RGB-D Salient Object Detection via Disentangled Representations

Mingfeng Zha[a], Feiyang Fu[a], Guoqing Wang[a,*], Tianyu Li[a], Ningjuan Ruan[c], Qiao Liu[a], Xiongxin Tang[b], Yang Yang[a], Heng Tao Shen[d]

[a] *University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, Chengdu, 611731, China*
[b] *Institute of Software, Chinese Academy of Science, No.4, South Fourth Street, Zhongguancun, Haidian District, Beijing, 100190, China*
[c] *Beijing Institute of Space Mechanics and Electricity, Beijing, 100094, China*
[d] *Tongji University, No.4800, Cao'an Highway, Shanghai, 201804, China*

Abstract

Existing RGB-D salient object detection methods design fusion strategies to integrate multimodal information but lack exploration of modal characteristics. To address this, we separately leverage the RGB and depth branches to learn disentangled representations. Specifically, to reduce modal discrepancies, we propose the focus-align-refine (FAR) module, which achieves fine-grained alignment guided by the semantic prototypes, and then refines with complementary modalities. Since depth maps with poor quality inevitably introduce noises, we design the uncertainty feature perception (UFP) module to explore high uncertainty regions by establishing pixel-level probability distributions. And we propose the progressive scale aware (PSA) module to handle object scale variations. Finally, we introduce the adjacency shrinkage (AS) decoder to generate high quality masks by aggregating continuous and adjacent features. Extensive experiments on five RGB-D SOD datasets demonstrate that our method surpasses the state-of-the-art methods, and results on underwater and mirror scenes datasets further show the impressing generalization.

Keywords: Multimodal fusion, Salient object detection, Feature disentanglement

*Corresponding author

1. Introduction

alient object detection (SOD) aims to locate the most visually appealing regions or objects in an image and can facilitate other tasks such as segmentation [1, 2], visual-language navigation [3, 4], image generation [5, 6]. However, it is usually susceptible to interference from complex backgrounds. To deal with this, some works attempt to introduce other visual cues to improve detection robustness. Inspired by the fact that RGB images are rich in visual appearance (*e.g.,* texture, colour) and depth maps can indicate the spatial geometry of objects, thus highlighting the edges or shapes. Some works [7, 8] have explored various image or feature fusion methods to facilitate complementary information. As shown in Figure 1, despite the state-of-the-art (SOTA) detection methods, the performance still needs to be improved when coping with some challenging scenarios. Inspired by [9], we disentangle salient objects into two parts, *e.g.,* trunk and details, where the trunk emphasises the central region of the object and the details indicate the edges and their surrounding regions. RGB images are rich in visual information, but edge regions are not clear. On the contrary, depth maps can highlight the object contours based on the depth differences. Therefore, the trunk and detail parts of salient objects can be derived from the feature decoding of RGB images and depth maps, respectively. Note that we do not directly employ edge maps as supervision for the depth branch, mainly due to two considerations: 1) Depth maps are subject to estimation errors, making it challenging to accurately reflect the position and contours of salient objects, especially in extreme cases. 2) The proportion of edge pixels to nearby background pixels is highly imbalanced, thus directly utilizing edge supervision would hinder the optimization process of the training.

We further rethink the existing detection frameworks and propose the disentangled representation learning architecture. As illustrated in Figure 2, existing RGB-D SOD methods can be roughly categorized into single-stream [10], two-stream [11] and three-stream [12] architectures by the fusion strategy. Single-stream networks typically employ early fusion, which involves fusing information at the image level. However, they lack the ability to explore modal differences, resulting in the introduction of much interference. Dual-stream networks consist of separate encoding and decoding branches

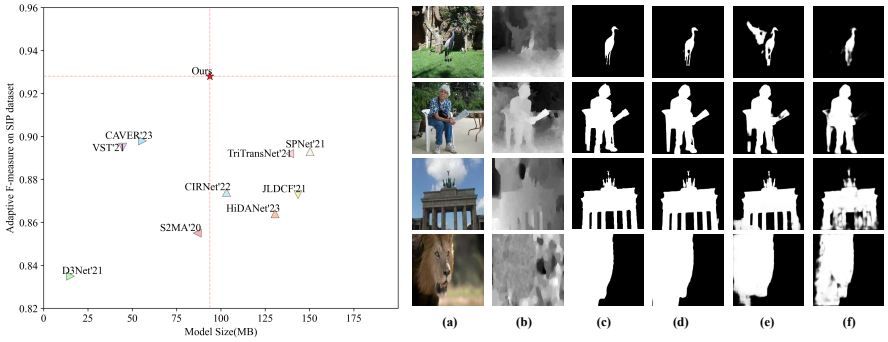

Figure 1: Our method achieves the best trade-off between efficiency and performance. (a) RGB images. (b) Depth maps. (c) Ground truths. (d) Our results. (e)-(f) Saliency maps produced by PICRNet [8] and CoNet [10], respectively.

for RGB images and depth maps with feature interactions. Triple-stream networks construct modal-shared branches at the image or feature level, which are learned jointly with two modality-specific branches. However, this significantly increases computational complexity. These frameworks are developed based on the advantages of fusing different modalities, without exploring the appropriate supervision information for each modality. Our approach combines the strengths of dual-stream and triple-stream networks. We facilitate feature interactions at the encoding stage and integrate modality-specific features at the decoding stage. We employ the trunk, detail, and binary ground truth maps as supervision for the RGB, depth, and fusion branches, respectively.

Based on the above discussions, we propose the **P**rogressive **D**isentanglement network based on **U**ncertainty perception, *i.e.,* (PDUNet). Our method consists of four parts. Specifically, the existing challenges in feature fusion mainly arise from three aspects: a) The complex visual relationships in RGB images; b) The imaging quality of depth maps; c) The alignment of corresponding features. To address these challenges, we introduce the FAR module that achieves calibration from local-global-local perspectives. We focus on the RGB and depth information of salient objects through prototype clustering, mining common features to reduce the interference of complex visual representations and low-quality depth information. These features are then fused with the initial RGB or depth features to achieve pixel-wise modality alignment, en-

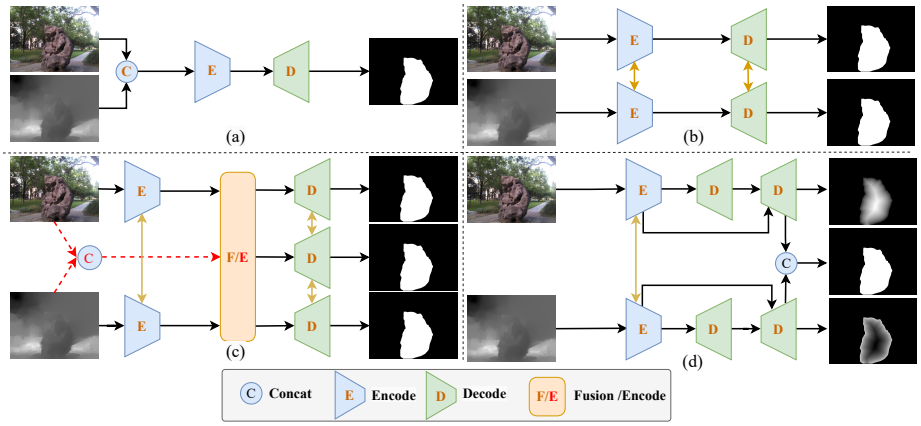

Figure 2: Architecture comparisons of RGB-D SOD models. (a) Single-stream model based on image-level fusion; (b) Dual-stream model based on feature interaction; (c) Triple-stream model based on intermediate modality or features; (d) Our proposed PDUNet, exploring the trunk and detail features, respectively.

abling complementary feature calibration from coarse to fine levels. Subsequently, we refine the features to generate cross-modal representations with different emphasis. Through the FAR module, we can effectively learn modality-shared and modality-specific knowledge. However, the introduction of depth errors is particularly harmful to weak salient feature representation or fine-grained regions, such as small objects and irregular regions. To overcome this, we design the UFP module that employs Bayesian learning to model the probability distribution of objects and incorporates into the feature maps. By exploring high uncertainty regions, we enhance the RGB representation and suppress depth errors, enabling fine-grained discrimination. In addition, we observe significant scale variations of salient objects, ranging from occupying only a tenth of the image to larger than half. Therefore, we formulate the PSA module, which gradually expands the feature perception range through cascaded convolution kernels with increasing size and dilation rates, capturing objects at different scales. Finally, unlike the U-shaped decoders [13, 14] that progressively propagate high-level semantic information to low-level detail information, we introduce the AS decoder that aggregates features from continuous and adjacent layers with similar semantic and detail representations, generating more refined masks. We further propose the refined AS (RAS) decoder that combines the outputs of the AS decoder, and feature maps gener-

ated by encoders to refine the initial features. Note that we employ the UFP module to further enhance the output of the RAS decoder and obtain the final predicted masks.

In summary, our contributions are as follows:

- We propose the novel disentangled representation learning framework that explores the characteristics of RGB and depth images and their correspondence with the trunk and detail parts.
- We propose the PDUNet by designing the FAR and UFP modules to narrow the modality gap, emphasize beneficial feature representations, and introduce the PSA module to handle variations of object scale. Furthermore, we formulate the AS decoder to obtain high quality masks.
- Our proposed method surpass SOTA methods on five widely used RGB-D datasets of natural scenes, as well as underwater and mirror detection datasets (RGB-D type), achieving better balance between performance and efficiency.

2. Related work

2.1. Salient Object Detection

Based on development process, these methods can be divided into two categories. The first category involves handcrafted feature design or the incorporation of prior features, such as background prior [15], center prior [16], context and shape prior [17]. However, these methods heavily rely on expert knowledge and lack generalization. The second category utilizes convolutional neural networks, Transformer series, or combination of both to extract high-dimensional features from images. Various enhancement strategies, such as multi-scale perception, edge supervision, and hybrid losses, are designed to explore complete salient objects. Although these methods have achieved promising results, there is still significant space for improvement in challenging scenarios such as low contrast, multiple objects, and complex backgrounds. Therefore, some recent works focus on two main aspects: 1) Improve image resolution to enrich visual information [18]; 2) Attempt to introduce additional visual [19, 20] cue as guidance or supplementation to improve detection performance and robustness. For example,

leveraging the depth differences in depth maps can effectively locate the approximate edges or contours of people or objects, laying the foundation for further SOD.

2.2. Cross-Modal Feature Fusion

Current RGB-D SOD methods mainly focus on exploiting the strengths of RGB and depth maps, and thus design various cross-modal fusion strategies. Sun *et al.* [21] leveraged depth-sensitive attention and automatic architecture search to refine cross-modal feature integration, using geometric priors. Liao *et al.* [22] enhanced feature fusion with a cross-collaborative encoder and cross-modal decoder for multi-scale complementary information aggregation, improving detection robustness. Unlike the above approaches, we propose the FAR module to achieve coarse- and fine-grained alignment and refinement from the local-global-local perspective.

2.3. Uncertainty Estimation

Due to the introduction of mismatched depth information, fine-grained features tend to weaken or disappear. Therefore, it is necessary to estimate uncertainty of the feature maps. Uncertainty can be categorized into aleatoric uncertainty and epistemic uncertainty, which describe the confidence in the input data and the predictions, respectively. The former can be learned implicitly by the network, while the latter often requires additional methods for measurement, such as Gaussian mixture and deep ensemble [23]. Unlike the above methods that cannot use gradient optimization, we use Bayesian inference to learn the posterior distribution of network weights and obtain epistemic uncertainty through the variance of sampling points. We further propose the UFP module to mine problem (high uncertainty) regions and thus enhance fine-grained perception.

3. Proposed method

3.1. Overview

The motivation of our proposed method is to disentangle salient objects into trunk and detail components, which are separately predicted by RGB and depth branches and

then integrated together. By employing the disentanglement-integration process, we aim to explore the characteristics of different modalities and leverage their respective advantages.

As shown in Figure 3, our method follows the overall paradigm of an encoder-decoder architecture and consists of four key elements. Given an image $I \in \mathbb{R}^{3\times H\times W}$ and corresponding depth map $D_o \in \mathbb{R}^{1\times H\times W}$ (adjusted to $D \in \mathbb{R}^{3\times H\times W}$ by convolution), we feed them to the respective encoders (weights are not shared) to obtain multiscale feature maps $X^i \in \mathbb{R}^{C_i\times \frac{H}{2^i}\times \frac{W}{2^i}}$ and $F^i \in \mathbb{R}^{C_i\times \frac{H}{2^i}\times \frac{W}{2^i}}$, where C, H, and W denote the number of channels, height, and width, respectively, $i \in \{1, 2, 3, 4\}$, $C \in \{64, 128, 320, 512\}$. We further utilize the FAR module to align and fuse the same-level features of X^i and F^i to generate X^i_f and F^i_f. We construct probability distributions for X^4_f via the UFP module to locate the high uncertainty regions and fuse with the features generated by the AS decoder to form M^i. M^i is then augmented with scale-awareness via the PSA module to generate M^i_r and M^i_d. Finally, we input M^i_r, M^i_d and X^i, F^i to the RAS decoder to obtain the trunk and detail maps, respectively, and fuse the outputs to form the final prediction maps. Note that we use different supervisions for the RGB, depth, and fusion branches, respectively.

3.2. Focus-Alignment-Refinement Module

Based on the complex visual relationships in RGB images and the imaging quality of depth maps, different regions of salient objects may have weakened features, while the most prominent parts still maintain good representations. Inspired by [24, 25], we generate salient prototypes by clustering to obtain common feature representations that focuses on the core regions. We further utilize generated prototypes as complements to the initial features, achieving feature alignment from coarse to fine levels. Note that, unlike previous works [24, 25, 26] that align prototypes or pixels, leading to under- or over-alignment, our approach combines the advantages of both by aligning details guided by critical features. In addition, since the representations learned by the RGB and depth branches are different, we refine them to obtain features with different emphases. The structure details of the FAR module are illustrated in Figure 4.

Coarse-to-fine alignment. We first use 3x3 convolution followed by softmax

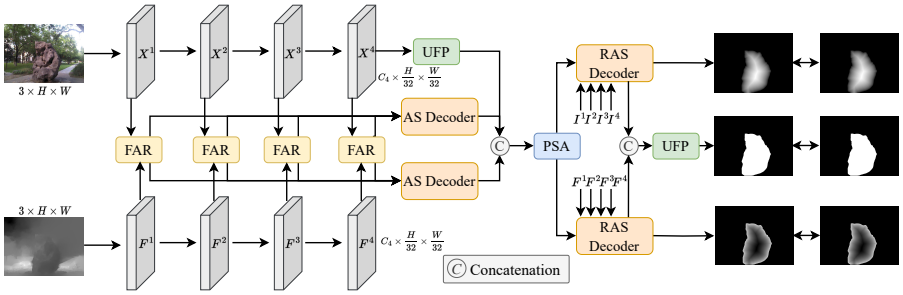

Figure 3: The overview of our PDUNet. We feed RGB and depth images separately into the encoder to generate multi-scale feature maps $\{X\}_i^4$ and $\{F\}_i^4$. Subsequently, we apply the FAR module to fuse the corresponding level feature maps. When $i = 4$, we utilize the UFP module to explore regions of high uncertainty and obtain initial output features by combining the generated feature maps with $\{X\}_i^4$ and $\{F\}_i^4$ using the AS decoder. Furthermore, we enhance scale sensitivity with the PSA module, combining the generated feature maps with $\{X\}_i^4$ and $\{F\}_i^4$ to obtain trunk and detail maps through the RAS decoder. We then refine the fused features using the UFP module to generate the final saliency map. Different supervisions are applied to the three branches.

for $X^i \in \mathbb{R}^{C_i \times \frac{H}{2^i} \times \frac{W}{2^i}}$ and $F^i \in \mathbb{R}^{C_i \times \frac{H}{2^i} \times \frac{W}{2^i}}$ to generate attention $X_a^i \in \mathbb{R}^{m \times 1 \times \frac{H \times W}{2^{i+1}}}$ and $F_a^i \in \mathbb{R}^{m \times 1 \times \frac{H \times W}{2^{i+1}}}$, where m represents the number of prototypes. The process can be formulated as:

$$X_a^i = Att(X^i), F_a^i = Att(F^i) \tag{1}$$

where Att represents the combination of regularization, convolution, softmax, and reshape. We randomly initialize the cluster centers $K_{x/f} \in \mathbb{R}^{m \times C_i \times \frac{H \times W}{2^{i+1}}}$ and generate the pixel difference $D_{x/f} \in \mathbb{R}^{m \times C_i \times \frac{H \times W}{2^{i+1}}}$ with X' and F' (reshape from X and F). We then obtain the saliency prototypes $P_{x/f} \in \mathbb{R}^{C_i \times \frac{H}{2^i} \times \frac{W}{2^i}}$ by summation and regularization, which can be written as:

$$P_{x/f} = \frac{\sum_{j=1}^{H \times W} X'/F' - K_{x/f}}{\| \sum_{j=1}^{H \times W} X'/F' - K_{x/f} \|} \tag{2}$$

where $\| \cdot \|$represents l2 norm. Note that / means or, and we omit reshape. Finally, we fuse $P_{x/f}$ with initial features X^i and F^i to generate prototype-guided X_p^i and F_p^i, which can be expressed as:

$$X_p^i = \Gamma(P_x, X^i), F_p^i = \Gamma(P_f, F^i) \tag{3}$$

where $\Gamma(\cdot, \cdot)$ represents the combination of channel concatenation and 1×1 convolution followed by ReLU.

We use 1×1 followed by 3×3 depthwise separable convolution for X^i/F^i and X_p^i/F_p^i to encode local features and generate $\hat{X}^i/\hat{F}^i$, $\hat{X}_p^i/\hat{F}_p^i$. We then apply $\hat{X}_p^i/\hat{F}_p^i$ as query and $\hat{F}^i/\hat{X}^i$ as key (and value) to calculate the correlation matrix $corr \in \mathbb{R}^{h\times\frac{C_i}{h}\times\frac{C_i}{h}}$ from channel dimension (h is number of attention heads). Note that we omit reshape. The process can be expressed as:

$$corr_{r2d} = \sigma(\frac{\hat{X}_p^i \hat{F}^{iT}}{\tau}), corr_{d2r} = \sigma(\frac{\hat{F}_p^i \hat{X}^{iT}}{\tau}) \tag{4}$$

where T represents transpose, σ is softmax function, and τ is a learnable scaling factor. Thus, we can obtain the enhanced feature $X_e^i, F_e^i \in \mathbb{R}^{C_i\times\frac{H}{2^i}\times\frac{W}{2^i}}$ by:

$$X_e^i = corr_{r2d}\hat{X}^i + \hat{X}^i, F_e^i = corr_{d2r}\hat{F}^i + \hat{F}^i \tag{5}$$

Refinement. We apply spatial compression on X_e^i and F_e^i respectively, and then fuse them to obtain the channel map $S_r, S_d \in \mathbb{R}^{C_i\times1\times1}$, which can be expressed as:

$$\begin{aligned} S_r &= \Gamma(Concat(Avg(X_e^i), Avg(F_e^i)))^{0\to C_i} \\ S_d &= \Gamma(Concat(Avg(F_e^i), Avg(F_e^i)))^{C_i\to 2\times C_i} \end{aligned} \tag{6}$$

where Avg indicates average pooling, $Concat$ is channel concatenation. Thus, refined features X_r^i is generated by:

$$X_r^i = \Gamma(Concat(X_e^i \odot S_r, F_e^i \odot S_d)) + X^i \tag{7}$$

where $\odot$ denotes element-by-element multiplication. We omit the fusion coefficients. F_r^i can be generated in the same way.

3.3. Uncertainty Feature Perception Module

By aligning from coarse to fine levels, we reduce the overall interference of depth errors on salient objects. However, for regions with weak feature representations or fine-grained details, the introduction of error information can degrade the original representations or even obscure, resulting in increased uncertainty. To address this issue, we propose the UFP module, which introduces probability modeling to locate regions with high uncertainty and further reduce depth noise. The structure details of the UFP module are illustrated in Figure 5.

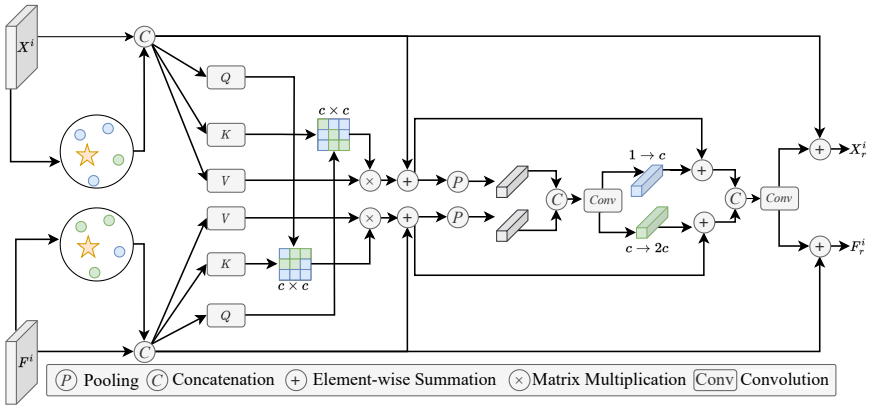

Figure 4: Structure of the FAR module. We apply coarse-to-fine alignment and refinement to obtain cross-modal representations with different emphases.

We establish Laplace distribution for each pixel of the feature maps to construct the uncertainty maps. We first obtain the mean $\mu \in \mathbb{R}^{C_4 \times \frac{H}{16} \times \frac{W}{16}}$ and variance $b \in \mathbb{R}^{C_4 \times \frac{H}{16} \times \frac{W}{16}}$ separately by using two different 1×1 convolutions on X_r^4, which can be expressed as:

$$\mu = \Gamma^{'}(X_r^4), b = \Gamma^{''}(X_r^4) \tag{8}$$

However, gradients cannot directly optimize random samples. Following [27], we randomly sample several times to generate variable ξ from standard Laplace distribution to obtain new uncertain distribution of pixels, *i.e,* $L = \mu + \xi b$. We further calculate the variance and normalize to yield uncertainty maps $U \in \mathbb{R}^{C_4 \times \frac{H}{16} \times \frac{W}{16}}$, which can be formulated as:

$$U = Norm(Var(\phi(sample(L)))) \tag{9}$$

where $Norm(\cdot)$, $Var(\cdot)$, and $\phi(\cdot)$ denote the normalization, sample variance, and sigmoid function, respectively.

We feed X_r^i and F_r^i into the AS decoder to generate $X_d, F_d \in \mathbb{R}^{C_1 \times \frac{H}{4} \times \frac{W}{4}}$, and then concatenate to output M, which can be expressed as:

$$M = Concat(AS\,decoder(X_r^i), AS\,decoder(F_r^i)) \tag{10}$$

Similar with the FAR module, we apply 1×1 followed by 3×3 convolution on M to obtain the query M_q, key M_k, and value M_v. We further apply U to the query and key

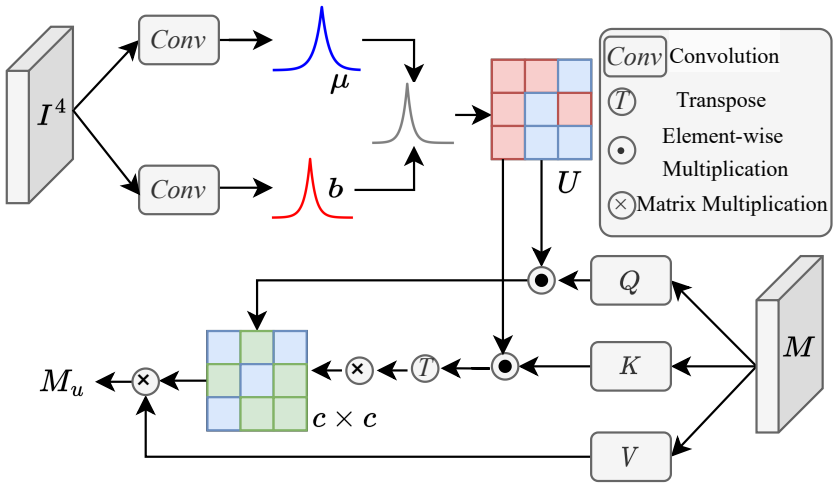

Figure 5: Structure of the UFP module. We employ pixel probability distribution to localize and enhance regions with high uncertainty.

to generate the correlation matrix, which can be formulated as:

$$corr_u = \frac{(M_q \odot U) \otimes (M_k \odot U)^T}{\tau} \tag{11}$$

Therefore, we can obtain the uncertainty perception feature $M_u \in \mathbb{R}^{2C_1 \times \frac{H}{4} \times \frac{W}{4}}$ by:

$$M_u = M_v \otimes corr_u \tag{12}$$

3.4. Progressive Scale-Aware Module

The size of salient regions varies significantly, making it difficult to locate small objects and detect complete large objects. Therefore, we propose the PSA module, which incorporates varying sizes dilated convolutions to progressively enlarge the receptive field, thus capturing different scales objects. The structure details of the PSA module are shown in Figure 5.

The main branch consists of four parts, each part containing two 3x3 convolution block branches, generating $M_u^{r_i}$ and $M_u^{d_i}$. For simplicity, we omit branch convolutions. For part one, we apply standard 1×1 followed by 3×3 convolution to generate:

$$M_u^{r_1} = Conv_{3\times3,d=1}(Conv_{1\times1}(M_u)) \tag{13}$$

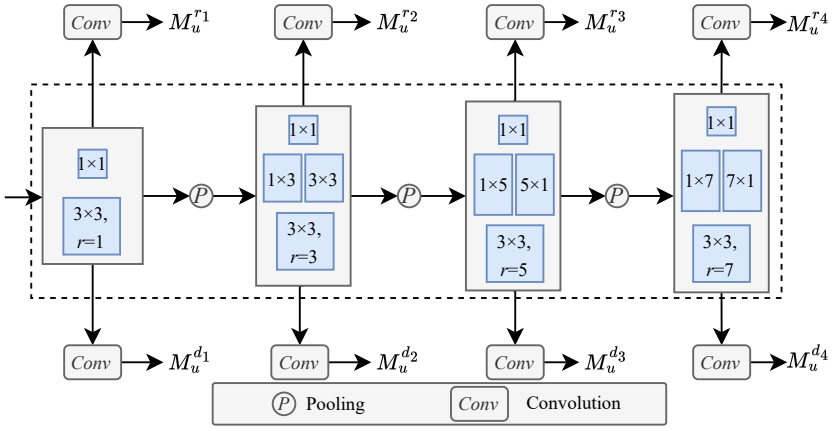

Figure 6: Structure of the PSA module. We cascade convolution blocks with different scales and dilation rates in the main branch.

For section two, we apply 1x1, followed by 1x3 and 3x1 convolutions, and 3x3 convolution with dilation rate of 3 to generate:

$$M_u^{r_2} = Conv_{3\times3,d=3}(Conv_{3\times1}(Conv_{1\times3}(Conv_{1\times1}(M_u^{r1})))) \tag{14}$$

Similarly, we can also generate $M_u^{r_3}$, $M_u^{r_4}$:

$$M_u^{r_3} = Conv_{3\times3,d=5}(Conv_{5\times1}(Conv_{1\times5}(Conv_{1\times1}(M_u^{r2})))) \tag{15}$$

$$M_u^{r_4} = Conv_{3\times3,d=7}(Conv_{7\times1}(Conv_{1\times7}(Conv_{1\times1}(M_u^{r3})))) \tag{16}$$

Similarly, we can generate $M_u^{d_i}$.

3.5. (Refined) Adjacent Shrinkage Decoder

The existing decoding paradigms mainly include: 1) Top-down, which may cause semantic information loss; 2) Dense interaction, incurring high computational cost and feature redundancy; 3) High-low level separation, leading to conflicts between details and semantic features. Different from [28], which aggregate adjacent features with intervals, we propose the AS decoder to progressively aggregate continuous and adjacent layers, reducing the gap between high and low-level information, thus capturing features focused at different layers to decode high-quality masks.

Assuming that T_l^i and T_l^{i+1} are adjacent feature maps, the aggregation can be expressed as:

$$T_{l+1}^{i+1} = CBR(Concat(UP(T_l^i), T_l^{i+1})) \tag{17}$$

where CBR represents the combination of convolution, BN layer, and ReLU function, $UP(\cdot)$ represents upsampling. l ($l \leq 4$) is the number of layers of shrinkage pyramid, and j ($j \leq l$) is the level of current layer.

Note that unlike the AS decoder, which only aggregates X_r^i/F_r^i, the refined AS decoder simultaneously aggregates X_r^i/F_r^i and $M_u^{r_i}/M_u^{d_i}$.

3.6. Loss Function

We apply supervision to the output of RGB, depth and fusion branches. Each branch loss $\mathcal{L}_R$, $\mathcal{L}_D$, $\mathcal{L}_F$ consists of weighted BCE loss, weighted IoU loss and SSIM loss. The total loss can be expressed as:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_R + \beta\mathcal{L}_D + \gamma\mathcal{L}_F \tag{18}$$

where α, β, γ are hyperparameters.

4. Experiment

4.1. Experimental Settings

Benchmark Datasets. We conduct experiments on seven datasets. Natural scenes RGB-D datasets: STERE1000 [38], SIP [39], NJU2K [40], NLPR [41] and DUT [42]. Underwater dataset: USOD10K [43]. Mirror dataset: RGBD-Mirror [34].

Implementation Details. We implement our method using Pytorch and conduct all experiments on a single NVIDIA RTX A100. Following [44], our training set consists of 1485 images from the NJU2K dataset and 700 images from the NLPR dataset. We evaluate on the corresponding test sets, as well as the STERE1000 and SIP datasets. For the DUT dataset, following [44], we employ original training set along with the training sets of NJU2K and NLPR, and the original test set for evaluation. For the USOD10K and RGBD-Mirror dataset, we follow the settings of [43, 34]. We utilize

Table 1: Quantitative comparison of S_α, F_β, E_ξ, and MAE on five benchmark datasets. The best performances are bolded.

Methods	Pub.	STERE(1000)				SIP(929)				NJU2K(500)				NLPR(300)				DUT(400)			
		$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$
HRTransNet	TCSVT22	.921	.904	.930	.030	.909	.916	.943	.035	.933	.928	.928	.026	.942	.919	.969	.016	.951	.952	.972	.019
CAVER	TIP23	.917	.901	.935	.032	.904	.898	.937	.037	.926	.916	.929	.029	.934	.906	.964	.021	.937	.936	.962	.026
CCFENet	TCSVT22	.906	.898	.945	.035	.882	.889	.920	.047	.916	.912	.949	.032	.926	.908	.959	.021	.934	.936	.959	.027
LSNet	TIP23	.871	.853	.908	.055	.886	.881	.920	.050	.911	.898	.939	.038	.919	.880	.951	.025	-	-	-	-
DCFNet	CVPR21	.903	.889	.927	.039	.873	.877	.920	.051	.905	.896	.924	.035	.920	.889	.957	.021	.924	.926	.952	.030
DSA2F	CVPR21	.904	.892	.933	.036	.860	.868	.908	.057	.903	.897	.923	.037	.918	.890	.950	.024	.922	.923	.954	.031
VST	ICCV21	.908	.878	.942	.041	.879	.895	.922	.055	.921	.899	.949	.035	.930	.885	.961	.023	.942	.930	.962	.025
C2DFNet	TMM22	.902	.880	.927	.038	.871	.866	.912	.052	.899	.897	.919	.038	.899	.894	.958	.021	.933	.932	.958	.026
DCMF	TIP22	.910	.866	.946	.043	-	-	-	-	.913	.881	.948	.043	.922	.853	.954	.029	.928	.906	.943	.035
SPSN	ECCV22	.906	.874	.941	.035	.891	.884	.932	.043	.918	.887	.949	.032	.923	.891	.956	.023	-	-	-	-
HiDANet	TIP23	.911	.897	.944	.035	.892	.864	.925	.043	.926	.922	.951	.029	.930	.908	.959	.021	-	-	-	-
PICRNet	MM23	.921	.905	.951	.031	-	-	-	-	.927	.919	.952	.029	.935	.911	.965	.019	.943	.943	.967	.020
PopNet	ICCV23	.917	.906	.947	.033	.897	.893	.937	.040	.924	.919	.952	.030	.932	.911	.963	.019	-	-	-	-
DCTNet	TIP24	.920	.890	.941	.035	.915	.911	.945	.034	.929	.912	.945	.031	.933	.889	.952	.023	.944	.940	.960	.024
Ours	-	**.927**	**.918**	**.957**	**.025**	**.917**	**.928**	**.953**	**.029**	**.934**	**.933**	**.960**	**.023**	**.939**	**.926**	**.969**	**.016**	**.951**	**.958**	**.975**	**.016**

Table 2: Quantitative comparison on USOD10K dataset

Methods	Pub.	USOD10K(1026)			
		$S_\alpha\uparrow$	$F_\omega\uparrow$	$E_\xi\uparrow$	$M\downarrow$
CDINet [29]	MM21	.705	.581	.712	.091
SVAM-Net [30]	RSS22	.747	.581	.747	.092
TC-USOD [31]	TIP23	.922	.897	.963	.020
Ours	-	**.932**	**.918**	**.969**	**.017**

the PVT network pretrained on ImageNet as the encoder for RGB images and depth maps. All input images are scaled to 384×384. For training, we use Adam as the optimiser with 200 epochs, the initial learning rate of 1e-4 and the batch size of 16. For testing, we do not use tricks (*e.g.,* test-time data augmentation) and post-processing (*e.g.,* CRF).

Table 3: Quantitative comparison on RGBD-Mirror dataset

Methods	Pub.	**RGBD-Mirror(1049)**			
		$S_\alpha \uparrow$	$F_\omega \uparrow$	$E_\xi \uparrow$	$M \downarrow$
JL-DCF [32]	TPAMI21	.814	.750	.861	.057
VST [33]	ICCV21	.814	.750	.858	.053
PDNet [34]	CVPR21	.855	.825	.906	.041
SANet [35]	CVPR22	.853	.833	.910	.040
SATNet [36]	AAAI23	.857	.829	.901	**.034**
ADRNet [37]	INFFUS24	-	.835	-	.040
Ours	**-**	**.867**	**.840**	**.918**	.036

Evaluation Metrics. We adopt five evaluation metrics: S-measure (S_α), mean E-measure (E_ξ), weighted F-measure (F_w) , adaptive F-measure (F_β), and Mean Absolute Error (M). Note that the higher the better for the first four.

4.2. Comparisons with the State-of-the-arts

We compare our method with ten SOTA RGB-D SOD models, including DCFNet [45], DSA2F [21], VST [33], C2DFNet [46], DCMF [47], SPSN [24], HiDANet [48], PICRNet [8], PopNet [49], DCTNet [50], two RGB-T SOD models, *i.e.,* CCFENet [22], LSNet [51], and two bi-modal methods *i.e.,* HRTransNet [52], CAVER[53] on five natural scenes datasets.

To validate the generalization of our method, we compare with twenty seven SOTA methods on the USOD10K dataset. We also compare with seven RGB-D SOD SOTA methods and three mirror detection methods on the RGBD-Mirror dataset. To ensure the fairness, we obtain evaluation results by utilizing the saliency maps provided by the authors or retraining using the provided official source codes. Refer to the supplemental material for further experimental results.

4.2.1. Quantitative Evaluation

As shown in Table 1, our method achieves the best performance across five widely used datasets for all metrics. Particularly, on the large-scale and challenging STERE

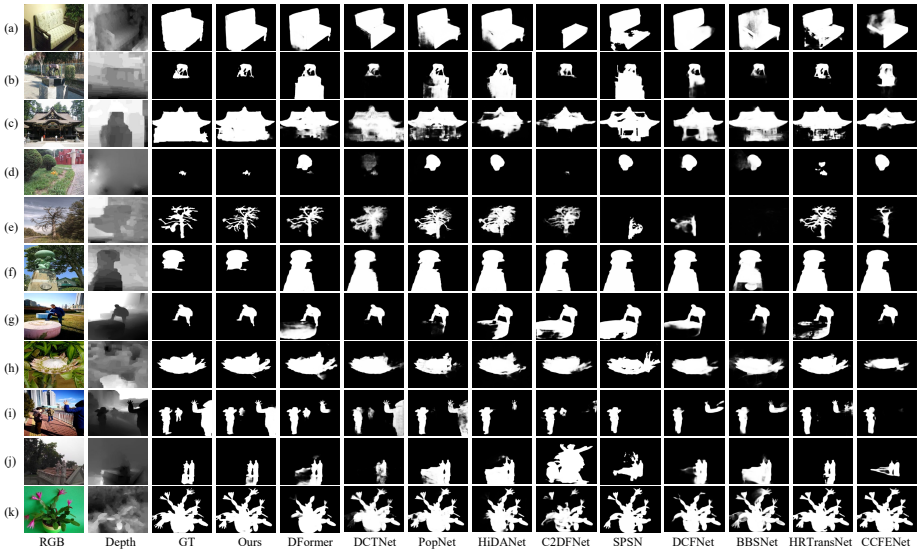

Figure 7: Qualitative comparison on natural scenes.

dataset, our method outperforms the second-best RGB-D SOD method, *i.e.,* DCTNet, 0.7%, 2.8%, 1.6%, 1.0% on S_α, F_β, E_ξ, and M. Moreover, we surpass the SOTA bimodal SOD method, *i.e.,* CAVER, 1.0%, 1.7%, 2.2%, 0.7% and the RGB-T SOD method, *i.e.,* LSNet, 5.6%, 6.5%, 4.9%, 3.0%, respectively. As shown in Table 2 and Table 3, our method also demonstrates superior generalization, surpassing the SOTA methods TC-USOD and SATNet by 1.0%, 2.1%, 0.6%, 0.3% and 1.0%, 1.1%, 1.7%, -0.2% on the USOD10K and RGBD-Mirror datasets, respectively. The results validate that our method can effectively address various and complex scenarios.

4.2.2. Qualitative Evaluation

As shown in Figure 7, we provide some comparison cases in challenging scenarios. Our method effectively establishes foreground-background differences in low contrast (rows a and b) and low-quality depth map (rows b, c, and d) scenarios, leveraging beneficial depth information while reducing the interference of depth errors. Despite significant variations in object scales (rows c and d), our method captures long- and short-range feature dependencies, enabling accurate localization and detection. In scenes with combined objects (rows f and g), where the depth differences between target and

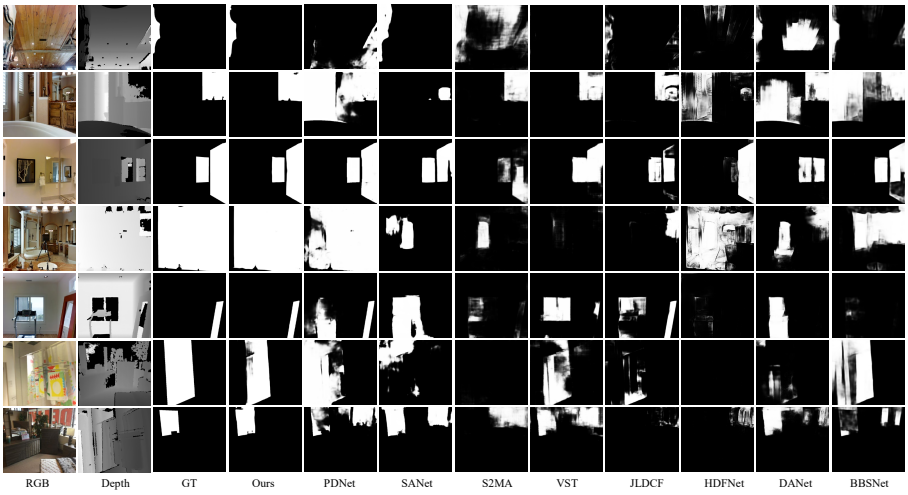

Figure 8: Qualitative comparison on mirror scenes.

non-target regions are not prominent, most methods mistakenly detect trapezoidal and circular rocks due to erroneous guidance from the depth map. In contrast, our method utilizes RGB information to achieve separation. When dealing with irregular-shaped targets (rows e, h and k), our method employs uncertainty modeling to exploit fine-grained feature cues and enhance detail information. Moreover, our method demonstrates complete detection without omissions when handling multiple targets (rows i and j). As shown in Figure 8 and Figure 9, we also provide visual comparisons of underwater and mirror scenes. Our method can address task-specific challenges, *i.e.,* camouflage and reflection.

4.3. Ablation Studies

To validate the effectiveness of each key module and hyperparameter, we conduct ablation experiments on the NJU2K and DUT datasets, including quantitative (Tables 4, 5 and 6) and qualitative analyses (Figure 10).

4.3.1. The effectiveness of the FAR module

As shown in Table 4, integrating the FAR module into the Baseline, we obtain an average improvement of 1.6%, 1.1%, and 1.5% in S_α, F_β, and E_ξ, respectively,

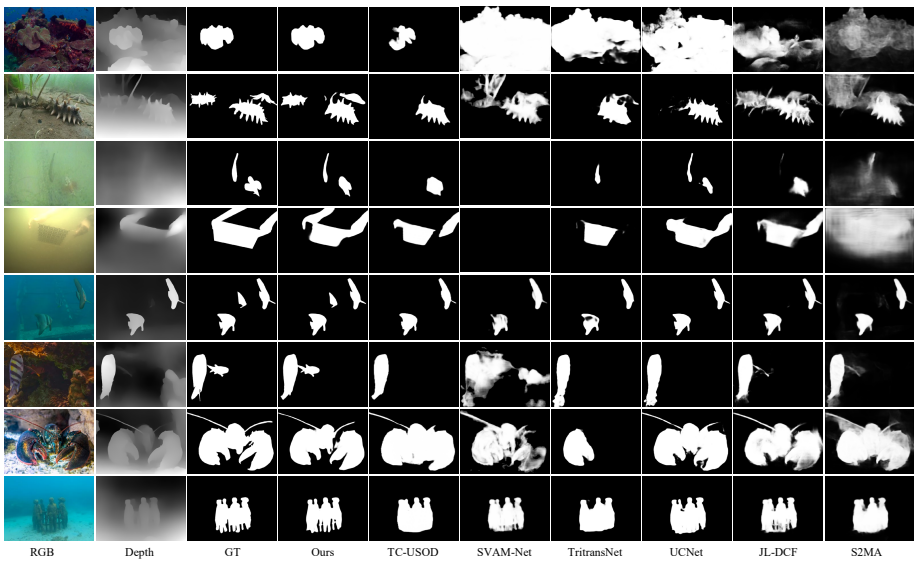

Figure 9: Qualitative comparison on underwater scenes.

with an average reduction of 0.5% on M. In Figure 10 (a, b, c), the low quality of the depth maps results in noticeable deficiencies in the baseline detection results. After integrating the FAR module, the detection of the horse tail in (a), the left mirror in (b), and the small fish in (c) have been improved significantly, demonstrating the effective feature alignment mechanism of the FAR module, which preserves the essence and discards the harmful information.

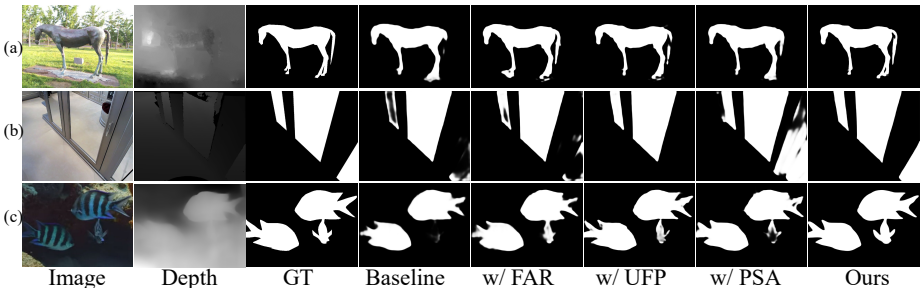

Figure 10: Visualization results of ablation study.

Table 4: Ablation study on proposed modules. B, I1, I2, and I3 indicate Baseline, FAR, UFP, and PSA, respectively.

	NJU2K				DUT			
Methods	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
B	.905	.908	.930	.033	.919	.928	.944	.025
+I1	.919	.917	.945	.028	.937	.941	.959	.020
+I2	.912	.914	.939	.030	.928	.935	.951	.021
+I3	.909	.920	.940	.027	.934	.942	.955	.021
+I1+I2	.921	.925	.944	.025	.941	.948	.967	.020
+I1+I3	.924	.922	.946	.026	.942	.946	.963	.019
+I2+I3	.920	.923	.942	.027	.940	.944	.965	.018
Ours	**.934**	**.933**	**.960**	**.023**	**.951**	**.958**	**.975**	**.016**

4.3.2. The effectiveness of the UFP module

As shown in Table 4, the average improvements after adding the UFP module are 0.8%, 0.7% (0.65%), and 0.8%, and the average decrease is 0.4% (0.35%) on M. In Figure 10, after incorporating the UFP module, we expand the detection regions (the horse tail in (a) and the left mirror in (b)), and also refine the outlines of the targets (fish tails in (c)). This indicates that the UFP module can enhance perception of fine-grained features and edge regions, highlighting detailed information.

4.3.3. The effectiveness of the PSA module

As shown in Table 4, the addition of the PSA module results in an average improvement of 1% (0.95%), 1.3% and 1.1% (1.05%), and an average reduction of 0.5% on M. Furthermore, with the addition of FAR module, the overall performance is further enhanced, indicating the complementary of them. In Figure 10 (b), by incorporating the PSA module to improve the distance dependency of features, we can simultaneously locate three mirrors, although the localization of the mirror on the right is still not accurate.

Table 5: Ablation study on different decoders. U, D, F, H, S denote top-down, dense connection, feedback, heterogeneity, shrinkage, respectively.

	NJU2K				DUT			
Types	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
U	.924	.922	.949	.026	.941	.952	.965	.019
D	.923	.923	.954	.025	.943	.949	.968	.020
F	.927	.924	.948	.023	.945	.952	.969	.018
H	.925	.927	.950	.024	.947	.950	.973	.017
S	.929	.926	.954	.024	.950	.951	.974	.018
Ours	**.934**	**.933**	**.960**	**.023**	**.951**	**.958**	**.975**	**.016**

Table 6: Ablation study of the weights of each loss function on DUT dataset

α	β	γ	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
1	.25	.25	.949	.958	.974	.016
	.50	.50	.950	.957	.973	.017
	.75	.75	.949	.959	.973	.016
	1	1	**.951**	**.958**	**.975**	**.016**

4.3.4. Types of Decoder

To validate the superiority of the AS and RAS decoder, as shown in Table 5, we apply different types of decoders, including top-down, dense connections, feature feedback, high-low level heterogeneous (separable), and feature shrinkage. Among these, we find that the AS and RAS decoder achieves the best performance, indicating the effectiveness of progressive aggregation of adjacent and continuous features.

4.3.5. The Weights of Each Loss Function

To find the best combination of weights for the triple branch loss functions, we set the weight of the fusion branch as 1, and the trunk and detail branches weights $\beta, \gamma \in$

$[0.25, 0.50, 0.75, 1]$. Through multiple iterations of different weight combinations, we discover that the best combination is $\alpha = \beta = \gamma = 1$.

5. Conclusion

In this paper, we rethink existing RGB-D SOD frameworks and propose the PDUNet. Our approach focuses on exploring the modality characteristics by disentangling salient objects and learning heterogeneous representations through different branches. To overcome the deficiencies of current modality fusion methods, we introduce the FAR module, which aligns and fuses features in a local-global-local manner. We propose the UFP module to locate regions with high uncertainty and extract fine-grained information. Furthermore, we introduce the PSA module to handle scale variations of the salient objects. Finally, we propose the AS and RAS decoders, which progressively aggregate adjacent and continuous features to reduce the semantic and detail gaps during the fusion process. Although our method has good performance on multiple benchmarks, there are still challenges in extending it to other modalities, such as events. Learning disentangled representations for multimodal feature fusion, such as infrared and visible image fusion, is very beneficial, as it achieves complementary advantages. Designing disentanglement methods based on the characteristics of auxiliary modalities is a question worth considering. In the future, we will attempt to use disentanglement learning to achieve a unified RGB-X SOD framework and ensure its lightweight nature for deployment on mobile devices.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant U23B2011, 62102069, U20B2063 and 62220106008, the Key R&D Program of Zhejiang under grant 2024SSYS0091.

References

[1] S. Lee, M. Lee, J. Lee, H. Shim, Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation, in: Proceedings

of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5495–5505.

[2] T. Chen, Y. Yao, L. Zhang, Q. Wang, G. Xie, F. Shen, Saliency guided inter- and intra-class relation constraints for weakly supervised semantic segmentation, IEEE Transactions on Multimedia (2022).

[3] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, S. Song, Clip on wheels: Zero-shot object navigation as object localization and exploration, arXiv preprint arXiv:2203.10421 3 (4) (2022) 7.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[5] L. Jiang, M. Xu, X. Wang, L. Sigal, Saliency-guided image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16509–16518.

[6] S. M. H. Miangoleh, Z. Bylinskii, E. Kee, E. Shechtman, Y. Aksoy, Realistic saliency guided image enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 186–194.

[7] H. Zhou, C. Tian, Z. Zhang, C. Li, Y. Xie, Z. Li, Frequency-aware feature aggregation network with dual-task consistency for rgb-t salient object detection, Pattern Recognition 146 (2024) 110043.

[8] R. Cong, H. Liu, C. Zhang, W. Zhang, F. Zheng, R. Song, S. Kwong, Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 406–416.

[9] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, Q. Tian, Label decoupling framework for salient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13025–13034.

[10] W. Ji, J. Li, M. Zhang, Y. Piao, H. Lu, Accurate rgb-d salient object detection via collaborative learning, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16, Springer, 2020, pp. 52–69.

[11] G. Li, Z. Liu, H. Ling, Icnet: Information conversion network for rgb-d based salient object detection, IEEE Transactions on Image Processing 29 (2020) 4873–4884.

[12] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, Y. Zhao, Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection, IEEE Transactions on Image Processing 31 (2022) 6800–6815.

[13] H. Bi, R. Wu, Z. Liu, H. Zhu, C. Zhang, T.-Z. Xiang, Cross-modal hierarchical interaction network for rgb-d salient object detection, Pattern Recognition 136 (2023) 109194.

[14] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, H. Lu, Zoom in and out: A mixed-scale triplet network for camouflaged object detection, in: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2022, pp. 2160–2170.

[15] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, F. Wu, Background prior-based salient object detection via deep reconstruction residual, IEEE Transactions on Circuits and Systems for Video Technology 25 (8) (2014) 1309–1321.

[16] R. Liu, J. Cao, Z. Lin, S. Shan, Adaptive partial differential equation learning for visual saliency detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 3866–3873.

[17] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, S. Li, Automatic salient object segmentation based on context and shape prior., in: BMVC, Vol. 6, 2011, p. 9.

[18] X. Deng, P. Zhang, W. Liu, H. Lu, Recurrent multi-scale transformer for high-resolution salient object detection, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 7413–7423.

[19] J. Deng, J. Zhang, Z. Hu, L. Wang, J. Jiang, X. Zhu, X. Chen, Y. Yuan, C. Wang, Rgb-d salient object ranking based on depth stack and truth stack for complex indoor scenes, Pattern Recognition 137 (2023) 109251.

[20] Y. Fang, H. Zhang, J. Yan, W. Jiang, Y. Liu, Udnet: Uncertainty-aware deep network for salient object detection, Pattern Recognition 134 (2023) 109099.

[21] P. Sun, W. Zhang, H. Wang, S. Li, X. Li, Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1407–1417.

[22] G. Liao, W. Gao, G. Li, J. Wang, S. Kwong, Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection, IEEE Transactions on Circuits and Systems for Video Technology 32 (11) (2022) 7646–7661.

[23] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in neural information processing systems 30 (2017).

[24] M. Lee, C. Park, S. Cho, S. Lee, Spsn: Superpixel prototype sampling network for rgb-d salient object detection, in: European conference on computer vision, Springer, 2022, pp. 630–647.

[25] Z. Zhang, J. Wang, Y. Han, Saliency prototype for rgb-d and rgb-t salient object detection, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 3696–3705.

[26] J. Wu, F. Hao, W. Liang, J. Xu, Transformer fusion and pixel-level contrastive learning for rgb-d salient object detection, IEEE Transactions on Multimedia (2023).

[27] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, D.-P. Fan, Uncertainty-guided transformer reasoning for camouflaged object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4146–4155.

[28] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, H. Xiong, Feature shrinkage pyramid for camouflaged object detection with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 5557–5566.

[29] C. Zhang, R. Cong, Q. Lin, L. Ma, F. Li, Y. Zhao, S. Kwong, Cross-modality discrepant interaction network for rgb-d salient object detection, in: Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 2094–2102.

[30] M. J. Islam, R. Wang, J. Sattar, Svam: Saliency-guided visual attention modeling by autonomous underwater robots, arXiv preprint arXiv:2011.06252 (2020).

[31] L. Hong, X. Wang, G. Zhang, M. Zhao, Usod10k: A new benchmark dataset for underwater salient object detection, IEEE Transactions on Image Processing (2023) 1–1doi:10.1109/TIP.2023.3266163.

[32] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, C. Zhu, Siamese network for rgb-d salient object detection and beyond, IEEE transactions on pattern analysis and machine intelligence 44 (9) (2021) 5541–5559.

[33] N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 4722–4732.

[34] H. Mei, B. Dong, W. Dong, P. Peers, X. Yang, Q. Zhang, X. Wei, Depth-aware mirror segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3044–3053.

[35] H. Guan, J. Lin, R. W. Lau, Learning semantic associations for mirror detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5941–5950.

[36] T. Huang, B. Dong, J. Lin, X. Liu, R. W. Lau, W. Zuo, Symmetry-aware transformer-based mirror detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 935–943.

[37] W. Zhou, Y. Cai, X. Dong, F. Qiang, W. Qiu, Adrnet-s*: Asymmetric depth registration network via contrastive knowledge distillation for rgb-d mirror segmentation, Information Fusion 108 (2024) 102392.

[38] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 454–461.

[39] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, M.-M. Cheng, Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks, IEEE Transactions on neural networks and learning systems 32 (5) (2020) 2075–2089.

[40] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in: 2014 IEEE international conference on image processing (ICIP), IEEE, 2014, pp. 1115–1119.

[41] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, Rgbd salient object detection: A benchmark and algorithms, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13, Springer, 2014, pp. 92–109.

[42] Y. Piao, W. Ji, J. Li, M. Zhang, H. Lu, Depth-induced multi-scale recurrent attention network for saliency detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7254–7263.

[43] L. Hong, X. Wang, G. Zhang, M. Zhao, Usod10k: a new benchmark dataset for underwater salient object detection, IEEE transactions on image processing (2023).

[44] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, Q. Hou, Dformer: Rethinking rgbd representation learning for semantic segmentation, arXiv preprint arXiv:2309.09668 (2023).

[45] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, et al., Calibrated rgb-d salient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9471–9481.

[46] M. Zhang, S. Yao, B. Hu, Y. Piao, W. Ji, C˜{2} dfnet: Criss-cross dynamic filter network for rgb-d salient object detection, IEEE Transactions on Multimedia (2022).

[47] F. Wang, J. Pan, S. Xu, J. Tang, Learning discriminative cross-modality features for rgb-d saliency detection, IEEE Transactions on Image Processing 31 (2022) 1285–1297.

[48] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, C. Demonceaux, Hidanet: Rgb-d salient object detection via hierarchical depth awareness, IEEE Transactions on Image Processing 32 (2023) 2160–2173.

[49] Z. Wu, D. P. Paudel, D.-P. Fan, J. Wang, S. Wang, C. Demonceaux, R. Timofte, L. Van Gool, Source-free depth for object pop-out, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1032–1042.

[50] H. Chen, F. Shen, D. Ding, Y. Deng, C. Li, Disentangled cross-modal transformer for rgb-d salient object detection and beyond, IEEE Transactions on Image Processing (2024).

[51] W. Zhou, Y. Zhu, J. Lei, R. Yang, L. Yu, Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images, IEEE Transactions on Image Processing 32 (2023) 1329–1340.

[52] B. Tang, Z. Liu, Y. Tan, Q. He, Hrtransnet: Hrformer-driven two-modality salient object detection, IEEE Transactions on Circuits and Systems for Video Technology 33 (2) (2022) 728–742.

[53] Y. Pang, X. Zhao, L. Zhang, H. Lu, Caver: Cross-modal view-mixed transformer for bi-modal salient object detection, IEEE Transactions on Image Processing 32 (2023) 892–904.