

Joint Medical Image Enhancement and Segmentation with Diffusion-based Symbiotic Information Interaction

Ying Chen¹, Jinyue Li², Qiankun Li^{3†}

¹Shenzhen Research Institute, The Chinese University of Hong Kong

²University of Science and Technology of China

³Imperial Global Singapore, Imperial College London
q.li2@imperial.ac.uk

Abstract

Image quality is critical for accurate medical diagnosis. However, MRI, CT, and ultrasound images are often of low resolution and quality due to cost constraints, complicating the visualization of key anatomical structures and lesions. While such limitations are common in practice, traditional methods treat image enhancement as a separate pre-processing step, failing to fully leverage its potential synergy with image segmentation. To address this, we propose DiSIINet (Diffusion-based Symbiotic Information Interaction Network), which is built on the principle that enhancement and segmentation should mutually reinforce each other in a unified model. Based on Denoising Diffusion Implicit Models (DDIM), DiSIINet integrates an enhancement branch and a segmentation branch. These branches interact through a novel Symbiotic Information Interaction (SII) module, which facilitates dynamic, feature-level information exchange via cross-attention during the reverse diffusion process. This design enables both tasks to iteratively improve each other. The DDIM backbone ensures high-quality output and efficient inference through deterministic sampling. Experiments on multi-modal medical datasets (MRI, CT, ultrasound) show that DiSIINet achieves significant performance improvements compared to sequential or independent enhancement and segmentation approaches. The code is available at: <https://github.com/Reconsider80/DiSIINet>.

1 Introduction

Medical images acquired under suboptimal conditions may encounter quality issues, including blurriness, poor lighting, low resolution, and noise, which can result in misdiagnosis. While advanced medical image enhancement techniques often face challenges in improving high-resolution image quality while preserving clear local anatomical features. Traditional image enhancement methods, particularly those

based on transformers, typically operate under the assumption of known degradation types, which often leads to overly smoothed results. StillGAN [Ma *et al.*, 2021] introduces an innovative and versatile bi-directional GAN for enhancing medical image quality. Although there have been significant advancements in improving visual perception, methods based on generative adversarial networks (GANs) [Liang *et al.*, 2022; Park *et al.*, 2023] still encounter challenges in balancing perceptual quality with fidelity, often causing artifacts.

Recently, the emergence of diffusion models (DMs) [Ho *et al.*, 2020] has demonstrated remarkable capabilities in approximating intricate distributions and generating realistic images. Stable Diffusion [Rombach *et al.*, 2022] is a conditional diffusion model, whose core concept lies in being a generation system centrally guided by conditional input. Among diffusion-based super-resolution techniques, StableSR [Wang *et al.*, 2023] employs the latent representation of low-quality (LQ) images as the guiding condition for StableDiffusion to achieve super-resolution. In contrast, DiffBIR [Lin *et al.*, 2024] first restores LQ images before using generative priors to maintain a balance between quality and fidelity during the diffusion process. These approaches highlight the significant potential of generative priors in SR tasks; however, relying solely on LQ image data without additional semantic control may lead to inaccurate content reconstruction. Given the inherent advantages of textual prompts in directing generation within pretrained text-to-image models, recent approaches have begun to utilize semantic descriptions extracted from LQ images to enhance control over the generation process and to improve the semantic accuracy of image restoration. For example, PASD [Yang *et al.*, 2024] and SeeSR [Wu *et al.*, 2024b] leverage pretrained captioning or tagging models to derive image content, using textual descriptions as prompts to guide generation. However, these methods may encounter two primary limitations. First, they can occasionally overlook significant components include key objects, textures, or regions that contribute to the overall semantic understanding of the image, resulting in inaccurate semantic details in those areas. Second, a single region may respond strongly to multiple prompts, creating semantic ambiguity in the context of image super-resolution.

In this paper, we aim to address these challenges by utilizing semantic segmentation as an additional control condition. Unlike textual prompts, semantic segmentation provides

[†]Corresponding author[†] (q.li2@imperial.ac.uk).

a more detailed understanding of prominent objects within an image by assigning class labels to individual pixels. However, predicting segmentation masks from heavily degraded images presents significant challenges, and inaccurate masks can lead to erroneous semantic content and distorted spatial structures in the restoration output. Motivated by the fact that image enhancement and segmentation can benefit each other, we propose the joint medical image enhancement and segmentation with diffusion-based Symbiotic Information Interaction (DiSIINet) framework as Fig.1 which introduces a dual diffusion framework to facilitate interaction between the medical image enhancement and segmentation these two closely related tasks.

The key contributions can be summarized as follows:

- We propose DiSIINet, the first leverages a dual-branch diffusion architecture where enhancement and segmentation mutually reinforce each other through continuous information exchange.
- We design an innovative SII module by computing cross-attention between branch features that enables bidirectional information flow between enhancement (DiEnh) and segmentation (DiSeg) branches during reverse diffusion.
- Extensive experiments demonstrate that our method consistently improves performance across diverse mainstream medical imaging modalities (MRI, CT, ultrasound) and multiple datasets.

2 Related Work

2.1 Joint Medical Image Enhancement and Segmentation

Traditionally, research has addressed these image enhancement and segmentation task independently, but recent efforts [Xiao *et al.*, 2024] explore their synergistic integration. Existing joint approaches can be categorized into two main paradigms: segmentation-guided enhancement and enhancement-guided segmentation. Segmentation-guided enhancement methods utilize segmentation information to guide the denoising process while preserving structural integrity. Liu *et al.* [Liu *et al.*, 2018] proposed a deep neural network framework that integrates image denoising with high-level tasks, using a joint loss function to propagate semantic information to low-level visual processing. More recently, the SDABN framework [Xu *et al.*, 2023] employs segmentation and denoising blocks (SDBs) in a cascading manner, where each block’s segmentation probability map refines subsequent denoising. However, these methods typically establish unidirectional information flow, limiting their ability to fully exploit the mutual benefits between tasks. Enhancement-guided segmentation focuses on improving segmentation robustness by incorporating noise resilience during training. DenoiSeg [Buchholz *et al.*, 2020] achieved simultaneous segmentation and denoising within a UNet architecture through a combined self-supervised denoising loss. Recent studies in medical imaging have begun to explore deeper synergies. For instance, Wang *et al.* [Wang *et al.*, 2024b] proposed an

Instance-aware Embedding Module (IEM) to leverage segmentation features for denoising, though they did not utilize denoising features to improve segmentation.

However, There remains a significant gap in methods that facilitate deep feature interaction between two tasks, particularly within the powerful generative framework of diffusion models. Our DiSIINet framework addresses this gap by introducing a symbiotic information interaction mechanism that enables continuous bidirectional information flow between enhancement and segmentation during the diffusion process.

2.2 Diffusion Models for Medical Image Analysis

Diffusion models have shown promise in medical image segmentation by leveraging their generative capabilities. Scaling Vision Transformers [Zhai *et al.*, 2022] employ stochastic sampling to create an implicit ensemble of segmentations, but suffer from convergence issues and require numerous time-intensive iterations. MedSegDiff-V2 [Wu *et al.*, 2024a] introduces dual conditioning mechanisms—Anchor Condition and Semantic Condition—along with SS-Former to integrate segmentation features into the diffusion process. However, these approaches treat segmentation as an isolated task without considering its potential synergy with image enhancement. Medical image enhancement faces unique challenges due to acquisition-related noise, artifacts, and low signal-to-noise ratios (SNR). Diffusion models have been adapted for various denoising applications. For MRI, DDM² [Xiang *et al.*, 2023] incorporates self-supervised denoising into diffusion models. For low-dose CT (LDCT), CoreDiff [Gao *et al.*, 2023] replaces random Gaussian noise with LDCT images as initial states and introduces a mean-preserving degradation operator to reduce sampling steps while preventing structural distortion.

Despite these advancements, when conditioning is employed such as using segmentation to guide enhancement or vice versa, it is often unidirectional failing to establish symbiotic relationships where both tasks mutually benefit. Our work addresses these gaps by proposing DiSIINet, which integrates enhancement and segmentation within a unified DDIM framework with bidirectional information exchange, enabling mutual reinforcement while maintaining efficient sampling.

3 Methodology

3.1 Dual-Diffusion Joint Medical Image Enhancement and Segmentation

This framework includes two branches, the diffusion-based image enhancement branch (DiEnh) and the diffusion-based segmentation branch (DiSeg), which are linked by the Symbiotic Information Interaction (SII) module. Inspired by DDIMs [Song *et al.*, 2021], which is an efficient variant of DDPMs with accelerated sampling and deterministic generation, DiEnh executes both forward and reverse diffusion processes within the latent space utilizing a pre-trained Variational Autoencoder (VAE) [Rombach *et al.*, 2022], specifically the Image Encoder and Image Decoder components. DDIM retains the same forward process as DDPM but intro-

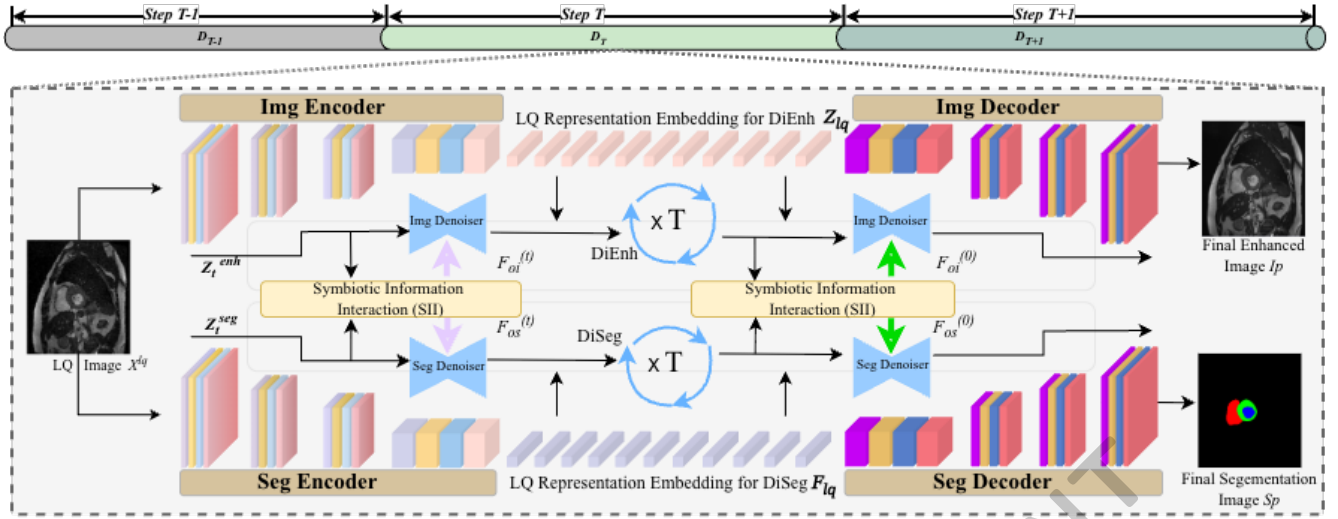


Figure 1: The framework of DiSiNet comprises three key parts: (a)DiEnh branch performs image enhancement diffusion process; (b)DiSeg branch conducts semantic segmentation diffusion process; (c)DiEnh and DiSeg collaborate through SII module to achieving mutual benefits.

duces a non-Markovian reverse process that enables accelerated sampling.

In the DiEnh branch, we leverage the DDIM framework for efficient image enhancement. The branch operates in two phases: training and inference.

During training, we have paired data consisting of low-quality (LQ) images $\mathbf{X}^{lq}$ and their corresponding ground truth (GT) images $\mathbf{I}_{gt}$. We first encode both images into latent space using a VAE encoder:

$$\mathbf{Z}_{lq} = \text{Image Encoder}(\mathbf{X}^{lq}). \quad (1)$$

$$\mathbf{Z}^{gt,enh} = \text{Encoder}(\mathbf{I}_{gt}). \quad (2)$$

The forward diffusion process adds noise to the ground truth latent representation $\mathbf{Z}^{gt}$ according to the DDIM schedule:

$$\mathbf{Z}_t^{enh} = \sqrt{\bar{\alpha}_t} \mathbf{Z}^{gt,enh} + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (3)$$

The reverse process employs a conditional UNet network ϵ_{θ}^{enh} to predict the noise, conditioned on the LQ latent $\mathbf{Z}_{lq}$ and SII guidance:

$$\epsilon_{\theta}^{enh} = \text{UNet}^{enh}(\mathbf{Z}_t^{enh}, t, \mathbf{Z}_{lq}, \mathbf{F}_{oi}^{(t)}). \quad (4)$$

where t is the timestep, and $\mathbf{F}_{oi}^{(t)}$ is the guided feature from the SII module at step t .

During inference, given only an LQ image $\mathbf{X}^{lq}$, we encode it to obtain $\mathbf{Z}_{lq}$. We then initialize with random noise $\mathbf{Z}_T^{enh} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and perform DDIM sampling:

$$\begin{aligned} \mathbf{Z}_{t-1}^{enh} = & \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{Z}_t^{enh} - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\theta}^{enh}}{\sqrt{\bar{\alpha}_t}} \right) \\ & + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2 \epsilon_{\theta}^{enh}} + \sigma_t \epsilon_t. \end{aligned} \quad (5)$$

where $\sigma_t = \eta \sqrt{(1 - \bar{\alpha}_{t-1}) / (1 - \bar{\alpha}_t)} \sqrt{1 - \bar{\alpha}_t / \bar{\alpha}_{t-1}}$ and $\eta \in [0, 1]$ controls stochasticity.

The final enhanced image is obtained by decoding the denoised latent:

$$\mathbf{I}_p = \text{Decoder}(\mathbf{Z}_0^{enh}). \quad (6)$$

In the DiSeg branch, we employ DDIM for efficient segmentation mask generation.

During training, given an LQ image $\mathbf{X}^{lq}$ and its ground truth segmentation mask $\mathbf{S}_{gt}$, we encode the segmentation mask:

$$\mathbf{Z}^{gt,seg} = \text{Encoder}^{seg}(\mathbf{S}_{gt}). \quad (7)$$

The forward diffusion process adds noise to the ground truth segmentation latent:

$$\mathbf{Z}_t^{seg} = \sqrt{\bar{\alpha}_t} \mathbf{Z}^{gt,seg} + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (8)$$

The segmentation denoiser ϵ_{θ}^{seg} predicts noise conditioned on LQ image features and SII guidance:

$$\mathbf{F}_{lq} = \text{Seg Encoder}(\mathbf{X}^{lq}). \quad (9)$$

$$\epsilon_{\theta}^{seg} = \text{UNet}^{seg}(\mathbf{Z}_t^{seg}, t, \mathbf{F}_{lq}, \mathbf{F}_{os}^{(t)}). \quad (10)$$

where $\mathbf{F}_{lq}$ are LQ image features and $\mathbf{F}_{os}^{(t)}$ is the guided feature from SII.

During inference, we initialize with random noise $\mathbf{Z}_T^{seg} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and perform DDIM sampling:

$$\begin{aligned} \mathbf{Z}_{t-1}^{seg} = & \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{Z}_t^{seg} - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\theta}^{seg}}{\sqrt{\bar{\alpha}_t}} \right) \\ & + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2 \epsilon_{\theta}^{seg}} + \sigma_t \epsilon_t. \end{aligned} \quad (11)$$

The final segmentation mask is obtained by decoding:

$$\mathbf{S}_p = \text{Sigmoid}(\text{Decoder}^{seg}(\mathbf{Z}_0^{seg})). \quad (12)$$

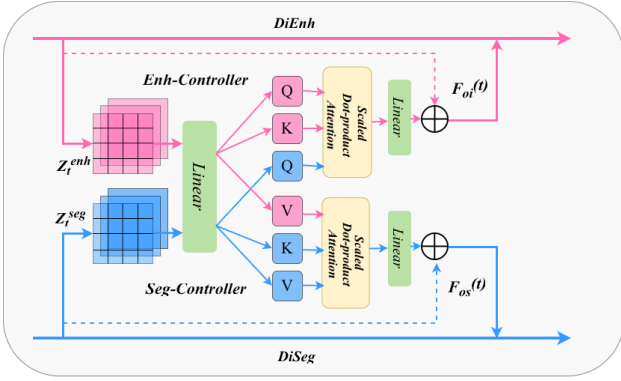


Figure 2: Diagram illustrating the Symbiotic Information interaction module within the DDIM framework. Enh-Controller treats $\mathbf{Z}_t^{seg}$ as the query matrix and $\mathbf{Z}_t^{enh}$ as the key and value matrices to compute cross-attention, generating refined image feature $\mathbf{F}_{oi}^{(t)}$ for DiEnh branch. The Seg-Controller treats $\mathbf{Z}_t^{enh}$ as the query matrix and $\mathbf{Z}_t^{seg}$ as the key and value matrices to compute cross-attention, generating guided feature $\mathbf{F}_{os}^{(t)}$ for DiSeg branch.

At each reverse step t during the DDIM sampling process, $\mathbf{Z}_t^{enh}$ and $\mathbf{Z}_t^{seg}$ are processed through the SII module. The SII generates guided features $\mathbf{F}_{oi}^{(t)}$ for the DiEnh branch and $\mathbf{F}_{os}^{(t)}$ for the DiSeg branch, enabling mutual information exchange. The DDIM framework allows for flexible scheduling of the SII interactions. We employ a strategy where SII is applied at key steps $\tau \subset \{1, 2, \dots, T\}$, with denser interactions at critical denoising stages.

3.2 Symbiotic Information Interaction Module

To achieve information interaction between the DiEnh and DiSeg branches within the DDIM framework, we propose a Symbiotic Information Interaction (SII) module for joint optimization. SII consists of Enh-Controller and Seg-Controller, as shown in Fig.2.

At time step t during the DDIM reverse process, the Enh-Controller extracts updated image information from $\mathbf{Z}_t^{enh}$ in DiEnh and segmentation information from $\mathbf{Z}_t^{seg}$, generating a guided feature $\mathbf{F}_{oi}^{(t)}$ for DiEnh. Specifically, the Enh-Controller adopts the Multi-Head Attention mechanism to compute refined image features:

$$\mathbf{F}_{oi}^{(t)} = \text{MultiHeadAttention}(\mathbf{Z}_t^{seg}, \mathbf{Z}_t^{enh}, \mathbf{Z}_t^{enh}), \quad (13)$$

where $\mathbf{Z}_t^{seg}$ serves as the query, and $\mathbf{Z}_t^{enh}$ serves as both key and value matrices.

We then apply $\text{ReLU}(\cdot)$ to the enhanced image feature and combine it with the input image feature:

$$\mathbf{F}_{oi}^{(t)} = \text{ReLU}(\mathbf{F}_{oi}^{(t)}) + \mathbf{Z}_t^{enh}. \quad (14)$$

The Seg-Controller extracts updated segmentation information from $\mathbf{Z}_t^{seg}$ in DiSeg and image information from $\mathbf{Z}_t^{enh}$, generating a guided feature $\mathbf{F}_{os}^{(t)}$ for DiSeg:

$$\mathbf{F}_{os}^{(t)} = \text{MultiHeadAttention}(\mathbf{Z}_t^{enh}, \mathbf{Z}_t^{seg}, \mathbf{Z}_t^{seg}), \quad (15)$$

where $\mathbf{Z}_t^{enh}$ serves as the query, and $\mathbf{Z}_t^{seg}$ serves as key and value matrices.

Then, we apply $\text{ReLU}(\cdot)$ to the segmentation feature and add it to the input segmentation feature:

$$\mathbf{F}_{os}^{(t)} = \text{ReLU}(\mathbf{F}_{os}^{(t)}) + \mathbf{Z}_t^{seg}. \quad (16)$$

The final image feature $\mathbf{F}_{oi}^{(t)}$ and segmentation feature $\mathbf{F}_{os}^{(t)}$ are then used as conditioning inputs for the next DDIM sampling step.

3.3 Joint Training Mechanism

For DiSiNet, we extend the DDIM training objective to incorporate symbiotic interactions:

$$\mathcal{L}_{\text{DiEnh}}^{\text{DDIM}} = \mathbb{E}_{t, \mathbf{z}_0^{enh}, \epsilon} [\|\epsilon - \epsilon_\theta^{enh}(\mathbf{Z}_t^{enh}, t, \mathbf{Z}_{lq}, \mathbf{F}_{oi}^{(t)})\|^2]. \quad (17)$$

$$\mathcal{L}_{\text{DiSeg}}^{\text{DDIM}} = \mathbb{E}_{t, \mathbf{z}_0^{seg}, \epsilon} [\|\epsilon - \epsilon_\theta^{seg}(\mathbf{Z}_t^{seg}, t, \mathbf{F}_{lq}, \mathbf{F}_{os}^{(t)})\|^2]. \quad (18)$$

The final enhanced image $\mathbf{I}_p$ from DiEnh is supervised by a restoration loss:

$$\mathcal{L}_{\text{enh}} = \|\mathbf{I}_p - \mathbf{I}_{gt}\|^2. \quad (19)$$

where $\mathbf{I}_{gt}$ is the ground truth high-quality image.

The final segmentation mask $\mathbf{S}_p$ from DiSeg is optimized using a weighted binary cross-entropy loss:

$$\mathcal{L}_{\text{seg}} = -\frac{1}{N} \sum w_i [\mathbf{S}_{gt} \log(\mathbf{S}_p) + (1 - \mathbf{S}_{gt}) \log(1 - \mathbf{S}_p)]. \quad (20)$$

where $\mathbf{S}_{gt}$ is the ground truth mask.

The overall loss integrates both DDIM training losses and final output supervision:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{DiEnh}}^{\text{DDIM}} + \beta \mathcal{L}_{\text{DiSeg}}^{\text{DDIM}} + \mathcal{L}_{\text{enh}} + \lambda \mathcal{L}_{\text{seg}}. \quad (21)$$

where β balances the DDIM training losses between branches, and λ weights the segmentation loss relative to enhancement loss.

4 Experiments and Results

4.1 Datasets

To evaluate our DiSiNet model's performance, we conduct experiments on several medical image segmentation datasets including ACDC [Bernard *et al.*, 2018], KiTS19 [Heller *et al.*, 2021], and TN3K [Gong *et al.*, 2021]. ACDC dataset is a heart dataset featuring segmentation annotations for the LV and RV volumes and the MYO. ACDC contains 150 patients, 100 training patients, and 50 test patients. The publicly accessible KiTS19 dataset comprises CT data from 300 patients. Specifically, cases indexed from 0 to 209 constitute the training set, while cases from 210 to 299 are designated as the test set. The TN3K (Thyroid Nodule Region Segmentation Dataset) consists of 3493 ultrasound images from 2421 patients. TN3K dataset contains 2879 images for training and 614 images for testing. For low-quality and high-quality image pairs, we use the original image as high-quality images, and we degraded the image as STS-SR [Li *et al.*, 2025] method through down sampling to obtain low-quality images.

4.2 Experimental Setup

All experiments were performed using the PyTorch framework and executed on 4 NVIDIA GeForce RTX 3090 GPUs. For 3D datasets like MRI ACDC or CT KiTS19, we split them into 2D with a slice depth of one and corresponding labels were resized to dimensions of 320×320 pixels. In contrast, images from the TN3K datasets were uniformly resized to a resolution of 256×256 pixels. The networks were trained in an end-to-end manner utilizing the AdamW optimizer with a batch size of 32. The starting learning rate was established at 1×10^{-4} . We conducted a joint loss function was employed to optimize the network.

4.3 Implementation Details

Our DiSiNet follows the standard DDIM framework: the model is trained with the full $T = 1000$ diffusion steps (as in DDPM) to learn the complete denoising process, while during inference we adopt DDIM’s accelerated sampling with only $S = 50$ steps. This is achieved by selecting a subsequence of timesteps $\{\tau_1, \tau_2, \dots, \tau_S\}$ where $\tau_S = T$, allowing efficient generation while preserving distribution fidelity. We employ a cosine noise schedule [Nichol and Dhariwal, 2021] for $\bar{\alpha}_t$:

$$\bar{\alpha}_t = \frac{\cos(\pi t / (2T + s))}{1 + s}, \quad s = 0.008 \quad (22)$$

The SII module is applied at all sampling steps to ensure continuous information exchange. For training, we use $\eta = 1$ (fully stochastic) to improve robustness, while for inference we use $\eta = 0$ (deterministic) for reproducibility. The weighting factors are set through ablation studies: $\beta = 1.0$ to equally weight both DDIM training losses.

4.4 Joint multi-task learning improves medical image enhancement

The quantitative results on medical image enhancement task are shown in Table 1. Among the methods compared on ACDC datasets, DiSiNet demonstrates superior performance, achieving a PSNR of 32.15 and an SSIM of 0.932. This performance positions DiSiNet at the forefront of the methodologies evaluated in the ACDC dataset. Figure 3 example (a) presents a qualitative comparison for medical image enhancement using the MRI ACDC dataset. The enlarged area, emphasized by red rectangular boxes, illustrates DiSiNet’s effectiveness in maintaining fine structural details in the denoised images relative to other techniques. Remarkably, the ventricular region is depicted with the greatest clarity and sharpness in the DiSiNet output, while competing methods demonstrate significant blurring and loss of detail in this area. The pulmonary fissure serves as the boundary between the ventricular lobes. Within the methods evaluated on the KiTS19 datasets, DiSiNet achieves a PSNR of 34.68 and an SSIM of 0.961, marking it as the leading approach for this dataset. These results not only surpass the performance of other methods but also indicate a superior ability to maintain both pixel accuracy and perceptual quality in the enhanced images. Figure 3 example (b) presents a qualitative comparison for medical image enhancement using the CT KiTS19 dataset. The enlarged red area contains the kidney and kidney tumor. The kidney organ and tumor are distinctly shown

with contour boundaries in the DiSiNet enhanced images, whereas they cannot be differentiated by other methods. In the comparison of methods using the TN3K datasets, DiSiNet stands out as the top performer on the TN3K dataset, achieving a PSNR of 34.26 and an SSIM of 0.942. This remarkable performance underscores the model’s effectiveness in enhancing image quality while preserving structural integrity, making it a benchmark in the field. Figure 3 example (c) presents a qualitative comparison for medical image enhancement using the ultrasound TN3K. The enlarged red area shows the thyroid nodule, which has a clear contour against the surrounding background in the DiSiNet enhanced images. In contrast, it appears more blurred using other methods.

4.5 Joint multi-task learning improves medical image segmentation

In analyzing the performance of various methods on the ACDC dataset as presented in Table 2, DiSiNet emerges as the top-performing model, achieving an avg-mIoU of 84.81 and an avg-Dice of 91.78. The strong results indicate that the model excels in capturing the intricate features necessary for effective segmentation. We extensively evaluated the deep models on MRI ACDC datasets with visual results, as shown in Figure 4. It is visible that the DiSiNet can produce clean medical images without any visually disturbing artifacts for denoising on ACDC dataset. We can observe the rich hierarchical structure and distinct boundaries among the different heart organizations, containing the LV, RV, and MYO. The tiny but vital details of the heart become gradually obscured. This phenomenon indicates that the proposed DiSiNet can capture the tiny but vital structures hidden under the noise even if the noise level is high, thanks to the ability to abundantly collect information and preserve intricate structures through SII modules.

In evaluating the performance of various methods on the KiTS19 dataset, it is evident that joint multi-task learning significantly enhances segmentation capabilities. DiSiNet emerges as the leading model, achieving an mIoU of 75.25 and a Dice coefficient of 85.87. These results not only position DiSiNet at the forefront of the evaluated methods but also demonstrate its effectiveness in accurately segmenting images within the KiTS19 dataset. The high performance indicates that DiSiNet is adept at addressing the intricate challenges associated with this type of imagery.

In assessing the performance of different methods on the TN3K dataset, it is clear that joint multi-task learning significantly enhances segmentation outcomes, as evidenced by the metrics of mIoU and Dice coefficient. DiSiNet stands out as the top-performing model, achieving an mIoU of 75.18 and a Dice coefficient of 85.82. These exceptional results highlight DiSiNet’s ability to effectively segment images within the TN3K dataset, demonstrating a robust performance that surpasses that of competing methodologies. The high Dice coefficient, in particular, indicates a strong alignment between the predicted and ground truth segmentation, reflecting the model’s superior capability in capturing intricate details.

Method	ACDC		KiTS19		TN3K	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
PCENet [Liu <i>et al.</i> , 2022]	19.86 _{↓12.29}	0.832 _{↓0.100}	23.63 _{↓11.05}	0.703 _{↓0.258}	24.78 _{↓9.48}	0.894 _{↓0.048}
ArcNet [Li <i>et al.</i> , 2022b]	22.85 _{↓9.30}	0.824 _{↓0.108}	22.05 _{↓12.63}	0.802 _{↓0.159}	19.88 _{↓14.38}	0.851 _{↓0.091}
SCRNet [Li <i>et al.</i> , 2022a]	24.05 _{↓8.10}	0.886 _{↓0.046}	23.46 _{↓11.22}	0.854 _{↓0.107}	23.62 _{↓10.64}	0.895 _{↓0.047}
DeepRFT [Mao <i>et al.</i> , 2023]	26.35 _{↓5.80}	0.902 _{↓0.030}	30.45 _{↓4.23}	0.947 _{↓0.014}	26.44 _{↓7.82}	0.924 _{↓0.018}
I-SECRET [Cheng <i>et al.</i> , 2021]	27.18 _{↓4.97}	0.896 _{↓0.036}	32.60 _{↓2.08}	0.934 _{↓0.027}	29.04 _{↓5.22}	0.898 _{↓0.044}
CHLNet [Wang <i>et al.</i> , 2024a]	30.41 _{↓1.74}	0.921 _{↓0.011}	33.65 _{↓1.03}	0.953 _{↓0.008}	34.11 _{↓0.15}	0.929 _{↓0.013}
StillGAN [Ma <i>et al.</i> , 2021]	24.97 _{↓7.18}	0.845 _{↓0.087}	27.34 _{↓7.34}	0.912 _{↓0.049}	25.89 _{↓8.87}	0.885 _{↓0.057}
LED [Cheng <i>et al.</i> , 2023]	25.88 _{↓6.27}	0.918 _{↓0.014}	33.26 _{↓1.42}	0.923 _{↓0.038}	27.58 _{↓6.68}	0.846 _{↓0.096}
Fast-DDPM [Jiang <i>et al.</i> , 2025]	30.65 _{↓1.50}	0.922 _{↓0.010}	31.25 _{↓3.43}	0.948 _{↓0.013}	34.12 _{↓0.14}	0.935 _{↓0.007}
Wang [Wang <i>et al.</i> , 2024b]	30.81 _{↓1.34}	0.916 _{↓0.016}	32.87 _{↓1.81}	0.936 _{↓0.025}	33.88 _{↓0.38}	0.926 _{↓0.016}
DiSiINet	32.15	0.932	34.68	0.961	34.26	0.942

Table 1: Joint multi-task learning improves image enhancement Model comparison performance.

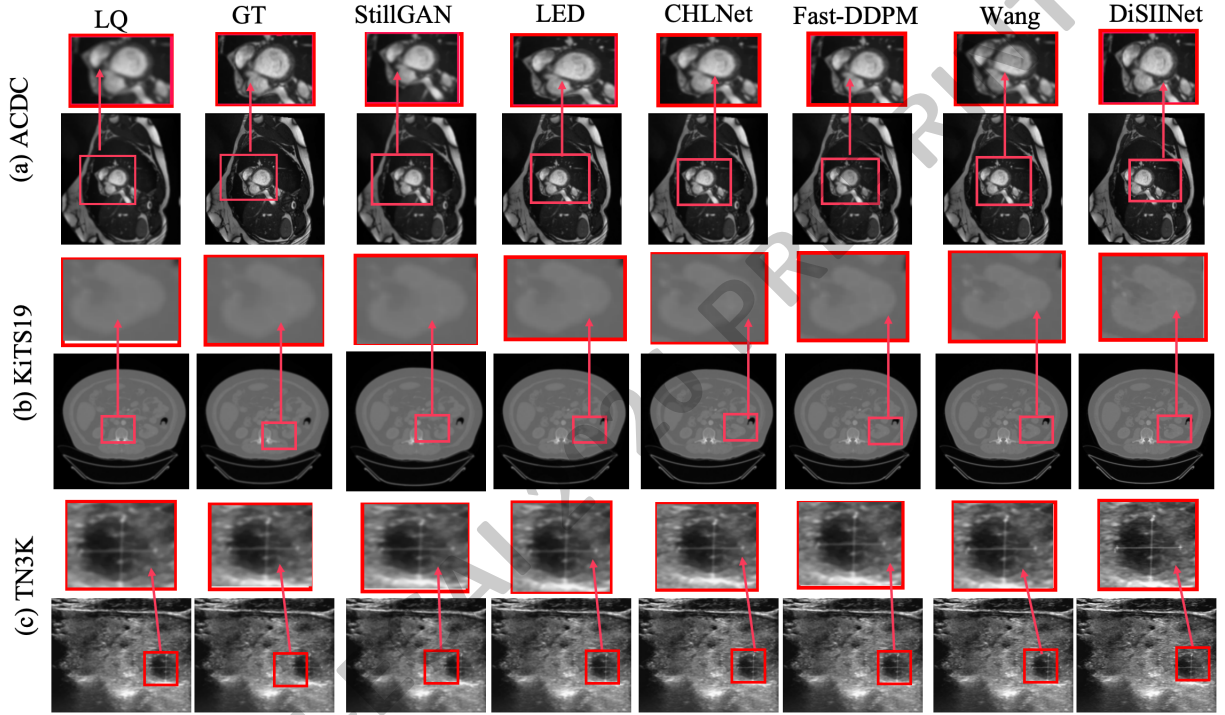


Figure 3: Medical image enhancement task visual comparison results on three datasets. Example (a) row correspond to MRI ACDC images, example (b) row correspond to CT KiTS dataset and example (c) row correspond to ultrasound TN3K dataset.

Method	ACDC		KiTS19		TN3K	
	avg-mIoU	avg-Dice	avg-mIoU	avg-Dice	mIoU	Dice
UNet [Ronneberger <i>et al.</i> , 2015]	70.07 _{↓15.05}	82.87 _{↓8.96}	59.60 _{↓17.65}	74.68 _{↓11.19}	64.20 _{↓10.98}	78.20 _{↓7.62}
nnUNet [Isensee <i>et al.</i> , 2021]	82.26 _{↓2.55}	91.40 _{↓0.38}	68.40 _{↓6.85}	81.20 _{↓4.67}	66.10 _{↓9.08}	79.60 _{↓6.22}
ResU-Net++ [Cao <i>et al.</i> , 2022]	75.00 _{↓9.81}	85.73 _{↓6.05}	64.20 _{↓11.05}	78.20 _{↓7.67}	67.80 _{↓7.38}	80.90 _{↓4.92}
TransUNet [Jha <i>et al.</i> , 2021]	77.60 _{↓7.21}	87.46 _{↓4.32}	66.80 _{↓8.45}	80.16 _{↓5.71}	69.80 _{↓5.38}	82.16 _{↓3.66}
SwinUNet [Chen <i>et al.</i> , 2021]	80.20 _{↓4.61}	89.15 _{↓2.63}	68.60 _{↓6.65}	81.38 _{↓4.49}	70.00 _{↓5.18}	82.40 _{↓3.42}
DenoiSeg [Buchholz <i>et al.</i> , 2020]	83.27 _{↓1.54}	90.15 _{↓1.63}	67.80 _{↓7.45}	80.88 _{↓4.99}	70.40 _{↓4.78}	82.60 _{↓3.22}
MedSegDiff-V2 [Wu <i>et al.</i> , 2024a]	84.68 _{↓0.13}	91.70 _{↓0.08}	75.20 _{↓0.05}	85.84 _{↓0.03}	74.82 _{↓0.36}	85.60 _{↓0.22}
Wang [Wang <i>et al.</i> , 2024b]	82.40 _{↓2.41}	90.33 _{↓1.45}	73.70 _{↓1.55}	84.78 _{↓1.09}	73.30 _{↓1.95}	84.60 _{↓1.22}
DiSiINet	84.81	91.78	75.25	85.87	75.18	85.82

Table 2: Joint multi-task learning improves image segmentation model comparison performance.

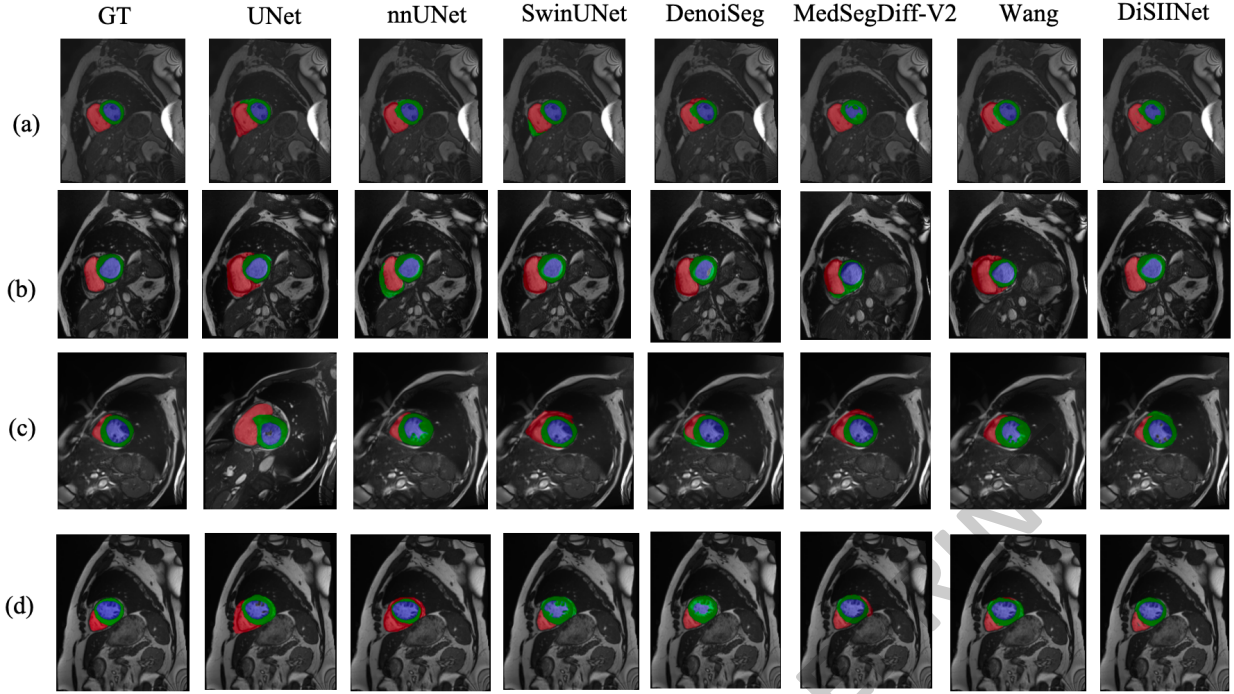


Figure 4: Medical image segmentation task visual comparison results on MRI ACDC dataset.

4.6 Ablation Study

Effectiveness of Key Component

Method	ACDC		KiTS19		TN3K	
	PSNR	avg-Dice	PSNR	avg-Dice	PSNR	Dice
Cascaded	23.08	84.51	24.46	81.03	23.73	82.62
w/o Enh-Controller	26.28	88.34	26.38	84.87	25.13	85.48
w/o Seg-Controller	27.75	86.12	28.34	82.61	27.33	81.06
Joint Enh-Controller & Seg-Controller (Ours DiSiINet)	32.15	91.83	34.68	86.89	34.26	86.62

Table 3: Ablation study of each component of the SII module.

Table 3 highlights the contributions of different components in DiSiINet. Our proposed DiSiINet, which integrates both the Enh-Controller and Seg-Controller, achieves the highest performance across all datasets. For comparison, a baseline cascaded method where low-quality images are first enhanced with Fast-DDPM and then segmented using MedSegDiff-V2 exhibits a significant performance decline. This result underscores the critical contribution of each controller to optimal outcomes. Specifically, removing the Enh-Controller from DiSiINet leads to decreased metrics. Conversely, removing the Seg-Controller results in even lower PSNR values and a more pronounced decline in Dice coefficients. These ablation studies confirm that the synergistic integration of both controllers is essential for simultaneously maximizing image quality and segmentation accuracy, making DiSiINet model the most effective configuration.

Sensitivity analysis of loss weights

Finally, we evaluate the impact of different weighting parameter λ between the DiEnh and DiSeg branches during training on performance using three dataset. The choice of loss weighting parameter λ is entirely result-driven. Table

λ	ACDC		KiTS19		TN3K	
	PSNR	avg-Dice	PSNR	avg-Dice	PSNR	Dice
$\lambda = 3 \times 10^{-1}$	31.55	90.74	34.53	85.32	34.23	85.12
$\lambda = 4 \times 10^{-1}$	31.26	91.39	34.31	85.68	34.18	85.36
$\lambda = 5 \times 10^{-1}$	32.15	91.83	34.68	86.89	34.26	86.62
$\lambda = 6 \times 10^{-1}$	30.62	91.58	33.26	86.18	33.89	86.33

Table 4: Ablation study of weighting parameter λ .

4 shows that our DiSiINet achieves an optimal result when $\lambda = 5 \times 10^{-1}$. Therefore, we use this setting in all experiments.

5 Conclusion

This paper presents DiSiINet, a dual-DDIM diffusion symbiotic network that joint medical image enhancement and segmentation. By coupling two diffusion processes via the Symbiotic Information Interaction (SII) module, DiSiINet enables continuous, feature-level information exchange that allows enhancement and segmentation tasks to mutually reinforce each other. Extensive experiments on multi-modal medical datasets demonstrate that DiSiINet yields superior performance compared to sequential or independent approaches. However, the performance gains come at the cost of increased computational complexity. Future work will focus on exploring distillation techniques to smaller models, and investigating adaptive mechanisms to reduce the required denoising steps without compromising the symbiotic benefits.

Acknowledgements

This research is part of the IN-CYPHER programme and is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

- [Bernard *et al.*, 2018] Olivier Bernard, Alain Lalande, and Zotti. *et al.* Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.
- [Buchholz *et al.*, 2020] Tim-Oliver Buchholz, Mangal Prakash, Deborah Schmidt, Alexander Krull, and Florian Jug. Denoiseg: joint denoising and segmentation. In *European Conference on Computer Vision*, pages 324–337. Springer, 2020.
- [Cao *et al.*, 2022] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCVW*. Springer, 2022.
- [Chen *et al.*, 2021] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- [Cheng *et al.*, 2021] Pujin Cheng, Li Lin, Yijin Huang, Junyan Lyu, and Xiaoying Tang. I-secret: Importance-guided fundus image enhancement via semi-supervised contrastive constraining. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 87–96. Springer, 2021.
- [Cheng *et al.*, 2023] Puijin Cheng, Li Lin, Yijin Huang, Huaqing He, Wenhan Luo, and Xiaoying Tang. Learning enhancement from degradation: A diffusion model for fundus image enhancement. *arXiv preprint arXiv:2303.04603*, 2023.
- [Gao *et al.*, 2023] Qi Gao, Zilong Li, Junping Zhang, Yi Zhang, and Hongming Shan. Corediff: Contextual error-modulated generalized diffusion model for low-dose ct denoising and generalization. *IEEE Transactions on Medical Imaging*, 43(2):745–759, 2023.
- [Gong *et al.*, 2021] Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 257–261. IEEE, 2021.
- [Heller *et al.*, 2021] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, *et al.* The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021.
- [Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [Isensee *et al.*, 2021] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- [Jha *et al.*, 2021] Debesh Jha, Pia H Smedsrud, Dag Johansen, Thomas De Lange, Håvard D Johansen, Pål Halvorsen, and Michael A Riegler. A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation. *IEEE Journal of Biomedical and Health Informatics*, 2021.
- [Jiang *et al.*, 2025] Hongxu Jiang, Muhammad Imran, Teng Zhang, Yuyin Zhou, Muxuan Liang, Kuang Gong, and Wei Shao. Fast-ddpm: Fast denoising diffusion probabilistic models for medical image-to-image generation. *IEEE Journal of Biomedical and Health Informatics*, 2025.
- [Li *et al.*, 2022a] Heng Li, Haofeng Liu, Huazhu Fu, Hai Shu, Yitian Zhao, Xiaoling Luo, Yan Hu, and Jiang Liu. Structure-consistent restoration network for cataract fundus image enhancement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 487–496. Springer, 2022.
- [Li *et al.*, 2022b] Heng Li, Haofeng Liu, Yan Hu, Huazhu Fu, Yitian Zhao, Hanpei Miao, and Jiang Liu. An annotation-free restoration network for cataractous fundus images. *IEEE Transactions on Medical Imaging*, 41(7):1699–1710, 2022.
- [Li *et al.*, 2025] Yinghua Li, Weiao Hao, Hao Zeng, Longguang Wang, Jian Xu, Sidheswar Routray, Rutvij H. Jhaveri, and Thippa Reddy Gadekallu. Cross-scale texture supplementation for reference-based medical image super-resolution. *IEEE Journal of Biomedical and Health Informatics*, pages 1–15, 2025.
- [Liang *et al.*, 2022] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, pages 574–591. Springer, 2022.
- [Lin *et al.*, 2024] Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, pages 430–448. Springer, 2024.
- [Liu *et al.*, 2018] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S Huang. When image denoising meets high-level vision tasks: A deep learning approach. In *Proceedings of the International Joint Conference on Artificial Intelligence*, page 842–848, 2018.
- [Liu *et al.*, 2022] Haofeng Liu, Heng Li, Huazhu Fu, Ruoxiu Xiao, Yunshu Gao, Yan Hu, and Jiang Liu. Degradation-invariant enhancement of fundus images via pyramid constraint network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 507–516. Springer, 2022.
- [Ma *et al.*, 2021] Yuhui Ma, Jiang Liu, Yonghuai Liu, Huazhu Fu, Yan Hu, Jun Cheng, Hong Qi, Yufei Wu, Jiong

- Zhang, and Yitian Zhao. Structure and illumination constrained gan for medical image enhancement. *IEEE Transactions on Medical Imaging*, 40(12):3955–3967, 2021.
- [Mao et al., 2023] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1905–1913, 2023.
- [Nichol and Dhariwal, 2021] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021.
- [Park et al., 2023] JoonKyu Park, Sanghyun Son, and Kyoung Mu Lee. Content-aware local gan for photo-realistic super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10585–10594, 2023.
- [Rombach et al., 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [Ronneberger et al., 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015.
- [Song et al., 2021] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.
- [Wang et al., 2023] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2555–2563, 2023.
- [Wang et al., 2024a] Yaqi Wang, Leqi Chen, Qingshan Hou, Peng Cao, Jinzhu Yang, Xiaoli Liu, and Osmar R Ziaiane. A clinical-oriented lightweight network for high-resolution medical image enhancement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer, 2024.
- [Wang et al., 2024b] Zhicheng Wang, Jiacheng Li, Yinda Chen, Jiateng Shou, Shiyu Deng, Wei Huang, and Zhiwei Xiong. Joint em image denoising and segmentation with instance-aware interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–413. Springer, 2024.
- [Wu et al., 2024a] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6030–6038, 2024.
- [Wu et al., 2024b] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024.
- [Xiang et al., 2023] Tiange Xiang, Mahmut Yurt, Ali B Syed, Kawin Setsompop, and Akshay Chaudhari. Ddm²: Self-supervised diffusion mri denoising with generative diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.
- [Xiao et al., 2024] Jiahua Xiao, Jiawei Zhang, Dongqing Zou, Xiaodan Zhang, Jimmy Ren, and Xing Wei. Semantic segmentation prior for diffusion-based real-world super-resolution. *arXiv preprint arXiv:2412.02960*, 2024.
- [Xu et al., 2023] Shunxin Xu, Ke Sun, Dong Liu, Zhiwei Xiong, and Zheng-Jun Zha. Synergy between semantic segmentation and image denoising via alternate boosting. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–23, 2023.
- [Yang et al., 2024] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European conference on computer vision*, pages 74–91. Springer, 2024.
- [Zhai et al., 2022] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.