Dual Domain Perception and Progressive Refinement for Mirror Detection

Mingfeng Zha, Feiyang Fu, Yunqiang Pei, Guoqing Wang, *Member, IEEE*, Tianyu Li, Xiongxin Tang, Yang Yang, *Senior Member, IEEE*, and Heng Tao Shen, *Fellow, IEEE*

***Abstract*— Mirror detection aims to discover mirror regions in images to avoid misidentifying reflected objects. Existing methods mainly mine clues from spatial domain. We observe that the frequencies inside and outside the mirror region are distinctive. Besides, the low-frequency representing the feature semantics can help to locate the mirror region, and the high-frequency representing the details can refine it. Motivated by this, we introduce frequency guidance and propose the dual domain perception progressive refinement network (DPRNet) to mine dual-domain information. Specifically, we first decouple the images into high-frequency and low-frequency components by Laplace pyramid and vision Transformer, respectively, and design the frequency interaction alignment (FIA) module to integrate frequency features to initially localize the mirror region. To handle scale variations, we propose the multi-order feature perception (MOFP) module to adaptively aggregate adjacent features with progressive and gating mechanisms. We further propose the separation-based difference fusion (SDF) module to establish associations between entities and imagings and discover the correct boundary to mine the complete mirror region. Extensive experiments show that DPRNet outperforms the state-of-the-art method by an average of 3% with only about one-fifth of the parameters and FLOPs on four datasets. Our DPRNet also achieves promising performance on remote sensing and camouflage scenarios, validating its generalization. The code is available at https://github.com/winter-flow/DPRNet.**

***Index Terms*— Mirror detection, dual domain, reflection perception, lightweight model.**

I. Introduction

MIRROR detection (MD) is a challenging task that aims to determine imagings and entities in images and thus correctly identify and segment the mirror regions. The reflective property of mirrors can seriously affect other tasks such as segmentation [1], [2], image restoration [3], [4], depth estimation [5], vision-language navigation [6]. Therefore, accurate MD is fundamental to downstream tasks.

As shown in Fig. 2 (a), MD mainly has three issues. 1) Reflection misdirection (example 1). Reflection makes the physical properties of objects inside and outside the mirror the same, causing the model easy to confuse imagings and entities, how to establish the correlation between these two regions is the core of many methods. For example, SATNet [7] utilizes the symmetry property of mirrors. 2) Large scale variation (example 2-4). the variation of the mirror area is very large, either close to the whole image or occupying only less than 30×30 pixels (small target), making it difficult to find the corresponding entities. Utilizing additional cue guidance may be a solution. For example, Mei et al. [8] constructed a new RGBD mirror dataset and proposed the PDNet to guide MD by fusing depth information. 3) Occlusion and irregular shapes (example 5). The general shapes of mirrors are circle-like and rectangular. However, the camera angle and entities occlusion can lead to the mirror regions with various shapes and unclear boundaries. To resolve this, HetNet [9] utilizes edge generation as an auxiliary task to better handle irregular mirror regions.

Vision tasks similar to MD include salient object detection (SOD) and camouflage object detection (COD), but they can not be applied directly. Due to the interference of reflections, SOD methods are likely to detect both imagings and entities. COD aims at segmenting the foreground that is similar to the background, which can result in misidentified targets if being directly applied to mirror images. Therefore, some structures need to be customized according to the reflection mechanism. As shown in Fig. 1, our proposed efficient DPRNet achieves the best performance on F_{β}^{w} and E_m with low computational complexity, outperforming the state-of-the-art (SOTA) methods of SOD, COD, and MD, *i.e.,* VST [11], FPNet, and SATNet, respectively.

As shown in Fig. 2 (b), existing MD frameworks can be broadly categorized into three classes. 1) Plain framework. Following the U-Net encoding and decoding paradigm, the extracted features are processed by the same customized modules and then progressively fed into the decoder.

Manuscript received 19 April 2024; revised 26 May 2024; accepted 3 July 2024. Date of publication 11 July 2024; date of current version 27 November 2024. 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, the Sichuan Science and Technology Program under Grant 2024NSFTD0034. This article was recommended by Associate Editor S. Cheng. *(Corresponding author: Guoqing Wang.)*

Mingfeng Zha, Feiyang Fu, Yunqiang Pei, Guoqing Wang, and Tianyu Li are with the Center for Future Media and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: gqwang0420@uestc.edu.cn).

Xiongxin Tang is with the Institute of Software, Chinese Academy of Science, Beijing 100190, China.

Yang Yang is with the Center for Future Media and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Institute of Electronic and Information Engineering, University of Electronic Science and Technology of China, Guangdong 523808, China.

Heng Tao Shen is with the School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3426673.

Digital Object Identifier 10.1109/TCSVT.2024.3426673

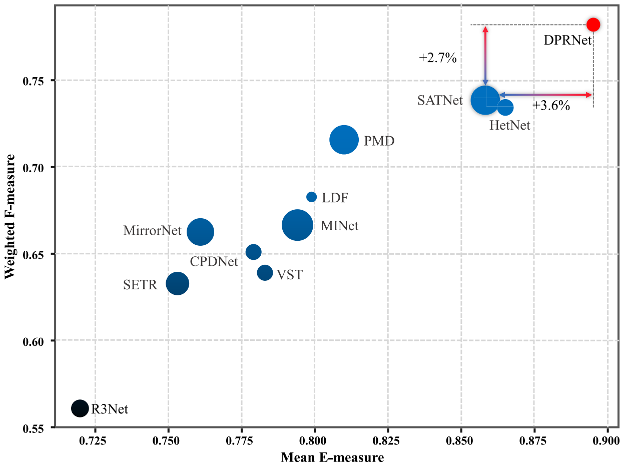


Fig. 1. Comparison of our DPRNet with ten SOTA detection methods on weighted F-measure (F_β^w), mean E-measure (E_m), and parameters using PMD dataset [10]. Larger circle indicates higher parameters. DPRNet demonstrates better performance with low computational complexity.

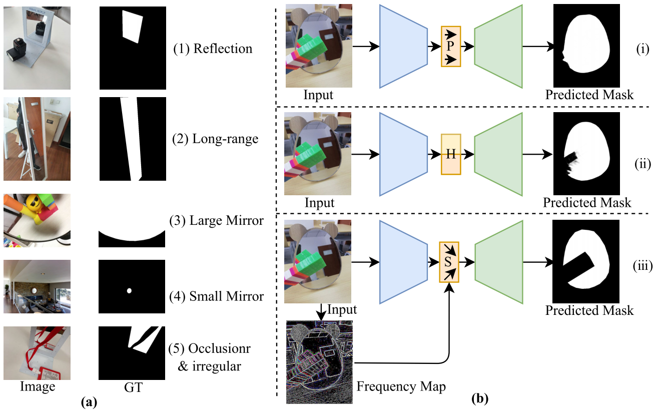


Fig. 2. Description of (a) MD task issues and (b) existing detection frameworks. (i) is **p**lain architecture represented by MirrorNet [12], PMD [10], which passes all the features of the encoder to the decoder through the same modules. (ii) is **h**eterogeneous architecture represented by SANet [13], HetNet [9], where customized modules are designed for low and high level features, respectively. (iii) is our **s**hrink architecture, which introduces frequency guidance and aggregates adjacent features.

2) Hierarchical framework. Low-level features contain more details, while high-level features have rich semantics, thus modules need to be designed separately and then integrated carefully for better results. 3) Our frequency guidance and feature shrinkage framework. Unlike the PDNet [8], which takes spatial and depth images as multimodal inputs, our approach takes only RGB images as inputs and obtains frequency features as guidance through specific transformation. In addition, we explore cross-layer correlations by aggregating adjacent features to achieve contextual information coupling without designing multiple modules.

Based on the above discussion, we propose the frequency-guided progressive refinement network (DPRNet). Since high frequency indicates details (*e.g.,* edges, textures) and low frequency indicates semantics, it is beneficial to distinguish the mirror and non-mirror regions by using the frequency information. Motivated by [14] and [15], we use vision Transformer to capture low-frequency components and Laplace pyramid to acquire multiple high-frequency components and adaptively fuse them, respectively. We further propose the frequency interaction alignment (FIA) module to eliminate frequency representation differences and align common features. The scale of mirrors varies greatly, which may result in images with only reflections but no corresponding entities, so it is insufficient to establish imagings and entities associations alone. To this end, we propose the multi-order feature perception (MOFP) module to aggregate adjacent features, and achieve deep mining of matched features through channel splitting, progressive fusion and gating mechanisms. Finally, we propose the separation-based difference fusion (SDF) module to fuse the foreground-background mask and difference map respectively to explore the correlations between mirror and non-mirror regions, imagings and entities. Thanks to these careful designs, we can mine clear mirror boundaries and detect complete mirror regions.

In summary, our main contributions are as follows:

- We introduce frequency guidance and propose the DPRNet model based on the frequency aware differences of mirror and non-mirror regions.
- We propose the FIA module to align different frequency representations. To handle scale variations, we design the MOFP moudle to aggregate adjacent features. In addition, we formulate the SDF moudle to establish object correlations in and out of the mirrors.
- Extensive experiments show that DPRNet outperforms the SOTA method on mirror detection task by an average of 3%, with only about one-fifth of the parameters and FLOPs. Our method also performs well on other tasks such as remote sensing and camouflage detection.

II. Related Work

A. Mirror Detection

Existing MD methods can be roughly divided into three categories. The first category aims to achieve more accurate detection performance. Yang et al. [12] proposed the first MD method, called MirrorNet, which explores the correlation between internal and external features of mirrors. Lin et al. [10] introduced the PMDNet, which compares mirror features with context for correspondence and incorporates edge information. Guan et al. [13] constructed semantic associations among objects based on graph representation. Huang et al. [7] built a dual-stream network based on Transformer to explore the symmetry property of mirrors. The second category involves incorporating prior information, such as depth. Mei et al. [8] proposed the PDNet, which utilizes the difference of depth information between mirrors and other objects to guide detection. However, this inevitably introduces noise. The third category focuses on constructing lightweight detection models. He et al. [9] presented the HetNet, which explores low-level and high-level features in heterogeneous manner. Without introducing prior information, our proposed DPRNet achieves the promising performance with low computational complexity.

B. (Remote Sensing) Salient/Camouflage Object Detection

SOD is a rapidly developing field aimed at discovering salient objects in images. Xie et al. [16] proposed the one-stage

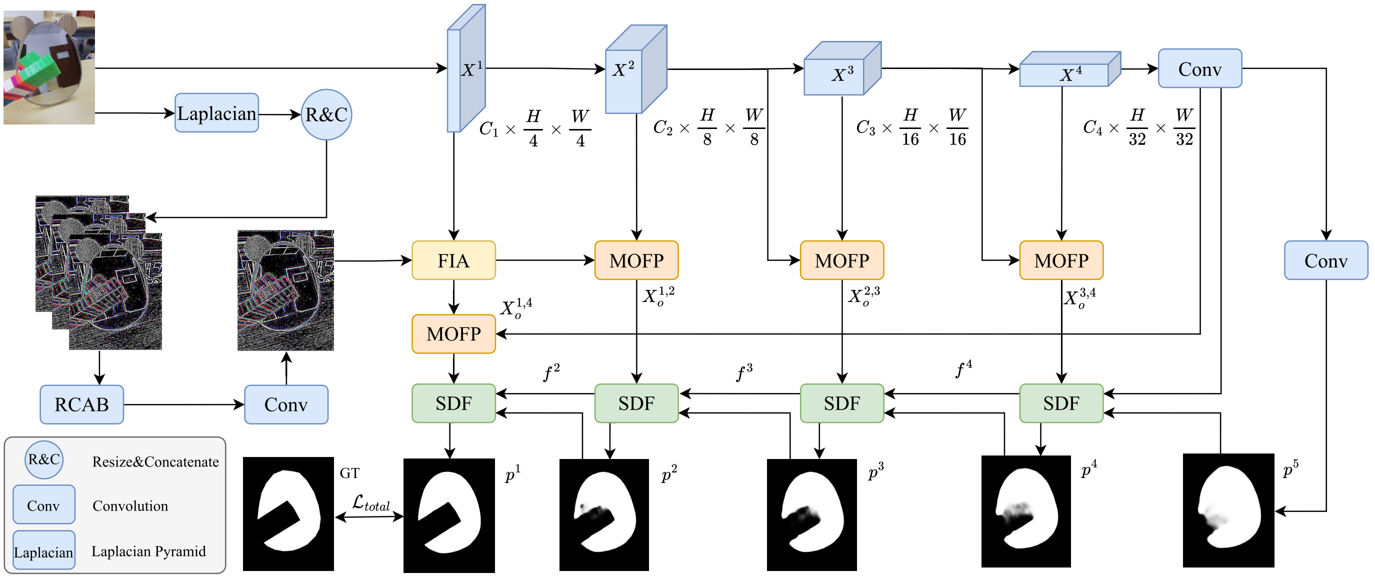


Fig. 3. The overview of our DPRNet. We first use PVT network (vision Transformer architecture) as encoder to extract spatial (low-frequency) features, and Laplace pyramid to acquire multiple high-frequency features and adaptively fuse them using RCAB. We further apply the FIA module to align the high and low frequency information. Then, we aggregate the adjacent features by the MOFP moudle. Finally, we utilize the SDF module to progressively decode to obtain the detection maps and compute the losses during the training process.

detection method for handling high-resolution images using combination of convolutional neural networks (CNN) and Transformer. Li et al. [17] introduced the ICNet through information conversion for RGB-D SOD. Remote Sensing SOD (RSSOD) aims at identifying salient objects in optical remote sensing images. Bai et al. [18] presented global-local-global context-aware network. Li et al. [19] introduced LVNet, an end-to-end network using a two-stream pyramid module. COD can be considered as the inverse task of SOD, aiming to identify camouflage (non-salient) objects. Fan et al. [20] presented the SINet, which decouples COD into search and recognition. Li et al. [21] proposed the PENet, which enhances feature representation in progressive manner. However, SOD/RSSOD/COD [22], [23], [24], [25], [26], [27], [28] and general segmentation methods [29], [30], [31], [32] do not establish correlations between entities and imagings, making them susceptible to reflection interference, thus cannot be directly applied to MD task.

C. Vision Transformer and Laplacian Pyramid

Unlike CNN that extract local features, Transformer, with self-attention (SA) [33] as its core architecture, is good at capturing long-range dependencies of features and has been applied to many vision tasks. However, Park et al. [34] have demonstrated that Transformer is insensitive to capturing high-frequency information.

Laplace pyramid [15] can decompose an image into high and low frequency bands. For arbitrary image/feature map $I^i \in \mathbb{R}^{C\times H\times W}$, the computed low-pass prediction $I_l^i \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$ is utilized to obtain the high frequency feature $I_h^i \in \mathbb{R}^{C\times H\times W}$ using $I^i - \hat{I}_l^i$, where $\hat{I}_l^i$ is upsampled from I^i, i is the number of iterations.

Inspired by the above discussion, we decouple the mirror and non-mirror regions features from frequency domain and design the FIA module for frequency alignment.

III. Proposed Method

A. Overall Architecture

The overview of our DPRNet is shown in Fig. 3. It contains three key components, *i.e.,* FIA moudule, MOFP module and SDF module. Given an image $I \in \mathbb{R}^{3\times H\times W}$, we pass it to the PVT encoder [35] and obtain multi-scale feature maps $X^i \in \mathbb{R}^{C_i\times \frac{H}{4^i}\times \frac{W}{4^i}}$, where C, H, W denote channels, height and width respectively, $i \in \{1, 2, 3, 4\}$. Meanwhile, we utilize Laplace pyramid to obtain high-frequency maps $F^i \in \mathbb{R}^{3\times \frac{H}{2^j}\times \frac{W}{2^j}}$ ($j \in \{0, 1, 2\}$) and residual channel attention block (RCAB) [36] for adaptive feature fusion, generating F_f (adjusted to the same size as X^1). We then use the FIA module to obtain better frequency representation, MOFP module and SDF module to aggregate and fuse features. Finally, we compute the loss of multiple prediction maps decoded during training.

B. Frequency Interaction Alignment Module

The high-frequency components of images represent details such as edges, textures, while the low-frequency components represent semantics. Therefore, mirror and non-mirror regions can be separated from the frequency domain. To extend the difference between these two regions, we design the FIA module for modeling better frequency representation. And we introduce frequency guidance only at X^1 to ensure the efficiency of our model. As shown in Fig. 4, FIA module can be divided into two parts: interaction and alignment.

1) Interaction: We first encode local features using 1×1 and 3×3 convolutions for $X^1 \in \mathbb{R}^{C_1\times \frac{H}{4}\times \frac{W}{4}}$ and generate query, key and value, *i.e.,* Q_r, K_r, and $V_r \in \mathbb{R}^{\sqrt{C_1}\times\sqrt{C_1}\times N}$ by dimension transformation and normalization, where $N = \frac{HW}{16}$. The process can be formulated as:

$$Q_r, K_r, V_r = \Phi(X_1) \tag{1}$$

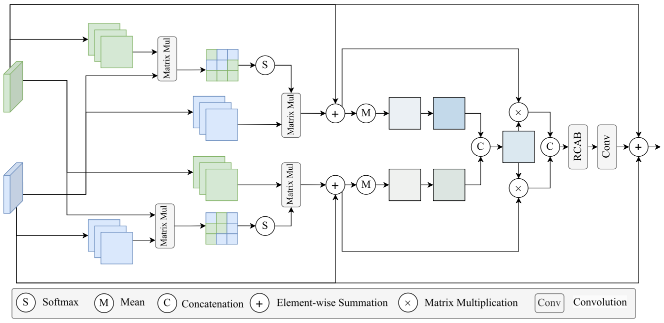


Fig. 4. Structure of the FIA module. We use it to eliminate differences for better frequency representation and divide into two parts: interaction and alignment.

where Φ denotes the combination of 3×3 depth-wise convolution [37] and feature map shape adjustment, normalization. Similarly, we can obtain Q_f, K_f, and $V_f \in \mathbb{R}^{\sqrt{C_1}\times\sqrt{C_1}\times N}$ from F_f. Then we compute the correlation map of the low-to-high frequency features to realize the interaction by:

$$corr_{l2h} = softmax(\frac{Q_r K_h^T}{\tau}) \tag{2}$$

where τ is a learnable scaling factor. Similarly, we can obtain the correlation map of high-to-low frequency $corr_{h2l} \in \mathbb{R}^{\sqrt{C_1}\times\sqrt{C_1}\times\sqrt{C_1}}$. Plain self-attention computes the correlation map $corr_{plain} \in \mathbb{R}^{C_1\times N\times N}$ from the spatial perspective (mainly related to width and height) and thus requires more computational complexity. We instead save at least three times the resource consumption (memory usage) from the channel perspective with essentially no impact on the performance. We further generate post-interaction low-frequency features by:

$$X_I^1 = corr_{h2l} V_r + X^1 \tag{3}$$

Similarly, we can obtain post-interaction high-frequency features F_I. Note that shape adjustment is required for sum.

2) Alignment: We perform channel compression on X_I^1 and F_I respectively, and then fuse them to obtain the spatial map $S \in \mathbb{R}^{2\times\frac{H}{4}\times\frac{W}{4}}$, which can be expressed as:

$$S = \sigma(CR(Concat(Avg(X_I^1), Avg(F_I)))) \tag{4}$$

where Avg indicates that pixels are averaged along the channel dimension, σ expresses the sigmoid, and CR expresses the convolution and ReLU. Thus, the alignment feature $X_f^1 \in \mathbb{R}^{C_1\times\frac{H}{4}\times\frac{W}{4}}$ is generated by:

$$X_f^1 = CR(RCAB(Concat(S^1 \odot X_I^1, S^2 \odot F_I))) + X^1 + F_f \tag{5}$$

where S^1 and $S^2 \in \mathbb{R}^{1\times\frac{H}{4}\times\frac{W}{4}}$ are the first and second channel maps of S, respectively. $\odot$ denotes element-by-element multiplication.

As shown in Fig. 5, we provide high and low frequency maps obtained by Laplace pyramid and corresponding frequency statistics, as well as the original image and statistics. The frequency distributions of the mirror (blue) and non-mirror regions (red) differ remarkably, especially in the high-frequency part.

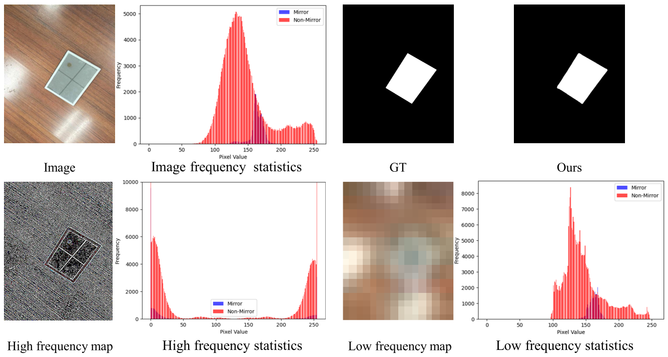


Image Image frequency statistics GT Ours

High frequency map High frequency statistics Low frequency map Low frequency statistics

Fig. 5. Frequency visualizations of mirror and non-mirror regions.

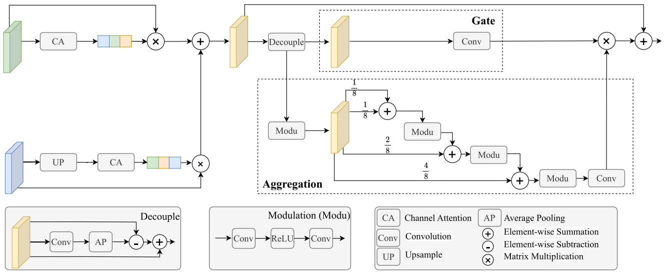


Fig. 6. Structure of the MOFP module. We concat the adjacent feature maps and then progressively explore the correlation.

C. Multi-Order Feature Perception Module

The size of the mirror region varies greatly, *e.g.,* large targets may be more than ten times the size of small targets. In such scenario, entities may not exist that correspond to imagings, or only partially correspond, making it difficult to infer the complete mirror region by establishing associations between them. To this end, we propose the MOFP module to explore the correlation of neighboring features and integrate contextual information. The structure details of the MOFP (abbr. M^k) are shown in Fig. 6, where $k \in \{1, 2, 3, 4\}$.

Specifically, when $k = 2, 3$, we adjust X^{k+1} and X^k to the same shape. When $k = 1$ and $k = 4$, we resize X_f^1 and X_2, the output of M_1 and X_{conv}^4 (generated based on X^4 by convolution), respectively. Using $k = 2$ as an example, we compute the weights of the neighboring feature maps by channel attention (CA) [38] for adaptive fusion, generating $X_f^{2,3}$, which can be formulated as:

$$X_f^{2,3} = CA(X^2) \odot X^2 + CA(UP(X^3)) \odot UP(X^3) \tag{6}$$

where UP denotes up-sampling. We then decouple $X_f^{2,3}$ by:

$$X_d = \delta(X^{2,3} + \Psi(X_f^{2,3} - AP(X_f^{2,3}))) \tag{7}$$

where AP denotes adaptive average pooling. Ψ is a scaling function, we set the coefficient to 1e-5. δ denotes the GELU activation function [39].

To explore the multiorder relationships within the features, we generate X_d^l by splitting X_d by the ratio 1:1:2:4, where $l \in \{1, 2, 3, 4\}$. Taking the example of aggregating X_d^1 and X_d^2, the process can be interpreted as:

$$X_p^{1,2} = Proj(\delta(DWConv_{5\times5}(X_d^1 + X_d^2))) \tag{8}$$

where $DWConv_{5\times5}$ means depth-wise convolution of size 5×5. We map channels of the current feature map to its

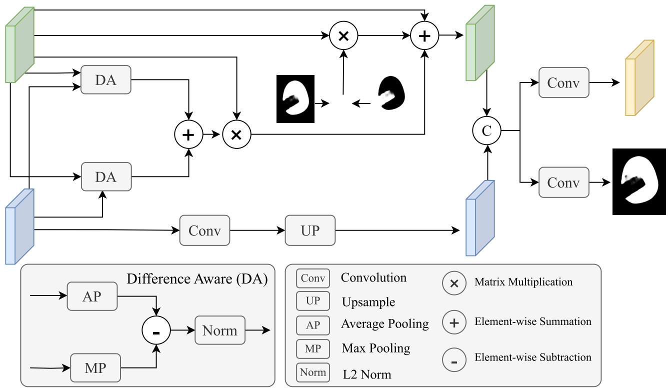


Fig. 7. Structure of the SDF module. We use foreground or background maps and combine with difference map to fuse features.

double by $Proj$. By progressively aggregating, we can get $X_f^{3,4}$. We then enhance $X_f^{3,4}$ and generate X_e by:

$$X_e = \delta(Conv_{1\times1}(X_f^{3,4})) \tag{9}$$

We further apply gating mechanism to X_a,generating X_g by:

$$X_g = \delta(Conv_{1\times1}(X_f^{2,3})) \odot X_a \tag{10}$$

Therefore, the final output X_o is:

$$X_o^{2,3} = X_f^{2,3} + X_g \tag{11}$$

D. Separation-Based Difference Fusion Module

Previous work mainly focus on the foreground (mirror region) while ignoring the fact that people determine the object attributes by comparing surroundings [40]. To establish the connection between mirror and non-mirror regions, imagings and entities, we formulate the SDF module. We can obtain clearer boundaries, better robustness of the model when dealing with occlusions, and thus complete mirror regions. The details of the SDF are shown in Fig. 7, just like a parallel-series structure.

Specifically, MOFP (M^k) shares correspondence with SDF (D^k), where $k \in \{1, 2, 3, 4\}$. When $k = 4$, the SDF focuses on the background mask. When $k \leq 3$, it focuses on the foreground mask and has three inputs: one from the ouput of M^k, two from the ouput of D^{k+1}. As with the illustration of MOFP, we take $k = 2$ as an example. Thus, the input of D^2 are f^3, p^3 (generated by D^3) and $X_o^{2,3}$. We use max pooing (MP) and average pooling (AP) to obtain difference maps Dif_m, which can be expressed as:

$$f_{up}^3 = UP(Conv_{7\times7}(f^3)) \tag{12}$$

$$\begin{aligned} Dif_m = \mathcal{L}_2(AP(X_o^{2,3}) - MP(f_{up}^3))) \\ + \mathcal{L}_2(AP(f_{up}^3)) - MP(X_o^{2,3})) \end{aligned} \tag{13}$$

We then fuse the feature maps, foreground mask and difference map, thus generating X_o', using the following formula:

$$X_o' = Concat((X_o \odot Dif_m \odot p^3), f_{up}^3) \tag{14}$$

Note that when $k = 1$, we use $1 - p^2$, not p^2. Finally, the output of D^2 can be expressed as:

$$f^2 = Conv_{3\times3}(X_o'), \; p^2 = Conv_{7\times7}(f^2) \tag{15}$$

E. Loss Function

We apply supervision to each predicted mask generated by the decoder. Following [9], we employ weighted BCE loss (wBCE) $\mathcal{L}_{BCE}^W$ [41] and weighted IoU loss (wIoU) $\mathcal{L}_{IoU}^W$ [42] to help the model mine difficult and then improve robustness. Thus, the total loss can be expressed as:

$$\mathcal{L}_{total} = \sum_{i=1}^{5} \mathcal{L}_{BCE}^W + \mathcal{L}_{IoU}^W \tag{16}$$

IV. Experiments

A. Datasets

We conduct experiments on four datasets. **Mirror Datasets:** 1) MSD dataset [12] primarily focuses on indoor scenes and consists of 3,063 training and 955 testing images. 2) PMD dataset contains various scenes and multiple objects, with 5,096 training and 571 testing images. 3) Mirror-RGBD dataset [8] includes depth images with higher resolutions and comprises 2,000 training and 1,049 testing images. We use the following dataset to validate the model's generalization ability. **Glass Dataset:** GSD dataset [43] contains 3,285 training and 813 testing images. **Remote Sensing Datasets**: ORSSD [19] consists of 600 training images and 200 testing images. EORSSD [24] is its extended version, with 1400 images for training and 600 for testing. ORSI4199 [44] contains various complex scenarios, such as small objects, with 2000 images for training and 2199 images for testing. **Camouflage Datasets**: CAMO [45] consists of 1000 training images and 250 testing images. COD10K [20], [46] has 3040 images for training and 2026 images for testing. NC4K [47] is only used for testing.

B. Implementation Details

We implement the model and conduct experiments on an A100 GPU via Pytorch. Following [48], we utilize PVT network [35] pretrained on ImageNet as the encoder and employ some image augmentation methods, *e.g.,* horizontal flip. For mirror, glass, and camouflage datasets, following [10], [43], [48], [49], and [50], all inputs are scaled to 384×384. For remote sensing datasets, following [18], all inputs are scaled to 352×352. Note that no post-processing (*e.g.,* CRF) is applied. For training on mirror, remote sensing, and camouflage scenes, we use AdamW [51] as the optimizer with 200 epochs, the initial learning rate of 1e-4 and the batch size of 28. For training on the GSD dataset, we follow [43] and set the epoch to 80 and the batch size to 6. For testing, we do not use tricks such as test-time augmentation.

C. Evaluation Metrics

We adopt six evaluation metrics: S-measure (S_m) [52], mean E-measure (E_m) [53], weighted F-measure, (F_β^w), adaptive F-measure(F_β^{adp}) [54], Intersection over union (IoU), and Mean

TABLE I

QUANTITATIVE COMPARISON ON MSD AND PMD DATASETS WITH FIVE EVALUATION METRICS. S, C, G, M DENOTE SOD, COD, GENERAL SEGMENTATION, MD METHODS RESPECTIVELY. THE BEST PERFORMANCES ARE BOLDED

Methods	Att.	MSD					PMD				
		MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
CPDNet	S	0.116	0.725	0.770	0.625	0.576	0.041	0.779	0.817	0.651	0.600
R3Net	S	0.111	0.723	0.743	0.615	0.554	0.045	0.720	0.756	0.561	0.496
EGNet	S	0.096	0.771	0.776	0.668	0.630	0.088	0.617	0.593	0.362	0.210
PoolNet	S	0.094	0.804	0.831	0.717	0.691	0.089	0.588	0.532	0.313	0.192
MINet	S	0.088	0.792	0.819	0.715	0.664	0.038	0.794	0.822	0.667	0.601
SETR	S	0.071	0.797	0.840	0.750	0.690	0.035	0.753	0.775	0.633	0.564
LDF	S	0.068	0.821	0.867	0.773	0.729	0.038	0.799	0.833	0.683	0.633
VST	S	0.054	0.861	0.901	0.818	0.791	0.036	0.783	0.814	0.639	0.591
FPNet	C	0.042	0.883	0.917	0.849	0.827	0.033	0.823	0.874	0.717	0.673
SAM	G	0.124	–	–	–	0.515	0.052	–	–	–	0.647
MirrorNet	M	0.065	0.850	0.891	0.812	0.790	0.043	0.761	0.841	0.663	0.585
PMDNet	M	0.047	0.875	0.908	0.845	0.815	0.032	0.810	0.859	0.716	0.660
HetNet	M	0.043	0.881	0.921	0.854	0.824	0.029	0.828	0.865	0.734	0.690
SATNet	M	0.033	0.887	0.916	0.865	0.834	0.025	0.826	0.858	0.739	0.684
CSFwinformer	M	0.045	–	–	–	0.821	**0.024**	–	–	–	0.700
Ours	M	**0.033**	**0.904**	**0.934**	**0.888**	**0.866**	0.026	**0.844**	**0.894**	**0.766**	**0.721**

TABLE II

QUANTITATIVE COMPARISON ON ORSSD, EORSSD, AND ORSI4199 DATASETS WITH FIVE EVALUATION METRICS. THE BEST PERFORMANCES ARE BOLDED

Methods	ORSSD				EORSSD				ORSI4199			
	MAE↓	S_m↑	E_m↑	F_β^{adp}↑	MAE↓	S_m↑	E_m↑	F_β^{adp}↑	MAE↓	S_m↑	E_m↑	F_β^{adp}↑
ICON	0.012	0.926	0.964	0.844	0.007	0.918	0.962	0.806	0.028	0.875	0.944	0.853
HFANet	0.009	0.940	0.971	0.882	0.007	0.938	0.968	0.836	0.031	0.877	0.934	0.832
ASTT	0.009	0.935	0.970	–	0.006	0.925	0.958	–	–	–	–	–
GLGCNet	0.007	0.949	0.982	0.893	0.006	0.937	0.976	0.850	0.027	0.884	0.947	0.867
Ours	**0.006**	**0.952**	**0.986**	**0.910**	**0.004**	**0.942**	**0.978**	**0.872**	**0.024**	**0.891**	**0.954**	**0.885**

TABLE III

QUANTITATIVE COMPARISON ON CAMO, COD10K, AND NC4K DATASETS WITH FIVE EVALUATION METRICS. THE BEST PERFORMANCES ARE BOLDED

Methods	CAMO				COD10K				NC4K			
	MAE↓	S_m↑	E_m↑	F_β^w↑	MAE↓	S_m↑	E_m↑	F_β^w↑	MAE↓	S_m↑	E_m↑	F_β^w↑
PoPNet	0.073	0.806	0.859	0.743	0.031	0.827	0.910	0.757	0.043	0.852	0.909	0.802
ZoomNet	0.066	0.820	0.892	0.752	0.029	0.838	0.911	0.729	0.043	0.853	0.912	0.784
FEDER	0.066	0.836	0.897	–	0.029	0.844	0.911	–	0.042	0.862	0.913	–
FSPNet	0.050	0.856	0.899	0.799	0.026	0.851	0.895	0.735	0.035	0.879	0.915	0.816
Ours	**0.046**	**0.865**	**0.930**	**0.830**	**0.025**	**0.854**	**0.926**	**0.772**	**0.033**	**0.880**	**0.934**	**0.838**

Absolute Error (MAE). Note that the higher the better for the first five. We also utilize more intuitive Precision-Recall (P-R) and F-measure curves.

D. Comparison With SOTA Methods

1) Quantitative Comparison: We compare with SOTA methods on three classes of four datasets to validate the superiority and generalization of our method, as shown in Tables I, II, III, IV, and V. Specifically, for MSD and PMD datasets, we select eight SOD methods, *i.e.,* CPDNet [55], R3Net [56], EGNet [57], PoolNet [58], MINet [59], SETR [60], LDF [61], VST [11], one COD method, *i.e.,* FPNet [49], three general segmentation model, *i.e.,* SAM [31] and five MD methods, *i.e.,* MirrorNet [12], PMDNet [10], HetNet [9], SATNet [7], and CSFwinformer [62]. Our method outperforms various types of detection models, and

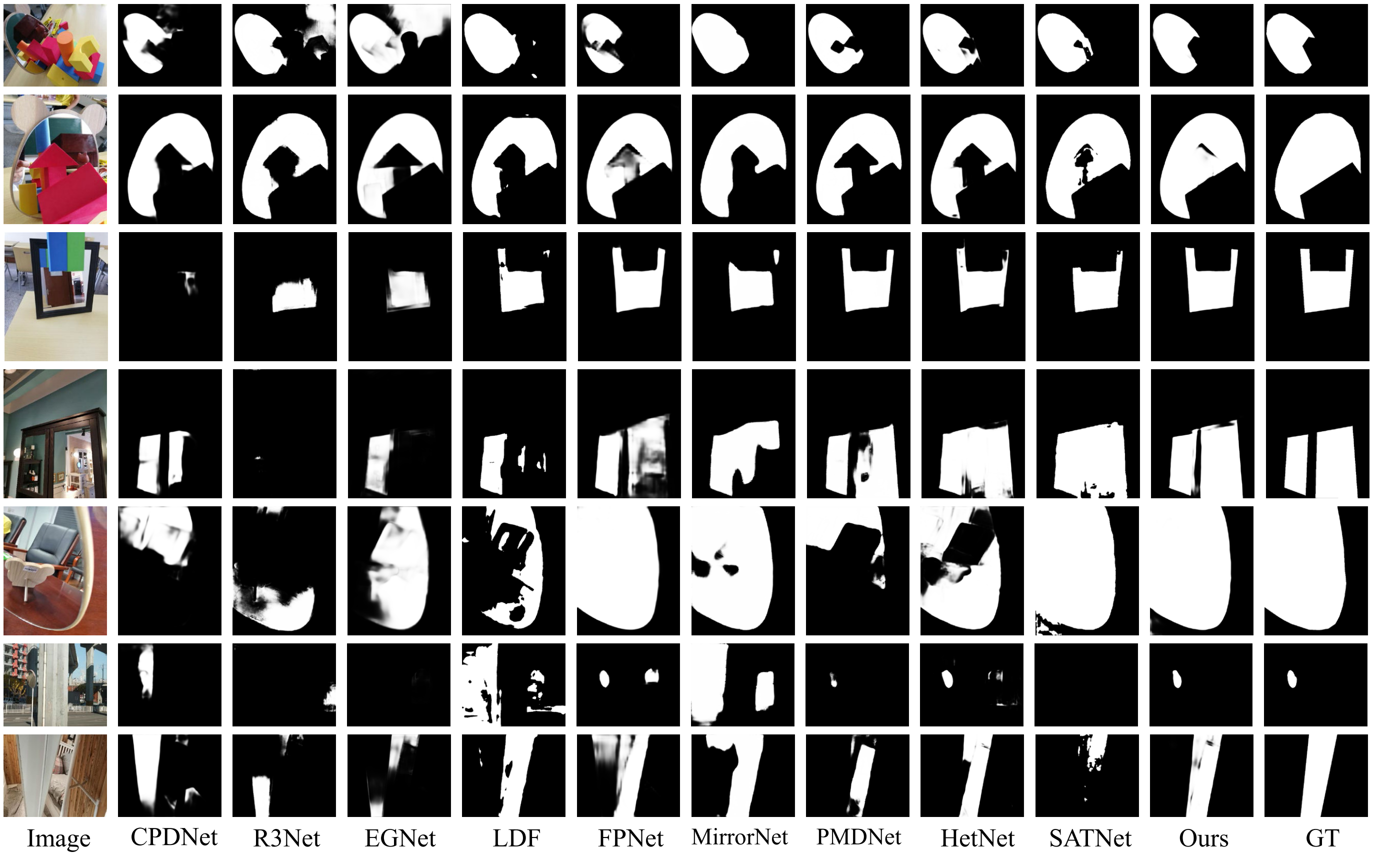


Fig. 8. Qualitative comparison on MSD and PMD datasets. The first three rows show occlusion scenes, the fourth row presents the multi-mirror regions and close to each other scene, and the last three rows show scale variation scenes.

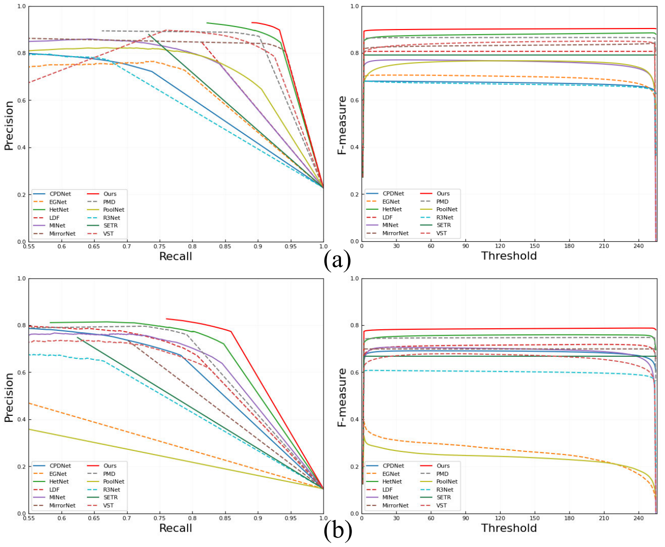


Fig. 9. P-R curves and F-measure curves tested by our method, eight SOD models, and three MD methods on two datasets, *i.e.,* (a) MSD and (b) PMD datasets. Zoom in for better view.

in particular outperforms the frequency-aware FPNet by a significant margin, demonstrating that our frequency-guided DPRNet considers the task attributes of MD better. For Mirror-RGBD dataset, seven multimodal SOD methods *i.e.,* A2dele [63], HDFNet [64], S2MA [65], DANet [66], JL-DCF [67], VST [11], BBSTNet [68], and one multimodal MD method *i.e.,* PDNet [8] are used for comparison. Although our method does not use depth information, it still outperforms some multimodal SOD methods and is close to the PDNet. Note that our DPRNet is not a multimodal method. For GSD dataset, we choose one SOD method *i.e.,* BASNet [69],

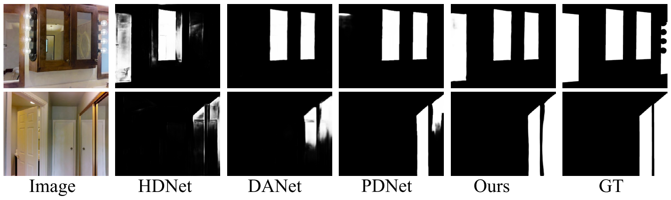


Fig. 10. Qualitative comparison on Mirror-RGBD, showing multi-mirror regions and close to each other scenes.

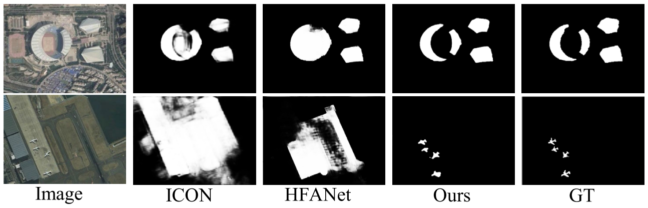


Fig. 11. Qualitative comparison on remote sensing scenes.

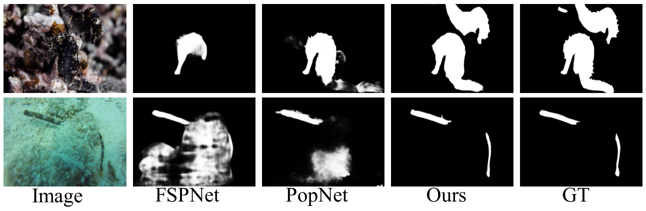


Fig. 12. Qualitative comparison on camouflage scenes.

TABLE IV

Quantitative Comparison on Mirror-RGBD Dataset With Five Evaluation Metrics. D. Represents Depth Map. The Best Performances Are Bolded

Methods	D.	Mirror-RGBD				
		MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
A2dele	✓	0.120	0.641	0.730	0.505	0.428
HDFNet	✓	0.095	0.671	0.663	0.521	0.447
S2MA	✓	0.075	0.765	0.797	0.646	0.609
DANet	✓	0.063	0.800	0.842	0.728	0.678
JL-DCF	✓	0.057	0.815	0.861	0.750	0.696
VST	✓	0.054	0.815	0.859	0.751	0.702
BBSTNet	✓	0.048	0.840	0.881	0.786	0.743
PDNet	✓	**0.042**	**0.856**	**0.906**	**0.825**	**0.778**
Ours		0.047	0.845	0.899	0.811	0.761

TABLE V

Quantitative Comparison on GSD Dataset With Three Evaluation Metrics. The Best Performances Are Bolded

Methods	GSD		
	MAE↓	F_β↑	IoU↑
BASNet	0.106	0.808	69.79
SINet	0.077	0.875	77.04
TransLab	0.088	0.837	74.05
GDNet	0.069	0.869	79.01
GSDNet	0.055	0.903	83.64
PGSNet	0.054	0.868	83.65
GlassSemNet	**0.044**	**0.920**	**85.60**
Ours	0.049	0.918	84.91

TABLE VI

Model Efficiency Comparison. We Compare With Four MD SOTA Models on Parameters (M), FLOPs (GMAC). The Best Performances Are Bolded

Methods	Input Size	FLOPs↓	Params.↓
MirrorNet	384×384	77.73	121.77
PMDNet	384×384	101.54	147.66
SATNet	512×512	153.00	139.36
CSFwinformer	512×512	139.45	150.54
Ours	384×384	**30.47**	**31.19**

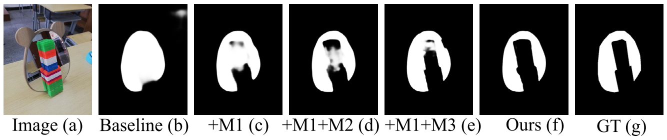


Fig. 13. Qualitative ablation. With more modules are added, our model can remove occlusion and generate accurate boundary.

one COD method *i.e.,* SINet [20] and five glass detection methods *i.e.,* TransLab [70], GDNet [71], GSDNet [43], PGSNet [50], and GlassSemNet [72]. For ORSSD, EORSSD and ORSI4199 datasets, we choose four remote sensing SOD methods, *i.e.,* ICON [73], HFANet [74], ASTT [75], and GLGCNet [18]. For CAMO, COD10K and NC4K datasets, ZoomNet [76], FEDER [77], FSPNet [78], and PopNet are chosen. Our method still achieves impressive results and proves to generalize well. In addition, compared with MD methods such as SATNet, our method has lower model and computational complexity, as shown in Table VI. Note that larger input size indicates higher FLOPs.

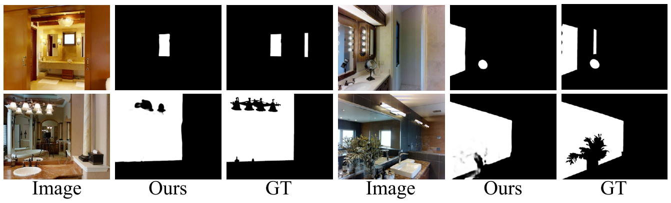


Fig. 14. Failure cases.

2) Qualitative Comparison: Some representative examples on the three mirror datasets are selected for visualization and comparison. As shown in Fig. 8, the first three rows represent irregular and regular occlusion scenarios. Our method can effectively distinguish entities and imagings to obtain complete mirror regions. The fourth row represents multi-target scenes, Our method can recognize all mirror regions and establish clear boundaries. The last three rows show scale variation scenarios. Our method can capture local and global dependencies with good robustness. As shown in Fig. 9, our method also exhibits promising performance in terms of the P-R and F-Measure curves. As shown in Fig. 10, we do not utilize depth information and use small input size, but achieve promising results in challenging scenarios. We also provide visualizations on remote sensing and camouflage scenarios, as shown in Fig. 11 and 12.

E. Ablation Study

We validate the effect of each module on MSD and PMD datasets and provide visualization, as shown in Tables VII, VIII, IX, X, and Fig. 13.

1) Effect of the FIA Module: As shown in Fig. 13(b) and (c), Baseline initially locates the position of mirror region, but can not handle occlusion and has some mislocalized pixels (upper right corner). By adding the FIA moudle, our model distinguishes mirror and non-mirror regions, initially recognizes occlusion entities, lays the foundation for subsequent refinement, and eliminates erroneous pixels. Compared with Baseline, it improves by 0.5%, 0.9%, 1.2%, 1.4%, and 1.6% for the S_m, E_m, F^w_β and IoU metrics on the MSD dataset, respectively.

2) Effect of the MOFP Module: We evaluate the performance of MOFP module based on "Baseline+FIA" model, with 0.4%, 1.4%, 0.5%, 1.1%, and 1.4% improvement on the five metrics for the MSD dataset, respectively. As shown in Fig. 13 (d), with the addition of the MOFP module, occlusion problem is further alleviated, and other mirror regions are not affected, indicating that the MOFP moudule can establish long and short-range feature dependencies to handle irregular foregrounds and explore the correlation of adjacent features.

TABLE VII

QUANTITATIVE ABLATION ON MSD AND PMD DATASETS. M1, M2, AND M3 INDICATE THE FIA MODULE, MOFP MODULE, AND SDF MODULE, RESPECTIVELY

Method	Component			MSD					PMD				
	M1	M2	M3	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
I				0.048	0.873	0.905	0.849	0.817	0.045	0.812	0.866	0.738	0.678
II	✓			0.043	0.882	0.917	0.863	0.833	0.039	0.823	0.875	0.747	0.695
III		✓		0.042	0.880	0.919	0.857	0.835	0.040	0.825	0.871	0.743	0.697
IV			✓	0.044	0.886	0.911	0.864	0.829	0.038	0.821	0.870	0.748	0.694
V	✓	✓		0.039	0.896	0.922	0.874	0.847	0.032	0.831	0.881	0.754	0.709
VI	✓		✓	0.038	0.891	0.924	0.871	0.850	0.030	0.829	0.885	0.751	0.713
VII		✓	✓	0.040	0.895	0.925	0.873	0.848	0.030	0.833	0.883	0.755	0.708
Ours	✓	✓	✓	**0.033**	**0.904**	**0.934**	**0.888**	**0.866**	**0.026**	**0.844**	**0.894**	**0.766**	**0.721**

TABLE VIII

QUANTITATIVE ABLATION ON MSD AND PMD DATASETS. M1, M2, M3, AND M4 INDICATE THE WAVELET, FOURIER, DISCRETE COSINE, AND LAPLACE TRANSFORMS, RESPECTIVELY

Method	Component				MSD					PMD				
	M1	M2	M3	M4	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
I	✓				0.031	0.897	0.928	**0.891**	0.859	**0.022**	0.838	0.891	0.761	0.717
II		✓			0.035	0.901	0.933	0.884	0.861	0.030	0.840	**0.898**	0.763	0.719
III			✓		**0.030**	0.900	0.927	0.881	0.857	0.028	0.840	0.888	0.759	**0.723**
Ours				✓	0.033	**0.904**	**0.934**	0.888	**0.866**	0.026	**0.844**	0.894	**0.766**	0.721

TABLE IX

QUANTITATIVE ABLATION ON MSD DATASET. M1, M2 INDICATE THE INTERACTION MODULE AND ALIGNMENT MODULE, RESPECTIVELY

Method	M1	M2	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
I	✓		0.045	0.880	0.909	0.858	0.827
II		✓	0.048	0.876	0.900	0.853	0.823
Ours	✓	✓	**0.043**	**0.882**	**0.917**	**0.863**	**0.833**

TABLE X

QUANTITATIVE ABLATION ON MSD DATASET. M1, M2 INDICATE THE DIFFERENCE MODULE AND SEPARATION MODULE, RESPECTIVELY

Method	M1	M2	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
I	✓		0.046	0.879	0.901	0.855	0.823
II		✓	0.045	0.882	0.909	**0.865**	0.826
Ours	✓	✓	**0.044**	**0.886**	**0.911**	0.864	**0.829**

3) Effect of the SDF Module: Similar to evaluating the performance of the MOFP module, we conduct further experiments based on the "Baseline+FIA" model to validate the effect of the SDF module. As shown in Fig. 13 (e), with the SDF module separating foreground and background and establishing the association between entities and imagings, occlusion problem has been basically resolved. The five metrics have shown improvements of 0.5%, 0.9%, 0.7%, 0.8%, and 1.7% on the MSD dataset. However, the mirror region in the upper right corner has been affected, thus it is necessary to combine the MOFP module.

4) Combination of Three Modules: Each module has its own role, and only by combining them can we maintain accurate and complete mirror regions in various scenarios, such as occlusions and scale changes, as shown in Fig. 13 (f). When combined, the performance improves by 1.5%, 3.1%, 2.9%, 3.9%, and 4.9% on the MSD dataset, respectively, and the detection result is close to the ground truth (GT).

5) Different Frequency Decomposition Methods: As shown in Table VIII, we use various frequency decomposition methods to obtain the high-frequency components of an image, and the Laplacian method achieves better performance. By combining features from different scales, we can obtain better high-frequency representations.

6) Effect of Interaction Component of the FIA Module: As shown in Table IX, the interaction component achieves information complementarity by exchanging pixel-level information between high-frequency and low-frequency features. However, the frequency differences inevitably result in feature misalignment.

7) Effect of Aliment Component of the FIA Module: As shown in Table IX, we utilize the alignment component to reduce the differences between high-frequency and low-frequency domains, but its improvement is weaker compared to using the interaction component alone. By combining both components, the performance is further enhanced. It can be observed that frequency interaction forms the basis for achieving alignment.

8) Effect of Difference-Awre Component of the SDF Module: As shown in Table X, we apply difference-aware component to emphasizes selective difference perception through pooling, capturing differences from both low-high and high-low perspectives. Despite the overall performance improvement compared to the baseline, there is a decrease of 0.4% in E_m. This may be attributed to the neglect of some insignificant (detail) but critical differences between features.

9) Effect of Separation-Awre Component of the SDF Module: As shown in Table X, we utilize the separation component to achieve foreground-background separation. Based on this, we further incorporate the difference component to hierarchically capture different feature differences while compensating for fine-grained perception. It can be observed that the combination of both components enables complementarity and further improves performance.

F. Failure Samples and Futuer Work

As shown in Fig. 14, our method does not perform well in scenes with severe irregular occlusions such as chandeliers, flowers and slender mirrors. In the future, we will address these from two aspects: 1) By applying uncertainty perception, we can establish probability pixel distributions to locate problematic areas, enabling us to fully exploit fine-grained or irregular features. 2) We will design more appropriate loss functions to enhance the ability of modeling elongated features.

V. Conclusion

In this paper, we rethink MD frameworks and propose the DPRNet. Motivated by our findings of the frequency difference of mirror and non-mirror regions, we use Laplace pyramid and vision Transformer to decouple and obtain high and low frequency features, respectively. Then the FIA module is proposed to eliminate the frequency differences and locate the target region initially. We further design the MOFP module to aggregate adjacen features to tackle mirror size variations. Finally, we propose the SDF module to facilitate the separation of foreground and background and establish connections between entities and imagings. Extensive experiments demonstrates that our DPRNet outperforms the SOTA method by an average of 3%, with only about one-fifth of the parameters and FLOPs.

References

[1] L. Chen, M. Li, Y. Duan, J. Zhou, and J. Lu, "Uncertainty-aware representation learning for action segmentation," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, vol. 2, Jul. 2022, p. 6.

[2] Y. Tang, T. Chen, X. Jiang, Y. Yao, G.-S. Xie, and H.-T. Shen, "Holistic prototype attention network for few-shot video object segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jul. 18, 2023, doi: 10.1109/TCSVT.2023.3296629.

[3] B. Song, J. Zhou, X. Chen, and S. Zhang, "Real-scene reflection removal with RAW-RGB image pairs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3759–3773, Aug. 2023.

[4] X. Zhang, K. Xing, Q. Liu, D. Chen, and Y. Yin, "Single image reflection removal based on dark channel sparsity prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6431–6442, Nov. 2023.

[5] Z. He et al., "Learning depth representation from RGB-D videos by time-aware contrastive pre-training," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4143–4158, Jun. 2024.

[6] L. Wang, Z. He, R. Dang, H. Chen, C. Liu, and Q. Chen, "RES-StS: Referring expression speaker via self-training with scorer for goal-oriented vision-language navigation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3441–3454, Jul. 2023.

[7] T. Huang, B. Dong, J. Lin, X. Liu, R. W. Lau, and W. Zuo, "Symmetry-aware transformer-based mirror detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 935–943.

[8] H. Mei et al., "Depth-aware mirror segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3044–3053.

[9] R. He, J. Lin, and R. W. H. Lau, "Efficient mirror detection via multi-level heterogeneous learning," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 1, pp. 790–798.

[10] J. Lin, G. Wang, and R. W. H. Lau, "Progressive mirror detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3694–3702.

[11] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4722–4732.

[12] X. Yang, H. Mei, K. Xu, X. Wei, B. Yin, and R. Lau, "Where is my mirror?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8808–8817.

[13] H. Guan, J. Lin, and R. W. H. Lau, "Learning semantic associations for mirror detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5931–5940.

[14] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 2, 2022, pp. 2071–2081.

[15] J. Liang, H. Zeng, and L. Zhang, "High-resolution photorealistic image translation in real-time: A Laplacian pyramid translation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9387–9395.

[16] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li, "Pyramid grafting network for one-stage high resolution saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11707–11716.

[17] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, 2020.

[18] Z. Bai, G. Li, and Z. Liu, "Global–local–global context-aware network for salient object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 198, pp. 184–196, Apr. 2023.

[19] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9156–9166, Nov. 2019.

[20] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2020, pp. 2777–2787.

[21] X. Li, J. Yang, S. Li, J. Lei, J. Zhang, and D. Chen, "Locate, refine and restore: A progressive enhancement network for camouflaged object detection," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, 2023, pp. 1116–1124.

[22] Y. Qiu, Y. Liu, L. Zhang, H. Lu, and J. Xu, "Boosting salient object detection with transformer-based asymmetric bilateral U-Net," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2332–2345, Apr. 2024.

[23] Z. Xie et al., "Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4149–4163, Aug. 2023.

[24] Q. Zhang et al., "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2020.

[25] R. Yan et al., "Global–local semantic interaction network for salient object detection in optical remote sensing images with scribble supervision," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[26] B. Xu, H. Liang, W. Gong, R. Liang, and P. Chen, "A visual representation-guided framework with global affinity for weakly supervised salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 248–259, Jan. 2024.

[27] K. Wang, Z. Tu, C. Li, C. Zhang, and B. Luo, "Learning adaptive fusion bank for multi-modal salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Mar. 14, 2024, doi: 10.1109/TCSVT.2024.3375505.

[28] Y. Wang, H. Chen, Y. Zhang, G. Li, and T. Gao, "Joint space–frequency for saliency detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[29] H. Zhang et al., "MP-former: Mask-piloted transformer for image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18074–18083.

[30] F. Li et al., "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3041–3050.

[31] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[32] J. Hao, M. Liu, and K. F. Hung, "GEM: Boost simple network for glass surface segmentation via segment anything model and data synthesis," 2024, *arXiv:2401.15282*.

[33] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[34] N. Park and S. Kim, "How do vision transformers work?" in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–26.

[35] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[36] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[39] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[40] J. Wagemans et al., "A century of gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization," *Psychol. Bull.*, vol. 138, no. 6, pp. 1172–1217, 2012.

[41] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[42] G. Mattyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3438–3446.

[43] J. Lin, Z. He, and R. W. H. Lau, "Rich context aggregation with reflection prior for glass surface detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13410–13419.

[44] Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, and B. Luo, "ORSI salient object detection via multiscale joint region and boundary model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[45] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Comput. Vis. Image Understand.*, vol. 184, pp. 45–56, Jul. 2019.

[46] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, Oct. 2022.

[47] Y. Lv et al., "Simultaneously localize, segment and rank the camouflaged objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 11591–11601.

[48] M. Zha et al., "Weakly-supervised mirror detection via scribble annotations," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 7, pp. 6953–6961.

[49] R. Cong, M. Sun, S. Zhang, X. Zhou, W. Zhang, and Y. Zhao, "Frequency perception network for camouflaged object detection," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 1179–1189.

[50] L. Yu et al., "Progressive glass segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 2920–2933, 2022.

[51] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[52] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

[53] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," 2018, *arXiv:1805.10421*.

[54] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

[55] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3907–3916.

[56] Z. Deng et al., "R3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Palo Alto, CA, USA, Jul. 2018, pp. 684–690.

[57] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8779–8788.

[58] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3917–3926.

[59] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9413–9422.

[60] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[61] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13025–13034.

[62] Z. Xie, S. Wang, Q. Yu, X. Tan, and Y. Xie, "CSFwinformer: Cross-space-frequency window transformer for mirror detection," *IEEE Trans. Image Process.*, vol. 33, pp. 1853–1867, 2024.

[63] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 9060–9069.

[64] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 235–252.

[65] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13756–13765.

[66] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 646–662.

[67] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3052–3062.

[68] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[69] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

[70] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, "Segmenting transparent objects in the wild," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Aug. 2020, pp. 696–711.

[71] H. Mei et al., "Don't hit me! Glass detection in real-world scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3684–3693.

[72] J. Lin, Y.-H. Yeung, and R. Lau, "Exploiting semantic relations for glass surface detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 22490–22504.

[73] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3738–3752, Mar. 2022.

[74] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624915.

[75] L. Gao, B. Liu, P. Fu, and M. Xu, "Adaptive spatial tokenization transformer for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602915.

[76] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2160–2170.

[77] C. He et al., "Camouflaged object detection with feature decomposition and edge reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22046–22055.

[78] Z. Huang, H. Dai, S. Wang, T. Xiang, H. Chen, and J. Qin, "Feature shrinkage pyramid for camouflaged object detection with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5557–5566.

Mingfeng Zha is currently pursuing the master's degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include computer vision and image/video processing.

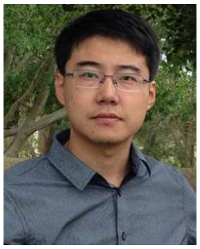
Feiyang Fu is currently pursuing the bachelor's degree with the Yingcai Honors College, University of Electronic Science and Technology of China. His research interests include deep learning, computer vision, and cross-modal learning.

Yunqiang Pei is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include mixed reality, physiological computing, and human–computer interaction.

Guoqing Wang (Member, IEEE) received the Ph.D. degree from The University of New South Wales, Australia, in 2021. He is currently with the School of Computer Science and Engineering, University of Electronic of Science and Technology of China. He has authored or co-authored more than 40 scientific articles at top venues, including IJCV, IEEE TRANSACTIONS ON IMAGE PROCESSING (IEEE TIP), IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (IEEE TIFS), ICCV, and ACM MM. His research work with UNSW has been recognized as the Australian Dean's Award for Outstanding Ph.D. Theses. His research interests include machine learning and unmanned systems, with a special emphasis on cognition and embodied agents.

Tianyu Li received the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, USA, in 2023. He is currently a Post-Doctoral Researcher with the Center for Future Media, University of Electronic Science and Technology of China, Chengdu, China. His research focuses on image processing using statistical models and deep learning models and time series forecasting and anomaly detection in time series.

Xiongxin Tang received the Ph.D. degree from the University of Chinese Academy of Sciences in 2013. She currently holds the position of a Vice Professor with the Institute of Software, Chinese Academy of Sciences. With extensive experience in optical simulation calculations and optical software development, she has actively engaged in research and development in this field. Her involvement includes National Key Research and Development Projects, the 173 Project, Key Project Sub-Project of the National Natural Science Foundation, Key Laboratory Fund for Equipment Pre-Research, and Innovation Projects of Chinese Academy of Sciences. She has made significant contributions to her field with over 20 published academic papers in reputable domestic and international journals and conferences. Furthermore, she has also been granted 25 invention patents.



Yang Yang (Senior Member, IEEE) received the bachelor's degree in computer science from Jilin University, Changchun, China, in 2006, the master's degree in computer science from Peking University, Beijing, China, in 2009, and the Ph.D. degree in computer science from The University of Queensland, Brisbane, QLD, Australia, in 2012. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Heng Tao Shen (Fellow, IEEE) received the B.Sc. (Hons.) and Ph.D. degrees from the Department of Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively. His research interests include multimedia search, computer vision, artificial intelligence, and big data management. He is a member of Academia Europaea and a fellow of ACM and OSA.