DiMSOD: A Diffusion-Based Framework for Multi-Modal Salient Object Detection

Shuo Zhang[1], Jiaming Huang[2], Wenbing Tang[3], Yan Wu[2], Terrence Hu[2], Xiaogang Xu[4], Jing Liu[1]*

[1]Shanghai Key Laboratory of Trustworthy Computing, East China Normal University
[2]Technology Center, Huolala
[3]College of Computing and Data Science, Nanyang Technological University
[4]The Chinese University of Hong Kong
shuo_zhang_ecnu@163.com, ucashjm@gmail.com, jliu@sei.ecnu.edu.cn

Abstract

Multi-modal salient object detection (SOD) through the integration of additional data such as depth or thermal information has become a significant task in computer vision during recent years. Traditionally, the challenges of identifying salient objects in RGB, RGB-D (Depth), and RGB-T (Thermal) images are tackled separately. However, without intricate cross-modal fusion strategies, such approaches struggle to effectively integrate multi-modal information, often resulting in poorly defined object edges or overconfident inaccurate predictions. Recent studies have shown that designing a unified end-to-end framework to handle all three types of SOD tasks simultaneously is both necessary and difficult. To address this need, we propose a novel approach that treats multi-modal SOD as a conditional mask generation task utilizing diffusion models. We introduce DiMSOD, which enables the concurrent use of local (depth maps, thermal maps) and global controls (original images) within a unified model for progressive denoising and refined prediction. DiMSOD only requires fine-tuning newly introduced modules on a pretrained stable diffusion trained on RGB images, which not only reduces fine-tuning costs for practical applications but also enhances the integration of multi-modal conditional controls. Specifically, we have developed modules including SOD-ControlNet, Feature Adaptive Network (FAN), and Feature Injection Attention Network (FIAN) to enhance the model's performance. Extensive experiments demonstrate that DiMSOD efficiently detects salient objects across RGB, RGB-D, and RGB-T datasets, achieving superior performance compared to previous well-established methods.

Introduction

Salient Object Detection (SOD) aims to accurately detect and locate the most salient objects in a given image, mimicking the human visual perception system (Gao et al. 2023b). It serves as an essential preliminary step for numerous other computer vision applications, including object detection (Cheng et al. 2023), visual tracking (Li et al. 2023), image segmentation (Chen et al. 2024), and quality assessment (Zhai and Min 2020). Despite the recent progress in SOD (Cai et al. 2024), most of these approaches primarily focus on processing individual RGB images. Moreover, achieving accurate SOD results in challenging background and complex scenes remains difficult. In recent years, extensive use of depth cameras and infrared imaging devices have shown that the depth and thermal data gathered can significantly improve the performance of salient object detection. Nevertheless, the task of effectively combining multi-modal complementary information without overestimating point estimates remains a considerable challenge, and it greatly influences the achievement of robust detection performance.

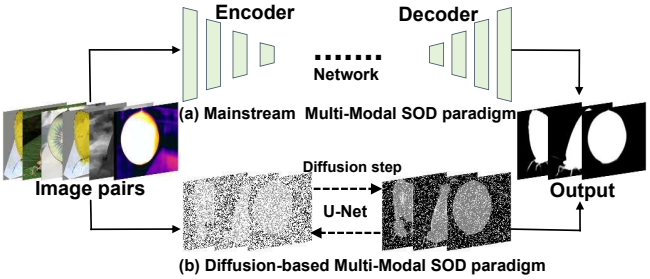

Figure 1: (a) The current discriminative paradigm feeds images into the network for one-way prediction, producing a deterministic segmentation mask. (b) In contrast, we propose a novel generative paradigm that reformulates multi-modal SOD as a forward-and-reverse diffusion process.

Currently, benefiting from the advancements of deep neural networks, SOD methods have evolved from designing ingenious low-level features (Jian and Yu 2023) to learning high-level models. By incorporating the advantages of dense feature interaction, diverse attention module (Gu et al. 2023), and multi-task learning pipeline (Zong et al. 2023), deep learning-based methods have emerged as a promising technology for the SOD task. However, existing deep learning-based methods only focus on one specific type of input data. Considering the detection needs of RGB, RGB-D and RGB-T data, it is essential to develop a unified and comprehensive method to accommodate different data types. In response, a few recent studies (Gao et al. 2021; Luo et al. 2024; Wang et al. 2024) have focused on this direction. Although these methods have paved the way for the much-needed unified approach to multi-modal salient object detection, they still struggle to achieve precise localization and

*Corresponding Author.

segmentation in complex scenarios. This limitation arises from their adherence to the paradigm illustrated in Fig. 1 (a), where a deterministic discriminative network solution generates a single output for a given input image and lacks adaptable cross-modal fusion strategies. As a result, they fail to effectively integrate multi-modal complementary information while avoiding overconfident incorrect predictions.

Considering the unique challenges of multi-modal SOD, we propose adopting a generative paradigm based on diffusion models. As illustrated in Fig. 1 (b), we reformulate the multi-modal SOD task into a generative process, training the model to produce salient object masks by constructing a conditional noise-to-mask paradigm. Diffusion models (Ho, Jain, and Abbeel 2020a) have recently shown exceptional efficacy in generative modeling, particularly in conditional generation tasks (Dhariwal and Nichol 2021). Their inherent iterative denoising mechanism replaces the need for complex refinement modules in popular multi-modal SOD models, allowing for gradual distinction between salient object boundaries and background context. The random sampling process enables the generation and evaluation of multiple predictions, which further reduces the risk of the model making overconfident and erroneous estimations. The integration of ControlNet (Zhang, Rao, and Agrawala 2023) with diffusion models introduces cross-modal information, thereby providing better guidance for the denoising process. However, applying diffusion models and ControlNet to multi-modal SOD directly still faces several shortcomings, including limited discriminative ability, sparse features, inadequate mask refinement, high fine-tuning costs, and relatively unstable controllability. To address these, we have tailored our method, DiMSOD, which leverages the denoising process of diffusion models to progressively correct the discrepancies between the initial noise and the ground truth. Depth maps and thermal maps are utilized as local auxiliary control conditions, while the image itself serves as a global auxiliary control condition. We have also improved upon ControlNet and proposed SOD-ControlNet, embedding our proposed Feature Adaptive Network (FAN) into the ControlNet and altering the method of conditional injection. By employing a multi-scale conditional injection strategy, we inject the introduced cross-modal complementary information from depth maps and thermal maps into all resolutions. This significantly enhances the expressive power of DiMSOD, reduces the model size and fine-tuning costs, and improves the accuracy of salient objects detection. Furthermore, to effectively bridge the gap between the diffusion noise embeddings and the conditional semantic features when integrating global image control into the model, we designed a Feature Injection Attention Network (FIAN). This network enhances the denoising process by aggregating the conditional semantic features of the image with the features from the diffusion model encoder through an improved cross-attention module.

To summarize, our main contributions are as follows:

- We are the first to formulate the multi-modal SOD as a generative denoising process and propose a diffusion-based model, DiMSOD. It detects salient objects via a noise-to-mask paradigm, using input images alone or combined with complementary depth or thermal maps.
- We introduce SOD-ControlNet, a module specifically designed for multi-modal SOD. It integrates our proposed Feature Adaptive Network (FAN), which injects complementary depth or thermal information at all resolutions, avoids interference between different modalities, and enables efficient adaptive cross-modal information fusion.
- We have also designed a Feature Injection Attention Network (FIAN) to facilitate the interaction between diffusion noise embeddings and salient image semantic features, thereby further integrating global guiding information from the image to enhance the denoising process.

Related Work

SOD is a fundamental task in computer vision (Li et al. 2024; Xia et al. 2024). There are currently numerous methods focused on SOD for individual types of data, such as RGB, RGB-D, or RGB-T (Pang et al. 2023; Lee et al. 2023; Konwer et al. 2023). However, as solving multi-modal SOD using one single model is a relatively novel field of study, there are only a limited number of methods available currently, among which the most notable ones are MMNet (Gao et al. 2021), LAFB (Wang et al. 2024), and VSCode (Luo et al. 2024). In MMNet (Gao et al. 2021), a cross-modal multi-stage fusion module (CMFM) is proposed, which consists of two stages: feature response and adversarial combination. This module explores the complementarity of information from different modalities. In LAFB (Wang et al. 2024), a novel adaptive fusion bank is introduced to fully leverage the complementary advantages of a set of basic fusion schemes, enabling the simultaneous handling of various challenges for robust multi-modal SOD. VSCode (Luo et al. 2024) utilizes visual saliency transformer as the foundational model and incorporates 2D prompts and discrimination loss within the encoder-decoder architecture. This approach facilitates the learning of both domain and task-specific knowledge, as well as critical shared knowledge.

Diffusion models (Cao et al. 2024) sample noisy images using a forward Gaussian diffusion process and refine them iteratively through a backward denoised process to generate images. Diffusion models have demonstrated significant potential across various practical fields, such as image super-resolution (Gao et al. 2023a), image synthesis (Gu et al. 2022), image inpainting (Zhang et al. 2023), depth estimation (Ke et al. 2023), medical image segmentation (Zhang, Rao, and Agrawala 2023), and semantic segmentation (Ji et al. 2023). Different from these works, we propose the first diffusion-based model for multi-modal SOD. Moreover, our proposed SOD-ControlNet, Feature Adaptive Network, and Feature Injection Attention Network modules are perfectly aligned with the specific requirements of multi-modal SOD, effectively addressing unique challenges that other multi-modal SOD approaches have been unable to overcome.

Method

Network Architecture

We have developed DiMSOD, a model built upon a pre-trained text-to-image LDM, Stable Diffusion (SD) (Rombach et al. 2022), which leverages the excellent image pri-

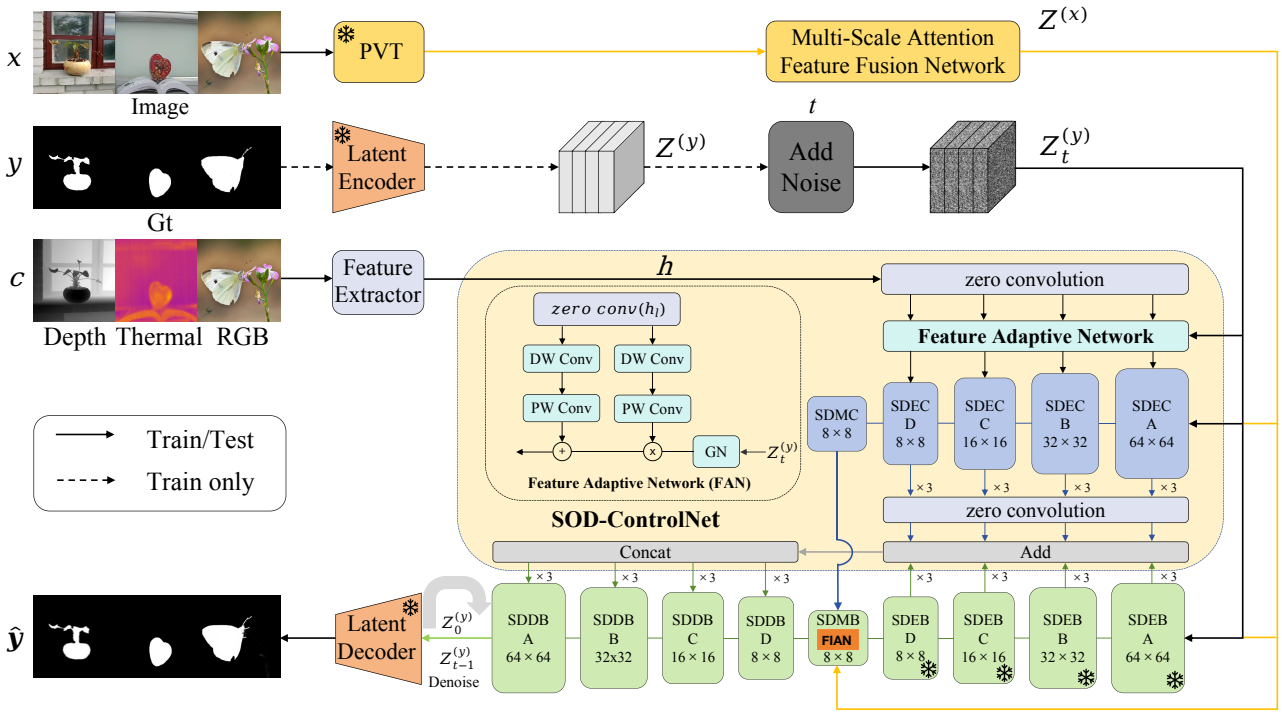

Figure 2: **Overview of the Architecture of DiMSOD**. We encode the original image $\mathbf{x}$ using PVT, and the saliency map $\mathbf{y}$ using the pre-trained Stable Diffusion latent encoder. After that, we fine-tune the U-Net by optimizing the self-diffusion noise loss relative to the saliency latent code. The SOD-ControlNet with Feature Adaptive Network (FAN) is designed to efficiently harness cross-modal information, such as depth maps and thermal maps, to control the precise generation of saliency masks. The Feature Injection Attention Network (FIAN) is introduced to explicitly guide the diffusion process by facilitating interactions between the diffusion information and image features, extracted from the PVT backbone. These interactions generate more direct conditional semantic features, which play a crucial role in providing more nuanced guidance for the diffusion process.

ors obtained from LAION-5B (Schuhmann et al. 2022). With minimal modifications to the model components, we have converted it into a salient object detector fine-tuned on RGB images. By further introducing our proposed additional components along with depth and thermal information, it is capable of achieving multi-modal salient object detection. Fig. 2 provides an overview of the network architecture of DiMSOD. We introduced SOD-ControlNet with a Feature Adaptive Network (FAN) to extract conditional semantic features and cross-modality interaction features from RGB images and depth maps or thermal maps, resulting in condition features rich in multi-scale details. Moreover, we developed a Feature Injection Attention Network (FIAN) with a cross-attention mechanism to extract and utilize salient information from global conditional semantic features for guiding the denoising process. These modules effectively integrate cross-modal information, bridging the gap between diffusion and image features, and helping the denoising network generate more precise object boundaries.

Saliency Encoder, Decoder and Feature Extractor. We start by separately encoding the input image using a pyramid vision transformer (Wang et al. 2021) backbone and the associated saliency map using the pre-trained latent encoder. To make the saliency map compatible with the model, we expand it to three channels to mimic an RGB image. This adjustment is essential because the latent encoder was originally designed for 3-channel (RGB) inputs, whereas the saliency map contains only a single channel. The saliency maps can be reconstructed from encoded latent codes without modifying the latent space or decoder. During inference, the predicted saliency map is obtained by averaging the three channels of the saliency latent code after it has been decoded at the end of the diffusion process. Our feature extractor is composed of a stack of convolutional layers and SiLU activations, enabling the extraction of conditional features at various resolutions. Batch Normalization (BN) computes normalization statistics across all local controls within a batch, which inevitably diminishes the unique characteristics of each individual local control. While BN is well-suited for cases involving relatively large mini-batches with uniform data distributions, it becomes less effective in multi-modal SOD tasks. Because data distributions across different modalities vary significantly, making batch-wide normalization unsuitable for multi-modal scenarios. As the outputs produced by stable diffusion are largely determined by the input image and its corresponding local control instance, Instance Normalization (Ulyanov 2016) (IN) emerges as a more appropriate choice. IN not only facilitates faster model convergence but also ensures that the independence of each local control instance is preserved. The Feature Adaptive Network retrieves features from various modalities while avoiding batch normalization interference in the learning

process. In addition, the feature extractor projects the extra conditions into the corresponding latent spaces of different encoding layers, thereby enhancing the alignment between the local control features and diffusion noise features.

SOD-ControlNet. SD employs a U-Net-like (Ronneberger, Fischer, and Brox 2015) architecture as its denoising model, consisting primarily of an encoder, a middle block, and a decoder. Each encoder and decoder contains 12 corresponding blocks. Inspired by ControlNet (Zhang, Rao, and Agrawala 2023), we introduced SOD-ControlNet, as depicted in Fig. 2. In this illustration, each SDEB and SDDB represents an encoder block and a decoder block, respectively, while each SDEC represents the copied versions of SDEB. The diagram shows four blocks, each of which needs to be repeated three times. Additionally, SDMB denotes the middle block, and SDMC is the copied version of SDMB. For brevity, we denote the output of the i-th block in SDEB and SDDB as e_i and d_i . Similarly, $e^{'}$ denotes the output of the i-th block in SDEC, while m and m' represent the output of SDMB and SDMC, respectively. Due to the skip connections in the U-Net and the need to incorporate the local control information from the SOD-ControlNet during the decoding process, as shown in Fig. 2, we modify the input for the first decoder block as $concat(m + m', e_{12} + Z(e^{'}_{12}))$. For subsequent i-th decoder blocks, the input is modified as $concat(d_{i-1}, e_j + Z(e^{'}_j))$. Here, the sum of i and j equals 13, which represents the total number of either encoder or decoder blocks in the U-Net, along with the middle block. Z represents a zero convolutional layer whose weights progressively increase from zero, gradually embedding salient global and local control information into stable diffusion.

Our SOD-ControlNet differs from ControlNet in the way it handles conditions. While ControlNet directly adds conditions to the input noise and injects them into copied encoders, we employ a multi-scale condition injection strategy, adapting the conditions to all resolutions before injecting them into the copied encoders. Specifically, we begin by extracting multi-resolution local conditional control features (Depth or Thermal map) using a feature extractor. We then select the first block from each resolution level (i.e., 64×64, 32×32, 16×16, and 8×8) within the copied encoder (i.e., the SDEC in Fig. 2) for condition injection. Inspired by the Feature Denormalization technique in SPADE (Park et al. 2019) , we develop the Feature Adaptive Network (FAN) for the injection process. By substituting instance normalization for batch normalization in the feature extractor, FAN is able to learn features adaptively from various modalities, thus eliminating the potential interference that batch normalization may cause when learning features from different modalities. Additionally, FAN can modulate the normalization (i.e., Group Normalization) of the input self-diffusion noise features using conditional features. As shown in Fig. 2, $Z_t^{(y)}$ represents self-diffusion noise features, with a resolution of l, for simplicity, l has been omitted from the notation. The h_l denotes the features obtained from local conditions c (Depth map, Thermal map, RGB image) after passing through the feature extractor, with a resolution of l. Learnable Depthwise Separable Convolution layers, consisting of Depthwise (DW) and Pointwise (PW) convolutions, are employed to map extra local condition features into spatially-sensitive scale and shift modulation coefficients, enabling dynamic and flexible adaptive feature modulation.

Multi-scale Attention Feature Fusion Network and Feature Injection Attention Network. For an RGB image $\mathbb{R}^{H \times W \times 3}$, we use the Pyramid Vision Transformer (Wang et al. 2021) as our visual backbone to extract the top three high-level image features $Z_i^{(x)}$, $i \in \{1, 2, 3\}$, with resolutions of $\frac{H}{s} \times \frac{W}{s}$, where $s \in \{8, 16, 32\}$. These features are combined using a trainable multi-scale feature fusion network, as proposed in SeeCoder (Xu et al. 2024), which results in image features $Z^{(x)}$. Due to space limitations, please refer to the attention module design in SeeCoder for more detailed information, where we have adopted its simplified version. To incorporate salient information and semantic details from the original input features during the denoising process, we propose the Feature Injection Attention Network (FIAN), which is integrated into the the UNet-based denoising network. We use the multi-scale attention feature $Z^{(x)}$ and the deepest diffusion feature e_{12} from SDEB as inputs to FIAN. In detail, we utilize e_{12} to generate the query $\mathbf{Q}$, key $\mathbf{K}$, and the value $\mathbf{V_1}$ by linear projection of a square matrix. Similarly, we use $Z^{(x)}$ to generate the intermediary $\mathbf{P}$ and value $\mathbf{V_2}$. To reduce computational complexity and perform information weighting and fusion, we do not generate $\mathbf{Q}$ and $\mathbf{K}$ for $Z^{(x)}$. Instead, we use $\mathbf{P}$ as an intermediary to connect with e_{12}. We define M_1 and M_2 as the intermediate results of the feature maps, and m as the final SDMB output feature map, where $M_1 = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$, $M_2 = \text{softmax}\left(\frac{KP^T}{\sqrt{d}}\right)$, $m = M_1 \times M_2 \times (V_1 + V_2)$.

Training

We formulate multi-modal SOD as a conditional denoising diffusion generation task and train DiMSOD to fit the conditional distribution $D(\mathbf{y}|\mathbf{x}, c)$ over saliency $\mathbf{y} \in \mathbb{R}^{H \times W}$, where the global condition x is input image and the local condition c is the corresponding depth map or thermal map. In the *forward* process, which begins at $\mathbf{y}_0 := \mathbf{y}$ from the conditional distribution, Gaussian noise is gradually added to the ground-truth $\mathbf{y}_t$ at levels $t \in \{1, ..., T\}$ to obtain noisy mapping $\mathbf{y}_t$ as $\mathbf{y}_t = \sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, $\bar{\alpha}_t := \prod_{s=1}^{t} 1 - \beta_s$, and $\{\beta_1, \ldots, \beta_T\}$ denotes the variance schedule of a process with T steps. However, during actual training, the sparsity of certain sparse ground truth (GT) can easily lead to model training failures. To address this issue, we first add noise to the GT, transforming it into a form of random noise. We then apply a self-diffusion method (Duan, Guo, and Zhu 2025) to process it. Simultaneously, we reorganize the entire sparse ground truth during training using techniques such as random cropping, jittering, and flipping augmentations. This approach helps ensure that the model does not focus solely on the inherent erroneous parts of the prediction. In the *reverse* process, the noise present in $\mathbf{y}_t$ is progressively eliminated to produce $\mathbf{y}_{t-1}$ using the conditional denoising model $\boldsymbol{\epsilon}_\theta(\cdot)$, which is param-

eterized by learned parameters. To allow the input image $\mathbf{x}$ to conditionally guide the latent denoiser $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{(\mathbf{y})}, \mathbf{z}^{(\mathbf{x})}, c, t)$, we leverage the image features $\mathbf{z}^{(\mathbf{x})}$ to interact with the noisy saliency map $\mathbf{z}_t^{(\mathbf{y})}$ through a cross-attention mechanism.

During training, the parameters θ are updated based on the parameters fine-tuned using the RGB training data. The salient mask $\mathbf{y}$ is then noised via self-diffusion with sampled multi-resolution noise $\boldsymbol{\epsilon}$ at a randomly selected timestep t. The noise estimate $\hat{\boldsymbol{\epsilon}}$ is computed using $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{y}_t, \mathbf{x}, c, t)$. Finally, the denoising diffusion objective function $\mathbb{E}_{\mathbf{y}_0, c, \boldsymbol{\epsilon}\sim\mathcal{N}(0,I), t\sim\mathcal{U}(T)} |\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}|_2^2$ is minimized. Latent diffusion models improve computational efficiency and image generation by performing diffusion in a low-dimensional latent space, which is created within the latent encoder's bottleneck and is trained separately from the denoiser (Rombach et al. 2022). To translate our formulation into the latent space, a latent code is defined as $\mathcal{E}$: $\mathbf{z}^{(\mathbf{y})} = \mathcal{E}(\mathbf{y})$, which is generated by the encoder. Given a saliency latent code, the saliency mask can be reconstructed using the decoder $\mathcal{D}$ as follows: $\hat{\mathbf{y}} = \mathcal{D}(\mathbf{z}^{(\mathbf{y})})$. The conditioning image $\mathbf{x}$ is transformed into $\mathbf{z}^{(\mathbf{x})}$ via PVT. Subsequently, the denoiser is trained in the latent space as $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{(\mathbf{y})}, \mathbf{z}^{(\mathbf{x})}, c, t)$. The adapted inference procedure introduces an additional step in which the decoder $\mathcal{D}$ reconstructs the data $\hat{\mathbf{y}}$ from the estimated clean saliency latent code $\mathbf{z}_0^{(\mathbf{y})}$: $\hat{\mathbf{y}} = \mathcal{D}(\mathbf{z}_0^{(\mathbf{y})})$.

Inference

During inference phase, the target saliency mask $\mathbf{y} := \mathbf{y}_0$ is reconstructed by iteratively applying the denoiser $\boldsymbol{\epsilon}_\theta(\mathbf{y}_t, \mathbf{x}, c, t)$ to a normally distributed variable $\mathbf{y}_T$. We begin by initializing the saliency latent code with standard Gaussian noise and encoding the input image via PVT and a multi-scale attention feature fusion network. We then progressively denoise this latent code following the same schedule used during training. From our experience, we have observed that initializing with standard Gaussian noise yields better results compared to using multi-resolution noise, despite the model being trained with the latter. To expedite the inference process, we adopt the non-Markovian sampling with recalibrated steps as described in DDIM (Song, Meng, and Ermon 2020). Finally, using the latent decoder, we generate the ultimate saliency map from the latent code and apply channel-wise averaging for further post-processing.

Experiments

Experimental Setup

DiMSOD is trained jointly using three different types of SOD datasets, following recent work, our training dataset consists of the following subsets and resize it to 512×512 : the RGB dataset DUTS-TR (Wang et al. 2017) with 10,553 images, the RGB-T dataset VT5000 (Tu et al. 2022b) with 2,500 images, the RGB-D dataset NJUD (Ju et al. 2014) with 1,485 image, NLPR (Peng et al. 2014) with 700 images, and DUTLF-Depth (Piao et al. 2019) with 800 images. Stable Diffusion is used as our backbone when implementing DiMSOD. The initial pre-training configurations with a v-objective (Salimans and Ho 2022) are adhered to our experiments. In training, we implement the DDPM noise scheduler (Ho, Jain, and Abbeel 2020b) with 1,000 diffusion steps. For inference, we employ DDIM scheduler and sample 20 steps. For the final prediction, we combine outcomes from 10 inference iterations initiated with diverse initial noise. Training our method takes 100 epochs with a batch size of 32 on 4 Nvidia A100 GPU cards. We adopt the Adam optimizer with a learning rate of 3×10^{-5}. We also implement training data augmentation strategies through the application of random horizontal and vertical flips.

Evaluation Datasets and Metrics

For RGB datasets, we evaluate DiMSOD on 5 widely used benchmark datasets that are not seen during training, including DUT-OMRON (5,168 images), ECSSD (1,000 images), PASCAL-S (850 images), HKU-IS (4,447 images), and DUTS-TE (5,019 Images). For RGB-D datasets , we use the test sets of DUTLF-Depth (400 images), NJUD (500 images), NLPR (300 images), SIP (929 images), LFSD (100 images). For RGB-T datasets , we use the testset of VT5000 (2,500 images) ,VT821 (821 images) , VT1000 (1,000 images). Four widely accepted evaluation metrics are thoroughly assessed on each dataset: **F_β-measure**, **MAE**, **E-measure** (Fan et al. 2018), **S-measure** (Fan et al. 2017).

Comparisons with state-of-the-art

For all the RGB, RGB-D, and RGB-T experiments, we conducted comprehensive comparisons with state-of-the-art multi-modal SOD methods, including MMNet, LAFB, and VSCode. Additionally, for RGB SOD, we compared DiMSOD against other specialized methods , namely, F^{3}Net, MINet (Pang et al. 2020), U^{2}Net, PFSNet (Ma, Xia, and Li 2021), VST(Liu et al. 2021), EDN (Wu et al. 2022), SHNet (Zhang et al. 2022), SRfor (Yun and Lin 2022), USOD (Zhou et al. 2023a), M^{3}Net (Yuan, Gao, and Tan 2023), and UTD (Huo et al. 2024). For RGB-D SOD, the compared methods are HDF-Net, CMWNet (Li et al. 2020), DANet, PGA-Net, BBS-Net, DDCNN (Wang et al. 2022) and PICR. For RGB-T SOD, the compared methods are R3Net, SGDL, M3S-NIR, ADF (Tu et al. 2022b), DCNet (Tu et al. 2022a), and LSNet (Zhou et al. 2023b). For fair comparisons, all results either come directly from the authors or are reproduced using the model retrained on the identical training dataset with the suggested settings. The code for evaluating the model is derived from F^{3}Net.

Quantitative Evaluation. For RGB SOD results, see Table 1. We can find that DiMSOD outperformed in almost all metrics across the three benchmark datasets. It demonstrates the excellent performances of the proposed DiMSOD. Due to space limitations, we present only a subset of the results. On the three larger datasets, DUTS-TE, HKUIS, and PASCAL-S, we achieved average improvements of 1.05%, 5%, 1.43%, and 1.07% in F_β, MAE, E_ξ, and S_α, respectively, compared to the second-best method. For RGB-D SOD, as shown in Table 2, our method also outperforms the second-best model by 6.48% and 1.15% in MAE and F_β, respectively. See Table 3 for RGB-T SOD, our DiMSOD achieves top-tier performance, despite some wins and losses against VSCode on VT1000. Compared to all other

Methods	DUTS-TE				HKU-IS				PASCAL-S			
	$F_\beta\uparrow$	$M\downarrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$
F^3Net_{20}	.891	.035	.901	.888	.936	.028	.952	.917	.871	.061	.858	.854
$MINet_{20}$	.883	.037	.897	.884	.934	.028	.953	.918	.866	.063	.850	.849
$U2Net_{20}$	.872	.044	.886	.873	.935	.031	.948	.915	.859	.073	.842	.838
$PFSNet_{21}$	.896	.036	.902	.892	.943	.026	.956	.924	.875	.063	.856	.854
VST_{21}	.890	.037	.891	.896	.942	.029	.952	.928	.875	.060	.837	.865
EDN_{22}	.893	.035	.904	.891	.939	.027	.948	.922	.879	.062	.857	.855
$SHNet_{22}$	.883	.030	.938	.908	.926	.025	.959	.926	.855	.056	.910	.884
$SRfor_{22}$	.905	.029	.919	.904	.943	.025	.956	.928	.890	.052	.870	.873
M^3Net_{23}	.909	.026	.921	.908	.947	**.024**	.958	.931	.897	.050	.878	.879
UTD_{24}	.904	.043	.926	.900	.933	.028	.926	.921	.854	.063	.845	.851
Ours	**.918**	**.025**	**.967**	**.919**	**.949**	.025	**.961**	**.937**	**.915**	**.045**	**.919**	**.896**

Table 1: Quantitative comparisons between DiMSOD and other methods on three RGB SOD benchmark datasets.

Methods	DUTD				LFSD				NJUD			
	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$
$PGANet_{20}$	.048	.889	.894	.898	.071	.868	.865	.874	.035	.927	.925	.903
$DANet_{20}$	.029	.936	.931	.933	.074	.854	.858	.829	.037	.914	.915	.917
HDF_{20}	.025	.944	.933	.893	.066	.870	.866	.872	.030	.926	.919	.897
$MMNet^*_{21}$	.032	.918	.920	.951	.057	.875	.878	.913	.038	.899	.910	.922
$BBSNet_{21}$	.029	.934	.930	.916	.061	.879	.877	.857	.036	.918	.918	.912
CMW_{21}	.036	.915	.905	.903	.087	.832	.827	.831	.037	.911	.910	.915
$DDCNN_{23}$	.024	.947	.941	.923	.054	.892	**.890**	.898	.035	.919	.922	.915
$PICR_{23}$	.029	.938	.932	.924	.068	.883	.875	.848	.035	.928	.925	.881
$LAFB^*_{24}$	.027	.930	.928	.957	.072	.859	.853	.864	.033	.906	.904	.918
$VSCode^*_{24}$	.034	.959	.952	**.974**	.072	.862	.859	.869	.038	.945	.941	.967
OURS	**.023**	**.967**	**.957**	.951	**.050**	**.914**	.888	**.917**	**.028**	**.947**	**.947**	**.969**

Table 2: Quantitative comparisons between DIMSOD and other methods on three RGB-D SOD benchmark datasets.

Method	VT821				VT1000				VT5000			
	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$M\downarrow$
$R3Net_{18}$	.786	.809	.660	.073	.842	.859	.761	.055	.757	.790	.615	.083
$M3S_{19}$	.723	.859	.734	.140	.726	.827	.717	.145	.652	.780	.575	.168
$SGDL_{20}$	.765	.847	.731	.085	.787	.856	.764	.090	.750	.824	.672	.089
$MMNet^*_{21}$	.874	.894	.799	.040	.905	.926	.874	.032	.850	.896	.796	.045
ADF_{22}	.810	.842	.717	.077	.910	.921	.847	.034	.864	.891	.778	.048
$DCNet_{22}$	.877	.913	.822	.033	.923	.949	.902	.021	.872	.921	.819	.035
$LSNet_{23}$	.877	.911	.827	.033	.924	.936	.887	.022	.876	.916	.827	.036
$LAFB^*_{24}$	.904	.915	.843	.034	.841	.945	.905	**.018**	.896	.931	.857	.030
$VSCode^*_{24}$	.921	**.951**	.906	.029	.949	**.981**	**.944**	.024	.918	.954	.892	.033
Ours	**.923**	.949	**.917**	**.025**	**.953**	.955	.935	.020	**.921**	**.959**	**.898**	**.029**

Table 3: Quantitative comparisons between DiMSOD and other methods on three RGB-T SOD benchmark datasets.

competing methods, the average improvements of our results across the three datasets in terms of the S_α, E_ξ, and F_β metrics are 10.4%, 7.3%, and 15.3%, respectively. Our approach performs better on RGB and RGB-D datasets compared to RGB-T, indicating a high level of consistency between salient objects and depth information in many cases. In summary, we can conclude that by leveraging the proposed framework, DiMSOD demonstrates a competitive advantage across RGB, RGB-D, and RGB-T datasets.

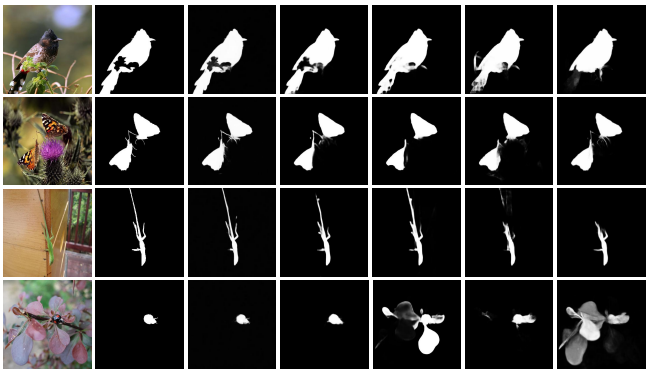

Figure 3: Detailed visual comparison of saliency map results generated by various advanced methods for RGB SOD.

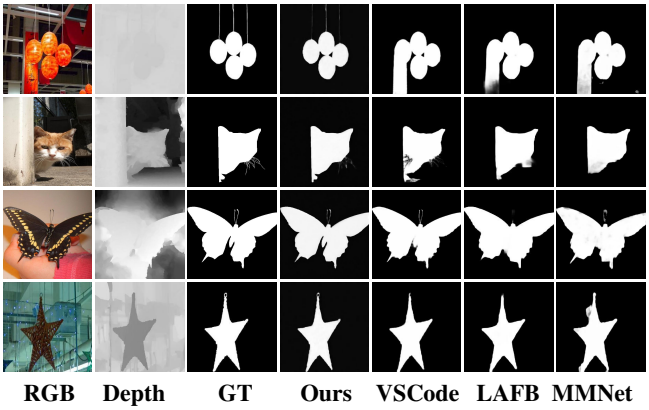

Figure 4: Detailed visual comparison of saliency map results generated by various advanced methods for RGB-D SOD.

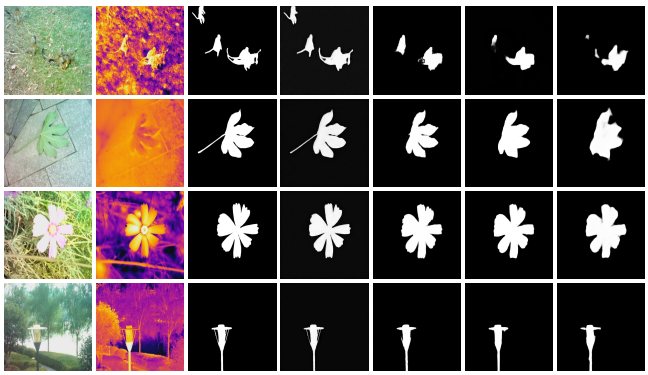

Figure 5: Detailed visual comparison of saliency map results generated by various advanced methods for RGB-T SOD.

Qualitative Evaluation. Fig. 3, Fig. 4, and Fig. 5 show the comprehensive visual comparison of many challenging samples, including complex backgrounds, rich edge details, small objects, and multiple salient objects. From the results,

we can find that compared with other methods, our method exhibits good structural completeness and has more intricate details. Besides, previous models have muddled the identification of edge components, even when accurately pinpointing the object's location. Nevertheless, DiMSOD captures intricate object textures effectively in an incredibly detailed way, addressing the segmentation mask blurring issue presented in other methods. More detailed results can be seen in the last column of Fig. 6, where DiMSOD shows excellent handling of the texture and edge details of the targets.

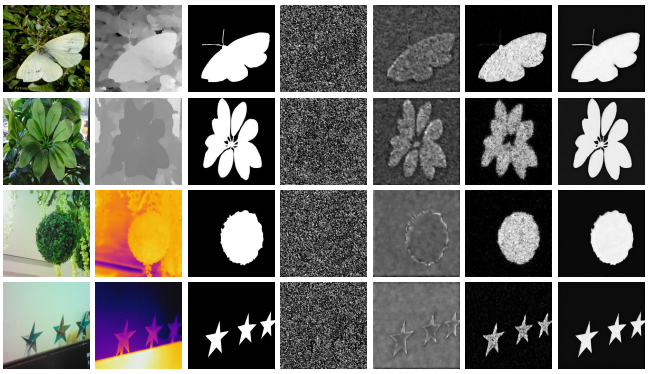

Figure 6: Visual results of the DiMSOD sampling process.

Ablation Studies

We conduct ablation experiments on key components of DiMSOD to validate their effectiveness and scrutinize their influence on model performance, As depicted in Table 4, experimental results demonstrate that PVT with multi-scale attention feature fusion network, SOD-ControlNet (SOD-CNet), Feature Injection Attention Network (FIAN) can all improve multi-model SOD performance very well. Combining them with Stable Diffusion (SD) led to significant improvements across all evaluation metrics in all datasets.

Effectiveness of PVT backbone with multi-scale attention feature fusion. From Table 4, SD(VAE)+CNet represents the use of Stable Diffusion's native VAE as a global feature extractor for RGB images, which is then directly combined with ControlNet to extract local features (e.g., Depth and Thermal). No.2 achieves an average improvement of 5.33% and 17.39% over No.1, which only uses VAE, in terms of F_β and MAE, respectively, across three types of datasets. Fig. 6 illustrates how the critical clues identified by PVT are seamlessly incorporated into the diffusion process through the assistance of mutil-scale attention FF and FIAN.

Effectiveness of SOD-ControlNet. The purpose of SOD-ControlNet is to address the issue of cross-modal information fusion in SOD. As shown in Table 4, compared to No.2, No.3 has an average improvement of 3% and 29.82% for F_β and MAE on the three types of datasets, respectively.

Effectiveness of Feature Injection Attention Network. From Table 4, wo can find that FIAN plays a key role in improving the performance of the model. The average improvement of DiMSOD with FIAN over No.3 without FIAN for F_β and MAE on the three types of datasets is 1.2% and 8.75%, respectively. It indicates that FIAN effectively integrates both diffusion features and salient features from the trainable multi-scale attention feature fusion network.

The core of DiMSOD lies in leveraging SOD-ControlNet to effectively integrate the rich visual priors stored in Stable Diffusion and cross-modal auxiliary information. This enhanced integration facilitates more accurate guidance in the generation of saliency masks. Despite being trained on relatively coarse SOD benchmark datasets, our model effectively segments the edges of salient objects, thanks to its robust visual priors. The final results of DiMSOD exhibit even greater precision and refinement compared to the GT masks. Moreover, the incredible generalization capability of DiMSOD also benefits from the visual priors provided by SD.

No.	Settings	DUTS-TE		NJUD		VT1000	
		$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$
1	SD(VAE)+CNet	.851	.045	.844	.049	.856	.044
2	SD+PVT+CNet	.890	.037	.924	.038	.873	.039
3	SD+PVT+SOD-CNet	.915	.026	.935	.030	.917	.024
Ours	SD+PVT+SOD-CNet+FIAN	**.918**	**.025**	**.947**	**.028**	**.935**	**.020**

Table 4: Ablation studies to evaluate the core components of DiMSOD. The best results are highlighted in boldface.

Limitations and Future Work

Despite the consistently strong performance of DiMSOD across various metrics in multi-model SOD, its inference time does not show a notable advantage over other methods. This inherent limitation of diffusion models in generative tasks is mitigated by our inference code, which optimizes the trade-off between inference time and detection accuracy, achieving reasonable inference time with only a slight and acceptable reduction in accuracy. However, accelerating inference time through engineering optimization is merely a temporary solution. The fundamental approach to achieving true acceleration involves optimizing the diffusion model theory itself. This represents both a future research area for us and a prevalent challenge in image generation.

Conclusion

In this paper, we presented DiMSOD, a diffusion-based framework for RGB, RGB-D and RGB-T images. To the best of our knowledge, this is the first framework to apply a denoising diffusion model to multi-modal SOD. DiMSOD decomposes multi-modal SOD into a series of forward and reverse diffusion processes, leveraging key details from the semantic features under both global condition (original image) and local condition (depth map or thermal map) to guide the processes. Extensive quantitative and qualitative experiments demonstrate that DiMSOD outperforms other state-of-the-art methods across various benchmark datasets. Additionally, ablation studies confirm the effectiveness of the SOD-ControlNet and FIAN we introduced for multi-modal SOD. While our current model offers a strategy to balance accuracy and inference time, this is not a long-term solution. In the future, we will conduct additional research and refinement to improve the model's inference efficiency.

Acknowledgments

This work was supported by the National Key Research and Development Project under Grant No. 2022YFC3302603. Additionally, this paper received funding from the Special Fund for International Conferences for Graduate Students at East China Normal University.

References

Cai, X.; Wang, G.; Lou, J.; Jian, M.; Dong, J.; Chen, R.-C.; Stevens, B.; and Yu, H. 2024. Perceptual loss guided Generative adversarial network for saliency detection. *Information Sciences*, 654: 119625.

Cao, H.; Tan, C.; Gao, Z.; Xu, Y.; Chen, G.; Heng, P.-A.; and Li, S. Z. 2024. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*.

Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; and Shi, Z. 2024. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*.

Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; and Han, J. 2023. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Duan, Y.; Guo, X.; and Zhu, Z. 2025. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In *European Conference on Computer Vision*, 432–449. Springer.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, 4548–4557.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023a. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10021–10030.

Gao, W.; Fan, S.; Li, G.; and Lin, W. 2023b. A Thorough Benchmark and a New Model for Light Field Saliency Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gao, W.; Liao, G.; Ma, S.; Li, G.; Liang, Y.; and Lin, W. 2021. Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 2091–2106.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10696–10706.

Gu, Y.; Xu, H.; Quan, Y.; Chen, W.; and Zheng, J. 2023. Orsi salient object detection via bidimensional attention and full-stage semantic guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Ho, J.; Jain, A.; and Abbeel, P. 2020a. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Jain, A.; and Abbeel, P. 2020b. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Huo, F.; Liu, Z.; Guo, J.; Xu, W.; and Guo, S. 2024. UTDNet: A unified triplet decoder network for multimodal salient object detection. *Neural Networks*, 170: 521–534.

Ji, Y.; Chen, Z.; Xie, E.; Hong, L.; Liu, X.; Liu, Z.; Lu, T.; Li, Z.; and Luo, P. 2023. Ddp: Diffusion model for dense visual prediction. *arXiv preprint arXiv:2303.17559*.

Jian, M.; and Yu, H. 2023. Towards reliable object representation via sparse directional patches and spatial center cues. *Fundamental Research*.

Ju, R.; Ge, L.; Geng, W.; Ren, T.; and Wu, G. 2014. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE international conference on image processing (ICIP)*, 1115–1119. IEEE.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2023. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145*.

Konwer, A.; Hu, X.; Bae, J.; Xu, X.; Chen, C.; and Prasanna, P. 2023. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21415–21425.

Lee, Y.-L.; Tsai, Y.-H.; Chiu, W.-C.; and Lee, C.-Y. 2023. Multimodal Prompting with Missing Modalities for Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14943–14952.

Li, G.; Liu, Z.; Ye, L.; Wang, Y.; and Ling, H. 2020. Cross-modal weighting network for RGB-D salient object detection. *ECCV*.

Li, J.; Ji, W.; Wang, S.; Li, W.; et al. 2024. DVSOD: RGB-D Video Salient Object Detection. *Advances in Neural Information Processing Systems*, 36.

Li, J.; Qiao, S.; Zhao, Z.; Xie, C.; Chen, X.; and Xia, C. 2023. Rethinking lightweight salient object detection via network depth-width tradeoff. *IEEE Transactions on Image Processing*.

Liu, N.; Zhang, N.; Wan, K.; Shao, L.; and Han, J. 2021. Visual saliency transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4722–4732.

Luo, Z.; Liu, N.; Zhao, W.; Yang, X.; Zhang, D.; Fan, D.-P.; Khan, F.; and Han, J. 2024. VSCode: General Visual Salient and Camouflaged Object Detection with 2D Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17169–17180.

Ma, M.; Xia, C.; and Li, J. 2021. Pyramidal feature shrinking for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2311–2318.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-scale interactive network for salient object detection. In *CVPR*, 9413–9422.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2023. CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing*, 32: 892–904.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346.

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

Piao, Y.; Ji, W.; Li, J.; Zhang, M.; and Lu, H. 2019. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7254–7263.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tu, Z.; Li, Z.; Li, C.; and Tang, J. 2022a. Weakly alignment-free RGBT salient object detection with deep correlation network. *IEEE Transactions on Image Processing*, 31: 3752–3764.

Tu, Z.; Ma, Y.; Li, Z.; Li, C.; Xu, J.; and Liu, Y. 2022b. RGBT salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*.

Ulyanov, D. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Wang, K.; Tu, Z.; Li, C.; Zhang, C.; and Luo, B. 2024. Learning Adaptive Fusion Bank for Multi-modal Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 136–145.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wang, X.; Zhu, L.; Tang, S.; Fu, H.; Li, P.; Wu, F.; Yang, Y.; and Zhuang, Y. 2022. Boosting RGB-D saliency detection by leveraging unlabeled RGB images. *IEEE Transactions on Image Processing*, 31: 1107–1119.

Wu, Y.-H.; Liu, Y.; Zhang, L.; Cheng, M.-M.; and Ren, B. 2022. EDN: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, 31: 3125–3136.

Xia, C.; Sun, Y.; Li, K.-C.; Ge, B.; Zhang, H.; Jiang, B.; and Zhang, J. 2024. RCNet: Related Context-Driven Network with Hierarchical Attention for Salient Object Detection. *Expert Systems with Applications*, 237: 121441.

Xu, X.; Guo, J.; Wang, Z.; Huang, G.; Essa, I.; and Shi, H. 2024. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8682–8692.

Yuan, Y.; Gao, P.; and Tan, X. 2023. M^3Net: Multilevel, Mixed and Multistage Attention Network for Salient Object Detection. *arXiv preprint arXiv:2309.08365*.

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

Zhai, G.; and Min, X. 2020. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63: 1–52.

Zhang, G.; Ji, J.; Zhang, Y.; Yu, M.; Jaakkola, T. S.; and Chang, S. 2023. Towards coherent image inpainting using denoising diffusion implicit models.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, W.; Zheng, L.; Wang, H.; Wu, X.; and Li, X. 2022. Saliency hierarchy modeling via generative kernels for salient object detection. In *European Conference on Computer Vision*, 570–587. Springer.

Zhou, H.; Qiao, B.; Yang, L.; Lai, J.; and Xie, X. 2023a. Texture-Guided Saliency Distilling for Unsupervised Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7257–7267.

Zhou, W.; Zhu, Y.; Lei, J.; Yang, R.; and Yu, L. 2023b. LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images. *IEEE Transactions on Image Processing*, 32: 1329–1340.

Zong, M.; Wang, R.; Ma, Y.; and Ji, W. 2023. Spatial and temporal saliency based four-stream network with multi-task learning for action recognition. *Applied Soft Computing*, 132: 109884.