Projection, Interaction and Fusion: A Progressive Difference Fusion Network for Salient Object Detection

Xiao Ke[1,2] , **Weijie Zhou**[1,2] , **Yuzhen Niu**[*1,2]

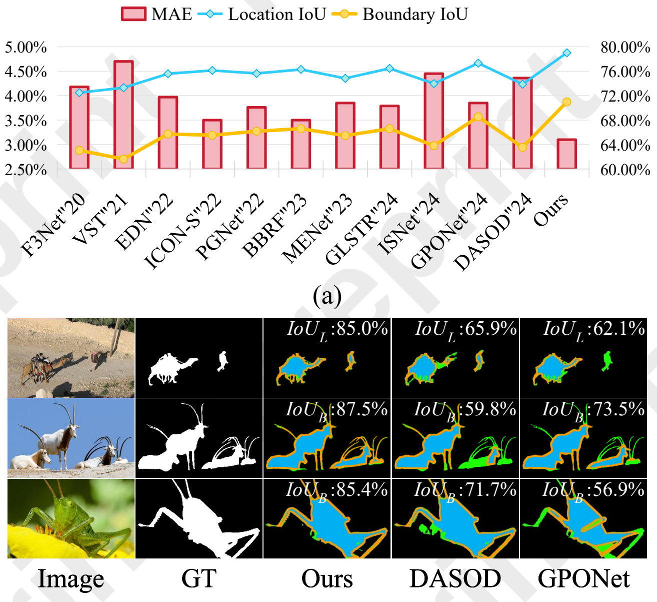
[1]Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing,
College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China
[2]Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350116, China
kex@fzu.edu.cn, lovthero@gmail.com, yuzhenniu@gmail.com

Abstract

In recent years, deep learning-based Salient Object Detection (SOD) methods have made tremendous progress; however, their performance in complex scenarios has reached a bottleneck. In this paper, we propose a novel Progressive Difference Fusion Network (PDFNet) based on fine-grained feature fusion. First, to address the scale variability of salient objects, we introduce a Self-Guided Module (SGM) with dynamic receptive fields. Second, to tackle the shape variability of salient objects, we design a Feature Aggregation Module (FAM) incorporating cross convolutions and a feedback loop. Finally, to alleviate the issue of confusion between global and detail information during multi-scale feature fusion in existing models, we develop a Progressive Difference Fusion Unit (PDFU) to project multi-scale features into fine-grained nodes and enhance them through node interaction based on difference features. Additionally, we propose a Conditional Random Field Based on Patch (CRFbp), which focuses on handling discrete points, further improving the model's performance. Extensive experiments demonstrate that our method achieves state-of-the-art (SOTA) performance on five benchmark datasets. Code is available at: https://github.com/pdfnet2025/PDFNet.git.

1 Introduction

Salient Object Detection (SOD) [Jiang *et al.*, 2013; Zhou *et al.*, 2024], also known as saliency detection, aims to mimic the mechanism of the human visual system by identifying and highlighting the most prominent or important objects and regions in natural images. In recent years, deep learning-based salient object detection methods have made significant progress, but the performance of these methods has gradually reached a saturation point. To delve deeper into this issue, we adopted the CIEDE2000 [Luo *et al.*, 2001] color difference formula to classify five commonly used datasets [Wang *et al.*, 2017; Yang *et al.*, 2013; Yan *et al.*, 2013; Li and Yu, 2015; Li *et al.*, 2014] in the SOD field and selected the top 30% of the most challenging samples to form a subset named TOP-30. TOP-30 can be considered as encapsulating the most challenging cases in the SOD field. Building upon this, we introduced two metrics, Location IoU (IoU_L) and Boundary IoU (IoU_B) [Cheng *et al.*, 2021], to measure a model's localization accuracy and segmentation quality. As shown in Figure 1(a), starting in 2022, the localization and segmentation capabilities of methods in challenging and complex scenarios have started to plateau. As shown in Figure 1(b), existing SOTA methods perform poorly in complex scenarios, which can be attributed to three main reasons: (i) the scale vari-



(a)

(b)

Figure 1: Comparison of our PDFNet with other SOTA methods in complex scenarios. **(a)** MAE, Location IoU (IoU_L), and Boundary IoU (IoU_B) on the challenging examples dataset TOP-30. The methods are sorted from left to right according to their publication dates. **(b)** Visualization of IoU_L and IoU_B computed on the TOP-30 dataset for recent SOTA methods. Orange pixels represent the predicted boundary pixels of the salient objects, blue pixels represent the predicted salient object body pixels, and green pixels represent misclassified pixels.

[*] Corresponding author

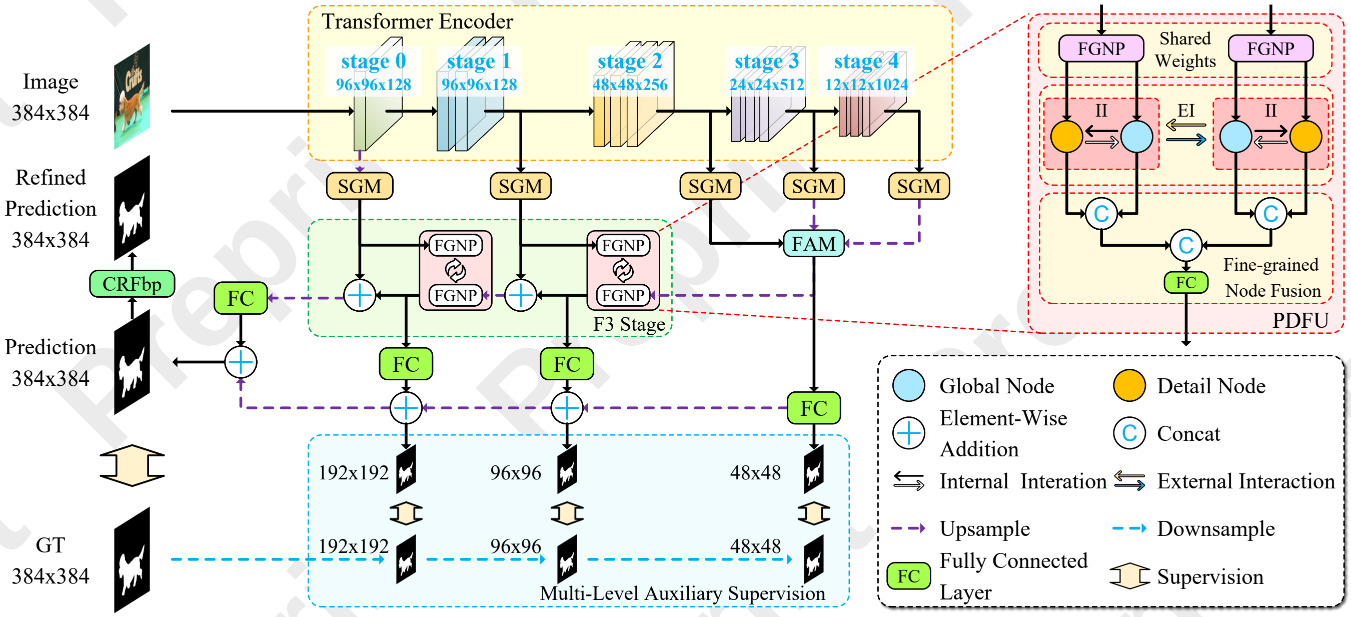


Figure 2: The pipeline of the peoposed method. We choose Swin-B as the backbone network to extract multi-level features. Unlike the traditional U-shaped structure, we use the patch embedding layer output of Swin Transformer as the 0th layer of the backbone network, serving as a transition to avoid information loss caused by excessive upsampling.

ability of salient objects (row 1); (ii) the shape variability of salient objects (row 2); and (iii) the confusion between global and detail information during the multi-scale feature fusion stage (row 3).

To address the aforementioned challenges, we propose a novel architecture—Progressive Difference Fusion Network (PDFNet). First, to overcome the scale variability of salient objects, we design the Self-Guided Module (SGM) with dynamic receptive fields using large kernel convolutions. This allows the model to dynamically adjust the receptive field during the multi-scale feature extraction phase, enhancing the model's scale adaptability. Second, to address the shape variability of salient objects, we use cross convolutions to design the Feature Aggregation Module (FAM) with a feedback loop. This module can preliminarily aggregate features from the deepest three layers of the encoder to generate initial global features, thereby improving the model's shape adaptability. Third, to alleviate the confusion between global and detail information during multi-scale feature fusion, we propose a Fine-Grained Feature Fusion Stage (F3 Stage) composed of Progressive Difference Fusion Units (PDFU). This stage models multi-scale features as fine-grained nodes and enhances the feature representations of nodes at different scales through the interaction of difference features between the nodes.

In summary, our main contributions can be summarized as follows:

1) We propose a novel SOD architecture—Progressive Difference Fusion Network (PDFNet). Our method achieves state-of-the-art performance on five commonly used datasets.
2) We designed the Self-Guided Module (SGM) and Feature aggregation module (FAM), which enhance feature representation and improve the model's adaptability to the size and shape variations of salient objects.
3) We propose a Progressive Difference Fusion Unit (PDFU), which projects multi-scale features into fine-grained nodes to facilitate fine-grained fusion across multi-scale features.

Additionally, we optimize the traditional post-processing technique, the fully connected CRF [Lafferty *et al.*, 2001], and propose the Conditional Random Field based on Patch (CRFbp) that focuses on handling discrete points, further improving the accuracy of predictions.

2 Related Works

2.1 Vision Transformer

When the Transformer [Vaswani, 2017] was first introduced, it was primarily used for natural language processing tasks. ViT [Dosovitskiy, 2020] was the first to apply the pure Transformer architecture to computer vision, achieving remarkable results in image classification, semantic segmentation, and other tasks. VST [Liu *et al.*, 2021a] proposed the first pure Transformer-based saliency object detection model, proving the effectiveness and potential of Transformer-based models for saliency detection. Swin Transformer [Liu *et al.*, 2021c], a variant of Vision Transformer, innovatively introduced the sliding window mechanism, which efficiently extracts local features while effectively capturing global context information.

Owing to their superior global context modeling capability compared to CNNs, Transformer-based methods often outperform CNN-based methods in SOD.

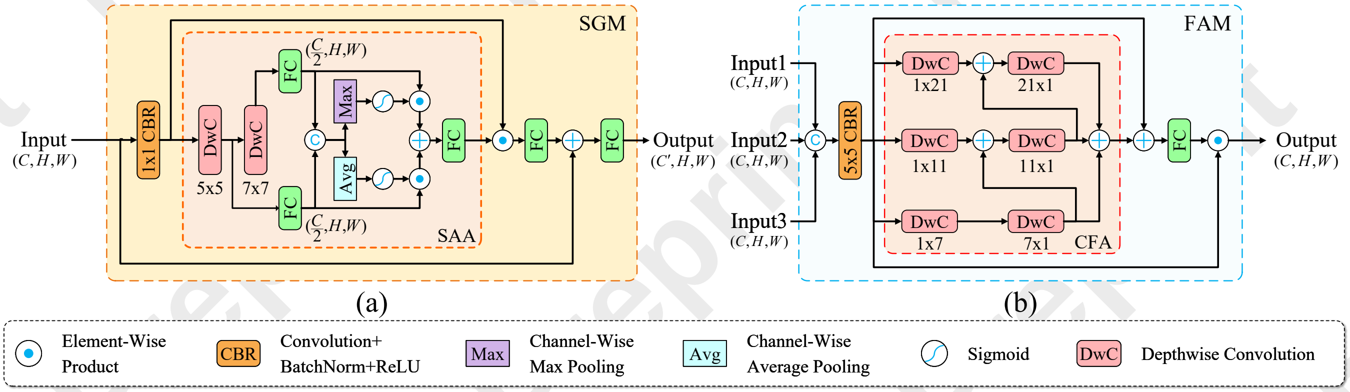


Figure 3: **(a)** The proposed Self-Guided Module (SGM). **(b)** The proposed Feature Aggregation Module (FAM).

2.2 Multi-Scale Feature Fusion

The method of integrating multi-scale features has made significant progress in SOD. For example, U-Net [Ronneberger *et al.*, 2015] proposed a U-shaped encoder-decoder architecture, which became the foundational framework for many subsequent SOD works. VST [Liu *et al.*, 2021a] was the first to use a Transformer for saliency object detection, constructing global features from a patch set to obtain a global perspective that guides feature fusion. PAKRN [Xu *et al.*, 2021] proposed the first dual-stream architecture of ”localization first, then segmentation,” where the localization branch first obtains global context information and then progressively refines the prediction results through feature fusion. EDN [Wu *et al.*, 2022] deepens the encoder layers further to obtain global context information, which is used as an attention map for channel attention to guide feature fusion. GPONet [Yi *et al.*, 2024] uses a gated recurrent network to filter redundant information during feature fusion.

Despite significant progress in existing methods for SOD, confusion between global and detail information remains a key challenge, particularly in complex scenes.

3 Method

3.1 Overall Architecture

As shown in Figure 2, for a given input image $I \in \mathbb{R}^{3\times H\times W}$, we represent its multi-level outputs generated by the backbone encoder as a set $F_{En} = \{F_{En}^{(i)} | i \in \{0, 1, 2, 3, 4\}\}$. Then, we enhance the feature set F_{En} through the SGM, reducing the number of channels to 64 to obtain the enhanced feature set $F_{SG} = \{F_{SG}^{(i)} | i \in \{0, 1, 2, 3, 4\}\}$. Following this, the FAM aggregates $F_{SG}^{(2)}$, $F_{SG}^{(3)}$ and $F_{SG}^{(4)}$ to generate the initial global feature map. Next, the F3 Stage, composed of multiple PDFUs, performs fine-grained fusion of multi-scale features from adjacent levels. Finally, the predictions are refined using CRFbp to enhance the overall results.

3.2 Self-Guided Module

Previous research [Luo *et al.*, 2016] has shown that increasing the size of the convolution kernel can enlarge the effective receptive field (ERF). However, larger convolution kernels increase computational costs and hinder the deepening of the model. Therefore, we combine large kernel dilated convolutions with depthwise separable mechanisms [Chollet, 2017] and introduce Scale-Aware Attention (SAA) to further enhance the network’s scale adaptability.

As shown in Figure 3(a), suppose the input of SSA is $X \in \mathbb{R}^{C\times H\times W}$. We first use two depthwise separable convolutions with different receptive fields to expand the global receptive field of the input features, resulting in receptive field enhanced features $X_{ds}^{(1)}$ and $X_{ds}^{(2)}$:

$$\begin{aligned} X_d^{(1)} &= DwC_{(5,1)}(X), \\ X_{ds}^{(1)} &= FC(X_d^{(1)}), \\ X_{ds}^{(2)} &= FC(DwC_{(7,3)}(X_d^{(1)})). \end{aligned} \tag{1}$$

Where $X_{ds}^{(i)} \in \mathbb{R}^{\frac{C}{2}\times H\times W}$; and $DwC_{(k,d)}(\cdot)$ denotes a depthwise convolution layer with a kernel size of $k \times k$ and a dilation rate of d; and $FC(\cdot)$ represents a fully connected layer, which is technically implemented by a 1×1 convolution layer.

Next, channel-wise average pooling and max pooling are applied to obtain A_{avg} and A_{max}, which are then used to weight and sum with $X_{ds}^{(1)}$ and $X_{ds}^{(2)}$ to produce the output of SSA:

$$\begin{aligned} S &= Cat(X_{ds}^{(1)}, X_{ds}^{(2)}) \in \mathbb{R}^{C\times H\times W}, \\ A_{avg} &= \sigma(Avg(S)) \in \mathbb{R}^{1\times H\times W}, \\ A_{max} &= \sigma(Max(S)) \in \mathbb{R}^{1\times H\times W}, \\ A_{SSA} &= FC(A_{avg} \odot X_{ds}^{(1)} \oplus A_{max} \odot X_{ds}^{(2)}). \end{aligned} \tag{2}$$

Where $A_{SSA} \in \mathbb{R}^{C\times H\times W}$, which is consistent with the input feature X; $Cat(\cdot)$ denotes the concatenation operation, $\oplus$ represents the element-wise addition operation, $\odot$ indicates the element-wise multiplication operation, and $\sigma(\cdot)$ represents the sigmoid activation function.

Assuming the input feature of the encoder is $F_{En} \in \mathbb{R}^{C\times H\times W}$, the process of SGM can be described as follows:

$$\begin{aligned} F_{En}^{'} &= CBR_{1\times 1}(F_{En}), \\ F_{SG} &= FC(F_{En} \oplus FC(F_{En}^{'} \odot SSA(F_{En}^{'}))). \end{aligned} \tag{3}$$

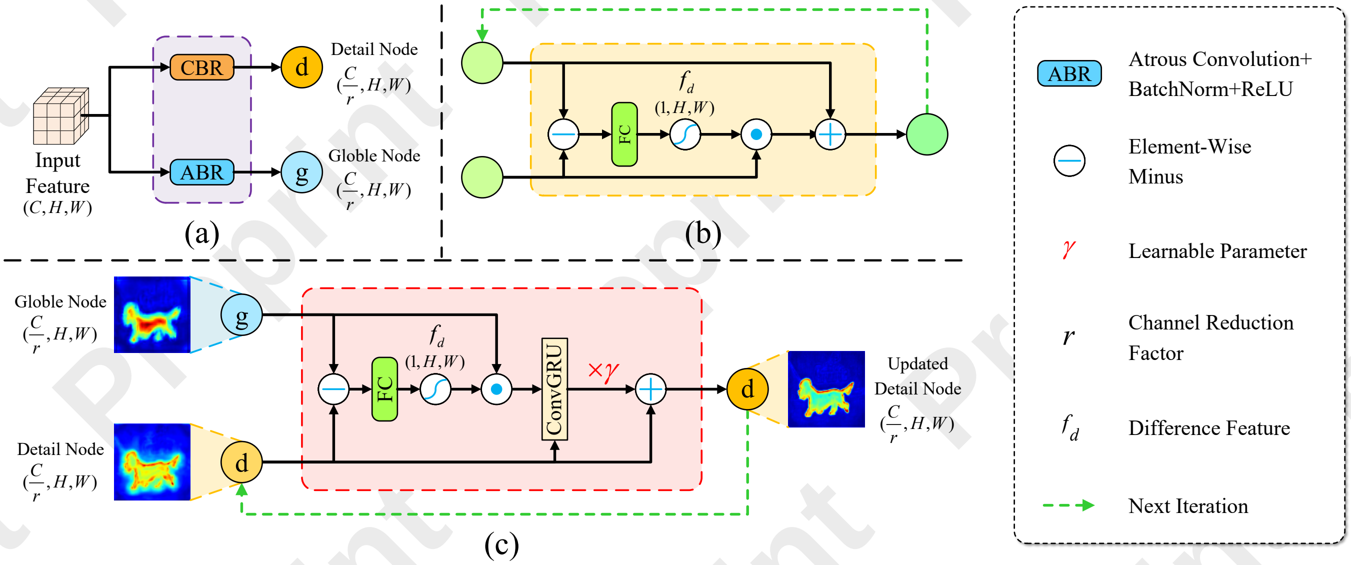


Figure 4: **(a)** The proposed Fine-grained Node Projection (FGNP). **(b)** The proposed External Interaction (EI). **(c)** The proposed Internal Interaction (II).

Where $F_{SG} \in \mathbb{R}^{C' \times H \times W}$, the SGM compresses the feature channels to C' in the final FC layer and C' is set to 64 in this paper; $CBR_{k\times k}(\cdot)$ denotes a sequence of operations consisting of a convolution layer, batch normalization, and ReLU activation; and k represents the size of the convolution kernel as $k \times k$.

3.3 Feature Aggregation Module

The structure of the Feature Aggregation Module (FAM) is shown in Figure 3(b). Since the first two stages of the encoder (Stage 0 and 1) contain excessive low-level detail information, we select the features $F_{SG}^{(2)}$, $F_{SG}^{(3)}$ and $F_{SG}^{(4)}$ from the three deepest stages of the encoder for preliminary aggregation:

$$F_{fuse} = CBR_{5\times5}(Cat(F_{SG}^{(2)}, U_2(F_{SG}^{(3)}), U_4(F_{SG}^{(4)}))) \tag{4}$$

Where $F_{fuse} \in \mathbb{R}^{64\times H\times W}$; $U_i(\cdot)$ represents the dynamic upsampling operation [Liu *et al.*, 2023], with i denoting the upsampling factor.

Previous works have demonstrated [Ding *et al.*, 2019; Guo *et al.*, 2022] that a cross-shaped asymmetric convolution can serve as a complement to square convolutions, helping capture elongated objects, enhancing the model's shape adaptability. Based on this, we design the Cross-Feedback Attention (CFA). The process can be represented as:

$$F_M^{(i)} = \begin{cases} DwC_{1\times k_i}(DwC_{k_i\times 1}(F_{fuse})), & i = 0 \\ DwC_{1\times k_i}(F_M^{(i-1)} \oplus DwC_{k_i\times 1}(F_{fuse})), & i > 0 \end{cases} \tag{5}$$

$$A_{CFA} = \sum_{i=1}^{3} F_M^{(i)} \in \mathbb{R}^{64\times H\times W} \tag{6}$$

Where $DwC_{k\times k}(\cdot)$ represents the depthwise convolution operation, and k denotes the size of the convolution kernel.

CFA captures multi-scale contextual information through multi-path asymmetric depthwise convolutions and aggregates local information from other branches via feedback connections.

Finally, multi-scale dynamic weights are used to enhance the fused features:

$$\begin{aligned} A'_{CFA} &= FC(A_{CFA} \oplus F_{fuse}) \in \mathbb{R}^{64\times H\times W}, \\ F_{FA} &= A'_{CFA} \cdot F_{fuse}. \end{aligned} \tag{7}$$

Where $F_{FA} \in \mathbb{R}^{64\times H\times W}$ and represents the initial aggregated features.

3.4 Progressive Difference Fusion Unit

Fine-grained Node Projection

Typically, researchers assume that the deepest features from the encoder represent global features [Zhao *et al.*, 2017; Zhao *et al.*, 2021; Wu *et al.*, 2022]. However, high-level features may contain some detail information, and low-level features may also carry some global information. As shown in Figure 4(a), to extract feature information at a finer granularity, we design the Fine-grained Node Projection (FGNP), which projects the features into detail nodes and global nodes based on different scales. In general, for an input feature $X \in \mathbb{R}^{C\times H\times W}$, the FGNP process can be expressed as:

$$\begin{aligned} d^{(i)} &= CBR_{3\times3}(X^{(i)}) \in \mathbb{R}^{\frac{C}{r}\times H\times W}, \\ g^{(i)} &= ABR_{(3,4)}(X^{(i)}) \in \mathbb{R}^{\frac{C}{r}\times H\times W}. \end{aligned} \tag{8}$$

Where i denotes the layer of the input feature from the encoder; r is the channel reduction factor used to decrease the computational load during node interactions, which is set to 2 in this paper; $ABR_{(k,d)}(\cdot)$ represents an operation sequence consisting of an atrous convolution with kernel size $k \times k$, dilation rate d, followed by BatchNorm and ReLU layers.

Fine-grained Node Interaction

For convenience in discussion, we define two different nodes to be interacted as x and y. Then, the difference feature from x to y, denoted as $f_d^{x\to y}$, is defined as:

$$f_d^{x\to y} = \sigma(FC(x-y)) \in \mathbb{R}^{1\times H\times W} \tag{9}$$

As shown in Figure 4(b) and (c), we define the interaction between nodes of the same level but different scales as internal interaction $I(\cdot,\cdot)$, while the interaction between nodes of different levels but the same scale is defined as external interaction $E(\cdot,\cdot)$. This process can be expressed as:

$$\begin{aligned} x_{(k+1)} &= E(x_{(k)}, y) \\ &= x_{(k)} + y \odot f_d^{y\to x}, \\ x_{(k+1)} &= I(x_{(k)}, y) \\ &= x_{(k)} + \gamma \cdot CG(y \odot f_d^{x\to y}, x_{(k)}). \end{aligned} \tag{10}$$

Where k denotes the number of iterations; CG($\cdot$) represents convGRU [Ballas *et al.*, 2015]; and γ represents the learnable parameter, and its initial value is set to 0 in this paper.

Let the low-level feature be $X^{(i)} \in \mathbb{R}^{C\times H\times W}$ and the high-level feature be $X^{(i+1)} \in \mathbb{R}^{C\times H\times W}$. The initial fine-grained node set generated from $X^{(i)}$ is denoted as $V^{(i)} = \{d^{(i)}_{(0,0)}, \mathfrak{g}^{(i)}_{(0,0)}\}$. Here, $v^{(i)}_{(t_1,t_2)}$ represents the fine-grained node obtained by projecting the feature $X^{(i)}$, which has undergone t_1 rounds of internal interactions and t_2 rounds of external interactions. For the fine-grained node set $V^{(i)}$, one round of external interaction can be represented as:

$$\begin{aligned} d^{(i)}_{(t_1+1,0)} &= E(d^{(i)}_{(t_1,0)}, d^{(i+1)}_{(t_1,0)}), \\ g^{(i)}_{(t_1+1,0)} &= E(g^{(i)}_{(t_1,0)}, g^{(i+1)}_{(t_1,0)}). \end{aligned} \tag{11}$$

One round of internal interaction can be represented as:

$$\begin{aligned} d^{(i)}_{(K,t_2+1)} &= I(d^{(i)}_{(K,t_2)}, g^{(i)}_{(K,t_2)}), \\ g^{(i)}_{(K,t_2+1)} &= I(g^{(i)}_{(K,t_2)}, d^{(i)}_{(K,t_2)}). \end{aligned} \tag{12}$$

Where K is the total number of iterations, set to 2 in this paper. After K interactions, all nodes in $V^{(i)}$ and $V^{(i+1)}$ undergo fine-grained fusion to obtain the output of the i-th layer PDFU:

$$F^{(i)}_{PDFU} = FC(Cat(d^{(i)}_{(K,K)}, g^{(i)}_{(K,K)}, d^{(i+1)}_{(K,K)}, g^{(i+1)}_{(K,K)})) \tag{13}$$

Where $F^{(i)}_{PDFU} \in \mathbb{R}^{C\times H\times W}$, which is consistent with the input features.

Fine-grained Feature Fusion Stage

We designed the Fine-Grained Feature Fusion Stage (F3 Stage) based on PDFUs to progressively perform fine-grained fusion of multi-scale features layer by layer and output intermediate prediction results for multi-scale supervision:

$$\begin{aligned} P^{(2)}_M &= FC(F_{FA}) \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}}, \\ P^{(1)}_M &= FC(F^{(1)}_{PDFU}) \oplus U_2(P^{(2)}_M) \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}}, \\ P^{(0)}_M &= FC(F^{(0)}_{PDFU}) \oplus U_2(P^{(1)}_M) \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}}. \end{aligned} \tag{14}$$

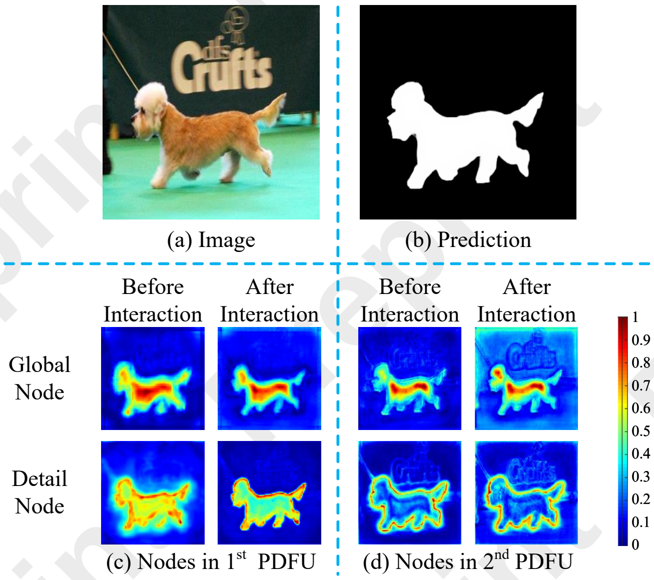

(a) Image
(b) Prediction
(c) Nodes in 1st PDFU
(d) Nodes in 2nd PDFU

Figure 5: Visualization of fine-grained nodes within the PDFU at different levels.

Where $P^{(i)}_M$ represents the intermediate prediction results, with i indicating the decoder's layer. The model's generated prediction results can be expressed as:

$$P = FC(U_2(F^{(0)}_{PDFU} + \mathrm{F}^{(0)}_{\mathrm{SG}})) + U_2(P^{(0)}_M) \in \mathbb{R}^{H\times W} \tag{15}$$

The visualization results of the fine-grained nodes are shown in Figure 5. It can be observed that for the same level of PDFU, global nodes mainly process global localization information, while the weights of detail nodes are concentrated around the edges of the salient objects, processing detailed segmentation information. Specifically, as interactions with other nodes progress, the model enhances its learning of edge information, which is crucial for segmentation in SOD.

3.5 Supervision Strategy

For training, we adopted the widely used multi-level supervision strategy in this field. For the loss function, we use the combination of $\mathcal{L}_{wbce}$ and $\mathcal{L}_{wiou}$ proposed by F3Net [Wei *et al.*, 2020]. The saliency loss and total loss can then be defined as:

$$\begin{aligned} &\mathcal{L}_{sal} = \mathcal{L}_{wbce} + \mathcal{L}_{wiou}, \\ &\mathcal{L}_{total} = \mathcal{L}_{sal}(P, GT) + \beta \sum_{i=0}^{2} \mathcal{L}_{sal}(P^{(i)}_M, D_{2^{i+1}}(GT)). \end{aligned} \tag{16}$$

Where $GT \in \mathbb{R}^{H\times W}$ represents the ground truth; $D_i(\cdot)$ denotes the downsampling operation implemented by bilinear interpolation, with i indicating the downsampling factor; β is a hyperparameter and in this work, and it is set to 1.

4 Experiment and Analysis

4.1 Datasets

Following the setup of most existing studies, our model will be trained on the DUTS-TR [Wang *et al.*, 2017] dataset

Method	Soure	DUTS-TE				DUT-OMRON				ECSSD				HKU-IS				PASCAL-S			
		$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$
F3Net	AAAI"20	.888	.835	.920	.035	.838	.747	.864	.053	.924	.912	.948	.033	.917	.900	.952	.028	.861	.816	.898	.061
TSPOANet	TPAMI"21	.860	.767	.885	.049	.818	.697	.840	.061	.907	.876	.927	.046	.902	.862	.931	.038	.842	.775	.871	.077
VST*	ICCV"21	.896	.828	.919	.037	.850	.755	.871	.058	.932	.910	.951	.033	.928	.897	.952	.029	.872	.816	.902	.061
PAKRN	AAAI"21	.901	.861	.935	.033	.853	.779	.888	.050	.928	.918	.950	.032	.923	.909	.956	.027	.858	.817	.839	.067
PoolNet+	TPAMI"22	.887	.817	.910	.037	.831	.725	.848	.054	.926	.904	.945	.035	.919	.888	.945	.030	.865	.809	.896	.065
EDN	TIP"22	.909	.867	.937	.030	.865	.792	.890	.045	.938	.929	.958	.027	.934	.918	.960	.023	.877	.842	.909	.056
ICON-S*	TPAMI"22	.917	.886	.954	.025	.869	.804	.900	.043	.941	.936	.966	.023	.935	.925	.968	.022	.885	.854	.924	.048
PGNet*	CVPR"22	.911	.874	.942	.027	.855	.775	.879	.045	.938	.929	.959	.027	.929	.916	.959	.024	.880	.844	.916	.052
SelfReformer*	TMM"23	.911	.872	.943	.027	.861	.784	.884	.043	.936	.926	.957	.027	.931	.915	.960	.024	.881	.848	.919	.051
BBRF*	TIP"23	.909	.886	.949	.025	.861	.803	.896	.044	.939	.944	.969	.022	.932	.932	.969	.020	.878	.856	.923	.049
MENet	CVPR"23	.905	.870	.938	.028	.850	.771	.871	.045	.928	.920	.951	.031	.927	.917	.960	.023	.872	.838	.910	.054
MGuidNet	TOMM"23	.888	.818	.908	.037	.836	.751	.865	.056	.927	.900	.956	.036	.922	.890	.944	.031	.869	.812	.897	.061
DASOD	IMAVIS"24	.893	.856	.933	.034	.850	.782	.884	.053	.932	.930	.961	.027	.924	.916	.962	.025	.865	.833	.909	.059
GLSTR*	TETCI"24	.919	.873	.944	.027	.868	.787	.890	.046	.942	.930	.961	.025	.936	.914	.961	.024	.886	.846	.919	.052
ISNet	PR"24	.896	.849	.929	.034	.848	.764	.877	.052	.929	.917	.951	.032	.922	.905	.954	.027	.864	.822	902	.062
GPONet*	PR"24	.919	.872	.941	.028	.874	.799	.898	.045	.945	.932	.962	.025	.937	.918	.961	.023	.880	.839	.912	.055
Ours*	——	**.931**	**.918**	**.963**	**.019**	**.875**	**.824**	**.906**	**.040**	**.948**	**.952**	**.970**	**.019**	**.942**	**.942**	**.973**	**.017**	**.889**	**.872**	**.930**	**.044**

Table 1: Quantitative comparisons between our proposed method and other 16 methods on five benchmark datasets under metrics of S-measure (S_α), Weighted F-measure (F_β^ω), Average E-measure (E_ξ^a), Mean Absolute Error (M). Methods marked with an asterisk (*) are based on the Transformer encoder. Text in bold indicates the best performance.

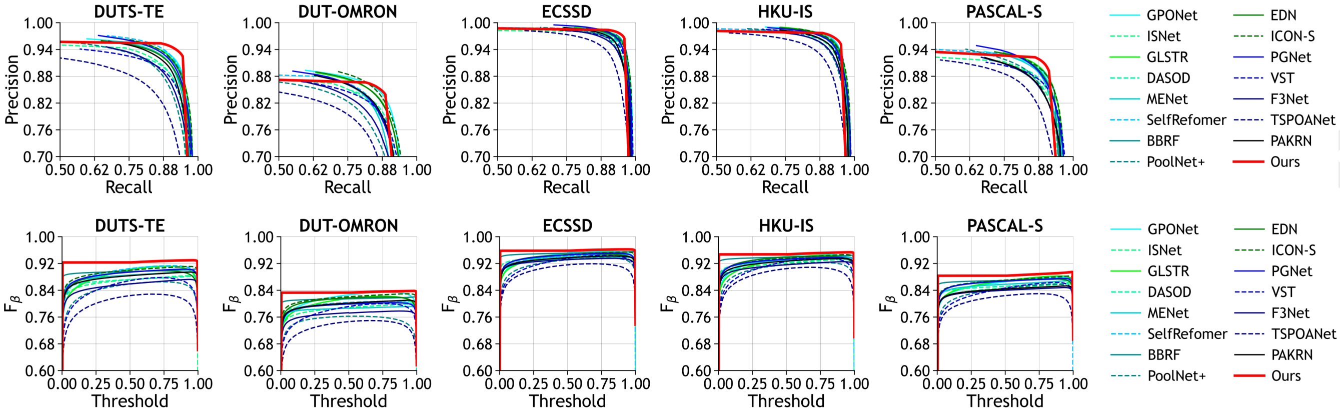


Figure 6: Precision-Recall Curves (row 1) and F-measure Curves (row 2) comparison on five saliency benchmark datasets. As shown above, our network achieved the best results among all networks across five datasets.

and evaluated on five widely recognized benchmark datasets: DUT-OMRON [Yang *et al.*, 2013], DUTS-TE [Wang *et al.*, 2017], ECSSD [Yan *et al.*, 2013], HKU-IS [Li and Yu, 2015], and PASCAL-S [Li *et al.*, 2014].

4.2 Implementation Details

Our PDFNet employs Swin-B as the backbone network. Training and testing were conducted on an NVIDIA 2080 Ti, with input images resized to 384x384. During model training, the Adam optimizer is employed with an initial learning rate set to 1.0×10^{-5}. The learning rate adjustment follows a polynomial decay (PolyLr) strategy, with the entire training process spanning 80 epochs. The batch size is set to 4 to balance memory consumption and training efficiency.

4.3 Evaluation Metric

We will evaluate all methods using the following widely used evaluation metrics in the SOD field: Mean Absolute Error (M) [Perazzi *et al.*, 2012], S-measure (S_α) [Fan *et al.*, 2017], Weighted F-measure (F_β^ω) [Margolin *et al.*, 2014], and Average E-measure (E_ξ^a) [Fan *et al.*, 2018].

4.4 Comparisons with State-of-the-Art

This paper compares the proposed method with the top 16 state-of-the-art models in recent years, including: F3Net [Wei *et al.*, 2020], TSPOANet [Liu *et al.*, 2021b], VST [Liu *et al.*, 2021a], PAKRN [Xu *et al.*, 2021], PoolNet+ [Liu *et al.*, 2022], EDN [Wu *et al.*, 2022], ICON-S [Zhuge *et al.*, 2022], PGNet [Xie *et al.*, 2022], SelfReformer [Yun and Lin, 2023], BBRF [Ma *et al.*, 2023], MENet [Wang *et al.*, 2023], MGuidNet [Hui *et al.*, 2023], DASOD [Asheghi *et al.*, 2024], GLSTR [Ren *et al.*, 2024], ISNet [Zhu *et al.*, 2024], and GPONet [Yi *et al.*, 2024]. Specifically, some models have two implementations, one based on CNNs and the other on Transformers. We have chosen the Transformer-based implementations for comparison.

No.	Components				DUTS-TE				DUT-OMRON				ECSSD				HKU-IS				PASCAL-S			
	SGM	FAM	PDFU	CRFbp	$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^\omega\uparrow$	$E_\xi^a\uparrow$	$M\downarrow$
0					.913	.873	.934	.031	.862	.789	.881	.057	.930	.923	.947	.029	.928	.914	.954	.030	.874	.844	.905	.059
1		✓	✓		.926	.903	.950	.024	.872	.811	.895	.045	.940	.939	.962	.022	.936	.926	.963	.024	.885	.860	.921	.048
2	✓		✓		.928	.905	.953	.024	.873	.813	.896	.044	.941	.941	.963	.022	.937	.928	.963	.023	.885	.861	.922	.046
3	✓	✓			.925	.895	.948	.026	.870	.804	.891	.047	.938	.935	.959	.024	.934	.925	.959	.026	.883	.857	.916	.049
4	✓	✓	✓		.933	.912	.961	.020	.877	.818	.904	.041	.949	.948	.968	.020	.942	.936	.970	.018	.891	.868	.928	.044
5	✓	✓	✓	✓	.931	.918	.963	.019	.875	.824	.906	.040	.948	.952	.970	.019	.942	.942	.973	.017	.889	.872	.930	.044

Table 2: Ablation study on each component of PDFNet on SGM, FAM, PDFU and CRFbp. Text in red represents the best result, while blue indicates the second-best result.

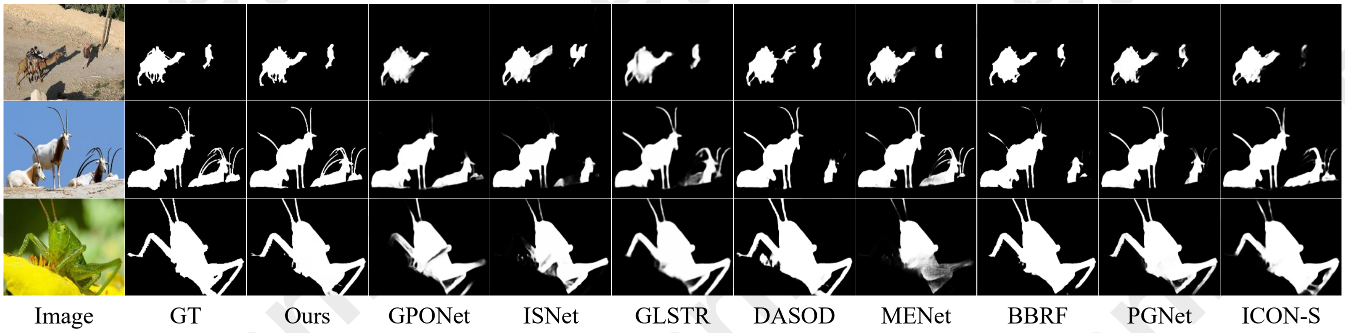


Figure 7: Visual comparisons between our proposed method and 8 state-of-the-art networks.

Quantitative Comparison

As shown in Table 1, our PDFNet achieves the best performance across five classic benchmark datasets, outperforming other state-of-the-art methods. Furthermore, Figure 6 presents the PR curves and F-measure curves for the aforementioned networks. From the figure, it is clearly observed that our method (denoted by the red curve) outperforms other methods.

Visual Comparison

The visual comparison results are shown in Figure 7. Compared to other methods, our predictions perform excellently in complex scenes: for scenes with salient objects of inconsistent scales (row 1), our method achieves better localization. For elongated and irregularly shaped objects (row 2), our method generates more refined boundaries. In scenes where salient objects are difficult to identify (row 3), our method is able to accurately locate and segment.

4.5 Ablation Study

Effectiveness of SGM

As shown in Table 2, to validate the effectiveness of SGM, we removed it from model (4) to obtain model (1). It can be observed that on five datasets, the performance metrics have all decreased. Specifically, E_ξ^a decreased by 1.1%, 0.9%, 0.6%, 0.7%, and 0.7%, respectively. This indicates that SGM significantly enhances the model's scale adaptability.

Effectiveness of FAM

To validate the effectiveness of FAM, we removed FAM from model (4) to obtain model (2). It can be observed that on five datasets, the performance metrics have all decreased. Specifically, E_ξ^a decreased by 0.8%, 0.8%, 0.5%, 0.7%, and 0.6%, respectively. This indicates that FAM enhances the model's shape adaptability.

Effectiveness of PDFU

To further validate the effectiveness of PDFU, we removed PDFU from model (4) to obtain model (3). It can be observed that the performance metrics have all decreased on five datasets. Specifically, F_β^ω decreased by 1.7%, 1.4%, 1.3%, 1.1%, and 1.1%, respectively. This demonstrates that PDFU effectively alleviates the issue of confusion between global and detail information.

Effectiveness of CRFbp

From (4) and (5) in Table 2, we can observe that after adding the CRFbp post-processing, the results of model (5) generally outperform those of model (4). With the inclusion of CRFbp, the F_β^ω of (5) increased by 0.6%, 0.6%, 0.4%, 0.6% and 0.4% on five datasets compared to (4). At the same time, we notice that S_α in (5) slightly decreased compared to (4). This may be because CRFbp focuses on processing local discrete points, which may disrupt the structural integrity of the predicted map. Therefore, we recommend using CRFbp as an optional step during the model prediction phase.

5 Conclusion

In this paper, to address the performance bottleneck of SOD models in complex scenes, we propose a new SOD architecture—PDFNet. Specifically, we first design the Self-Guided Module (SGM) to enhance the model's scale adaptability. Then, we design the Feature Aggregation Module (FAM), which enhances the model's shape adaptability. Finally, we design the F3 Stage based on Progressive Difference Fusion Units (PDFU) to alleviate the issue of information confusion within features. Additionally, we optimize the fully connected CRF and propose CRFbp, which further improves the model's performance. Our proposed PDFNet achieves state-of-the-art results on five widely used datasets.

Acknowledgments

This work was supported in part by the National Key Research and Development Plan of China under Grant 2021YFB3600503, in part by the National Natural Science Foundation of China under Grant 61972097 and U21A20472, in part by the Major Scientific Research Project for Technology Promotes Police under Grant 2024YZ040001, in part by the Natural Science Foundation of Fujian Province under Grant 2021J01612 and 2020J01494, in part by the Major Science and Technology Project of Fujian Province under Grant 2021HZ022007, in part by the Fujian Collaborative Innovation Center for Big Data Application in Governments, and in part by the Fujian Engineering Research Center of Big Data Analysis and Processing.

References

[Asheghi *et al.*, 2024] Bahareh Asheghi, Pedram Salehpour, Abdolhamid Moallemi Khiavi, Mahdi Hashemzadeh, and Amirhassan Monajemi. Dasod: Detail-aware salient object detection. *Image and Vision Computing*, 148:105154, 2024.

[Ballas *et al.*, 2015] Nicolas Ballas, L. Yao, Christopher Joseph Pal, and Aaron C. Courville. Delving deeper into convolutional networks for learning video representations. *CoRR*, abs/1511.06432, 2015.

[Cheng *et al.*, 2021] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15334–15342, 2021.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[Ding *et al.*, 2019] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1911–1920, 2019.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fan *et al.*, 2017] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017.

[Fan *et al.*, 2018] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018.

[Guo *et al.*, 2022] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1140–1156. Curran Associates, Inc., 2022.

[Hui *et al.*, 2023] Shuaixiong Hui, Qiang Guo, Xiaoyu Geng, and Caiming Zhang. Multi-guidance cnns for salient object detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–19, 2023.

[Jiang *et al.*, 2013] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA, 2001.

[Li and Yu, 2015] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015.

[Li *et al.*, 2014] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 280–287, 2014.

[Liu *et al.*, 2021a] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4722–4732, 2021.

[Liu *et al.*, 2021b] Yi Liu, Dingwen Zhang, Qiang Zhang, and Jungong Han. Part-object relational visual saliency. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3688–3704, 2021.

[Liu *et al.*, 2021c] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2022] Jiang-Jiang Liu, Qibin Hou, Zhi-Ang Liu, and Ming-Ming Cheng. Poolnet+: Exploring the potential of pooling for salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):887–904, 2022.

[Liu *et al.*, 2023] Wenze Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. Learning to upsample by learning to sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6027–6037, 2023.

[Luo *et al.*, 2001] M Ronnier Luo, Guihua Cui, and Bryan Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(5):340–350, 2001.

[Luo *et al.*, 2016] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[Ma *et al.*, 2023] Mingcan Ma, Changqun Xia, Chenxi Xie, Xiaowu Chen, and Jia Li. Boosting broader receptive fields for salient object detection. *IEEE Transactions on Image Processing*, 32:1026–1038, 2023.

[Margolin *et al.*, 2014] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2014.

[Perazzi *et al.*, 2012] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–740, 2012.

[Ren *et al.*, 2024] Sucheng Ren, Nanxuan Zhao, Qiang Wen, Guoqiang Han, and Shengfeng He. Unifying global-local representations in salient object detection with transformers. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Wang *et al.*, 2017] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017.

[Wang *et al.*, 2023] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10031–10040, 2023.

[Wei *et al.*, 2020] Jun Wei, Shuhui Wang, and Qingming Huang. F^{3}net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12321–12328, 2020.

[Wu *et al.*, 2022] Yu-Huan Wu, Yun Liu, Le Zhang, Ming-Ming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, 31:3125–3136, 2022.

[Xie *et al.*, 2022] Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11717–11726, 2022.

[Xu *et al.*, 2021] Binwei Xu, Haoran Liang, Ronghua Liang, and Peng Chen. Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3004–3012, 2021.

[Yan *et al.*, 2013] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1155–1162, 2013.

[Yang *et al.*, 2013] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.

[Yi *et al.*, 2024] Yugen Yi, Ningyi Zhang, Wei Zhou, Yanjiao Shi, Gengsheng Xie, and Jianzhong Wang. Gponet: A two-stream gated progressive optimization network for salient object detection. *Pattern Recognition*, 150:110330, 2024.

[Yun and Lin, 2023] Yi Ke Yun and Weisi Lin. Towards a complete and detail-preserved salient object detection. *IEEE Transactions on Multimedia*, 2023.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[Zhao *et al.*, 2021] Zhirui Zhao, Changqun Xia, Chenxi Xie, and Jia Li. Complementary trilateral decoder for fast and accurate salient object detection. In *Proceedings of the 29th acm international conference on multimedia*, pages 4967–4975, 2021.

[Zhou *et al.*, 2024] Huajun Zhou, Yang Lin, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Benchmarking deep models on salient object detection. *Pattern Recognition*, 145:109951, 2024.

[Zhu *et al.*, 2024] Ge Zhu, Jinbao Li, and Yahong Guo. Separate first, then segment: An integrity segmentation network for salient object detection. *Pattern Recognition*, 150:110328, 2024.

[Zhuge *et al.*, 2022] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3738–3752, 2022.