Pushing the Boundaries of Salient Object Detection: A Denoising-Driven Approach

Mengke Song[1*], Luming Li[1*], Xu Yu[1] and Chenglizhao Chen[1,2,3,†]
[1]China University of Petroleum (East China)
[2]Jiangsu Key Laboratory of Image and Video Understanding for Social Safety
[3]Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education

***Abstract*—Salient Object Detection (SOD) aims to identify the most attention-grabbing regions in an image and focuses on distinguishing salient objects from their backgrounds. Current SOD methods primarily use a discriminative approach, which works well for clear images but struggles in complex scenes with similar colors and textures between objects and backgrounds. To address these limitations, we introduce the diffusion-based salient object detection model (DiffSOD), which leverages a noise-to-image denoising process within a diffusion framework, enhancing saliency detection in both RGB and RGB-D images. Unlike conventional fusion-based SOD methods that directly merge RGB and depth information, we treat RGB and depth as distinct conditions, i.e., the appearance condition and the structure condition, respectively. These conditions serve as controls within the diffusion UNet architecture, guiding the denoising process. To facilitate this guidance, we employ two specialized control adapters: the appearance control adapter and the structure control adapter. Moreover, conventional denoising UNet models may struggle when handling low-quality depth maps, potentially introducing detrimental cues into the denoising process. To mitigate the impact of low-quality depth maps, we introduce a quality-aware filter. This filter selectively processes only high-quality depth data, ensuring that the denoising process is based on reliable information. Comparative evaluations on benchmark datasets have shown that DiffSOD substantially surpasses existing RGB and RGB-D saliency detection methods, improving average performance by 1.5% and 1.2% respectively, thus setting a new benchmark for diffusion-based dense prediction models in visual saliency detection.**

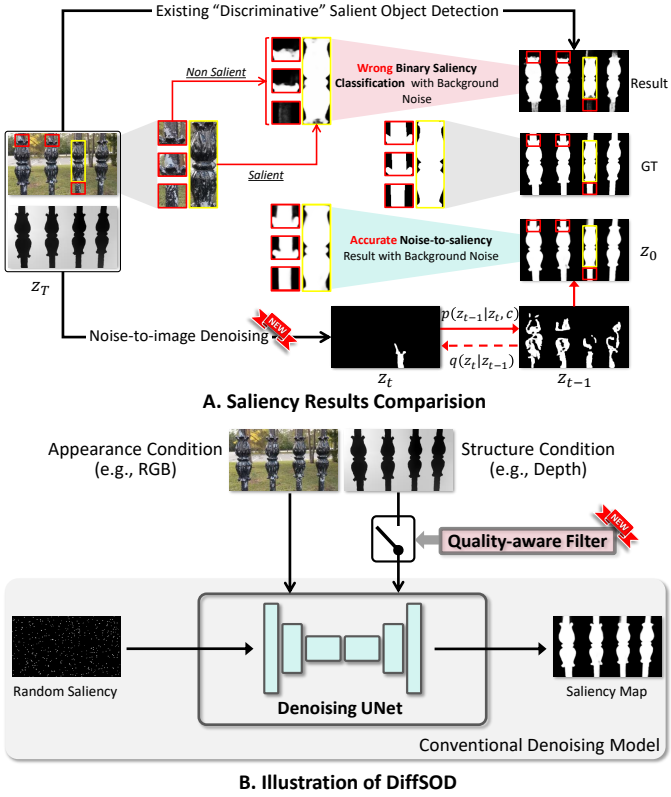


Fig. 1. **A.** Saliency results comparison between existing discriminative classification and our noise-to-image denoising method when facing scenes with background noise. **B.** Illustration of DiffSOD. The model learns an iterative denoising process to transform a randomly distributed saliency pattern into a saliency map with two types of conditions to guide the denoising process, and the quality-aware filter to filter out low-quality depth maps.

I. Introduction

Visual saliency detection[1] (VSD) is a computer vision task that aims to identify the most visually noticeable regions or objects in an image or video. Salient object detection (SOD), a specific task within VSD, focuses on distinguishing salient objects from their backgrounds. It replicates the selective attention mechanism of the human visual system, which instinctively focuses on the most relevant or salient parts of a scene. This task is crucial for downstream visual applications such as image retrieval [1], group activity recognition [2], and segmentation [3].

In recent years, the field of salient object detection (SOD) [4] has witnessed significant advancements with the introduction of deep learning techniques. State-of-the-art (SOTA) methods, including RGB and RGB-D saliency detection, have primarily adopted a "discriminate" perspective. These approaches focus on performing pixel-level "discriminative" binary saliency classification, aiming to distinguish "salient" regions from "non-salient" regions directly at the pixel level. However, it has been observed that these discriminate-based models face challenges when dealing with complex scenes and background noise. The impact of background noise on SOD methods primarily manifests in several aspects. Firstly, background noise may share similar colors, textures, or other visual features with salient objects, making it challenging for traditional feature-based methods to distinguish between salient objects and the background. Secondly, in some cases, noise elements in the background may mistakenly attract the model's attention, thereby reducing the accuracy of identifying actual salient objects. Thirdly, due to the diversity and complexity of background noise, SOD models relying solely on discriminative learning may struggle to maintain good generalization performance on unseen noisy

* The first two authors contribute equally to this work.
† Corresponding author: Chenglizhao Chen, cclz123@163.com.
[1]It is mainly referred to RGB and RGB-D salient object detection (SOD) in this paper.

backgrounds. As shown in Fig. 1-A (top of the figure), in a complex scene with background noise, these models may classify the trunk of a black cast copper column (yellow box) as salient, while considering the ends of the column (red box) as non-salient. These methods typically rely on learning discriminative features or employing complex neural network architectures to classify pixels as salient or non-salient. Although they have shown promising results in some cases, their effectiveness in handling noisy images and background distractions is limited.

In contrast to existing discriminate-based methods, this paper aims to enhance model resilience against complex background noise. Drawing inspiration from the recent successes of diffusion models [5], [6] in various generative tasks, we propose a novel "noise-to-image denoising" approach to optimize visual saliency detection. The denoising diffusion framework enhances visual saliency detection by progressively refining features through iterative denoising, making it robust to noise and complex backgrounds. Unlike single-step methods, this approach leverages contextual information to better separate salient objects from ambiguous regions, resulting in improved accuracy and adaptability in challenging scenes. As shown in Fig. 1-A (bottom of the figure), our method performs well even under the interference of complex background noise. To achieve this, we employ existing noise-to-image denoising models, which transforms random noise masks into saliency maps through an iterative denoising process. This allows the model to gradually refine saliency predictions at each iteration, effectively distinguishing salient objects from background noise. However, conventional noise-to-image denoising models that follow a random denoising process present challenges in achieving our goal. The random denoising process in these models lacks control, which can lead to the generation of saliency maps with inferior quality. Therefore, our insight is to devise a controlled diffusion model for saliency conditions, where our proposed conditions regulate latent saliency features in the noise-to-saliency process. This allows our models to transform the "discrimination-based" perspective into a "noise-to-image denoising[2]" perspective.

In our study, we introduce two main innovations that enhance the generation of saliency maps through a noise-to-image denoising process. We guide this process with two key conditions (Fig. 1-B): the appearance condition and the structure condition. These are managed by two specialized adapters. The appearance control adapter (Sec. III-D1) examines the visual aspects of an image to help distinguish salient objects from complex backgrounds effectively. The structure control adapter (Sec. III-D2), on the other hand, uses depth data to outline the salient objects' shapes more precisely. These adapters collectively ensure a more controlled denoising process, yielding accurate and coherent saliency maps. Additionally, conventional denoising UNet models may struggle when handling low-quality depth maps, to address this issue, we've introduced a quality-aware filter (Sec. III-C) approach. This selectively processes only the depth data that meets our quality standards, e.g., high-quality depth maps, maintaining the robustness and reliability of our saliency detection method. In summary, our contributions can be summarized as follows:

- We introduce an insightful perspective by advocating for a shift in visual saliency methods towards a "noise-to-image denoising" approach. This perspective challenges the conventional "discriminate" viewpoint and opens up new possibilities for visual saliency detection.
- To control the noise-to-saliency process, we propose to leverages two types of conditions — appearance and structure, which is achieved by two specialized adapters.
- We propose a quality-aware filter to selectively use the high-quality depth data, preventing low-quality input from weakening the detection process and ensuring reliable saliency outcomes.
- Experimental results suggest DiffSOD achieves state-of-the-art performance on both RGB and RGB-D SOD benchmark datasets, which demonstrates its effectiveness. Codes, datasets, and results are available at https://github.com/MengkeSong/DiffSOD.

II. Related Work

A. CNN-based Visual Saliency Detection Models

Visual Saliency Detection can be categorized into two types: RGB/RGB-D Salient Object Detection (SOD). In recent years, significant progress has been made in image saliency object detection using CNN-based approaches. These methods [7]–[10] leverage the powerful feature representation capabilities of CNNs to capture both low-level and high-level visual information. Various CNN architectures, such as VGGNet [11], ResNet [12], and DenseNet [13], have been employed to extract discriminative features from images. Additionally, some techniques [14] combine deep learning with traditional methods, further enhancing the performance of image saliency object detection. Certain works [15], [16] have introduced attention mechanisms to learn more discriminative features, including spatial and channel attention and pixel-wise contextual attention. Other approaches [17], [18] have explored the use of recurrent networks to refine the saliency map progressively. Furthermore, multi-task learning has been utilized to incorporate fixation prediction, image captioning, and edge detection, leading to improved SOD performance.

The integration of color (RGB) and depth data, known as RGB-D SOD, has gained considerable attention due to the availability of depth sensors and the additional geometric information they provide. CNN-based methods for RGB-D SOD [19]–[24] have proven to be effective in accurately capturing the interactions between color and geometry. Conventional methods [25], [26] often fuse RGB and depth features through middle fusion strategies. This enables them to model better the complex relationships between color contrast and

[2]Indeed, in images captured under natural lighting conditions, background information contains rich scene details and environmental cues, which can be considered as an important information dimension rather than just noise interference. However, from the perspective of visual saliency detection, the objective is to highlight salient objects and reduce attention to non-salient elements (which may include certain background information). Therefore, the term "denoising" may not be entirely appropriate in some contexts, but its usage is more of a simplification for the sake of addressing the impact of background complexity on saliency detection.

depth discontinuity, resulting in improved detection performance. Some approaches utilize depth cues to generate spatial or channel attention for enhancing RGB features. Dynamic convolution, graph neural networks, and knowledge distillation have also been adopted for multi-modal feature fusion. Moreover, the cross-attention mechanism has been utilized to facilitate long-range cross-modal interactions between RGB and depth cues.

However, many current saliency detection methods heavily rely on CNN architectures, which limits their ability to capture long-range dependencies. Certain techniques aim to integrate global and local information to achieve accurate salient region detection. For example, Zhang *et al.* [27] proposed a framework that considers the complementary nature of global positions and local details from two modalities, yielding favorable results. Nevertheless, these methods still face challenges in fully exploiting the advantageous relationships between features.

B. Transformer-based Visual Saliency Detection Models

As Vaswani *et al*. [27] first proposed a Transformer encoder-decoder architecture for machine translation, where multi-head self-attention and point-wise feed-forward layers are stacked multiple times to capture long-range global dependencies, more and more works have introduced the Transformer model to various computer vision tasks and achieved excellent results. For the visual saliency detection task, some recent works [28]–[35] also adopt the Transformer structure. Some first use CNNs to extract image features and then leverage the Transformer to incorporate long-range dependencies [29], [36]. Others combine CNNs and Transformers into hybrid architectures [33]. Also, some use pure Transformer-based models for feature representation learnings [28], [37].

These CNN/Transformer-based models are primarily focused on discrimination, learning to distinguish between salient and non-salient regions at a pixel level. While they can achieve impressive performance in relatively simple scenes, these discrimination-based models often struggle when confronted with complex scenes containing background noise and distractions. To address this limitation, our research proposes an alternative approach: a shift from discrimination-based classification to a noise-to-image denoising framework. This novel framework incorporates advanced techniques and saliency conditions to enhance the generation process of saliency maps. By adopting this perspective, we aim to overcome the challenges posed by complex scenes, providing a more effective solution.

C. Diffusion-based Models in Computer Vision

The diffusion model is a powerful generative model that uses a forward Gaussian diffusion process to sample a noisy image and then refines it using a backward generative process. Diffusion models have shown great potential in various fields such as image synthesis [38], image editing [39], and image super-resolution [40] tasks due to their ability to capture high-level semantic information. Several works have explored the application of the image diffusion model in different areas. MedSegDiff [6] proposes the first DPM-based medical segmentation model, and MedSegDiff-V2 [41] further improves the performance based on it using a Transformer. In diffCOD [42], the model learns to reverse the diffusion process that transforms ground-truth masks into random distributions. CamDiff [43] utilizes a latent diffusion model to synthesize salient objects within camouflaged scenes and DiffusionDepth [44] learns an iterative denoising process to refine depth maps.

However, there are no studies that demonstrate the effectiveness of diffusion models in the SOD task. In this paper, we propose to use the diffusion model for denoising the input RGB (or depth) as a conditioned saliency refinement process instead of adopting it as a typical generative head. To our knowledge, this is the first work introducing the diffusion model into the SOD task.

III. The Proposed Method

A. Method Overview

The key insight of our method is to enhance the generation of saliency maps through a noise-to-image denoising process. Fig. 2 illustrates the method pipeline of our denoising diffusion-based visual saliency detection model. Our approach comprises three main components: 1) denoising UNet (Sec. III-B), 2) quality-aware filter (Sec. III-C), and 3) control adapter (Sec. III-D).

Initiated with a random noise mask Z_T, the model employs a state-of-the-art denoising UNet (e.g., [45]), which iteratively refines the saliency map. This process is meticulously guided by two specialized control adapters — the appearance control adapter, informed by RGB image characteristics, and the structure control adapter, shaped by depth information, ensuring the saliency map accurately outline object contours. A quality-aware filter (QAF) is integral to the model, selectively incorporating only high-quality depth maps, thus enhancing the fidelity of saliency detection. Next, we will introduce each component in detail.

B. Preliminaries: Denoising UNet

In this work, we approach visual saliency detection from a noisy-to-image perspective and formulate it as a diffusion model. Specifically, we reformulate SOD, which includes RGB SOD and RGB-D SOD, as a diffusion model that utilizes the denoising UNet architecture, similarly to [5], consisting of an encoder and a decoder (depicted in Fig. 2-"E" and "D").

A diffusion model consists of two fundamental processes: the forward noise process and the backward diffusion process. During the training phase, the forward noise process is trained iteratively (for each step *t*) to serve as a prior for the backward diffusion process. In the testing phase, the backward diffusion process reverses the forward noise process and generates the desired image as the output. The forward noise process, denoted as $q(\cdot)$, introduces noise to the desired image distribution z_0 from a noise variance schedule β_s within the Gaussian space $\mathcal{N}(\cdot)$. This process generates a latent noisy sample z_t.

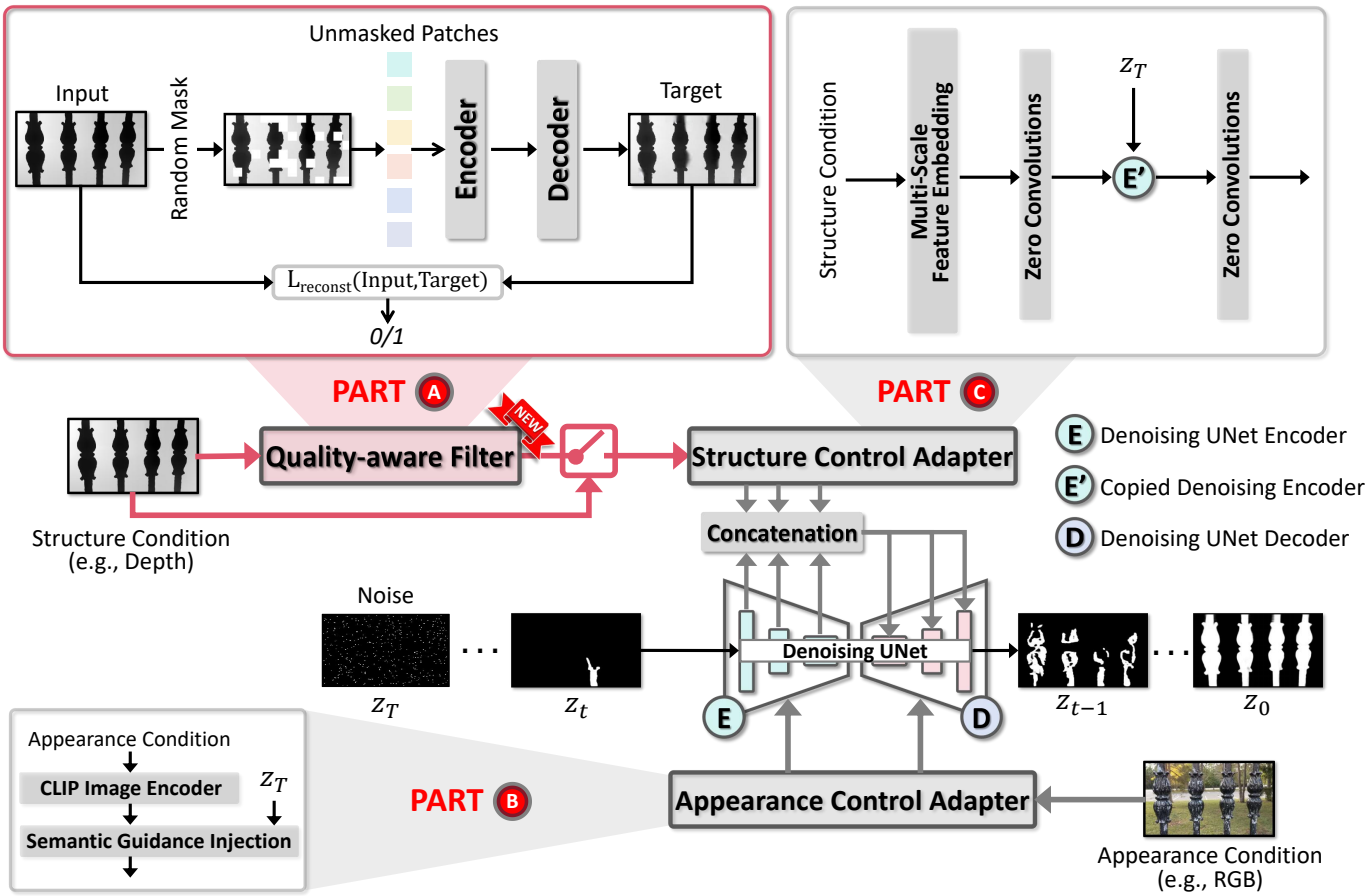


Fig. 2. Our proposed diffusion-based visual saliency detection model is adeptly structured to handle both RGB and RGB-D SOD tasks. The model initiates with a quality-aware filter (**PART A**, Sec. III-C), which rigorously processes only high-quality depth maps to ensure the robustness and reliability of the subsequent denoising process. Furthermore, to enhance the precision of the denoising phase, two specialized control adapters (**PART B** and **PART C**) are implemented. These adapters guide the denoising process by leveraging "appearance" and "structure" conditions, which are detailed in Sec. III-D1 and Sec. III-D2, respectively. Notably, the structure condition, involving both **PART A** and **PART C**, is specifically excluded in RGB SOD task to optimize task-specific processing efficiency. During testing, when using RGB-D data as input, if the depth map quality is low, the network will ignore the depth and rely solely on RGB for prediction.

For each iterative step $t \in 0, 1, ..., T$, the forward noise process can be expressed by:

$$q(z_t \mid z_0) = \mathcal{N}(z_t \mid \sqrt{\bar{\alpha}_t} \cdot z_0, (1 - \overbrace{\bar{\alpha}_t}^{\Uparrow} \cdot \mathrm{I}), \quad \overbrace{\textstyle\prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t}(1-\beta_s)} \tag{1}$$

where $q(\cdot)$ denotes the forward noise process, z_0 represents the desired image distribution and z_t represents the latent noisy images. The forward noise process introduces noise to the desired image distribution within the Gaussian space denoted by $\mathcal{N}(\cdot)$. I is the denoised image. α_s represents the signal retention ratio for a single step s, defined as $1 - \beta_s$, where β_s is the noise variance introduced at each step. $\bar{\alpha}_t$ is the cumulative product of α_s from step 0 to t, representing the total signal retention ratio up to step t. $\bar{\alpha}_t \cdot z_0$ scales the signal component, while $(1 - \bar{\alpha}_t) \cdot \mathrm{I}$ represents the variance of the noise component, showing how signal and noise evolve through the diffusion steps.

The backward diffusion process, based on the forward diffusion process, aims to reverse the effects of noise to iteratively recover the desired image distribution from a latent noisy sample, denoted as z_t. This iterative process involves sampling each iteration, denoted as $p_\theta(z_{t-1} \mid z_t)$, from a Gaussian distribution $\mathcal{N}(\mu_\theta(z_t, t), \sigma_\theta(z_t, t))$. A network predicts the mean and variance of the Gaussian distribution. During each iterative step $t \in 0, 1, ..., T$, the forward noise process can be described by:

$$p_\theta(z_{t-1} \mid z_t) = \mathcal{N}(\mu_\theta(z_t, t), \sigma_\theta(z_t, t)), \tag{2}$$

We reformulate the saliency detection as a saliency-condition-guided denoising process. Saliency-condition-guided denoising saliency detector inputs $\mathbf{c}$ as latent noise saliency distribution z_t in the conventional diffusion model, and outputs desired saliency map z_0, The noise-to-saliency process can be presented as:

$$p_\theta(z_{t-1} \mid z_t, \mathbf{c}) = \mathcal{N}(\mu\theta(z_t, t, \mathbf{c}), \textstyle\sum\theta(z_t, t, \mathbf{c})), \tag{3}$$

where model $\mu\theta(z_t, t, c)$ is trained to refine latent z_t to z_{t-1}. To accelerate the denoising process, we utilized the improved inference process from DDIM [46], where it set $\sum\theta(z_t, t, \mathbf{c})$ as 0 to make the prediction output deterministic.

C. Quality-aware Filter

In typical CNN/Transformer-based multi-modality visual saliency detection tasks [47], [48], such as RGB-D SOD,

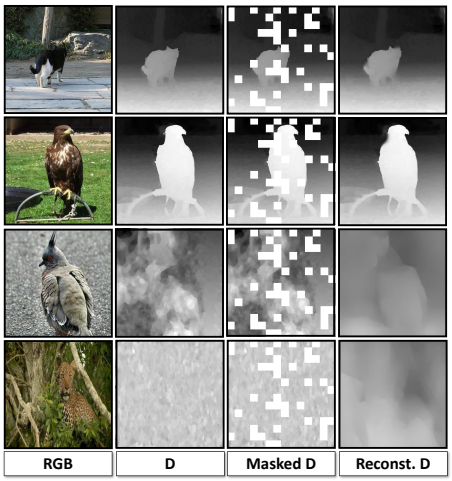


Fig. 3. Visualization of the reconstruction results (Sec. III-C). "D": depth maps, "Masked D": depth maps masked with certain masking ratio (here the masking ratio is 20%), "Reconst. D": reconstruted depth maps.

depth maps provide essential spatial details that augment RGB images. The quality of these depth maps is crucial for improving saliency prediction accuracy. Variations in depth sensor quality can lead to noisy or blurred depth maps, which impede the learning process in RGB-D models and degrade overall performance. Conventional denoising UNet models often struggle with this challenge, since they lack mechanisms to assess the accuracy and clarity of depth data, treating all inputs uniformly without distinguishing between high-quality and low-quality depth maps. Consequently, these models may inadvertently incorporate erroneous depth cues into the saliency detection process.

Building upon existing fusion-based RGB-D SOD methods that focus on evaluating and optimizing depth map quality [19], [20], we introduce a quality-aware filter (QAF) approach. This method dynamically selects high-quality depth maps, ensuring the structure condition (Sec. III-D2) incorporates only reliable depth data, thereby filtering out lower-quality inputs that could impair model performance. Drawing inspiration from masked auto-encoder architectures (e.g., [45]), as shown in Fig. 2-**PART A**, our QAF employs a self-supervised strategy to reconstruct the original depth map, assessing its quality based on the reconstruction error relative to the original depth.

The implementation involves segmenting the original depth map into small blocks. Random masks are applied to some blocks, followed by an encoder-decoder process that reconstructs the depth map. For more details on the methodology, refer to [45]. During the training phase, only high-quality depth maps are used to improve the model's reconstruction capabilities. The visualization of these results is shown in Fig. 3.

Our approach is predicated on the assumption that high-quality original depth maps will yield low reconstruction errors, whereas low-quality maps will result in high errors. The use of depth information in our DiffSOD model is determined by the magnitude of the reconstruction loss, formulated as follows:

$$\Delta = \begin{cases} 1, & \text{if } \mathrm{L}_{\mathrm{reconst}}(\mathrm{D}_{input}, \mathrm{D}_{target}) \leq \sigma, \\ 0, & \text{if } \mathrm{L}_{\mathrm{reconst}}(\mathrm{D}_{input}, \mathrm{D}_{target}) > \sigma, \end{cases} \tag{4}$$

where $\mathrm{L}_{\mathrm{reconst}}$ represents the reconstruction loss, with cosine similarity used as the metric. The variables D_{input} and D_{target} refer to the original and reconstructed depth maps, respectively. The hyper-parameter σ is set to define the acceptable quality threshold for depth maps, the effectiveness of which is discussed in the ablation analysis presented in Table IV. The binary indicator Δ determines the usage of depth information in our DiffSOD model: $\Delta = 1$ signifies inclusion of the depth map in the denoising process, while $\Delta = 0$ indicates exclusion.

In summary, by pre-training the model using only high-quality depth maps from existing datasets such as the KITTI [49] and the NYU Depth [50] datasets, the model effectively learns the structural information that depth maps should contain. During testing, regardless of the quality of the input depth map, the pre-trained model uses this structural knowledge to guide reconstruction. When the input quality is high, the reconstruction loss is small; when the input quality is low, the reconstruction loss is large. Through the variation in reconstruction loss, the model can indirectly assess the quality of the input depth map, while also demonstrating its adaptability in handling depth maps of varying quality.

Note that QAF is a preprocessing stage and is pre-trained before the denoising process begins. With QAF in place, if a depth map is of low quality, the structure condition is omitted, and only the appearance condition (i.e., RGB) is utilized in the denoising process.

D. Control Adapter

In deep learning, particularly within diffusion models utilized for image generation tasks, a "control adapter" is employed to introduce additional control information during the generation process. This additional information guides the model in generating images with specific styles or features. For instance, in conditional generation tasks, a control adapter aids the model in producing images that meet particular requirements based on provided conditional information, such as category labels or textual descriptions.

Building on this idea, we propose the concept of a "control adapter" for visual saliency detection tasks, consisting of two specialized types: the appearance control adapter, which responds to appearance conditions, i.e., RGB image characteristics, and the structure control adapter, which utilizes structure conditions, i.e., depth information. By integrating both adapters, our goal is to achieve precise control over the saliency generation process. Details of the appearance and structure control adapters are illustrated in Fig. 2-**PART B** and **PART C**. We will now describe these adapters in further detail.

1) Appearance Control Adapter: The inherent visual complexity of natural images requires subtlety interpretation to generate accurate saliency maps effectively. Therefore, as shown in Fig. 2-**PART B**, we leverage the CLIP image

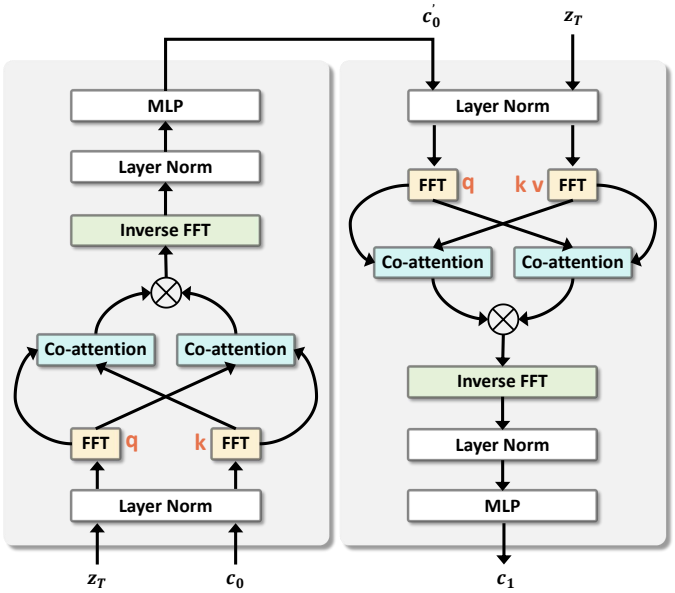


Fig. 4. Illustration of semantic guidance injection (Eq. 5 and Eq. 6). The semantic condition is integrated with the noise through a hierarchical FFT, co-attention mechanism, and inverse FFT.

encoder (e.g., CLIP ViT-B/32) to extract appearance condition embeddings. Since CLIP's training on a vast array of images paired with textual descriptions equips it with the capability to grasp and represent complex semantic concepts found across both visual and textual domains. This profound semantic understanding makes CLIP embeddings particularly beneficial for tasks requiring intricate interpretations of visual content.

As previously mentioned, our method aims to predict the desired saliency map from a noisy mask, posing a domain gap between its embedding and the initial appearance condition generated by CLIP image encoder. This high-level initial appearance condition features have different feature distributions compared to the saliency distributions during the noise-to-saliency process. This discrepancy can lead to confusion and incorrect representation of latent saliency features.

To bridge this gap, we propose an innovative solution called semantic guidance injection (SGI), which effectively combines initial appearance condition embedding and diffusion embedding. The SGI, illustrated in Fig. 4, enables our approach to learn the interaction between latent noise and initial appearance condition features, resulting in more robust representations of latent saliency features. The SGI consists of two blocks that share the same structure. Each block includes a discrete Fourier transform (FFT($\cdot$)), two co-attention modules, element-wise multiplication, and an inverse Fourier transform (IFFT[$\cdot$]). The main motivation for introducing the Fast Fourier Transform (FFT) into the appearance control adapter is to enhance the alignment and interaction between saliency features and noise features, thereby better guiding the model in the denoising and saliency prediction processes.

In the first block, given the noise embedding z_T and the initial appearance condition embedding c_0, we first perform FFT($\cdot$) along the spatial dimensions to convert z_T and c_0 to the frequency domain:

$$Z_T = \mathrm{FFT}(z_T), C_0 = \mathrm{FFT}(c_0), \tag{5}$$

where FFT($\cdot$) denotes the discrete Fourier transform, and Z_T and C_0 are the features in the frequency domain of z_T and c_0. After obtaining the frequency representations, we utilize two co-attention modules. Specifically, the embedding C_0 serves as the key (k) and the noise embedding Z_T serves as the query (q), allowing us to perform a cross-attention operation between the features from both embeddings, ensuring effective interaction between the initial appearance condition and the latent noise features. The features are then combined using element-wise multiplication to produce a modulated representation. Next, an inverse FFT (IFFT) is applied to convert the modulated frequency representation $f(Z_T, C_0)$ back to the spatial domain. After applying Layer Normalization, the result is further refined using a Multi-Layer Perceptron (MLP), resulting in the updated appearance condition embedding $c_0^{'}$:

$$c_0^{'} = \mathrm{MLP}(\mathrm{IFFT}[f(Z_T, C_0)]), \tag{6}$$

where IFFT[$\cdot$] represents the inverse Fourier transform.

In the second block, $c_0^{'}$ is used as the query (q) and the noise embedding z_T serves as both the key (k) and value (v). After applying Layer Normalization and FFT, the embeddings are processed through two co-attention modules, similar to the first block, ensuring alignment between noise and the updated condition embedding. The output features are then modulated through element-wise multiplication, converted back to the spatial domain via inverse FFT, and normalized again. Finally, another MLP is applied to produce the final transformed embedding c_1 which serves as the refined appearance condition and is injected into all encoder-decoder blocks in the denoising UNet model via cross-attention, effectively guiding the denoising process.

2) Structure Control Adapter: Unlike conventional cross-modality fusion-based RGB-D SOD methods [16], [51], our noise-to-image denoising technique introduces a novel structure control adapter, drawing inspiration from ControlNet [52], as shown in Fig. 2-**PART C**. This adapter capitalizes on the integration of structural and depth information to precisely define the contours of salient objects during the denoising process.

Instead of adding conditions directly to the input noise as in ControlNet, our approach employs a Feature Pyramid Network (FPN) as the backbone for feature extraction. This backbone processes the input structure condition — specifically, the depth map — to extract multi-scale feature embeddings. These feature embeddings capture both coarse and fine details of the scene, crucial for effective denoising.

The multi-scale feature embedding process begins by upsampling each feature level (MF_i) to a uniform scale. Convolution operations are then applied to each upsampled feature. We utilize a straightforward feature-level concatenation ($\mathcal{C}$), followed by multiple convolution layers to produce the structured condition $\mathrm{Con_s}$:

$$\mathrm{Con_s} = \mathrm{Conv}(\mathcal{C}(\sum_{i=1}^{4} \mathrm{Conv}(\mathrm{UP}(\overbrace{\mathrm{MF}_i}^{\Uparrow \mathrm{FPN(D)}})))), \tag{7}$$

where Conv denotes the 3×3 convolution operation, and UP is the upsample operation. D is the input structure condition, i.e., depth.

After establishing the backbone and feature embedding, we fix the original weights of the denoising UNet (Sec. III-B) and replicate the encoder (E) structure and weights, creating E$'$. This copied encoder integrates the structural information during the decoding phase to enhance the detail and accuracy of the denoising results. To achieve it, we ensure that all other elements in the denoising UNet remain unchanged while modifying the input of the i-th block of the decoder as:

$$\begin{cases} \mathcal{C}(m + \mathrm{Con_s}, e_j + zero(e_j^{'})), & \text{if } \ i = 1, i + j = 13, \\ \mathcal{C}(d_{i-1}, e_j + zero(e_j^{'})), & \text{if } \ 2 \leq i \leq 12, i + j = 13. \end{cases} \tag{8}$$

Here, m denotes the output from the final block of the encoder. This output is combined with the output of the structure control adapter, $\mathrm{Con_s}$, and then supplied to the first block of the decoder. $\mathrm{Con_s}$ is crucial as it adjusts the decoder's ability to accurately render the contours of salient objects. The variable d_{i-1} refers to the output from the $(i-1)$-th block of the decoder and serves as the input for the succeeding block. The terms e_j and $e_j^{'}$ represent the outputs from the j-th block of the original encoder E and the copied encoder E$'$, respectively. The function $zero(\cdot)$ denotes a zero-weight convolutional layer, which is utilized to incrementally incorporate structural control information into the main denoising UNet framework.

E. Loss Functions

The diffusion process $q\left(z_t \mid z_0\right)$ and denoising process $p_\theta\left(z_{t-1} \mid z_t\right)$ are respectively defined in Eq. 1 and Eq. 2. Trainable parameters are mainly the conditioned denoising model $\mu_\theta(z_t, t, c)$ and saliency feature defined above. The objective of the denoising UNet is defined as:

$$\mathrm{L}_{Diffusion} = \parallel z_{t-1} - \mu_\theta(z_t, t, c) \parallel^2 . \tag{9}$$

The DiffSOD is trained by combining losses through a weighted sum and the total loss L_{total} can be defined by:

$$\mathrm{L}_{total} = \alpha_1 \cdot \mathrm{L}_{Diffusion} + \alpha_2 \cdot \mathrm{L}_{BCE}, \tag{10}$$

where α_1 and α_2 are hyper-parameters. L_{BCE} is the binary cross-entropy loss.

IV. Experiments

A. Implementation Details

We implemented our proposed method using PyTorch with an NVIDIA GeForce RTX 3090 GPU. The model training is divided into two parts: QAF (Quality-Aware Filter) and DiffSOD (Diffusion-based Visual Saliency Detection). QAF and DiffSOD were trained separately, with QAF's output guiding the structure control used by DiffSOD during the denoising process.

For QAF training, we utilized the MultiMAE framework for self-supervised learning, reconstructing depth maps from masked inputs. Random masking is applied to depth maps, and the network reconstructs the original ones. The architecture has an encoder and a decoder, minimizing reconstruction error measured by cosine similarity. High-quality depth maps from KITTI and NYU Depth datasets were used, with Adam optimizer and an initial learning rate of 1e-4, decayed by a factor of 10 at the 60th epoch, and a batch size of 16. For DiffSOD, it's a diffusion-based visual saliency detection model. Following [16], [53], it used widely accepted RGB and RGB-D saliency detection benchmarks, resizing all inputs to 352×352. Trained with Adam optimizer starting at 1e-4, decayed at the 60th epoch. The training involved forward and backward diffusion processes with $T = 1000$ and a linear noise schedule, batch size of 2. The loss function combines diffusion loss and binary cross-entropy loss. It had a two-stage training: pre-training with high-quality depth maps and then using QAF to select reliable depth data. Each epoch took about 5 hours on the GPU.

To comprehensively demonstrate the effectiveness of the model, we also conducted component evaluations and ablation studies on RGB SOD datasets, using depth maps generated by existing depth estimation methods (e.g., DepthFormer [56]) for comparison.

B. Datasets and Evaluation Metrics

In our experiments, we follow the prevalent settings of different SOD tasks. Specifically, for RGB SOD, we use the training subsets of DUTS [57] to train our method, and evaluate the effectiveness of our method on five widely used public benchmark datasets, *i.e.*, DUT-OMRON [58] with 5,168 images, ECSSD [59] with 1,000 images, HKU-IS [60] with 4,447 images, PASCAL-S [61] with 850 images, DUTS-TE [57] with 5,019 images. For RGB-D SOD, we choose 2185 samples from NLPR [55] and NJUD [54] as the training set. The testing sets are seven widely used benchmark datasets: STEREO [62] (797 image pairs), LFSD (100 image pairs), SSD (80 image pairs), NJUD [54] (1,985 image pairs), NLPR [55] (1,000 image pairs), SIP [25] (929 image pairs), and ReDWeb-S [63] (3,179 image pairs).

Four metrics are adopted for quantitative evaluation, including S-measure (Sm) [64], F-measure (Fm) [65], E-measure (Em) [66] and mean absolute error (MAE).

C. Comparison with the SOTA Models

To demonstrate the effectiveness of the proposed method, we compare it with the SOD models, which are widely used and well recognized in the community. For RGB-D SOD, we compare our DiffSOD with 12 state-of-the-art (SOTA) CNN-based RGB-D SOD methods, *i.e.*, DMRA [67], CPFP [68], S2MA [69], A2dele [70], D3Net [25], UCNet [26], BBS [15], CIRNet [21], SPNet [22], SSL [71], MIRV [72], DIM [73], and 6 Transformer-based RGB-D SOD methods, HFMD [74], *i.e.*, CPNet [75], DCM [76], HTrans [32], CAT [33], and VST [28] (a Transformer-based unified model for both RGB/RGB-D SOD). For RGB SOD, we compare our DiffSOD with 19 SOTA RGB SOD models, including 7 CNN-based RGB SOD methods, *i.e.*, SGL [77], PA-KRN [77], DAD [78], EDN [79], MENet [80], TSD [81], LARNet [82] and 9 Transformer-based RGB SOD methods, *i.e.*, VST [28], SDETR [83],

TABLE I
QUANTITATIVE COMPARISON OF OUR PROPOSED DIFFSOD WITH OTHER 18 SOTA CNN-BASED AND TRANSFORMER-BASED RGB-D SOD METHODS ON 7 BENCHMARK DATASETS. THE TOP-2 RESULTS ARE MARKED IN RED AND GREEN.

		CNN-based Models												Transformer-based Models						
Model		DMRA	CPFP	S2MA	A2dele	D3Net	UCNet	BBS	CIRNet	SPNet	SSL	MIRV	DIM	VST	HTrans	CAT	CPNet	DCM	HFMD	Ours
Year		2019	2020	2020	2020	2021	2021	2021	2022	2022	2022	2023	2023	2021	2023	2023	2024	2024	2024	
FLOPs (G)		-	-	70.5	21	198.5	8.8	31.32	145.1	135.8	-	30.8	-	31	17.12	341.8	186.7	-	-	352.6
Params (M)		59.7	69.5	86.7	30.3	43.2	33.3	49.8	60.2	175.3	74.2	186.3	31.6	83.8	58.9	262.6	216.5	80.3	431.6	279.1
Speed (FPS)		16.2	6.5	9.3	120	65.2	17.1	26.1	14	12.4	52.4	18.5	90.0	67.8	77.3	22.8	65.5	84.2	9.9	22.9
NJUD	Sm↑	.886	.878	.894	.871	.892	.897	.919	.915	.914	.909	.890	.902	.922	.930	.928	.935	.932	.934	.938
	Fm↑	.872	.877	.889	.874	.863	.886	.899	.897	.890	.884	.880	.918	.914	.927	.925	.933	.928	.935	.937
	Em↑	.908	.906	.930	.897	.913	.915	.919	.922	.920	.928	.929	.921	.899	.931	.933	.935	.937	.935	.938
	M↓	.051	.053	.053	.051	.047	.043	.037	.035	.036	.038	.046	.036	.034	.028	.027	.025	.031	.024	.022
NLPR	Sm↑	.899	.888	.915	.898	.902	.920	.826	.923	.926	.909	.914	.896	.931	.938	.934	.940	.934	.938	.942
	Fm↑	.855	.822	.902	.878	.857	.890	.878	.914	.901	.884	.895	.899	.886	.919	.916	.924	.923	.925	.927
	Em↑	.942	.924	.953	.945	.943	.953	.949	.952	.954	.939	.953	.957	.954	.962	.961	.965	.961	.961	.967
	M↓	.031	.036	.030	.028	.030	.025	.028	.023	.024	.038	.025	.023	.023	.020	.021	.016	.023	.017	.015
SIP	Sm↑	.806	.850	.872	.829	.860	.873	.876	.888	.869	.871	.876	.866	.904	.909	.908	.907	.911	.886	.913
	Fm↑	.819	.818	.849	.825	.835	.868	.874	.885	.872	.875	.863	.887	.895	.910	.905	.917	.923	.896	.919
	Em↑	.863	.899	.911	.892	.902	.913	.915	.923	.908	.921	.924	.910	.937	.940	.934	.941	.937	.925	.941
	M↓	.085	.064	.058	.070	.063	.051	.056	.052	.055	.046	.049	.052	.040	.037	.038	.035	.033	.044	.032
STEREO	Sm↑	.886	.871	.890	.879	.885	.903	.909	.913	.899	.886	.890	.888	.909	.918	.917	-	-	-	.922
	Fm↑	.868	.827	.882	.874	.855	.885	.886	.896	.883	.875	.892	.894	.905	.905	.908	-	-	-	.914
	Em↑	.920	.897	.932	.915	.920	.922	.927	.930	.924	.919	.917	.927	.929	.932	.935	-	-	-	.935
	M↓	.047	.054	.051	.044	.046	.039	.041	.038	.043	.045	.045	.038	.039	.035	.035	-	-	-	.032
LFSD	Sm↑	.847	.828	.837	.834	.825	.854	.856	.865	.847	.834	.849	.873	.884	.887	.892	.892	-	.880	.898
	Fm↑	.849	.813	.835	.832	.810	.845	.850	.852	.843	.819	.844	.877	.871	.879	.881	.890	-	.871	.879
	Em↑	.899	.867	.873	.871	.862	.891	.889	.901	.887	.888	.889	.904	.903	.905	.908	.919	-	.910	.921
	M↓	.075	.088	.094	.077	.095	.076	.074	.068	.078	.080	.072	.060	.061	.064	.052	.049	-	.059	.048
SSD	Sm↑	.857	.807	.868	.802	.847	.865	.870	.868	.865	.855	.871	.881	.889	.883	.879	-	-	.885	.891
	Fm↑	.821	.725	.848	.776	.815	.854	.832	.829	.830	.833	.828	.862	.871	.874	.872	-	-	.871	.879
	Em↑	.892	.832	.909	.861	.888	.907	.904	.895	.899	.896	.891	.907	.913	.922	.931	-	-	.915	.934
	M↓	.058	.082	.052	.070	.058	.049	.049	.048	.047	.050	.047	.049	.045	.045	.046	-	-	.040	.043
ReDweb-S	Sm↑	.592	.685	.711	.641	.689	.713	.692	.703	.709	.710	.699	.725	.759	.763	.764	-	.765	.747	.769
	Fm↑	.579	.645	.675	.603	.673	.710	.647	.708	.712	.706	.684	.716	.755	.762	.761	-	.761	.744	.767
	Em↑	.721	.744	.750	.678	.768	.794	.709	.748	.759	.754	.746	.760	.813	.815	.818	-	.821	.798	.825
	M↓	.188	.142	.140	.160	.149	.130	.150	.132	.129	.136	.143	.122	.113	.115	.116	-	.115	.119	.111

SDG [84], SelfRe [85], BBRF [31], TIGAN [86], UGLR [87], ELSA [88], Prior [89] and 3 diffusion-based methods MDiff [41], dCOD [42], and Camo [90]. Note that the MDiff and dCOD are proposed for medical image segmentation and camouflaged object detection, respectively. We retrain them using RGB SOD training sets. The compared results are from the codes or saliency maps provided by the authors.

1) Quantitative Evaluation: The quantitative comparison results for RGB-D and RGB SOD are shown in Table I and Table II. Our method outperforms previous CNN, Transformer, and diffusion-based SOD models on both benchmarks, achieving top performance on datasets such as NJUD, NLPR, DUTS, and ECSSD, while maintaining competitive results on others. For RGB SOD, we provide two comparisons: (1) using RGB only ("Ours") and (2) using RGB with estimated depth ("Ours+"). In the "ours" method, we disable the depth branch by dynamically excluding low-quality depth information using the Quality-Aware Filter (QAF), allowing the model to rely solely on RGB input. In contrast, the "Ours+" method incorporates both RGB and depth information. Instead of removing the depth branch entirely, the model selectively ignores depth information when deemed unreliable, preventing network mismatch and ensuring robust saliency detection. This selective use of depth is managed through the Structure Control Adapter (SCA).

As shown in Table II, our method demonstrates strong performance with RGB data alone, highlighting its robustness in complex scenes. When incorporating estimated depth information, the performance further improves, emphasizing the adaptability of our approach to additional depth cues. This improvement is driven by the iterative denoising process in our diffusion framework, which effectively suppresses background noise, enhances salient regions, and integrates complementary depth information. Additionally, our model performs strongly on challenging datasets like ReDWeb-S and DUT-OMRON, further validating the effectiveness of DiffSOD.

2) Qualitative Evaluation: Fig. 5 showcases the visual comparison results of our proposed DiffSOD model against state-of-the-art representative models. The first row highlights the outstanding performance of our model in detecting multiple objects. In the second, third, and fourth rows, we observe that our model surpasses others in capturing salient regions with more complex objects, resulting in clear boundaries. These visual results demonstrate the effectiveness of DiffSOD in saliency detection, particularly in scenarios involving complex backgrounds and objects of various shapes.

TABLE II
QUANTITATIVE COMPARISON OF OUR PROPOSED DIFFSOD WITH OTHER 19 SOTA CNN-BASED, TRANSFORMER-BASED, AND DIFFUSION-BASED RGB SOD METHODS ON 5 BENCHMARK DATASETS. THE TOP-2 RESULTS ARE MARKED IN RED AND GREEN. "OURS": USING RGB ONLY; "OURS+": USING RGB WITH ESTIMATED DEPTH.

		CNN-based Models							Transformer-based Models									Diffusion-based				
Model		SGL	PA-KRN	DAD	EDN	MENet	TSD	LARNet	VST	SDETR	SDG	SelfRe	BBRF	TIGAN	UGLR	ELSA	Prior	MDiff	dCOD	Camo	Ours	Ours+
Year		2021	2021	2022	2022	2023	2023	2024	2021	2022	2022	2022	2023	2024	2024	2024	2024	2023	2023	2024		
FLOPs (G)		-	197.4	-	74.6	85.3	-	-	50	-	-	-	92	-	-	-	152.9	983	375.3	389.4	279.3	352.6
Params (M)		-	102.2	-	83.6	51.5	-	66	44.5	56.7	-	-	74.4	-	-	31.9	104.6	379.1	253.7	247.2	178.5	279.1
Speed (FPS)		-	-	-	6.7	45.0	-	98.1	86.2	26.9	-	-	21.2	-	15	52	36.5	7.4	14.1	25.6	28.7	22.9
ECSSD	Sm↑	.923	.928	-	.927	.928	.909	-	.932	.937	.935	.933	.939	.941	.940	-	.931	.930	.935	.931	.941	**.945**
	Fm↑	.924	.930	-	.930	.940	.916	.907	.920	.939	.940	.953	.942	.936	.953	.941	.952	.936	.940	.941	.952	**.955**
	Em↑	.946	.950	.953	.951	.954	.942	-	.918	.956	.959	.926	.950	.960	-	.958	-	.950	.953	.955	.958	**.961**
	M↓	.036	.032	.032	.032	.030	.044	.041	.033	.025	.025	.029	.024	.025	.028	.030	.031	.027	.028	.026	.024	**.023**
DUTS	Sm↑	.893	.901	-	.892	.903	.894	-	.896	.903	.905	.904	.908	.912	.912	-	.897	.906	.909	.911	.913	**.917**
	Fm↑	.865	.876	-	.863	.892	.810	.793	.818	.873	.878	.905	.893	.873	.901	.882	.899	.904	.905	.904	.908	**.910**
	Em↑	.928	.927	.925	.925	.937	.901	-	.892	.937	.938	.919	.927	.937	-	.934	-	.918	.924	.917	.938	**.939**
	M↓	.034	.031	.035	.035	.028	.047	.052	.037	.028	.027	.029	.025	.026	.029	.034	.033	.030	.028	.027	.026	**.024**
HKU-IS	Sm↑	.921	.924	-	.924	.927	.923	-	.920	.922	.927	.928	.931	.929	.932	-	.921	.929	.934	.931	.934	**.936**
	Fm↑	.915	.920	-	.920	.922	.902	.895	.900	.918	.924	.933	.923	.922	**.941**	.928	.928	.907	.921	.911	.928	.934
	Em↑	.954	.956	.953	.955	.955	.947	-	.953	.956	.962	.956	.958	**.964**	-	.956	-	.938	.951	.947	.951	.959
	M↓	.028	.027	.028	.026	.025	.037	.036	.029	.026	.024	.025	.024	**.023**	.026	.025	.027	.031	.028	.033	.025	**.023**
PASCAL-S	Sm↑	.857	.858	-	.865	.872	.870	-	.865	.869	.870	.873	.871	.879	.881	-	.865	.865	.874	.872	.878	**.884**
	Fm↑	.837	.839	-	.849	.860	.830	.801	.829	.855	.858	.878	.862	.869	**.882**	.862	.884	.855	.858	.854	.877	.879
	Em↑	.894	.896	.901	.916	.913	.882	-	.837	.911	.913	.870	.867	**.919**	-	.912	-	.887	.902	.893	.912	.914
	M↓	.068	.067	.060	.062	.053	.074	.082	.061	.055	.054	.052	.052	.053	.054	.059	.058	.058	.053	.059	.051	**.050**
DUT-OMRON	Sm↑	.846	.853	-	.849	.850	.858	-	.850	.865	.865	.848	.855	.861	.865	-	.848	.860	.866	.864	.865	**.868**
	Fm↑	.783	.796	-	.788	.813	.745	.745	.756	.811	.806	**.829**	.814	.796	.819	.794	.821	.803	.811	.807	.816	.822
	Em↑	.878	.888	.867	.877	.891	.863	-	.861	**.902**	.898	.877	.887	.890	-	.891	-	.888	.883	.888	.891	.897
	M↓	.049	.050	.052	.049	.045	.061	.065	.058	.044	.043	.043	.042	.047	.045	.050	.051	.042	.045	.045	.043	**.040**

To provide an intuitive understanding of how the denoising process refines the saliency prediction step by step, we present a visualization in Fig. 6. This figure illustrates how our proposed approach learns the location information, shapes, and edges of salient objects in the first stage. Subsequently, it utilizes the guided denoising model to refine the saliency masks further, resulting in a more accurate segmentation of the salient objects. This approach can be likened to first locating the desired scenery and then segmenting its salient objects. We believe that this step-by-step process closely aligns with the functioning of the human visual system. Moreover, integrating guided clues from the two control conditions is seamlessly incorporated into the diffusion process using two control adapters, further enhancing the effectiveness of our approach.

To further demonstrate that the progressive denoising mechanism can enhance salient features, we provide visualizations. As shown in in Fig. 7, for the green-object and horse examples, as the time parameter T increases from 200 to 1000, the boundaries of salient regions gradually clarify. Initially, noise causes misclassifications (red-circled areas for the green object and red-boxed area for the horse). But with iterative denoising, the model filters out noise, making the salient objects' contours approach the ground - truth, thus validating the mechanism's effectiveness.

D. Component Evaluation

To validate the efficacy of our proposed components within the RGB-D DiffSOD architecture, we conducted a comprehensive ablation study, the results of which are summarized in Table III. Depth maps of ECSSD and PASCAL-S datasets are generated by existing depth estimation methods (e.g., DepthFormer [56]). The baseline model, outlined in the first row, utilizes a basic diffusion model that combines noise with RGB-D data to generate a saliency map.

The effectiveness of the Structure Control Adapter (SCA, Sec. III-D2) is demonstrated in lines 2-5. When compared to the baseline, incorporating the structure condition (SC, line 2) consistently enhanced all evaluation metrics. Further improvements were achieved by integrating Multi-Scale Feature Embedding (MFE, line 3) into the appearance condition generation. A significant contribution was also made by the Copied Denoising Encoder (E', line 5), which was instrumental in refining the diffusion and denoising stages, as evidenced by the Fm metric on the NJUD test set improving from 0.898 to 0.918 (line 1 vs. line 5). The complete implementation of SCA (line 4) yielded the best results.

The Appearance Control Adapter (ACA, Sec. III-D1) is evaluated in lines 6-9. Incorporating the appearance condition (AC, line 6) consistently improved all evaluation metrics relative to the baseline. The inclusion of a CLIP image encoder (CLIP, line 9) further enhanced the performance. Notably, the Semantic Guidance Injection (SGI, line 8) proved vital in steering the diffusion and denoising processes, with the Fm metric on the NJUD test set increasing from 0.898 to 0.937 (row 1 vs. row 8). The full deployment of ACA (line 7) achieved the highest performance metrics.

As shown in lines 10-11, our Quality-Aware Filter (QAF, Sec. III-C, line 11) outperformed the approach of directly utilizing all depth maps (line 10). This was particularly evident in metrics such as Sm on the PASCAL-S dataset, where performance increased from 0.875 (line 10) to 0.884 (line 11).

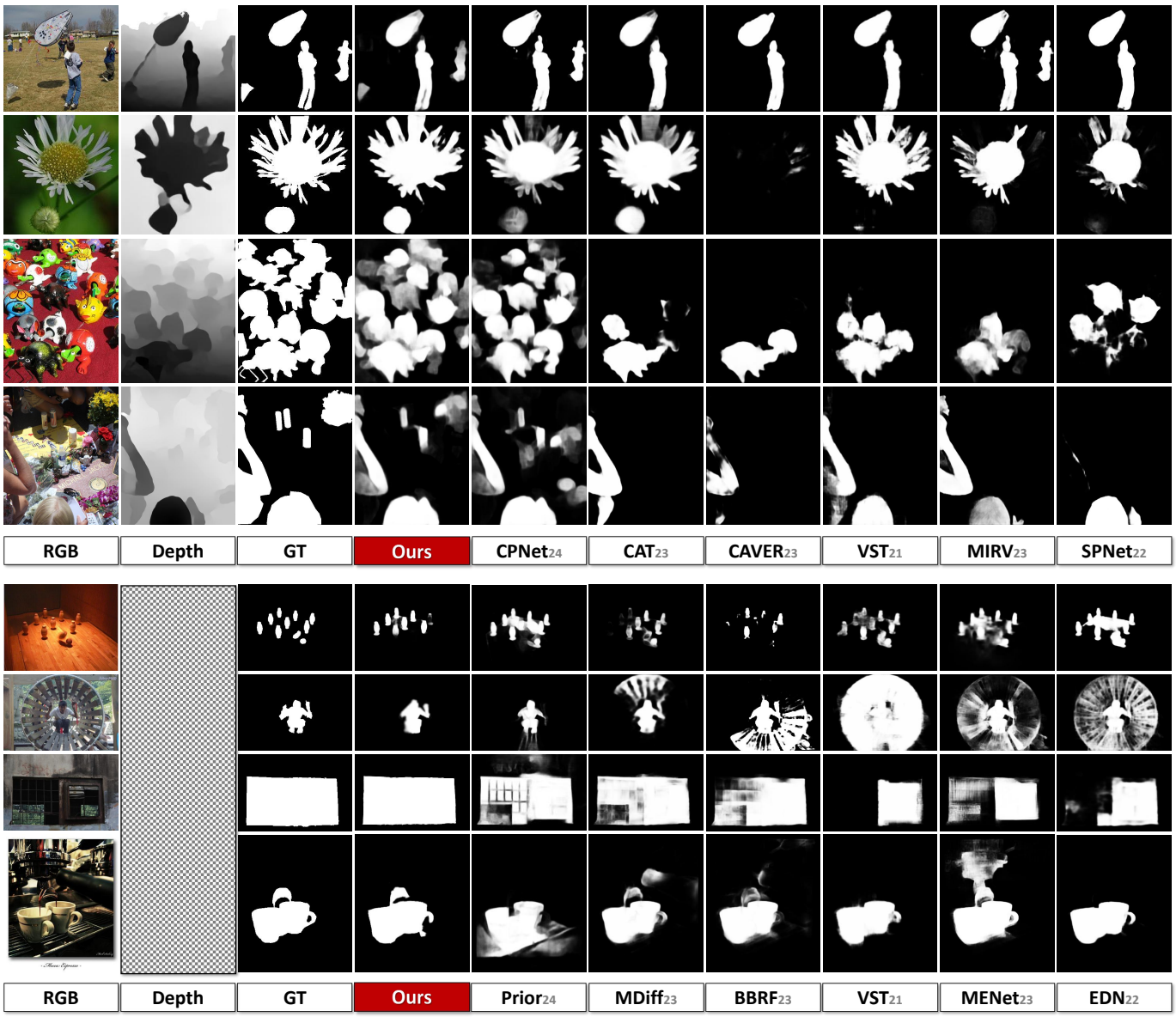


Fig. 5. Visual comparison between our method and several most representative SOTA RGB-D SOD models (top) and RGB SOD models (bottom).

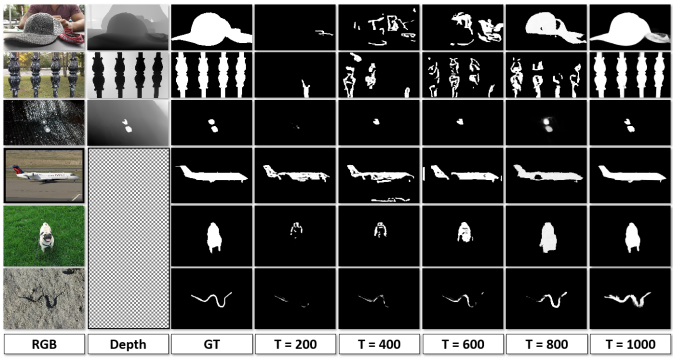


Fig. 6. Visualization of the denoising process with 1000 inference steps, where T denotes the current step.

The advantage of the QAF lies in its selective use of high-quality depth data, which prevents low-quality inputs from compromising the detection process and ensures more reliable saliency detection outcomes.

E. Ablation Study

1) Effectiveness of Threshhold σ in Eq. 4: We conducted an ablation study on the threshold σ, a hyper-parameter that determines the quality level of depth maps as described in Eq. 4. We tested σ values of 0.1, 0.2, 0.3, and 0.4, where the reconstruction error is normalized between zero and one. Depth maps of ECSSD dataset are generated by existing depth estimation methods (e.g., DepthFormer [56]). Results presented in Table IV indicate that the model's performance is moderately sensitive to changes in σ. The optimal performance was observed at $\sigma = 0.2$, while setting σ = 0.1 resulted in noticeable performance degradation. Conversely, a σ of 0.4 led to a decrease in performance (e.g., .769 vs. .757 in the Sm metric on the ReDweb-S set). This sensitivity can be attributed to excessive loss of depth information at lower σ values and inclusion of low-quality depth maps at higher values, which disrupts model training. Therefore, $\sigma = 0.2$ is established as the optimal setting to balance accuracy and efficiency.

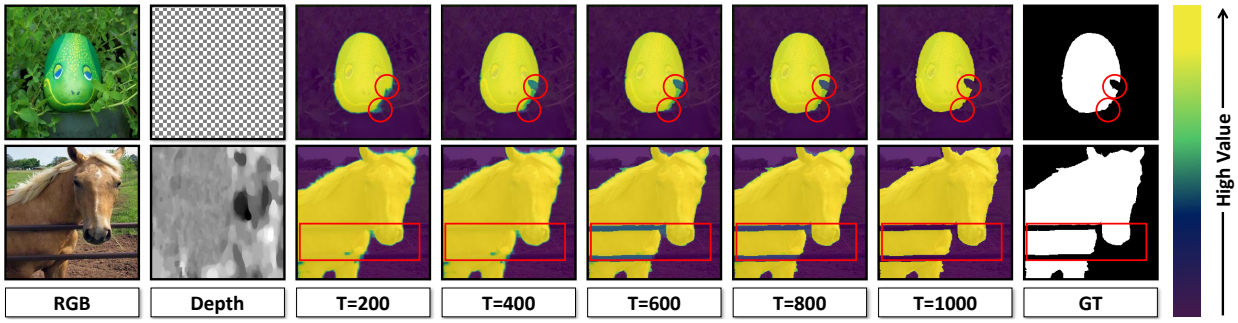


Fig. 7. Visualization of the progressive denoising process for salient feature enhancement.

TABLE III
QUANTITATIVE EVALUATION OF MAJOR COMPONENTS USED IN OUR APPROACH ON BOTH RGB-D BENCHMARK DATASETS NJUD, REDWEB-S AND RGB BENCHMARK DATASETS ECSSD, PASCAL-S. THE BEST RESULTS ARE MARKED IN **BOLD**.

	Major Components							Datasets and Metrics															
	QAF	ACA			SCA			NJUD				ReDweb-S				ECSSD				PASCAL-S			
		AC	MFE	E'	SC	CLIP	SGI	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓
1	✘	✘	✘	✘	✘	✘	✘	.908	.898	.909	.042	.740	.744	.801	.133	.921	.918	.929	.041	.859	.846	.897	.072
2	✔	✔	✘	✘	✘	✘	✘	.918	.916	.927	.032	.758	.755	.807	.119	.924	.937	.944	.032	.866	.856	.906	.059
3	✔	✔	✔	✘	✘	✘	✘	.925	.923	.927	.027	.747	.752	.807	.119	.930	.935	.947	.037	.871	.856	.909	.058
4	✔	✔	✔	✔	✘	✘	✘	.931	.925	.932	.026	.764	.763	.814	.116	.939	.945	.955	.027	.878	.864	.911	.053
5	✔	✔	✘	✔	✘	✘	✘	.921	.918	.925	.028	.751	.756	.811	.117	.929	.932	.949	.035	.865	.852	.907	.061
6	✘	✘	✘	✘	✔	✘	✘	.926	.919	.925	.032	.752	.753	.805	.120	.928	.935	.947	.031	.865	.853	.907	.055
7	✘	✘	✘	✘	✔	✔	✔	.935	.934	.935	.023	.763	.762	.821	.114	.941	.948	.958	.026	.879	.871	.912	.052
8	✘	✘	✘	✘	✔	✘	✔	.933	.931	.932	.024	.762	.761	.818	.115	.938	.946	.955	.025	.876	.869	.910	.052
9	✘	✘	✘	✘	✔	✔	✘	.928	.926	.929	.030	.761	.761	.814	.116	.937	.943	.953	.027	.867	.862	.906	.055
10	✘	✔	✔	✔	✔	✔	✔	.932	.929	.933	.026	.759	.764	.818	.115	.939	.942	.950	.027	.875	.869	.908	.053
11	✔	✔	✔	✔	✔	✔	✔	**.938**	**.937**	**.938**	**.022**	**.769**	**.767**	**.825**	**.111**	**.945**	**.955**	**.961**	**.023**	**.884**	**.879**	**.914**	**.050**

QAF: Quality-aware Filter
ACA: Appearance Control Adapter
SCA: Structure Control Adapter
AC: Appearance Condition
MFE: Multi-scale Feature Embedding
E': Copied Denoising Encoder
SC: Structure Condition
CLIP: CLIP Image Encoder
SGI: Semantic Guidance Injection

TABLE IV
ABLATION STUDY ON THRESHHOLD σ IN EQ. 4.

Sets	ReDweb-S				ECSSD			
Metrics	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓
σ = 0.1	.760	.758	.815	.114	.938	.949	.955	.026
σ = 0.2	**.769**	**.767**	**.825**	**.111**	**.945**	**.955**	**.961**	**.023**
σ = 0.3	.765	.766	.819	.112	.942	.952	.957	.024
σ = 0.4	.757	.753	.814	.118	.933	.943	.951	.027

TABLE V
ABLATION STUDY ON MASKING RATIOS IN QUALITY-AWARE FILTER (SEC. III-C). "50%-90%" MEANS THE PROPORTION OF AREAS NOT MASKED.

Sets	ReDweb-S				ECSSD			
Metrics	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓
50%	.751	.755	.814	.115	.927	.939	.950	.025
60%	.753	.759	.819	.114	.936	.943	.953	.022
70%	.765	.762	.823	.112	.941	.950	.958	**.021**
80%	**.769**	**.767**	**.825**	**.111**	**.945**	**.955**	**.961**	.023
90%	.742	.745	.810	.117	.925	.936	.946	.027

2) Effectiveness of Masking Ratio: To assess the impact of masking ratios on the quality-aware filter (QAF, Sec. III-C), we experimented with different ratios. Depth maps of ECSSD dataset are generated by existing depth estimation methods (e.g., DepthFormer [56]). As shown in Table V, A non-masking ratio of 80% yielded the best performance across nearly all metrics. Higher masking ratios can cause significant information loss in even high-quality depth reconstructions, while lower ratios may lead to model overfitting, as the auto-encoder tends to memorize rather than learn meaningful representations. Thus, a non-masking ratio of 80% effectively balances the trade-off between preserving information integrity and avoiding overfitting.

3) Effectiveness of Masked Auto-encoder: We further evaluated the robustness of our QAF (Sec. III-C) through an ablation study comparing three masked auto-encoder approaches: MultiMAE [45], MAE [91], and MIM-Depth [92], as detailed in Table VI. Depth maps of ECSSD dataset are generated by existing depth estimation methods (e.g., DepthFormer [56]). Although MultiMAE outperformed the other models in all metrics, the differences were marginal, indicating that our QAF maintains robust performance across various masked auto-encoder frameworks.

4) Effectiveness of Semantic Guidance Injection: The semantic guidance injection component in our DiffSOD framework, as described in Sec. III-D1, plays a crucial role in combining the semantic condition embedding and diffusion embedding. It enables the model to learn the interaction between the noise and semantic condition features, resulting in

TABLE VI
ABLATION ON MASKED AUTO-ENCODERS IN QUALITY-AWARE FILTER (SEC. III-C).

Sets	ReDweb-S				ECSSD			
Metrics	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓
MIM-Depth	.763	.761	.818	.114	.939	.951	.955	.026
MAE	.765	.764	.822	.112	.941	.954	.957	.025
MultiMAE	**.769**	**.767**	**.825**	**.111**	**.945**	**.955**	**.961**	**.023**

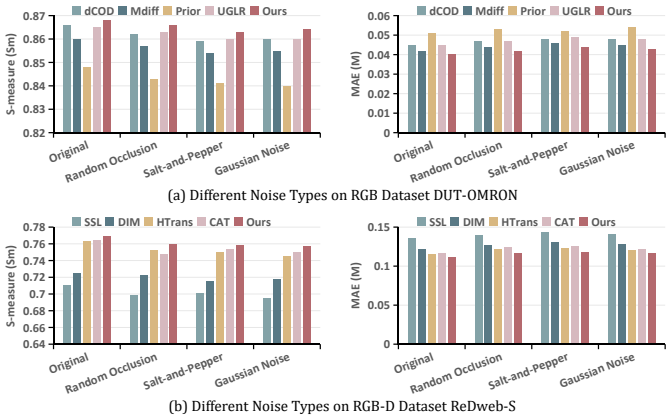
TABLE VII
ABLATION ON SEMANTIC CONDITION INJECTION. "OURS" DENOTES THE COMPLETE SCI VERSION USED IN DIFFSOD.

Sets	ReDweb-S				ECSSD			
Metrics	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓
w/o Co-attention	.761	.760	.815	.119	.938	.942	.953	.028
w/o FFT, + MHA	.765	.763	.818	.115	.941	.948	.957	.025
Ours	**.769**	**.767**	**.825**	**.111**	**.945**	**.955**	**.961**	**.023**

a more robust representation. To substantiate its effectiveness, we conducted experiments to assess the impact of different components. Initially, we replaced the Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT) with multi-head self-attention ("w/o FFT, + MHA"). The performance decline observed in Table VII demonstrates the importance of converting the features into the Fourier space. This conversion helps mitigate the domain gap between the noise and semantic condition features in the Euler space. Furthermore, we conducted experiments where the co-attention mechanism was removed. The results showed a significant performance decline. For instance, in the ReDweb-S dataset, the F-measure dropped from 0.767 (line 3) to 0.760 (line 1). This decline underscores the significant role of the co-attention mechanism in addressing correlation and alignment issues between the noise and semantic condition features.

5) Robustness to Noise: We evaluated our method's robustness to noise by adding Gaussian, random occlusion, and salt-and-pepper noise to the DUT-OMRON (RGB) and ReDweb-S (RGB-D) datasets. As shown in Fig. 8, our method consistently achieved the best S-measure (Sm) and lowest Mean Absolute Error (MAE) across all noise types, outperforming other RGB saliency detection methods (like dCOD [42], Mdiff [41], Prior [89], and UGLR [87]) on DUT-OMRON, and RGB-D methods (like SSL [71], DIM [73], HTrans [32], and CAT [33]) on ReDweb-S. While other methods showed significant performance drops, particularly under salt-and-pepper noise, our approach demonstrated strong resilience, with minimal declines in Sm and slight mean absolute error (MAE) increases. This indicates that our method effectively handles complex noise and challenging scenes, especially in RGB-D datasets where overall scores are typically lower.

6) Denoising Inference and Speed: We conducted an ablation study to analyze the impact of different inference settings during the denoising process. Reducing the number of inference steps can be advantageous for practical applications as it decreases GPU memory consumption and speeds up the inference process. However, as shown in Table VIII, while reducing the number of inference steps does increase FPS (Frames Per Second), it also results in a significant



(a) Different Noise Types on RGB Dataset DUT-OMRON

(b) Different Noise Types on RGB-D Dataset ReDweb-S

Fig. 8. Ablation study on different noises by adding Gaussian, random occlusion, and salt-and-pepper noise to the DUT-OMRON (RGB) and ReDweb-S (RGB-D) datasets.

TABLE VIII
ABLATION STUDY ON DIFFERENT INFERENCE SETTINGS AND INFERENCE SPEED, T DENOTES THE INFERENCE STEP.

Sets	FPS	ECSSD				PASCAL-S			
Metrics		Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓
T = 200	38.9	.924	.931	.936	.039	.859	.852	.887	.065
T = 400	34.5	.928	.939	.943	.033	.868	.861	.895	.062
T = 600	30.8	.934	.943	.950	.029	.876	.867	.899	.059
T = 800	27.1	.939	.950	.955	.025	.880	.872	.909	.052
T = 1000	22.9	**.945**	**.955**	**.961**	**.023**	**.884**	**.879**	**.914**	**.050**

performance drop. Specifically, as the number of inference steps (T) decreases, FPS increases from 22.9 to 38.9, but the performance metrics decrease. In DiffSOD, although using 1000 inference steps results in increased inference time, it significantly improves the performance across all metrics, demonstrating that more inference steps are worthwhile in scenarios where high accuracy is required. At the same time, our model can achieve 38.9 FPS with only 200 inference steps, making it feasible for real-world applications where faster inference is prioritized. With further optimization, we can continue to enhance the speed while maintaining good performance.

7) In-depth Analysis between Unsupervised Methods Using Background Information: Unsupervised RGB-D saliency detection methods, which utilize background cues as supplementary information, encounter several limitations:

Firstly, unsupervised methods that use background as a supplementary information source can be severely limited by the complexity or variability of the background. These methods may fail in scenes where the background and foreground do not have clear contrasting features, leading to inaccurate saliency detection. Secondly, relying on background information primarily for salient object detection often results in a lack of semantic understanding. This approach may neglect important contextual and semantic cues that are crucial for accurate identification and delineation of salient regions, a gap that DiffSOD addresses through its innovative use of saliency conditions and semantic optimization. Thirdly, unsupervised methods may not generalize well across diverse imaging conditions or different types of scenes due to their

reliance on specific background characteristics. In contrast, the denoising diffusion-based approach of DiffSOD, guided by saliency conditions, offers a more adaptable and robust framework for saliency detection across varied scenarios.

Our proposed visual saliency diffusion model addresses these issues with its "noise-to-image denoising" approach, enhancing robustness against background noise and ensuring accurate detection even in challenging conditions. By introducing two specific control conditions — appearance and structure — DiffSOD provides a controlled denoising process that maintains the semantic relevance of salient objects. This methodology offers a robust and adaptable framework for saliency detection across varied scenarios, overcoming the limitations of unsupervised methods.

F. Limitations

While the DiffSOD model shows promise for practical implementation, it is essential to acknowledge the computational challenges associated with its iterative nature. The increased complexity of computations may pose limitations on real-time applications, particularly in resource-constrained environments or scenarios requiring rapid processing. Furthermore, diffusion-based models tend to smooth out image details and edges during denoising. While this effectively reduces noise, it can result in a loss of fine texture and sharpness in the denoised image. Striking a balance between noise reduction and preserving important structural details remains challenging for diffusion models.

V. Conclusions

This paper presents a novel and unified approach to visual saliency detection by adopting a noise-to-image denoising perspective applicable to both RGB and RGB-D salient object detection. The proposed Visual Saliency Diffusion model (DiffSOD) leverages a denoising diffusion-based framework to predict saliency while effectively preserving spatial interactions between pixels. Additionally, we incorporate two distinct control conditions to guide the denoising process, enhancing the accuracy and detail of the saliency maps. Experimental results demonstrate the superior performance of our model compared to existing visual saliency detection methods on benchmark datasets for both RGB and RGB-D SOD tasks.

Looking ahead, our framework introduces a fresh paradigm for diffusion-based dense prediction models and provides a new perspective in visual saliency detection. Future research can focus on adapting our model to handle diverse noise types and integrating additional visual conditions to improve the denoising learning process further.

References

[1] J.-M. Guo, A. W. H. Prayuda, H. Prasetyo, and S. Seshathiri, "Deep learning-based image retrieval with unsupervised double bit hashing," *IEEE TCSVT*, vol. 33, no. 11, pp. 7050–7065, 2023.

[2] X. Zhu, Y. Zhou, D. Wang, W. Ouyang, and R. Su, "Mlst-former: Multi-level spatial-temporal transformer for group activity recognition," *IEEE TCSVT*, vol. 33, no. 7, pp. 3383–3397, 2023.

[3] T. Sun, G. Zhang, W. Yang, J.-H. Xue, and G. Wang, "Trosd: A new rgb-d dataset for transparent and reflective object segmentation in practice," *IEEE TCSVT*, vol. 33, no. 10, pp. 5721–5733, 2023.

[4] M. Song, L. Li, D. Wu, W. Song, and C. Chen, "Rethinking object saliency ranking: A novel whole-flow processing paradigm," *IEEE TIP*, vol. 33, pp. 338–353, 2024.

[5] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*, pp. 8162–8171, 2021.

[6] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, "Medsegdiff: Medical image segmentation with diffusion probabilistic model," in *MIDL*, pp. 1623–1639, PMLR, 2024.

[7] Z. Bai, G. Li, and Z. Liu, "Global–local–global context-aware network for salient object detection in optical remote sensing images," *ISPRS P&RS*, vol. 198, pp. 184–196, 2023.

[8] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion," *IEEE TCYB*, vol. 48, no. 11, pp. 3171–3183, 2018.

[9] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *CVPR*, 2020.

[10] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y.-C. Gu, and M.-M. Cheng, "Mobilesal: Extremely efficient rgb-d salient object detection," *IEEE TPAMI*, 2021.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, pp. 770–778, 2016.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, pp. 4700–4708, 2017.

[14] J. Liu, R. Dian, S. Li, and H. Liu, "Sgfusion: A saliency guided deep-learning framework for pixel-level image fusion," *Information Fusion*, vol. 91, pp. 205–214, 2023.

[15] Y. Zhai, D.-P. Fan, J. Yang, A. Borji, L. Shao, J. Han, and L. Wang, "Bifurcated backbone strategy for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8727–8742, 2021.

[16] M. Song, W. Song, G. Yang, and C. Chen, "Improving rgb-d salient object detection via modality-aware decoder," *IEEE TIP*, vol. 31, pp. 6124–6138, 2022.

[17] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE TPAMI*, vol. 41, no. 7, pp. 1734–1746, 2019.

[18] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE TIP*, vol. 27, no. 7, pp. 3264–3274, 2018.

[19] C. Chen, J. Wei, C. Peng, and H. Qin, "Depth-quality-aware salient object detection," *IEEE TIP*, vol. 30, pp. 2350–2363, 2021.

[20] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion," *IEEE TIP*, vol. 29, pp. 4296–4307, 2020.

[21] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, "Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection," *IEEE TIP*, vol. 31, pp. 6800–6815, 2022.

[22] T. Zhou, H. Fu, G. Chen, Y. Zhou, D. Fan, and L. Shao, "Specificity-preserving rgb-d saliency detection," in *ICCV*, 2021.

[23] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for rgb-d salient object detection," *IEEE TIP*, vol. 30, pp. 3528–3542, 2021.

[24] G. Li, Z. Liu, and H. Ling, "Icnet: Information conversion network for rgb-d based salient object detection," *IEEE TIP*, vol. 29, pp. 4873–4884, 2020.

[25] 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 TNNLS*, vol. 32, no. 5, pp. 2075–2089, 2021.

[26] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *IEEE Conference on Computer Vision and Pattern Recognition*, p. 8579–8588, IEEE, 2020.

[27] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *IEEE CVPR*, pp. 3469–3478, 2020.

[28] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," *arXiv preprint arXiv:2104.12099v2*, 2021.

[29] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "Tritransnet rgb-d salient object detection with a triplet transformer embedding network," in *ACM MM*, 2021.

[30] J. Yuan, A. Zhu, Q. Xu, K. Wattanachote, and Y. Gong, "Ctif-net: A cnn-transformer iterative fusion network for salient object detection," *IEEE TCSVT*, pp. 1–1, 2023.

[31] M. Ma, C. Xia, C. Xie, X. Chen, and J. Li, "Boosting broader receptive fields for salient object detection," *IEEE TIP*, vol. 32, pp. 1026–1038, 2023.

[32] B. Tang, Z. Liu, Y. Tan, and Q. He, "Hrtransnet: Hrformer-driven two-modality salient object detection," *IEEE TCSVT*, pp. 1–1, 2022.

[33] 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 TMM*, pp. 1–14, 2023.

[34] X. Fang, J. Zhu, X. Shao, and H. Wang, "Grouptransnet: Group transformer network for rgb-d salient object detection," *arXiv preprint arXiv:2203.10785v1*, 2022.

[35] X. Zhou, K. Shen, and Z. Liu, "Admnet: Attention-guided densely multi-scale network for lightweight salient object detection," *IEEE TMM*, pp. 1–14, 2024.

[36] K. Shen, X. Zhou, and Z. Liu, "Minet: Multiscale interactive network for real-time salient object detection of strip steel surface defects," *IEEE TII*, vol. 20, no. 5, pp. 7842–7852, 2024.

[37] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Caver: Cross-modal view-mixed transformer for bi-modal salient object detection," *IEEE TIP*, vol. 32, pp. 892–904, 2023.

[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE CVPR*, pp. 10684–10695, 2022.

[39] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *IEEE CVPR*, pp. 6007–6017, 2023.

[40] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[41] J. Wu, R. Fu, H. Fang, Y. Zhang, and Y. Xu, "Medsegdiff-v2: Diffusion based medical image segmentation with transformer," *arXiv preprint arXiv:2301.11798*, 2023.

[42] Z. Chen, R. Gao, T.-Z. Xiang, and F. Lin, "Diffusion model for camouflaged object detection," *arXiv preprint arXiv:2308.00303*, 2023.

[43] X.-J. Luo, S. Wang, Z. Wu, C. Sakaridis, Y. Cheng, D.-P. Fan, and L. Van Gool, "Camdiff: Camouflage image augmentation via diffusion model," *arXiv preprint arXiv:2304.05469*, 2023.

[44] Y. Duan, X. Guo, and Z. Zhu, "Diffusiondepth: Diffusion denoising approach for monocular depth estimation," *arXiv preprint arXiv:2303.05021*, 2023.

[45] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *ECCV*, pp. 348–367, 2022.

[46] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2020.

[47] J. Li, W. Ji, Q. Bi, C. Yan, M. Zhang, Y. Piao, H. Lu, and L. cheng, "Joint semantic mining for weakly supervised rgb-d salient object detection," in *NeurIPS*, vol. 34, 2021.

[48] Z. Zhang, Z. Lin, J. Xu, W. Jin, S. Lu, and D. Fan, "Bilateral attention network for rgb-d salient object detection," *IEEE TIP*, 2021.

[49] M. Moritz, H. Christian, and G. Andreas, "Object scene flow," *ISPRSJPRS*, 2018.

[50] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, pp. 746–760, 2012.

[51] M. Zhang, S. Yao, B. Hu, Y. Piao, and W. Ji, "C^2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection," *IEEE TMM*, pp. 1–13, 2022.

[52] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, pp. 3836–3847, 2023.

[53] C. Chen, M. Song, W. Song, L. Guo, and M. Jian, "A comprehensive survey on video saliency detection with auditory information: the audio-visual consistency perceptual is the key!," *IEEE TCSVT*, 2022.

[54] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *ICIP*, 2014.

[55] H. Peng, L. Bing, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *ECCV*, 2014.

[56] Z. Li, Z. Chen, X. Liu, and J. Jiang, "Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation," *Machine Intelligence Research*, vol. 20, no. 6, pp. 837–854, 2023.

[57] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE CVPR*, pp. 3796–3805, 2017.

[58] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE CVPR*, pp. 3166–3173, 2013.

[59] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE CVPR*, pp. 1155–1162, 2013.

[60] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE CVPR*, pp. 5455–5463, 2015.

[61] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *IEEE CVPR*, pp. 280–287, 2014.

[62] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *CVPR*, 2012.

[63] N. Liu, N. Zhang, L. Shao, and J. Han, "Learning selective mutual attention and contrast for rgb-d saliency detection," *IEEE TPAMI*, 2021.

[64] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," *IJCV*, vol. 129, p. 2622–2638, 2021.

[65] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009.

[66] D. Fan, C. Gong, Y. Cao, B. Ren, M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *IJCAI*, 2018.

[67] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *IEEE International Conference on Computer Vision*, pp. 7253–7262, 2019.

[68] J. Zhao, C. Y, D.-P. Fan, M. Cheng, X. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3922–3931, 2019.

[69] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *CVPR*, 2020.

[70] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection," in *CVPR*, 2020.

[71] X. Zhao, Y. Pang, L. Zhang, , H. Lu, and X. Ruan, "Self-supervised pretraining for rgb-d salient object detection," in *AAAI*, 2022.

[72] A. Li, Y. Mao, J. Zhang, and Y. Dai, "Mutual information regularization for weakly-supervised rgb-d salient object detection," *IEEE TCSVT*, pp. 1–1, 2023.

[73] S. Yao, M. Zhang, Y. Piao, C. Qiu, and H. Lu, "Depth injection framework for rgbd salient object detection," *IEEE TIP*, vol. 32, pp. 5340–5352, 2023.

[74] Y. Luo, F. Shao, Z. Xie, H. Wang, H. Chen, B. Mu, and Q. Jiang, "Hfmdnet: Hierarchical fusion and multilevel decoder network for rgb-d salient object detection," *IEEE TIM*, vol. 73, pp. 1–15, 2024.

[75] X. Hu, F. Sun, J. Sun, F. Wang, and H. Li, "Cross-modal fusion and progressive decoding network for rgb-d salient object detection," *IJCV*, pp. 1–19, 2024.

[76] H. Chen, F. Shen, D. Ding, Y. Deng, and C. Li, "Disentangled cross-modal transformer for rgb-d salient object detection and beyond," *IEEE TIP*, vol. 33, pp. 1699–1709, 2024.

[77] B. Xu, H. Liang, R. Liang, and P. Chen, "Slocate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *AAAI*, vol. 35, p. 3004–3012, 2021.

[78] J. Li, W. He, and H. Zhang, "Towards complex backgrounds: A unified difference-aware decoder for binary segmentation," *arXiv preprint arXiv:2210.15156*, 2022.

[79] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren, "Edn: Salient object detection via extremely-downsampled network," *IEEE TIP*, vol. 31, pp. 3125–3136, 2022.

[80] Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, "Pixels, regions, and objects: Multiple enhancement for salient object detection," in *IEEE CVPR*, pp. 10031–10040, 2023.

[81] H. Zhou, B. Qiao, L. Yang, J. Lai, and X. Xie, "Texture-guided saliency distilling for unsupervised salient object detection," in *IEEE CVPR*, pp. 7257–7267, 2023.

[82] Z. Wang, Y. Zhang, Y. Liu, C. Qin, S. A. Coleman, and D. Kerr, "Larnet: Towards lightweight, accurate and real-time salient object detection," *IEEE TMM*, vol. 26, pp. 5207–5222, 2024.

[83] G. Liu, B. Xu, H. Huang, C. Lu, and Y. Guo, "Sdetr: Attention-guided salient object detection with transformer," in *IEEE ICASSP*, pp. 1611–1615, 2022.

[84] B. Xu, G. Liu, H. Huang, C. Lu, and Y. Guo, "Semantic distillation guided salient object detection," *arXiv preprint arXiv:2203.04076*, 2022.

[85] Y. K. Yun and W. Lin, "Selfreformer: Self-refined network with transformer for salient object detection," *arXiv preprint arXiv:2205.11283*, 2022.

[86] Y. Mao, J. Zhang, Z. Wan, X. Tian, A. Li, Y. Lv, and Y. Dai, "Generative transformer for accurate and reliable salient object detection," *IEEE TCSVT*, 2024.

[87] S. Ren, N. Zhao, Q. Wen, G. Han, and S. He, "Unifying global-local representations in salient object detection with transformers," *IEEE TETCI*, pp. 1–10, 2024.

[88] L. Zhang and Q. Zhang, "Salient object detection with edge-guided learning and specific aggregation," *IEEE TCSVT*, vol. 34, pp. 534–548, 2024.
[89] G. Zhu, J. Li, and Y. Guo, "Priornet: Two deep prior cues for salient object detection," *IEEE TMM*, vol. 26, pp. 5523–5535, 2024.
[90] Z. Chen, K. Sun, and X. Lin, "Camodiffusion: Camouflaged object detection via conditional diffusion models," in *AAAI*, vol. 38, pp. 1272–1280, 2024.
[91] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, pp. 16000–16009, 2022.
[92] Z. Xie, Z. Geng, J. Hu, Z. Zhang, H. Hu, and Y. Cao, "Revealing the dark secrets of masked image modeling," in *CVPR*, pp. 14475–14485, 2023.