FSCDiff: Frequency-Spatial Entangled Conditional Diffusion model for Underwater Salient Object Detection

Anonymous Author(s)

Abstract

Salient object detection (SOD) plays a crucial role in image understanding and visual guidance. However, due to the complexity of underwater environments, the accuracy of underwater salient object detection is often low. To enhance the accuracy and robustness of underwater salient object detection, existing RGB-D multimodal methods typically represent scene features from the spatial domain, but rarely explore the characteristics of different modalities in the frequency domain. Different from the existing methods that perceive RGB-D in the spatial domain, we propose a novel Frequency-Spatial Entangled Conditional Diffusion(FSCDiff) framework for underwater salient object detection. The FSCDiff aims to address the insufficient representation and boundary shift issues in underwater salient object detection by leveraging frequency-domain information and the powerful multi-step iterative generation capability of diffusion models. The FSCDiff framework consists of two key components: the Frequency-Spatial Entanglement Enhancement Block (DTEB) and the Stable Time-step Mask Prediction Module (STMP). DTEB utilizes frequency-spatial entanglement learning to fully exploit the frequency and spatial domain information of RGB images and depth maps, thereby optimizing feature representation. STMP takes advantage of the excellent multi-step iterative mechanism of diffusion models to enhance the accuracy and robustness of the segmentation results. Comprehensive experimental results indicate that our FSCDiff method outperforms the state-of-the-art approaches on the USOD10K and USOD datasets. The code will be made publicly available upon the acceptance of the paper.

CCS Concepts

• **Computing methodologies** → **Computer vision problems**;

Keywords

Underwater Salient Object Detection, MultiModal, Diffusion, Fourier Frequency Information

1 Introduction

Inspired by the human visual system, the objective of Salient Object Detection(SOD) is to identify the visually most prominent objects within a given scene, which holds significant value in image understanding and visual guidance. Relevant technologies can serve various practical application scenarios, including object tracking [1], image editing [2], action recognition [3], image understanding [4], virtual reality [5], and so on. In recent years, deep learning-based methods have achieved remarkable progress in SOD task within terrestrial natural scenes. However, in complex scenes such as underwater scenes, the performance of SOD still faces many challenges and needs further improvement. Compared with SOD in terrestrial natural images, underwater images usually have quality defects such as low contrast and low visibility due to light absorption, refraction and scattering in the underwater scenes. In addition,

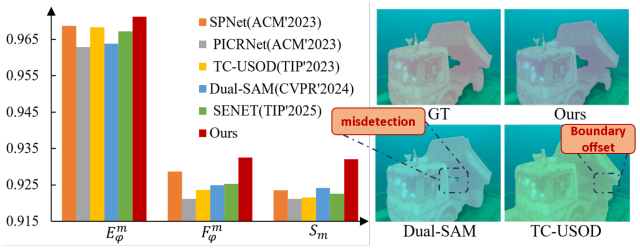

Figure 1: A simple comparison of FSCDiff and other state-of-the-art methods trained on the USOD10K dataset.

the presence of plankton and turbulence in water introduces varying degrees of noise interference, further reducing the quality of underwater images [6]. With the rapid development of depth sensing technology, the RGB-D SOD has shown excellent performance in various complex scenes. The purpose of multimodal image fusion lies in effectively integrating the unique scene-related information and details from each source image to generate a more comprehensive representation. Recently, many RGB-D feature fusion methods for SOD have been proposed, and their fusion strategies mainly fall into two paradigms: the first is to inject depth features as cues into RGB features; the second is to employ attention mechanisms to fuse the two modalities. For example, Yao et al. [7] proposed directly injecting depth cues into RGB features to enhance the semantic representation of RGB features; Wu et al. [8] used spatial attention and channel attention to fuse multi-modal and multi-level features for subsequent use. However, the above works have not paid attention to the problem that RGB images and depth maps in complex scenes present different degrees of quality degradation, which often leads to the reduction of object segmentation accuracy and misidentification when the model encounters difficult samples. Further analysis reveals that current SOD models mainly focus on fusing these two modalities from the spatial domain perspective, which has obvious limitations in representational capacity and makes it difficult to mine more potential useful information. Although spatial domain information is beneficial to the SOD task to a certain extent, its local dependence on pixel-level information makes it vulnerable to the interference of complex backgrounds [7]. Moreover, in underwater scenes, the interference of background information often leads to boundary shift issues in SOD results. Figure 1 shows the comparison results of our method with other state-of-the-art methods. To highlight these problems, we convert the binary masks generated by the models into semi-transparent color masks and overlay them on the RGB images. As shown in Figure 1, due to background interference, Dual-SAM fails to obtain sufficient representation in the spatial domain and mistakenly considers the gap of the truck as part of the truck. Additionally, due to the boundary blurring caused by noise interference, TC-USOD experiences boundary shift at the blurred boundaries. Therefore, how to overcome the limitations of

single spatial domain features has become one of the key issues in obtaining accurate SOD results in underwater scenes. Recently, the combination of frequency domain information and spatial domain information has been proven to be an effective means of obtaining more potential information [9][10][11]. The frequency domain features of images obtained through Fourier transform have global characteristics. Specifically, the Fourier transform can decompose an image into phase and amplitude components. The phase component contains the structural information of the image, while the amplitude component is closely related to the contrast and texture details of the image [12]. We observe that in underwater scenes, the amplitude information of RGB images helps to distinguish objects from the background; the phase information in depth maps provides clear structure and contours, enhancing the ability to distinguish object shapes and boundaries. For more details, please refer to Part A of the supplementary material. Please see more details in Section A in the supplementary materials. Recently, methods based on diffusion models have received great attention due to their excellent performance in image-generation tasks [13][14][15]. These methods utilize the iterative mechanism of diffusion models to achieve high-quality mapping from randomly sampled Gaussian noise to the target image or latent distribution. Some studies have shown that diffusion models can also achieve promising results in tasks such as object detection, semantic segmentation, and instance segmentation[16][17][18]. For instance, MedSegDiffV2 [19] integrates the conditional features generated by the transformer into the diffusion model to guide the segmentation process; DiffusionDet [20] applies the diffusion model to the object detection framework, defining object detection as a denoising diffusion process from noise to bounding boxes. Inspired by this, we introduce the diffusion model into our proposed underwater RGB-D SOD architecture, using it as a mask generation task rather than a traditional segmentation task. We leverage its excellent generative patterns and denoising generalization to address the interference caused by the underwater environment. Moreover, we observe that during the multi-step iterative denoising process of the diffusion model, multiple intermediate results are produced, each aiming to generate an accurate segmentation mask. These intermediate results are highly random in the early stages of the iterative process due to the influence of noise distribution but gradually stabilize as the iterations progress. By integrating these diverse intermediate results, especially their different performances on object boundaries and contours, we can provide richer information for the final mask generation, thereby improving the accuracy and robustness of the segmentation results.

Based on the above analysis, this paper proposes a novel Underwater Salient Object Detection (USOD) framework, named FSCDiff. This framework achieves effective detection of salient objects in complex underwater scenes by integrating spatial-frequency domain information and the characteristics of diffusion models. Specifically, this method employs a Dual-Domain Entanglement Enhancement mechanism to extract more comprehensive feature representations, which are then used as conditional features input to the diffusion model, thereby guiding the model to generate target masks more stably. In terms of frequency domain modeling, we design a Frequency Perception Enhancement Block to capture the global complementary representations of depth maps and RGB images in the frequency domain. Additionally, in the spatial branch, we conduct spatial feature entanglement learning from a multi-scale perspective for depth maps and RGB images to adapt to objects of different sizes and enhance feature expression capabilities. To address the unique representations in the spatial and frequency domains, we introduce an attention mechanism for adaptive learning in both channel and spatial dimensions, further strengthening the discriminative power of features and effectively suppressing redundant information. In the diffusion model branch, the fused features optimized via spatial-frequency entanglement serve as conditional inputs, guiding the diffusion model to concentrate on the mask generation task for the target region while enhancing the efficiency of the overall prediction process. Moreover, to better utilize the intermediate results generated by the multi-step iterations of the diffusion model, we propose a Stable Time-Step Mask Prediction strategy based on change rate, aiming to capture more meaningful prediction information and refine the results.

Our main contributions can be concluded as follows:

(1) We propose a novel underwater salient object detection framework, FSCDiff, which utilizes frequency domain information supplementation and the multi-step iterative properties of diffusion models to address two existing issues in underwater RGB-D image salient object detection: insufficient representation and boundary shift.

(2) To effectively exploit the modal superiority of RGB images and depth maps, we have devised a frequency-space entangled differential attention module. This module permits the entangled learning of RGB features and depth map features within the frequency-space domain and employs a de-redundant spatial-channel attention to accentuate key information, thereby attaining a more comprehensive comprehension of the data.

(3) We analyzed the theoretical basis for the validity of the intermediate results of multi-step iterative prediction in diffusion models and proposed a Time-Step Mask Prediction strategy (STMP) based on stable changes. This strategy can filter out beneficial prediction maps for integration, improve the accuracy and stability of the prediction results, and enhance the model's robustness in complex scenarios.

(4) The experimental results demonstrate that our method achieves state-of-the-art performance in RGB-D salient object detection on multiple benchmark tests. Additionally, ablation experiments verify the effectiveness of the key modules proposed in our method.

2 Related Work

2.1 RGB-D Salient Object Detection

Traditional RGB-D salient object detection (SOD) methods mainly rely on handcrafted features, which have inherent limitations in capturing the complex relationships between modalities, often resulting in poor detection performance. In recent years, deep learning-based RGB-D SOD methods have begun to emerge and have demonstrated extraordinary performance. In 2022, Cong et al. [21] proposed CIR-Net, which achieved cross-modal interaction through progressive attention guidance and gated fusion units in the encoder and decoder stages, demonstrating the importance of integrating multi-modal information at multiple scale levels. In 2023, Qiu et al. [22] proposed the asymmetric bilateral U-Net model - ABiUNet,

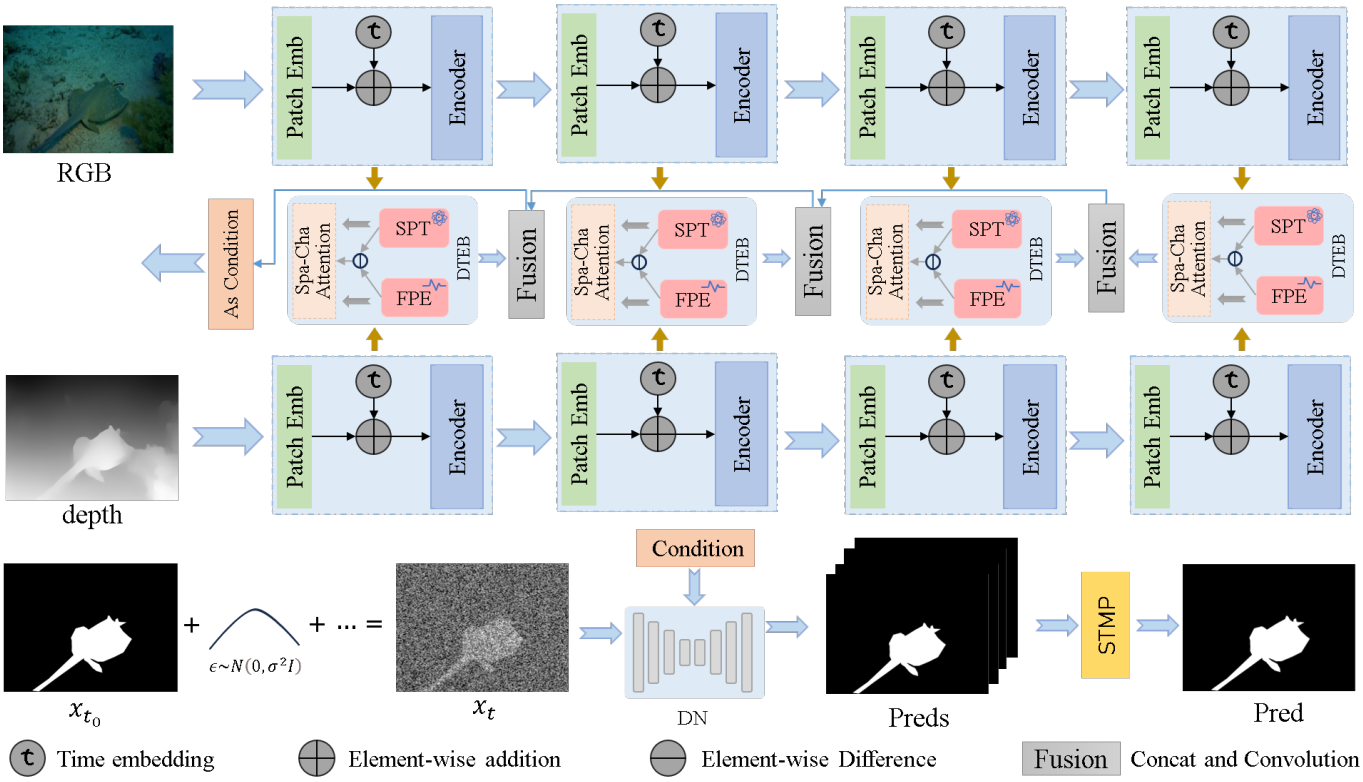

Figure 2: An overview of the FSCDiff model framework proposed in SOD. The proposed FSCDiff method generates rich and representative conditional features by incorporating the Spatial Perception Transformer Block (SPT), Frequency Perception Enhancement Block (FPE), and Redundancy-Free Spatial Channel Attention (RSCA) into the Dual-domain Entanglement Enhancement Block (DTEB) to guide the Diffusion Model(DN). The Stable Time-Step Mask Prediction Module (STMP) further integrates and optimizes the multi-step prediction results.

combining the advantages of Transformer and CNN to jointly handle the task of Salient Object Detection. In 2023, Zhang et al. [7] proposed SPNet, which utilized the clearer definition of salient regions provided by saliency prototypes to address the quality issue of the auxiliary modality. In 2023, Wu et al. [23] proposed HiDAnet, using a granularity-based attention scheme to enhance the discriminative power of each modality branch, respectively. In 2024, Liu et al. [24] proposed VST++, dividing the foreground into fine-grained fragments and aggregating background information into a single coarse-grained label, thereby reducing computational costs. In 2025, Zhan et al. [8] proposed MambaSOD, leveraging the linear long-range modeling capability of Mamba to perform module-specific feature enhancement. However, these previous methods mainly improved detection accuracy from the spatial domain perspective, lacking exploration of the frequency domain between RGB images and depth maps, resulting in an inability to effectively utilize the representational power of deep models.

2.2 Diffusion Model

In recent years, diffusion probability models have been extensively employed in numerous domains, such as image generation [15][25][26], image segmentation [19][27][28], object detection [29][30], and super-resolution tasks [31][32][33], et al. These successful applications have attested to the potent capability of diffusion models in generating high-quality and diversified data samples. Inspired by these applications, we have discerned that diffusion models exhibit an extremely high degree of suitability for underwater salient object detection. In the underwater environment, owing to the inherent environmental interferences, boundary prediction has consistently constituted a significant challenge. The multi-step iterative property of diffusion models allows for the attainment of multiple iterative results during the continuous refinement process. These intermediate results are all gradual approximations to the expected outcome. Integrating these intermediate results helps to better identify object features and more stably predict boundaries. Nevertheless, assessing the reliability of the intermediate prediction results generated by diffusion models during the iterative process remains a crucial concern.

3 Method

3.1 Network Architecture

In this paper, we present a novel underwater salient object detection (USOD) framework, namely FSCDiff, to resolve two core problems in underwater RGB-D image salient object detection,

namely, insufficient representation and boundary shift, through the supplementation of frequency domain representation and the powerful multi-step iterative generation ability of the diffusion model. The architecture of FSCDiff is shown in Figure 2, mainly consisting of two major parts: the Dual-Domain Entanglement Enhancement Block (DTEB) and the diffusion model with a Stable Time-step Mask Prediction (STMP) module. For the given input image $I_c \in \mathbb{R}^{H\times W\times 3}$, depth map $I_d \in \mathbb{R}^{H\times W\times 3}$ and x_t, we first use the basic encoders (i.e., $PVTv4-m$ and $PVTv1$ [34]) to extract the initial features $P_c = \{\mathcal{P}_i\}_{i=1}^4$ and $P_d = \{\mathcal{P}_i\}_{i=1}^4$ from the RGB image and depth map respectively at the resolution of $\frac{W}{2^{i+1}} \times \frac{H}{2^{i+1}}$. To fully integrate the characteristics of spatial domain information and Fourier domain information, we designed the Dual-Domain Entanglement Enhancement Block (DTEB). The DTEB consists of the Spatial Perception Transformer (SPT), Frequency Perception Enhancement (FPE), and the Redundancy-Free Spatial Channel Attention (RSCA). The SPT block utilizes the spatial attention mechanism to highlight the features of the target region and suppress background noise; the FPE block maps the features to the frequency domain through Fourier transform to extract global structural information and enhance the edge features of the target. The RSCA Module further optimizes the feature representation by calculating the difference between RGB features and depth map features to suppress redundant information and highlight useful features. Through this design, DTEB achieves entangled learning in the spatial and frequency domains, fully leveraging the complementary information of the two modalities. After extracting the optimized features, we pass these features as conditional inputs to the diffusion model (DN) to guide it in accurately detecting salient objects. The diffusion model gradually generates the mask of salient objects from noise. During the iterative process of the diffusion model, multiple intermediate results are produced, which are highly random in the early stages but gradually stabilize as the iterations proceed. To effectively utilize the diverse predictions of these intermediate results, we propose a Stable Time-step Mask Prediction (STMP) strategy. By integrating these diverse intermediate results, more abundant information can be provided for the final mask generation, thereby improving the accuracy and robustness of the segmentation results.

3.2 Dual-Domain Entanglement Enhancement Block (DTEB)

Unlike previous methods[35][36][37] that only learn the feature dependencies between modalities based on spatial modeling, our DTEB combines different relationships from the frequency domain and the spatial domain. This module exploits the characteristics of both RGB and depth maps in the spatial and frequency domains, enabling the entangled learning of features from the two modalities in different domains and allowing the integration of information such as color, texture, edge, spectrum, amplitude, and energy. This approach is beneficial for learning discriminative representations by considering the features of RGB images and depth maps in different domains and using various types of information. Moreover, we also suppress the redundant information existing in the fusion process and reduce feature interference through a differential attention strategy. As shown in Figure 3, our DTEB consists of three key parts: Frequency Perception Enhancement(FPE), Spatial Perception

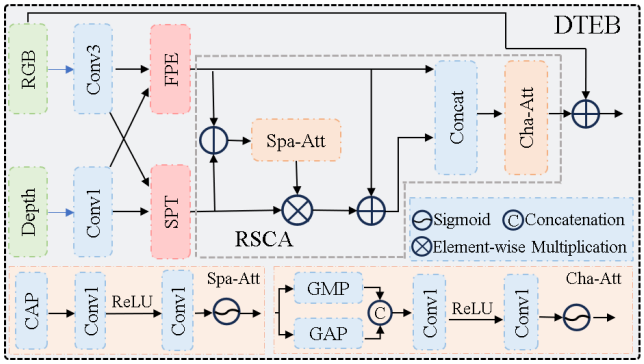

Figure 3: Details of the DTEB. FPE accurately captures global fine details through Fourier fusion. SPT enhances the spatial features of RGB images and depth maps using global attention. Finally, RSCA precisely highlights effective information by calculating the differences in spatial and Fourier information.

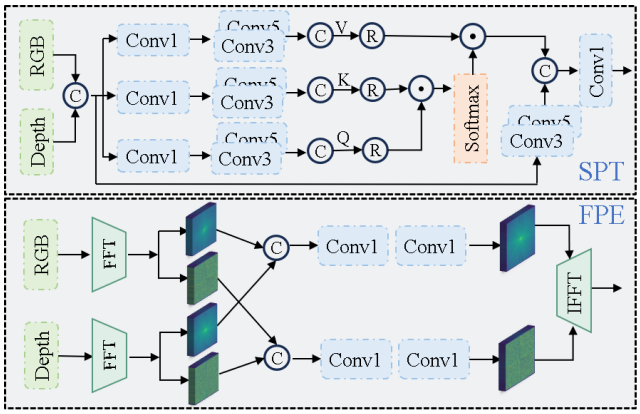

Figure 4: The specific processes of the SPT module and the FPE module.

Transformer(SPT), and Redundancy-Free Spatial Channel Attention (RSCA).

Frequency Perception Enhancement Block (FPE). Considering that in underwater scenes, the amplitude information of RGB images helps identify the bright and dark areas of the image, while the phase information of depth maps contains clearer structural and contour information, we achieve modal information complementarity between the two in the Fourier domain. As shown in Figure 4, we perform a Fast Fourier Transform (FFT) on the modal perception features of the depth map and the RGB image to obtain their amplitude components and phase components, respectively. Specifically, the RGB feature and depth feature of the $i-th$ layer are represented as $\mathbf{F}_c^i$ and $\mathbf{F}_d^i$, respectively, and their Fourier transforms can be expressed as follows:

$$\mathcal{A}\left(\mathbf{F}_c^i\right), \mathcal{P}\left(\mathbf{F}_c^i\right) = \mathcal{F}\left(\mathbf{F}_c^i\right), \tag{1}$$

$$\mathcal{A}\left(\mathbf{F}_d^i\right), \mathcal{P}\left(\mathbf{F}_d^i\right) = \mathcal{F}\left(\mathbf{F}_d^i\right), \tag{2}$$

The $\mathcal{A}(\cdot)$ represents the amplitude, and the $\mathcal{P}(\cdot)$ represents the phase components. Then, we concatenate the $\mathcal{A}(\cdot)$ and the $\mathcal{P}(\cdot)$ of the two respectively to obtain the concatenation(Cat) result of the channel-level frequency-domain joint features($\mathcal{A}\left(\mathbf{F}_{cd}^i\right)$ and $\mathcal{P}\left(\mathbf{F}_{cd}^i\right)$), and employ 1×1 convolution and activation functions to generate the globally optimized frequency-domain representation after entangled learning for the joint features. Finally, for the optimized amplitude and phase components($\left(\mathcal{A}'\left(\mathbf{F}_{cd}^i\right)\right.$ and $\left.\mathcal{P}'\left(\mathbf{F}_{cd}^i\right)\right)$), we convert them back to the original domain using the inverse fast Fourier transform ($ifft(\cdot)$), which is formulated as:

$$\mathbf{X}_f = ifft(\mathcal{A}'(\cdot), \mathcal{P}'(\cdot)) \tag{3}$$

where $\mathbf{X}_f$ represents the result after optimization in the Fourier domain.

Spatial Perception Transformer Block (SPT). In the spatial branch, as shown in Figure 4, we concatenate the modality-aware features $\mathbf{F}_c^i$ and $\mathbf{F}_d^i$ of the depth map and RGB image to obtain the feature $\mathbf{F}_{c+d}^i$. Considering the diversity in the size of underwater objects, we use two depthwise separable convolutions, namely a 3×3 deconvolution and a 5×5 deconvolution, to obtain the query Q, key K, and value V required for self-attention. Then, through the operation of the self-attention module, we obtain V', which contains global spatial information. Additionally, to increase spatial local information, we perform additional multi-scale depthwise separable convolutions on the input feature $\mathbf{F}_{c+d}^i$, and then we use it as a supplement to V' to obtain the final spatial entanglement-optimized feature X_s, as shown in the following formula:

$$\mathbf{X}_s = C_{1\times1}(Concat(V', (D_{35}(F_{c+d})))) \tag{4}$$

where $C_{1\times1}$ denotes the 1×1 convolution operation,D_{35} represents two depthwise separable convolutions with sizes of 3×3 and 5×5 respectively, as well as a concatenation operation.

Redundancy-Free Spatial Channel Attention (RSCA). In multi-modal image fusion, information redundancy often exists between the spatial and frequency branches, which restricts the model's ability to utilize information effectively. To address this problem, we have designed a Redundancy-Free Spatial Channel Attention module aiming to suppress redundant information and highlight useful features. Specifically, we initially conduct a subtraction operation on the features X_f of the frequency branch and X_s of the spatial branch to capture the differential information between them. Subsequently, we apply a spatial attention mechanism $SA(\cdot)$ to the differential features to capture the spatial dependency attention map (S_{s-f} , $S_{s-f} = SA(X_s - X_f)$). The generated spatial attention map S_{s-f} is multiplied by X_s, thereby obtaining $\tilde{S}_{s-f}$, which highlights the advantageous information in the two branches of the spatial dimension. Eventually, we add this result to X_f to obtain the spatially enhanced fusion feature S_{fs}. To further enhance the discriminative ability of feature effectiveness, we concatenate the spatially enhanced fused feature S_{fs} with X_f, and apply the channel attention mechanism $CA(\cdot)$ to it to obtain the attention map of specific channels (C_{fs}, $C_{fs} = CA([S_{fs}, X_f])$). Ultimately, we add the output of the channel attention module C_{fs} to the result in the image domain X_s to prevent gradient vanishing and ensure the stability and convergence of the model. This process not only effectively suppresses redundant information but also enhances the model's ability to capture useful features, thereby improving overall performance. The specific process is as follows:

$$\mathbf{S}_{f-s} = C_{3\times3}([\mathrm{GMP}(X_{f-s}), \mathrm{GAM}(X_{f-s})]) \tag{5}$$

$$\mathbf{C}_{fs} = \sigma(C_{1\times1}(\mathrm{GAP}([\mathbf{S}_{fs}, \mathbf{X}_f]))) \otimes ([\mathbf{S}_{fs}, \mathbf{X}_f]) \tag{6}$$

$$\mathbf{X}_{final} = \mathbf{C}_{fs} + \mathbf{X}_s \tag{7}$$

where GAP denotes global average pooling, GAP denotes global max pooling, σ denotes Sigmoid activation function and $\mathbf{X}_{final}$ denotes the final result after DTEB.

3.3 Stable Time-Step Mask Prediction (STMP)

In the diffusion model branch, we adopt the spatial frequency entanglement-optimized features as the conditional input. This conditional feature will guide the model to focus on the mask generation of specific regions, thereby accelerating the mask generation process. Additionally, according to (Denoising Diffusion Probabilistic Models), we know that in the iterative denoising process of the diffusion model, the early iterations are mainly responsible for extracting the rough features related to the target data from the noise, while the later iterations focus on refining and optimizing the generated results. Moreover, based on the mutual information theory, we recognize that the mask at each time step of the diffusion model contains useful information related to the target data. A stable rather than highly variable iterative process means that the result of the next step is based on the result of the previous step, and it implies that the information transmission is continuous. Therefore, after the rough features of the target are formed, since the noise influence has relatively stabilized, the later stable iterative process can effectively guide the edge generation process. Please see more details in Section B in the supplementary materials. Based on the above theory and discovery, we propose a mask prediction module based on stable time steps. Firstly, by calculating the predicted change rate C between each step, select the predicted values of the n time steps with the lower change rates as the valid predicted values. For the given multiple prediction values, a binary mask xx is first generated using an adaptive threshold, and then the weighted average of these mask results is taken to vote for the most reliable result. The specific process is as follows:

$$\mathbf{C}_{t+1} = |\mathbf{P}_{t+1} - \mathbf{P}_t| \qquad t \in (1, 2, \ldots, \mathrm{T}) \tag{8}$$

$$\mathrm{S} = \mathrm{argmin}_n\left(\mathbf{C}_t\right) \quad t \in (1, 2, \ldots, \mathrm{T}) \tag{9}$$

$$\mathbf{P}_{\text{ensemble}} = \frac{1}{n} \sum_{t \in \mathrm{S}} \mathbf{P}_t \tag{10}$$

where T denotes the total time steps, C_t denotes the degree of change between step t and the previous step, and n denotes the number of masks used for the final refinement, defaulting to $T/2$. argmin_n denotes the selection of the n results with the smallest change rate, and S is the array recording the indices of these results. P_{ensemble} is the final aggregated result.

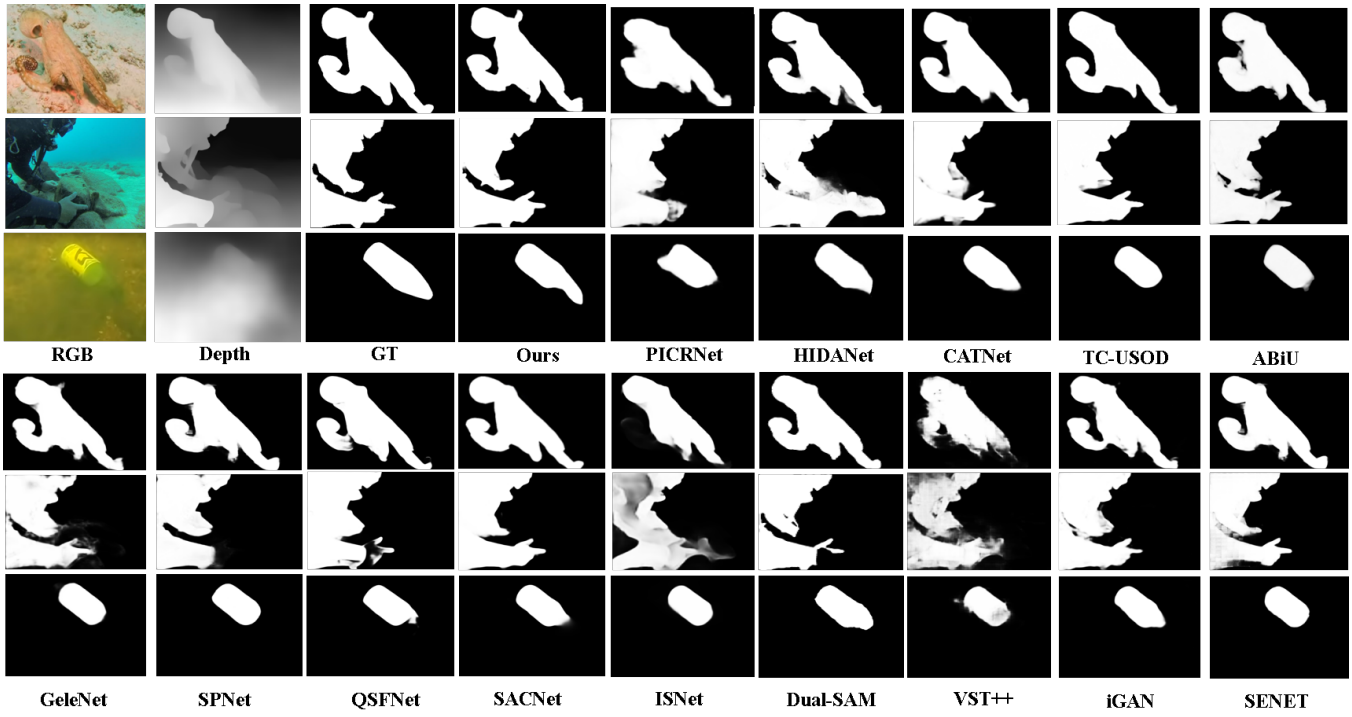

Figure 5: Qualitative comparisons of the proposed FSCDiff and fourteen SOD methods. More details and qualitative comparisons are in Section C in the supplementary materials.

3.4 Loss function

In the developed FSCDiff method, we guide the diffusion model to generate precise prediction maps. More specifically, we utilize the weighted intersection-over-union (IoU) loss and the weighted binary cross-entropy (BCE) loss [38] as optimization functions to refine the model according to the ground truth (G). This loss function can be expressed as:

$$\mathcal{L} = \mathcal{L}_{IoU}^{w}\left(\mathrm{X}_{pred}, \mathrm{G}\right) + \mathcal{L}_{BCE}^{w}\left(\mathrm{X}_{pred}, \mathrm{G}\right), \tag{11}$$

where $\mathcal{L}_{IoU}^{w}$ and $\mathcal{L}_{BCE}^{w}$ denote the weighted IoU and BCE functions. X_{pred} means the prediction.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our FSCDiff on two benchmark datasets: USOD10K[35] and USOD[39]. For the USOD task, we follow the standard protocol established in previous studies. USOD10K is the first large-scale dataset in the field of underwater salient object detection (USOD), containing 10,255 underwater salient images, of which 7,178 images are used as the training set, 2,051 images as the validation set, and 1,026 images as the test set. The USOD dataset is a dedicated test dataset, consisting of 300 salient images, each with corresponding labels provided.

Implementation details. In the Pytorch framework, the experiments were conducted on a 24GB NVIDIA RTX 4090D GPU to implement the proposed xxx model. The backbone of the RGB branch and depth branch of the conditional generation network is PVT, which is initialized with pre-trained parameters from ImageNet. We used an Adam optimizer with an initial learning rate of 1e-5 and gradually reduced the learning rate to 1e-6 over the training period using CosineAnnealingLR. During training and inference, the output image size was adjusted to 352×352, with a batch size of 16, for 150 epochs of training.

Evaluation Metrics. Evaluation Metrics: Evaluation metrics: We use six widely used evaluation metrics, namely Mean Absolute Error ($\mathcal{M}$), Maximum F-measure (F_{φ}^{m}), Average F-measure (F_{φ}^{a}), Maximum E-measure (E_{φ}^{m}), Average E-measure (E_{φ}^{a}), and S-measure (S_m)

4.2 Comparisons with the SOTAs.

We compare FSCDiff with the SOD method in 21, including UCNet [40], S2MA [41], BBSNet [42], CDINet [43], CIRNet [21], PICRNet [44], HIDANet [23], CATNet [45], TC-USOD [35], ABiU [22], GeleNet [46], SPNet [7], QSFNet [47], SACNet [48], ISNet [36], Dual-SAM [49], CoralsCOP [50], VST++ [24], SPDE [51], iGAN [37], SENET [52]. Note that the predicted maps from all methods are provided by the authors or obtained from open-source codes.

Quantitative Evaluation. Table 1 comprehensively compares the quantitative results for our FSCDiff model and 21 other state-of-the-art (SOTA) models. As shown in Table 1, the FSCDiff model demonstrates superior performance across all metrics on the USOD10K dataset and exhibits outstanding capabilities on the USOD dataset as

Table 1: Quantitative comparisons of different SOD methods on the USOD10K and USOD datasets. The best results are highlighted in bold, and the second-best results are underlined.

Method	Year,Pub.	USOD10K(1026 images)						USOD(300 images)					
		$\mathcal{M}\downarrow$	$F_\varphi^m\uparrow$	$F_\varphi^a\uparrow$	$E_\varphi^m\uparrow$	$E_\varphi^a\uparrow$	$S_m\uparrow$	$\mathcal{M}\downarrow$	$F_\varphi^m\uparrow$	$F_\varphi^a\uparrow$	$E_\varphi^m\uparrow$	$E_\varphi^a\uparrow$	$S_m\uparrow$
UCNet	2020,CVPR	0.0335	0.8984	0.8893	0.9443	0.9416	0.8968	0.0493	0.9032	0.8969	0.9245	0.9032	0.9021
S2MA	2020,CVPR	0.0561	0.8589	0.7996	0.9223	0.8706	0.8669	0.0721	0.8714	0.8645	0.9197	0.8725	0.8724
BBSNet	2021,TIP	0.0347	0.9021	0.8597	0.9378	0.9163	0.9003	0.0574	0.9076	0.8946	0.9258	0.8946	0.8974
CDINet	2021,ACM	0.0291	0.9029	0.8809	0.9513	0.9385	0.9083	0.0551	0.8959	0.8812	0.9222	0.9079	0.8858
CIRNet	2022,TIP	0.0294	0.9107	0.8785	0.9568	0.9344	0.9127	0.0542	0.9044	0.8832	0.9285	0.9091	0.8927
PICRNet	2023,ACM	0.0226	0.9212	0.9077	0.9629	0.9578	0.9212	0.0465	0.9087	0.9007	0.9296	0.9235	0.8959
HIDANet	2023,TIP	0.0261	0.9111	0.8958	0.9567	0.9522	0.9109	0.0448	0.9107	0.9034	0.9353	0.9296	0.8942
CATNet	2023,TMM	0.0214	0.9237	0.9021	0.9657	0.9583	0.9207	0.0463	0.9057	0.8957	0.9291	0.9241	0.8954
TC-USOD	2023,TIP	0.0201	0.9236	0.9098	0.9683	0.9634	0.9215	0.0447	0.9103	0.9029	0.9335	0.9277	0.8941
ABiU	2023,TCSVT	0.0266	0.9156	0.8754	0.9613	0.9338	0.9169	0.0502	0.9098	0.8807	0.9342	0.9078	0.8949
GeleNet	2023,TIP	0.0216	0.9249	0.9125	0.9655	0.9605	0.9238	0.0442	0.9168	0.9098	0.9368	0.9298	0.8999
SPNet	2023,ACM	0.0196	0.9287	0.9133	0.9687	0.9641	0.9235	0.0442	0.9107	0.9034	0.9332	0.9284	0.8976
QSFNet	2024,TIP	0.0235	0.9175	0.9064	0.9574	0.9541	0.9158	0.0422	0.9201	**0.9149**	0.9407	**0.9372**	0.9023
SACNet	2024,TMM	0.0191	0.9287	0.9174	0.9693	0.9654	0.9261	0.0417	0.9159	0.9101	0.9363	0.9315	0.8979
ISNet	2024.PR	0.0328	0.9031	0.8764	0.9506	0.9384	0.9029	0.0516	0.9054	0.8859	0.9305	0.9161	0.8916
Dual-SAM	2024,CVPR	0.0183	0.9249	0.9178	0.9638	0.9663	0.9242	0.0412	0.9196	0.9098	0.9373	0.9348	0.9045
CoralsCOP	2024,CVPR	0.0316	0.9307	–	0.9338	–	0.8884	–	–	–	–	–	–
VST++	2024,TPAMI	0.0467	0.8584	0.8388	0.9208	0.9047	0.8702	0.063	0.8846	0.8665	0.9179	0.8981	0.8699
SPDE	2025,TCSVT	0.0199	0.9273	–	0.9688	–	0.9233	–	–	–	–	–	–
iGAN	2025,TCSVT	0.0212	0.9249	0.9079	0.9682	0.9629	0.9195	0.0429	0.9151	0.9061	0.9383	0.9326	0.8989
SENET	2025,TIP	0.0199	0.9253	0.9092	0.9672	0.9619	0.9226	0.0414	0.9173	0.9118	0.9351	0.9309	0.9024
Ours	–	**0.0172**	**0.9325**	**0.9187**	**0.9712**	**0.9674**	**0.9321**	**0.0404**	**0.9213**	0.9127	**0.9408**	0.9359	**0.9058**

well. Notably, when compared with the recently introduced Dual-SAM framework (a dual SAM-based underwater object feature learning approach), our model achieves an overall improvement of 11.05% and 1.98% in the $\mathcal{M}$ metric across the two public datasets. Furthermore, relative to the TC-USOD model, which is specifically tailored for the USOD task, our model attains performance enhancements of 14.43% and 9.62% in the $\mathcal{M}$ metric on the respective datasets. Additionally, the superior F_φ performance of our model reflects its optimal balance between precision and recall, while its leading E_φ performance highlights its ability to robustly and accurately capture the shape and boundary information of targets. These results collectively demonstrate that our method effectively mitigates the adverse effects of the underwater environment. The superiority in performance is attributed to the joint perception optimization of the generated conditional features in the frequency domain and spatial domain by the DTEB module, as well as the excellent diverse generation capability of the diffusion model.

Qualitative Evalation. Figure 5 provides an intuitive comparison between our FSCDiff method and recent salient object detection (SOD) methods across various underwater scenarios. As illustrated in Figure 5, compared to existing SOD methods, the proposed FSCDiff method exhibits superior salient object perception and edge detection capabilities for underwater scenes with varying degrees of degradation. In the first row, the imperfect depth information introduced by the degraded depth map in the spatial domain prevents other models from correctly segmenting the tentacles near the octopus's head. In contrast, our FSCDiff method not only accurately segments these tentacles but also precisely delineates the overall shape of the octopus, demonstrating its exceptional ability to locate significant targets and define boundaries in complex underwater environments. In the second row, due to the low contrast of the RGB image and insufficient depth map information, other models struggle to fully segment the diver's two hands. However, our FSCDiff method not only achieves accurate segmentation but also produces clearer edge details, further highlighting its superior performance in handling complex underwater images. Overall, the FSCDiff method significantly enhances feature representation through dual-domain optimization perception, enabling more precise identification and segmentation of salient objects and their components. Additionally, the incorporation of the STMP mechanism improves the stability and quality of detection results, contributing to the exceptional performance of the FSCDiff method in challenging underwater conditions.

4.3 Ablation Study.

Effectiveness of each component. We present the quantitative results of the ablation experiments regarding the effectiveness of each component in FSCDiff, as depicted in Table 2. Specifically, (a) and (b) respectively denote the utilization of the FPE module and the SPT module within the DTEB module. The outcomes reveal that the performance of the frequency domain and the spatial

Table 2: Comparison of different structure settings on USOD10K and USOD datasets

NO.	Structure Setting					USOD10K (1026 images)						USOD (300 images)					
	FEC	STA	S-F	Spa-Cha	STMP	$\mathcal{M}\downarrow$	$F_\varphi^m\uparrow$	$F_\varphi^a\uparrow$	$E_\varphi^m\uparrow$	$E_\varphi^a\uparrow$	$S_m\uparrow$	$\mathcal{M}\downarrow$	$F_\varphi^m\uparrow$	$F_\varphi^a\uparrow$	$E_\varphi^m\uparrow$	$E_\varphi^a\uparrow$	$S_m\uparrow$
(a)	✓					0.0214	0.9251	0.9148	0.9646	0.9607	0.9238	0.0449	0.9149	0.9094	0.9301	0.9316	0.9002
(b)		✓				0.0232	0.9229	0.9133	0.9621	0.9577	0.9203	0.0457	0.9143	0.9092	0.9315	0.9323	0.8983
(c)	✓	✓				0.0193	0.9288	0.9164	0.9685	0.9646	0.9287	0.0427	0.9176	0.9103	0.9365	0.9341	0.9037
(d)	✓	✓	✓			0.0186	0.9293	0.9169	0.9692	0.9652	0.9293	0.0421	0.9188	0.9114	0.9378	0.9345	0.9043
(e)	✓	✓		✓		0.0183	0.9289	0.9171	0.9689	0.9554	0.9288	0.0416	0.9192	0.9112	0.9383	0.9342	0.9039
(f)	✓	✓	✓	✓		0.0177	0.9314	0.9182	0.9698	0.9663	0.9309	0.0409	0.9202	0.9119	0.9401	0.9352	0.9053
(g)	✓	✓	✓	✓	✓	0.0172	0.9325	0.9187	0.9712	0.9674	0.9321	0.0404	0.9213	0.9127	0.9408	0.9359	0.9058

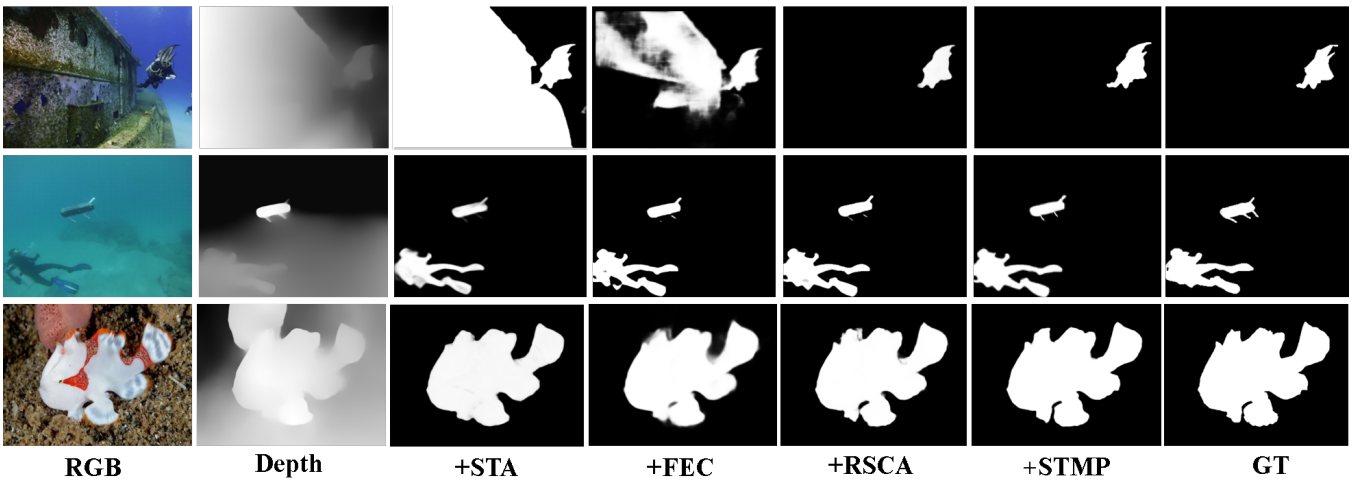

Figure 6: Visual results of the effectiveness of our modules.

domain alone exhibits certain disparities. (c) demonstrates the result of integrating the spatial and frequency features, and it can be observed that the performance of the predicted mapping has been significantly enhanced, affirming that the entangled learning of the two can indeed augment the model's reasoning ability for salient objects. (d) The results of using the elimination of redundant features are presented. It can be seen that the model, after reducing redundant features, can improve the upper limit. (e) indicates the operation of spatial channel attention. (f) presents the result of implementing the attention operation on the result after redundancy reduction. In contrast to (e), it can be discerned that after reducing redundant noise, the result after the attention operation is more precise. (g) shows the situation where STMP is added. It is evident that the model performance has improved, with predictions becoming more stable and refined. Furthermore, as depicted in Figure 6, through the gradual introduction of the proposed components (namely FPE, SPT, RSCA, and STMP. We have observed that the prediction results are gradually approaching the ground truth (GT). The spatial-frequency entanglement attention strengthens the model's representational capacity for salient objects, and STMP further optimizes the outcome. The aforementioned experimental results have fully verified the efficacy of the proposed modules in the underwater salient object detection task.

5 Conclusion

In this paper, we propose a novel underwater salient object detection method, named Frequency-Spatial Entangled Conditional Diffusion (FSCDiff). The key of FSCDiff lies in extracting more critical information from the spatial and frequency domains of RGB and depth images as conditional features to guide the diffusion model to accurately detect salient objects. To this end, we develop a Dual-domain Entanglement Enhancement Block, which acquires a more comprehensive representation ability through global perception in the Fourier domain and multi-scale correlation Spatial Perception. Furthermore, we have designed a prediction strategy named STMP that is consistent with the multi-step iterative mechanism of the diffusion model to optimize the prediction results. Through these operations, we effectively address two key challenges in underwater salient object detection (USOD): insufficient representation and boundary shift. We conduct extensive comparative experiments on two commonly used test datasets, and the experimental results show that the performance of the FSCDiff algorithm outperforms 21 state-of-the-art salient object detection (SOD) methods.

References

[1] Levi Cai, Nathan E McGuire, Roger Hanlon, T Aran Mooney, and Yogesh Girdhar. 2023. Semi-supervised visual tracking of marine animals using autonomous underwater vehicles. IJCV 131 (2023), 1406–1427.

[2] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In CVPR. 6007–6017.

[3] Håkon Måløy, Agnar Aamodt, and Ekrem Misimi. 2019. A spatio-temporal recurrent network for salmon feeding action recognition from underwater videos in aquaculture. Computers and Electronics in Agriculture 167 (2019), 105087.

[4] Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, and Sam Kwong. 2023. WaterMask: Instance Segmentation for Underwater Imagery. In ICCV. 1305–1315.

[5] Christian Sinnott, James Liu, Courtney Matera, Savannah Halow, Ann Jones, Matthew Moroz, Jeffrey Mulligan, Michael Crognale, Eelke Folmer, and Paul MacNeilage. 2019. Underwater virtual reality system for neutral buoyancy training: Development and evaluation. In ACM MM. 1–9.

[6] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. 2019. An underwater image enhancement benchmark dataset and beyond. IEEE TIP 29 (2019), 4376–4389.

[7] Zihao Zhang, Jie Wang, and Yahong Han. 2023. Saliency prototype for RGB-D and RGB-T salient object detection. In ACM MM. 3696–3705.

[8] Yue Zhan, Zhihong Zeng, Haijun Liu, Xiaoheng Tan, and Yinli Tian. 2025. MambaSOD: Dual Mamba-driven cross-modal fusion network for RGB-D Salient Object Detection. Neurocomputing 631 (2025), 129718.

[9] C Li, CL Guo, M Zhou, Z Liang, S Zhou, R Feng, and CC Loy. 2023. Embedding fourier for ultra-high-definition low-light image enhancement. ICLR (2023).

[10] Hemkant Nehete, Amit Monga, Partha Kaushik, and Brajesh Kumar Kaushik. 2024. Fourier Prior-Based Two-Stage Architecture for Image Restoration. In CVPR. 6014–6023.

[11] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. 2023. Efficient frequency domain-based transformers for high-quality image deblurring. In CVPR. 5886–5895.

[12] Daerji Suolang, Jiahao He, Wangchuk Tsering, Keren Fu, Xiaofeng Li, and Qijun Zhao. 2025. Lightweight Multi-Frequency Enhancement Network for RGB-D Video Salient Object Detection. In ICASSP. 1–5.

[13] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. 2024. Diffit: Diffusion vision transformers for image generation. In ECCV. 37–55.

[14] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. 2023. Shifted diffusion for text-to-image generation. In CVPR. 10157–10166.

[15] Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. 2023. Conditional text image generation with diffusion models. In CVPR. 14235–14245.

[16] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. IEEE TPAMI 45, 9 (2023), 10850–10869.

[17] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. 2023. Diffusion models in medical imaging: A comprehensive survey. Medical image analysis 88 (2023), 102846.

[18] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. 2024. A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering (2024).

[19] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. 2024. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In AAAI.

[20] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. 2023. Diffusiondet: Diffusion model for object detection. In ICCV. 19830–19843.

[21] Runmin Cong, Qinwei Lin, Chen Zhang, Chongyi Li, Xiaochun Cao, Qingming Huang, and Yao Zhao. 2022. CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. IEEE TIP 31 (2022), 6800–6815.

[22] Yu Qiu, Yun Liu, Le Zhang, Haotian Lu, and Jing Xu. 2023. Boosting salient object detection with transformer-based asymmetric bilateral U-Net. IEEE TCSVT 34, 4 (2023), 2332–2345.

[23] Zongwei Wu, Guillaume Allibert, Fabrice Meriaudeau, Chao Ma, and Cédric Demonceaux. 2023. Hidanet: Rgb-d salient object detection via hierarchical depth awareness. IEEE TIP 32 (2023), 2160–2173.

[24] Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. 2024. Vst++: Efficient and stronger visual saliency transformer. IEEE TPAMI 46, 11 (2024), 7300–7316.

[25] Kangfu Mei, Mauricio Delbracio, Hossein Talebi, Zhengzhong Tu, Vishal M Patel, and Peyman Milanfar. 2024. CoDi: conditional diffusion distillation for higher-fidelity and faster image generation. In CVPR. 9048–9058.

[26] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In ICCV. 3836–3847.

[27] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. 2024. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In Medical Imaging with Deep Learning. 1623–1639.

[28] Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. 2021. Conditional diffusion for interactive segmentation. In ICCV. 7345–7354.

[29] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. Dire for diffusion-generated image detection. In ICCV. 22445–22455.

[30] Zhongxi Chen, Ke Sun, and Xianming Lin. 2024. CamoDiffusion: Camouflaged object detection via conditional diffusion models. In AAAI, Vol. 38. 1272–1280.

[31] Axi Niu, Trung X Pham, Kang Zhang, Jinqiu Sun, Yu Zhu, Qingsen Yan, In So Kweon, and Yanning Zhang. 2024. ACDMSR: Accelerated conditional diffusion models for single image super-resolution. IEEE Transactions on Broadcasting (2024).

[32] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. 2023. Implicit diffusion models for continuous super-resolution. In CVPR. 10021–10030.

[33] Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. 2024. Diffusion models, image super-resolution, and everything: A survey. IEEE Transactions on Neural Networks and Learning Systems (2024), 1–21.

[34] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2022. Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media 8, 3 (2022), 415–424.

[35] Lin Hong, Xin Wang, Gan Zhang, and Ming Zhao. 2025. Usod10k: a new benchmark dataset for underwater salient object detection. IEEE TIP 34 (2025), 1602–1615.

[36] Ge Zhu, Jinbao Li, and Yahong Guo. 2024. Separate first, then segment: An integrity segmentation network for salient object detection. PR 150 (2024), 110328.

[37] Yuxin Mao, Jing Zhang, Zhexiong Wan, Xinyu Tian, Aixuan Li, Yunqiu Lv, and Yuchao Dai. 2025. Generative Transformer for Accurate and Reliable Salient Object Detection. IEEE TCSVT 35, 2 (2025), 1041–1054.

[38] Md Atiqur Rahman and Yang Wang. 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. In International symposium on visual computing. 234–244.

[39] Md Jahidul Islam, Ruobing Wang, and Junaed Sattar. 2022. SVAM: Saliency-guided visual attention modeling by autonomous underwater robots. (2022).

[40] Nian Liu, Ni Zhang, Ling Shao, and Junwei Han. 2021. Learning selective mutual attention and contrast for RGB-D saliency detection. IEEE TPAMI 44, 12 (2021), 9026–9042.

[41] Nian Liu, Ni Zhang, and Junwei Han. 2020. Learning selective self-mutual attention for RGB-D saliency detection. In CVPR. 13756–13765.

[42] Yingjie Zhai, Deng-Ping Fan, Jufeng Yang, Ali Borji, Ling Shao, Junwei Han, and Liang Wang. 2021. Bifurcated Backbone Strategy for RGB-D Salient Object Detection. IEEE TIP 30 (2021), 8727–8742.

[43] Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam Kwong. 2021. Cross-modality discrepant interaction network for RGB-D salient object detection. In ACM MM. 2094–2102.

[44] Runmin Cong, Hongyu Liu, Chen Zhang, Wei Zhang, Feng Zheng, Ran Song, and Sam Kwong. 2023. Point-aware interaction and cnn-induced refinement network for RGB-D salient object detection. In ACM MM. 406–416.

[45] Fuming Sun, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. 2023. CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection. IEEE TMM 26 (2023), 2249–2262.

[46] Gongyang Li, Zhen Bai, Zhi Liu, Xinpeng Zhang, and Haibin Ling. 2023. Salient object detection in optical remote sensing images driven by transformer. IEEE TIP 32 (2023), 5257–5269.

[47] Liuxin Bao, Xiaofei Zhou, Xiankai Lu, Yaoqi Sun, Haibing Yin, Zhenghui Hu, Jiyong Zhang, and Chenggang Yan. 2024. Quality-aware selective fusion network for VDT salient object detection. IEEE TIP 33 (2024), 3212–3226.

[48] Kunpeng Wang, Danying Lin, Chenglong Li, Zhengzheng Tu, and Bin Luo. 2024. Alignment-free rgbt salient object detection: Semantics-guided asymmetric correlation network and a unified benchmark. IEEE TMM 26 (2024), 10692–10707.

[49] Pingping Zhang, Tianyu Yan, Yang Liu, and Huchuan Lu. 2024. Fantastic animals and where to find them: Segment any marine animal with dual sam. In CVPR. 2578–2587.

[50] Ziqiang Zheng, Haixin Liang, Binh-Son Hua, Yue Him Wong, Put Ang, Apple Pui Yi Chui, and Sai-Kit Yeung. 2024. CoralSCOP: segment any coral image on this planet. In CVPR. 28170–28180.

[51] Jianhui Jin, Qiuping Jiang, Qingyuan Wu, Binwei Xu, and Runmin Cong. 2025. Underwater Salient Object Detection via Dual-stage Self-paced Learning and Depth Emphasis. IEEE TCSVT 35, 3 (2025), 2147–2160.

[52] Chao Hao, Zitong Yu, Xin Liu, Jun Xu, Huanjing Yue, and Jingyu Yang. 2025. A simple yet effective network based on vision transformer for camouflaged object and salient object detection. IEEE TIP 34 (2025), 608–622.