SimpleDiffusion: A Lightweight and Efficient Conditional Diffusion Model for Multi-Modal Salient Object Detection

Shuo Zhang[1]*, Jiaming Huang[3], Wenbing Tang[4], Jing Liu[2], LI HAN[2], Jiandun Li[1], Hongchun Yuan[5], Zizhu Fan[6]

[1]School of Electronic Information Engineering, Shanghai Dianji University
[2]Shanghai Key Laboratory of Trustworthy Computing, East China Normal University
[3]Technology Center, Huolala
[4]College of Information Engineering, Northwest A&F University
[5]College of Information Technology, Shanghai Ocean University
[6]College of Computer Science and Technology, Shanghai University of Electric Power

Abstract

Multi-modal salient object detection (MSOD), which integrates complementary modalities such as depth or thermal data, primarily faces two challenges: accurately preserving salient object details and effectively aligning cross-modal features. Recent advances in using Stable Diffusion to generate images with fine edge details have inspired researchers to reformulate MSOD as a conditional mask generation process guided by salient features, which has achieved excellent visual results. However, these approaches often overlook the high computational cost and large-scale architecture of Stable Diffusion, both of which render it unsuitable for real-world MSOD applications. Therefore, we propose SimpleDiffusion, the first lightweight and efficient conditional diffusion model for MSOD that does not rely on Stable Diffusion. Specifically, we propose an Adaptive Cross-Modal Fusion Conditional Network and a Latent Denoising Network to reduce the complexity of diffusion models. Furthermore, we design a Multi-modal Feature Rectification and Fusion Module to enhance the representational capacity of cross-modal salient features. Customized training and sampling strategies are also developed to improve inference efficiency and reduce erroneous object segmentations. Experiments on multiple MSOD datasets demonstrate that SimpleDiffusion reduces model size by over tenfold and improves inference speed by more than fivefold compared to other diffusion-based methods, while maintaining comparable or superior performance.

Introduction

The human visual system exhibits remarkable efficiency in identifying salient objects within complex environments, owing to its sophisticated mechanisms of visual perception and cognition(Hao et al. 2025). Salient Object Detection (SOD) aims to replicate this unique capability by pinpointing visually distinctive regions in images and videos(Gao et al. 2024). By incorporating complementary modalities such as depth or thermal data, Multi-modal SOD (MSOD) enhances geometric spatial representations and effectively addresses challenging scenarios that are difficult for traditional SOD methods(Yuan, Song, and Li 2025). With recent advancements in depth-sensing and thermal monitoring technologies, MSOD has achieved impressive performance gains in areas such as medical imaging, robotic vision, and surveillance systems(Chen et al. 2024b; Li et al. 2024; Liu et al. 2025).

Previous MSOD methods, designed to distinguish salient objects from complex backgrounds and integrate complementary multimodal data, can be broadly divided into three categories (Wang et al. 2024): 1) *Multi-stream frameworks* (Liao et al. 2022), which utilize multiple input sub-networks to directly capture specific complementary representations. However, these approaches typically involve a complex multi-stage feature fusion process and often fail to effectively integrate auxiliary information from both modalities, as they do not fully account for cross-modal distributional discrepancies; 2) *Bottom-up (top-down) frameworks* (Liu et al. 2021b), which combine low-level detailed features with high-level semantic features for foreground-background segmentation. Nevertheless, these methods often struggle to capture the distinct characteristics of different feature levels, thereby limiting the exploration of the complementarity and diversity between high-level and low-level features(Zhang et al. 2025a). Moreover, the direct integration of multi-level features may compromise the preservation of their original information; 3) *Branched frameworks* (Pei et al. 2024), which adopt a dual-branch cross-modal feature calibration architecture comprising kernel and mask branches for auxiliary information integration. However, negative feedback introduced by cross-modal feature mapping may cause the generated kernels to over-segment salient targets(Song et al. 2023).

The root cause of these issues is that most existing MSOD methods are based on the discriminative paradigm of conventional semantic segmentation, which uses a learning-based encoder for salient feature extraction and a decoder to output corresponding object masks. Without advanced multi-modal fusion and refined post-processing, this paradigm often results in overconfident yet inaccurate predictions, especially in complex or low-quality scenarios(Hu et al. 2024). Therefore, the latest studies have proposed solutions based on the generative paradigm of Stable Diffusion (Rombach et al. 2022), aiming to address the challenges of imprecise edge prediction and incomplete cross-modal feature fusion in MSOD.

*Corresponding Author.

Stable Diffusion comprises two autoencoders, a UNet-based denoising network, a CLIP text encoder, and a noise scheduler. One autoencoder projects images into a latent space for noising, the other reconstructs images from this space. The UNet predicts and removes noise during the reverse diffusion process, while the CLIP text encoder provides semantic guidance. The DDPM/DDIM scheduler modulates noise levels at each step to maintain generation stability and coherence. Trained on the 2-petabyte LAION-5B dataset, Stable Diffusion incorporates rich visual priors and contains over 2 GB of weights. The diffusion-based MSOD approaches typically adapt Stable Diffusion by: (1) replacing the CLIP text encoder with a customized transformer for saliency feature extraction to effectively guide salient mask generation (e.g., SOD-Diffusion(Zhang et al. 2024)); (2) improving attention modules in the denoising network and fusing modalities via the improved ResNets to accelerate the mask generation process (e.g., diffSOD(Zhang et al. 2025a)); or (3) integrating ControlNet for multi-modal fusion to reduce the computational cost associated with fine-tuning (e.g., DiMSOD(Zhang et al. 2025b)). These methods, all derived by fine-tuning Stable Diffusion, owe their success to the built-in iterative denoising mechanism and the image priors embedded in the pre-trained model. The iterative process enables a gradual separation of subtle boundary details from surrounding contexts, while the rich image prior knowledge further fosters a deeper semantic understanding. Moreover, the stochastic sampling process allows for diverse predictions and uncertainty quantification, thereby mitigating overconfident errors in model outputs. However, integrating Stable Diffusion for MSOD inevitably increases inference time, and its large parameter size and high computational demands during inference significantly limit its practicality in real-world applications.

In this work, we aim to develop a lightweight and efficient conditional diffusion model for MSOD by leveraging its inherent iterative denoising mechanism and flexible conditional modeling to learn multimodal salient visual features, rather than relying heavily on the visual priors of Stable Diffusion, thereby enhancing the real-world applicability of diffusion-based MSOD methods. Additionally, we seek to simultaneously address several inherent limitations of diffusion models for MSOD, including restricted discriminative capability, insufficient feature fusion, suboptimal mask refinement, time-consuming inference, and an unstable training process(Cao et al. 2024). To this end, we propose a tailored solution named SimpleDiffusion. Specifically, we propose an Adaptive Cross-Modal Fusion Conditional Network (ACFCN) with Feature Rectification and Fusion Module (FRFM) to strengthen the conditional expressive capability by incorporating a cross-modal guiding feature, which serves as a critical component for segmenting salient objects with detailed boundaries. Furthermore, we design customized learning schedules and output strategies for both training and sampling phases: SNR-driven Variance Adaptation (SVA), Sparse Structure Reorganization (SSR), and Weighted Confidence Ensemble (WCE). The aforementioned components and strategies, when appropriately integrated with our proposed Latent Denoising Network (LDN), can fully leverage the advantages of diffusion models while significantly enhancing computational efficiency and reducing parameter space requirements.

Related Work

Multi-Modal Salient Object Detection

Previous CNN-based approaches have made some progress in MSOD by incorporating sophisticated multi-scale cross-modal fusion mechanisms and multi-task learning(Fang et al. 2023; Huo et al. 2021; Wu et al. 2023). Recent Transformer-based methods (Tang et al. 2022; Pang et al. 2023; Lv et al. 2024; Chen et al. 2024b) have achieved reasonable performance in MSOD. This is mainly attributed to their innovative self-attention mechanisms, which effectively model long-range dependencies and simultaneously integrate multi-scale and multimodal features. By providing global contextual information of salient regions, these methods strive to capture object boundaries and semantic cues. Nevertheless, constrained by the conventional discriminative paradigm, these MSOD methods still face challenges such as inaccurate salient object localization and overconfident erroneous boundary segmentation(Chen et al. 2024a). Moreover, they rely heavily on intricate fusion strategies and sophisticated refinement modules, which also constrain their efficiency and scalability. The latest advancements in MSOD are primarily built upon Stable Diffusion, such as SOD-Diffusion(Zhang et al. 2024), which transforms Stable Diffusion into a saliency mask generator by modifying it and employing a Swin Transformer to extract salient features that guide the UNet denoising process. Similarly, diffSOD(Zhang et al. 2025a) leverages Stable Diffusion as its backbone while incorporating auxiliary multimodal information through a tailored attention-based feature fusion network. This design enables diffSOD to generate saliency masks for both RGB-D and RGB-T data. Moreover, DiMSOD(Zhang et al. 2025b) employs an enhanced ControlNet architecture to incorporate supplementary modal information, leveraging the visual priors of Stable Diffusion with minimal fine-tuning overhead. However, Stable Diffusion-based methods are primarily limited by their large model sizes and prolonged inference times, confining them to exploratory solutions within the generative MSOD research paradigm rather than practical applications.

Diffusion Models for Image Segmentation

Despite the groundbreaking progress diffusion models have achieved in image generation (Rombach et al. 2022; Zhang, Rao, and Agrawala 2023), their suitability for discriminative tasks remains largely underexplored. Only a limited number of studies have explored diffusion models for image segmentation tasks at a preliminary stage (Baranchuk et al. 2021; Brempong 2022; Chen et al. 2023b; Kim, Oh, and Ye 2023). Such segmentation tasks are approached in a manner akin to image-to-image translation. DiffusionDet (Chen et al. 2023a) pioneers the extension of the diffusion process to the generation of detection box proposals. Seg-diffusion (Zhang et al. 2025c) also leverages the rich visual and linguistic priors of Stable Diffusion, and combines them with a content attention module and text embeddings to perform open-vocabulary semantic segmentation, thereby enhancing the model's generalization ability. diffCOD (Chen et al. 2023c) proposes a

diffusion model-based framework for camouflaged object detection, which formulates the segmentation task as a denoising diffusion process from noisy masks to object masks. By integrating image priors and an Injection Attention Module (IAM), the framework enhances the model's denoising learning capability and detection performance. Distinct from previous works, we propose utilizing the diffusion model to denoise multi-modal inputs through a conditioned saliency refinement process, rather than using it as a traditional generator. To the best of our knowledge, this is the first work to introduce a lightweight and efficient conditional diffusion model for multi-modal salient object detection.

Proposed Method

Task Reformulation

Diffusion models, as a class of latent variable models, have been widely adopted in image generation tasks (Croitoru et al. 2023; Cao et al. 2024). These models are trained to reconstruct images corrupted by Gaussian noise through learning the reverse diffusion process, thereby generating high-fidelity outputs. Our proposed SimpleDiffusion is built on diffusion models, consisting of a forward process that gradually adds noise to a salient mask and a reverse process that reconstructs the target distribution from the noise mask. Given a training mask sample $\mathbf{I}$, we first encode it into latent space to obtain features $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. Then, the latent noisy samples $\{\mathbf{x}_t\}_{t=1}^{T}$ are generated via the following Markov process:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathcal{I}), \tag{1}$$

where $t \in \{0, 1, ..., T\}$, and the noise variance schedule $\beta_t \in (0, 1)$. The latent diffusion process of $\mathbf{x}_t$ is:

$$q(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathcal{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_t = \prod_{i=1}^{T}(1-\beta_t)$. Beginning with $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathcal{I})$, the reverse process uses a trained neural network f_θ to incrementally denoise and recover the clean latent mask. The reverse distribution can be defined as :

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathcal{I}). \tag{3}$$

Although directly modeling the latent variable $\mathbf{x}_{t-1}$ is feasible, (Ho, Jain, and Abbeel 2020) demonstrated that maintaining a consistent output space for the network significantly enhances performance. Accordingly, we adopt a hybrid strategy that integrates both approaches in a complementary manner. In SimpleDiffusion, we train a network $f_\theta(\mathbf{x}_t, \mathcal{C}, t)$ to estimate the denoised latent mask $\hat{\mathbf{x}}_0$ conditional on cross-modal image feature $\mathcal{C}$. Here, $\mathcal{C}$ denotes the result of multi-scale emphasis aggregation, integrating the RGB image with its corresponding depth or thermal map. In practice, to accelerate the denoising process, we adopt the enhanced inference technique proposed in DDIM (Song and Ermon 2020), which sets $\sigma_t^2\mathcal{I}$ to zero, thereby yielding deterministic predictions. Therefore, $\mu\theta(\mathbf{x}_t, \mathcal{C}, t)$ can be defined as:

$$\mu_\theta(\mathbf{x}_t, \mathcal{C}, t) = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0}{\sqrt{1-\bar{\alpha}_t}}, \tag{4}$$

Initially, the predicted latent variable $\hat{\mathbf{x}}_0$ is obtained by the trained network $f_\theta(\mathbf{x}_t, \mathcal{C}, t)$. Subsequently, the noisy latent variable $\mathbf{x}_t$ is refined to $\mathbf{x}_{t-1}$ based on Eq.(3) and Eq.(4). Notably, the primary trainable network $f_\theta(\mathbf{x}_t, \mathcal{C}, t)$ comprises two components: ACFCN and LDN. All of these processes take place within our customized latent space, which is primarily designed to improve the training and inference efficiency of model, while generating more refined salient object masks. The latent diffusion process $q(\mathbf{x}_t|\mathbf{x}_0)$ and denoising process $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathcal{C})$ are respectively formulated in Eq. (2) and the expanded version of Eq. (3). The model $f_\theta(\mathbf{x}_t, \mathcal{C}, t)$ is trained by minimizing the discrepancy between diffusion results and the latent denoising prediction, defined as:

$$L_{\text{ddim}} = \left\|\boldsymbol{x}_{t-1} - \mu_\theta(\boldsymbol{x}_t, \mathcal{C}, t)\right\|^2, \tag{5}$$

where the diffusion result $\boldsymbol{x}_{t-1}$ can be obtained via the diffusion process outlined in Eq. (2) by sampling at a given timestep t. It essentially leverages the reversed diffusion process to guide and supervise the latent saliency at each refined step. Meanwhile, to ensure that the final saliency mask generated by the model remains highly consistent with the ground truth, we introduce an additional control loss as follows. Specifically, we employ a simple decoder to decode the predicted latent variable $\hat{\mathbf{x}}_0$, generating the predicted mask $\hat{\mathbf{I}}$ at each DDIM denoising step. The loss between the predicted mask $\hat{\mathbf{I}}$ and the ground truth $\mathbf{I}$ is then computed to guide model training. The loss $\ell(\hat{\mathbf{I}}, \mathbf{I})$ is formulated as follows:

$$L_{pixel}(\hat{\mathbf{I}}, \mathbf{I}) = \ell_{IoU}(\hat{\mathbf{I}}, \mathbf{I}) + \ell_{BCE}(\hat{\mathbf{I}}, \mathbf{I}), \tag{6}$$

where ℓ_{IoU} denotes the weighted intersection-over-union (IoU) loss, while ℓ_{BCE} refers to the weighted binary cross-entropy (BCE) loss. Hence, SimpleDiffusion is optimized by minimizing the weighted sum of hybrid losses, defined as:

$$L = \lambda_1 L_{ddim} + \lambda_2 L_{pixel}(\hat{\mathbf{I}}, \mathbf{I}). \tag{7}$$

Network Architecture

We propose SimpleDiffusion, an iterative framework that progressively refines its segmentation predictions by conditioning on adaptively fused features derived from multi-modal salient priors. Unlike traditional discriminative segmentation approaches, SimpleDiffusion utilizes conditional diffusion mechanisms to generate high-quality segmentation results. The superior performance of SimpleDiffusion stems from three critical components: architectural design, training procedure, and sampling strategy. As shown in Figure 1, SimpleDiffusion comprises an Adaptive Cross-Modal Fusion Conditional Network (ACFCN) and a tailored Latent Denoising Network (LDN). The ACFCN extracts cross-modal visual features at multiple scales to preserve both the overall structure and fine details of the salient prior. Drawing on the extracted multi-scale features, it incorporates hierarchical aggregation and heterogeneous interaction (Li et al. 2022) to strengthen feature connectivity across different scales. Additionally, a feature pyramid neck is employed to effectively fuse these features into a unified cross-modal visual condition. The cross-modal visual condition is subsequently injected into the downstream LDN, where guided denoising is performed in the latent space prior to decoding the expected saliency mask. SimpleDiffusion is compatible with most visual backbones capable of extracting multi-scale features.

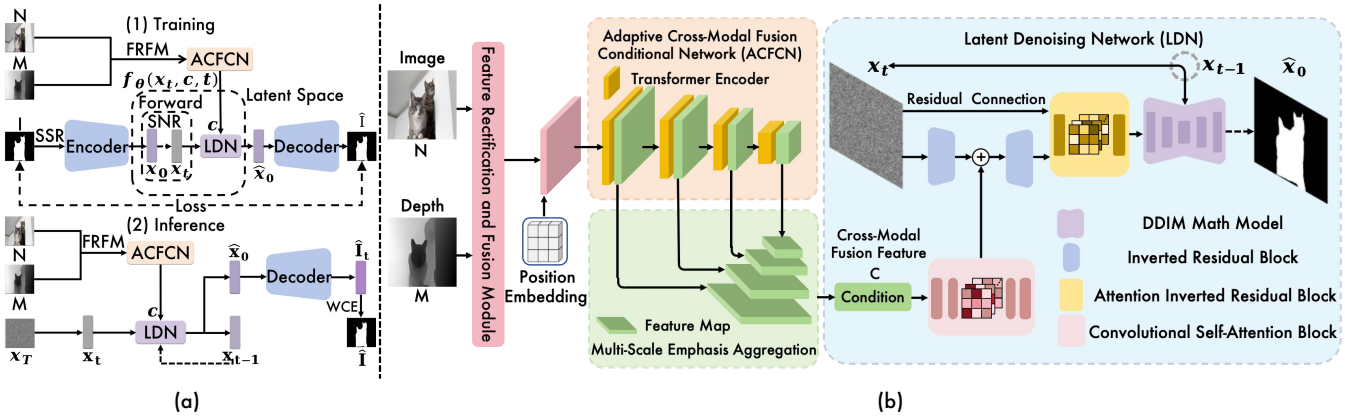

Figure 1: The architecture design of our SimpleDiffusion consists of an Adaptive Cross-Modal Fusion Conditional Network for extracting features as conditions, and a Latent Denoising Network for recovering clear mask predictions from the noisy mask.

Adaptive Cross-Modal Fusion Conditional Network (ACFCN). In SimpleDiffusion, ACFCN is designed to assist the downstream LDN at each iteration in accurately identifying the multi-modal salient regions, thereby isolating the desired salient target. We identify three key challenges in designing a lightweight and efficient conditional network: (1) extracting more distinctive features from salient image priors; (2) mitigating the adverse effects of low-quality cross-modal auxiliary information on feature fusion; and (3) reducing model complexity while accelerating inference. To address these challenges, we introduce ACFCN, which dynamically and comprehensively extracts cross-modal image features.

As illustrated in Figure 1, the ACFCN employs Pyramid Vision Transformer (PVT)(Wang et al. 2022) layers to hierarchically extract multi-scale feature representations from both the input image $\mathcal{N}$ and its corresponding depth or thermal map $\mathcal{M}$. The PVT layer, comprising a Position Embedding module and a Transformer Encoder, is well recognized for its robust feature extraction capabilities. Leveraging multi-scale emphasis aggregation(Zhang et al. 2023), it facilitates comprehensive integration of salient image features, progressively refining representations from coarse to fine granularity at each stage. Furthermore, whereas previous studies (Amit et al. 2021; Chen et al. 2022) relied on a conditional network that exclusively took images as input and reused the same image features throughout the reverse process, we argue that the features required for denoising differ across various stages. When the Signal-to-Noise Ratio (SNR) of the noised mask is low, the conditional network should prioritize focusing on the global structure of the image, but as the mask's contour becomes clear, its attention should transition to finer details. Simultaneously, addressing the detrimental effects of low-quality cross-modal auxiliary information on the model is crucial. To this end, we propose the Feature Rectification and Fusion Module, designed to assist ACFCN in adaptively incorporating cross-modal auxiliary information $\mathcal{M}$.

Feature Rectification and Fusion Module (FRFM). Structural details are strongly correlated with salient features, yet noise in low-quality depth or thermal maps can significantly hinder accurate saliency detection. To address this with minimal complexity and maximal efficiency, we propose a Feature Rectification and Fusion Module (FRFM) that adaptively weights cross-modal features based on the structural similarity between the RGB images and its corresponding depth or thermal data. FRFM employs a dual-branch architecture to achieve shared cross-modal calibration. Essentially, it exploits depth or thermal cues to enhance multi-scale salient features within both branches, thereby maintaining strong internal consistency during feature fusion. Simultaneously, it suppresses low-quality structural information from auxiliary modalities and facilitates efficient fusion of valuable spatial cues with RGB features. In SimpleDiffusion, the FRFM is integrated into the first Overlapping Embedding of PVT backbone to calibrate and enhance the cross-modal features.

First, the original image and its corresponding cross-modal complementary information are individually processed by separate convolutional layers to generate feature maps n and m. After that, these feature maps are flattened, and their similarity score is computed using cosine similarity(Islam, Zunair, and Mohammed 2024), which guides the suppression of redundant noise in the complementary information and facilitates the selection of informative hierarchical features. Both feature maps are then passed through a spatial attention (SA) mechanism to compute spatial attention weights, from which shared affinity weights are derived. These shared affinity weights are employed to calibrate and fuse the cross-modal features, effectively capturing strong correlations and mutual information between the two modalities. Subsequently, the shared weights inform two successive convolutional layers and a softmax function to generate modality-specific weights. The RGB features and the filtered complementary features are integrated with their respective spatial weights through matrix multiplication and residual connections, thereby enhancing modality-specific spatial attention while preserving the intrinsic characteristics of each modality. Finally, the fused cross-modal features are further refined via additional convolutional layers to produce the final feature output.

Latent Denoising Network (LDN). As stated earlier, we define multi-modal salient object detection as a denoising process $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathcal{C})$, which progressively refines the salient latent representation $\mathbf{x_t}$ and enhances prediction accuracy under the guidance of visual information extracted from cross-modal images. Specifically, this is accomplished using a neural network $f_\theta(\mathbf{x}_t, t, \mathcal{C})$ which takes cross-modal

fused visual condition $\mathcal{C}$ and current salient latent $\mathbf{x}_t$ as inputs to predict the distribution $\mathbf{x}_{t-1}$. The cross-modal visual condition $\mathcal{C}$ is generated via a multi-scale visual feature emphasis aggregation process. As shown in Figure 1 (b), we incorporate and combine several widely used modules to accomplish this process. Given that the MSOD task typically demands low inference time for practical applications, we implement the denoising head in a lightweight design. The visual condition $\mathcal{C}$ is essentially a lower-resolution aggregated feature map that maintains a strong local correlation with the salient latent $\mathbf{x}_t$ to be denoised. Initially, we employ a local projection layer to upsample the condition $\mathcal{C}$ to match the dimensions of the salient latent $\mathbf{x}_t$ while preserving the local relation between features. The projected condition is integrated with the salient latent $\mathbf{x}_t$ through element-wise addition, followed by a CNN block and a self-attention layer. The fused salient latent undergoes processing through a Inverted Residual Block (Sandler et al. 2018) CNN layer and channel-wise attention, incorporating a residual connection. The denoised output $\mathbf{x}_{t-1}$ is derived using the DDIM (Song and Ermon 2020) inference method, based on the predefined diffusion schedule β, α applied to the model's outputs. The input size is $384 \times 384 \times 3$. After FRFM, it reduces to $96 \times 96 \times 64$. Inside ACFCN, the final size is $12 \times 12 \times 512$. Following Aggregation, it becomes $96 \times 96 \times 256$. In LDN, combined with the noisy GT, the size is $192 \times 192 \times 256$, and the generated mask size is $384 \times 384 \times 1$.

Training Strategy

During training, we start a diffusion process from the ground truth mask to the noisy mask and train the model to reverse this procedure. Although training is effective, certain issues persist. First, the model struggles to reconstruct a clear mask from a low-SNR mask solely based on image features. To address this, an SNR-driven Variance Adaptation (SVA) is introduced to enhance the model's stability. The second issue stems from the insufficiency of Gaussian noise added to the latent mask feature during the forward process, which is resolved by employing Sparse Structure Reorganization (SSR) to enable the model to learn structure-aware denoising. These approaches form the basis of the diffusion process.

SNR-driven Variance Adaptation (SVA). Masking induces excessively high Signal-to-Noise Ratio (SNR) during training, hindering the model's ability to reconstruct masks from low-SNR inputs (Chen et al. 2022; Chen, Sun, and Lin 2024). The widely adopted cosine variance schedule, originally developed for low-resolution image generation tasks, is inadequate for effectively injecting noise into binary masks. At high SNR levels, the mask remains clearly discernible, allowing the model to recover it easily without relying on inherent image features. To tackle this issue, we refine an SNR-driven variance adaptation (Hoogeboom, Heek, and Salimans 2023) to better regulate the SNR during training. Specifically, following the SNR definition (Kingma et al. 2021): $\mathrm{SNR}(t) = \bar{\alpha}_t/(1-\bar{\alpha}_t)$, we devise an SNR schedule strategy, where $\log \mathrm{SNR}(t) = -2\log(\tan(\frac{\pi t}{2}))$ to which we apply an tunable offset. Reducing this offset results in fewer simple MSOD cases being reversed at the same time step t.

Sparse Structure Reorganization (SSR). A significant challenge in integrating generative methods into saliency prediction is the issue of sparse ground truth saliency values. This sparsity often causes mode collapse during conventional generative training(Duan, Guo, and Zhu 2024). Additionally, existing diffusion models typically rely on pixel-level corruption to generate noisy masks directly from the ground truth (GT), causing the model to wrongly assume that the reconstructed contour from the noised mask is consistently correct. This flawed assumption, however, does not necessarily hold during sampling. To overcome this limitation, we introduce Sparse Structure Reorganization during forward diffusion. In this approach, we randomly disrupt the GT contour and introduce additive Gaussian noise into the corrupted GT features within latent space. This strategy allows the model to learn how to accurately reconstruct the saliency mask from the noised input, even when bias is present. Furthermore, by leveraging standard random cropping, jittering, and flipping augmentations during training, the model is encouraged to reorganize the entire saliency map rather than merely relying on regressing known regions. This significantly enhances the overall visual quality of the generated saliency results.

Sampling Strategy

Our denoising model performs gradual refinement on a sample $\mathbf{x}_T$ drawn from a standard normal distribution across T steps in latent space. This step-by-step denoising process progressively reduces the discrepancy between the predicted mask and the ground truth, ultimately resulting in a more accurate output. While the final denoised mask displays sharp and distinct boundaries, we argue that intermediate predictions generated throughout the denoising process also carry meaningful information. Nevertheless, not all intermediate predictions contribute equally to the final result, as those closer to the final denoising step have a more significant influence. Therefore, we adopt a Weighted Confidence Ensemble (WCE) strategy aggregating predictions across sampling steps, enhancing precision and robustness of the final output.

Weighted Confidence Ensemble (WCE). Motivated by the annotation process in saliency detection (Fan et al. 2020; Zhang et al. 2021), we propose Weighted Confidence Ensemble (WCE) method, which aggregates predictions across different sampling steps without introducing additional computational overhead. Specifically, for each sampling step at time t, the denoised latent image $\hat{\mathbf{x}}_0$ is decoded into $\mathbf{I}_t$. Given a set of predictions $\{\mathbf{I}_t\}_{t=1}^T$, binary masks $\{\mathbf{I}_t^b\}_{t=1}^T$ are first generated using adaptive thresholding. Subsequently, these binary masks $\{\mathbf{I}_t^b\}_{t=1}^T$ collectively vote on the position of each point to produce a candidate mask, where the probability value at each selected point is computed as the average of all corresponding predictions. Mathematically,

$$P_{emb} = round(\frac{\sum_{t=1}^T w_t \cdot \mathbf{I}_t^b}{\sum_{t=1}^T w_t}) * \frac{\sum_{t=1}^T w_t \cdot \mathbf{I}_t}{\sum_{t=1}^T w_t}. \quad (8)$$

where

$$w_t = \frac{1}{\|\hat{x}_t - \hat{x}_{t-1}\|_2^2 + \epsilon} \quad (9)$$

ϵ is introduced to prevent numerical instability, such as division by zero, and to enhance computational robustness.

Experiments

Experiment Settings

Datasets and Evaluation Metrics. For RGB-D SOD, we use six datasets, including **STERE**, **NJUD** (Ju et al. 2014), **NLPR**, **DUTLF-Depth** (Piao et al. 2019), **SIP**, and **ReDWeb-S** (Liu et al. 2021a). For RGB-T SOD, we adopt three datasets: **VT821** (Wang et al. 2018), **VT1000** (Tu et al. 2019), and **VT5000** (Tu et al. 2022). To ensure consistency in evaluating all methods, we employ seven widely adopted metrics to measure their performance: Parameters, Frames Per Second (FPS), Giga Floating Point Operations Per Second (GFLOPs), structure-measure S_m (Fan and Cheng 2017), maximum enhanced-alignment measure E_m (Fan et al. 2018), maximum F-measure F_m, and Mean Absolute Error M.

Implementation Details. Drawing upon previous studies (Liu et al. 2021c; Tu et al. 2022; Luo et al. 2024), we utilize the training sets of **NLPR**, **NJUD**, and **DUTLF-Depth** for RGB-D SOD, and the training set of **VT5000** for RGB-T SOD. For a fair comparison with prior studies (Lee et al. 2022; Luo et al. 2024), we resize each image to 384×384 pixels during training. Our model is implemented in PyTorch and both trained and evaluated on an NVIDIA A100 GPU. The ACFCN is initialized with PVTv2-B4 and optimized using AdamW with a batch size of 32. The learning rate starts at 0.001 and is scheduled by cosine annealing over 150 epochs. By default, we set $T = 10$ for sampling.

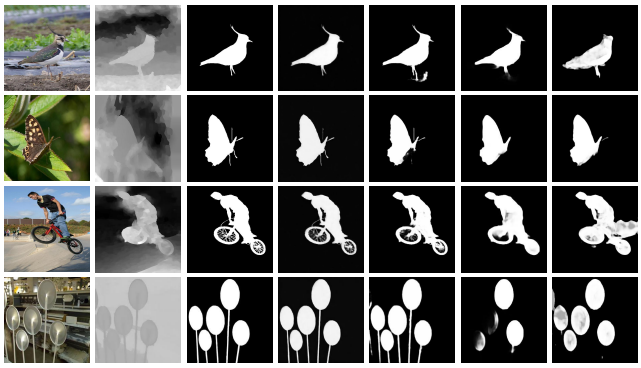

Figure 2: Diffusion-based RGB-D SOD Comparison.

Comparisons with Advanced Methods

Due to space constraints, we present a subset of results in Table 1. SimpleDiffusion clearly outperforms other diffusion-based methods in terms of parameters, FPS, and GFLOPS. It also remains highly competitive against non-diffusion-based methods and achieves state-of-the-art results on other general metrics. As illustrated in Figure 2 and Figure 3, it demonstrates superior performance in challenging scenarios, such as those involving small or multiple objects, by significantly mitigating errors like over-segmentation and misclassification. Therefore, it generates salient object segmentation masks with well-defined boundaries and clear structures. These improvements lead to more reliable predictions across diverse and complex real-world scenes, enhancing overall robustness.

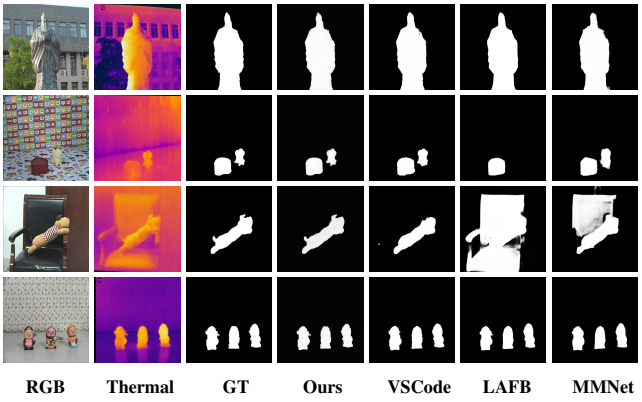

Figure 3: Diffusion-based RGB-T SOD Comparison.

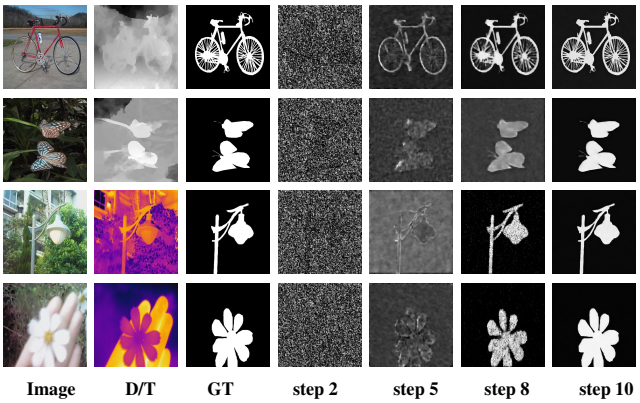

Figure 4: Visual results of the sampling process.

Ablation Study

Ablation of key Model Components. We conduct an ablation study on the individual components of SimpleDiffusion, and the detailed results are shown in Table 2. Compared to using the ResNet backbone, our adopted PVT backbone demonstrates superior performance. Additionally, multi-scale emphasis aggregation (MEA) proves to be more suitable for our model compared to the FPN approach. The FRFM contributes significantly to performance improvement, which also highlights the adverse impact of low-quality depth maps on MSOD. The noise scheduling strategy based on SNR (SVA) and the sparsity reconstruction method (SSR) effectively restructure the GT mask and the signal-to-noise ratio of the noise, enabling the model to capture more discriminative and dynamic features during training, thereby enhancing prediction accuracy. Furthermore, the adoption of the WCE strategy allows for better weighting and effective utilization of our generated sampling results, ultimately leading to improved precision in generating salient object masks.

Ablation of Training Strategies. We assess the impact of SNR-driven Variance Adaptation (SVA) and Sparse Structure Reorganization (SSR) techniques integrated into SimpleDiffusion (see Table 2). These strategies enhance the model's ability to explore and correct features. The results show a clear boost when combining SVA and SSR in the framework.

Method	Params↓ (M)	FPS↑	GFLOPs↓	NJUD			NLPR			DUTLF-Depth			ReDWeb-S			VT821			VT1000		
				$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$
LSNet	**17.2**	**314**	**3.1**	.899	.885	.922	.881	.882	.955	.775	.831	.891	.458	.566	.691	.809	.825	.911	.887	.885	.935
CoLA	184.2	43	281.3	.934	.913	.947	.935	.909	.957	.941	.947	.971	.759	.763	.826	.905	.860	.930	.927	.895	.955
VST-T++	100.5	54	325.1	.928	.929	.958	.933	.921	.964	.944	.948	.969	.756	.757	.819	.894	.861	.923	.941	.931	.972
CAVER	55.8	67	137.6	.920	.924	.953	.929	.921	.964	.931	.939	.962	.730	.724	.802	.891	.874	.933	.936	.927	.970
MMNet	-	-	-	.910	.899	.922	.925	.892	.955	.920	.918	.951	-	-	-	.873	.795	.892	.913	.863	.923
LAFB	398.1	37	552.1	.910	.919	.924	.902	.905	.958	.919	.930	.957	.664	.727	.757	.817	.843	.915	.905	.905	.945
VSCode	54.1	-	484.2	.941	.945	.967	.938	.930	.966	.952	.959	.974	.766	.771	.831	.921	.906	.951	.949	.944	**.981**
SOD-diff	890.2	3	1862.4	.908	.904	.913	.917	.919	.941	.924	.929	.935	.731	.742	.803	.913	.894	.922	.925	.930	.952
diffSOD	860.5	5	1275.3	.929	.931	.937	.917	.921	.938	.922	.928	.935	.741	.752	.812	.914	.889	.927	.921	.926	.947
DiMSOD	984.7	8	784.1	.947	.947	.969	.923	.922	.934	.957	.967	.951	.748	.757	.821	.923	.917	.949	**.953**	.935	.955
Ours	92.4	26	251.8	**.949**	**.951**	**.973**	**.947**	**.935**	**.969**	**.957**	**.963**	**.977**	**.787**	**.779**	**.837**	**.932**	**.928**	**.964**	.945	**.948**	.959

Table 1: Quantitative comparison of our SimpleDiffusion with other SOTA RGB-D/T SOD methods on benchmark datasets.

Settings	STERE		VT5000	
	$F_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$M\downarrow$
(a) Res50+FPN	.803	.097	.789	.118
(b) PVT+FPN	.819	.084	.811	.093
(c) PVT+MEA	835	.071	.829	.081
(d) PVT+MEA+FRFM	884	.053	.857	.058
(e) PVT+MEA+FRFM+SSR	.913	.044	.879	.047
(f) PVT+MEA+FRFM+SVA+SSR	.920	.035	.887	.034
PVT+MEA+FRFM+SVA+SSR+WCE	**.932**	**.031**	**.901**	**.029**

Table 2: Ablation of key components and training strategies.

Ablation of Sampling Strategy. As shown in Table 2, integrating WCE leads to a clear enhancement in our model's performance. To assess its ability to mitigate overconfident mis-segmentation, we calculated the number of misclassified pixels with extreme confidence levels and applied dilation to minimize the influence of insufficient model precision. The findings demonstrate a significant decrease in overconfident incorrect pixels when adopting the WCE sampling strategy.

Analysis. As shown in Figure 4, to further demonstrate the ability of SimpleDiffusion to reduce noise and progressively focus on intricate salient details, we present the prediction results at various sampling stages. Initially, the model generates a coarse mask with high uncertainty in regions where boundaries are blurry. However, as the sampling steps increase, the model gradually focuses on the target objects embedded in the surrounding environment and continuously refines the object mask, establishing clear and precise object boundaries based on the subtle yet critical details of the foreground.

Furthermore, SimpleDiffusion achieves outstanding results in RGB-D salient instance segmentation (SIS). To the best of our knowledge, only a few methods have been specifically developed for SIS to date. Nevertheless, with minimal modifications, SimpleDiffusion can be readily adapted to perform SIS. During training, SimpleDiffusion employs a lightweight encoder and decoder, each constructed by stacking convolutional layers. The encoder extracts GT features for subsequent noising, while the decoder functions as the mask generator. For MSOD, where the input is a binary GT mask, the encoder's input channel is set to 1, with all other parameters remaining unchanged. In contrast, for SIS, which uses instance-level GT masks, the loss is adjusted to a multi-class version, and the encoder's input channel is set to 3. The corresponding decoder is configured similarly to accommodate this change. Specifically, in the decoder's output layer, we modify the dimensionality of the generated mask using a 1×1 convolution. For binary masks, the number of 1×1 kernels is set to 1, for instance-level masks, it is set to 3.

Similar to the MSOD setting, we adapt all baseline methods, including CalibNet(Pei et al. 2024), M2For(Cheng et al. 2022), and RDPNet, to the RGB-D SIS task, and train and evaluate all models on the COME15K dataset using their official implementations. The results clearly demonstrate that SimpleDiffusion achieves more precise and complete boundary segmentation of salient instances. Furthermore, owing to its integration of multiple strategies and efficient modules, SimpleDiffusion also shows competitive inference speed for RGB-D SIS, making it a strong and practical solution.

Conclusion

In this work, we propose SimpleDiffusion, the first diffusion-based method for MSOD that does not rely on Stable Diffusion. SimpleDiffusion is a lightweight and efficient conditional diffusion model primarily composed of ACFCN and LDN modules. It substantially reduces model complexity and greatly improves inference speed compared to previous diffusion-based MSOD approaches. Furthermore, we introduce a dedicated Rectification and Fusion Module along with optimized training and sampling strategies to enhance feature fusion and fine detail generation, enabling more robust mask prediction. Experimental results demonstrate that our method markedly outperforms existing diffusion-based approaches.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC330260, in part by the Fundamental Research Funds for the Central Universities of China, in part by the National Natural Science Foundation of China under Grants 62202168 and 62571310, in part by the Shanghai Educational Science Research General Project under Grant C2025084, and in part by the Shanghai Science and Technology Program under Grant 23010501000.

References

Amit, T.; Nachmani, E.; Shaharbany, T.; and Wolf, L. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Baranchuk, D.; Rubachev, I.; Voynov, A.; Khrulkov, V.; and Babenko, A. 2021. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*.

Brempong, E. A. 2022. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4175–4186.

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, G.; Wang, Q.; Dong, B.; Ma, R.; Liu, N.; Fu, H.; and Xia, Y. 2024a. : Edge-Aware Multimodal Transformer for RGB-D Salient Object Detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Chen, H.; Shen, F.; Ding, D.; Deng, Y.; and Li, C. 2024b. Disentangled cross-modal transformer for RGB-d salient object detection and beyond. *IEEE Transactions on Image Processing*.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023a. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19830–19843.

Chen, T.; Li, L.; Saxena, S.; Hinton, G.; and Fleet, D. J. 2022. A generalist framework for panoptic segmentation of images and videos. *arXiv preprint arXiv:2210.06366*.

Chen, T.; Li, L.; Saxena, S.; Hinton, G.; and Fleet, D. J. 2023b. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, 909–919.

Chen, Z.; Gao, R.; Xiang, T.-Z.; and Lin, F. 2023c. Diffusion model for camouflaged object detection. In *ECAI 2023*, 445–452. IOS Press.

Chen, Z.; Sun, K.; and Lin, X. 2024. CamoDiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1272–1280.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *IEEE CVPR*, 1290–1299.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869.

Duan, Y.; Guo, X.; and Zhu, Z. 2024. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In *European Conference on Computer Vision*, 432–449. Springer.

Fan, D.-P.; and Cheng, M.-M. 2017. Structure-measure: A new way to evaluate foreground maps. 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. 698–704.

Fan, D.-P.; Lin, Z.; Zhang, Z.; Zhu, M.; and Cheng, M.-M. 2020. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems*, 32(5): 2075–2089.

Fang, X.; Jiang, M.; Zhu, J.; Shao, X.; and Wang, H. 2023. M2RNet: Multi-modal and multi-scale refined network for RGB-D salient object detection. *Pattern Recognition*, 135: 109139.

Gao, S.; Zhang, P.; Yan, T.; and Lu, H. 2024. Multi-scale and detail-enhanced segment anything model for salient object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9894–9903.

Hao, C.; Yu, Z.; Liu, X.; Xu, J.; Yue, H.; and Yang, J. 2025. A simple yet effective network based on vision transformer for camouflaged object and salient object detection. *IEEE Transactions on Image Processing*.

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

Hoogeboom, E.; Heek, J.; and Salimans, T. 2023. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*.

Hu, X.; Sun, F.; Sun, J.; Wang, F.; and Li, H. 2024. Cross-modal fusion and progressive decoding network for RGB-D salient object detection. *International Journal of Computer Vision*, 132(8): 3067–3085.

Huo, F.; Zhu, X.; Zhang, L.; Liu, Q.; and Shu, Y. 2021. Efficient context-guided stacked refinement network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5): 3111–3124.

Islam, M.; Zunair, H.; and Mohammed, N. 2024. CosSIF: Cosine similarity-based image filtering to overcome low inter-class variation in synthetic medical image datasets. *Computers in Biology and Medicine*, 172: 108317.

Ju, R.; Ge, L.; Geng, W.; Ren, T.; and Wu, G. 2014. Depth saliency based on anisotropic center-surround difference. 1115–1119.

Kim, B.; Oh, Y.; and Ye, J. C. 2023. Diffusion Adversarial Representation Learning for Self-supervised Vessel Segmentation. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. The International Conference on Learning Representations.

Kingma, D.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational diffusion models. *Advances in neural information processing systems*, 34: 21696–21707.

Lee, M.; Park, C.; Cho, S.; and Lee, S. 2022. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *ECCV*, 630–647. Springer.

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, Z.; Chen, Z.; Liu, X.; and Jiang, J. 2022. DepthFormer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*.

Liao, G.; Gao, W.; Li, G.; Wang, J.; and Kwong, S. 2022. Cross-collaborative fusion-encoder network for robust RGB-thermal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7646–7661.

Liu, N.; Zhang, N.; Shao, L.; and Han, J. 2021a. Learning selective mutual attention and contrast for RGB-D saliency detection. 44(12): 9026–9042.

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

Liu, N.; Zhang, N.; Wan, K.; Shao, L.; and Han, J. 2021c. Visual Saliency Transformer. 4722–4732.

Liu, Y.; Li, C.; Xu, S.; and Han, J. 2025. Part-whole relational fusion towards multi-modal scene understanding. *International Journal of Computer Vision*, 133(7): 4483–4503.

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.

Lv, C.; Zhou, X.; Wan, B.; Wang, S.; Sun, Y.; Zhang, J.; and Yan, C. 2024. Transformer-Based Cross-Modal Integration Network for RGB-T Salient Object Detection. *IEEE Transactions on Consumer Electronics*.

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.

Pei, J.; Jiang, T.; Tang, H.; Liu, N.; Jin, Y.; Fan, D.-P.; and Heng, P.-A. 2024. CalibNet: Dual-branch cross-modal calibration for RGB-D salient instance segmentation. *IEEE Transactions on Image Processing*.

Piao, Y.; Ji, W.; Li, J.; Zhang, M.; and Lu, H. 2019. Depth-induced Multi-scale Recurrent Attention Network for Saliency Detection. 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.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Song, X.; Guo, F.; Zhang, L.; Lu, X.; and Hei, X. 2023. Salient object detection with dual-branch stepwise feature fusion and edge refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4): 2832–2844.

Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33: 12438–12448.

Tang, B.; Liu, Z.; Tan, Y.; and He, Q. 2022. HRTransNet: HRFormer-driven two-modality salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2): 728–742.

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

Tu, Z.; Xia, T.; Li, C.; Wang, X.; Ma, Y.; and Tang, J. 2019. RGB-T image saliency detection via collaborative graph learning. 22(1): 160–173.

Wang, G.; Li, C.; Ma, Y.; Zheng, A.; Tang, J.; and Luo, B. 2018. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *IJIG*, 359–369. Springer.

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, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. 8(3): 415–424.

Wu, Z.; Allibert, G.; Meriaudeau, F.; Ma, C.; and Demonceaux, C. 2023. Hidanet: Rgb-d salient object detection via hierarchical depth awareness. *IEEE Transactions on Image Processing*, 32: 2160–2173.

Yuan, G.; Song, J.; and Li, J. 2025. IF-USOD: Multimodal information fusion interactive feature enhancement architecture for underwater salient object detection. *Information Fusion*, 117: 102806.

Zhang, G.; Luo, Z.; Tian, Z.; Zhang, J.; Zhang, X.; and Lu, S. 2023. Towards efficient use of multi-scale features in transformer-based object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6206–6216.

Zhang, J.; Fan, D.-P.; Dai, Y.; Anwar, S.; Saleh, F.; Aliakbarian, S.; and Barnes, N. 2021. Uncertainty inspired RGB-D saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5761–5779.

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, S.; Huang, J.; Chen, S.; Wu, Y.; Hu, T.; and Liu, J. 2024. SOD-diffusion: Salient Object Detection via Diffusion-Based Image Generators. In *Computer Graphics Forum*, volume 43, e15251. Wiley Online Library.

Zhang, S.; Huang, J.; Tang, W.; Tian, L.; Wei, Y.; and Liu, J. 2025a. Multi-modal Salient Object Detection via a Unified Diffusion Model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhang, S.; Huang, J.; Tang, W.; Wu, Y.; Hu, T.; Xu, X.; and Liu, J. 2025b. DiMSOD: A Diffusion-Based Framework for Multi-Modal Salient Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10103–10111.

Zhang, S.; Huang, J.; Wu, Y.; Hu, T.; Tang, W.; and Liu, J. 2025c. Seg-diffusion: Text-to-Image Diffusion Model for Open-Vocabulary Semantic Segmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.