Heterogeneous Experts and Hierarchical Perception for Underwater Salient Object Detection

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

***Abstract*—Existing underwater salient object detection (USOD) 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, formulating the heterogeneous experts and hierarchical perception network (HEHP). Specifically, to reduce modal discrepancies, we propose the hierarchical prototype guided interaction (HPI), which achieves fine-grained alignment guided by the semantic prototypes, and then refines with complementary modalities. We further design the mixture of frequency experts (MoFE), where experts focus on modeling high- and low-frequency respectively, collaborating to explicitly obtain hierarchical representations. To efficiently integrate diverse spatial and frequency information, we formulate the four-way fusion experts (FFE), which dynamically selects optimal experts for fusion while being sensitive to scale and orientation. Since depth maps with poor quality inevitably introduce noises, we design the uncertainty injection (UI) to explore high uncertainty regions by establishing pixel-level probability distributions. We further formulate the holistic prototype contrastive (HPC) loss based on semantics and patches to learn compact and general representations across modalities and images. Finally, we employ varying supervision based on branch distinctions to implicitly construct difference modeling. Extensive experiments on two USOD datasets and four relevant underwater scene benchmarks validate the effect of the proposed method, surpassing state-of-the-art binary detection models. Impressive results on seven natural scene benchmarks further demonstrate the scalability.**

***Index Terms*—Multimodal fusion, underwater perception, expert learning, uncertainty guidance.**

Received 18 December 2024; revised 21 April 2025; accepted 14 May 2025. Date of publication 2 June 2025; date of current version 16 June 2025. This work was supported in part by the National Natural Science Foundation of China under Grant U23B2011, Grant 62102069, Grant U20B2063, and Grant 62220106008; in part by the Key Research and Development Program of Zhejiang under Grant 2024SSYS0091; and in part by Sichuan Science and Technology Program under Grant 2024NSFTD0034. The associate editor coordinating the review of this article and approving it for publication was Dr. Tiesong Zhao. *(Corresponding author: Guoqing Wang.)*

Mingfeng Zha, Guoqing Wang, Yunqiang Pei, Tianyu Li, and Yang Yang are with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: zhamf1116@gmail.com; gqwang0420@uestc.edu.cn; simonyqpei@qq.com; cosmos.yu@hotmail.com; yang.yang@uestc.edu.cn).

Xiongxin Tang is with the Institute of Software, Chinese Academy of Science, Beijing 100190, China (e-mail: xiongxin@iscas.ac.cn).

Chongyi Li is with the School of Computer Science, Nankai University, Tianjin 300350, China (e-mail: lichongyi25@gmail.com).

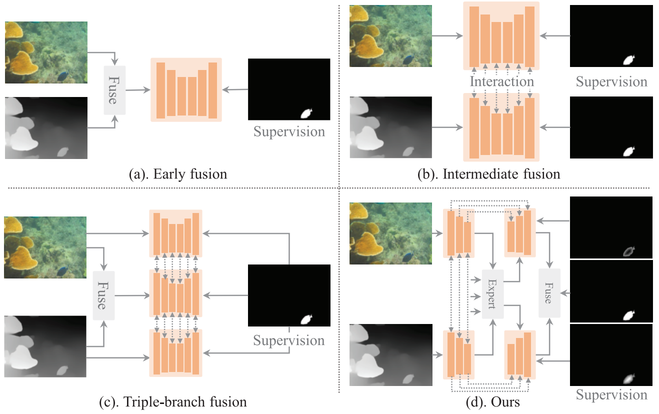
Heng Tao Shen is with the School of Computer Science and Technology, Tongji University, Shanghai 201804, China, also with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: shenhengtao@hotmail.com).

Digital Object Identifier 10.1109/TIP.2025.3572760

I. Introduction

DIFFERENT from common semantic segmentation or object detection, underwater salient object detection (USOD) aims to overcome complex hydrological interferences, such as low light and uneven illumination, to locate visually compelling regions, which is crucial for tasks like image restoration/generation [1], [2] and path planning [3], [4]. Currently, mainstream USOD approaches rely on depth map assistance, *i.e.,* multi-modal learning. We categorize them into four types based on fusion strategies: 1) Single-stream frameworks [5], relying on image-level fusion, although lightweight, lack consideration of modal differences, inevitably introducing significant noise and feature misalignment. 2) Dual-stream frameworks [6], setting up separate encoding-decoding networks for each modality with interactions at each stage. Clearly, excessive interaction may lead to feature redundancy and assimilation, making it difficult to highlight modal representation differences and escalating model complexity. 3) Triple-branch frameworks [7], integrating strategies 1 and 2, generate additional modality through image-level fusion and establish three interacting sub-networks aiming to align at both image and feature levels. Despite achieving impressive performance, the high computational cost raises obstacles for deployment on underwater devices. Therefore, we propose the heterogeneous experts and hierarchical perception (HEHP) framework, *i.e.,* strategy 4, where: a) We abandon coarse-grained image fusion in favor of fine-grained feature interactions; 2) We only consider encoding-side interactions to obtain high-dimensional representations and maximize semantic difference preservation; 3) We decouple features in frequency and achieve modal alignment based on the heterogeneous mixture of experts (MoE); 4) We assign different supervision to each branch, generating diverse representations by adjusting learning directions, reducing learning pressure by having branches responsible for specific regions only, rather than all. Based on the above, our strategy can achieve a better balance between performance and efficiency. Hence, we wonder two questions. *1) Why utilize heterogeneous experts learning? 2) Why and how to design hierarchical perception?*

We answer the first question. Underwater scenes are more complex and diverse compared to natural scenes. Some approaches often involve customizing functional components to address specific issues, requiring domain knowledge and limiting generalizability. By combining components in a



(a). Early fusion

(b). Intermediate fusion

(c). Triple-branch fusion

(d). Ours

Fig. 1. Architecture comparisons of USOD/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 HEHP, exploring the low- and high-frequency, trunk and detail features, respectively.

sequential or parallel manner, dividing representations somewhat expands the solution space. Limited resources and potential functional overlap or conflicts among components may lead to degradation, *i.e.,* performance does not increase proportionally with the number of components and even decrease. Thus, we turn to leveraging MoE to avoid intricate component design and combinations. On one hand, MoE can dynamically adjust based on scene-specific characteristics, such as allocating experts for lighting adjustment and scale perception according to the input, *e.g.,* low-light conditions with large objects. On the other hand, during the inference stage, we only need to select a subset of highly responsive experts, reducing computational costs. Due to modal differences, we require heterogeneous experts to handle domain-specific representations and fusion. Heterogeneity is reflected in: 1) Expert inputs; 2) The internal design of experts; 3) The balanced weight distribution in the routing network. Intuitively, MoE significantly increases the model complexity. Therefore, we map representations to the rank space to alleviate.

We answer the second question. Previous works [6], [8], [9] focus more on exploring how to effectively fuse modal information but overlook the inherent differences between modalities. The RGB modal contains rich low-level information such as textures and edges, while the depth modal highlights positions based on differences in pixel depths, leaning towards semantics. Moreover, most works tend to independently utilize the overall data of single image (modal), lacking consideration for decoupling, pairwise, and historical representations. We aim to construct the fine-grained and comprehensive alignment. Specifically, our hierarchical modeling comprises four aspects: 1) Each representation consists of high- and low- frequency, where we decouple features and generate hierarchical region prototypes to aggregate dual-domain pixels; 2) We modulate the high- and low- frequency of RGB and depth features separately to explicitly learn modality differences; 3) We construct three-scale alignment for foreground and background, encompassing global semantic, regional semantic, and patch dimensions, as well as inter-modality and intra-batch dimensions, to acquire compact and highly generalizable representations; 4) We utilize binary, detail, and trunk maps to supervise three branches, implicitly differentiating modal representations.

In summary, our insights are based on three observations: 1) RGB and depth modalities differ inherently, allowing decoupling in frequency characteristics and supervision signals; 2) Fixed models learning general representations may struggle to capture novel, scene-specific visual distribution patterns; 3) Under imaging noise conditions, direct and coarse-grained feature interaction may hinder rather than facilitate cross-modal complementarity.

Technically, 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. We propose the hierarchical prototype guided interaction (HPI) that achieves calibration from local-global-local perspectives. We leverage prototype clustering, mining common features to reduce the noise interference. Through progressive calibration and refinement of dual spaces (space and channel), we can effectively learn modality-shared and modality-specific knowledge. We further decode and decouple the frequency, combining separately with RGB and depth features, modulated by the mixture of frequency experts (MoFE). To dynamically and comprehensively integrate multi-domain features, we introduce the four-way fusion experts (FFE). Within the experts, we propose the continuous and adjacent feature aggregation approach to combine representations (this approach can narrow the gap between details and semantics, and is also applied in our decoder), along with scale and orientation modeling. The introduction of depth errors is particularly harmful to weak or fine-grained regions representations, *e.g.,* small objects and irregular regions. We design the uncertainty injection (UI) that employs Bayesian learning to model the probability distributions and thus determine high uncertainty regions. Compared to pixel-level unstable contrasts, we construct paired and unpaired prototype contrasts, bringing similar prototypes closer and distancing dissimilar/other sample prototypes, forming the holistic prototype contrastive (HPC) loss. Inspired by [10], we disentangle ground truth maps 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. We utilize them as supervision for three separate branches to prevent identical supervision from inducing representations homogenization.

Our contributions are as follows:

- We rethink existing frameworks and propose the HEHP based on expert and hierarchical learning to address complex underwater scenarios and efficiently achieve modality alignment and fusion.
- We propose the HPI to enhance region representations and achieve two-stage feature interactions. We design the MoFE and the FFE to dynamically select the optimal modulation and fusion expert groups. We introduce the UI to locate highly uncertain areas, and formulate the HPC loss to generate accurate and information-dense

representations. We decouple supervision based on branch characteristics.
- Without relying on additional data or large models, our proposed HEHP surpasses state-of-the-art approaches on six underwater and seven natural benchmarks.

II. Related Work

A. Salient Object Detection

Based on development process, SOD methods can be divided into two categories. The first category involves handcrafted feature design or the incorporation of prior features, *e.g.,* background [11] and center [12]. Despite the effective performance of expert knowledge in specific scenarios, it requires reconfiguration for new contexts and struggles to transfer to broader and more complex visual patterns. The second category of data-driven paradigms utilizes convolutional neural networks, Transformer series, or combination of both to extract high-dimensional features. 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) Improving image resolution to enrich initial visual information [13]; 2) Attempting to introduce additional visual cues [14], [15] as guidance or supplementation to improve detection performance and robustness. For example, leveraging the depth differences can effectively locate the approximate edges or contours of people or objects, laying the foundation for further SOD. However, existing methods primarily rely on full-stage and high-dimensional embedding spaces for lossless and sufficient feature interaction, making them susceptible to interference from low-quality samples and challenging for practical deployment. We aim to implement at the encoder level in the high information density rank space to ensure higher noise tolerance and computational efficiency.

B. RGB-D Salient Object Detection

RGB-D SOD has been extensively explored in natural scenes, but research in underwater scenarios is limited. In natural scenes, Cong et al. [8] leveraged the advantages of CNN-based local modeling and Transformer's long-range dependencies. Hu et al. [6] designed a unified feature encoding, fusion, and decoding network to avoid integrating additional components. Yin et al. [16] pre-trained the backbone network using image-depth pairs from ImageNet-1K, endowing with the ability to encode RGB-D representations. These methods excel with well-aligned, high-quality data but struggle when depth maps are noisy or misaligned. Moreover, the paired underwater datasets are limited, causing pre-training to be inefficient. In underwater scenes, challenges such as poor lighting remain. Islam et al. constructed [17] the first USOD dataset, but due to its small scale, it is ineffective in validating the efficacy of methods. To address this, Hong et al. [18] built the first large-scale dataset, *i.e.,* USOD10K, encompassing various scenes and equipped with depth maps. Recently, Jin et al. [19] devised a curriculum learning-based framework, considering the difficulty differences of different training samples in two phases. However, current methods lack consideration of modality characteristics and struggle to adjust based on the input (dominance of information from different modalities or equal weighting) by designing fixed components for modality fusion. Our HEHP leverages heterogeneous experts and hierarchical perception to adapt architecture and fuse representations for different contexts dynamically, enabling unified modeling of natural and underwater scenes.

C. Mixtures of Experts

MoE [20], [21] partitions the representation space into multiple subspaces, assigning them to different experts for processing, and finally consolidates the processing results through integration strategies, with weighted fusion being the most widely used. Building upon the principle, large language models extensively adopt it as a foundational framework, continually increasing parameters, *i.e.,* expert scale, to learn more complex representations. During inference, a subset of high-performance experts is dynamically selected based on the input. Inspired by this, we apply MoE to modality and dual domain space fusion. Through heterogeneous learning representations, we can enhance beneficial information between modalities while disregarding harmful information. Theoretically, the number of experts and representation space are proportionate, yet inevitably introduce numerous parameters, which is disadvantageous for deploying models on underwater mobile devices with high real-time requirements. Recently, low-rank adaptation (LoRA) [22], [23] has garnered widespread attention and is employed in efficient machine learning systems. By reducing data dimensions and noise to acquire highly informative representations, it effectively reduces complexity without significantly compromising core features. Hence, we transform the complex high-dimensional space into the low-rank space across multiple components and experts in our proposed framework. In addition, pixel space is susceptible to interference from complex underwater scenes, such as lighting distribution, which even expert collaboration struggles to mitigate. We further convert spatial experts into spectral experts to achieve frequency disentanglement.

D. Uncertainty Estimation

Mismatched depth information can degrade fine-grained features, making uncertainty estimation crucial. Uncertainty is divided into aleatoric and epistemic types. Aleatoric uncertainty arises from noise in the data, *e.g.,* depth map errors, and can be learned implicitly by the network. Epistemic uncertainty reflects the model's confidence in its predictions and often requires external techniques like Gaussian mixture models (GMMs) [24] or deep ensembles [25]. GMMs are effective in modeling complex, multimodal uncertainty but require significant computational resources to update and maintain multiple Gaussian components. Deep ensembles combine multiple models to predict, but are computationally expensive, as they require independent training of several models. Both methods cannot be directly integrated into

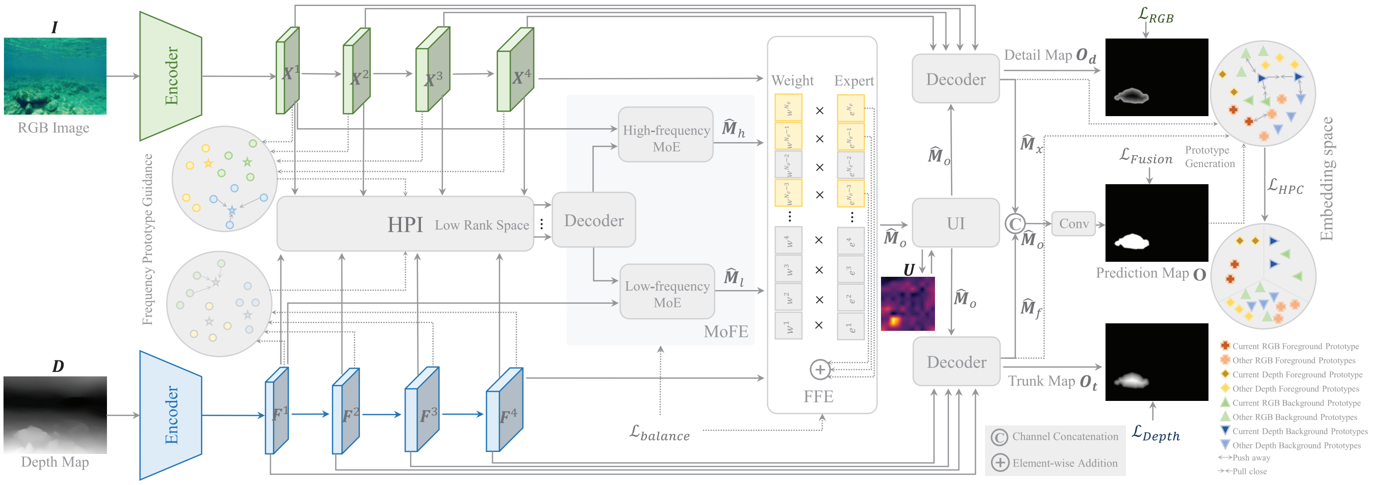


Fig. 2. The overview of our HEHP. We use separate encoders to generate RGB and depth features, which are then interacted with by the HPI. Subsequently, we decode and decouple the fused features, equipped with RGB and depth representations, and modulate through hierarchical MoFE to leverage the strengths of each modality. The FFE is utilized to merge heterogeneous features and dynamically adjust strategies based on the input. Furthermore, we enhance sensitivity to fine-grained regions using the UI. Leveraging modality differences, we employ hierarchical supervision with detail and trunk maps and introduce contrastive learning to generate compact and universal representations.

an end-to-end training pipeline, which complicates in large-scale, gradient-based deep learning models. To overcome this, we leverage Bayesian inference [26], [27] to estimate epistemic uncertainty by learning the posterior distribution of network weights. This approach enables gradient-based optimization, allowing it to operate without step-by-step training. Furthermore, we introduce the UI, which identifies high-uncertainty regions and enhances feature representations, improving fine-grained perception, especially in noisy or complex scenes such as low-quality depth maps and cluttered backgrounds.

III. Proposed Method

A. Motivation and Overview

The motivation of our proposed method is to decouple salient objects into detail and trunk, high- and low-frequency components, which are separately predicted by RGB and depth branches and then integrated together. And we maximize the differences between salient and non-salient regions based on contrastive learning. By employing the decoupling-integration-contrast process, we aim to explore the characteristics of different modalities and leverage their respective advantages.

As shown in Figure 2 and Algorithm 1, our method follows the overall paradigm of an encoder-decoder architecture and consists of five key elements. Given an image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ and corresponding depth map $\mathbf{D} \in \mathbb{R}^{3\times H\times W}$, we feed them to the respective encoders (weights are not shared) to obtain multiscale feature maps $\mathbf{X^i} \in \mathbb{R}^{C_i\times H_i\times W_i}$ and $\mathbf{F^i} \in \mathbb{R}^{C_i\times \ H_i\times W_i}$, where C, H, and W denote the number of channels, height, and width, respectively. We further utilize the HPI to align the same-level features and generate $\mathbf{M}^i$. We decouple the frequency, input into the MoFE for modulation, and obtain high and low frequency $\hat{\mathbf{M}}_h$ and $\hat{\mathbf{M}}_l$. We then combine the last features of the two branches, and $\hat{\mathbf{M}}_h$, $\hat{\mathbf{M}}_l$ via the FFE, generating $\hat{\mathbf{M}}_o$. By modeling the uncertainty through the UI, $\hat{\mathbf{M}}_o$ is decoded in combination with the encoded features of the two branches separately. Finally, we apply three supervisions to achieve implicit hierarchical learning.

Algorithm 1 Ours HEHP

```
Input: RGB and depth image $\mathbf{I}$ and $\mathbf{D}$; Parameters: Number of prototypes, experts and rank $N_p$, $N_{\mathcal{E}}$ and $R$
$\mathbf{X}^i, \mathbf{F}^i \leftarrow \mathsf{Encoder}(\mathbf{I}, \mathbf{D})$
$\mathbf{P}^i_l, \mathbf{P}^i_h \leftarrow \mathsf{FrequencyPrototypeGeneration}(\mathbf{X}^i, \mathbf{F}^i)$
$\mathbf{M}^i \leftarrow \mathsf{CalibrationAndRefinement}(\mathbf{P}^i_l, \mathbf{P}^i_h, \mathbf{X}^i, \mathbf{F}^i, R)$
$\hat{\mathbf{M}}_h, \hat{\mathbf{M}}_l \leftarrow \mathsf{DecoderAndDecouple}(\mathbf{M}^i)$
Frequency Expert Weight $\mathbf{W}_h, \mathbf{W}_l \leftarrow \mathcal{G}(\hat{\mathbf{M}}_h, \hat{\mathbf{M}}_l)$
if training then
    for $e_l \in \mathcal{E}_l$ and $e_h \in \mathcal{E}_h$ do
        $\hat{\mathbf{M}}_h, \hat{\mathbf{M}}_l \leftarrow \mathsf{ScaleAndOrientation}(\hat{\mathbf{M}}_h, \hat{\mathbf{M}}_l, R)$
    end for
    $\hat{\mathbf{M}}_h \leftarrow \sum_{e\in\mathcal{E}_h} \mathbf{W}_h \cdot e(\hat{\mathbf{M}}_h) + \hat{\mathbf{M}}_h,$
    $\hat{\mathbf{M}}_l \leftarrow \sum_{e\in\mathcal{E}_l} \mathbf{W}_l \cdot e(\hat{\mathbf{M}}_l) + \hat{\mathbf{M}}_l$
else
    $\mathbf{W}_h, \mathbf{W}_l \leftarrow \mathsf{Top}^2(\mathbf{W}_h, \mathbf{W}_l)$
end if
Fusion Expert Weight $\mathbf{W} \leftarrow \mathcal{G}(\mathbf{X}^{\text{Final}})$
if training then
    for $e \in \mathcal{E}$ do
        $\hat{\mathbf{M}}_\mathbf{o} \leftarrow \mathsf{Aggregation}(\hat{\mathbf{M}}_h, \hat{\mathbf{M}}_l, \mathbf{X}^{\text{Final}}, \mathbf{F}^{\text{Final}})$
    end for
    $\hat{\mathbf{M}}_\mathbf{o} \leftarrow \sum_{e\in\mathcal{E}} \mathbf{W}_o \cdot e(\hat{\mathbf{M}}_o) + \hat{\mathbf{M}}_o$
else
    $\mathbf{W} \leftarrow \mathsf{Top}^2(\mathbf{W})$
end if
$\hat{\mathbf{M}}_\mathbf{o} \leftarrow \mathsf{UncertaintyInjection}(\hat{\mathbf{M}}_\mathbf{o})$
Prediction Maps $\mathbf{O}_d, \mathbf{O}, \mathbf{O}_t \leftarrow \mathsf{Decoder}(\hat{\mathbf{M}}_\mathbf{o}, \mathbf{X}^i, \mathbf{F}^i)$
```

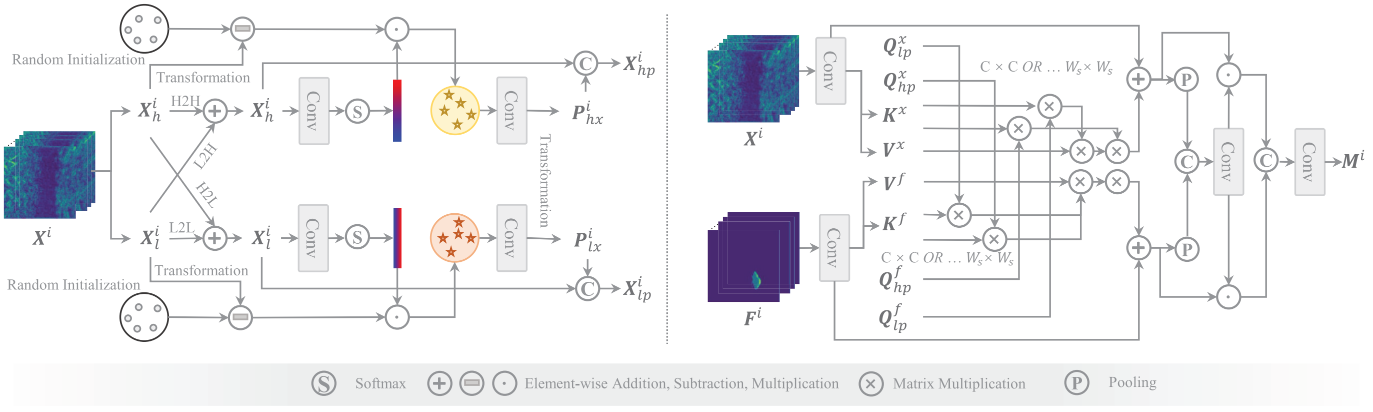


Fig. 3. Structure of the HPI. We obtain compact representations based on local frequency prototypes and apply calibration and refinement to obtain cross-modal representations with different emphases.

B. Hierarchical Prototype Guided Interaction

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 [28] and [29], we generate prototypes by clustering to obtain common feature representations that focus on the core regions. We further utilize generated frequency prototypes as complements to the initial features, achieving feature alignment from coarse to fine levels. Note that, unlike previous works [28], [29], [30] 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 details are in Figure 3.

Calibration. Based on octave convolution [31], we efficiently decouple features into high and low-frequency components in an end-to-end manner:

$$\begin{aligned}\mathbf{X}^i_l &= \mathcal{F}^{L\to L}\left(\mathbf{X}^i\right) + \mathsf{AP}\left(\mathcal{F}^{H\to L}\left(\mathbf{X}^i\right)\right)\\ \mathbf{X}^i_h &= \mathcal{F}^{H\to H}\left(\mathbf{X}^i\right) + \mathsf{UP}\left(\mathcal{F}^{L\to H}\left(\mathbf{X}^i\right)\right)\end{aligned} \tag{1}$$

where $\mathcal{F}$, UP and AP represent convolution, upsampling and average pooling, respectively. To perceive the differences of content, we first apply convolution followed by softmax for $\mathbf{X}^i_l$ and $\mathbf{X}^i_h$ to generate attention weight $\mathbf{W}^i_{lx}$ and $\mathbf{W}^i_{hx}$:

$$\mathbf{W}^i_{lx} = \mathsf{Softmax}\left(\mathcal{F}\left(\mathbf{X}^i_l\right)\right),\ \mathbf{W}^i_{hx} = \mathsf{Softmax}\left(\mathcal{F}\left(\mathbf{X}^i_h\right)\right) \tag{2}$$

For convenience, we omit the feature transformation. We randomly initialize the n_p cluster centers $\mathbf{C}^i_{lx}, \mathbf{C}^i_{hx} \in \mathbb{R}^{n_p\times C_i}$, and then generate the pixel difference $\mathbf{D}^i_{lx}, \mathbf{D}^i_{hx}$ by:

$$\mathbf{D}^i_{lx} = \mathbf{W}^i_{lx}\cdot(\hat{\mathbf{X}}^i_l - \hat{\mathbf{C}}^i_{lx}), \mathbf{D}^i_{hx} = \mathbf{W}^i_{hx}\cdot(\hat{\mathbf{X}}^i_h - \hat{\mathbf{C}}^i_{hx}) \tag{3}$$

where $(\cdot)$ denotes element-by-element multiplication, $\mathbf{X}^i_l$ and $\mathbf{C}^i_{lx}$ underwent dimensionality transformations to obtain $\hat{\mathbf{X}}^i_l$ and $\hat{\mathbf{C}}^i_{lx}$, respectively. We then obtain the high- and low-frequency prototypes $\mathbf{P}^i_{lx}, \mathbf{P}^i_{hx}$ by aggregating along the spatial dimensions:

$$\mathbf{P}^i_{lx} = \frac{\sum_{j=1}^{H\times W}\mathbf{D}^{i(j)}_{lx}}{\|\sum_{j=1}^{H\times W}\mathbf{D}^{i(j)}_{lx}\|},\ \mathbf{P}^i_{hx} = \frac{\sum_{j=1}^{H\times W}\mathbf{D}^{i(j)}_{hx}}{\|\sum_{j=1}^{H\times W}\mathbf{D}^{i(j)}_{hx}\|} \tag{4}$$

where $\|\cdot\|$ represents l2 norm. Intuitively, by reducing the differences in random parameters and features, clusters representing the region representation's centers can be generated, similar to superpixel clustering, although the latter operates in the original pixel space. We further fuse $\mathbf{P}^i_{lx}, \mathbf{P}^i_{hx}$ with corresponding initial features and then generate hierarchical prototype-guided features $\mathbf{X}^i_{lp}$ and $\mathbf{X}^i_{hp}$:

$$\mathbf{X}^i_{lp} = \mathcal{F}\left(\mathsf{Concat}(\mathbf{X}^i_l, \mathbf{P}^i_{lx})\right),\ \mathbf{X}^i_{hp} = \mathcal{F}\left(\mathsf{Concat}(\mathbf{X}^i_h, \mathbf{P}^i_{hx})\right) \tag{5}$$

where Concat represents the channel concatenation. Similarly, we can obtain prototype-enhanced depth features $\mathbf{F}^i_{lp}$ and $\mathbf{F}^i_{hp}$.

Calibrating across multiple scales introduces a significant number of parameters and computational complexity. The most straightforward approach is to reduce the scales, but this may introduce noise due to insufficient interactions. Therefore, we transfer the interaction space to the low-rank space with high information density. We use 1 × 1 followed by 3 × 3 depthwise separable convolution for $\mathbf{X}^i$ to encode local features and generate key $\mathbf{K}^x$ and value $\mathbf{V}^x$.[1] For queries, we perform the same operations on $\mathbf{X}^i_{lp}$ and $\mathbf{X}^i_{hp}$, generating $\mathbf{Q}^x_{lp}$ and $\mathbf{Q}^x_{hp}$. For depth features, similar operations are leveraged. We provide two interaction spaces, *i.e.,* channel and spatial, for comprehensive fusion and efficient computation. We can obtain the correlation matrix $\mathcal{M}_{x_{lp}\to k}$:

$$\mathcal{M}_{x_{lp}\to k} = \mathsf{Softmax}\left(\frac{\mathbf{Q}^x_{lp}\otimes\mathbf{K}^f_{lp}}{\tau_{x_{lp}\to k}}\right) \in \{\mathbb{R}^{C\times C} \vee \mathbb{R}^{\cdots\times W_s\times W_s}\} \tag{6}$$

where τ, W_s denote learnable scaling factor and window size, respectively. Similarly, we can obtain $\mathcal{M}_{x_{hp}\to k}$, $\mathcal{M}_{f_{lp}\to k}$, and $\mathcal{M}_{f_{hp}\to k}$. The interacted features are:

$$\begin{aligned}\hat{\mathbf{X}}^i &= \mathcal{M}_{f_{lp}\to k}\otimes\mathbf{K}^x\otimes\mathcal{M}_{f_{hp}\to k},\\ \hat{\mathbf{F}}^i &= \mathcal{M}_{x_{lp}\to k}\otimes\mathbf{K}^f\otimes\mathcal{M}_{x_{hp}\to k}\end{aligned} \tag{7}$$

where $\otimes$ is matrix multiplication.

Refinement. We apply spatial compression on $\hat{\mathbf{X}}^i$ and $\hat{\mathbf{F}}^i$ respectively, and then fuse them to obtain the channel map $\mathbf{S}_r, \mathbf{S}_d$:

$$\mathbf{S}_r, \mathbf{S}_d = \mathsf{Split}\left(\mathcal{F}\left(\mathsf{Concat}\left(\mathsf{AP}\left(\hat{\mathbf{X}}^i\right), \mathsf{AP}\left(\hat{\mathbf{F}}^i\right)\right)\right)\right) \tag{8}$$

[1] For simplicity, we omit the indices here.

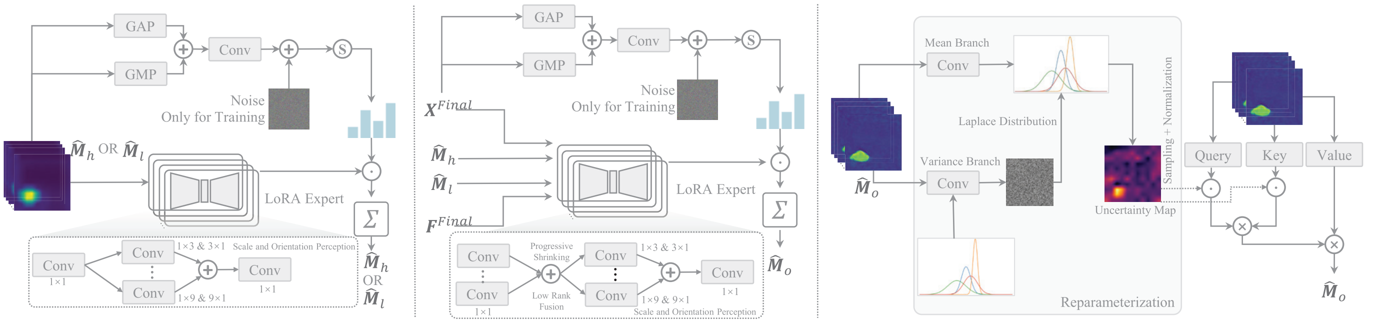


Fig. 4. Structure of the MoFE, FFE, and UI. We utilize the MoFE to modulate frequency signals and the FFE to dynamically fuse heterogeneous information, minimizing error interference to the maximum extent. We employ pixel probability distribution to localize regions with high uncertainty.

where Split indicates channel split. Thus, refined mixed feature $\mathbf{M}^i$ is generated by:

$$\mathbf{M}^i = \mathcal{F}\left(\mathsf{Concat}\left(\hat{\mathbf{X}}^i \cdot \mathbf{S}_r, \hat{\mathbf{F}}^i \cdot \mathbf{S}_d\right)\right) \tag{9}$$

For channel refinement, the operation is similar.

C. Mixture of Frequency Experts/Four-Way Fusion Experts

RGB features carry more details, corresponding to high frequency; depth features reflect pixel positions, corresponding to low frequency. Based on the characteristics, we decouple features and allocate the hierarchical MoFE. Due to the differences in frequency and modality, direct fusion is suboptimal and challenging to modulate based on each sample. Therefore, we introduce the FFE to dynamically merge heterogeneous features, allowing experts to automatically learn the optimal combination strategy, reducing manual design and improving efficiency. The number of experts does not directly correlate with performance improvement and has an upper limit; introducing too many experts may lead to redundancy and performance degradation. Similarly, the spatial capacity of expert handling should not be excessive; we adjust all interaction spaces to low rank. The details are in Figure 4.

Leveraging the aggregation strategy $\mathcal{A}$, we progressively shrink $\mathbf{M}^i$ in the decoder and obtain $\hat{\mathbf{M}}$:

$$\mathcal{A}(\mathbf{M}^1, \ldots, \mathbf{M}^i) = \begin{cases} \mathbf{M}^1 & \text{if } i = 1 \\ \mathcal{F}(\mathcal{A}(\mathbf{M}^1, \ldots, \mathbf{M}^{i-1}) + \mathbf{M}^i) & \text{if } i > 1 \end{cases} \tag{10}$$

Similar to Eq. 1, we decouple $\hat{\mathbf{M}}$ to generate low-frequency $\hat{\mathbf{M}}_l$ and high-frequency $\hat{\mathbf{M}}_h$ based on octave convolution. We consider the following: 1) The high-frequency information of RGB images is stronger in shallow feature representations; 2) Depth map information is constrained, making its low-frequency variations less pronounced. Therefore, we fuse $\hat{\mathbf{M}}_h$ and $\mathbf{X}^1$, $\hat{\mathbf{M}}_l$ and $\mathbf{F}^1$, then input into mixture of high and low-frequency experts for modulation:

$$\hat{\mathbf{M}}_h := \mathcal{F}(\hat{\mathbf{M}}_h + \mathbf{X}^1), \ \hat{\mathbf{M}}_l := \mathcal{F}(\hat{\mathbf{M}}_l + \mathbf{F}^1) \tag{11}$$

1) MoFE. MoFE consists of two components, *i.e.,* high- and low- frequency. Taking high frequency as an example, the crucial components include a routing mechanism (implemented by a gating network $\mathcal{G}$) and experts $\mathcal{E}$. $\mathcal{G}$ allocates weights to different $\mathcal{E}$. To enrich representations, we use average and max pooling (MP) to compress space and then formulate:

$$\mathcal{G}(\hat{\mathbf{M}}_h) = \mathsf{Softmax}\left(\mathcal{F}\left(\mathsf{AP}(\hat{\mathbf{M}}_h) + \mathsf{MP}(\hat{\mathbf{M}}_h)\right) + \mathbf{z}\right), \quad \mathbf{z} \sim \mathbb{N}(0,1) \tag{12}$$

where $\mathbf{z}$ follows the standard Gaussian distribution aimed at enhancing robustness. Therefore, we remove this term during the inference stage.

Similar to the HPI, to maintain a balance between computational costs and performance, we place experts in the low-rank space to handle features. Considering variations in object scale and position, we design kernels of multiple sizes for the experts and decompose them into horizontal and vertical directions (also optimizing computational complexity). Therefore, the final output $\hat{\mathbf{M}}_h$ is:

$$\hat{\mathbf{M}}_h := \hat{\mathbf{M}}_h + \sum_{n=0}^{N_{\mathcal{E}}} \mathcal{G}(\hat{\mathbf{M}}_h) \cdot \mathcal{E}^i(\hat{\mathbf{M}}_h) \tag{13}$$

where N is the number of experts. Similarly, we can obtain the modulated $\hat{\mathbf{M}}_l$.

2) FFE. To introduce semantic representation, we utilize the final layer features of two branches, *i.e.,* $\mathbf{X}^{\text{Final}}$ and $\mathbf{F}^{\text{Final}}$. FFE and MoFE share the same essence. The difference is that FFE receives four heterogeneous outputs, *i.e.,* different modalities and spaces. Therefore, utilizing the aggregation method in Eq. 10, we fuse within the experts and generate $\hat{\mathbf{M}}_o$. We further perceive scales and orientations:

$$\hat{\mathbf{M}}_o := \mathcal{F}^{\text{up}}\left(\sum_{i=1}^{L}\left(\mathcal{F}^{1\times(2l+1)}\left(\mathcal{F}^{\text{down}}(\hat{\mathbf{M}}_o)\right) + \mathcal{F}^{(2l+1)\times 1}\left(\mathcal{F}^{\text{down}}(\hat{\mathbf{M}}_o)\right)\right)\right) \tag{14}$$

where $\mathcal{F}^{\text{down}}$ and $\mathcal{F}^{\text{up}}$ are used to reduce the original dimensionality to the low rank and to restore, respectively. L is the convolution levels. The final output $\hat{\mathbf{M}}_o$ is:

$$\hat{\mathbf{M}}_o := \mathbf{X}^{\text{Final}} + \sum_{n=0}^{N_{\mathcal{E}}} \mathcal{G}(\mathbf{X}^{\text{Final}}) \cdot \mathcal{E}(\mathbf{X}^{\text{Final}}, \mathbf{F}^{\text{Final}}, \hat{\mathbf{M}}_h, \hat{\mathbf{M}}_l) \tag{15}$$

During training iterations, the gating function may tend to frequently activate experts with larger weights $\mathbf{W}_{\mathcal{E}}$, leaving other experts idle, leading to decreased expert utilization

TABLE I

Quantitative Comparison on USOD10K and USOD Benchmarks. Best Performances Are Highlighted in Bold, Followed by Performances Indicated With Underlines

Method	USOD10K [18]				USOD [17]			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	MAE ↓	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	MAE ↓
DCF [34] CVPR21	0.912	0.905	0.954	0.031	0.893	0.902	0.931	0.053
SPNet [35] ICCV21	0.909	0.909	0.955	0.028	0.888	0.903	0.925	0.052
TC-USOD [18] TIP23	0.922	0.923	0.968	0.021	0.895	0.910	0.934	0.046
HiDANet [36] TIP23	0.911	0.922	0.959	0.028	0.889	0.907	0.930	0.042
PopNet [37] ICCV23	0.926	0.927	0.963	0.027	0.902	0.915	0.944	0.038
CATNet [38] TMM23	0.890	0.888	0.949	0.030	0.878	0.895	0.925	0.052
TPCL [30] TMM23	0.898	0.904	0.954	0.029	0.902	0.908	0.923	0.044
MSNet [39] TASE24	0.889	0.899	0.956	0.032	0.888	0.896	0.927	0.054
PICR-Net [8] MM23	0.921	0.919	0.964	0.023	0.897	0.912	0.934	0.045
SPDE [19] TCSVT24	0.923	0.927	0.969	0.020	0.900	0.913	0.934	0.044
UniTR [40] TMM24	0.918	0.928	0.963	0.019	0.911	0.920	0.938	0.035
CPNet [6] IJCV24	0.920	0.932	0.963	0.023	0.909	0.918	0.942	0.031
Dual-SAM [41] CVPR24	0.924	0.931	‡	0.019	0.917	‡	‡	0.026
DFormer [16] ICLR24	0.925	0.924	0.968	0.022	0.913	0.927	0.953	0.024
VSCode [42] CVPR24	0.921	0.928	0.971	0.018	0.920	0.938	0.959	0.025
Ours	**0.935**	**0.940**	**0.983**	**0.014**	**0.931**	**0.950**	**0.978**	**0.022**

and diversity. To achieve load balancing, we introduce the balancing loss $\mathcal{L}_{\text{balance}}$ as a constraint,

$$\mathcal{L}_{\text{balance}} = -\log\left(\prod_{i=0}^{N_{\mathcal{E}}} \frac{\exp\left(\mathbf{W}_{\mathcal{E}_i}/\tau\right)}{\sum_{j=1}^{N_{\mathcal{E}}}\exp\left(\mathbf{W}_{\mathcal{E}_j}/\tau\right)}\right) \tag{16}$$

D. Uncertainty Injection

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. Thus, we propose the UI, which introduces probability modeling to locate regions with high uncertainty and further reduce depth noise, shown in Figure 4.

We establish Laplace distribution for each pixel to construct the uncertainty maps. We first obtain the mean $\boldsymbol{\mu}$ and variance $\mathbf{b}$ separately by using two different projections (towards channels of 1) on $\hat{\mathbf{M}}_o$:

$$\boldsymbol{\mu} = \mathcal{F}(\hat{\mathbf{M}}_o), \mathbf{b} = \mathcal{F}(\hat{\mathbf{M}}_o) \tag{17}$$

Why not choose Gaussian distribution modeling? The Laplace distribution is easier to optimize and better highlights the details of features. In Table IV, we further validate through quantitative results. Gradients cannot directly optimize random samples. Following [32], we randomly sample several times to generate variable $\boldsymbol{\xi}$ from standard Laplace distribution to obtain new uncertain distribution of pixels, *i.e,* $\mathcal{L} = \boldsymbol{\mu} + \boldsymbol{\xi}\mathbf{b}$. We further calculate the variance and normalize to yield uncertainty maps $\mathbf{U}$:

$$\mathbf{U} = \frac{\mathcal{F}\left(\mathsf{Var}\left(\phi\left(\mathsf{Sample}(\mathcal{L})\right)\right)\right)}{\left\|\mathcal{F}\left(\mathsf{Var}\left(\phi\left(\mathsf{Sample}(\mathcal{L})\right)\right)\right)\right\|} \tag{18}$$

where $\mathsf{Var}(\cdot)$ and $\phi(\cdot)$ denote the sample variance and sigmoid function, respectively.

Similar with the HPI, we apply 1×1 followed by 3×3 convolution on $\hat{\mathbf{M}}_o$ to obtain the query $\mathbf{Q}_m$, key $\mathbf{K}_m$, and value $\mathbf{V}_m$. We further apply $\mathbf{U}$ to the query and key to generate the correlation matrix $\mathcal{M}_u$, thus, we obtain uncertainty-aware features generated based on self-attention by:

$$\mathcal{M}_u = \mathsf{Softmax}\left(\frac{(\mathbf{Q}_m\mathbf{U}) \otimes (\mathbf{K}_m\mathbf{U})}{\tau_u}\right), \ \hat{\mathbf{M}}_o := \mathbf{V}_m \otimes \mathcal{M}_u \tag{19}$$

E. Loss Function

We enhance features using frequency prototypes but lack exploration of semantic prototypes, which is crucial for comprehensive understanding of salient regions. Therefore, we utilize the HPC loss for modeling, which includes two dimensions: global and patch. We first employ masked average pooling (MAP) for RGB modality to generate foreground and background prototypes based on decoder features and corresponding binary outputs. We have:

$$\mathbf{P}_r^f = \frac{\sum_{h=1,w=1}^{H,W}\mathbf{O}^{hw}\cdot\hat{\mathbf{M}}_x}{\sum_{h=1,w=1}^{H,W}\mathbf{O}^{hw}}, \mathbf{P}_r^b = \frac{\sum_{h=1,w=1}^{H,W}(1-\mathbf{O}^{ij})\cdot\hat{\mathbf{M}}_x}{\sum_{h=1,w=1}^{H,W}1-\mathbf{O}^{hw}} \tag{20}$$

Similarly, we can generate global foreground and background prototypes for deep modality. We align prototypes between modalities and batches by pulling similar prototypes closer and pushing dissimilar prototypes apart, formulating the global prototype contrastive loss. We have $\mathcal{L}_{\text{global}}$ by:

$$\mathcal{L}_{\text{global}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{k\in\mathcal{K}} -\log\frac{\exp(s_k^i)}{\exp(s_k^i) + \sum_{n\in\mathcal{N}_k^i}\exp(s_{k,n}^i)} \tag{21}$$

where B is the batch size, $\mathcal{K}$ is the total number of classes, *i.e.,* 2. $\mathcal{N}$ is the set of negative samples, $s_k = \cos(\mathbf{P}_r^k, \mathbf{P}_d^k)/\tau$.

Due to limitations in batch size and GPU memory, the number of global prototypes is insufficient to support comprehensive contrast. Therefore, we apply the MAP at the patch level to generate prototypes. Expanding on Eq. 21, we establish intra-class and inter-class patch contrast to formulate patch prototype contrastive loss $\mathcal{L}_{\text{patch}}$,

$$\begin{aligned}\mathcal{L}_{\text{patch}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{k\in\mathcal{K}}\Bigg(&-\log\frac{\exp(s_k^i)}{\sum_{p\in\mathcal{P}_k^i}\exp(s_{k,p}^i)}\\ &-\log\frac{\exp(s_k^i)}{\exp(s_k^i)+\sum_{n\in\mathcal{N}_k^i}\exp(s_{k,n}^{\bar{k},i})}\Bigg)\end{aligned} \tag{22}$$

where $\mathcal{P}$ is the set of positive samples. Different from local prototypes with more noise from random parameters in the HPI, the patch prototypes we obtain can more accurately capture fine-grained representations.

We apply supervision to the output of the RGB, depth, and fusion branches. Each branch loss $\mathcal{L}_{\text{RGB}}$, $\mathcal{L}_{\text{Depth}}$, $\mathcal{L}_{\text{Fusion}}$ consists of structure loss $\mathcal{L}_{\text{structure}}$ (combination of binary cross entropy loss, intersection over union loss, and structural similarity index loss) and three auxiliary losses. We have $\mathcal{L}_{\text{RGB}}$:

$$\mathcal{L}_{\text{RGB}} = \mathcal{L}_{\text{structure}} + \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{patch}} + 0.1 \times \mathcal{L}_{\text{balance}} \tag{23}$$

Similarly, we can obtain $\mathcal{L}_{\text{Depth}}$ and $\mathcal{L}_{\text{Fusion}}$. The total loss $\mathcal{L}_{\text{Total}}$ can be expressed as:

$$\mathcal{L}_{\text{Total}} = \alpha\mathcal{L}_{\text{RGB}} + \beta\mathcal{L}_{\text{Depth}} + \gamma\mathcal{L}_{\text{Fusion}} \tag{24}$$

where α, β, γ are hyperparameters. We empirically set α, β, γ to 1.2, 0.3, and 0.3.

TABLE II

QUANTITATIVE COMPARISON ON MAS3K, RMAS, UFO120, AND RUWI BENCHMARKS

Method	MAS3K [42]				RMAS [43]				UFO120 [44]				RUWI [45]			
	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$E_\phi^m\uparrow$	MAE ↓	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$E_\phi^m\uparrow$	MAE ↓	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$E_\phi^m\uparrow$	MAE ↓	$S_\alpha\downarrow$	$F_\beta^w\uparrow$	$E_\phi^m\uparrow$	MAE ↓
OCENet [46] WACV22	0.824	0.703	0.868	0.052	0.836	0.752	0.900	0.030	0.725	0.668	0.773	0.161	0.791	0.798	0.863	0.115
ZoomNet [47] CVPR22	0.862	0.780	0.898	0.032	0.855	0.795	0.915	0.022	0.702	0.670	0.815	0.174	0.753	0.771	0.817	0.137
MASNet [43] IJOE23	0.864	0.788	0.906	0.032	0.862	0.801	0.920	0.024	0.827	0.820	0.879	0.083	0.880	0.913	0.944	0.047
SETR [48] CVPR21	0.855	0.789	0.917	0.030	0.818	0.747	0.933	0.028	0.811	0.796	0.871	0.089	0.864	0.895	0.924	0.055
H2Former [49] TMI23	0.865	0.810	0.925	0.028	0.844	0.799	0.931	0.023	0.844	0.845	0.901	0.070	0.884	0.919	0.945	0.045
SAM [50] ICCV23	0.763	0.656	0.807	0.059	0.697	0.534	0.790	0.053	0.768	0.745	0.827	0.121	0.855	0.907	0.929	0.057
SAM-Adapter [51] ICCVw23	0.847	0.782	0.914	0.033	0.816	0.752	0.927	0.027	0.829	0.834	0.884	0.081	0.878	0.913	0.946	0.046
CPNet [6] IJCV24	0.869	0.805	0.918	0.028	0.837	0.791	0.939	0.024	0.840	0.847	0.898	0.069	0.881	0.927	0.951	0.037
Dual-SAM [40] CVPR24	0.884	0.838	0.933	0.023	0.860	0.812	0.944	0.022	0.856	0.864	0.914	0.064	**0.903**	0.939	0.959	0.035
MAS-SAM [52] IJCAI24	0.887	0.840	0.938	0.025	0.865	0.819	0.948	0.021	0.861	0.864	0.914	0.063	0.894	0.941	0.961	0.035
Ours	**0.906**	**0.862**	**0.961**	**0.017**	**0.885**	**0.841**	**0.978**	**0.018**	**0.875**	**0.869**	**0.930**	**0.052**	0.890	**0.959**	**0.972**	**0.024**

TABLE III

QUANTITATIVE ABLATION OF CRUCIAL COMPONENTS. B1, B2, C1, C2, C3, C4, C5, AND C6 INDICATE PVT-V2, SWIN-S, THE HPI, MOFE, FFE, UI, AND HPC LOSS, HIERARCHICAL SUPERVISION, RESPECTIVELY. THE PRECEDING COMPONENTS PROVIDE THE FEATURE FOUNDATION FOR SUBSEQUENT COMPONENTS. WHEN LACKING, WE UTILIZE CONVOLUTION AND CHANNEL CONCATENATION TO COMPENSATE

Method	Backbone		Component						USOD10K				USOD			
	B1	B2	C1	C2	C3	C4	C5	C6	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	MAE↓
I	✓								0.859	0.873	0.896	0.027	0.855	0.868	0.904	0.035
II	✓		✓						0.871	0.880	0.909	0.026	0.867	0.877	0.915	0.032
III	✓			✓					0.874	0.882	0.915	0.026	0.870	0.880	0.919	0.031
IV	✓				✓				0.868	0.884	0.918	0.025	0.864	0.884	0.913	0.032
V	✓					✓			0.865	0.881	0.908	0.027	0.862	0.875	0.911	0.034
VI	✓		✓	✓					0.892	0.897	0.933	0.023	0.883	0.885	0.930	0.029
VII	✓		✓		✓				0.885	0.890	0.931	0.026	0.876	0.893	0.925	0.028
VIII	✓		✓			✓			0.887	0.892	0.927	0.025	0.879	0.890	0.922	0.031
IX	✓			✓	✓				0.877	0.887	0.928	0.026	0.867	0.884	0.928	0.030
X	✓			✓		✓			0.880	0.885	0.919	0.026	0.870	0.881	0.924	0.032
XI	✓				✓	✓			0.879	0.893	0.931	0.024	0.875	0.894	0.927	0.032
XII	✓		✓			✓		✓	0.895	0.900	0.936	0.024	0.886	0.895	0.930	0.030
XIII	✓		✓	✓	✓	✓			0.907	0.915	0.952	0.020	0.905	0.914	0.945	0.028
XIV	✓		✓	✓	✓	✓	✓		0.926	0.931	0.974	0.016	0.922	0.938	0.967	0.025
XV	✓		✓	✓	✓	✓		✓	0.920	0.926	0.965	0.017	0.916	0.933	0.961	0.025
XVI	✓		✓	✓	✓	✓	✓	✓	0.935	0.940	0.983	0.014	0.931	0.950	0.978	0.022
XVII		✓	✓	✓	✓	✓	✓	✓	0.938	0.941	0.974	0.015	0.928	0.953	0.971	0.021

TABLE IV

QUANTITATIVE ABLATION ON DISTRIBUTION IN THE UI

Method	Component		USOD10K				USOD			
	Gaussian	Laplace	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	MAE↓
I	✓		0.930	0.934	0.980	0.014	0.932	0.945	0.975	0.022
II		✓	0.935	0.940	0.983	0.014	0.931	0.950	0.978	0.022

IV. EXPERIMENT

A. Datasets

We conduct experiments on eleven benchmarks in underwater and natural scenes. We select two widely used USOD datasets equipped with depth maps, USOD10K [18] and USOD [17], where USOD is used for testing only. USOD10K contains 7178 training and validation pairs, as well as 1026 test images, with diverse size variations and rich scenes, while USOD comprises 300 images. Additionally, we employ four underwater datasets: MAS3K [42], RMAS [43], UFO120 [44], and RUWI [45]. Due to the lack of corresponding depth maps, we utilize a monocular depth estimation method, *i.e.,* DPT [63], for generation. Furthermore, seven natural scene datasets, STERE [53], SIP [54], NJU2K [55], NLPR [56], DUT [57], DUTS-TE [61] and VizWiz-SO [62] are chosen.

B. Implementation Details

We implement our method using PyTorch and conduct all experiments on NVIDIA RTX A100s. For the USOD10K and USOD datasets, following [18], inputs are scaled to 224 × 224. For the MAS3K, RMAS, UFO120, RUWI, following [40], we adjust to 512 × 512. For natural scenes datasets, following [16], 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 [16], we employ the original training set along with the training sets of NJU2K and NLPR, the original test set for evaluation, and resize the input to 384 × 384. We utilize the different backbone networks as the encoder. For training, we use AdamW as the optimizer with 150 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 augmentation) and post-processing (*e.g.,* CRF).

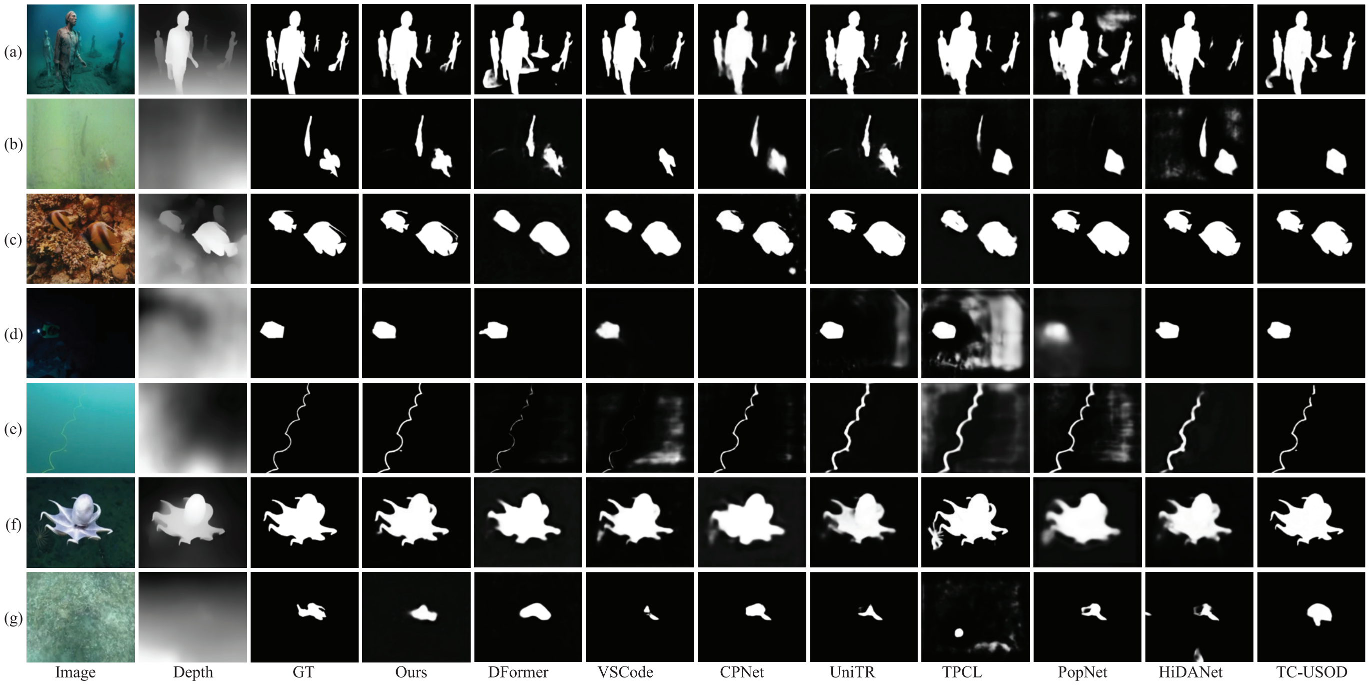


Fig. 5. Qualitative comparison on underwater scenes.

To ensure fairness, we obtain evaluation results by utilizing the saliency maps provided by the projects or retraining using the official source codes.

C. Evaluation Metrics

We adopt eight evaluation metrics: S-measure (S_α) [64], mean and maximum E-measure (E_ϕ^m, E_ϕ) [65], mean, weighted, adaptive, and maximum F-measure (F_β^m, F_β^w, F_β^a, F_β) [66],[2] and Mean Absolute Error (MAE). Note that the higher the better for the first seven.

D. Comparison on Underwater Benchmarks

1) Methods on USOD10K and USOD: We compare our method with fifteen different types of methods, including DCF [33], SPNet [34], TC-USOD [18], PopNet [36], CATNet [37], TPCL [30], HiDANet [35], MSNet [38], PICR-Net [8], SPDE [19], UniTR [39], CPNet [6], Dual-SAM [40], DFormer [16] and VSCode [41].

2) Methods on MAS3K, RMAS, UFO120, and RUWI: Ten models, *i.e.,* OCENet [46], ZoomNet [47], MASNet [43], SETR [48], H2Former [49], SAM [50], SAM-Adapter [51], CPNet [6], Dual-SAM [40], and MAS-SAM [52] are selected.

3) Quantitative Comparison: In Table I, the HEHP achieves the best performance across all metrics on USOD10K and USOD, surpassing existing methods, including those relying on additional data (*e.g.,* VSCode, DFormer) or large foundation models (*e.g.,* DualSAM). For example, the HEHP outperforms DualSAM by +1.1% in S_α and +0.9% in F_β on USOD10K, demonstrating the effectiveness of domain-specific design over general pre-training. From the model design perspective, the HEHP integrates lightweight yet adaptive components such as the MoFE and the FFE, which decompose and fuse features across frequency domains. This design captures both local details and global semantics, crucial for handling low-contrast and structure-ambiguous underwater scenes. Compared to CPNet's fixed and heavily coupled fusion modules, our expert-based routing adapts more flexibly to different inputs while maintaining efficiency, leading to stronger performance (*e.g., +1.5%* S_α on USOD10K). The HPI further alleviates cross-modal misalignment by guiding interactions through hierarchical prototypes, particularly benefiting scenarios with noisy depth or cluttered backgrounds. In contrast, methods like TPCL lack structured alignment and suffer under such conditions. The HEHP improves upon TPCL by +4.2% in F_β and +5.5% in E_ϕ on USOD, highlighting the advantages of guided interaction and global-patch contrast. From the data characteristics perspective, underwater scenes introduce strong domain shifts, including low visibility, severe color distortion, and inaccurate or synthetic depth. These factors limit the generalization of models trained on natural images. SAM-based approaches, though trained on massive datasets, struggle under these conditions, and adapter tuning alone is insufficient to bridge the domain gap. The HEHP explicitly addresses these challenges through uncertainty-aware learning, which suppresses unreliable signals, and hierarchical supervision that encourages diverse, modality-specific representation learning. As shown in Table II, the HEHP delivers even more notable gains, surpassing MAS-SAM and CPNet by an average of +2.2% in S_α and +2.7% in E_ϕ^m, confirming its robustness across varied underwater environments. These results underline the importance of aligning model structure with data characteristics to achieve both adaptability and generalization.

[2] The mainstream evaluation metrics vary slightly across benchmarks.

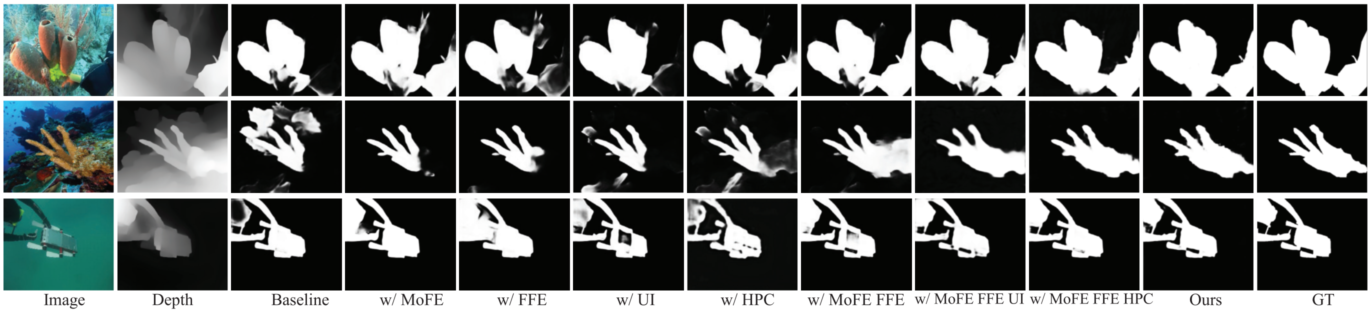


Fig. 6. Quantitative ablation of models at different stages with the inclusion of proposed components.

TABLE V
QUANTITATIVE COMPARISON ON COMPUTATION AND MODEL COMPLEXITY. WE COMPARE ON PARAMETERS (M), FLOPs (GMAC), AND FPS

Method	Backbone	FLOPs.↓	Params.↓	FPS↑
XMSNet [9] MM23	PVT-v2	30.32	53.89	42
CPNet [6] IJCV24	Swin-S	129.34	216.50	36
Ours-P	PVT-v2	36.74	46.55	48
Ours-S	Swin-S	75.63	99.74	44

TABLE VI
QUANTITATIVE ABLATION ON THE HPC LOSS

Method	USOD10K				USOD			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	MAE ↓	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	MAE ↓
CPNet IJCV24 [6]	0.920	0.932	0.963	0.023	0.909	0.918	0.942	0.031
CPNet + $\mathcal{L}_{\text{local}}$	0.925	0.936	0.968	0.021	0.914	0.923	0.947	0.028
CPNet + $\mathcal{L}_{\text{global}}$	0.927	0.934	0.970	0.021	0.917	0.927	0.949	0.027
CPNet + $\mathcal{L}_{\text{local}}$ + $\mathcal{L}_{\text{global}}$	0.930	0.938	0.973	0.019	0.921	0.931	0.953	0.027

4) Qualitative Comparison: In Figure 5, we offer visual comparisons across different scenarios. Our method excels in perceiving multiple objects and details comprehensively (row a). In low-contrast environments (row b), false negatives are mitigated. When dealing with complex backgrounds (row c) and low-light conditions (row d), our approach effectively tackles noise and imaging interferences. For elongated (row e), irregular (row f), and camouflaged targets (row g), our approach achieves long-range modeling, global localization, local calibration through hierarchical perception, and avoids false positives.

E. Ablation Studies

To validate the effect of proposed components and hyper-parameters, we conduct quantitative and qualitative ablation analyses on the USOD10K and USOD datasets.

1) Effect of the Crucial Components: In Table III, we categorize the components into two types: internal and external to the model. Internally, we use Swin and PVT as the backbone, with the performance of the former being inferior. Based on this, we further validate the correlation of performance by incrementally adding components in one-step and two-step manners. We observe that the most significant performance improvements occur with the introduction of expert learning. We argue that the MoFE and the FFE are tightly coupled to improve feature fusion. The MoFE separates high- and low-frequency features, capturing modality-specific details, while the FFE dynamically selects experts based on spatial and orientation differences, ensuring that features from each modality are effectively aligned and fused. The HPI aligns features at different abstraction levels, reducing modality discrepancies and improving the quality of the fused features. The UI directs attention to high-uncertainty regions, focusing on fine-grained or noisy parts that may otherwise disrupt feature fusion. Together, the HPI and the UI ensure that the interaction process is both semantically and spatially accurate, especially in complex underwater scenes. Externally, HPC loss and hierarchical supervision enhance the model's final representations. HPC loss enforces compact prototype learning at both global and patch levels, improving foreground-background distinctions. However, when comparing Models VIII, XII, and XV, despite hierarchical supervision implicitly leveraging modality characteristics to shape distinctive representations, the performance improvement is not as prominent in the absence of heterogeneous expert handling. Therefore, when representations are mixed, explicit disentanglement proves to be more crucial. In Table VI, we gradually equip CPNet with $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$. We find that the gain from $\mathcal{L}_{\text{local}}$ is not as significant as $\mathcal{L}_{\text{global}}$, and when both work together, the combined effect surpasses that of each one. We analyze that the lack of alignment in global prototypes may lead to inaccurate foreground-background representations, resulting in prototypes generated from patch divisions containing more noise and consequently reducing the quality of contrasts. Through the supplementation of fine-grained prototypes, global contrasts are promoted, thereby enhancing regional divisions. In Figure 6, we present the quantitative impact of these components at each stage.

2) Effect of Backbone Network and Model Efficiency Analysis: In Table III and Table V, we use PVT-v2 [67] and Swin-S [68] as backbone networks, with similar performance but higher model efficiency for the former. Compared to CPNet, our Swin-S version surpasses by +1.8%, +0.9%, +2.0%, −0.8%, and +1.9%, +3.5%, +2.9%, −1.0% on four indicators, respectively. Despite incorporating several components, our method achieves a balance between performance and efficiency through low-rank optimization and reduction of

TABLE VII

Quantitative Ablation on Components of the HPI and the MoFE. C1, C2, C3, C4, C5, C6, C7, C8, C9 Indicate the Cross-Attention, Spatial-Calibration, Channel-Calibration, Spatial-Refinement, Channel-Refinement, Laplace, Fourier, Discrete Cosine, Octave Convolution, Respectively

Method	Component									USOD10K				USOD			
	C1	C2	C3	C4	C5	C6	C7	C8	C9	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	MAE$\downarrow$
I	✓									0.905	0.911	0.945	0.023	0.901	0.922	0.946	0.027
II		✓								0.908	0.916	0.949	0.021	0.904	0.918	0.949	0.024
III			✓							0.903	0.917	0.952	0.020	0.906	0.923	0.943	0.025
IV		✓	✓							0.916	0.921	0.957	0.020	0.913	0.931	0.957	0.025
V				✓						0.904	0.900	0.941	0.023	0.894	0.914	0.947	0.028
VI					✓					0.901	0.905	0.949	0.024	0.896	0.913	0.945	0.027
VII				✓	✓					0.907	0.913	0.947	0.022	0.903	0.920	0.945	0.027
VIII	✓	✓	✓	✓	✓					0.927	0.932	0.972	0.017	0.924	0.942	0.967	0.023
IX	✓	✓	✓	✓	✓	✓				0.930	0.929	0.978	0.015	0.933	0.947	0.970	0.022
X	✓	✓	✓	✓	✓		✓			0.931	0.933	0.981	0.014	0.931	0.952	0.973	0.023
XI	✓	✓	✓	✓	✓			✓		0.937	0.935	0.977	0.015	0.927	0.945	0.976	0.022
XII	✓	✓	✓	✓	✓				✓	0.935	0.940	0.983	0.014	0.931	0.950	0.978	0.022

TABLE VIII

Quantitative Ablation on Components of the MoFE, FFE, and Supervision Strategy. C1, C2, C3, C4, C5, C6, C7, C8, C9, C10 Indicate the Spatial Expert, Low-Frequency Expert, High-Frequency Expert, Inverse Frequency, Balanced, Scale, Direction, Single Supervision, Inverse Supervision, Ours, Respectively

Method	Component										USOD10K				USOD			
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	MAE$\downarrow$
I	✓							✓			0.907	0.910	0.949	0.021	0.904	0.913	0.937	0.029
II		✓						✓			0.904	0.907	0.947	0.022	0.902	0.916	0.931	0.030
III			✓					✓			0.901	0.902	0.953	0.021	0.900	0.908	0.933	0.029
IV		✓	✓					✓			0.915	0.922	0.963	0.018	0.918	0.926	0.950	0.027
V		✓	✓	✓				✓			0.919	0.917	0.958	0.019	0.913	0.920	0.944	0.028
VI		✓	✓		✓			✓			0.920	0.927	0.969	0.018	0.922	0.931	0.954	0.026
VII		✓	✓		✓	✓		✓			0.924	0.933	0.966	0.016	0.917	0.936	0.963	0.024
VIII		✓	✓		✓		✓	✓			0.925	0.927	0.970	0.017	0.920	0.934	0.965	0.025
IX		✓	✓		✓	✓	✓	✓			0.930	0.931	0.975	0.015	0.925	0.940	0.973	0.022
X		✓	✓		✓	✓	✓		✓		0.926	0.932	0.972	0.016	0.923	0.941	0.969	0.023
XI		✓	✓		✓	✓	✓			✓	0.935	0.940	0.983	0.014	0.931	0.950	0.978	0.022

interaction space, significantly reducing model and computational complexity.

3) Analysis of Alignment Strategies: As shown in Table VII, we employ the HPI for aligning modal features, which is based on cross attention (CA). To emphasize the differences and advantages of the proposed component, we divide it into four subparts based on interaction space and process. We find that both spatial and channel calibration outperform vanilla CA. We analyze the key lies in: 1) Introducing local prototype aggregation pixels to provide references for reducing mismatch fusion; 2) Decomposing features into different frequencies to provide more fine-grained interaction elements. Moreover, the computational complexity of vanilla CA is $\mathcal{O}(H^2W^2)$, while ours is $\mathcal{O}(W_s^2)$ or $\mathcal{O}(C^2)$ (more accurately proportional to the rank), significantly reduced. Correspondingly, we integrate spatial and channel refinements, focusing on high-interest regions from two dimensions, further enhancing overall performance. To further explore the differences in frequency decomposition methods, we compare with three traditional methods. We analyze that the performance differences stem from: 1) Traditional methods lacking sensitivity to changes in feature content, typically based on fixed mathematical transformations; 2) Complex transformations in underwater scenes, where RGB and depth maps carry massive noises, interfering with decomposition. Additionally, since the processed features are high-dimensional data, their complexity is considerably higher compared to octave convolutions.

4) Analysis of Expert Design: In Table VIII, when we do not decouple $\hat{\mathbf{M}}$, the MoFE transforms into spatial experts to modulate RGB and depth features, *i.e.,* Model I. In comparison to single-type frequency experts, spatial experts exhibit superiority; however, when combined, the opposite holds true, as in Models II-IV. Reversing the frequency order *i.e.,* high-frequency and low-frequency experts modulate depth and RGB representations separately—yields results inferior to spatial experts. Our analysis reveals: 1) Modal representation differences dictate expert attributes *i.e.,* RGB features emphasize details, while depth features highlight semantics, and they are non-interchangeable; 2) Frequencies possess combinatory characteristics, making it challenging for a single frequency to comprehensively characterize feature content. By leveraging $\mathcal{L}_{\text{balance}}$ for constraint to balance routing, we further enhance performance. The fusion of convolutions across different scales in horizontal and vertical perceptual directions outperforms vanilla convolutions. We analyze that modeling the motion direction, position of underwater objects, and scale variations

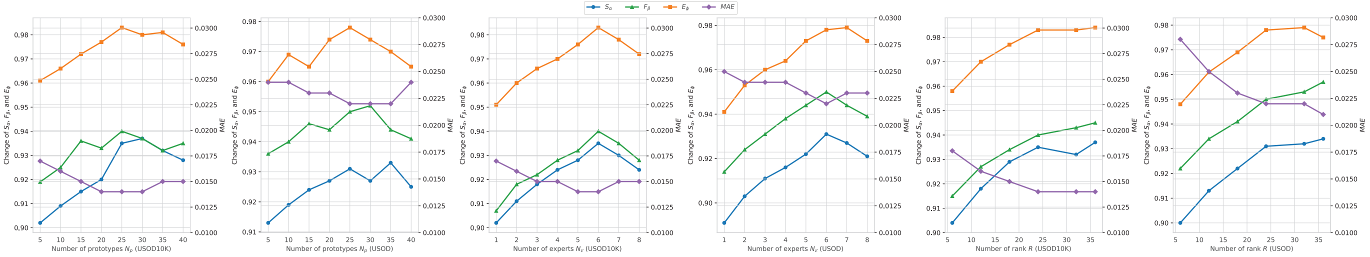


Fig. 7. Quantitative ablation study on hyperparameters: number of prototypes N_p, number of experts $N_\mathcal{E}$, and low-rank approximation R.

due to shooting angles is particularly crucial, unlike natural scenes.

5) Analysis of Supervision: In Table VIII, we employ three supervision strategies: 1) Binary maps for all branches; 2) Detail and trunk maps for RGB and depth branches, with fusion branches corresponding to binary maps; 3) Supervision reversal for RGB and depth branches in Strategy 2. We observe that Strategy 2 performs the best while Strategy 3 performs the worst. Our analysis suggests that misalignment between supervision and representation may disrupt gradient optimization, and conversely, can promote.

6) Analysis of Hyperparameter Settings: In Figure 7, we consider three key parameters: the number of frequency prototypes N_p, experts $N_\mathcal{E}$, and the rank R. As N_p increases from 1 to 25, the model benefits from finer spatial partitioning, which enhances boundary sensitivity and captures localized structures in underwater scenes. However, beyond $N_p = 25$, the performance declines, aligning with the classic granularity-noise trade-off in superpixel clustering [69]: excessive prototypes fragment semantically coherent regions, amplifying noise from depth inaccuracies or texture artifacts. Thus, $N_p = 25$ emerges as an optimal point where spatial resolution is maximized without compromising structural coherence. Similarly, increasing $N_\mathcal{E}$ from 1 to 6 improves performance by encouraging functional specialization among the experts. Each expert focuses on distinct modalities (*e.g.,* RGB textures or depth semantics) or frequency domains (high or low), enabling richer and disentangled representations. When $6 < N_\mathcal{E}$, redundancy arises as overlapping feature subspaces dilute specialization, while increased parameters risk overfitting under limited training data, just like diminishing returns in ensemble learning [70]. For rank selection, the model performs best at $R = 24$, in line with the information bottleneck [71]. This setting preserves the most informative components while suppressing irrelevant variations. A smaller rank ($R < 24$) fails to retain key high-variance features crucial for distinguishing objects from complex backgrounds, whereas a larger rank ($R > 24$) captures noise-dominated subspaces, harming generalization. This aligns with singular value decay patterns [72] in underwater feature spaces, where the first 24 principal components encapsulate the majority of discriminative information.

F. Broader Impacts

We compare across seven natural scene benchmarks.

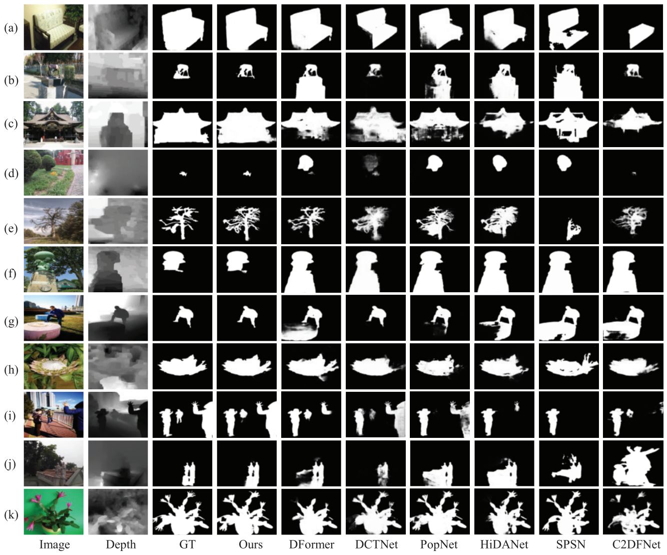


Fig. 8. Qualitative comparison on natural scenes.

1) Methods for Comparisons: Ten models, *i.e.,* C2DFNet [58], SPSN [28], HiDANet [35], PICRNet [8], XMSNet [9], PopNet [36], DCTNet [59], DFormer [16], CPNet [6], and VSCode [60] are selected.

2) Quantitative Comparison: In Table IX, our method achieves the best performance across natural scene benchmarks. On the large-scale STERE dataset, it outperforms CPNet, *i.e.,* the strongest method without additional data, by +1.4%, +1.9%, and +1.4% in S_α, F_β^a, and E_ϕ^m, respectively. Compared with underwater, natural scenes typically contain clearer structures, richer textures, and more diverse semantic content. The HEHP effectively separates and integrates multi-scale, multi-frequency cues, enabling more accurate localization and boundary perception under such complex visual patterns. The incorporation of contrastive learning and uncertainty modeling further enhances robustness, leading to consistent improvements across datasets without relying on external supervision or large-scale pretraining. In Table X, we provide pseudo-depth maps for RGB images and also achieve promising results on two large SOD benchmarks.

3) Qualitative Comparison: In Figure 8, 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

TABLE IX

QUANTITATIVE COMPARISON ON FIVE RGB-D NATURAL SCENES BENCHMARKS

Methods	STERE [53]				SIP [54]				NJU2K [55]				NLPR [56]				DUT [57]			
	$S_\alpha \uparrow$	$F_\beta^a \uparrow$	$E_\phi^m \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta^a \uparrow$	$E_\phi^m \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta^a \uparrow$	$E_\phi^m \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta^a \uparrow$	$E_\phi^m \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta^a \uparrow$	$E_m \uparrow$	MAE$\downarrow$
C2DFNet [58] TMM22	0.902	0.880	0.927	0.038	0.871	0.866	0.912	0.052	0.899	0.897	0.919	0.038	0.899	0.894	0.958	0.021	0.933	0.932	0.958	0.026
SPSN [28] ECCV22	0.906	0.874	0.941	0.035	0.891	0.884	0.932	0.043	0.918	0.887	0.949	0.032	0.923	0.891	0.956	0.023	‡	‡	‡	‡
HiDANet [35] TIP23	0.911	0.897	0.944	0.035	0.892	0.864	0.925	0.043	0.926	0.922	0.951	0.029	0.930	0.908	0.959	0.021	‡	‡	‡	‡
PICRNet [8] MM23	0.921	0.905	0.951	0.031	‡	‡	‡	‡	0.927	0.919	0.952	0.029	0.935	0.911	0.965	0.019	0.943	0.943	0.967	0.020
XMSNet [9] MM23	0.927	‡	‡	0.026	0.913	‡	‡	0.032	0.931	‡	‡	0.025	0.936	‡	‡	0.018	‡	‡	‡	‡
PopNet [36] ICCV23	0.917	0.906	0.947	0.033	0.897	0.893	0.937	0.040	0.924	0.919	0.952	0.030	0.932	0.911	0.963	0.019	‡	‡	‡	‡
DCTNet [59] TIP24	0.920	0.890	0.941	0.035	0.915	0.911	0.945	0.034	0.929	0.912	0.945	0.031	0.933	0.889	0.952	0.023	0.944	0.940	0.960	0.024
DFormer [16] ICLR24	0.925	0.908	0.954	0.029	0.908	0.908	0.942	0.035	0.933	0.925	0.958	0.025	0.936	0.913	0.963	0.019	0.940	0.943	0.964	0.023
CPNet [6] IJCV24	0.920	0.903	$\underline{0.954}$	0.029	0.908	$\underline{0.916}$	0.942	0.035	0.934	$\underline{0.933}$	0.959	0.025	$\underline{0.939}$	**0.924**	$\underline{0.969}$	$\underline{0.016}$	$\underline{0.951}$	$\underline{0.955}$	$\underline{0.972}$	$\underline{0.019}$
VSCode [60] CVPR24	$\underline{0.928}$	$\underline{0.914}$	0.951	$\underline{0.029}$	$\underline{0.917}$	**0.923**	$\underline{0.950}$	$\underline{0.031}$	$\underline{0.940}$	0.932	$\underline{0.961}$	$\underline{0.024}$	0.938	0.914	0.961	0.018	‡	‡	‡	‡
Ours	**0.934**	**0.922**	**0.968**	**0.023**	**0.928**	0.911	**0.968**	**0.025**	**0.943**	**0.939**	**0.975**	**0.018**	**0.946**	$\underline{0.921}$	**0.980**	**0.013**	**0.953**	**0.955**	**0.981**	**0.015**

TABLE X

QUANTITATIVE COMPARISON ON RGB SOD BENCHMARKS

Method	DUTS-TE [61]				VizWiz-SO [62]			
	$S_\alpha \uparrow$	$F_\beta^m \uparrow$	$E_\phi^m \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta^m \uparrow$	$E_\phi^m \uparrow$	MAE $\downarrow$
CPNet [6] IJCV24	0.912	0.918	0.930	$\underline{0.029}$	0.911	0.945	$\underline{0.951}$	0.040
Dual-SAM [40] CVPR24	$\underline{0.918}$	$\underline{0.923}$	**0.937**	0.030	$\underline{0.917}$	$\underline{0.952}$	0.948	$\underline{0.037}$
Ours	**0.927**	**0.930**	$\underline{0.935}$	**0.026**	**0.928**	**0.960**	**0.958**	**0.034**

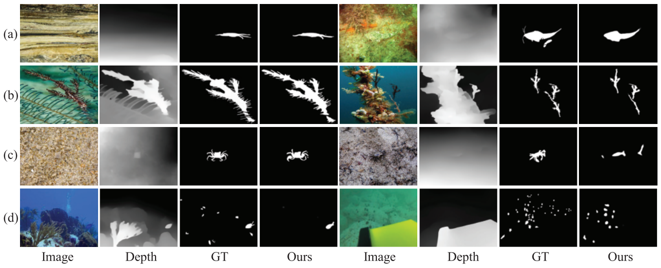


Fig. 9. Failure cases.

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 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).

G. Failure Samples and Future Work

In Figure 9, our method encounters significant limitations in handling composite scenarios of low-quality depth maps (row a), highly irregular shapes (row b), low contrast (row c), and multiple small objects (row d). Low-quality depth maps introduce spatial noise that distorts the uncertainty distribution, impairing the UI's ability to localize informative regions and suppress unreliable cues during fusion accurately. This distortion is particularly problematic in irregularly shaped objects, where the frequency-based experts struggle to maintain consistent contour continuity, resulting in fragmented attention and incomplete saliency detection. In low-contrast settings, both RGB and depth modalities are compromised, *i.e.,* RGB lacks sufficient gradient variation for effective boundary extraction, while noisy depth cues become less informative, hindering prototype alignment and expert modulation. Additionally, when multiple small objects are present, the degraded quality of depth maps further complicates fusion, making it difficult to isolate individual targets and accurately capture their features. These challenges reveal the limitations of both feature disentanglement and expert coordination under degraded input conditions, pointing to areas for improving robustness and modality reliability. Therefore, our future work focuses on: 1) Designing the unified paradigm for extremely complex scenes; 2) Optimizing the model efficiency, especially the expert strategies. 3) Expanding to more scenarios, *e.g.,* mirror detection [73] and aerial perception [74].

V. CONCLUSION

In this paper, we rethink the existing USOD and RGB-D SOD paradigms and propose the HEHP framework based on expert and hierarchical representation learning. We facilitate modal feature denoising and coupling through constructing frequency prototypes and fine-grained interactions. We observe the inherent differences between the RGB and depth branches, explicitly designing high-low frequency experts for modulation, and based on the FFE to fuse heterogeneous four-class representations, avoiding biases introduced by static fusion. Considering that noise carried by depth signals may affect fine-grained features, we perform uncertainty modeling. We further apply different supervisions to the three branches to implicitly learn differences. In addition, we propose prototype contrasts between modalities and images from both global and patch perspectives to learn aligned compact representations. Extensive experiments on eleven datasets validate the effectiveness and transferability of the proposed methods and components.

References

[1] L. Jiang, M. Xu, X. Wang, and L. Sigal, "Saliency-guided image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16509–16518.

[2] S. M. H. Miangoleh, Z. Bylinskii, E. Kee, E. Shechtman, and Y. Aksoy, "Realistic saliency guided image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 186–194.

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

[4] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[5] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D salient object detection via collaborative learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 52–69.

[6] X. Hu, F. Sun, J. Sun, F. Wang, and H. Li, "Cross-modal fusion and progressive decoding network for RGB-D salient object detection," *Int. J. Comput. Vis.*, vol. 132, no. 8, pp. 3067–3085, Aug. 2024.

[7] R. Cong et al., "CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 6800–6815, 2022.

[8] R. Cong et al., "Point-aware interaction and CNN-induced refinement network for RGB-D salient object detection," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 406–416.

[9] Z. Wu et al., "Object segmentation by mining cross-modal semantics," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3455–3464.

[10] 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.

[11] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.

[12] R. Liu, J. Cao, Z. Lin, and S. Shan, "Adaptive partial differential equation learning for visual saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3866–3873.

[13] X. Deng, P. Zhang, W. Liu, and H. Lu, "Recurrent multi-scale transformer for high-resolution salient object detection," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7413–7423.

[14] J. Deng et al., "RGB-D salient object ranking based on depth stack and truth stack for complex indoor scenes," *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109251.

[15] M. Zha et al., "Dual domain perception and progressive refinement for mirror detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 11, pp. 11942–11953, Nov. 2024.

[16] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou, "DFormer: Rethinking RGBD representation learning for semantic segmentation," 2023, *arXiv:2309.09668*.

[17] M. Jahidul Islam, R. Wang, and J. Sattar, "SVAM: Saliency-guided visual attention modeling by autonomous underwater robots," 2020, *arXiv:2011.06252*.

[18] L. Hong, X. Wang, G. Zhang, and M. Zhao, "USOD10K: A new benchmark dataset for underwater salient object detection," *IEEE Trans. Image Process.*, vol. 34, pp. 1602–1615, 2023.

[19] J. Jin, Q. Jiang, Q. Wu, B. Xu, and R. Cong, "Underwater salient object detection via dual-stage self-paced learning and depth emphasis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 3, pp. 2147–2160, Mar. 2025.

[20] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.

[21] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," 2017, *arXiv:1701.06538*.

[22] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[23] Y. Gu et al., "Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 15890–15902.

[24] D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia Biometrics*, vol. 741. NY, USA: Springer, 2009, pp. 659–663.

[25] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2016, pp. 6402–6413.

[26] G. E. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. Hoboken, NJ, USA: Wiley, 2011.

[27] A. Shapiro, "Monte Carlo sampling methods," in *Handbooks in Operations Research and Management Science*, vol. 10. Amsterdam, The Netherlands: Elsevier, 2003, pp. 353–425.

[28] M. Lee, C. Park, S. Cho, and S. Lee, "SPSN: Superpixel prototype sampling network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, 2022, pp. 630–647.

[29] Z. Zhang, J. Wang, and Y. Han, "Saliency prototype for RGB-D and RGB-T salient object detection," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3696–3705.

[30] J. Wu, F. Hao, W. Liang, and J. Xu, "Transformer fusion and pixel-level contrastive learning for RGB-D salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 1011–1026, 2023.

[31] Y. Chen et al., "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3435–3444.

[32] F. Yang et al., "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4146–4155.

[33] W. Ji et al., "Calibrated RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9471–9481.

[34] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving RGB-D saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4681–4691.

[35] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, and C. Demonceaux, "HiDAnet: RGB-D salient object detection via hierarchical depth awareness," *IEEE Trans. Image Process.*, vol. 32, pp. 2160–2173, 2023.

[36] Z. Wu et al., "Source-free depth for object pop-out," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1032–1042.

[37] F. Sun, P. Ren, B. Yin, F. Wang, and H. Li, "CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 2249–2262, 2023.

[38] W. Zhou, F. Sun, and W. Qiu, "MSNet: Multiple strategy network with bidirectional fusion for detecting salient objects in RGB-D images," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 4341–4353, 2025.

[39] R. Guo, X. Ying, Y. Qi, and L. Qu, "UniTR: A unified TRansformer-based framework for co-object and multi-modal saliency detection," *IEEE Trans. Multimedia*, vol. 26, pp. 7622–7635, 2024.

[40] P. Zhang, T. Yan, Y. Liu, and H. Lu, "Fantastic animals and where to find them: Segment any marine animal with dual SAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2578–2587.

[41] Z. Luo et al., "VSCode: General visual salient and camouflaged object detection with 2D prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17169–17180.

[42] L. Li, E. Rigall, J. Dong, and G. Chen, "MAS3K: An open dataset for marine animal segmentation," in *Proc. Int. Symp. Benchmarking, Measuring Optim.*, Jan. 2021, pp. 194–212.

[43] Z. Fu, R. Chen, Y. Huang, E. Cheng, X. Ding, and K.-K. Ma, "MASNet: A robust deep marine animal segmentation network," *IEEE J. Ocean. Eng.*, vol. 49, no. 3, pp. 1104–1115, Jul. 2023.

[44] M. Jahidul Islam, P. Luo, and J. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," 2020, *arXiv:2002.01155*.

[45] P. Drews- Jr., I. D. Souza, I. P. Maurell, E. V. Protas, and S. S. C. Botelho, "Underwater image segmentation in the wild using deep learning," *J. Brazilian Comput. Soc.*, vol. 27, no. 1, pp. 1–14, Dec. 2021.

[46] J. Liu, J. Zhang, and N. Barnes, "Modeling aleatoric uncertainty for camouflaged object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1445–1454.

[47] 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.

[48] 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.

[49] A. He, K. Wang, T. Li, C. Du, S. Xia, and H. Fu, "H2Former: An efficient hierarchical hybrid transformer for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 9, pp. 2763–2775, Sep. 2023.

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

[51] T. Chen et al., "SAM fails to segment anything?-SAM-adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, medical image segmentation, and more," 2023, *arXiv:2304.09148*.

[52] T. Yan, Z. Wan, X. Deng, P. Zhang, Y. Liu, and H. Lu, "MAS-SAM: Segment any marine animal with aggregated features," 2024, *arXiv:2404.15700*.

[53] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.

[54] 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.

[55] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1115–1119.

[56] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.

[57] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7254–7263.

[58] M. Zhang, S. Yao, B. Hu, Y. Piao, and W. Ji, "C^{2}DFNet: Criss-cross dynamic filter network for RGB-D salient object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 5142–5154, 2023.

[59] H. Chen, F. Shen, D. Ding, Y. Deng, and C. Li, "Disentangled cross-modal transformer for RGB-D salient object detection and beyond," *IEEE Trans. Image Process.*, vol. 33, pp. 1699–1709, 2024.

[60] Z. Luo et al., "Vscode: General visual salient and camouflaged object detection with 2D prompt learning," 2023, *arXiv:2311.15011*.

[61] L. Wang et al., "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 136–145.

[62] J. Reynolds, C. K. Nagesh, and D. Gurari, "Salient object detection for images taken by people with vision impairments," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 8507–8516.

[63] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[64] 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.

[65] 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*.

[66] 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.

[67] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.

[68] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[69] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[70] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 241–258, Apr. 2020.

[71] A. Saxe et al., "On the information bottleneck theory of deep learning," *J. Stat. Mech., Theory Exp.*, vol. 2019, no. 12, Dec. 2019, Art. no. 124020.

[72] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A Practical Approach to Microarray Data Analysis*. Cham, Switzerland: Springer, 2003, pp. 91–109.

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

[74] M. Zha, W. Qian, W. Yang, and Y. Xu, "Multifeature transformation and fusion-based ship detection with small targets and complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.