






A Prior-Driven Lightweight Network for Endoscopic Exposure Correction

Zhijian Wu^{1,4} , Hong Wang^{2,†} , Yuxuan Shi³ , Dingjiang Huang⁴ , and Yefeng Zheng^{1,†} 

¹ Medical Artificial Intelligence Lab, Westlake University

² School of Life Science and Technology, Xi'an Jiaotong University

hongwang01@xjtu.edu.cn

³ ENT Institute and Department of Otorhinolaryngology, Eye and ENT Hospital, Fudan University

⁴ School of Data Sciences and Engineering, East China Normal University

zhengyefeng@westlake.edu.cn

[†] Corresponding Author

Abstract. Against this endoscopic exposure correction task, although some past studies have yielded promising results, these methods do not fully explore the task-specific priors, and they generally require a large number of parameters thus compromising their applications on resource-constrained devices. In this paper, we carefully explore that regardless of the exposure level degradation, the illumination information is usually contained in the low frequency part, and the relative smoothness of structures in captured endoscopic images generally lead to the sparse high-frequency representation. Motivated by such prior understandings, we specifically construct a lightweight wavelet transform-based hierarchical network structure for this correction task, called WTNet, which utilizes the inherent frequency decomposition characteristics of wavelet transform and makes the core of network learning focus on the modelling of low-frequency information. Based on four datasets and three different tasks, including exposure correction, low-light enhancement, and downstream segmentation, we comprehensively substantiate the superiority of our proposed WTNet. With only 1.41M model parameters, our WTNet achieves a better balance between performance and cost, and demonstrates favorable clinical application potential. The code will be available at <https://github.com/charonf/WTNet>.

Keywords: Endoscopic Exposure Correction · Lightweight · Prior

1 Introduction

Wireless capsule endoscopy (WCE) is becoming a favorable alternative to gastrointestinal (GI) examinations due to its non-invasive and painless advantages over traditional endoscopy [29]. However, the limitations of sensor hardware, coupled with the intricate internal structure of the GI tract, often lead to the issues of underexposure and overexposure in the captured images, which can

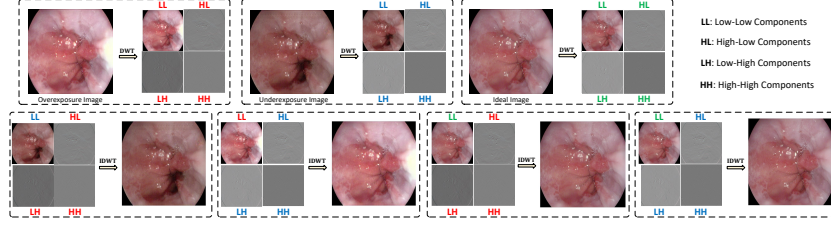


Fig. 1. *Top:* We perform wavelet transform on endoscopic images with different exposure levels and get four components, including low-frequency LL, and another three high-frequencies LH, HL, and HH. *Bottom:* Take the first virtual frame as an example, when performing inverse wavelet transform on the four components (LL of underexposure image, and HL, LH, HH of overexposure image), the image presents an underexposure effect. This suggests the illumination mainly exists in low-frequency.

adversely affect the subsequent diagnosis and treatment planning [25,15]. How to effectively enhance the quality of inappropriately exposed WCE images is gradually attracting the attention of the research community [3].

Against this exposure correction task, in recent years, with the rapid development of deep learning, diverse deep neural network structures have been constructed. For example, Gomez et al. [8] devised a multi-scale structure-aware network for laryngoscopic low-light enhancement while introducing adversarial loss to ensure the authenticity and realism of the enhanced images. Ma et al. [12] introduced a cyclic structure and illumination-constrained generative adversarial network for medical image enhancement. This innovative approach computed local structures and illumination constraints, enabling the model to comprehensively learn both the overall features and fine-grained local details of medical images. Very recently, drawing inspiration from the great success of denoising diffusion probabilistic models in learning data distribution, Bai et al. [3,2] developed different diffusion-based exposure correction methods tailored for WCE images. As seen, most of the existing exposure correction research techniques generally put more emphasis on designing various and complicated network structures for better restoration performance, but neglect the full exploration and the efficient incorporation of the inherent prior characteristics underlying the task. Besides, although the diffusion-based enhancement methods have achieved better performance, they typically require a substantial number of network parameters and involve multiple sampling steps, posing huge challenges to practical deployment.

Against the aforementioned limitations, in this paper, we carefully investigate the intrinsic prior property of endoscopic images under different exposure conditions and specifically construct a lightweight network framework for this exposure correction task. Specifically, as shown in the top row of Fig. 1, based on the classic discrete wavelet transform (DWT), we first analyze the frequency characteristics of endoscopic images with different lighting conditions, and observe that the endoscopic exposure correction task has specific prior properties: 1) The illumination information mainly reflects in the low-frequency component

(*i.e.*, LL); 2) Due to the smoothness of anatomical structures of endoscopic images [6], other high-frequency-involved components (*i.e.*, LH, HL, and HH) are generally relatively sparse. When we only exchange the low-frequency components between ill-exposed and ideally-exposed images, the lighting conditions can be finely transferred, as presented in the bottom row of Fig. 1. These inherent prior investigations show that in order to enhance the ill-exposed endoscopic images, we can make the core of network learning mainly focus on the restoration of low-frequency components without spending too much computational overhead on the recovery of other frequency components. To this end, motivated by the excellent advantages of DWT in decoupling different frequency components, and achieving the downsampling and upsampling for multi-scale learning without introducing extra parameters, we meticulously devise a wavelet transform-based hierarchical exposure correction network, called WTNNet. Specifically, as presented in Fig. 2, at the encoder stage, the deep features are hierarchically decomposed into low-frequency components and three different high-frequency-related components via DWT, which are then modeled based on the powerful Transformer block and the simple depth-wise convolution layer, respectively. At the decoder stage, it utilizes the inverse DWT to reconstruct the modeled features at the same level into a high-resolution feature which is treated as the input at the next level. Such careful designs not only naturally enable multi-order frequency decomposition and multi-scale network learning, but also largely reduce the parametric cost. Our main contributions are three-fold:

- We carefully explore the inherent prior characteristics underlying this endoscopic exposure correction task, and then construct a wavelet transform-based hierarchical network framework, called WTNNet.
- Considering the complicated degradation process of poorly exposed images, we specifically devise an exposure correction module for the coarse illumination calibration, which would further boost the subsequent multi-order frequency decomposition and multi-scale learning.
- Based on four publicly available datasets spanning three different tasks including exposure correction, low-light enhancement, and downstream segmentation, extensive experiments substantiate that our WTNNet can always achieve excellent performance with only 1.41M parameters.

2 Methodology

Attributed to the powerful capability in modeling the multi-scale contextual information, the UNet-like hierarchical network structure [14] has been widely adopted by different image restoration methods [20,19,2]. Among these existing techniques, during the encoder stage, for every level, most of them typically utilize the convolution operation to downsample the deep features while doubling the number of channels in order to mitigate the information loss. Clearly, for the hierarchical structure with multiple levels, successive downsampling procedures will cause the number of feature channels to increase exponentially, thus leading

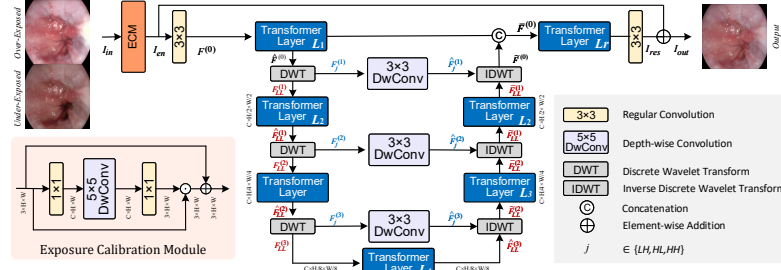


Fig. 2. The overall architecture of our proposed WTNet.

to a dramatic growth of the model parameters. Besides, they aim to construct complicated network modules for better performance without fully analyzing the inherent prior characteristics of the endoscopic exposure correction task.

Based on the prior observations from Fig. 1 as analyzed in Sec. 1, we find that for the endoscopic exposure correction task, our main goal should be reconstructing the favorable low-frequency component. To this end, we propose to utilize the wavelet transform to decouple different frequency components, and specifically construct a novel wavelet transform-based lightweight hierarchical framework, called WTNet. In the proposed method, the modeling procedures of low-frequency components and other frequency components are implemented through high-complexity and low-complexity operators (*i.e.*, a Transformer block and a depth-wise convolution layer), respectively. In such a manner, the core of network learning can be forced to mainly focus on the recovery of the low-frequency component which complies with the prior property well as stated in Sec. 1. The overall pipeline of the proposed WTNet is presented in Fig. 2.

Specifically, given an incorrectly exposed endoscopic image $I_{in} \in \mathbb{R}^{3 \times H \times W}$, considering that the exposure degradation process is extremely complicated due to improper techniques by operators and limited acquisition space, we first design an exposure calibration module (ECM) to perform a coarse correction of the lighting conditions for boosting the subsequent hierarchical modeling procedure. As shown in Fig. 2, the ECM first projects the input image to a high-dimensional feature space using a 1×1 convolution, which is followed by a 5×5 depth-wise convolution to encode the spatial context to model the lighting conditions in different regions. Next, the features are aggregated using a 1×1 convolution to estimate a light map $L \in \mathbb{R}^{3 \times H \times W}$ pixel by pixel. Finally, the coarsely enhanced image $I_{en} \in \mathbb{R}^{3 \times H \times W}$ is obtained by the element-wise multiplication between the original input image I_{in} and the light map L . Mathematically, the concrete computation procedure of ECM is expressed as:

$$\begin{aligned} L &= W_{1 \times 1}^{C \rightarrow 3} \otimes (W_{5 \times 5} \otimes_d (W_{1 \times 1}^{3 \rightarrow C} \otimes I_{in})), \\ I_{en} &= I_{in} \odot L, \end{aligned} \quad (1)$$

where \otimes denotes the conventional convolution; \otimes_d denotes the depth-wise convolution; $W_{1 \times 1}^{C \rightarrow 3}$ and $W_{1 \times 1}^{3 \rightarrow C}$ are two 1×1 convolutions carrying out channel

contraction and channel expansion, respectively; $W_{5 \times 5}$ is a 5×5 convolutional kernel; and \odot is element-wise multiplication.

By feeding I_{en} into a 3×3 convolutional layer, we can obtain an initial shallow feature $F^{(0)} \in \mathbb{R}^{C \times H \times W}$. Then by further modeling this feature through Transformer blocks, we can get the deep feature $\hat{F}^{(0)} \in \mathbb{R}^{C \times H \times W}$, which is processed through a three-level symmetric encoder-decoder with a bottleneck block. As shown in Fig. 2, at the encoder stage, the feature $\hat{F}^{(i-1)} \in \mathbb{R}^{C \times \frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}}$ at the level i ($i = 1, 2, 3$) is decomposed into four sub-bands via DWT, including low-frequency component $F_{LL}^{(i)}$, and three high-frequency-related components $F_{LH}^{(i)}$, $F_{HL}^{(i)}$, and $F_{HH}^{(i)}$, which have the same size of $C \times \frac{H}{2^i} \times \frac{W}{2^i}$. Then motivated by the prior analysis that the lighting information mainly presents in the low-frequency part, we propose to further model $F_{LL}^{(i)}$ and the other three components via a long-range Operator(\cdot) (*e.g.*, Transformer or Mamaba) and the low-complexity depth-wise convolutional layer, respectively. The powerful long-range modeling capability of the Operator(\cdot) can better guarantee the learning of low-frequency illumination content, and the structure-aware ability of depth-wise convolution can better help maintain high-frequency texture details with lower computational cost. Through this divide-and-conquer customized learning strategy, the synergy between different components can be better exerted, thereby achieving full feature modeling. For level i , the computation is formulated as:

$$\begin{aligned} [F_{LL}^{(i)}, F_{LH}^{(i)}, F_{HL}^{(i)}, F_{HH}^{(i)}] &= \text{DWT}(\hat{F}_{LL}^{(i-1)}), \\ \hat{F}_{LL}^{(i)} &= \text{Operator}(F_{LL}^{(i)}), \\ \hat{F}_j^{(i)} &= W_{3 \times 3} \otimes_d F_j^{(i)}, j \in \{LH, HL, HH\}, \end{aligned} \quad (2)$$

where $i = 1, 2, 3$; $\hat{F}_{LL}^{(0)} = \hat{F}^{(0)}$. For Operator(\cdot), in experiments, we adopt the Transformer block proposed in Restormer [26] to reduce the complexity. $\hat{F}_{LL}^{(i)} \in \mathbb{R}^{C \times \frac{H}{2^i} \times \frac{W}{2^i}}$ and $\hat{F}_j^{(i)} \in \mathbb{R}^{C \times \frac{H}{2^i} \times \frac{W}{2^i}}$ are four sub-bands at level i .

In the decoder stage, starting from the bottleneck layer, through a simple inverse DWT (IDWT), we can reconstruct the high-resolution feature from the modeled four sub-bands at the same level. Considering that the IDWT operation is non-parametric, at the decoder stage, we also introduce the Transformer layer to further optimize the low-frequency sub-band for better learning of high-resolution features. Mathematically, the concrete formulation is:

$$\begin{aligned} \tilde{F}_{LL}^{(k-1)} &= \text{IDWT}(\bar{F}_{LL}^{(k)}, \{\hat{F}_j^{(k)}\}_{j \in \{LH, HL, HH\}}), \\ \bar{F}_{LL}^{(k-1)} &= \text{Transformer}(\tilde{F}_{LL}^{(k-1)}), \end{aligned} \quad (3)$$

where $k = 3, 2, 1$; $\bar{F}_{LL}^{(3)} = \hat{F}_{LL}^{(3)}$; $\tilde{F}_{LL}^{(k)} \in \mathbb{R}^{C \times \frac{H}{2^k} \times \frac{W}{2^k}}$; and $\bar{F}_{LL}^{(k)} \in \mathbb{R}^{C \times \frac{H}{2^k} \times \frac{W}{2^k}}$. Please note that $\bar{F}_{LL}^{(0)} = \text{Concat}(\hat{F}_{LL}^{(0)}, \tilde{F}_{LL}^{(0)}) \triangleq \text{Concat}(\hat{F}^{(0)}, \tilde{F}^{(0)}) \triangleq \bar{F}^{(0)}$. Here Concat(\cdot) denotes the concatenation operation along the channel dimension. Finally, by feeding $\bar{F}^{(0)}$ into an extra Transformer layer for refinement followed by a 3×3 convolutional layer, we can obtain the residual image $I_{res} \in \mathbb{R}^{3 \times H \times W}$.

Table 1. Comparisons on the exposure correction task based on CEC and Endo4IE.

Methods	Ref.	CEC			Endo4IE			Params
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	
MIRNetv2 [27]	TPAMI 22	28.36	93.58	0.1080	23.85	82.33	0.2376	5.89M
LLCaps [3]	MICCAI 23	27.55	85.95	0.2366	-	-	-	119.73M
LA-Net [23]	IJCV 23	-	-	-	23.51	83.78	0.1186	0.56M
PyDiff [30]	IJCAI 23	28.18	95.79	0.0941	24.73	84.78	0.2148	32.00M
PromptIR [13]	NIPS 23	28.27	83.14	0.0717	23.73	79.57	0.2396	32.96M
LACT [1]	ICCV 23	28.40	93.09	0.1103	22.92	76.88	0.2671	-
PIP [11]	Arxiv 23	25.01	70.09	0.1527	25.28	81.94	0.2150	26.82M
Retinexformer [4]	ICCV 23	37.18	<u>97.22</u>	<u>0.0356</u>	27.22	<u>85.73</u>	0.2092	1.61M
MambaLLIE [18]	NIPS 24	33.92	95.56	0.0548	<u>27.43</u>	85.56	0.2132	2.28M
EndoUIC [2]	MICCAI 24	29.65	96.80	0.0655	25.49	85.20	<u>0.1937</u>	28.91M
WTNet (Ours)	-	<u>36.82</u>	97.43	0.0343	27.91	86.14	0.1954	1.41M

As seen, in our WTNet, low and high frequencies are finely decoupled through the DWT, which are separately processed in a multi-scale hierarchical structure. Most of the computational overheads occur only in the low-dimensional low-frequency components. This not only complies with our design motivation but also helps reduce the model complexity.

3 Experiments

3.1 Experimental Setup

Datasets. Following [2], we comprehensively evaluate our method on two exposure correction datasets and two low-light image enhancement (LLIE) datasets: **1) Capsule Endoscopy Exposure Correction (CEC) [2].** This dataset contains an equal number of overexposed and underexposed images. The training set consists of 800 images, while the test set includes 200 images. **2) Endo4IE [7].** This dataset consists of 956 underexposed and 1194 overexposed image pairs, of which 1552 images are used for training while 598 images are used for testing. **3) Kvasir-Capsule (KC) [16] and Red Lesion Endoscopy (RLE) [5].** These two datasets are composed of low-light images. The former contains 2000 training images and 400 testing images, and the latter contains 946 training images and 337 test images. Besides, we further conduct a downstream segmentation task based on the RLE dataset with pixel-wise annotations to evaluate the effectiveness of our method in clinical applications.

Evaluation Metrics. Following [3,2], the peak signal-to-noise ratio (PSNR) [9], the structural similarity index measure (SSIM) [17], and Learned Perceptual Image Patch Similarity (LPIPS) [28] are used for quantitative evaluation. For the downstream segmentation task, we follow [3] and utilize the mIoU metric.

Implementation Details. Our WTNet consists of four layers of encoder-decoder. The number of Transformer blocks is [4, 7, 7, 8] from Level 1 to Level 4, and the number of channels is [40, 40, 40, 40]. We extract patches with 128×128 pixels from training images, and the batch size is set to 8. The Adam optimizer

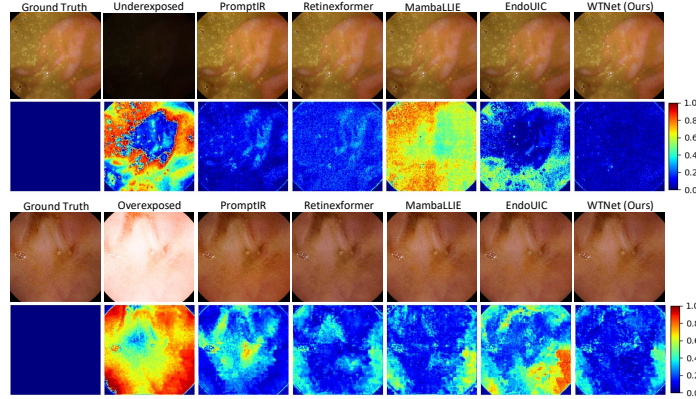


Fig. 3. Comparisons on two different images randomly selected from the CEC testing set. For every method, the upper row is the enhanced result and the lower row is the error map between the enhanced image and the ground truth.

with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is used. The initial learning rate is set to 2×10^{-4} and steadily decreased to 1×10^{-6} by the cosine annealing scheme. The model is trained in a total of $150k$ iterations. The experiments are conducted in the PyTorch framework with four NVIDIA 4090 GPUs. For the downstream segmentation task, we train the UNet using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, while the learning rate is set to 2×10^{-4} and the number of the total epochs is 100 with the batch size of 12. A combination of cross-entropy loss and Dice loss is utilized to supervise the segmentation training process.

3.2 Comparisons on the Exposure Correction Task

We first execute the evaluation on the exposure correction task. Table 1 reports the quantitative results of different state-of-the-art methods on these two datasets, including CEC and Endo4IE. As observed, our WtNet always performs competitively across all different metrics on both datasets. Fig. 3 shows the enhanced results and corresponding error maps of different methods on the CEC dataset. It clearly shows that the results recovered by our method under different lighting conditions are closer the ground truth images. It is worth mentioning that our method only requires 1.41M, which is extremely lightweight. These results show that the proposed WtNet makes a better trade-off between performance and cost.

3.3 Comparisons on the Low-Light Image Enhancement Task

To further evaluate the potential of our method, we add the comparisons on the LLIE task based on KC and RLE datasets. Table 2 shows that our WtNet achieves the best performance on almost all metrics on these two datasets, which

Table 2. Comparisons on the LLIE task based on the KC and RLE datasets.

Methods	Ref.	KC			RLE			Params
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	
MIRNetv2 [27]	TPMAI 22	31.67	95.22	0.0486	32.85	92.69	0.0781	5.89M
SNR-Aware [21]	CVPR 22	30.32	94.92	0.0521	27.73	88.44	0.1094	4.01M
LLCaps [3]	MICCA 23	35.24	96.34	0.0374	33.18	93.34	<u>0.0721</u>	119.73M
PIP [11]	Arxiv 23	33.60	95.09	0.0302	28.60	87.27	0.0977	26.82M
CFWD [22]	Arxiv 24	35.88	96.26	0.0467	30.14	90.25	0.1088	-
Dif-LOL [10]	ACM TOG 23	33.60	95.42	0.0847	28.46	82.52	0.1437	-
LA-Net [23]	IJCV 23	30.84	95.32	0.0562	25.92	85.72	0.1491	0.56M
CLE [24]	MM 23	26.55	87.87	0.0829	26.20	81.42	0.1134	37.01M
PyDiff [30]	IJCAI 23	35.07	96.60	0.0364	33.21	93.54	0.0774	32.00M
PromptIR [13]	NIPS 23	33.54	96.77	0.0377	32.07	93.30	0.0694	32.96M
Retinexformer [4]	ICCV 23	37.14	<u>97.67</u>	0.0219	32.81	93.23	0.0831	1.61M
MambaLLIE [18]	NIPS 24	<u>37.45</u>	97.25	<u>0.0213</u>	33.36	93.30	0.0784	2.28M
EndoUIC [2]	MICCAI 24	36.85	97.04	0.0255	<u>33.50</u>	<u>93.99</u>	0.0658	28.91M
WTNet (Ours)	-	38.64	97.89	0.0199	34.37	94.00	0.0772	1.41M

Table 3. Comparisons on the downstream segmentation task on the RLE dataset.

Metric	MIRNetv2 [29]	LLCaps [3]	PIP [11]	CFWD [22]	Dif-LOL [10]	LA-Net [23]	SNR-Aware [21]
mIoU \uparrow	63.14	66.47	59.46	51.47	62.46	52.57	58.95
Metric	CLE [24]	PyDiff [30]	PromptIR [13]	Retinexformer[4]	MambaLLIE[18]	EndoUIC[2]	WTNet (Ours)
mIoU \uparrow	45.33	62.56	59.92	65.91	68.87	<u>68.97</u>	69.73

substantiates its superior universality. To further validate the advantages of our approach, following [3], we compare the performance of different methods on the downstream segmentation task on the RLE dataset. As shown in Table 3, our method obtains the best segmentation performance with the highest mIoU score. This fully indicates the clinical value of our method.

3.4 Ablation Studies

We further conduct ablation experiments to verify the effectiveness of the proposed module in WTNet. As shown in Table 4, by only replacing the default downsampling/upsampling operations in baseline (*i.e.*, Restormer) with the proposed wavelet transform strategy, WTNet₁ can significantly reduce the number of model parameters while maintaining similar performance. In addition, the introduction of depth-wise convolution on high-frequency components and ECM can both further improve the performance with almost no additional parameters. These results fully demonstrate the effectiveness of our designs.

4 Conclusion

Against this endoscopic exposure correction task, we carefully investigated the task-specific priors and then utilized the wavelet transform to construct a simple yet effective lightweight network framework, called WTNet. Extensive experiments conducted on four datasets demonstrated the comprehensive effectiveness

Table 4. Ablation study based on Endo4IE. “WT” denotes the proposed wavelet transform based downsampling/upsampling operation. “DWC” means depth-wise convolution executed on high-frequency components. “ECM” is exposure calibration module.

Variant	UNet Backbone	WT	DWC	ECM	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	Params
Baseline	✓	✗	✗	✗	27.61	86.10	0.1992	18.78M
WTNet ₁	✓	✓	✗	✗	27.73	86.05	0.2023	1.40M
WTNet ₂	✓	✓	✗	✓	27.76	86.04	0.1999	1.41M
WTNet ₃	✓	✓	✓	✗	27.84	86.08	0.1973	1.41M
WTNet (Ours)	✓	✓	✓	✓	27.91	86.14	0.1954	1.41M

of our proposed WTNet in balancing performance and cost. Besides, through a downstream segmentation experiment, we further validated the application potential of our method. Especially, our model only requires 1.41M parameters, which is extremely friendly for the deployment on the resource-limited devices. Please note that Restormer is just a way for us to experiment. In the future, we will attempt to explore more configurations on $\text{Operator}(\cdot)$ in Eq. (2).

Acknowledgments. This work was supported by the Ministry of Science and Technology of the People’s Republic of China, Key Research and Development Program (Grant No. 2024YFA1012003-ZKT02).

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: IEEE International Conference on Computer Vision. pp. 6156–6165 (2023)
2. Bai, L., Chen, T., Tan, Q., Nah, W.J., Li, Y., He, Z., Yuan, S., Chen, Z., Wu, J., Islam, M., et al.: EndoUIC: Promptable diffusion transformer for unified illumination correction in capsule endoscopy. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 296–306 (2024)
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 (2023)
4. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In: IEEE International Conference on Computer Vision. pp. 12504–12513 (2023)
5. Coelho, P., Pereira, A., Leite, A., Salgado, M., Cunha, A.: A deep learning approach for red lesions detection in video capsule endoscopies. In: International Conference on Image Analysis and Recognition. pp. 553–561 (2018)
6. Esposito, G., Pimentel-Nunes, P., Angeletti, S., Castro, R., Libânio, D., Galli, G., Lahner, E., Di Giulio, E., Annibale, B., Dinis-Ribeiro, M.: Endoscopic grading of

- gastric intestinal metaplasia: a multicenter validation study. *Endoscopy* **51**(06), 515–521 (2019)
7. 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 (2022)
 8. 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: IEEE International Symposium on Biomedical Imaging. pp. 1–5 (2023)
 9. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of PSNR in image/video quality assessment. *Electronics Letters* **44**(13), 800–801 (2008)
 10. 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)
 11. Li, Z., Lei, Y., Ma, C., Zhang, J., Shan, H.: Prompt-in-prompt learning for universal image restoration. *arXiv preprint arXiv:2312.05038* (2023)
 12. 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: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 667–677 (2020)
 13. Potlapalli, V., Zamir, S.W., Khan, S.H., Shahbaz Khan, F.: PromptIR: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems* **36**, 71275–71293 (2024)
 14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 234–241 (2015)
 15. Shi, Y., Li, Z., Wang, L., Wang, H., Liu, X., Gu, D., Chen, X., Liu, X., Gong, W., Jiang, X., et al.: Artificial intelligence-assisted detection of nasopharyngeal carcinoma on endoscopic images: a national, multicentre, model development and validation study. *The Lancet Digital Health* (2025)
 16. 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)
 17. 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)
 18. Weng, J., Yan, Z., Tai, Y., Qian, J., Yang, J., Li, J.: MambaLLIE: Implicit retinex-aware low light enhancement with global-then-local state space. *Advances in Neural Information Processing Systems* (2024)
 19. Wu, H., Yang, Y., Aviles-Rivero, A.I., Ren, J., Chen, S., Chen, H., Zhu, L.: Semi-supervised video desnowing network via temporal decoupling experts and distribution-driven contrastive regularization. In: European Conference on Computer Vision. pp. 70–89 (2024)
 20. Wu, H., Yang, Y., Xu, H., Wang, W., Zhou, J., Zhu, L.: Rainmamba: Enhanced locality learning with state space models for video deraining. In: ACM International Conference on Multimedia. pp. 7881–7890 (2024)
 21. Xu, X., Wang, R., Fu, C.W., Jia, J.: SNR-aware low-light image enhancement. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 17714–17724 (2022)
 22. 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)

23. 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)
24. Yin, Y., Xu, D., Tan, C., Liu, P., Zhao, Y., Wei, Y.: Cle diffusion: Controllable light enhancement diffusion model. In: *ACM International Conference on Multimedia*. pp. 8145–8156 (2023)
25. 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)
26. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5728–5739 (2022)
27. 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)
28. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 586–595 (2018)
29. 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: *IEEE International Conference on Robotics and Biomimetics*. pp. 01–06 (2022)
30. Zhou, D., Yang, Z., Yang, Y.: Pyramid diffusion models for low-light image enhancement. In: *International Joint Conference on Artificial Intelligence*. pp. 1795–1803 (2023)