in the CFEM module. Obviously, the performance of conventional aggregation methods experiences a significant decline. Compared to these models (ID:6 and ID:7), the model performance (ID:5) on the PASCAL-S dataset is particularly notable, with M decreased by 6.8%, F_β^ω by 1.6%, E_ξ^m by

0.9%, and S_m by 0.7%. These findings underscore that 3D convolution is more efficient than 2D convolution in learning relative relationships among features.

5 Conclusion

We propose a novel Global-guided Weighted Enhancement Network, GWENet, to detect irregular-scale objects in complex scenes by utilizing a global sliding weighted enhancement technique. To effectively address issues such as semantic dilution, noise interference, and feature misalignment, we construct the Deep Semantic Feature Extractor (DSFE) to generate pure high-level semantics for top-down, the Global Perception Module (GPM) to extract global context for guidance from pixel-level correspondence, and the CFEM, employing 3D convolution to explore feature correlation. The three complement each other and jointly enhance the object integrity. Comprehensive experiments on five benchmarks demonstrate that GWENet achieves the new state-of-the-art for SOD.

Acknowledgements. This work is Supported by the National Natural Science Foundation of China under Grant 61672128.

References

1. Wang, W., et al.: Salient object detection in the deep learning era: an in-depth survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 3239–3259 (2019)
2. Wu, Z., et al.: Cascaded partial decoder for fast and accurate salient object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3902–3911 (2019)
3. Zhou, H., et al.: Interactive two-stream decoder for accurate and fast saliency detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9138–9147 (2020)
4. Liu, J., et al.: A simple pooling-based design for real-time salient object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3912–3921 (2019)
5. Zhao, X., et al.: Suppress and Balance: A Simple Gated Network for Salient Object Detection. ArXiv abs/2007.08074 (2020)
6. Zhao, Z., et al.: complementary trilateral decoder for fast and accurate salient object detection. In: Proceedings of the 29th ACM International Conference on Multimedia (2021)
7. Wu, Y.H., et al.: EDN: salient object detection via extremely-downsampled network. IEEE Trans. Image Proc. **31**, 3125-3136 (2020)
8. Wei, J., et al.: F3Net: Fusion, Feedback and Focus for Salient Object Detection. ArXiv abs/1911.11445 (2019)
9. Chen, Z., et al.: Global Context-Aware Progressive Aggregation Network for Salient Object Detection. ArXiv abs/2003.00651 (2020)
10. Xu, B., et al.: Locate globally, segment locally: a progressive architecture with knowledge review network for salient object detection. In: AAAI Conference on Artificial Intelligence (2021)

11. Liu, J., et al.: PoolNet+: exploring the potential of pooling for salient object detection. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 887–904 (2022)
12. Zhao, J., et al.: EGNet: edge guidance network for salient object detection. In: 2019 IEEE/CVF International Conference on Computer Vision, pp. 8778-8787 (2019)
13. Pang, Y., et al.: Multi-scale interactive network for salient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9410–9419 (2020)
14. Ma, M., et al.: Pyramidal feature shrinking for salient object detection. In: AAAI Conference on Artificial Intelligence (2021)
15. Zhuge, M., et al.: Salient object detection via integrity learning. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 3738–3752 (2021)
16. Qin, X., et al.: U2-Net: going deeper with nested u-structure for salient object detection. Pattern Recogniton. **106** , 107404 (2020)
17. Vaswani, A., et al.: Attention is all you need. Neural Inf. Proc. Syst. (2017)
18. Hu, J., et al.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2017)
19. Chen, L.C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision (2018)
20. Chen, Q., et al.: 3-D convolutional neural networks for RGB-D salient object detection and beyond. IEEE Trans. Neural Netw. Learn. Syst. **35**, 4309–4323 (2022)
21. Tan, Z., Xiaodong, G.: Feature recalibration network for salient object detection. In: International Conference on Artificial Neural Networks (2022). https://doi.org/10.1007/978-3-031-15937-4_6
22. Li, J., et al.: Rethinking lightweight salient object detection via network depth-width tradeoff. IEEE Trans. Image Proc. **32**, 5664–5677 (2023)