EndoIR: Degradation-Agnostic All-in-One Endoscopic Image Restoration via Noise-Aware Routing Diffusion

Tong Chen[1]*, Xinyu Ma[2]*, Long Bai[3]*, Wenyang Wang[1], Yue Sun[2]†, Luping Zhou[1]‡

[1]School of Electrical and Computer Engineering, The University of Sydney, Sydney, Australia
[2]Intelligent Medical Computing Laboratory, Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
[3]The Chinese University of Hong Kong, Hong Kong SAR, China
tong.chen1@sydney.edu.au
https://github.com/DavisMeee/EndoIR

Abstract

Endoscopic images often suffer from diverse and co-occurring degradations such as low lighting, smoke, and bleeding, which obscure critical clinical details. Existing restoration methods are typically task-specific and often require prior knowledge of the degradation type, limiting their robustness in real-world clinical use. We propose EndoIR, an all-in-one, degradation-agnostic diffusion-based framework that restores multiple degradation types using a single model. EndoIR introduces a Dual-Domain Prompter that extracts joint spatial–frequency features, coupled with an adaptive embedding that encodes both shared and task-specific cues as conditioning for denoising. To mitigate feature confusion in conventional concatenation-based conditioning, we design a Dual-Stream Diffusion architecture that processes clean and degraded inputs separately, with a Rectified Fusion Block integrating them in a structured, degradation-aware manner. Furthermore, Noise-Aware Routing Block improves efficiency by dynamically selecting only noise-relevant features during denoising. Experiments on SegSTRONG-C and CEC datasets demonstrate that EndoIR achieves state-of-the-art performance across multiple degradation scenarios while using fewer parameters than strong baselines, and downstream segmentation experiments confirm its clinical utility.

Introduction

Endoscopic imaging is indispensable for gastrointestinal diagnosis and surgical guidance. However, in real-world clinical settings, captured images are often degraded by multiple co-occurring factors such as low illumination, smoke, and bleeding (Bai et al. 2023; Pan et al. 2022; Ramírez et al. 2002). These degradations obscure fine anatomical structures, reduce visibility of surgical tools, and impair downstream computational analysis. Restoring such degraded images is therefore essential for improving both the diagnostic quality and the safety of endoscopic procedures.

A variety of restoration approaches have been developed over the years (García-Vega et al. 2023; Ma et al. 2021; Guo et al. 2024). Early methods largely targeted single degradations in isolation, such as illumination enhancement, smoke removal, or denoising, typically using task-specific CNNs, GANs, or transformer-based architectures (Wang et al. 2024; Pan et al. 2022; Su and Wu 2023). While effective for their intended degradation type, these models cannot handle complex, mixed degradations without retraining or explicit knowledge of the degradation type. More recent research in the broader computer vision community has explored all-in-one restoration frameworks, aiming to address multiple degradations with a single unified model.

Despite these advances, three limitations remain prominent. First, many unified methods rely on explicit

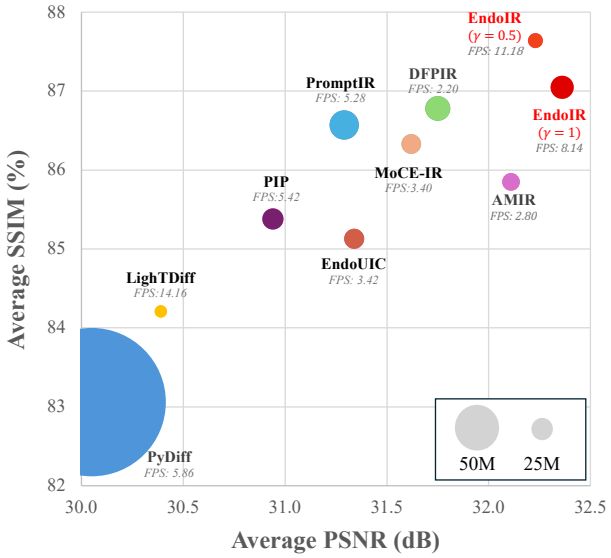


Figure 1: Comparison of average restoration accuracy (PSNR/SSIM), model size, and inference speed (FPS, annotated above each marker). Both our EndoIR ($\gamma = 0.5$) and ($\gamma = 1$) achieve state-of-the-art restoration quality, while the $\gamma = 0.5$ variant offers a compact model size and high inference speed, making it well-suited for real-time clinical use.

*These authors contributed equally.
†Co-Corresponding Author
‡Corresponding Author

degradation-type input or accurate prompt extraction. For example, prompt-based methods such as PromptIR (Potlapalli et al. 2023) and AutoDIR (Jiang et al. 2024) condition restoration on learned visual or language prompts, but these approaches require either prior knowledge of the degradation type or robust degradation classification, which is unreliable in dynamic surgical scenes. Second, routing-based (Ren et al. 2024) architectures such as AMIR [1]teyang2024all select restoration paths from degradation-specific cues. While effective for distinguishing degradation types, this can bias the model toward degradation modeling at the expense of preserving fine anatomical structures—a critical concern in medical images where subtle structural fidelity is essential. Third, in the case of diffusion-based methods (Chen et al. 2024; Xia et al. 2023), conventional concatenation-based conditioning merges degraded and noisy inputs early in the network. This can confuse the model with mixed feature distributions, particularly when degradations co-occur, leading to unstable optimization and reduced restoration quality. A detailed side-by-side comparison of these approaches and our EndoIR is provided in Supplementary Table 1, highlighting the differences in degradation handling, conditioning strategy, and efficiency.

To address these challenges, we introduce **EndoIR**, the first all-in-one, degradation-agnostic diffusion framework tailored for complex endoscopic image restoration. EndoIR is designed to operate without any prior knowledge of the degradation type, while effectively disentangling and leveraging both degradation-specific and content-preserving information. Central to our approach is a **Dual-Domain Prompter** that jointly extracts spatial and frequency features to produce informative restoration prompts, coupled with a **Task Adaptive Embedding** mechanism that separates shared anatomical content from degradation-specific cues. To overcome feature confusion in conventional conditioning, we design a **Dual-Stream Diffusion** architecture that processes clean and degraded inputs independently, followed by a **Rectified Fusion Block** that integrates them in a structured, degradation-aware manner. Finally, to improve computational efficiency, we propose a **Noise-Aware Routing Block** that dynamically selects only noise-relevant features for denoising, reducing redundant computation while preserving restoration quality.

Our main contributions are summarized as follows:

- We propose degradation-agnostic restoration without labels via a Dual-Domain Prompter and task adaptive embedding for joint spatial-frequency prompt learning.
- We introduce a disentangled dual-stream conditioning mechanism for stable diffusion restoration, with a Rectified Fusion Block for structured feature integration.
- We design an efficient, noise-aware decoding strategy using the Noise-Aware Routing Block that reduces computation without compromising quality.
- We conduct comprehensive validation on challenging endoscopic benchmarks, achieving state-of-the-art performance on SegSTRONG-C and CEC datasets and demonstrating strong generalization in downstream surgical tool segmentation tasks.

Related Work

In general scenarios for all-in-one restoration, early works addressed diverse degradation scenarios using corruption-agnostic frameworks, such as contrastive degradation encoders and degradation-guided decoders (Li et al. 2022). Subsequent methods like RAM (Qin et al. 2024) introduced masked image modeling for intrinsic feature extraction, while ADMS (Park, Lee, and Chun 2023) utilized adaptive filters to handle unknown degradations with minimal overhead. Recent studies explored adaptive prompts to guide restoration. PromptIR (Potlapalli et al. 2023) and PIP (Li et al. 2023) developed prompt-based modules to dynamically encode degradation information, combining high-level and low-level prompts through dedicated interaction mechanisms. Vision-language approaches also advanced the field by injecting semantic cues. AutoDIR (Jiang et al. 2024) integrated a vision-language module for open-vocabulary degradation recognition with a diffusion model guided by text prompts. MPerceiver (Ai et al. 2024) leveraged CLIP for improved zero-shot and few-shot performance. To mitigate task interference and data scarcity, DFPIR (Tian et al. 2025) introduced feature perturbations guided by degradation prompts, while FoundIR (Li et al. 2025) constructed a large real-world dataset and a two-stage diffusion-refinement framework. Defusion (Luo et al. 2025) further refined diffusion models with degradation-specific visual priors, focusing on visual rather than language cues.

In summary, unified frameworks and more recent prompt- and diffusion-based approaches have advanced all-in-one restoration with strong adaptability and robustness to diverse degradations. However, in medical imaging, degradation types and patterns often differ from those in natural images. Medical images typically have more regular structures and are highly sensitive to small changes, especially at lesion borders and in fine textures. Inaccurate restoration can cause errors in tasks like segmentation and diagnosis. Only a few studies address all-in-one restoration for medical images (Li et al. 2018). Recently, AMIR (Yang et al. 2024) introduced a task-adaptive routing network for multiple tasks, such as MRI super-resolution, CT denoising, and PET synthesis. Subsequent work further explores new paradigms in model architecture design (Yang et al. 2025; Chen et al. 2025). Nevertheless, in endoscopic scenarios, most work still targets specific tasks, like low-light enhancement (Bai et al. 2023; Chen et al. 2024), super-resolution (Chen et al. 2022; Liu et al. 2023), or desmoking (Pan et al. 2022; Wu et al. 2024), and optimizes each task in isolation. Few methods aim for a unified solution. For example, EndoUIC (Bai et al. 2024) provides unified enhancement for illumination only. Yet, degradations such as low light, smoke, and blood often occur together in real endoscopy (Ding et al. 2024), creating a clear need for an all-in-one model that can robustly restore images under multiple degradations for accurate diagnosis and safe clinical decision-making.

Methodology

Overview. The pipeline of our EndoIR framework is presented in Fig. 2. Our EndoIR framework is based on the

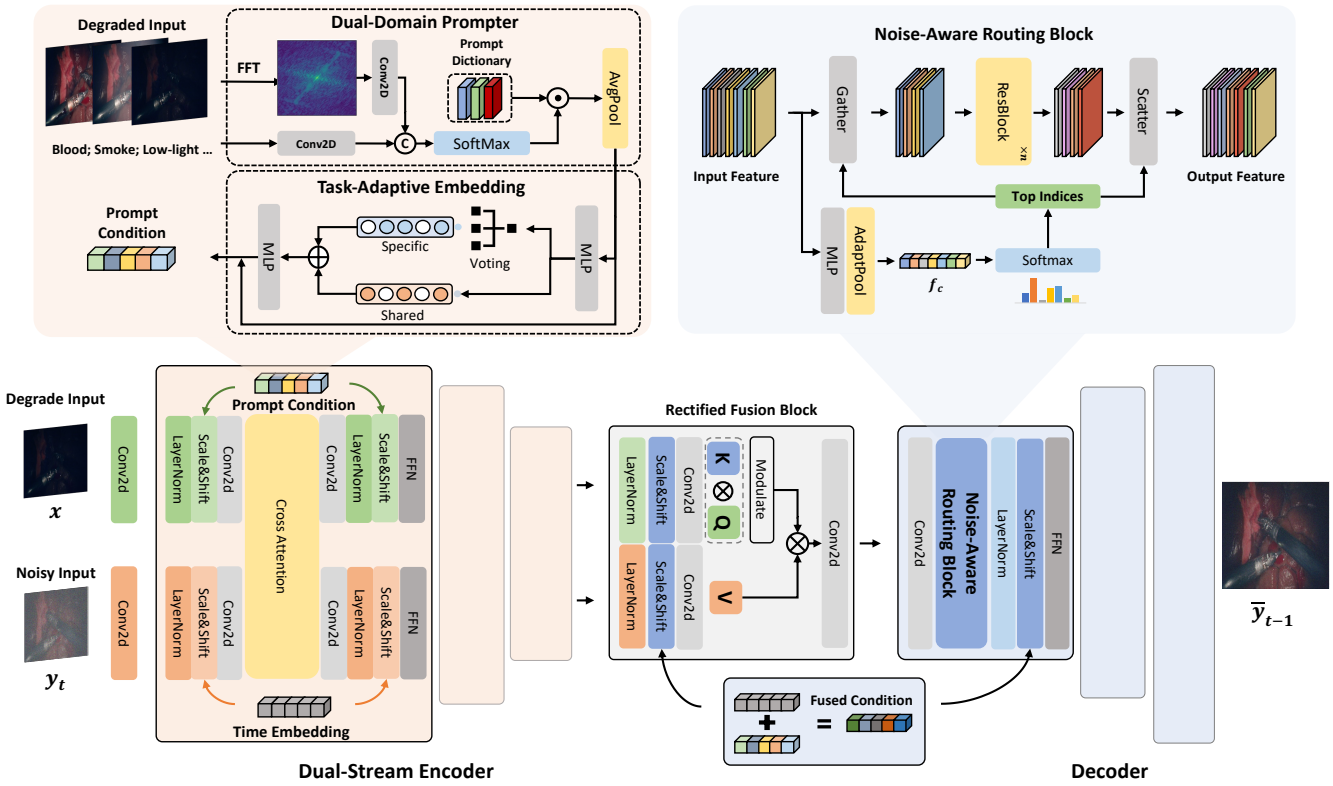


Figure 2: Overview of our EndoIR framework: Our denoising block consists of a Dual-Domain Prompter(DDP) to generate fine-grained prompt guidance; Task Adaptive Embedding(TAE) to dynamically learn task-specific conditions; Dual-Stream Encoder(DSE) to enable the dual-stream interaction for decoupled feature learning; Rectified Fusion Block(RFB) is designed to fuse dual-stream features; Noise-Aware Routing Block(NARB) for efficient feature refinement.

pyramid DDIM model (Zhou, Yang, and Yang 2023), which integrates a diffusion method to perform sampling in a pyramid resolution style to achieve faster sampling. Its processes can be expressed as follows:

$$q(y_t \mid y_{t-1}) := \mathcal{N}\left(y_t; \sqrt{\hat{\alpha}_t}(y_t \downarrow r_{t-1}/r_t), \beta_t \mathbf{I}\right), \quad (1)$$

$$p(y_{t-1} \mid y_t) := \mathcal{N}(y_{t-1}; \sqrt{\hat{\alpha}_{t-1}} f_\theta(y_t \uparrow r_t/r_{t-1}) + \frac{1-\hat{\alpha}_{t-1}}{1-\hat{\alpha}_t} y_t, \frac{\beta_t(1-\hat{\alpha}_{t-1})}{1-\hat{\alpha}_t} \mathbf{I}), \quad (2)$$

where y_t represents the image at time step t, β_t is the noise variance, and $\hat{\alpha}_t := \prod_{i=1}^{t}(1-\beta_i)$ is the cumulative product of noise schedules given by β_t. The denoising framework of our diffusion model adopts a Dual-Unet (Ronneberger, Fischer, and Brox 2015) architecture, consisting of our Dual-Stream Encoder and a Decoder with Noise-Aware Routing Blocks. The Rectified Fusion Block shall perform the fusion of dual-stream intermediate features. Additionally, the Dual-Domain Prompter and Task Adaptive Embedding are designed to achieve auxiliary feature representation for precision denoising guidance.

Dual-Domain Prompter. Prompt-based approaches often rely solely on spatial features, missing the complementary frequency cues that explicitly capture global structural integrity. Frequency features, obtained via Fast Fourier Transform (FFT), reveal global spectral signatures of degradation — for example, blood introduces localized low frequency dominance due to large homogeneous regions, smoke attenuates mid and high frequency components by veiling fine details, and low light shifts the overall spectrum. By fusing these complementary cues, we propose Dual-Domain Prompter to form richer degradation-aware prompts that encode both the spatial and the spectral characteristics of corruption. These prompts guide the restoration network to adapt effectively to diverse degradations without requiring explicit degradation labels. The effectiveness of this spatial frequency fusion is validated in our ablation results 4. First, we initialize a set of learnable parameters as the prompt dictionary $\mathbb{D}$. Then, we extract joint representations by integrating information from the spatial and frequency domains. After combining these representations with $\mathbb{D}$, we apply average pooling to obtain the prompt output. The Dual-Domain Prompter can be formulated as:

$$\begin{aligned} F_{img} &= Conv2D(x) \\ F_{freq} &= Conv2D(\mathbf{FFT}(x)) \\ \mathbb{P} &= \text{AvgPool}\{\text{Softmax}[F_{img} + F_{freq}] \cdot \mathbb{D}\} \end{aligned} \quad (3)$$

in which x represent the input image and $\mathbb{P}$ is the output prompt parameter. The output prompt $\mathbb{P}$ is further propa-

gated through the Task Adaptive Embedding to generate the conditions for controlling the diffusion process.

Task Adaptive Embedding. In Surgical imaging scenarios, the clean images often exhibit relatively consistent contents. When using a unified routing scheme, there is a risk that the model may overfit to degradation-specific cues, potentially leading to semantic drift or incorrect restoration. To mitigate this, we propose Task-Adaptive Embedding (TAE), which balances general content preservation with corruption-specific adaptation.

The task adaptive embedding module receives the DDP output $\mathbb{P}$ as input and processes it through two parallel branches. The shared branch captures global, content-preserving semantics that remain stable across different types of degradation. The specific branch consists of multiple modules, each specialized in handling a particular corruption pattern. A soft voting mechanism selects the most degraded-relevant branch based on the prompt itself. Formally, the task-aware embedding is given by:

$$E_{\text{task}} = \sum_{i=1}^{n} MLP_i(\mathbb{P}) + \sum_{k=1}^{K} MLP_k(\mathbb{P}) \cdot Top\mathbf{K}(\text{Softmax}(\mathbb{P})) \tag{4}$$

This disentangled embedding structure ensures that the model adapts flexibly to varying corruption types while maintaining fidelity to anatomical content. This hybrid embedding scheme enhances the model's ability for generate precise, condition-aware representations for guiding the denoising process within diffusion model.

Dual-Stream Encoder. Traditional conditional diffusion models often adopt a simple concatenation of the corrupted image as guidance, implicitly assuming that both follow similar distributions. However, since various corrupted inputs lie in distinct distribution domains, naïve concatenation can disrupt the intrinsic distribution of clean images, thereby introducing uncertainty and compromising the effectiveness of the denoising process. To address this, we propose a Dual-Stream Encoder(DSE) that separately encodes the corrupted image and the noise-added image, thereby disentangling degradation-specific information and facilitating more accurate guidance for restoration.

As illustrated in Fig. 2(d), given a corrupted input image x and its corresponding noise-added counterpart y_t, we first embed both images into feature representations using separate convolution layers, followed by Layer Normalization:

$$F'_n = \text{Conv2D}(\text{LN}(\text{Conv2D}(n))), \quad n \in \{x, y\} \tag{5}$$

These two features are then jointly processed via a cross-attention module. Specifically, the feature maps are concatenated along the channel dimension and treated as the query, key, and value in a self-attention block. Since the attention computation is spatially self-referential, the two different feature streams do not directly cross-interfere. The output is split back into two streams using a chunk operation, which are then passed through feed-forward convolution modules with residual connections to form the final encoded features.

The proposed architecture enables the model to capture both degradation-aware and content-aware representations in a disentangled manner. This design not only enhances the model's ability to characterise degradation patterns but also promotes the preservation of fine structures and textures during restoration.

Rectified Fusion Block. After the dual-stream encoder, both features from the degraded domain and the diffusion domain are supposed to be sufficiently extracted. To effectively integrate these two representations, we introduce the Rectified Fusion Block (RFB), which employs a modulated cross-attention. Specifically, we use the degraded-domain feature $F_{\overline{x}}$ as the source for query and key, and the diffusion-domain feature $F_{\overline{y}}$ as the value:

$$F'_{\overline{n}} = \text{Conv2D}(\text{LayerNorm}(F_{\overline{n}})), \quad n \in \{x, y\}, \tag{6}$$

where the output of $F'_{\overline{x}}$ is expanded to twice the channel size of $F'_{\overline{y}}$, allowing splitting into Q and K. The chunk function is applied to obtain:

$$Q, K = \text{chunk}(F'_{\overline{x}}), \quad V = F'_{\overline{y}}$$

To reduce the domain discrepancy and guide the attention map computed from the degraded domain towards the structure of the diffusion domain, we apply a modulation on the similarity scores:

$$\text{Attn} = \left[w_1 \cdot \text{Softmax}(QK^\top) + w_2 \cdot \text{GeLU}(QK^\top)\right] \cdot V, \tag{7}$$

where w_1 and w_2 are learnable scalars that balance sharp and smooth attention responses, aligning the degraded-domain attention with the structure of the diffusion domain. Finally, the output is processed through a residual convolution and a feed-forward network to yield the fused feature:

$$F_f = \text{FFN}\left(\text{Conv2D}(\text{Attn}) + F_{\overline{x}}\right). \tag{8}$$

This fusion block rectifies the features of both domains, producing a unified feature that captures both the degradation context and structural priors.

Noise-Aware Routing Block. Standard decoders process all feature channels at every denoising step, even though degraded and clean images share substantial structural similarity and many channels may contain features already close to clean. This wastes computation and, in multi-degradation restoration, allows degradation-irrelevant features to interfere with reconstruction.

We propose a Noise-Aware Routing Block (NARB) that dynamically selects only the top-k most relevant channels at each step, conditioned on the current noise embedding 1. Given a feature map $F \in \mathbb{R}^{C\times H\times W}$, NARB first computes a compact channel descriptor f_c via global average pooling. This descriptor is passed through a noise-conditioned gating network to estimate per-channel relevance a. Given the selection ratio γ, we retain only the most relevant channels $F_{\text{select}} \in \mathbb{R}^{\gamma\cdot C\times H\times W}$, indexed by $\mathcal{I}_{\text{select}}$, where $\mathcal{I}_{\text{select}}$ is determined by selecting the Top-k channels with the highest relevance scores, with $k := \gamma \cdot C$. These selected channels are refined by residual blocks to produce F'_{select}, which are then reinserted into their original positions to form the output feature F_{out}.

Through selective channel refinement, NARB dynamically allocates denoising capacity to the most noise-affected

Table 1: Single-mode image restoration (blood removal, low-light image enhancement, and desmoking) and all-in-one image restoration performance comparison of our EndoIR against SOTA restoration methods on the SegSTRONG-C dataset. The all-in-one restoration methods are trained only once, while the original single-task methods are trained separately for each task.

Models	Params(M)	FPS	Blood			Low-Light			Smoke			Average		
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
PyDiff (Zhou, Yang, and Yang 2023)	97.9	5.86	29.77	82.80	0.1309	30.72	85.63	0.0993	30.36	83.10	0.1406	30.05	83.06	0.1296
LighTDiff (Chen et al. 2024)	<u>18.6</u>	**14.16**	30.75	84.93	0.0858	31.79	86.47	0.07526	30.71	84.79	0.0773	30.39	84.21	0.1011
EndoUIC (Bai et al. 2024)	25.8	3.42	31.15	84.19	0.0763	32.48	86.22	0.0644	31.01	85.98	0.0616	31.34	85.13	0.0694
LLCaps (Bai et al. 2023)	120.0	2.40	30.46	84.13	0.1280	31.00	85.17	0.1530	31.60	86.37	0.1128	31.26	85.62	0.1268
Diff-LOL (Jiang et al. 2023)	22.1	2.29	30.38	82.24	0.1024	28.90	80.16	0.1790	23.06	72.43	0.1647	27.41	78.57	0.1496
LA-Net (Yang et al. 2023)	**0.57**	2.05	28.14	81.26	0.1549	27.15	79.68	0.1974	29.06	82.57	0.1536	26.81	78.80	0.1943
PromptIR (Potlapalli et al. 2023)	33.0	5.28	29.72	84.96	0.0992	32.27	87.13	0.0741	31.88	87.61	0.0615	31.29	86.57	0.0783
PIP (Li et al. 2023)	26.8	5.42	30.02	83.37	0.0905	31.32	86.60	0.0782	31.48	86.19	0.0646	30.94	85.38	0.0778
AMIR (Yang et al. 2024)	23.54	2.80	31.02	83.53	0.0795	<u>32.71</u>	87.19	0.0868	32.09	86.84	0.0669	32.11	85.85	0.0777
MoCE-IR (Zamfir et al. 2025)	25.35	3.40	30.31	84.78	0.1797	32.19	86.96	0.2054	32.37	87.26	0.1509	31.62	86.33	0.1786
DFPIR (Tian et al. 2025)	29.63	2.20	30.57	85.01	0.0932	32.38	<u>87.94</u>	0.0768	32.32	87.40	0.0615	31.75	86.78	0.0771
EndoIR ($\gamma = 0.5$)	21.3	<u>11.18</u>	<u>31.33</u>	**86.76**	<u>0.0759</u>	32.64	**87.96**	**0.0679**	<u>32.71</u>	**88.20**	<u>0.0554</u>	<u>32.23</u>	**87.64**	<u>0.0664</u>
EndoIR ($\gamma = 1$)	28.2	8.14	**31.53**	85.96	**0.0640**	**32.82**	87.67	<u>0.0693</u>	**32.74**	87.51	**0.0551**	**32.36**	87.05	**0.0628**

Algorithm 1: **Noise-Aware Routing Block**

Input: $F_{\text{in}} \in \mathbb{R}^{C\times H\times W}$;
Selection ratio $\gamma \in (0, 1]$;
Number of ResBlocks N_{res}
Output: Refined feature $F_{\text{out}} \in \mathbb{R}^{C\times H\times W}$

$f_c \leftarrow \text{AdaptiveAvgPool}(F_{\text{in}})$
$a \leftarrow \text{Softmax}(\text{Linear}(\text{Linear}(f_c))) \in \mathbb{R}^C$
// Top-k Selection
$\mathcal{I}_{\text{select}} \leftarrow \text{Top}k(a, \gamma \cdot C)$
$F_{\text{select}} \leftarrow \text{Gather}(F_{\text{in}}, \mathcal{I}_{\text{select}})$
// Feature Refinement
for n in Number of ResBlocks N_{res} **do**
    $F'_{\text{select}} \leftarrow \text{ResBlock}_n(F_{\text{select}})$
end for
$F_{\text{out}} \leftarrow \text{Scatter}(F'_{\text{select}}, \mathcal{I}_{\text{select}}, F_{\text{in}})$
return F_{out}

dimensions, improving efficiency while preserving clean semantic features. This targeted strategy maintains overall restoration fidelity with reduced computation.

Experiments

Experiment Settings

Datasets.

All-in-one Restoration: SegSTRONG-C dataset (Ding et al. 2024) was initially developed for robust segmentation tasks under non-adversarial corruptions. The validation and test sets contain images with three categories of corrupt scenarios: blood, low-brightness, and smoke, in which each category comprises 900 images. We merge and divide these images in 3:1 ratio, resulting in 1350 images per category for the training set and 450 images per category for the test set.

Unified Illumination Correction: CEC dataset (Bai et al. 2024) is a capsule endoscopy illumination restoration dataset consisting of underexposed and overexposed images. The annotations were generated by professional photographers using the Adobe SDK. The dataset includes 800 training images and 200 testing images.

Implementation Details. We compare our method against SOTA all-in-one and specialized methods, as shown in Table 1. The all-in-one restoration methods are trained only **once**, while the single-task methods are trained **separately** for each task. The experiments are performed on NVIDIA A40 GPUs. Our model is trained for 100 epochs using Adam with a learning rate of 2×10^{-4} and a batch size of 8. The evaluation metrics include PSNR, SSIM, and LPIPS. Moreover, to address how well the task-adaptive embedding enhances feature discrimination across different restoration tasks, we involve Wilks' Lambda, a statistical metric that quantifies class separability. It can be demonstrated as a significant difference when the value is lower than 0.05.

We further explore the clinical applicability by performing downstream surgical tool segmentation on the enhanced images of the SegSTRONG-C dataset (Ding et al. 2024). The segmentation model is based on a pre-trained ResNet-101 (He et al. 2016) backbone and DeeplabV3+ (Chen et al. 2018) decoder, trained for 200 epochs on the SegSTRONG-C training set, and evaluated on the test set. We assess the segmentation results using mean Intersection over Union (mIoU), Dice score (Dice), and Accuracy (ACC).

Experimental Results

Restoration Results. Table 1 quantitatively compares the restoration performance of our EndoIR framework against a range of SOTA methods across the unified all-in-one restoration setting on the SegSTRONG-C dataset. Our EndoIR consistently outperforms prior advanced restoration models across all evaluation metrics. Notably, even compared to the strong medical restoration baseline AMIR (Yang et al. 2023), which exhibits competitive results in all subtasks, indicating that our method produces more perceptually faithful reconstructions. Specifically, EndoIR achieves a peak PSNR of 32.82 dB and SSIM of 87.67% on low-light enhancement, slightly surpassing AMIR and clearly outperforming illumination restoration approaches such as EndoUIC (Bai et al. 2024) and LLCaps (Bai et al. 2023). In the smoke and blood removal task, EndoIR also attains excellent structural preservation and perceptual similarity, showcasing the model's robustness in handling various degradation sources. Besides,

Table 2: Performance comparison across different models with evaluation metrics on CEC Dataset (Bai et al. 2024).

Models	PromptIR	PIP	PyDiff	LighTDiff	EndoUIC	LLCaps	Diff-LOL	LA-Net	AMIR	MoCE-IR	DFPIR	**EndoIR** ($\gamma = 0.5$)	**EndoIR** ($\gamma = 1$)
PSNR(dB) ↑	28.27	25.01	30.54	29.23	26.39	27.55	28.07	15.86	32.27	29.44	26.34	**32.65**	32.61
SSIM(%) ↑	83.14	70.09	77.07	96.14	95.64	85.95	96.13	66.40	96.43	93.27	84.17	96.98	**97.48**
LPIPS ↓	0.0717	0.1833	0.0796	0.0752	0.1079	0.2366	0.0883	0.4233	0.0512	0.1257	0.0795	0.0481	**0.0451**

Table 3: Downstream task performance across different models on SegSTRONG-C Dataset (Ding et al. 2024).

Models	PromptIR	PIP	PyDiff	LighTDiff	EndoUIC	LLCaps	Diff-LOL	LA-Net	AMIR	MoCE-IR	DFPIR	**EndoIR** ($\gamma = 0.5$)	**EndoIR** ($\gamma = 1$)
Dice ↑	95.09	94.97	95.53	95.03	91.38	92.63	92.59	91.31	95.13	95.35	95.47	95.54	**96.22**
mIoU ↑	90.41	90.56	91.79	90.43	84.29	84.20	84.96	84.19	90.78	90.80	91.39	**91.51**	91.40
ACC ↑	98.15	98.06	98.75	98.45	96.11	96.43	96.25	95.40	98.16	98.18	98.27	98.31	**98.97**

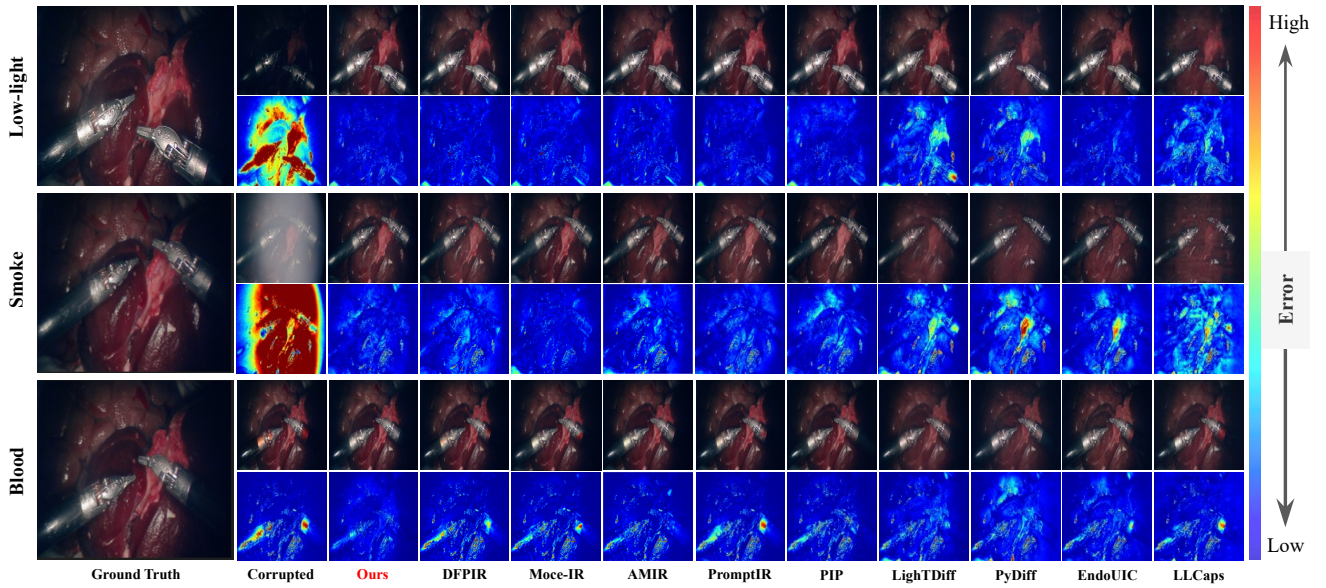


Figure 3: The quantitative visualizations and error maps on the SegSTRONG-C (Ding et al. 2024) dataset. Blue and red represent low and high error, respectively. (Zoom in to see details.)

EndoIR delivers the highest average performance, confirming its capability in capturing task-shared and task-specific features simultaneously. Our method demonstrates a clear edge in preserving fine details and perceptual quality under compound degradation scenarios. Figure 3 further supports our quantitative findings, where EndoIR exhibits the bluest error maps among all baselines, indicating the least pixel-wise deviation from the ground truth. We also evaluate EndoIR on a unified illumination correction benchmark (Table 2), where it achieves a significant PSNR improvement of 1.96 dB and a 0.68% SSIM gain over the SOTA, reaffirming the generalizability of our method across different datasets. These comprehensive results demonstrate the strong restoration capacity of EndoIR and highlight its potential for robust deployment in real clinical endoscopic applications.

Downstream Evaluation. Furthermore, downstream segmentation tasks are conducted to assess the effectiveness of EndoIR in joint medical image restoration. As presented in Table 3, EndoIR surpasses competing methodologies in surgical instrument segmentation, demonstrating its exceptional ability to preserve instrumental structures. It consistently outperforms all SOTA methods in mIoU and Dice scores (91.40/96.22), thereby emphasizing its advanced abilities in restoring images and preserving critical details.

Efficiency. Compared with other SOTA methods, our model also demonstrates strong efficiency in addition to the high restoration performance. As shown in Table 1, the FPS of our model is second only to LightDiff. Table 1 also demonstrates that EndoIR can achieve strong restoration performance with a relatively lower number of parameters, highlighting its excellent potential for real-world deployment.

Ablation Study

Components Ablation. Our ablation study in Table 5 further shows improvements introduced by our proposed modules. After involving the Dual-Stream Encoder, the performance greatly improved, while the Recified Fusion Block for fine-grained fuses the feature to further improve restoration quality. Our Task Adaptive Embedding provides task-related information for the frame and assists the final

Table 4: Ablation study on the Dual-Domain Prompter with employing frequency embedding F_{freq} and image embedding F_{img} on the SegSTRONG-C dataset.

F_{freq}	F_{img}	PSNR ↑	SSIM ↑	LPIPS ↓
×	×	31.19	85.98	0.0712
×	✓	32.20	86.44	0.0709
✓	×	32.11	86.85	0.0675
✓	✓	**32.36**	**87.05**	**0.0628**

Table 5: Ablation study on SegSTRONG-C (Ding et al. 2024). We remove (i) the Task Adaptive Embedding, (ii) the Dual-Stream Encoder, (iii) and the Rectified Fusion Block.

TAE	DSE	RFB	PSNR ↑	SSIM ↑	LPIPS ↓
×	×	×	30.39	84.21	0.1011
✓	×	×	31.40	85.09	0.0741
×	✓	×	31.96	86.83	0.0703
×	×	✓	31.56	86.03	0.0764
✓	✓	×	32.20	86.94	0.0649
✓	×	✓	31.51	86.22	0.0706
×	✓	✓	32.10	86.62	0.0654
✓	✓	✓	**32.36**	**87.05**	**0.0628**

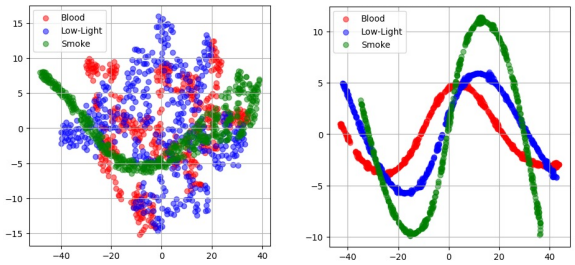
pipeline in gaining the best performance among all the evaluation metrics. Table 4 verifies the prompt effects and shows that the best performance occurs when both spatial and frequency domains are involved.

Frequency Transformation. We further conduct experiments to compare the effectiveness of different frequency transformations (FFT and Wavelet Transformation) for Dual-Domain Prompter. As shown in Table 6, FFT demonstrates superior performance, outperforming the Wavelet Transform by 0.40 dB in PSNR, 0.88% in SSIM, and 0.0048 in LPIPS. We attribute this improvement to the ability of FFT to capture global frequency patterns and provide more comprehensive frequency domain information, which is more beneficial compared to wavelet transformation.

Table 6: Ablation study on the Dual-Domain Prompter by employing different frequency transformation methods on the SegSTRONG-C dataset.

FFT	Wavelet	PSNR ↑	SSIM ↑	LPIPS ↓
–	✓	31.96	86.17	0.0676
✓	–	**32.36**	**87.05**	**0.0628**

Statistical Significance. Figure 4 shows the t-SNE visualization of images from different degradation types, before and after the Task Adaptive Embedding. It can be seen that the Task Adaptive Embedding clearly separates the features of the three types of degradations. In addition, Wilks' Lambda analysis (Rao 1951) reveals a significant difference of 40 times between inter-task discriminability ($\lambda = 0.0170$) and intra-task variability ($\lambda = 0.6907$). This large statistical separation shows that the embedding mechanism can preserve important task-specific information while suppressing task-irrelevant features. This ability is especially important in multi-task learning, where representations need to keep different task identities but still support knowledge sharing. Such controlled separation of features helps improve model generalization and reduces interference between tasks.



(a) (b)

Figure 4: (a) and (b) shows t-SNE visualization between different tasks: (a) Output feature after passing Dual-Domain Prompter; (b) Task Adaptive Embedding output condition.

Feature Selection Ratio. As shown in Fig. 5, we empirically chose 50% as the Top-k selection ratio to achieve better performance. Meanwhile, when 100% features are used, the performance drops significantly. This shows that not all features are noise-relevant, and using all of them leads to redundancy. It is better to select only a representative subset.

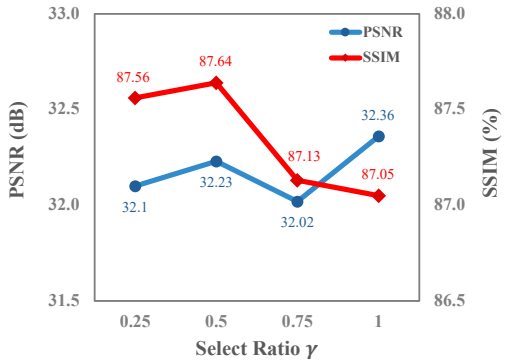


Figure 5: Ablation study on Noise-Aware Routing Block under different ratio γ on the SegSTRONG-C dataset.

Conclusion

In this work, we propose EndoIR, an all-in-one, degradation-agnostic diffusion framework for endoscopic image restoration. EndoIR integrates a Dual-Domain Prompter and task-adaptive embedding to provide robust conditioning, while a Dual-Stream Encoder and Rectified Fusion Block disentangle and integrate degraded and clean representations. Noise-Aware Routing Block further improves efficiency by selectively refining noise-relevant channels. EndoIR sets new benchmarks in endoscopic image restoration, with further tests in segmentation tasks underscoring its potential to enhance surgical precision. EndoIR shows promising performance across individual degradation types and frame-level

restoration, while real-world endoscopic scenes often involve compound degradations and require temporal coherence for reliable video analysis. Future efforts will aim to expand its capabilities to address multi-source compound degradations and incorporate temporal consistency for endoscopic video restoration.

References

Ai, Y.; Huang, H.; Zhou, X.; Wang, J.; and He, R. 2024. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25432–25444.

Bai, L.; Chen, T.; Tan, Q.; Nah, W. J.; Li, Y.; He, Z.; Yuan, S.; Chen, Z.; Wu, J.; Islam, M.; et al. 2024. Endouic: Promptable diffusion transformer for unified illumination correction in capsule endoscopy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 296–306. Springer.

Bai, L.; Chen, T.; Wu, Y.; Wang, A.; Islam, M.; and Ren, H. 2023. Llcaps: Learning to illuminate low-light capsule endoscopy with curved wavelet attention and reverse diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 34–44. Springer.

Chen, H.; Yang, Z.; Hou, H.; Zhang, H.; Wei, B.; Zhou, G.; and Xu, Y. 2025. All-in-One Medical Image Restoration with Latent Diffusion-Enhanced Vector-Quantized Codebook Prior. *arXiv preprint arXiv:2507.19874*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, T.; Lyu, Q.; Bai, L.; Guo, E.; Gao, H.; Yang, X.; Ren, H.; and Zhou, L. 2024. LighTDiff: Surgical endoscopic image low-light enhancement with T-diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 369–379. Springer.

Chen, W.; Liu, Y.; Hu, J.; and Yuan, Y. 2022. Dynamic Depth-Aware Network for Endoscopy Super-Resolution. *IEEE Journal of Biomedical and Health Informatics*, 26(10): 5189–5200.

Ding, H.; Lu, T.; Zhang, Y.; Liang, R.; Shu, H.; Seenivasan, L.; Long, Y.; Dou, Q.; Gao, C.; and Unberath, M. 2024. SegSTRONG-C: Segmenting Surgical Tools Robustly On Non-adversarial Generated Corruptions–An EndoVis' 24 Challenge. *arXiv preprint arXiv:2407.11906*.

García-Vega, A.; Espinosa, R.; Ramírez-Guzmán, L.; Bazin, T.; Falcón-Morales, L.; Ochoa-Ruiz, G.; Lamarque, D.; and Daul, C. 2023. Multi-scale structural-aware exposure correction for endoscopic imaging. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Guo, G.; Zhang, D.; Han, L.; Liu, N.; Cheng, M.-M.; and Han, J. 2024. Pixel Distillation: Cost-Flexible Distillation Across Image Sizes and Heterogeneous Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9536–9550.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770.

Jiang, H.; Luo, A.; Fan, H.; Han, S.; and Liu, S. 2023. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics*, 42(6): 1–14.

Jiang, Y.; Zhang, Z.; Xue, T.; and Gu, J. 2024. Autodir: Automatic all-in-one image restoration with latent diffusion. In *European Conference on Computer Vision*, 340–359. Springer.

Li, B.; Liu, X.; Hu, P.; Wu, Z.; Lv, J.; and Peng, X. 2022. All-in-One Image Restoration for Unknown Corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17452–17462.

Li, H.; Chen, X.; Dong, J.; Tang, J.; and Pan, J. 2025. Foundir: Unleashing million-scale training data to advance foundation models for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Li, H.; He, X.; Tao, D.; Tang, Y.; and Wang, R. 2018. Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. *Pattern Recognition*, 79: 130–146.

Li, Z.; Lei, Y.; Ma, C.; Zhang, J.; and Shan, H. 2023. Prompt-In-Prompt Learning for Universal Image Restoration. *arXiv preprint arXiv:2312.05038*.

Liu, T.; Chen, Z.; Li, Q.; Wang, Y.; Zhou, K.; Xie, W.; Fang, Y.; Zheng, K.; Zhao, Z.; Liu, S.; et al. 2023. MDA-SR: multi-level domain adaptation super-resolution for wireless capsule endoscopy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 518–527. Springer.

Luo, W.; Qin, H.; Chen, Z.; Wang, L.; Zheng, D.; Li, Y.; Liu, Y.; Li, B.; and Hu, W. 2025. Visual-Instructed Degradation Diffusion for All-in-One Image Restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12764–12777.

Ma, Y.; Liu, J.; Liu, Y.; Fu, H.; Hu, Y.; Cheng, J.; Qi, H.; Wu, Y.; Zhang, J.; and Zhao, Y. 2021. Structure and illumination constrained GAN for medical image enhancement. *IEEE Transactions on Medical Imaging*, 40(12): 3955–3967.

Pan, Y.; Bano, S.; Vasconcelos, F.; Park, H.; Jeong, T. T.; and Stoyanov, D. 2022. DeSmoke-LAP: improved unpaired image-to-image translation for desmoking in laparoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 17(5): 885–893.

Park, D.; Lee, B. H.; and Chun, S. Y. 2023. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5815–5824. IEEE.

Potlapalli, V.; Zamir, S. W.; Khan, S.; and Khan, F. S. 2023. PromptIR: Prompting for All-in-One Blind Image Restoration. *arXiv preprint arXiv:2306.13090*.

Qin, C.-J.; Wu, R.-Q.; Liu, Z.; Lin, X.; Guo, C.-L.; Park, H. H.; and Li, C. 2024. Restore anything with masks: Leveraging mask image modeling for blind all-in-one image restoration. In *European Conference on Computer Vision*, 364–380. Springer.

Ramírez, G.; Romero, A.; García-Vallejo, J. J.; and Munoz, M. 2002. Detection and removal of fat particles from post-operative salvaged blood in orthopedic surgery. *Transfusion*, 42(1): 66–75.

Rao, C. R. 1951. An asymptotic expansion of the distribution of Wilk's criterion. *Bulletin of the international statistical institute*, 33(2): 177–180.

Ren, Y.; Li, X.; Li, B.; Wang, X.; Guo, M.; Zhao, S.; Zhang, L.; and Chen, Z. 2024. Moe-diffir: Task-customized diffusion priors for universal compressed image restoration. In *European Conference on Computer Vision*, 116–134. Springer.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Su, X.; and Wu, Q. 2023. Multi-stages de-smoking model based on CycleGAN for surgical de-smoking. *International Journal of Machine Learning and Cybernetics*, 14(11): 3965–3978.

Tian, X.; Liao, X.; Liu, X.; Li, M.; and Ren, C. 2025. Degradation-Aware Feature Perturbation for All-in-One Image Restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28165–28175.

Wang, W.; Liu, F.; Hao, J.; Yu, X.; Zhang, B.; and Shi, C. 2024. Desmoking of the Endoscopic Surgery Images Based on A Local-Global U-Shaped Transformer Model. *IEEE Transactions on Medical Robotics and Bionics*.

Wu, R.; Zhang, Z.; Zhang, S.; Gou, L.; Chen, H.; Zhang, L.; Chen, H.; and Zuo, W. 2024. Self-supervised video desmoking for laparoscopic surgery. In *European Conference on Computer Vision*, 307–324. Springer.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13095–13105.

Yang, K.-F.; Cheng, C.; Zhao, S.-X.; Yan, H.-M.; Zhang, X.-S.; and Li, Y.-J. 2023. Learning to adapt to light. *International Journal of Computer Vision*, 131(4): 1022–1041.

Yang, Z.; Chen, H.; Qian, Z.; Yi, Y.; Zhang, H.; Zhao, D.; Wei, B.; and Xu, Y. 2024. All-in-one medical image restoration via task-adaptive routing. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 67–77. Springer.

Yang, Z.; Li, J.; Zhang, H.; Zhao, D.; Wei, B.; and Xu, Y. 2025. Restore-rwkv: Efficient and effective medical image restoration with rwkv. *IEEE Journal of Biomedical and Health Informatics*.

Zamfir, E.; Wu, Z.; Mehta, N.; Tan, Y.; Paudel, D. P.; Zhang, Y.; and Timofte, R. 2025. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12753–12763.

Zhou, D.; Yang, Z.; and Yang, Y. 2023. Pyramid Diffusion Models For Low-light Image Enhancement. *arXiv preprint arXiv:2305.10028*.