EndoUIC: Promptable Diffusion Transformer for Unified Illumination Correction in Capsule Endoscopy

Long Bai[1,2] *, Tong Chen[3] *, Qiaozhi Tan[2] *, Wan Jun Nah[2], Yanheng Li[4], Zhicheng He[2], Sishen Yuan[1], Zhen Chen[5], Jinlin Wu[5], Mobarakol Islam[6], Zhen Li[7], Hongbin Liu[5], and Hongliang Ren[1,2] **

[1] Department of Electronic Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong, China
[2] Shenzhen Research Institute, CUHK, Shenzhen, China
[3] The University of Sydney, Sydney, NSW, Australia
[4] City University of Hong Kong, Hong Kong, China
[5] Centre for Artificial Intelligence and Robotics, HKISI-CAS, Hong Kong, China
[6] University College London, London, UK
[7] Qilu Hospital of Shandong University, Jinan, China
b.long@link.cuhk.edu.hk, hlren@ee.cuhk.edu.hk

Abstract. Wireless Capsule Endoscopy (WCE) is highly valued for its non-invasive and painless approach, though its effectiveness is compromised by uneven illumination from hardware constraints and complex internal dynamics, leading to overexposed or underexposed images. While researchers have discussed the challenges of low-light enhancement in WCE, the issue of correcting for different exposure levels remains underexplored. To tackle this, we introduce EndoUIC, a WCE unified illumination correction solution using an end-to-end promptable diffusion transformer (DiT) model. In our work, the illumination prompt module shall navigate the model to adapt to different exposure levels and perform targeted image enhancement, in which the Adaptive Prompt Integration (API) and Global Prompt Scanner (GPS) modules shall further boost the concurrent representation learning between the prompt parameters and features. Besides, the U-shaped restoration DiT model shall capture the long-range dependencies and contextual information for unified illumination restoration. Moreover, we present a novel Capsule-endoscopy Exposure Correction (CEC) dataset, including ground-truth and corrupted image pairs annotated by expert photographers. Extensive experiments against a variety of state-of-the-art (SOTA) methods on four datasets showcase the effectiveness of our proposed method and components in WCE illumination restoration, and the additional downstream experiments further demonstrate its utility for clinical diagnosis and surgical assistance. The code and the proposed dataset are available at github.com/longbai1006/EndoUIC.

---

* Co-first authors.
** Corresponding author.

1   Introduction

Wireless Capsule Endoscopy (WCE) has revolutionized gastrointestinal (GI) diagnostics by offering a minimally invasive, painless way of examination in the GI tract [30]. However, the effectiveness of WCE can often be influenced due to factors such as limited battery capacity, camera performance, and the complexity of the GI tract [22]. Uneven illumination within the tract can significantly degrade image quality, thus affecting the accuracy and efficiency of diagnosis, screening, and the provision of timely feedback [28]. While the issue of low-light image enhancement (LLIE) in WCE images has received considerable attention, leading to various strategies to improve visibility in low-light areas [14,16], the challenge of overexposure remains less explored [26]. Various solutions [2,15,17,28] have been put forward to enhance low-light WCE images. Nevertheless, the complex and dynamic internal body environment will also result in overexposure, which obscures critical details with excessive brightness, as the brightness levels often extend beyond the dynamic range these techniques can adequately adjust [20].

Conventional approaches have been utilized to enhance the structure visibility in WCE images [20,24]. However, compared to deep learning methods, they tend to be less adaptive, less content-aware, and require manual intervention. Sequentially, García-Vega *et al.* [5] implemented a structure-aware deep neural network for exposure correction (EC), and employed CycleGAN [32] for EC dataset generation [4]. Presently, solutions for WCE unified illumination adaptation are still underexplored, lacking an end-to-end architecture that can unify illumination correction tasks. Furthermore, existing endoscopic EC datasets produced via generative models struggle to replicate the complexity encountered in real-world scenarios. This gap underscores the need for a unified light adaptation model capable of concurrently tackling over & underexposed images, which is crucial for the retention and enhancement of vital diagnostic details.

Denosing diffusion probabilistic models (DDPMs) have demonstrated outstanding performance in low-level vision tasks including denoising, super resolution, and low-light enhancement, owing to their ability to model complex data distributions and incorporate conditional information effectively [6,23]. In scenarios involving over & underexposed images, which typically demand different parameter spaces and optimization trajectories, directly training diffusion models might not be the best approach. Contrastive learning methods have already been employed to learn varying image degradation types, while an additional network would be needed [10]. To this end, we introduce a set of learnable parameters that act as our prompt. These prompt parameters are optimized through an end-to-end process, learning to adjust the model's prior for different image degradation. Then, it shall steer the model within the parameter space toward different low-level details essential for EC and LLIE. Thus, leveraging the task-specific knowledge acquired by the model, it dynamically adapts the input data according to different brightness levels. Additionally, the prompt module can act as an attention mechanism and increase the depth of the network. Therefore, even if the input exhibits only one type of degradation, the model can still maintain effective restoration performance. Moreover, to address the issue of data

scarcity, we have collected a WCE dataset and invited photography experts to annotate over & underexposed images manually. Specifically, our contributions to this work can be summarized as three-fold:

- We propose **EndoUIC** - **Endo**scopic **U**nified **I**llumination **C**orrection - a promptable diffusion model for unified WCE illumination correction. Specifically, the illumination prompt module is designed to navigate the diffusion model toward specific illumination conditions.
- In our proposed framework, we embed a diffusion process within a U-shape transformer to perceive global illumination and multiscale contextual information, and utilize prompts to guide the illumination restoration procedure. Our prompt module contains an Adaptive Prompt Integration (API) module, which dynamically produces and integrates prompt parameters with feature representations. Additionally, we incorporate the Global Prompt Scanner (GPS) module to enhance the interaction between prompts and features.
- To tackle the data shortage issue, we propose a novel WCE EC dataset, named **C**apsule-endoscopy **E**xposure **C**orrection (**CEC**) dataset, with normal and wrongly exposed image pairs. Extensive comparison, ablation, and downstream experiments on four datasets demonstrate the superior effectiveness of our EndoUIC, showcasing its potential in clinical applications.

2   Methodology

2.1   Preliminaries

Visual Prompt Learning introduces a set of learnable parameters that provide deep learning models with contextual information regarding the image degradation types in image restoration tasks [12, 18]. These prompts interact with the features of the input image, directing the model to adaptively adjust to different degradation types, thus restoring high-quality and clean images. This method enables a single unified model to address multiple image degradation challenges, enhancing the model's generalization capabilities.

Pyramid Diffusion Models (PyDiff) is an LLIE diffusion model that implements a pyramid diffusion strategy [31]. Unlike DDPMs, where image resolution remains constant throughout the reverse process, PyDiff starts with a lower resolution and progressively increases it to a higher resolution in the diffusion process. The forward and reverse process can be formulated with the given input $\mathbf{x}_0$, time step t, noise schedule $\{\alpha\}_{t=0}^{T}$, and scaling schedule $\{U\}_{t=0}^{T}$:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\left(\mathbf{x}_0 \downarrow_{U_t/U_{t-1}}\right), (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{1}$$

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) = \begin{cases} \mathcal{N}\left(\mathbf{x}_{t-1}; \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\mathbf{y}_\theta\left(\mathbf{x}_t\right) + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t, \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)\mathbf{I}\right), & \text{if } U_t = U_{t-1} \\ \mathcal{N}\left(\mathbf{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\left(\mathbf{y}_\theta\left(\mathbf{x}_t\right)\uparrow_{U_t/U_{t-1}}\right), (1-\bar{\alpha}_{t-1})\mathbf{I}\right), & \text{if } U_t > U_{t-1} \end{cases} \tag{2}$$
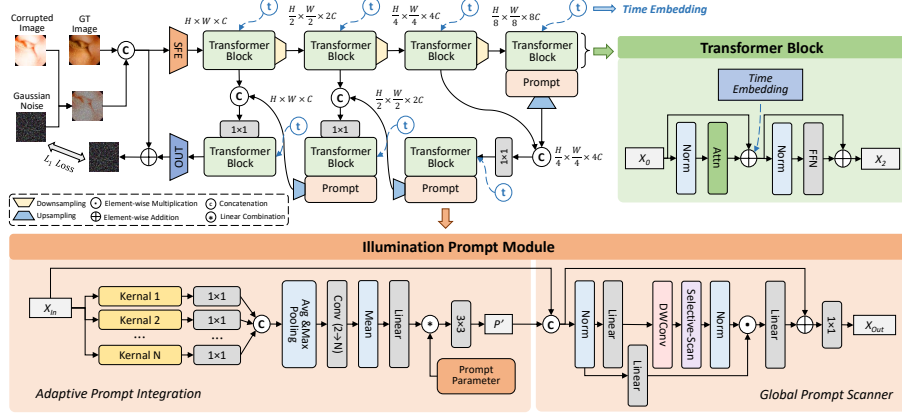
Fig. 1. The overview of our EndoUIC. Our network comprises the 4-level diffusion transformer (DiT), which is used to predict the noise. In each upsampling stage of the restoration DiT, the illumination prompt module is incorporated, which consists of the Adaptive Prompt Integration (API) and a Global Prompt Scanner (GPS) blocks. 'SFE' and 'OUT' denote the shallow feature extractor and the output block, respectively.

in which $\alpha_t \in (0,1)$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. While $a_t \geq a_{t+1}$ is getting bigger noise, $s_t \leq s_{t+1}$ is getting lower resolution. This approach optimizes the sampling speed and achieves improved image restoration quality by gradually refining image details with increasing resolution.

2.2   Proposed Method: EndoUIC

Our proposed EndoUIC framework is presented in Fig. 1, with the U-shape restoration DiT to estimate noises. The network is optimized with simple $\mathcal{L}_1$ loss and the noise sampling strategy follows [31].

Restoration Diffusion Transformer. Our network begins with a shallow feature extractor that transforms the image into the feature representation R. The features are then fed into the 4-level down-sampling transformer encoder and up-sampling transformer decoder, which is similar to the UNet structure [19]. The skip connections are executed at each level of the encoder-decoder. With $\mathcal{X}_0$ and $\mathcal{X}_2$ denoting the input and output feature respectively, each transformer block can be formulated as:

$$\mathcal{X}_1 = Attn(Norm(\mathcal{X}_0)), \quad \mathcal{X}_2 = FFN(Norm(\mathcal{X}_1)) \tag{3}$$

in which the time embedding t is injected with $\mathcal{X}_1$. FFN denotes the Feed-Forward Network. The decoder finally outputs a high-resolution image with normal illumination after propagating through the output block. Each up-sampling stage incorporates an illumination prompt module. Firstly, the output of the

prompt module is concatenated with the corresponding encoder's skip connection output. Then it will be passed through a 1×1 convolutional (Conv) layer before being input into the respective decoder.

Illumination Prompt Module. Our illumination prompt module shall utilize additional prompt parameters to encode key discriminative information about brightness degradation levels. The Adaptive Prompt Integration (API) block is used to learn discriminative illumination information, and the Global Prompt Scanner (GPS) integrates the learned prompt parameters with the image features, guiding the model's learning process. Given P as the learnable prompt parameters, $\mathcal{X}_{In}$ as the input features, and $\mathcal{X}_{Out}$ as the output features, the prompt module can be formulated:

$$P' = \mathcal{F}_{API}(\mathcal{X}_{In}; P), \quad \mathcal{X}_{Out} = \mathcal{F}_{GPS}(\mathcal{X}_{In}; P') \tag{4}$$

Adaptive Prompt Integration. The API module is designed to generate the prompt parameters and integrate them with the adaptively learned feature maps. We first define a set of learnable parameters P, which are designed to embed different illumination conditions into the features. This design can efficiently capture long-range dependencies to perceive global illumination information and address local-region uneven illumination conditions. Thus, our method can also effectively learn the illumination representation of features.

To achieve this, we refrain from directly multiplying P with the features as this could diminish the correlation between P and the features. Instead, we employ a multi-scale dynamic feature space for efficient feature learning. Specifically, we construct kernels of varying sizes to equip the model with different receptive fields [11], and fuse the feature representations using the 1×1 layer.

As illustrated in the left-bottom of Fig. 1, our dynamic kernel selection mechanism concatenates features from different receptive fields to obtain the combined feature $\mathcal{X}_{\mathcal{A}}$. We then utilize average and max pooling to extract spatial information from the feature space. The extracted features, after concatenation, are passed through a Conv layer, which expands the channel from 2 to $\mathbf{N}$:

$$\mathcal{X}'_A = \mathcal{F}_{2 \to \mathbf{N}}[AvgPool(\mathcal{X}_{\mathcal{A}}) \| MaxPool(\mathcal{X}_{\mathcal{A}})] \tag{5}$$

where $[\cdot \| \cdot]$ denotes concatenation. Subsequently, we apply the Sigmoid activation σ to obtain the weighted coefficient, which is then multiplied with P using the following equation, enabling dynamic feature learning to weight P adaptively:

$$P' = \mathcal{F}_{FCN}[Mean(\sigma(\mathcal{X}_A'))] \odot P \tag{6}$$

where $\odot$ denotes element-wise multiplication and $\mathcal{F}_{FCN}$ means the fully-connected layer. The obtained $Conv_{3 \times 3}(P')$ is then propagated through the GPS module for further correlation learning of the feature maps and prompt parameters.

Global Prompt Scanner. In the GPS module, the prompt parameter P' is firstly concatenated with the input feature $\mathcal{X}_{In}$, utilizing P' to guide the process of luminance restoration:

$$\mathcal{X}_P = [\mathcal{X}_{In} \| P'] \tag{7}$$
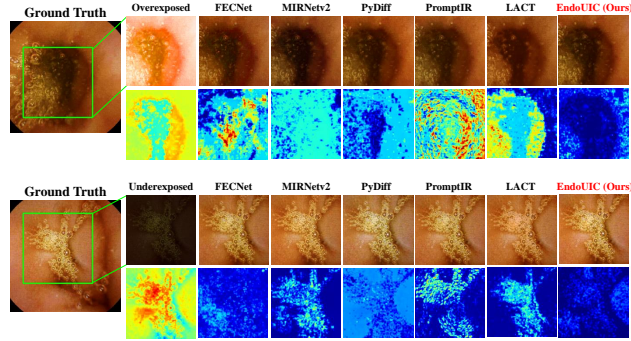
Fig. 2. The visualization results and error maps of our EndoUIC against SOTA methods on the CEC dataset. We present the enhanced images by SOTA exposure correction methods and their heat maps of reconstruction errors, with blue indicating lower errors and red denoting higher errors.

The selective-scan mechanism of VMamba [13], which captures long-range representations by scanning sequentially from four directions (top-left $\rightarrow$ bottom-right, bottom-right $\rightarrow$ top-left, top-right $\rightarrow$ bottom-left, bottom-left $\rightarrow$ top-right), has proven to be an effective approach for learning visual representations. To further enhance the global perception and foster the interaction between $\mathcal{X}_{In}$ and P', we conduct the cross-scan on $\mathcal{X}_P$. In this case, the scans in the same dimension as the concatenation can effectively facilitate the interaction between P' and $\mathcal{X}_{In}$, while the scans vertical to the concatenation dimension can promote the internal representation learning within P' and $\mathcal{X}_{In}$ themselves.

The GPS module is presented in the right-bottom of Fig. 1(c). Features that follow a skip connection are processed by a 1×1 Conv layer and a 3×3 Conv layer. The combined feature is then fused with features from higher spatial dimensions to facilitate the illumination restoration process of the overall model.

3 Experiments

3.1 Dataset

We conduct our experiments on two EC datasets and two LLIE datasets:

Capsule endoscopy Exposure Correction (CEC) dataset is collected by ANKON magnetically controlled WCEs of three patients. The training set includes 800 images from two patients, and the test set contains 200 images from another patient. The dataset comprises half overexposed and half underexposed images. We invite expert photographers to simulate the camera aperture settings with the Adobe Camera Raw SDK by adjusting exposure values (EVs). Changing the exposure value is equivalent to changing the camera aperture size. We render the raw image with varying numerical EVs to change the exposure range of highlights and shadows, emulating real exposure mistakes.

Endo4IE dataset is a public synthetic EC dataset of conventional endoscopy [4]. It was created by initially selecting public images without exposure issues. Then,

Table 1. EC comparison against existing and SOTA methods on our CEC dataset.

Methods	FECNet [8]	SID [7]	DRBN [7]	MIRv2 [29]	LLCaps [2]	PyDiff [31]	PromptIR [18]	LACT [1]	PIP [12]	EndoUIC
PSNR ↑	28.78	24.29	26.83	28.36	27.55	28.18	28.27	28.40	25.01	**29.65**
SSIM ↑	92.61	85.69	90.50	93.58	85.95	95.79	83.14	93.09	70.09	**96.80**
LPIPS ↓	0.1048	0.2111	0.1452	0.1080	0.2366	0.0941	0.0717	0.1103	0.1527	**0.0655**

Table 2. EC comparison against existing and SOTA methods on the Endo4IE dataset. '*' means we use the results from the previous works.

Methods	LMSPEC* [4]	LMSPEC+* [5]	FECNet [8]	MIRv2 [29]	LA-Net [26]	PyDiff [31]	PromptIR [18]	LACT [1]	PIP [12]	EndoUIC
PSNR ↑	23.97	23.62	24.72	23.85	23.51	24.73	23.73	22.92	25.28	**25.49**
SSIM ↑	80.34	79.97	81.84	82.33	83.78	84.78	79.57	76.88	81.94	**85.20**
LPIPS ↓	-	-	0.2031	0.2376	**0.1186**	0.2148	0.2396	0.2671	0.2150	0.1937

CycleGAN [32] was applied to generate paired over & underexposed synthetic images, and MSE and SSIM metrics were used to filter and finalize a dataset of 985 underexposed and 1231 overexposed images.

Kvasir-Capsule [21] and **Red Lesion Endoscopy (RLE)** [3] are originally two datasets utilized for WCE disease diagnosis. Bai *et al.* [2] have curated images from these datasets and synthesized two datasets specifically tailored for WCE LLIE by applying random Gamma correction and illumination reduction. Specifically, the Kvasir-Capsule dataset comprises 2000 training images and 400 test images. The RLE dataset contains 946 training images and 337 test images.

3.2 Implementation Details

The performance of our proposed EndoUIC is compared with a variety of state-of-the-art (SOTA) LLIE and EC methodologies, which are listed in Table 1, 2, 3, and supplementary materials. For methods marked with '*', we obtain results directly from previous works. For the remaining methods, we reproduce the results through their official repositories. We conduct our experiments with Python PyTorch on NVIDIA A100 GPUs. We train our model with Adam for 1000 epochs. The learning rate is set to 10^{-4}. We evaluate the image enhancement performance with Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). We also follow the previous work [2] to conduct a downstream medical diagnosis task on the RLE test set - the red lesion segmentation. The UNet [19] is trained using Adam with 20 epochs and a learning rate of 10^{-4}, and evaluated with mIoU.

3.3 Results

EC Comparison. As indicated in Tables 1 and 2, we initially perform the endoscopy exposure correction experiment in comparison with the existing SOTA methods. Our approach successfully surpasses various methods with different architectures (e.g., CNNs, Transformers, and DDPMs). Our method primarily restores images at the pixel level, leading to top-1 results in PSNR and SSIM. However, it falls behind LANet in feature perception quality, while our method

Table 3. LLIE comparison with existing and SOTA solutions on the Kvasir-Capsule [21] and RLE datasets [3]. The red lesion segmentation experiment is conducted on the RLE test set [3] by following the previous work [2]. LLCaps* means we use the results from the previous SOTA [2].

Methods		LLCaps* [2]	PIP [12]	CFWD [25]	Diff-LOL [9]	LA-Net [26]	CLE [27]	PyDiff [31]	PromptIR [18]	EndoUIC
Kvasir-Capsule	PSNR ↑	35.24	33.60	35.88	33.60	30.84	26.55	35.07	33.54	**36.85**
	SSIM ↑	96.34	95.09	96.26	95.42	95.32	87.87	96.60	96.77	**97.04**
	LPIPS ↓	0.0374	0.0302	0.0467	0.0847	0.0562	0.0829	0.0364	0.0377	**0.0255**
RLE	PSNR ↑	33.18	28.60	30.14	28.46	25.92	26.20	33.21	32.07	**33.50**
	SSIM ↑	93.34	87.27	90.25	82.52	85.72	81.42	93.54	93.30	**93.99**
	LPIPS ↓	0.0721	0.0977	0.1088	0.1437	0.1491	0.1134	0.0774	0.0694	**0.0658**
RLE Seg	mIoU ↑	66.47	59.46	51.47	62.46	52.57	45.33	62.56	59.92	**68.97**

Table 4. Ablation study of the proposed EndoUIC on the EndoUIC dataset. Specifically, we (i) replace the restoration DiT with the original U-Net architecture, (ii) remove the API block, and (iii) remove the GPS block.

Diffusion Trans		✗	✓	✗	✗	✓	✓	✗	✓
Prompt	API	✗	✗	✓	✗	✓	✗	✓	✓
	GPS	✗	✗	✗	✓	✗	✓	✓	✓
PSNR ↑		28.18	29.16	28.45	28.47	29.31	29.42	28.78	**29.65**
SSIM ↑		95.79	95.81	96.49	96.60	94.92	95.73	96.21	**96.80**
LPIPS ↓		0.0941	0.0735	0.858	0.0776	0.0710	0.0682	0.0727	**0.0655**

still ranks 2nd in LPIPS. Overall, our EndoUIC achieves the best results on our self-proposed CEC dataset and the public Endo4IE benchmark. The visualization of our image enhancement results and their corresponding error maps are presented in Fig. 2, where blue indicates fewer errors. It is observable that our method demonstrates the least errors and optimal results. Additionally, our EndoUIC demonstrates the FLOPs number of 14.36M and a standard deviation of 3.51 in PSNR on the CEC dataset. Future work will focus on enhancing the inference speed and robustness of our proposed framework.

LLIE Comparison. We also compare our method with SOTA LLIE techniques on two publicly available LLIE datasets, with the results presented in Table 3. We demonstrate that our method still achieves excellent results even when only one type of illumination degradation occurs. We also conduct a segmentation experiment on red lesions following LLCaps [2], and the superior results further prove the clinical applicability of our EndoUIC.

Ablation Study. In Table 4, we remove the API and GPS modules from the prompt architecture, and revert the DiT back to the UNet architecture. In all cases, we observed varying degrees of performance degradation, which further proves the effectiveness of the proposed components.

Feature Clustering. We also observe the feature clustering of over & underexposed images in the CEC test set with t-SNE. After passing through the 1st, 2nd, and 3rd prompt blocks, the feature clustering between over & underexposed images became more distinct and better clustered (Davies-Bouldin Index $1.68 \rightarrow 1.55 \rightarrow 0.53$). After removing the prompt blocks, we find worse clustering results ($2.24 \rightarrow 1.94 \rightarrow 1.87$), indicating that the prompt block can help optimize the feature clustering based on the discriminative illumination information.

4   Conclusion

This paper presents EndoUIC, a promptable DiT model for unified illumination correction for WCE. The model's ability to navigate through different regions of the parameter space allows for tailored adjustments that address the distinct challenges posed by either overexposed or underexposed images. Furthermore, with the assistance of photographer experts, we customize the CEC dataset tailored for the EC task in WCEs. Extensive experiments conducted on four datasets demonstrate that our EndoUIC surpasses existing SOTA techniques, validating its efficacy in performing endoscopic LLIE and EC tasks. Our proposed approach can be integrated with clinical endoscopy systems, greatly enhancing the visualization, diagnosis, screening, and treatment of GI diseases.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baek, J.H., Kim, D., Choi, S.M., Lee, H.j., Kim, H., Koh, Y.J.: Luminance-aware color transform for multiple exposure correction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6156–6165 (2023)
2. Bai, L., Chen, T., Wu, Y., Wang, A., Islam, M., Ren, H.: 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. pp. 34–44. Springer (2023)
3. Coelho, P., Pereira, A., Leite, A., Salgado, M., Cunha, A.: A deep learning approach for red lesions detection in video capsule endoscopies. In: Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15. pp. 553–561. Springer (2018)
4. García-Vega, A., Espinosa, R., Ochoa-Ruiz, G., Bazin, T., Falcón-Morales, L., Lamarque, D., Daul, C.: A novel hybrid endoscopic dataset for evaluating machine learning-based photometric image enhancement models. In: Mexican International Conference on Artificial Intelligence. pp. 267–281. Springer (2022)
5. García-Vega, A., Espinosa, R., Ramírez-Guzmán, L., Bazin, T., Falcón-Morales, L., Ochoa-Ruiz, G., Lamarque, D., Daul, C.: Multi-scale structural-aware exposure correction for endoscopic imaging. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
7. Huang, J., Liu, Y., Fu, X., Zhou, M., Wang, Y., Zhao, F., Xiong, Z.: Exposure normalization and compensation for multiple-exposure correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6043–6052 (2022)

8. Huang, J., Liu, Y., Zhao, F., Yan, K., Zhang, J., Huang, Y., Zhou, M., Xiong, Z.: Deep fourier-based exposure correction network with spatial-frequency interaction. In: European Conference on Computer Vision. pp. 163–180. Springer (2022)
9. Jiang, H., Luo, A., Fan, H., Han, S., Liu, S.: Low-light image enhancement with wavelet-based diffusion models. ACM Transactions on Graphics **42**(6), 1–14 (2023)
10. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17452–17462 (June 2022)
11. Li, Y., Hou, Q., Zheng, Z., Cheng, M.M., Yang, J., Li, X.: Large selective kernel network for remote sensing object detection. arXiv preprint arXiv:2303.09030 (2023)
12. Li, Z., Lei, Y., Ma, C., Zhang, J., Shan, H.: Prompt-in-prompt learning for universal image restoration. arXiv preprint arXiv:2312.05038 (2023)
13. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)
14. Long, M., Li, Z., Xie, X., Li, G., Wang, Z.: Adaptive image enhancement based on guide image and fraction-power transformation for wireless capsule endoscopy. IEEE transactions on biomedical circuits and systems **12**(5), 993–1003 (2018)
15. Ma, Y., Liu, Y., Cheng, J., Zheng, Y., Ghahremani, M., Chen, H., Liu, J., Zhao, Y.: Cycle structure and illumination constrained gan for medical image enhancement. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. pp. 667–677. Springer (2020)
16. Moghtaderi, S., Yaghoobian, O., Wahid, K.A., Lukong, K.E.: Endoscopic image enhancement: Wavelet transform and guided filter decomposition-based fusion approach. Journal of Imaging **10**(1), 28 (2024)
17. Mou, E., Wang, H., Yang, M., Cao, E., Chen, Y., Ran, C., Pang, Y.: Global and local enhancement of low-light endoscopic images (2023)
18. Potlapalli, V., Zamir, S.W., Khan, S., Khan, F.S.: Promptir: Prompting for all-in-one blind image restoration. arXiv preprint arXiv:2306.13090 (2023)
19. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)
20. Rukundo, O., Pedersen, M., Hovde, Ø., et al.: Advanced image enhancement method for distant vessels and structures in capsule endoscopy. Computational and mathematical methods in medicine **2017** (2017)
21. Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., et al.: Kvasir-capsule, a video capsule endoscopy dataset. Scientific Data **8**(1), 142 (2021)
22. Wang, G., Bai, L., Wu, Y., Chen, T., Ren, H.: Rethinking exemplars for continual semantic segmentation in endoscopy scenes: Entropy-based mini-batch pseudo-replay. Computers in Biology and Medicine **165**, 107412 (2023)
23. Wang, L., Yang, Q., Wang, C., Wang, W., Pan, J., Su, Z.: Learning a coarse-to-fine diffusion transformer for image restoration. arXiv preprint arXiv:2308.08730 (2023)
24. Wang, L., Wu, B., Wang, X., Zhu, Q., Xu, K.: Endoscopic image luminance enhancement based on the inverse square law for illuminance and retinex. International Journal of Medical Robotics and Computer Assisted Surgery **18**(4), e2396 (2022)

25. Xue, M., He, J., He, Y., Liu, Z., Wang, W., Zhou, M.: Low-light image enhancement via clip-fourier guided wavelet diffusion. arXiv preprint arXiv:2401.03788 (2024)
26. Yang, K.F., Cheng, C., Zhao, S.X., Yan, H.M., Zhang, X.S., Li, Y.J.: Learning to adapt to light. International Journal of Computer Vision **131**(4), 1022–1041 (2023)
27. Yin, Y., Xu, D., Tan, C., Liu, P., Zhao, Y., Wei, Y.: Cle diffusion: Controllable light enhancement diffusion model. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8145–8156 (2023)
28. Yue, G., Gao, J., Cong, R., Zhou, T., Li, L., Wang, T.: Deep pyramid network for low-light endoscopic image enhancement. IEEE Transactions on Circuits and Systems for Video Technology (2023)
29. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Learning enriched features for fast image restoration and enhancement. IEEE transactions on pattern analysis and machine intelligence **45**(2), 1934–1948 (2022)
30. Zhang, Y., Bai, L., Liu, L., Ren, H., Meng, M.Q.H.: Deep reinforcement learning-based control for stomach coverage scanning of wireless capsule endoscopy. In: 2022 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 01–06. IEEE (2022)
31. Zhou, D., Yang, Z., Yang, Y.: Pyramid diffusion models for low-light image enhancement. arXiv preprint arXiv:2305.10028 (2023)
32. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (2017)