S2DLight: Single Step Diffusion Network via Degradation-Aware Adaptation for Surgical Endoscopic Image Low-Light Enhancement

Song Fei[1], Guanyi Wang[1], Tian Ye[1], Sixiang Chen[1], Jianyu Lai[1], and Lei Zhu[1,2]

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] The Hong Kong University of Science and Technology
leizhu@hkust-gz.edu.cn
Code:https://github.com/FeiSong123/S2DLight

Abstract. Surgical endoscopy inherently involves constrained lighting conditions, which can obscure fine anatomical structures. Consequently, low-light image enhancement (LLIE) is crucial for improving visual clarity and structural discernibility. Recent diffusion-based methods improve restoration quality but are typically trained from scratch on small medical datasets, limiting their ability to exploit broader visual priors. Large-scale pretrained text-to-image (T2I) diffusion models encode powerful generative knowledge, yet directly applying them to surgical LLIE suffers from mismatched visual characteristics and high inference cost. We propose **S2DLight**, a single-step latent restoration framework built on a Stable Diffusion backbone for endoscopic LLIE. We introduce *Degradation-Aware Adaptation* to provide image-specific low-light guidance via contrastive degradation representations, and reformulate iterative sampling into one forward latent prediction to improve efficiency. To compensate for removing progressive refinement, we incorporate *Latent Structural Augmentation*, and further mitigate training–inference prompt mismatch using *Training-Time Semantic Alignment*. Experiments on public endoscopic benchmarks demonstrate that S2DLight achieves state-of-the-art performance while reducing the computational burden associated with large-scale diffusion inference.

Keywords: Endoscopic Imaging · Low-Light Image Enhancement · Single-Step Diffusion · Medical Image Processing

1 Introduction

Endoscopic imaging [4] is fundamental to modern minimally invasive surgery and diagnostic procedures, providing real-time visualization of internal anatomical structures. High-quality endoscopic imaging is critical for accurate tissue identification, precise surgical navigation, and safe intraoperative decision-making.

✉ Lei Zhu (leizhu@hkust-gz.edu.cn) is the corresponding author.

However, imaging within body cavities is inherently constrained by limited illumination and complex light scattering, often resulting in low contrast, severe noise, and attenuation of fine details [10,15]. Such degradation may obscure critical structures, increase operative difficulty, and elevate the risk of surgical complications [9]. Therefore, robust low-light image enhancement (LLIE) tailored for surgical endoscopy is of substantial clinical importance.

Early LLIE methods were primarily convolutional neural network (CNN)-based, such as SNR-Aware [27] and Zero-DCE++ [18]. Although effective in brightness correction, CNN pipelines are limited by local receptive fields and often oversmooth subtle textures under severe degradation. Transformer-based approaches were later introduced to capture long-range dependencies, but they generally require large-scale annotated data and substantial computational resources, which are often impractical in privacy-sensitive medical scenarios. More recently, diffusion-based image restoration methods [6,7,17,28,3,5,20,8] have demonstrated superior performance in fidelity. Nevertheless, due to privacy and ethical constraints, medical datasets are typically limited in scale. As a result, most existing approaches train diffusion models from scratch on relatively small datasets, restricting their ability to leverage large-scale visual priors.

Recently, large-scale pretrained text-to-image (T2I) diffusion models trained on web-scale image-text pairs encode rich visual priors and exhibit strong generative capability. This motivates a fundamental question: *Can large-scale pretrained T2I priors be effectively transferred to surgical endoscopic LLIE?* We use the diffusion backbone as a prior-transfer mechanism rather than as a sampling procedure. Addressing this question introduces two main challenges. First, pretrained T2I models are optimized for natural-image generation rather than medical restoration, resulting in discrepant visual patterns in anatomical structures and degradation characteristics. Second, directly adopting large-scale T2I diffusion backbones entails iterative multi-step denoising, introducing substantial computational overhead that complicates timely restoration in surgical scenarios.

To address these challenges, we propose **S2DLight**, a single-step diffusion framework for surgical endoscopic LLIE built on a Stable Diffusion 2.1 backbone [23]. We introduce *Degradation-Aware Adaptation*, where a contrastively trained degradation estimation network captures image-specific endoscopic low-light characteristics and injects degradation representations into the diffusion backbone to guide degradation-aware enhancement. We perform *Single-Step Reformulation* by converting iterative diffusion sampling into a single-step latent prediction, substantially reducing inference overhead. Since removing progressive refinement may cause a moderate performance drop, we further incorporate *Latent Structural Augmentation* through lightweight VAE encoder–decoder skip connections to reinforce structural preservation. Furthermore, since T2I backbones rely on textual conditioning for optimal performance, the difficulty of obtaining reliable captions for degraded endoscopic images in surgical low-light scenarios introduces a training–inference prompt mismatch. Inspired by the conditional control mechanism in [30], we propose *Training-Time Semantic Align-*

ment. During training, captions derived from corresponding normal-light images provide semantically accurate anatomical guidance, while a generic enhancement prompt is jointly used to expose the model to weak semantic conditioning. At inference, only the fixed enhancement prompt is applied, allowing the model to retain semantic prior utilization learned during training without relying on detailed caption input. Together, these components enable effective adaptation of large-scale pretrained T2I diffusion models to surgical LLIE. Experimental results on public endoscopic datasets demonstrate consistent improvement over state-of-the-art methods.

Our main contributions are summarized as follows:

- We present the first systematic exploration of adapting large-scale pretrained text-to-image diffusion models to surgical endoscopic LLIE, aiming to transfer pretrained visual priors rather than to perform iterative generation.
- We propose a unified framework integrating Degradation-Aware Adaptation, Single-Step Reformulation, Latent Structural Augmentation, and Training-Time Semantic Alignment, enabling effective adaptation of pretrained T2I diffusion models to endoscopic LLIE.
- Extensive experiments on public endoscopic LLIE datasets demonstrate that S2DLight consistently outperforms state-of-the-art methods.

2 Method

2.1 Preliminaries: Latent Diffusion

Latent diffusion models [23] operate in a compressed latent space defined by a pretrained VAE. Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, the encoder E maps it into a latent representation $\mathbf{z}_0 = E(\mathbf{x}) \in \mathbb{R}^{h\times w\times c}$, where the spatial resolution is reduced by a factor of 8. The forward diffusion process gradually corrupts $\mathbf{z}_0$ into a noisy latent $\mathbf{z}_t$ at timestep $t \in \{1, \dots, T\}$:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is determined by the noise schedule. A diffusion U-Net F_θ is trained to predict the noise component

$$\hat{\boldsymbol{\epsilon}} = F_\theta(\mathbf{z}_t; t), \tag{2}$$

and the clean latent can be reconstructed as

$$\hat{\mathbf{z}}_0 = \frac{\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\boldsymbol{\epsilon}}}{\sqrt{\bar{\alpha}_t}}. \tag{3}$$

In T2I diffusion models, the noise predictor is conditioned on textual embeddings $\mathbf{z}_y$, i.e., $F_\theta(\mathbf{z}_t; t, \mathbf{z}_y)$. Recent large-scale T2I diffusion models are trained on web-scale image–text pairs and have demonstrated remarkable generative capability, implicitly capturing rich visual priors from diverse natural images.

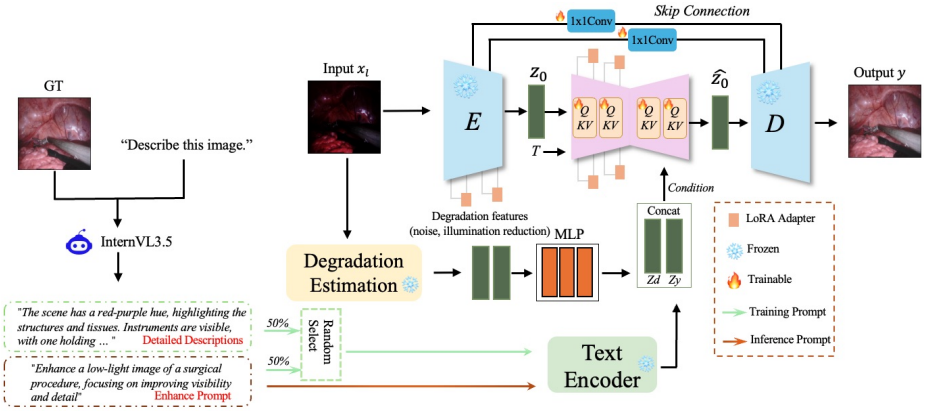

Fig. 1. Overall framework of S2DLight. The model integrates Degradation-Aware Adaptation, Single-Step Reformulation, Latent Structural Augmentation, and Training-Time Semantic Alignment within a Stable Diffusion backbone for surgical endoscopic low-light enhancement.

2.2 Proposed Methodology

Figure 1 illustrates the overall framework of S2DLight. Given a low-light endoscopic image $\mathbf{x}_l$, the VAE encoder produces a latent representation $\mathbf{z}_l = E(\mathbf{x}_l)$. A degradation estimation branch extracts degradation embeddings $\mathbf{z}_d$, while the text encoder produces semantic embeddings $\mathbf{z}_y$. Conditioned on these signals, the diffusion backbone predicts noise components in latent space, and the VAE decoder reconstructs the enhanced image from the estimated clean latent $\hat{\mathbf{z}}_0$.

Degradation-Aware Adaptation Pretrained T2I diffusion models are trained on natural-image distributions and do not explicitly model the severe underexposure and noise patterns observed in endoscopic imaging. Because degradation severity varies across frames, a fixed degradation code cannot adapt restoration strength to each input. We therefore condition the pretrained backbone on learned, image-specific degradation representations.

Contrastive degradation representation learning. As illustrated in Fig. 2, we train a lightweight degradation estimation network DENet$(\cdot)$ with a contrastive objective to capture low-light degradation characteristics. Two randomly cropped patches from the same degraded image form a positive pair because they share degradation statistics despite different anatomy, while patches from other images serve as negatives because degradation severity usually differs. Let d_i^1 and d_i^2 denote two degradation embeddings of sample i in a batch of size N. The contrastive loss is defined as follows:

$$\mathcal{L}_{\text{cl}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(d_i^1 \cdot d_i^2/\tau)}{\exp(d_i^1 \cdot d_i^2/\tau) + \sum_{j\neq i}\exp(d_i^1 \cdot d_j^1/\tau) + \sum_{j\neq i}\exp(d_i^1 \cdot d_j^2/\tau)}, \tag{4}$$

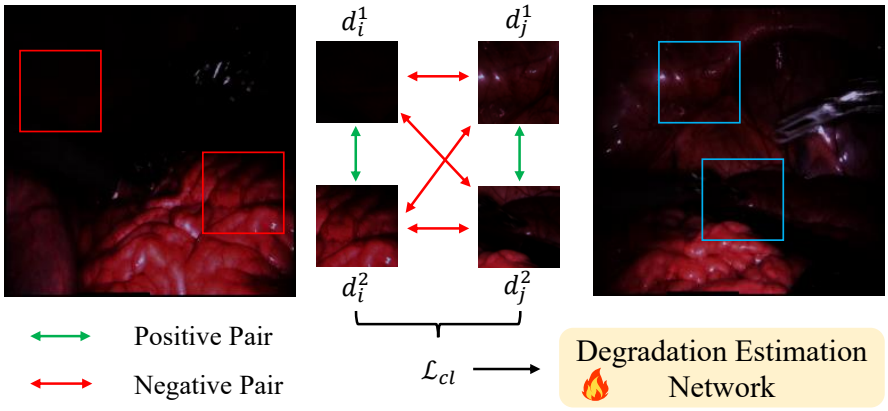

Fig. 2. Contrastive training scheme for the degradation estimation network (DENet). Two patches from the same low-light image form a positive pair, while patches from other images serve as negatives.

where τ is a temperature parameter. This objective emphasizes degradation cues over patch identity, so the diffusion backbone is guided by corruption level rather than local anatomy.

Degradation-conditioned diffusion. After pretraining, DENet extracts degradation embeddings $\mathbf{z}_d = \text{DENet}(\mathbf{x}_l)$, which are injected into the cross-attention layers of the diffusion U-Net (Fig. 1). Extending Eq. 2, the noise predictor is conditioned on both semantic and degradation embeddings, i.e., $\hat{\boldsymbol{\epsilon}} = F_\theta(\mathbf{z}_t; t, \mathbf{z}_y, \mathbf{z}_d)$. The clean latent $\hat{\mathbf{z}}_0$ is then reconstructed following Eq. 3. To preserve pretrained visual priors while enabling domain adaptation, LoRA modules are inserted into selected U-Net layers, and only the cross-attention and LoRA parameters are optimized, while the remaining backbone parameters remain frozen.

Single-Step Reformulation Standard diffusion models rely on iterative multi-step denoising, introducing substantial computational overhead. To advance the integration of large-scale pretrained diffusion models into surgical low-light enhancement, we reformulate the reverse diffusion process as a single-step latent prediction.

Inspired by recent findings that degraded inputs can serve as effective reverse diffusion initializations [29,19], we treat the encoded low-light latent $\mathbf{z}_l = E(\mathbf{x}_l)$ as an approximate noisy state at a fixed timestep T. Based on Eq. 3, we directly compute

$$\hat{\mathbf{z}}_0 = \frac{\mathbf{z}_l - \sqrt{1-\bar{\alpha}_T}\, F_\theta(\mathbf{z}_l; T, \mathbf{z}_y, \mathbf{z}_d)}{\sqrt{\bar{\alpha}_T}}. \tag{5}$$

This single-step reformulation removes progressive refinement while preserving the generative prior of the pretrained diffusion backbone. Unlike DDIM-style sampling [24], which follows a scheduled reverse trajectory across timesteps, S2DLight uses the encoded low-light latent as a degraded observation at a fixed timestep and predicts the restored latent in one pass.

Latent Structural Augmentation Removing progressive refinement may reduce high-frequency structural fidelity. To compensate for this effect, we incorporate *Latent Structural Augmentation* through lightweight VAE encoder–decoder skip connections, as shown in Fig. 1. Intermediate encoder features are projected through 1×1 convolutions and fused into corresponding decoder stages. These additional structural pathways propagate multi-scale information and reinforce structural preservation during reconstruction with negligible computational overhead.

Training-Time Semantic Alignment Pretrained T2I diffusion models are intrinsically text-conditioned and benefit from explicit semantic guidance. However, in surgical low-light scenarios, acquiring reliable captions for degraded endoscopic images is challenging, as illumination degradation obscures anatomical details and compromises the accuracy of both automatic caption generation and manual semantic description during intraoperative workflows. Training with descriptive captions while providing absent or unreliable semantic guidance at inference introduces a training–inference prompt mismatch that may impair enhancement performance.

Inspired by the conditional control mechanism in [30], we adopt *Training-Time Semantic Alignment*. During training, 50% of samples are conditioned on captions generated from corresponding normal-light images, providing semantically accurate anatomical guidance. The remaining samples use a generic enhancement prompt (e.g., *"enhance a low-light surgical endoscopic image"*), encouraging the model to learn enhancement behavior under weak semantic conditioning. At inference, only the fixed enhancement prompt is applied, enabling stable enhancement without requiring detailed caption input.

3 Experiments

Dataset and Implementation Details We evaluate S2DLight on EndoVis17 [2] and EndoVis18 [1]. EndoVis18 contains 2,400 images from 10 videos and EndoVis17 contains 2,235 images from 14 videos. Following [11], EndoVis18 is split into 1,800 training and 1,200 testing images, and EndoVis17 into 1,639 training and 596 testing images. All images are resized to 256×256. Low-light conditions are simulated following [7] using random gamma correction and illumination attenuation, consistent with prior LLIE studies [3,17,21]. This paired synthetic protocol enables controlled training and fair quantitative comparison, since well-aligned real low-light/normal-light surgical pairs are difficult to acquire during procedures. The simulated degradation is a proxy for exposure variation and attenuation. During training, images are padded to 320×320, randomly cropped to 256×256, and randomly flipped horizontally and vertically. The model is optimized with AdamW (learning rate 5×10^{-5}, batch size 1) using a weighted combination of L_1 and LPIPS losses (0.1 and 0.9, respectively). Training runs for 10,000 iterations on four NVIDIA RTX 4090 GPUs, and inference speed is measured on a single GPU.

Table 1. Performance comparison on EndoVis17 [2] and EndoVis18 [1]. Metrics include PSNR (↑), SSIM (↑), LPIPS (↓), and FPS (↑). For S2DLight, parameters are reported as total (trainable). The best results are shown in bold.

Models	FPS ↑	Parameter	EndoVis17[2]			EndoVis18[1]		
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-DCE++[18]	381.20	10.56K	12.65	0.3741	0.5046	11.24	0.3425	0.5321
RUAS[22]	197.39	3.44K	14.79	0.5979	0.3942	13.68	0.5612	0.4215
SNR-Aware[27]	24.20	39.12M	28.16	0.9103	0.1043	27.23	0.8995	0.1432
LLCaps[3]	4.48	119.83M	29.73	0.9386	0.1045	25.69	0.9122	0.1827
DiffLL[14]	6.52	20.76M	29.89	0.9349	0.0838	26.62	0.9226	0.1477
CLEDiff[28]	3.65	85.03M	13.85	0.3406	0.5355	10.76	0.2765	0.6468
LLFormer[25]	14.84	24.55M	30.24	0.9439	0.0750	26.83	0.9215	0.0987
PyDiff[32]	12.02	30.13M	28.98	0.9243	0.0953	27.67	0.9172	0.1047
ReDDiT[16]	14.96	17.43M	30.12	0.9310	0.0716	25.58	0.9198	0.0923
LLIEDiff[12]	0.01	859.52M	21.35	0.7520	0.3125	17.80	0.6945	0.3812
LighTDiff[7]	13.12	18.78M	30.47	0.9453	**0.0458**	26.94	0.9249	0.0861
S2DLight(4-Step)	1.39	0.86B(55M)	31.49	0.9615	0.0654	24.25	0.9423	0.1097
S2DLight(50-Step)	0.18	0.86B(55M)	29.86	0.9587	0.0752	24.57	0.9500	0.0908
S2DLight(Ours)	10.15	0.86B(55M)	**35.51**	**0.9681**	0.0495	**32.16**	**0.9791**	**0.0414**

Comparison With State-of-The-Art Methods Following the experimental protocol of LighTDiff [7], we compare S2DLight with representative state-of-the-art methods, including CNN-based Zero-DCE++ [18], RUAS [22], and SNR-Aware [27]; the transformer-based LLFormer [25]; and diffusion-based LLCaps [3], DiffLL [14], CLEDiffusion [28], ReDDiT [16], LLIEDiff [12], LighTDiff [7], and PyDiff [32]. For non-surgical LLIE methods, we retrain their official implementations on the same datasets for fair comparison. We also report ablation variants with 4-step and 50-step diffusion sampling. Performance is evaluated using PSNR [13], SSIM [26], and LPIPS [31], with runtime measured in FPS.

Quantitative Comparison Table 1 reports results on EndoVis17 and EndoVis18. On EndoVis17, S2DLight achieves 35.51 PSNR and 0.9681 SSIM, surpassing the strongest prior method (LighTDiff) by +5.04 dB in PSNR (30.47 → 35.51) and +0.0228 in SSIM (0.9453 → 0.9681), while maintaining competitive LPIPS (0.0495 vs. 0.0458). On EndoVis18, our method reaches 32.16 PSNR and 0.9791 SSIM, improving over the previous best PSNR of 27.67 by +4.49 dB and over the best SSIM of 0.9249 by +0.0497. LPIPS is reduced to 0.0414. Although built upon a larger diffusion backbone, S2DLight achieves competitive inference speed (10 FPS), surpassing several diffusion-based methods such as LLIEDiff [12], LLCaps [3], DiffLL [14], and CLEDiff [28], while remaining comparable to LighTDiff [7], PyDiff [32] and ReDDiT [16]. Compared with the 4-step and 50-step variants trained under the same setting, the proposed single-step formulation achieves the best overall accuracy while offering substantially higher FPS. This suggests that repeated refinement may introduce over-refinement or distribution drift.

Qualitative Comparison Fig. 3 presents representative visual comparisons with state-of-the-art methods. RUAS [22] improves overall brightness to some

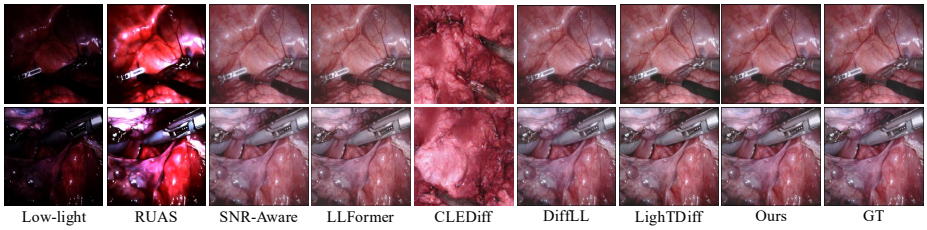

Fig. 3. Qualitative comparison of our method against state-of-the-art approaches on the EndoVis17 dataset[2], demonstrating superior performance in terms of visual fidelity and detail preservation.

Table 2. Ablation study on EndoVis17 [2].

Degradation-Aware Adaptation	Latent Structural Augmentation	Training-Time Semantic Alignment	PSNR↑	SSIM↑	LPIPS↓
×	×	×	15.82	0.4127	0.6821
✓	×	×	32.14	0.9581	0.0615
✓	✓	×	34.02	0.9647	0.0543
✓	✓	✓	**35.51**	**0.9681**	**0.0495**

extent, yet the restored results remain relatively dim in global tone. CLEDiff [28] produces severely degraded visual results under the same training budget. SNR-Aware [27], LLFormer [25], DiffLL [14], and LighTDiff [7] effectively enhance illumination, but subtle deviations in local structures and chromatic tone can still be observed. In contrast, S2DLight achieves higher consistency with the ground-truth references in both structural details and color fidelity. The gains in PSNR, SSIM, and LPIPS support improved fidelity and structure preservation.

Ablation Study Table 2 reports the ablation results on EndoVis17. Directly applying the pretrained diffusion backbone without any proposed modules yields limited performance (PSNR 15.82 / SSIM 0.4127). Introducing *Degradation-Aware Adaptation* increases PSNR by +16.32 dB to 32.14 and substantially improves structural similarity, highlighting the importance of modeling low-light degradation. Adding *Latent Structural Augmentation* further improves PSNR to 34.02 (+1.88 dB). Finally, incorporating *Training-Time Semantic Alignment* achieves the best performance (35.51 PSNR / 0.9681 SSIM / 0.0495 LPIPS), bringing an additional +1.49 dB gain. These results demonstrate the cumulative and complementary contributions of the proposed components.

Failure Case and Uncertainty Analysis Similar to previous methods, our approach may fail when the input surgical endoscopic image is extremely underexposed, where severe signal degradation leads to irreversible information loss and incomplete recovery of fine anatomical details. In addition, repeated inferences on selected test images show negligible variance across runs, indicating that the proposed framework produces stable and consistent restoration results with minimal uncertainty.

4 Conclusion

We presented S2DLight, a single-step latent restoration framework for surgical endoscopic low-light image enhancement built upon Stable Diffusion 2.1. The proposed framework transfers large-scale pretrained T2I priors to surgical LLIE through degradation-aware adaptation, reformulates iterative diffusion into one forward latent prediction for efficient inference, and compensates for reduced refinement via latent structural augmentation. In addition, training-time semantic alignment mitigates prompt inconsistency between training and inference. Extensive experiments demonstrate that S2DLight achieves superior enhancement quality with reduced computational demand on public paired benchmarks.

Limitations. The effectiveness of S2DLight is still influenced by the scale and diversity of available surgical low-light datasets, which remain limited. Our experiments use a synthetic paired low-light protocol for fair comparison, while validation on real clinical low-light videos and downstream tasks such as surgical instrument segmentation or recognition remain future work. We will further explore broader clinical data collection and unpaired or semi-supervised adaptation to improve generalization and clinical utility.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Project No.82572383), the Program of Guangdong Education Department, and AI Research and Learning Base of Urban Culture under Project 2023WZJD008.

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

References

1. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
3. 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)
4. Boese, A., Wex, C., Croner, R., Liehr, U.B., Wendler, J.J., Weigt, J., Walles, T., Vorwerk, U., Lohmann, C.H., Friebe, M., et al.: Endoscopic imaging technology today. Diagnostics **12**(5), 1262 (2022)
5. Chen, S., Ye, T., Lin, Y., Jin, Y., Yang, Y., Chen, H., Lai, J., Fei, S., Xing, Z., Tsung, F., et al.: Genhaze: Pioneering controllable one-step realistic haze generation for real-world dehazing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9194–9205 (2025)
6. Chen, S., Ye, T., Zhang, K., Xing, Z., Lin, Y., Zhu, L.: Teaching tailored to talent: Adverse weather restoration via prompt pool and depth-anything constraint. In: European Conference on Computer Vision. pp. 95–115. Springer (2024)

7. Chen, T., Lyu, Q., Bai, L., Guo, E., Gao, H., Yang, X., Ren, H., Zhou, L.: Lightdiff: Surgical endoscopic image low-light enhancement with t-diffusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 369–379. Springer (2024)
8. Fei, S., Ye, T., Wang, L., Zhu, L.: Lucidflux: Caption-free universal image restoration via a large-scale diffusion transformer. arXiv preprint arXiv:2509.22414 (2025)
9. 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 (2023). https://doi.org/10.1109/ISBI53787.2023.10230724
10. Gómez, P., Semmler, M., Schützenberger, A., Bohr, C., Döllinger, M.: Low-light image enhancement of high-speed endoscopic videos using a convolutional neural network. Medical & biological engineering & computing **57**, 1451–1463 (2019)
11. González, C., Bravo-Sánchez, L., Arbelaez, P.: Isinet: an instance-based approach for surgical instrument segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 595–605. Springer (2020)
12. Huang, Y., Liao, X., Liang, J., Quan, Y., Shi, B., Xu, Y.: Zero-shot low-light image enhancement via latent diffusion models. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25, AAAI Press (2025). https://doi.org/10.1609/aaai.v39i4.32398, https://doi.org/10.1609/aaai.v39i4.32398
13. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters **44**(13), 800–801 (2008)
14. Jiang, H., Luo, A., Fan, H., Han, S., Liu, S.: Low-light image enhancement with wavelet-based diffusion models. ACM Transactions on Graphics (TOG) **42**(6), 1–14 (2023)
15. Koohestani, F., Nabizadeh, Z., Karimi, N., Shirani, S., Samavi, S.: Brightvae: Luminosity enhancement in underexposed endoscopic images. arXiv preprint arXiv:2411.14663 (2024)
16. Lan, G., Ma, Q., Yang, Y., Wang, Z., Wang, D., Li, X., Zhao, B.: Efficient diffusion as low light enhancer. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 21277–21286 (June 2025)
17. Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.M., Gu, J., Loy, C.C.: Low-light image and video enhancement using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(12), 9396–9416 (2021)
18. Li, C., Guo, C.G., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). https://doi.org/10.1109/TPAMI.2021.3063604
19. Li, R., Zhou, Q., Guo, S., Zhang, J., Guo, J., Jiang, X., Shen, Y., Han, Z.: Dissecting arbitrary-scale super-resolution capability from pre-trained diffusion generative models (2023), https://arxiv.org/abs/2306.00714
20. Lin, Y., Ye, T., Chen, S., Fu, Z., Wang, Y., Chai, W., Xing, Z., Li, W., Zhu, L., Ding, X.: Aglldiff: Guiding diffusion models towards unsupervised training-free real-world low-light image enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5307–5315 (2025)
21. Lore, K.G., Akintayo, A., Sarkar, S.: Llnet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition **61**, 650–662 (2017)

22. Risheng, L., Long, M., Jiaao, Z., Xin, F., Zhongxuan, L.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
24. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
25. Wang, T., Zhang, K., Shen, T., Luo, W., Stenger, B., Lu, T.: Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2654–2662 (2023)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
27. Xu, X., Wang, R., Fu, C.W., Jia, J.: Snr-aware low-light image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17714–17724 (2022)
28. 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)
29. Yue, Z., Loy, C.C.: Difface: Blind face restoration with diffused error contraction. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9991–10004 (2024). https://doi.org/10.1109/TPAMI.2024.3432651
30. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
31. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
32. Zhou, D., Yang, Z., Yang, Y.: Pyramid diffusion models for low-light image enhancement. arXiv preprint arXiv:2305.10028 (2023)