IEEE | IEEE Transactions on Multimedia

Structure-guided Diffusion Transformer for Low-Light Image Enhancement

Journal:	*IEEE Transactions on Multimedia*
Manuscript ID	MM-020552.R1
Manuscript Type:	Regular Paper (S1)
Date Submitted by the Author:	25-Dec-2024
Complete List of Authors:	Yin, Xiangchen; University of Science and Technology of China, None; Yu, Zhenda; Anhui University, School of Artificial Intelligence JIANG, LONGTAO; University of Science and Technology of China School of Information Science and Technology, Department of Electronic Engineering and Information Science Yang, Xun; University of Science and Technology of China, School of Information Science and Technology Gao, Xin; China University of Mining and Technology - Beijing, School of Artificial Intelligence Sun, Xiao; Hefei University of Technology, School of Computer Science and Information Engineering Liu , Zhi; The University of Electro-Communications, Department of Computer and Network Engineering
Subject Category Please select at least one subject category that best reflects the scope of your manuscript:	SIGNAL PROCESSING FOR MULTIMEDIA APPLICATIONS, EMERGING TOPICS IN MULTIMEDIA
EDICS:	1-IVGAS Image/Video/Graphics Analysis and Synthesis < 1 SIGNAL PROCESSING FOR MULTIMEDIA APPLICATIONS, 9-DLMA Deep Learning for Multimedia Analysis < 9 EMERGING TOPICS IN MULTIMEDIA, 9-DLMP Deep Learning for Multimedia Processing < 9 EMERGING TOPICS IN MULTIMEDIA

Author's Responses to Custom Submission Questions

Is this manuscript a resubmission of, or related to, a previously rejected manuscript?	Yes
If "Yes", specify the publication venue and manuscript ID of the previous submission and upload a supporting document detailing how the resubmission has addressed the concerns raised during the previous review. If this does not apply, type N/A.	Submission: IEEE Transactions on Multimedia ID: MM-020367 In the previous submission, we did not clearly express the importance of our work to multimedia in the cover letter. In the latest submission, we revised the paper and elaborated on the enhancement of low-light images to improve the visual quality and availability of multimedia content, thus bringing a better user experience and more accurate analysis to applications that rely on visual information. By advanced diffusion model and transformer architecture, the methods we proposed greatly improve the quality of enhanced images, thus contributing to the latest development of multimedia technology.
Is this manuscript an extended version of a conference publication?	No
If "Yes", provide the full citation of the conference submission or publication. If this does not apply, type N/A.	N/A
Is this manuscript related to any other papers of the authors that are either published, accepted for publication, or currently under review, and that are not included among the references cited in the manuscript?	No
If "Yes", please list these papers below. Except for permitted preprints, explain why these papers are not included among the references cited in the manuscript and how they are different from the manuscript. Include any unpublished papers as "Supporting Documents". If this does not apply, type N/A.	N/A
What is the contribution of this paper, within the scope of Transactions on Multimedia?	By leveraging advanced diffusion models and transformer architectures, our proposed method offers a substantial improvement in the quality of enhanced images, thereby contributing to the state-of-the-art in multimedia technologies. Within the title of "Transactions on Multimedia", our

	paper contributes to Image Analysis and Synthesis. Recent low-light enhancement work included in "Transactions on Multimedia" is as follows: 1. TBEFN: A two-branch exposure-fusion network for low-light image enhancement 2. Retinex-based variational framework for low-light image enhancement and denoising However, most of these works are implemented on CNN or Transformer. Their method of learning complex mapping is very dependent on the dataset and is prone to overfitting. Our work uses diffusion transformer to learn the noise distribution to the data distribution, and obtains stunning image generation effects through reverse denoising.
Why is the contribution significant (What impact will it have)?	Enhancing low-light images improves the visual quality and usability of multimedia content, leading to better user experiences and more accurate analyses in applications relying on visual information. Our contributions are as follows: â—☐ Our method provides a new low-light enhancement paradigm based on diffusion transformer, while joint the advantages of diffusion and transformer to achieve outstanding enhancement performance. â—☐ By enhancing and fusing frequency bands in the wavelet domain, our method restores the lost details in low-light images and greatly improves the naturalness and contrast of the enhanced image. â—☐ we adopt structural priors to pay more attention to these areas and avoid influence from noisy areas. â—☐ Our model achieves SOTA performance on 8 benchmark datasets and improves accuracy in nighttime segmentation.
What are the three papers in the published literature most closely related to this paper? Please provide full citation details, including DOI references where possible.	1.Peebles W, Xie S. Scalable diffusion models with transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4195-4205

	2.Jiang H, Luo A, Fan H, et al. Low-light image enhancement with wavelet-based diffusion models[J]. ACM Transactions on Graphics (TOG), 2023, 42(6): 1-14. 3. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.
What is distinctive/new about the current paper relative to these previously published works?	1. We introduce the diffusion transformer into the low-light enhancement task for the first time, and proposed a diffusion-based enhancement framework named SDTL; 2. We design a Structure Enhancement Module (SEM) to enhance the frequency domain information in two steps: enhancement and fusion; 3. We propose a Structure-guided Attention Block (SAB) to guide the self attention between tokens, avoid influence from noisy areas.

Structure-guided Diffusion Transformer for Low-Light Image Enhancement

Xiangchen Yin, Zhenda Yu, Longtao Jiang, Xun Yang, Xin Gao, Xiao Sun*, Senior Member, IEEE, Zhi Liu, Senior Member, IEEE

***Abstract*—While the diffusion transformer (DiT) has become a focal point of interest in recent years, its application in low-light image enhancement remains a blank area for exploration. Current methods recover the details from low-light images while inevitably amplifying the noise in images, resulting in poor visual quality. In this paper, we firstly introduce DiT into the low-light enhancement task and design a novel Structure-guided Diffusion Transformer based Low-light image enhancement (SDTL) framework. We compress the feature through wavelet transform to improve the inference efficiency of the model and capture the multi-directional frequency band. Then we propose a Structure Enhancement Module (SEM) that uses structural prior to enhance the texture and leverages an adaptive fusion strategy to achieve more accurate enhancement effect. In Addition, we propose a Structure-guided Attention Block (SAB) to pay more attention to texture-riched tokens and avoid interference from noisy areas in noise prediction. Extensive qualitative and quantitative experiments demonstrate that our method achieves SOTA performance on several popular datasets, validating the effectiveness of SDTL in improving image quality and the great potential of DiT in low-light enhancement tasks.**

***Index Terms*—Diffusion Transformer; Low-Light Enhancement; Low-level Vision**

I. Introduction

Low-light images bring unpleasant visibility and result in complex detail damage (e.g., noise, color distortion, etc.). Therefore, low-light enhancement has received increasing attention in computer vision to improve the visual quality of images and support downstream vision applications [1]–[4].

Traditional enhancement methods are mainly based on image priors or physical models. Histogram equalization [5] adjusts the pixel distribution of the image through statistics.

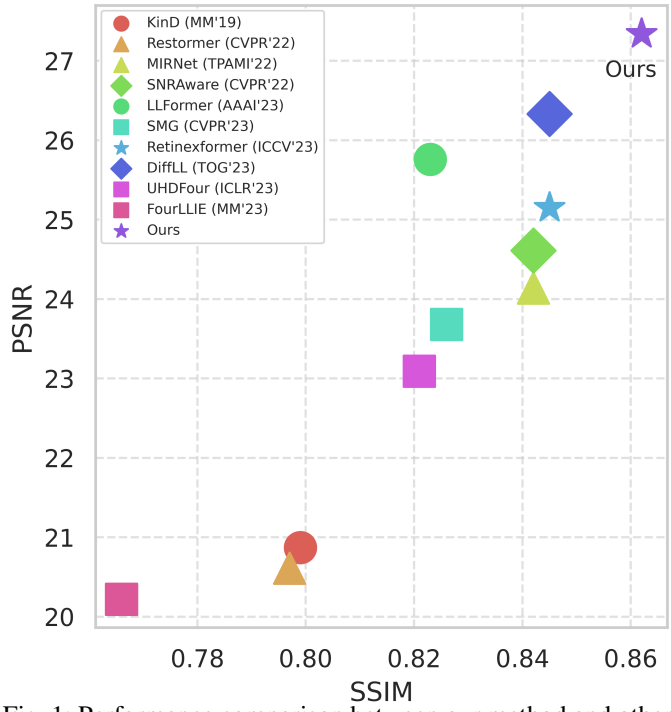


Fig. 1: Performance comparison between our method and other SOTA methods on the LOLv1 [8] dataset. Our method has achieved the best results on both PSNR and SSIM, reaching 27.34 and 0.862 respectively. The legend on the right shows each model and source.

Retinex-based methods [6], [7] decompose the image into illumination components and reflection components and adjust the reflection components. However, these methods often generate unnatural images and have poor scene robustness due to the use of artificially designed priors. Deep learning-based methods [8]–[12] fit complex mappings from low-light to normal-light on paired datasets. However, these methods have difficulty in recovering weak and missing details, even amplifying the noise to a certain extent inevitably.

Recently the denoising diffusion probabilistic model (DDPM) [13] has demonstrated remarkable success in the domain of image generation, which produces more realistic details through a series of refinement. Recognizing the prowess of DDPM in accurately capturing pixel distributions, it has been increasingly leveraged for enhancing image quality. In addition, there have been several successful attempts to introduce conditional DDPM in low-level vision tasks [14]–[16]. However, the diffusion model requires longer training time

* Xiao Sun is the corresponding author.

Xiangchen Yin, Longtao Jiang, Xun Yang is associated with the University of Science and Technology of China, and also with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China. Email: yinxiangchen@mail.ustc.edu.cn, taotao707@mail.ustc.edu.cn, xyang21@ustc.edu.cn

Zhenda Yu is associated with Anhui University, China, and also with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China. Email: wa22201140@stu.ahu.edu.cn

Xin Gao is associated with the School of Vehicle and Mobility, Tsinghua University, China, and also with the State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing, China. Email: gaoxin97@mail.tsinghua.edu.cn

Xiao Sun is associated with the School of Computer Science and Information Engineering, Hefei University of Technology, China, and also with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China. Email: sunx@hfut.edu.cn

Zhi Liu is associated with Department of Computer and Network Engineering, The University of Electro-Communications, Chofu-shi, Tokyo,1828585 Japan. Email: liu@ieee.org

and the image recovery effect is unstable. Transformer [17] obtains long-range dependencies in sequences and captures the global context in images. The diffusion model stands at the cutting edge of image generation, offering substantial benefits. However, the prevalent reliance on the rudimentary U-Net architecture fails to fully capitalize on the extensive potential of the diffusion model, suggesting a need for more sophisticated structural adaptations. Therefore, the research on the diffusion transformer (DiT) [18] for low-light enhancement needs to be deeply explored. The structure of DiT not only keeps the advantages of the transformer but also freely controls the size of the model by adjusting parameters such as the number of blocks and the embedded dimension.

To solve the above problems, this paper introduces the DiT technique for the low-light enhancement task for the first time and achieve impressive performance. We propose a novel **S**tructure-guided **D**iffusion **T**ransformer based **L**ow-light (SDTL) enhancement framework, making full use of generation capabilities of the diffusion model. Firstly, we transform the features into the frequency domain via wavelet transformation to improve the inference efficiency of the model, and the features are decomposed into the information in multiple directions, including horizontal, vertical, and diagonal. This decomposition captures local structures at different scales, providing richer information for subsequent enhancements. Then we design a Structure Enhancement Module (**SEM**) to enhance the frequency domain information in two steps: *enhancement* and *fusion*. SEM enhances the ability of network to refine details by incorporating structural priors, effectively directing the focus of network towards critical structural elements and simultaneously muting the influence of extraneous details. Additionally, we design an adaptive fusion strategy to complement the information in various directions. As for DiT, we propose a Structure-guided Attention Block (**SAB**) to guide the self-attention mechanism between tokens. We notice that only local areas of the image corresponding to texture-rich tokens are beneficial to the network, so we adopt structural priors to pay more attention to these areas and avoid influence from noisy areas. We conducted extensive experiments on 8 benchmark datasets, and the results show that our SDTL surpasses the current state-of-the-art (SOTA) models in low-light enhancement tasks without adopting any training strategy, as shown in Fig. 1. Experimental results demonstrate the effectiveness of our method in high image quality and model efficiency, while also reflect the great potential of DiT on low-light enhancement.

Our contributions can be summarized below:

- We devise a novel **S**tructure-guided **D**iffusion **T**ransformer based **L**ow-light image enhancement (**SDTL**) method, which achieves impressive performance.
- We design a Structure Enhancement Module (**SEM**) to exploit contextual information. By introducing structural priors, SEM enables the network to focus on important structural cues while suppressing other irrelevant information, thus improving the visual quality of the enhanced image. In addition, we also use an adaptive fusion strategy to complement information in various directions.
- We propose a Structure-guided Attention Block (**SAB**) in DiT to guide the self-attention mechanism between tokens so that the network pays more attention to texture-rich areas and avoids interference from noise areas.
- Extensive experiments on multiple datasets show that our SDTL surpasses current state-of-the-art models and improves the performance of night segmentation.

II. Related Work

A. Low-light Image Enhancement

Previous work used some traditional techniques to enhance images such as histogram equalization [5], gamma correction [19], and Retinex adjustment [7], [20]. With the prosperity of deep learning, many learning-based methods [21]–[24] have been proposed to learn the mapping from low-light to normal directly through CNN or Transformer. Wei et al. [8] proposed RetinexNet, which combines Retinex theory and CNN to achieve learnable image decomposition and adjustment of the reflection component. Xu et al. [25] jointed signal-noise ratio (SNR) and Transformer to achieve spatially varying image enhancement. LLFormer [26] was proposed to enhance ultra-high definition (UHD) images, and they published a UHD low-light dataset. Wang et al. [27] proposed FourLLIE, which estimates the amplitude transform in Fourier space to improve the brightness of low-light images.

B. Diffusion for Low-level Vision

Recently, with the advancement of the denoising diffusion probabilistic model (DDPM), great progress has been made in the diffusion model for low-level vision [28]–[30], which makes it gradually used in low-level vision tasks. Xia et al. [31] proposed an efficient DiffIR, which guides the convergence of the diffusion model through a compact IR prior representation. Jiang et al. [14] used the diffusion model to perform image enhancement in the wavelet domain, which significantly speeds up the inference speed, but ignores the prior knowledge of enhancement. Whang et al. [32] proposed a deblurring framework based on the conditional diffusion model, which trained a random sampler to refine the output of the predictor. Yin et al [33] proposed the Controllable Light Enhancement Diffusion model (CLE Diffusion), which allows users to input the required brightness level and apply the SAM model to achieve friendly interactive regional controllable brightness. However, the performance gap between this method and previous SOTA methods is not large and there is no great improvement in quantitative indicators. The previous method adopts U-Net shape denoising network, the flexibility of the model parameters is limited and the sampling speed is slow. Our method improves the speed through wavelet transform compression features and uses diffusion transformer to set network parameters at will according to the actual situation.

III. Methods

Given an image I, we select a pair of low-pass/high-pass filters to perform 2-dim Discrete Wavelet Transform (DWT) for frequency band decomposition as:

$$I_{\mathrm{LL}}, I_{\mathrm{HL}}, I_{\mathrm{LH}}, I_{\mathrm{HH}} = \mathcal{W}[I; G_{\mathrm{L}}(t), G_{\mathrm{H}}(t)], \tag{1}$$

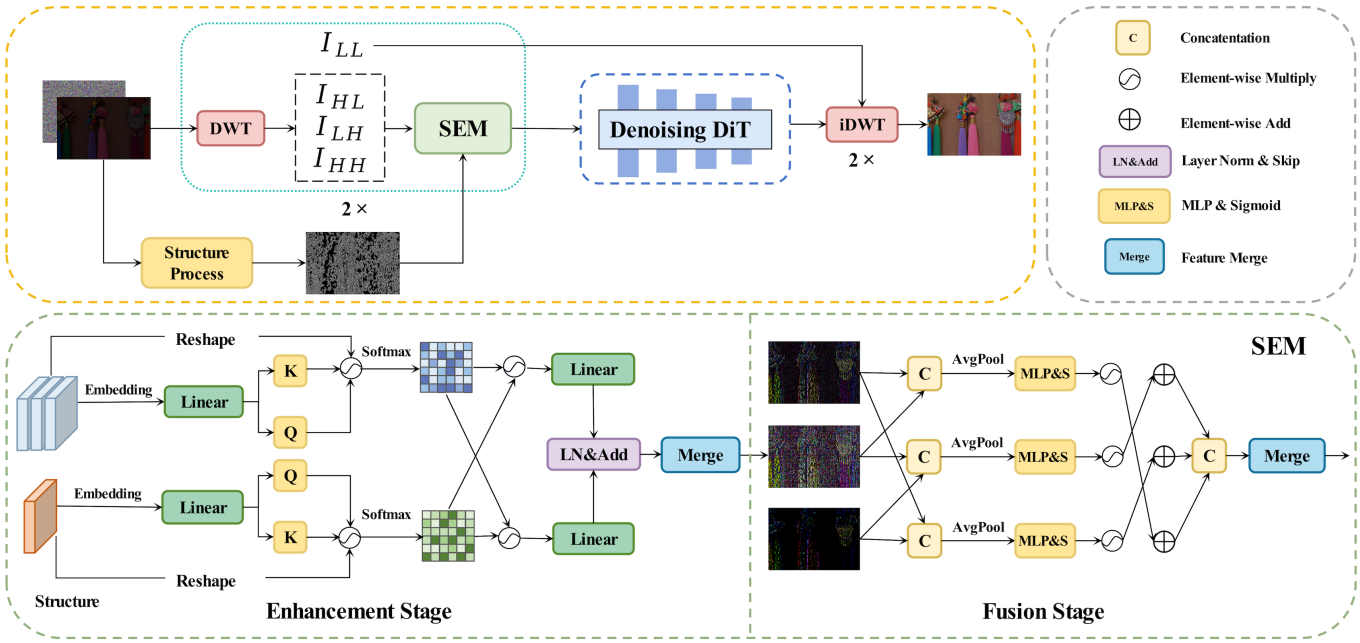


Fig. 2: Overview of SDTL. Our model compresses features to improve the efficiency of inference through wavelet transform. We design a Structure Enhancement Module (SEM) to enhance structural information under different frequency bands, which is divided into two stages: **Enhancement** and **Fusion**. In addition, we propose a Structure-guided Attention Block (SAB) to pay attention to the texture-riched tokens in the noise prediction network.

where $\mathcal{W}$ represents the 2-dim DWT, G_{L} represents a low-pass filter, G_{H} represents a high-pass filter, I_{LL}, I_{HL}, I_{LH}, I_{HH} represents low-frequency subband, vertical high-frequency subband, horizontal high-frequency subband, and diagonal high-frequency subband respectively. The low-frequency subband is home to the overarching structural essence of the image, capturing its fundamental layout and composition. In contrast, the high-frequency subband is where the intricate texture details reside, often accompanied by noise elements that are particularly prevalent in low-light images. We design a Structure Enhancement Module (SEM) to enhance structural information of high-frequency. After two wavelet transforms, we all adopted a SEM and the features will be reduced to 16 times of the original. Note that the low-frequency information doesn't participate in the network, and is only used for the final Inverse Wavelet Transform (IWT). We take the features as a condition after frequency processing and concatenate it with random noise as the input of the diffusion model. The overview of our SDTL is shown in Fig. 2.

A. Conditional Diffusion Model

The diffusion model gradually converts the distribution of Gaussian noise into a data distribution by learning the Markov chain, which is divided into two stages: forward diffusion and reverse denoising. Gaussian noise is gradually added to the forward diffusion. Firstly, we define a variance scheduler with T diffusion steps $\{\beta_1, \beta_2, \beta_3, \ldots, \beta_T\}$:

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) = \prod_{t=1}^{T} \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \tag{2}$$

where x_t is the damaged data after noising, $\mathcal{N}$ denotes the Gaussian noise, x_0 denotes the original image, and I denotes the condition of the diffusion model. The data of time t is only related to time $t-1$. The larger the t, the closer it is to pure noise. The reverse denoising process gradually denoises the randomly sampled Gaussian noise, and the model will predict the reverse distribution of noise p_θ as:

$$p_\theta\left(X_{0:T}\right) = p\left(x_T\right) \prod_{t=1}^{T} p_\theta\left(x_{t-1} \mid x_t\right), \tag{3}$$

$$p_\theta\left(x_{t-1} \mid x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \sigma_t^2 \mathcal{I}\right), \tag{4}$$

where $\boldsymbol{\epsilon}_\theta$ denotes a noise prediction network, θ is the parameters of $\boldsymbol{\epsilon}_\theta$ (here is DiT), $\mu_\theta\left(x_t, t\right)$ and σ_t^2 are the mean and variance of the distribution respectively. We adopt the frequency of low-light images as a condition, so the variance is constrained by $\mathcal{I}$, and the mean is computed as:

$$\boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right), \tag{5}$$

$$\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i. \tag{6}$$

The optimization goal of diffusion model is to maximize the logarithmic likelihood of the data distribution, and the loss function is defined as

$$\mathcal{L} = E_{\mathbf{x}_0, t, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}\left[\left\|\epsilon_t - \epsilon_\theta\left(\mathbf{x}_t, t\right)\right\|^2\right]. \tag{7}$$

B. Structure Enhancement Module

Give three frequency bands in different directions after the wavelet transform $\mathcal{I}_{\mathrm{HL}}$, $\mathcal{I}_{\mathrm{LH}}$, $\mathcal{I}_{\mathrm{HH}} \in R^{H \times W \times C}$ and

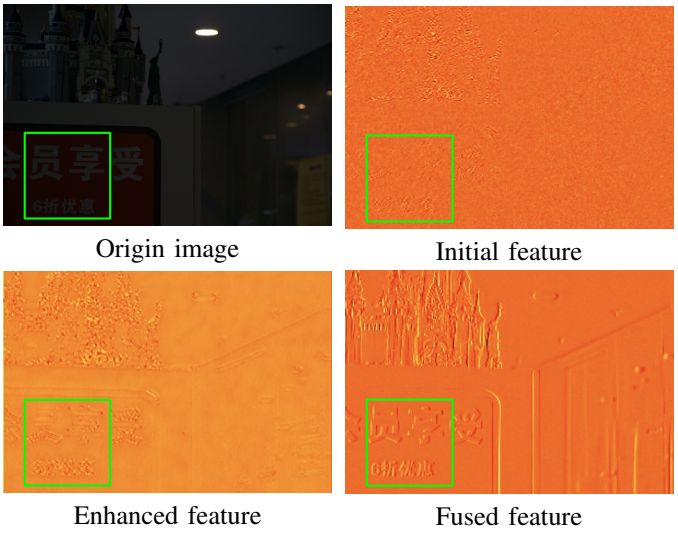


Fig. 3: Visualization of hot feature for SEM. We visualize the initial feature, enhanced feature, and fused feature in SEM respectively. It can be found that our SEM has a significant enhancement effect on features.

the structural information initially processed by convolution $\mathcal{S} \in R^{H\times W\times C}$. We design a Structure Enhancement Module (SEM) to achieve guided enhancement and multi-directional complementary fusion, as shown in Fig. 2. Then we introduce the two stages of enhancement and fusion of SEM.

(1) Enhancement: It should be noted that we perform the same operation for each frequency band. We firstly flatten the features and structure of a single frequency band to $R^{N\times C}(N = H \times W)$. Then two linear layers are used to transform the features as two vectors $\mathcal{X}_I$ and $\mathcal{X}_S$ $(R^{N\times \hat{C}})$, respectively. We further adopt a novel bi-directional guidance mechanism, which spreads the information of vectors with each other to make up for the uncertainty of structural information on the dark area, and provides more efficient structural guidance to the frequency band. Then we encode $\mathcal{X}_I$ and $\mathcal{X}_S$ into vectors respectively, which get the key $\mathcal{K}_I$ and query $\mathcal{Q}_I$ of the features, and the key $\mathcal{K}_S$ and query $\mathcal{Q}_S$ of the structure. We multiply the attention map with another sequence to realize the exchange of information between the sequence and the sequence. This process is described as:

$$\mathcal{E}_I = \text{Softmax}(\frac{\mathcal{K}_I^T \cdot \mathcal{Q}_I}{\alpha}) \cdot \mathcal{R}(\mathcal{X}_S), \tag{8}$$

$$\mathcal{E}_S = \text{Softmax}(\frac{\mathcal{K}_S^T \cdot \mathcal{Q}_S}{\alpha}) \cdot \mathcal{R}(\mathcal{X}_I), \tag{9}$$

where $\mathcal{R}$ denotes the reshape of the vector and α denotes the scale factor. After the information interaction, the vectors undergo processing via a linear layer, and then we implement layer normalization and skip connection. Finally, we perform the merge operation through a convolutional block and reshape the feature back to $R^{H\times W\times C}$ to get the enhanced frequency $\hat{\mathcal{I}}_t(t = \text{HL}, \text{LH}, \text{HH})$.

(2) Fusion: The frequency bands in different directions after enhancement are usually complementary. In this stage, we propose a new cross-band fusion mechanism, and each band is complementarily integrated after being corrected by the other two bands. Firstly, after the current frequency band is processed with convolution blocks, the other two frequency bands are concat on the channel, and the features are adjusted by the channel gate, then cross into the current frequency band by addition operation:

$$\begin{aligned} &\hat{\mathcal{I}}\delta = \hat{\mathcal{I}}\delta + \mathcal{G}(\text{Concat}(\hat{\mathcal{I}}\kappa, \hat{\mathcal{I}}\tau)), \\ &\delta, \kappa, \tau \in \{\text{HL}, \text{LH}, \text{HH}\}; \delta \neq \kappa \neq \tau \end{aligned} \tag{10}$$

where $\mathcal{G}(x) = \sigma(\text{MLP}(\text{AvgPool}(x))) * x$ represents channel gate, σ represents sigmoid function and AvgPool represents average pooling. Finally, we use a convolution block to summarize the information of the three frequency bands and make a skip connection. We visualize the hot features of the enhancement and fusion stages in SEM, as shown in Fig. 3. We can find the change of text in the rectangle part, by visualizing origin image, initial feature, enhanced feature and fused feature for comparison at different stages. Comparing the "Initial feature" and "Enhanced feature", shows that the structural prior significantly restores high-frequency information in the enhancement stage for subsequent noise prediction. Comparing "Enhanced feature" and "Fused feature" shows that fusion makes the frequency bands in different directions complementary and more balanced in color and texture.

C. Denoising of Diffusion Transformer

In this section, we introduce the structural design of DiT, as shown in Fig. 4. Give conditional input $C \in R^{H\times W\times 6}$, we convert the feature diagram of spatial input into N token sequences through patchify, N is determined by the size of each patch p, it is calculated as $N = HW/p^2$, the embedding dimension of each token is set to 384. In this paper, we set p to 4 and the head number of self-attention is set to 6. In addition, we adopt 6 Structure DiT (SDT) blocks, each of which mainly consists of a ViT [34] block and a Structure-guided Attention Block (SAB). SAB (Fig. 4 (c)) introduces structural information to guide the self-attention mechanism between tokens, which pays more attention to texture-riched token sequences and avoids influence by noisy areas. After patchify and projection the structure sequence is fused into the token sequence of features for self-attention calculation. This process is expressed as

$$\begin{aligned} &Q, K = \text{Fuse}(\mathcal{P}(x), \mathcal{P}(\mathcal{S})), \\ &V = \mathcal{R}(\mathcal{P}(x)), \\ &\text{SAB}(x) = \text{Softmax}(K^T Q/\alpha)V. \end{aligned} \tag{11}$$

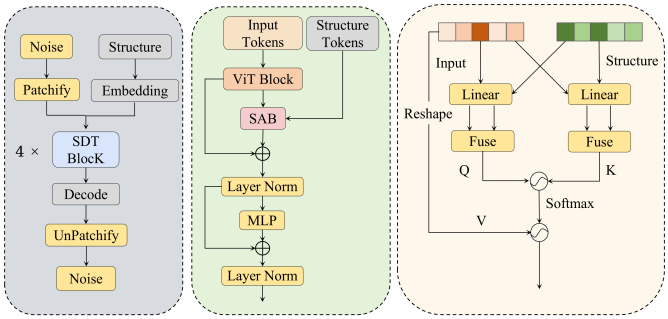
where $\mathcal{P}$ denotes the patchify and Fuse denotes the cross attention operation that fuses the two inputs. We use 4 SDT blocks as the encoding layer and use skip connection. Finally, the decoding layer uses two SDT blocks and a linear projection layer, and then rearranges the token to restore the same size as the original feature $R^{H\times W\times 3}$ to get the predicted noise.

IV. EXPERIMENTS

Datasets. We validate the effectiveness of our methods on the LOLv1 [8], LOLv2-real [35], LSRW [36] datasets. The

TABLE I: Quantitative comparisons of different methods on the LOLv1 [8], LOLv2 [35] and LSRW [36] datasets. The best results and second-best results are highlighted in blue and green, respectively. Note that we obtained these results either from the corresponding papers, or by running the officially released models, and the missing results on LSRW are marked as "-".

Methods	References	LSRW				LOLv1				LOLv2-real			
		PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
RetinexNet [8]	BMVC'18	15.61	0.414	0.454	108.350	16.77	0.462	0.474	126.266	18.37	0.723	0.365	133.905
KinD [37]	MM'19	-	-	-	-	20.87	0.799	0.207	104.632	17.54	0.669	0.375	137.346
DRBN [38]	CVPR'20	16.73	0.507	0.457	80.727	19.86	0.834	0.155	98.732	20.13	0.830	0.147	89.085
EnlightenGAN [9]	TIP'21	17.11	0.463	0.406	69.033	17.48	0.652	0.322	94.704	18.64	0.677	0.309	84.044
Restormer [39]	CVPR'22	16.30	0.453	0.427	69.219	20.61	0.797	0.288	72.998	24.91	0.851	0.264	58.649
URetinex-Net [40]	CVPR'22	18.27	0.518	0.419	66.871	19.84	0.824	0.237	52.383	21.09	0.858	0.208	49.836
Uformer [41]	CVPR'22	16.59	0.494	0.435	82.299	19.00	0.741	0.354	109.351	18.44	0.759	0.347	98.138
MIRNet [42]	TPAMI'22	16.47	0.477	0.430	93.811	24.14	0.842	0.131	69.179	20.36	0.782	0.317	49.108
SNR-Aware [25]	CVPR'22	16.49	0.505	0.419	65.807	24.61	0.842	0.152	55.121	21.48	0.849	0.157	54.532
IAT [43]	BMVC'22	20.81	0.565	0.467	80.499	23.38	0.809	0.134	67.412	23.50	0.824	0.191	62.153
LLFormer [26]	AAAI'23	20.69	0.560	0.518	96.782	25.76	0.823	0.167	65.271	26.20	0.819	0.209	63.432
FourLLIE [27]	MM'23	18.61	0.505	0.316	73.55	20.22	0.766	0.250	91.793	22.34	0.847	0.051	89.334
UHDFour [44]	ICLR'23	17.30	0.529	0.443	62.032	23.09	0.821	0.259	56.912	21.79	0.854	0.292	60.837
SMG [45]	CVPR'23	19.04	0.568	0.392	101.56	23.68	0.826	0.118	58.846	24.62	0.867	0.148	78.582
Retinexformer [46]	ICCV'23	20.15	0.534	0.336	70.36	25.15	0.845	0.131	71.148	22.80	0.840	0.171	62.439
DiffLL [14]	TOG'23	19.28	0.552	0.350	45.294	26.33	0.845	0.217	48.114	28.85	0.876	0.207	45.359
RSFNet [47]	CVPR'24	-	-	-	-	22.15	0.860	0.265	-	21.59	0.843	0.278	-
LightenDiffusion [48]	ECCV'24	18.55	0.539	0.311	-	20.45	0.803	0.192	-	-	-	-	-
ExpoMamba [49]	ICML'24	-	-	-	-	25.77	0.860	0.212	89.210	28.04	0.885	0.232	85.920
SDTL (Ours)	-	21.23	0.576	0.392	42.569	27.34	0.862	0.118	50.432	28.85	0.875	0.131	43.237



(a) Denoising DiT (b) SDT Block (c) SAB

Fig. 4: Details of Denoising Diffusion Transformer (a). SDT block consists of a ViT block and a SAB block (b). SAB integrates the original token with the structural token to make the network focus on the textured-riched token (c).

LOLv1 dataset has 485 paired images for training and 15 paired images for testing, both in sizes (400, 600). The LOLv2 dataset consists of a subset of real and synthetic scenes. LOLv2-real has 689 images for training and 100 images for testing respectively. LSRW is a large low-light dataset. It uses Huawei and Nikon cameras to collect 5,650 samples of different scenes, of which 5,600 are for training and 50 for testing. In addition, to verify the robustness of our model, we evaluate it on five datasets without reference: LIME [20], VV [50], DICM [51], NPE [52], and MEF [53]. The ACDC [54] dataset can be used to train semantic segmentation under adverse conditions. It contains 4,006 images, distributed in the four common bad weather of fog, night, rain, and snow, and each image has a high-quality pixel-level label. In this paper, we only use the part of the night scene to verify the performance of downstream night segmentation.

TABLE II: NIQE score on the LIME [20], VV [50], DICM [51], NPE [52], and MEF [53] datasets. The best results and second-best results are marked in blue and green respectively.

Methods	LIME	VV	DICM	NPE	MEF	Average
Zero-DCE [55]	5.820	4.810	4.580	4.530	4.930	4.934
RetinexNet [8]	5.750	4.320	4.330	4.950	4.930	4.856
KinD [37]	4.772	3.835	3.614	4.175	4.819	4.194
MIRNet [42]	6.453	4.735	4.042	5.235	5.504	5.101
SMG [45]	5.451	4.884	4.733	5.208	5.754	5.279
FECNet [56]	6.041	3.346	4.139	4.500	4.707	4.336
FourLLIE [27]	4.402	3.168	3.374	3.909	4.362	3.907
LLFormer [26]	4.796	5.030	4.080	3.809	3.647	4.272
DiffLL [14]	4.203	3.081	3.953	3.535	4.309	3.816
STDL (Ours)	4.004	2.927	3.552	3.977	3.729	3.565

Evaluation Metrics. For the LOLv1, LOLv2 and LSRW datasets, we adopt the four metrics of PSNR, SSIM [57], LPIPS [58] and FID [59] respectively to evaluate models. For the LIME, VV, DICM, NPE, and MEF datasets, we adopt the non-reference image quality metrics NIQE [60], BRISQUE [61] and PI [62] to evaluate them. On the nighttime segmentation dataset ACDC, we adopt $Precision$, $Recall$, $Fscore$ and $mIoU$ to test the performance of the model. $mIoU$ is the mean of IoU in all classes, IoU is described as

$$IoU = \frac{TP}{TP + FP + FN}, \tag{12}$$

where TP and FP are true positive and false positive respectively, FN is false negative. $Fscore$ takes into account

TABLE III: Application of night segmentation on the ACDC [54] dataset. We used DeepLabv3+ [63] as the segmentor (Baseline) to enhance the images. The best results are marked in blue.

Methods	Precision↑	Recall↑	Fscore↑	mIoU↑
Baseline	77.73	69.29	72.41	59.57
DeepLPF [21]	79.19	71.31	73.92	61.88
IAT [64]	79.02	70.67	73.64	61.54
FourLLIE [27]	78.61	69.79	72.84	60.44
DiffLL [14]	78.30	71.34	73.78	61.49
SDTL (Ours)	79.53	71.34	74.18	62.25

$Precision$ and $Recall$, which is described as:

$$Fscore = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{13}$$

where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$.

Implementation Details. We implemented SDTL through Pytorch [65], and this model was trained 1000 epochs on 8 NVIDIA V100-SXM2-32GB. We achieve multi-gpu parallel training through the accelerate framework [1]. The optimizer adopts Adam [66], the initial learning rate is set to 5e-4 and batch size is set to 8. We use StepLR to gradually adjust the learning rate during the training process, where the step size is set to 50 and gamma is set to 0.90. In addition, we randomly crop 256 patches as training samples in low-light/normal-light image pairs during training. It should be noted that we do not use any data augmentation and training strategies except for cropping, which is an end-to-end model. We use 200 diffusion steps for training and in the inference stage, we adopt the DDIM [67] and set 10 steps for image enhancement. In night segmentation, we use MMSegmentation [68]for experiments.

A. Comparison with SOTA Methods

1) Quantitative Results: Table I demonstrated the comparison on the LOLv1 [8], LOLv2-real [35] and LSRW [36] datasets. Obviously, our SDTL achieved advanced performance in three datasets in all the comparative SOTA methods. Specifically, compared with the second best model DiffLL, our method has been significantly improved in most of the metrics of most datasets. Our SDTL increased by 1.95dB, 0.024 and 2.725 respectively in PSNR, SSIM, and FID compared to DiffLL on the LSRW dataset. On the LOLv1 dataset, our method has achieved the best results on PSNR (27.34), SSIM (0.862) and LPIPS (0.118) and the second-best performance on FID (50.432). On the LOLv2 dataset, our method obtained the best results on PSNR (28.85) and FID (43.237). Due to the domain gap between the training set and the test set, many methods did not have stable performance. Our method has achieved the best or the second-best results in LPIPS and FID on the three datasets, which shows that our model produced satisfied visual quality. We also tested NIQE on five unpaired datasets DICM, LIME, MEF, MEF, NPE, and VV, as shown in Table II. The lower the values of NIQE, the better the visual quality of the enhanced image. We achieved the best results in LIME and MEF datasets and the second best results in VV and NPE datasets. We achieved the average best result of the five datasets, with an average NIQE of 3.565. This demonstrated that our SDTL has fully explored the potential of DiT and has better robustness in invisible real scenes.

2) Qualitative Results: The visualization of LOLv1 and LOLv2-real datasets is shown in Fig. 5. We enlarged part of the local area to the bottom left of each image. The first two lines in the figure are the results of the LOLv1 dataset and the last two lines are the results of the LOLv2-real dataset, our SDTL showed pleasant perceptual quality. Specifically, IAT [43] and Retinexformer [46] generated overly smooth textures and rich details and the exposure is unstable. Our method effectively improved the contrast, restored clearer details and suppressed noise. Retinexformer [46] and FourLLIE [27] had different degrees of color distortion or noise amplification. In contrast, our model can effectively enhance areas with low contrast and steadily restore color without introducing noise. In addition, we also compared the comparison between DICM and MEF of unpaired datasets, as shown in Fig. 6. LLFormer [69] and DiffLL [14] did not restore details and colors well and the texture is not rich. Obviously, our method is more natural in color, and the texture is more delicate than other models. This shows that the structure guide in DiT is effective and also performs well in quantitative evaluation.

Visualization of Reverse Denoising processing in SDTL: We visualize the process of SDTL under low light enhancement, which is the reverse denoising process of the diffusion model, as shown in Fig. 7. We use 10 sampling time steps in the process of reverse denoising and show the results for step 0, step 1, step 3, step 6, and step 10, respectively. The first line is the frequency information after wavelet transformation in a certain step, and the second line is the corresponding enhanced results. According to visualization, it can be observed that our SDTL gradually generates normal-light images during the reverse denoising process, and removes noise and artifacts. The iterative refinement process of the diffusion model is the benefit of detail recovery.

3) Application of Night Segmentation: We adopted DeepLabv3+ [63] as the segmentor, using SDTL to enhance the images of the ACDC dataset for the training of night segmentation. We only use the part of the night scene on the ACDC dataset, with about 1,000 images and corresponding high-quality pixel-level labels. We only use the part of the ACDC night scene, which contains about 1,000 images. We use the training set to learn the model and the valid set to evaluate the model. To compare the models fairly, we set the same hyperparameters and training strategies. The training input is the image obtained from the enhanced model, and the label does not change. The experimental results are shown in Table III, and our SDTL has greatly improved the performance of the segmentation model. Our method is 0.76% and 0.4% higher than DiffLL in mIoU and Fscore respectively. Compared with the baseline, our model has improved by 1.77 and 2.68 on Fscore (74.18%) and mIoU (62.25%) respectively. In addition, we visualize the enhanced ACDC and segmentation results, as shown in Fig. 8. From the first column of the figure, we compare the enhanced image effect. The contrast ratio of

[1]https://github.com/huggingface/accelerate

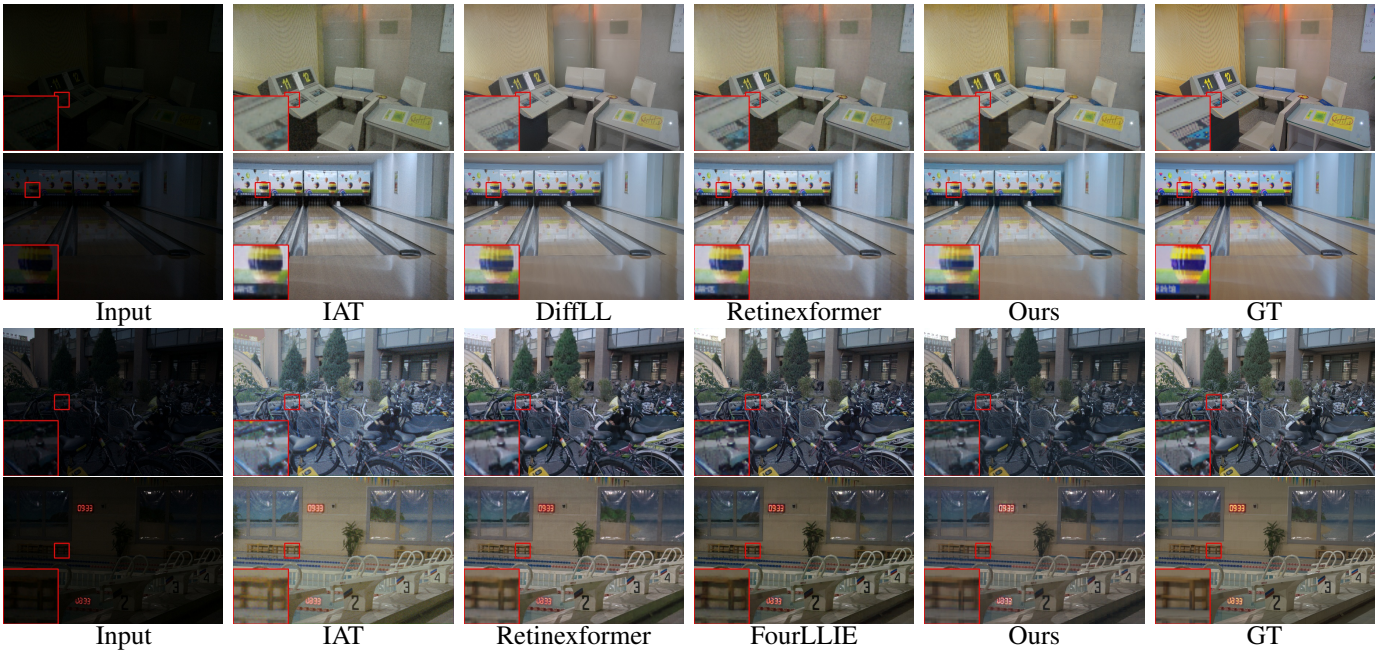


Fig. 5: Visual comparison of different methods on the LOLv1 [8] and LOLv2-real [35] datasets. The first two lines are the result of the LOLv1 dataset and the last two lines are the results of the LOLv2-real dataset. It is obvious that our model has more natural enhancement results and brings a pleasant visual experience.

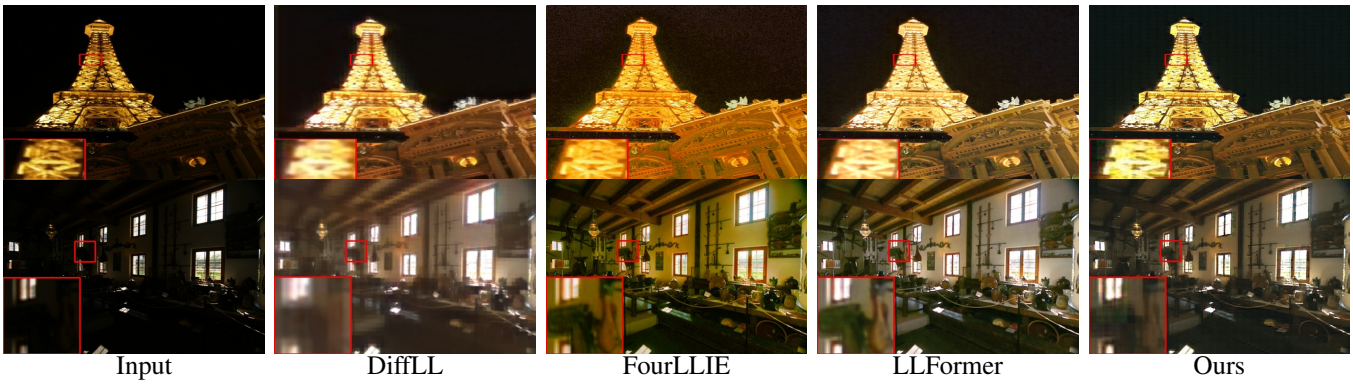


Fig. 6: Comparison on DICM [51] (row 1) and MEF [53] (row 2). SDTL still has good generalization in invisible real scenes.

DeepLPF is poor and the far area is dark. The increment and denoising of our model at the distance of the street view are significantly better than that of DeepLPF. In the comparison of the second column, ours is more accurate on some objects in segmentation results.

B. Ablation Studies

Effect of the number of SDT block: The impact of the depth (*i.e.,* the SDT block number) of Denoising DiT has been investigated on the LOLv2-real dataset, which is shown in Table IV. The results demonstrated that the larger the depth, the stronger the learning ability of the model, it will also bring more parameters. Our model still performed well even with a small number of parameters (depth=2). This demonstrated that we can adjust the depth of the network in time and adapt to different computing environments as needed.

TABLE IV: Effect of the depth (*i.e.,* the number of SDT block) of Denoising DiT on the LOLv2-real dataset.

Depth	PSNR↑	SSIM↑	LPIPS↓	Parameters (M)
2	26.67	0.822	0.215	11.08
4	27.66	0.866	0.148	18.77
6	28.85	0.875	0.131	24.84

Effect of each component in our SDTL: We carry out the following studies to investigate the effectiveness of the key modules in SDTL by deleting different components separately:

- ”w/o SEM-enhance”: Only remove the enhancement stage in SEM;
- ”w/o SEM-fusion”: Only remove the fusion stage in SEM;
- ”w/o SEM”: Remove all parts of SEM, including en-

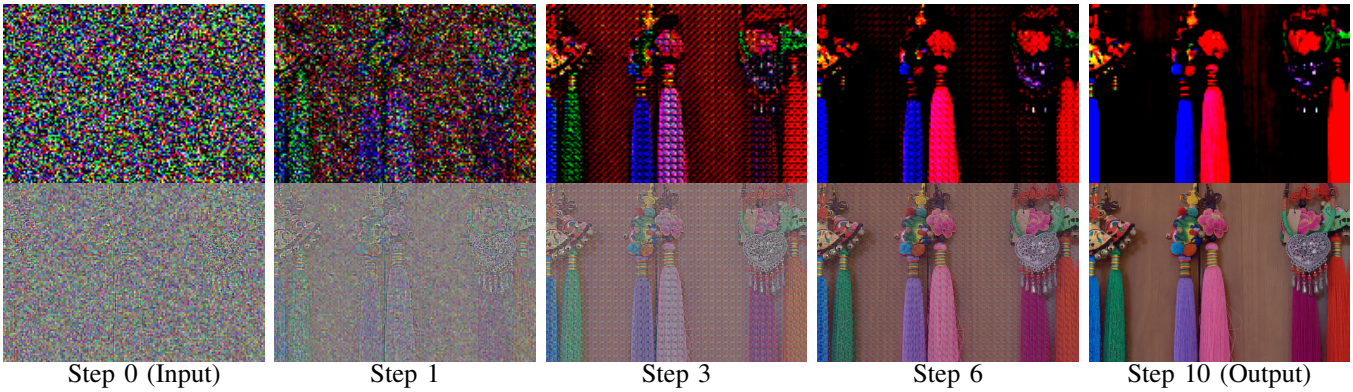


Fig. 7: Visualization of reverse denoising processing in SDTL. The first line represents the frequency information in the denoising step, and the second line represents the corresponding enhanced image. This process shows the powerful ability of the diffusion model in low-light enhancement and gradually removes the noise and artifacts in low-light images.

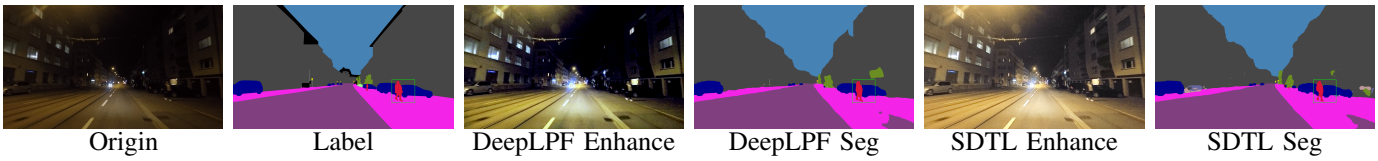


Fig. 8: Visualization of night segmentation on the ACDC [54] dataset. We visualized the enhanced image by the enhancement model and the corresponding predictive segmentation mask.

TABLE V: Ablation studies on the LOLv2-real dataset. The results show the effectiveness of each component in our model.

Methods	PSNR↑	SSIM↑	LPIPS↓
w/o SEM-enhance	26.89	0.819	0.212
w/o SEM-fusion	26.40	0.852	0.160
w/o SEM	25.92	0.763	0.291
w/o DiT	26.65	0.842	0.196
w/o SAB	27.41	0.852	0.160
w ALL	28.85	0.875	0.131

hancement and fusion;

- "w/o DiT": The structure of DiT is not adopted and replaced with the original DDPM;
- "w/o SAB": Remove SAB;
- "w ALL": Keep all components (SDTL)

We conducted experiments on ablation studies on the LOLv2-real dataset, and the results are shown in Table V. The comparison of "w/o SEM-enhance" and "w/o SEM" illustrates the necessity of complementary fusion of different frequency bands in the design of SEM, especially in SSIM and LPIPS, which have improved by 0.056 and 0.052 respectively. Through the comparison of "w/o SEM-fusion" and "w/o SEM", we found that SSIM and LPIPS increased by 0.089 and 0.131 respectively, it shows that the structural prior obviously enhances the texture of high-frequency information, which is beneficial to the recovery of information in the dark environment. Comparing "w/o DiT" and "w/o SAB" with "w ALL" highlights the advantages of the Transformer structure in the diffusion model, which not only delivers improved performance but also offers powerful flexibility and scalability due to the customizable number of blocks, embedding dimensions, and the number of self-attention heads in DiT. This allows for easy adjustments to the complexity of the model, making it more versatile than DDPM. From "w/o SAB", our method was improved by 0.023 and 0.029 respectively in SSIM and LPIPS. This result demonstrated that SAB effectively pays more attention to the texture-rich tokens while reducing the noise area. All in all, our framework brought superior improvement for low-light enhancement.

TABLE VI: Effect of the number of SEM on the LOLv2-real dataset.

SEM Number	PSNR↑	SSIM↑	LPIPS↓
0	25.92	0.763	0.291
1	26.79	0.813	0.237
2	28.85	0.875	0.131

TABLE VII: Analysis of the structure guidance in SEM on the LOLv2-real dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o SEM-enhance	26.89	0.819	0.212
Self-Guidance	27.06	0.833	0.189
Structure-Guidance	28.85	0.875	0.131

Effect of the SEM module: We analyzed the impact of different SEM numbers on the LOLv2 dataset, as shown in Table VI. The results showed that two SEMs have a better effect than one, which shows that the use of SEM in two wavelet transforms gradually restored the details of the frequency and achieved better results in the subsequent denoising network.

TABLE VIII: Effect of the patch size of Denoising DiT on the LOLv2-real dataset.

Patch Size	PSNR↑	SSIM↑	LPIPS↓
2	Out of Memory		
4	28.85	0.875	0.131
8	26.52	0.824	0.228

TABLE IX: Effect of the embedding size of Denoising DiT on the LOLv2-real dataset.

Embedding Size	PSNR↑	SSIM↑	LPIPS↓
192	27.39	0.863	0.162
384	28.85	0.875	0.131
768	27.77	0.863	0.154

Meanwhile, two wavelet transforms reduced the amount of calculation again. Meanwhile, we analyzed the impact of structure guidance in SEM, as shown in Table VII. Quantitative results demonstrated that structure guidance was necessary and achieved superior performance compared to self-guidance. Structure guidance adopts the details of the edge map, guiding the model to recover the high-frequency information from low-light images.

Effect of the patch size and embedding size of Denoising DiT: We constructed the analysis of patch size in the SDT block on the LOLv2-real dataset, as shown in Table VIII. It demonstrated that the smaller the patch size, the superior the network granularity, meanwhile the computational burden will also increase. Specially, the high patch size will be out of memory on the GPUs. Therefore, increasing the patch size within a range achieves better performance of our model. In addition, we also analyzed the embedding size of SDT block, as shown in Table IX. This demonstrated that the best embedding size is 384. Although the performance of our model from 192 to 384 has improved, the continuous improvement resulted in a decline of effect. We need to set the suitable embedding size to get stable model performance.

V. Conclusion

In this paper, we introduce firstly Diffusion Transformer (DiT) into low-light enhancement and design a framework named Structure-guided Diffusion Transformer for Low-light (SDTL). We design Structure Enhancement Module (SEM) and Structure-guided Attention Block (SAB) respectively to improve network performance by using structure prior. The experimental results show that our method achieves advanced performance on 8 benchmark datasets and improves the effect of the segmentation in the night dataset. In the future, we will extend our idea to other tasks for 3D engineering applications [70], [71].

References

[1] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.
[2] X. Yin, Z. Yu, Z. Fei, W. Lv, and X. Gao, "Pe-yolo: Pyramid enhancement network for dark object detection," in *International Conference on Artificial Neural Networks*. Springer, 2023, pp. 163–174.
[3] F. Wang, D. Guo, K. Li, and M. Wang, "Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5345–5353.
[4] F. Wang, D. Guo, K. Li, Z. Zhong, and M. Wang, "Frequency decoupling for motion magnification via multi-level isomorphic architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 984–18 994.
[5] S. M. Pizer, "Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group," in *Proceedings of the first conference on visualization in biomedical computing, Atlanta, Georgia*, vol. 337, 1990, p. 1.
[6] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.
[7] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2782–2790.
[8] W. Y. J. L. Chen Wei, Wenjing Wang, "Deep retinex decomposition for low-light enhancement," in *British Machine Vision Conference*. British Machine Vision Association, 2018.
[9] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE transactions on image processing*, vol. 30, pp. 2340–2349, 2021.
[10] X. Zhang, X. Yin, X. Gao, T. Qiu, L. Wang, G. Yu, Y. Wang, G. Zhang, and J. Li, "Adaptive entropy multi-modal fusion for nighttime lane segmentation," *IEEE Transactions on Intelligent Vehicles*, 2024.
[11] S. Hao, X. Han, Y. Guo, X. Xu, and M. Wang, "Low-light image enhancement with semi-decoupled decomposition," *IEEE transactions on multimedia*, vol. 22, no. 12, pp. 3025–3038, 2020.
[12] Q. Ma, Y. Wang, and T. Zeng, "Retinex-based variational framework for low-light image enhancement and denoising," *IEEE Transactions on Multimedia*, 2022.
[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
[14] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–14, 2023.
[15] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.
[16] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[18] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
[19] S.-C. Huang, F.-C. Cheng, and Y.-S. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE transactions on image processing*, vol. 22, no. 3, pp. 1032–1041, 2012.
[20] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on image processing*, vol. 26, no. 2, pp. 982–993, 2016.
[21] S. Moran, P. Marza, S. McDonagh, S. Parisot, and G. Slabaugh, "Deeplpf: Deep local parametric filters for image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 826–12 835.
[22] S. Lim and W. Kim, "Dslr: Deep stacked laplacian restorer for low-light image enhancement," *IEEE Transactions on Multimedia*, vol. 23, pp. 4272–4284, 2020.
[23] X. Wang, K. Chen, Z. Wang, and W. Huang, "Pmsnet: Parallel multi-scale network for accurate low-light light-field image enhancement," *IEEE Transactions on Multimedia*, 2023.
[24] K. Wu, J. Huang, Y. Ma, F. Fan, and J. Ma, "Cycle-retinex: Unpaired low-light image enhancement via retinex-inline cyclegan," *IEEE Transactions on Multimedia*, 2023.
[25] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 714–17 724.

[26] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu, "Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2654–2662.

[27] C. Wang, H. Wu, and Z. Jin, "Fourllie: Boosting low-light image enhancement by fourier frequency information," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7459–7469.

[28] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit diffusion models for continuous super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 021–10 030.

[29] H. Sahak, D. Watson, C. Saharia, and D. Fleet, "Denoising diffusion probabilistic models for robust image super-resolution in the wild," *arXiv preprint arXiv:2302.07864*, 2023.

[30] C. M. Nguyen, E. R. Chan, A. W. Bergman, and G. Wetzstein, "Diffusion in the dark: A diffusion model for low-light text recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4146–4157.

[31] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 095–13 105.

[32] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 293–16 303.

[33] Y. Yin, D. Xu, C. Tan, P. Liu, Y. Zhao, and Y. Wei, "Cle diffusion: Controllable light enhancement diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8145–8156.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[35] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE Transactions on Image Processing*, vol. 30, pp. 2072–2086, 2021.

[36] J. Hai, Z. Xuan, R. Yang, Y. Hao, F. Zou, F. Lin, and S. Han, "R2rnet: Low-light image enhancement via real-low to real-normal network," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103712, 2023.

[37] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1632–1640.

[38] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3063–3072.

[39] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[40] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5901–5910.

[41] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 683–17 693.

[42] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for fast image restoration and enhancement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 2, pp. 1934–1948, 2022.

[43] Z. Cui, K. Li, L. Gu, S. Su, P. Gao, Z. Jiang, Y. Qiao, and T. Harada, "Illumination adaptive transformer," *arXiv preprint arXiv:2205.14871*, 2022.

[44] C. Li, C.-L. Guo, M. Zhou, Z. Liang, S. Zhou, R. Feng, and C. C. Loy, "Embedding fourier for ultra-high-definition low-light image enhancement," *arXiv preprint arXiv:2302.11831*, 2023.

[45] X. Xu, R. Wang, and J. Lu, "Low-light image enhancement via structure modeling and guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9893–9903.

[46] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 504–12 513.

[47] S. Saini and P. J. Narayanan, "Specularity factorization for low-light enhancement," 2024. [Online]. Available: https://arxiv.org/abs/2404.01998

[48] H. Jiang, A. Luo, X. Liu, S. Han, and S. Liu, "Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models," in *European Conference on Computer Vision*. Springer, 2025, pp. 161–179.

[49] E. Adhikarla, K. Zhang, J. Nicholson, and B. D. Davison, "Expomamba: exploiting frequency ssm blocks for efficient and effective image enhancement," *arXiv preprint arXiv:2408.09650*, 2024.

[50] V. Vonikakis, I. Andreadis, and A. Gasteratos, "Fast centre–surround contrast modification," *IET Image processing*, vol. 2, no. 1, pp. 19–34, 2008.

[51] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation of 2d histograms," *IEEE transactions on image processing*, vol. 22, no. 12, pp. 5372–5384, 2013.

[52] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE transactions on image processing*, vol. 22, no. 9, pp. 3538–3548, 2013.

[53] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, 2015.

[54] C. Sakaridis, D. Dai, and L. Van Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 765–10 775.

[55] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1780–1789.

[56] J. Huang, Y. Liu, F. Zhao, K. Yan, J. Zhang, Y. Huang, M. Zhou, and Z. Xiong, "Deep fourier-based exposure correction network with spatial-frequency interaction," in *European Conference on Computer Vision*. Springer, 2022, pp. 163–180.

[57] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[59] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[60] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[61] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[62] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

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

[64] Z. Cui, K. Li, L. Gu, S. Su, P. Gao, Z. Jiang, Y. Qiao, and T. Harada, "You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction," *arXiv preprint arXiv:2205.14871*, 2022.

[65] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[67] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[68] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.

[69] H. Jie, X. Zuo, J. Gao, W. Liu, J. Hu, and S. Cheng, "Llformer: An efficient and real-time lidar lane detection method based on transformer," in *Proceedings of the 2023 5th International Conference on Pattern Recognition and Intelligent Systems*, 2023, pp. 18–23.
[70] L. Li, F. He, R. Fan, B. Fan, and X. Yan, "3d reconstruction based on hierarchical reinforcement learning with transferability," *Integrated Computer-Aided Engineering*, vol. 30, no. 4, pp. 327–339, 2023.
[71] P. Li, F. He, B. Fan, and Y. Song, "Tpnet: A novel mesh analysis method via topology preservation and perception enhancement," *Computer Aided Geometric Design*, vol. 104, p. 102219, 2023.

Xiangchen Yin received the B.Eng degree in computer science from Qufu Normal University (QFNU), Qufu, China, 2023. He is currently pursuing in the Master degree in computer technology with University of Science and Technology of China (USTC), Hefei, China. His research is mainly focused on computer vision, automatic driving and diffusion model.

Zhenda Yu receive the B.Eng degree in computer science and technology from Yangtze University, Jingzhou, China, 2022. He is currently pursuing in the Master degree in computer science and technology with Anhui University, Hefei, China. His research is mainly focused on time series, visual perception and multimodal.

Longtao Jiang received the B.Eng degree in communication Engineering from Hohai University (HHU), Nanjing, China, 2023. He is currently pursuing in the Master degree in information and communication engineering with University of Science and Technology of China (USTC), Hefei, China. His research is mainly focused on multimodal alignment, vision generation and video understanding.

Xun Yang Department of Electronic Engineering and Information Science, School of Information Science and Technology, University of Science and Technology of China, Hefei, China. Xun Yang received the Ph.D. degree from the Hefei University of Technology, Hefei, China, in 2017. He is currently a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC). From 2015 to 2017, he visited the University of Technology Sydney (UTS), Australia, as a Joint Ph.D. Student. He was a Research Fellow with the NExT++ Research Center, National University of Singapore (NUS), from 2018 to 2021. His current research interests include information retrieval, cross-media analysis and reasoning, and computer vision. He currently serves as an Associate Editor for the IEEE Transactions on Big Data journal and the Multimedia Systems journal.

Xin Gao received his bachelor degree in Computer Science and Technology from China University of Mining & Technology, Beijing in 2018. He is a Ph.D candidate majoring in Computer Science and Technology in China University of Mining & Technology, Beijing. In addition, he associate in the State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing, China. His research interests are pattern recognition, multi-modal fusion, image processing.

Xiao Sun (Senior Member, IEEE) was born in 1980. He received the M.E. degree from the Department of Computer Sciences and Engineering, Dalian University of Technology, Dalian, China, in 2004, and the dual Ph.D. degree from the University of Tokushima, Tokushima, Japan, in 2009, and the Dalian University of Technology, in 2010. His field of study was natural language processing. He is currently working as a Professor with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, Hefei University of Technology, Hefei, China. He is also a researcher with the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence. His research interests include affective computing, natural language processing, machine learning and human-machine interaction.

Zhi Liu Zhi Liu (S'11-M'14-SM'19) received the Ph.D. degree in informatics in National Institute of Informatics. He is currently an Associate Professor at The University of Electro-Communications. His research interest includes video network transmission, mobile edge computing and IoT. He is now an editorial board member of IEEE Transactions on Multimedia, IEEE Network and IEEE Internet of Things Journal. He is a senior member of IEEE.

Dear Editors and Reviewers

Paper: MM-020552 Structure-guided Diffusion Transformer for Low-Light Image Enhancement

First of all, we would like to thank the editor and two anonymous reviewers for the valuable comments. They have helped to improve the paper in a number of ways. We have carefully revised the paper by addressing all comments found in the reviews. Below is a point-by-point statement of changes and replies in response to the reviewers' comments.

Response to reviewer 1

• **Comment:** The paper discusses the challenges in low-light image enhancement, where current methods often amplify noise while recovering details, leading to poor visual quality. Which reports that the SDTL method achieves state-of-the-art (SOTA) performance on several popular datasets, demonstrating its effectiveness in improving image quality and the potential of DiT in low-light enhancement tasks. However， The author proposed a low illumination image enhancement method based on the network, but it is lack of theoretical and structural innovation, and the experimental results are not rich enough.

• **Response:** Thank you for your review and feedback on our paper. We understand your concerns about our work, especially in terms of theoretical innovation. In response to your comments, we have made improvements in the following aspects:

Theoretical innovation: We elaborated on the theoretical basis of the SDTL method in detail, and achieved effective low-light enhancement in the DiT-based model proposed the Structure Enhancement Module (SEM) and the Structure-guided Attention Block (SAB). The SEM innovatively introduces structural priors, optimizes the structural information in high-frequency subbands through a two-stage enhancement and fusion strategy, and significantly improves the effect of low-light image enhancement. The SAB guides the self-attention mechanism to pay more attention to texture-rich areas, reducing the interference of noise areas, thereby further improving the model's recovery of details. We explained how these modules use structural priors to improve the enhancement effect of low-light images, and compared them with existing methods to emphasize their innovation.

Rich experimental results: We provide more experimental results to verify the performance of SDTL in different datasets and under different low-light conditions. In addition, we have added more analytical experiments of the SDTL model to prove the effectiveness of our network structure in the new revised version. Specifically, we added the analysis of the effects of patch size and embedding size in SDTL and the analysis of structure guidance in SEM (change in page 8, 9). It demonstrated that the smaller the patchsize, the superior the network granularity, meanwhile the computational burden will also increase. Therefore, increasing the patch size within a range achieves the performance of our model. Quantitative results demonstrated that structure guidance was necessary and achieved superior performance compared to self-guidance. Structure guidance adopts the details of the edge map, guiding the model to recover the high-frequency information from low-light images.

Finally, we have further polished the language of the entire paper and revised the figures (Fig. 1, 2, 4). Meanwhile, we have enhanced the comparative experiments to further demonstrate the effectiveness of our model.

We hope that these revisions will satisfy your requirements for theoretical and experimental results and demonstrate the innovation and effectiveness of our work. Thank you again for your attention and suggestions to our work, and we look forward to your further feedback.

Changes in page 3:

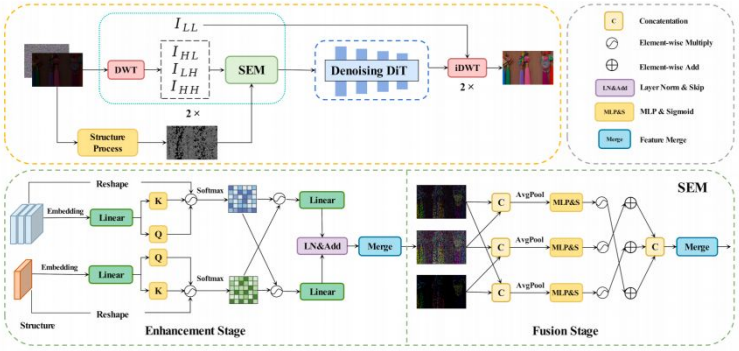


Fig. 2: Overview of SDTL. Our model compresses features to improve the efficiency of inference through wavelet transform. We design a Structure Enhancement Module (SEM) to enhance structural information under different frequency bands, which is divided into two stages: **Enhancement** and **Fusion**. In addition, we propose a Structure-guided Attention Block (SAB) to pay attention to the texture-riched tokens in the noise prediction network.

Changes in page 5:

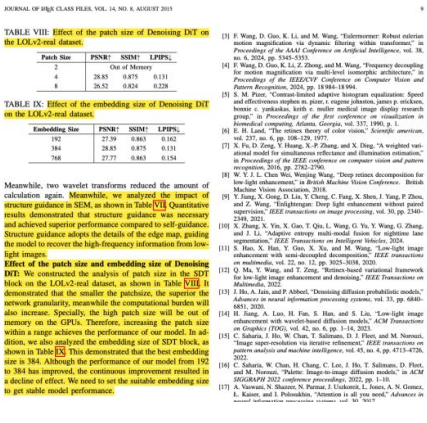
TABLE I: Quantitative comparisons of different methods on the LOLv1 [8], LOLv2 [35] and LSRW [36] datasets. The best results and second-best results are highlighted in blue and green, respectively. Note that we obtained these results either from the corresponding papers, or by running the officially released models, and the missing results on LSRW are marked as "-".

Methods	References	LSRW				LOLv1				LOLv2-real			
		PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
RetinexNet [8]	BMVC'18	15.61	0.414	0.454	108.350	16.77	0.462	0.474	126.266	18.37	0.723	0.365	133.905
KinD [37]	MM'19	-	-	-	-	20.87	0.799	0.207	104.632	17.54	0.669	0.375	137.346
DRBN [38]	CVPR'20	16.73	0.507	0.457	80.727	19.86	0.834	0.155	98.732	20.13	0.830	0.147	89.085
EnlightenGAN [9]	TIP'21	17.11	0.463	0.406	69.033	17.48	0.652	0.322	94.704	18.64	0.677	0.309	84.044
Restormer [39]	CVPR'22	16.30	0.453	0.427	69.219	20.61	0.797	0.288	72.998	24.91	0.851	0.264	58.649
URetinex-Net [40]	CVPR'22	18.27	0.518	0.419	66.871	19.84	0.824	0.237	52.383	21.09	0.858	0.208	49.836
Uformer [41]	CVPR'22	16.59	0.494	0.435	82.299	19.00	0.741	0.354	109.351	18.44	0.759	0.347	98.138
MIRNet [42]	TPAMI'22	16.47	0.477	0.430	93.811	24.14	0.842	0.131	69.179	20.36	0.782	0.317	49.108
SNR-Aware [25]	CVPR'22	16.49	0.505	0.419	65.807	24.61	0.842	0.152	55.121	21.48	0.849	0.157	54.532
IAT [43]	BMVC'22	20.81	0.565	0.467	80.499	23.38	0.809	0.134	67.412	23.50	0.824	0.191	62.153
LLFormer [26]	AAAI'23	20.69	0.560	0.518	96.782	25.76	0.823	0.167	65.271	26.20	0.819	0.209	63.432
FourLLIE [27]	MM'23	18.61	0.505	0.316	73.55	20.22	0.766	0.250	91.793	22.34	0.847	0.051	89.334
UHDFour [44]	ICLR'23	17.30	0.529	0.443	62.032	23.09	0.821	0.259	56.912	21.79	0.854	0.292	60.837
SMG [45]	CVPR'23	19.04	0.568	0.392	101.56	23.68	0.826	0.118	58.846	24.62	0.867	0.148	78.582
Retinexformer [46]	ICCV'23	20.15	0.534	0.336	70.36	25.15	0.845	0.131	71.148	22.80	0.840	0.171	62.439
DiffLL [14]	TOG'23	19.28	0.552	0.350	45.294	26.33	0.845	0.217	48.114	28.85	0.876	0.207	45.359
RSFNet [47]	CVPR'24	-	-	-	-	22.15	0.860	0.265	-	21.59	0.843	0.278	-
LightenDiffusion [48]	ECCV'24	18.55	0.539	0.311	-	20.45	0.803	0.192	-	-	-	-	-
ExpoMamba [49]	ICML'24	-	-	-	-	25.77	0.860	0.212	89.210	28.04	0.885	0.232	85.920
SDTL (Ours)	-	21.23	0.576	0.392	42.569	27.34	0.862	0.118	50.432	28.85	0.875	0.131	43.237

Change in page 8,9:

TABLE VII: Analysis of the structure guidance in SEM on the LOLv2-real dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o SEM-enhance	26.89	0.819	0.212
Self-Guidance	27.06	0.833	0.189
Structure-Guidance	28.85	0.875	0.131



JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015 9

TABLE VIII: Effect of the patch size of Denoising DiT on the LOLv2-real dataset.

Patch Size	PSNR↑	SSIM↑	LPIPS↓
2	Out of Memory		
4	26.85	0.875	0.131
8	26.52	0.824	0.228

TABLE IX: Effect of the embedding size of Denoising DiT on the LOLv2-real dataset.

Embedding Size	PSNR↑	SSIM↑	LPIPS↓
192	27.39	0.863	0.162
384	28.85	0.875	0.131
768	27.77	0.863	0.154

Meanwhile, two wavelet transforms reduced the amount of calculation again. Meanwhile, we analyzed the impact of structure guidance in SEM, as shown in Table VII. Quantitative results demonstrated that structure guidance was necessary and achieved superior performance compared to self-guidance. Structure guidance adopts the details of the edge map, guiding the model to recover the high-frequency information from low-light images.

Effect of the patch size and embedding size of Denoising DiT: We constructed the analysis of patch size in the SDT block on the LOLv2-real dataset, as shown in Table VIII. It demonstrated that the smaller the patchsize, the superior the network granularity, meanwhile the computational burden will also increase. Specially, the high patch size will be out of memory on the GPUs. Therefore, increasing the patch size within a range achieves the performance of our model. In addition, we also analyzed the embedding size of SDT block, as shown in Table IX. This demonstrated that the best embedding size is 384. Although the performance of our model from 192 to 384 has improved, the continuous improvement resulted in a decline of effect. We need to set the suitable embedding size to get stable model performance.

[3] F. Wang, D. Guo, K. Li, and M. Wang, "Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5345–5353.
[4] F. Wang, D. Guo, K. Li, Z. Zhong, and M. Wang, "Frequency decoupling for motion magnification via multi-level isomorphic architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 984–18 994.
[5] S. M. Pizer, "Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group," in *Proceedings of the first conference on visualization in biomedical computing, Atlanta, Georgia*, vol. 337, 1990, p. 1.
[6] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.
[7] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2782–2790.
[8] W. Y. J. L. Chen Wei, Wenjing Wang, "Deep retinex decomposition for low-light enhancement," in *British Machine Vision Conference*. British Machine Vision Association, 2018.
[9] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE transactions on image processing*, vol. 30, pp. 2340–2349, 2021.
[10] X. Zhang, X. Yin, X. Gao, T. Qiu, L. Wang, G. Yu, Y. Wang, G. Zhang, and J. Li, "Adaptive entropy multi-modal fusion for nighttime lane segmentation," *IEEE Transactions on Intelligent Vehicles*, 2024.
[11] S. Hao, X. Han, Y. Guo, X. Xu, and M. Wang, "Low-light image enhancement with semi-decoupled decomposition," *IEEE transactions on multimedia*, vol. 22, no. 12, pp. 3025–3038, 2020.
[12] Q. Ma, Y. Wang, and T. Zeng, "Retinex-based variational framework for low-light image enhancement and denoising," *IEEE Transactions on Multimedia*, 2022.
[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
[14] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–14, 2023.
[15] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.
[16] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in*

Response to reviewer 2

• **Comment (1):** In the framework illustrated in Figure 2, the "Fusion Stage" part is not consistent with the one described in Sec.3B.

• **Response:** Thanks for your careful review. We have corrected the Fusion Stage of SEM in Fig. 2. Meanwhile, we have beautified Fig. 1, Fig. 2 and Fig. 4.

Changes in page 3:

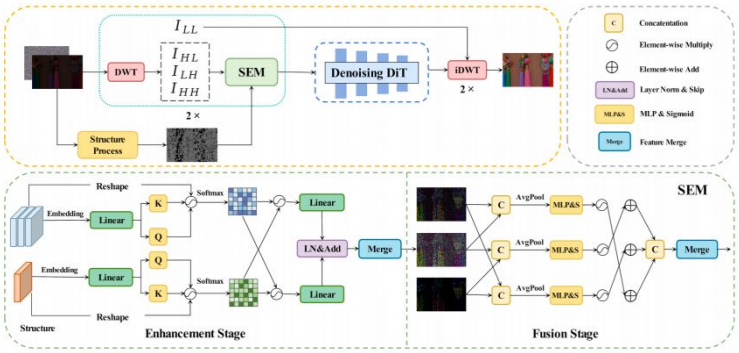


Fig. 2: Overview of SDTL. Our model compresses features to improve the efficiency of inference through wavelet transform. We design a Structure Enhancement Module (SEM) to enhance structural information under different frequency bands, which is divided into two stages: **Enhancement** and **Fusion**. In addition, we propose a Structure-guided Attention Block (SAB) to pay attention to the texture-riched tokens in the noise prediction network.

Changes in page 1:

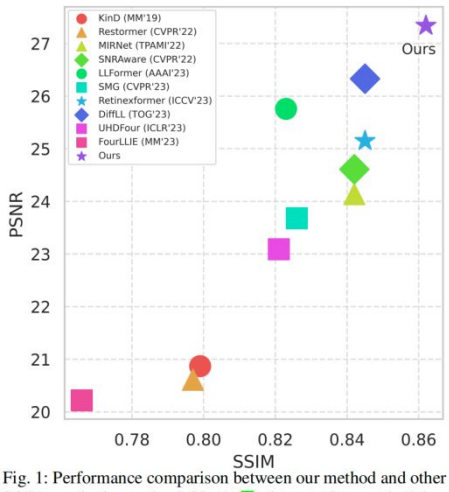


Fig. 1: Performance comparison between our method and other SOTA methods on the LOLv1 [8] dataset. Our method has achieved the best results on both PSNR and SSIM, reaching 27.34 and 0.862 respectively. The legend on the right shows each model and source.

Changes in Page 5:

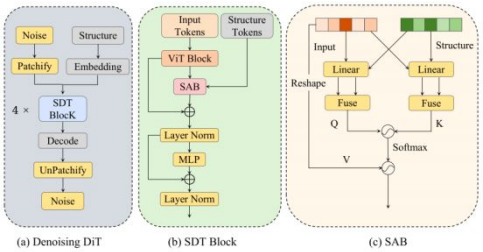


Fig. 4: Details of Denoising Diffusion Transformer (a). SDT block consists of a ViT block and a SAB block (b). SAB integrates the original token with the structural token to make the network focus on the textured-riched token (c).

• **Comment (2):** The low-frequency subband of the DWT output is involved in the final iDWT. There should be an arrow pointing from the DWT to the iDWT indicating the low-frequency subband.

• **Response:** We have added an arrow from LL(low-frequency subband) to iDWT in Fig.2 to fully represent iDWT.

• **Comment (3):** As described in Sec.3, after performing DWT on the input low-light image, the authors only use the neural network to process its high-frequency subbands. Before output, iDWT is performed on the three processed high-frequency subbands and the low-frequency subband of the input low-light image to obtain the output normal-light image. However, the low-frequency subband of the unprocessed low-light image still has the problem of insufficient brightness. How do the authors use iDWT to synthesize a normal-light image based on the low-frequency subband with insufficient brightness?

• **Response:** You have raised an interesting concern. In this paper, we adopt DWT as a filter to distrill high-frequency sub-band and an enhancement workflow is designed to process it, because high-frequency usb-band contains most semantic information of the image. The low-frequency sub-band contains the overall structure of the image and brightness has little influence on it, thus it can be directly used to rebuild the lightened image as default information.

• **Comment (4):** In the "effect of the number of SEM modules" part of the ablation experiments, the authors analyzed the data in Table 5 belonging to another ablation experiment, which seems to be an arrangement error.

• **Response:** We have adjusted the paper layout.

Changes in page 8:

hancement and fusion;
- "w/o DiT": The structure of DiT is not adopted and replaced with the original DDPM;
- "w/o SAB": Remove SAB;
- "w ALL": Keep all components (SDTL)

We conducted experiments on ablation studies on the LOLv2-real dataset, and the results are shown in Table V. The comparison of "w/o SEM-enhance" and "w/o SEM" illustrates the necessity of complementary fusion of different frequency bands in the design of SEM, especially in SSIM and LPIPS, which have improved by 0.056 and 0.052 respectively. Through the comparison of "w/o SEM-fusion" and "w/o SEM", we found that SSIM and LPIPS increased by 0.089 and 0.131 respectively, it shows that the structural prior obviously enhances the texture of high-frequency information, which is beneficial to the recovery of information in the dark environment. Comparing "w/o DiT" and "w/o SAB" with "w ALL" highlights the advantages of the Transformer structure in the diffusion model, which not only delivers improved performance but also offers powerful flexibility and scalability due to the customizable

Method	PSNR↑	SSIM↑	LPIPS↓
w/o SEM-enhance	26.89	0.819	0.212
Self-Guidance	27.06	0.833	0.189
Structure-Guidance	28.85	0.875	0.131

number of blocks, embedding dimensions, and the number of self-attention heads in DiT. This allows for easy adjustments to the complexity of the model, making it more versatile than DDPM. From "w/o SAB", our method was improved by 0.023 and 0.029 respectively in SSIM and LPIPS. This result demonstrated that SAB can effectively pay more attention to the texture-rich tokens while reducing the noise area. All in all, our framework bringed great improvements to low-light enhancement.

Effect of the SEM module: We analyzed the impact of different SEM numbers on the LOLv2 dataset, as shown in Table VI. The results showed that two SEMs have a better effect than one, which shows that the use of SEM in two wavelet transforms gradually restored the details of the frequency and achieved better results in the subsequent denoising network.

• **Comment (5):** In Table 1, there are a lot of blanks in the LSRW part. It is suggested that the authors reproduce these previous methods on the LSRW dataset and compare their results with the proposed SDTL to prove the superiority of the proposed SDTL on multiple datasets.

• **Response:** We evaluate several new methods on the LSRW dataset and the results are complemented in TABLE 1. Clearly, our SDTL still achieved advanced performance on LSRW.

Changes in page 5:

TABLE I: Quantitative comparisons of different methods on the LOLv1 [8], LOLv2 [35] and LSRW [36] datasets. The best results and second-best results are highlighted in blue and green, respectively. Note that we obtained these results either from the corresponding papers, or by running the officially released models, and the missing results on LSRW are marked as "-".

Methods	References	LSRW				LOLv1				LOLv2-real			
		PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
RetinexNet [8]	BMVC'18	15.61	0.414	0.454	108.350	16.77	0.462	0.474	126.266	18.37	0.723	0.365	133.905
KinD [37]	MM'19	-	-	-	-	20.87	0.799	0.207	104.632	17.54	0.669	0.375	137.346
DRBN [38]	CVPR'20	16.73	0.507	0.457	80.727	19.86	0.834	0.155	98.732	20.13	0.830	0.147	89.085
EnlightenGAN [9]	TIP'21	17.11	0.463	0.406	69.033	17.48	0.652	0.322	94.704	18.64	0.677	0.309	84.044
Restormer [39]	CVPR'22	16.30	0.453	0.427	69.219	20.61	0.797	0.288	72.998	24.91	0.851	0.264	58.649
URetinex-Net [40]	CVPR'22	18.27	0.518	0.419	66.871	19.84	0.824	0.237	52.383	21.09	0.858	0.208	49.836
Uformer [41]	CVPR'22	16.59	0.494	0.435	82.299	19.00	0.741	0.354	109.351	18.44	0.759	0.347	98.138
MIRNet [42]	TPAMI'22	16.47	0.477	0.430	93.811	24.14	0.842	0.131	69.179	20.36	0.782	0.317	49.108
SNR-Aware [25]	CVPR'22	16.49	0.505	0.419	65.807	24.61	0.842	0.152	55.121	21.48	0.849	0.157	54.532
IAT [43]	BMVC'22	20.81	0.565	0.467	80.499	23.38	0.809	0.134	67.412	23.50	0.824	0.191	62.153
LLFormer [26]	AAAI'23	20.69	0.560	0.518	96.782	25.76	0.823	0.167	65.271	26.20	0.819	0.209	63.432
FourLLIE [27]	MM'23	18.61	0.505	0.316	73.55	20.22	0.766	0.250	91.793	22.34	0.847	0.051	89.334
UHDFour [44]	ICLR'23	17.30	0.529	0.443	62.032	23.09	0.821	0.259	56.912	21.79	0.854	0.292	60.837
SMG [45]	CVPR'23	19.04	0.568	0.392	101.56	23.68	0.826	0.118	58.846	24.62	0.867	0.148	78.582
Retinexformer [46]	ICCV'23	20.15	0.534	0.336	70.36	25.15	0.845	0.131	71.148	22.80	0.840	0.171	62.439
DiffLL [14]	TOG'23	19.28	0.552	0.350	45.294	26.33	0.845	0.217	48.114	28.85	0.876	0.207	45.359
RSFNet [47]	CVPR'24	-	-	-	-	22.15	0.860	0.265	-	21.59	0.843	0.278	-
LightenDiffusion [48]	ECCV'24	18.55	0.539	0.311	-	20.45	0.803	0.192	-	-	-	-	-
ExpoMamba [49]	ICML'24	-	-	-	-	25.77	0.860	0.212	89.210	28.04	0.885	0.232	85.920
SDTL (Ours)	-	21.23	0.576	0.392	42.569	27.34	0.862	0.118	50.432	28.85	0.875	0.131	43.237

Response to reviewer 3

• **Comment (1):** For technical novelty, the authors are suggested to further emphasize new or novelty of this proposed method through theory part, in order to show the contributions of this method. For example, this work skillfully combines several concepts, approaches, techniques and components, such as: Diffusion Transformer; Low-Light Enhancement; Low-level Vision. It is a typical combination novelty and/or increment novelty, which can be further highlighted to show the contributions and/or advantages of the proposed method.

• **Response:** Thank you for your valuable comments, pointing out the need to further emphasize the novelty and technical contributions of our method. We understand that it is crucial to clearly demonstrate the combined innovation of new concepts, methods, techniques, and components of our method to highlight the contribution of this study. We have described the innovations of our approach in details in the new revision. The following is our further elaboration and emphasis on the technical novelty:

Our work introduces the Diffusion Transformer (DiT) to the field of low-light image enhancement for the first time. We not only inherit the long-range dependency capture capability of the Transformer, but also flexibly control the model size by adjusting model parameters such as the number of blocks and embedding dimension. At the same time, the diffusion model gradually recovers a clear image from noisy data through an iterative process that simulates the data distribution. In the context of low-light image enhancement, this iterative refinement process is particularly important because it can maintain the detail clarity of the image while enhancing the image brightness and contrast, avoiding the problems of over-smoothing and noise amplification. The structure enhancement module (SEM) and structure-guided attention block (SAB) we proposed are part of the innovations of this method. The combination of SEM, which enhances high-frequency information through structural priors, and SAB, which guides the self-attention mechanism to focus on texture-rich regions, is novel in low-light image enhancement. By introducing wavelet transform, we not only improve the inference efficiency of the model, but also capture the multi-directional frequency information of the image, which is an innovative attempt in the field of low-light image enhancement. We conduct extensive experiments on 8 popular benchmark datasets to verify the effectiveness of our method. In addition, we apply SDTL to the task of nighttime semantic segmentation to demonstrate its potential and advantages in practical applications.

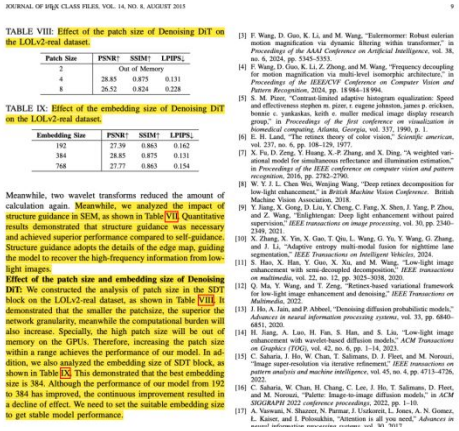
In addition, we added SDTL analysis experiments to further demonstrate the effectiveness of our innovation. Specifically, we analyzed the impact of SDTL on patch size and embedding size and the

necessity of SEM structure innovation in SDTL. It demonstrated that the smaller the patchsize, the superior the network granularity, meanwhile the computational burden will also increase. Therefore, increasing the patch size within a range achieves the performance of our model. Quantitative results demonstrated that structure guidance was necessary and achieved superior performance compared to self-guidance. Structure guidance adopts the details of the edge map, guiding the model to recover the high-frequency information from low-light images.

Change in page 8,9:

TABLE VII: Analysis of the structure guidance in SEM on the LOLv2-real dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o SEM-enhance	26.89	0.819	0.212
Self-Guidance	27.06	0.833	0.189
Structure-Guidance	28.85	0.875	0.131



JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015 9

TABLE VIII: Effect of the patch size of Denoising DiT on the LOLv2-real dataset.

Patch Size	PSNR↑	SSIM↑	LPIPS↓
2	Out of Memory		
4	28.85	0.875	0.131
8	26.52	0.824	0.228

TABLE IX: Effect of the embedding size of Denoising DiT on the LOLv2-real dataset.

Embedding Size	PSNR↑	SSIM↑	LPIPS↓
192	27.39	0.863	0.162
384	28.85	0.875	0.131
768	27.77	0.863	0.154

Meanwhile, two wavelet transforms reduced the amount of calculation again. Meanwhile, we analyzed the impact of structure guidance in SEM, as shown in Table VII. Quantitative results demonstrated that structure guidance was necessary and achieved superior performance compared to self-guidance. Structure guidance adopts the details of the edge map, guiding the model to recover the high-frequency information from low-light images.

Effect of the patch size and embedding size of Denoising DiT: We constructed the analysis of patch size in the SDT block on the LOLv2-real dataset, as shown in Table VIII. It demonstrated that the smaller the patchsize, the superior the network granularity, meanwhile the computational burden will also increase. Specially, the high patch size will be out of memory on the GPUs. Therefore, increasing the patch size within a range achieves the performance of our model. In addition, we also analyzed the embedding size of SDT block, as shown in Table IX. This demonstrated that the best embedding size is 384. Although the performance of our model from 192 to 384 has improved, the continuous improvement resulted in a decline of effect. We need to set the suitable embedding size to get stable model performance.

[3] F. Wang, D. Guo, K. Li, and M. Wang, "Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5345–5353.

[4] F. Wang, D. Guo, K. Li, Z. Zhong, and M. Wang, "Frequency decoupling for motion magnification via multi-level isomorphic architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 984–18 994.

[5] S. M. Pizer, "Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group," in *Proceedings of the first conference on visualization in biomedical computing, Atlanta, Georgia*, vol. 337, 1990, p. 1.

[6] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.

[7] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2782–2790.

[8] W. Y. J. L. Chen Wei, Wenjing Wang, "Deep retinex decomposition for low-light enhancement," in *British Machine Vision Conference*. British Machine Vision Association, 2018.

[9] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE transactions on image processing*, vol. 30, pp. 2340–2349, 2021.

[10] X. Zhang, X. Yin, X. Gao, T. Qiu, L. Wang, G. Yu, Y. Wang, G. Zhang, and J. Li, "Adaptive entropy multi-modal fusion for nighttime lane segmentation," *IEEE Transactions on Intelligent Vehicles*, 2024.

[11] S. Hao, X. Han, Y. Guo, X. Xu, and M. Wang, "Low-light image enhancement with semi-decoupled decomposition," *IEEE transactions on multimedia*, vol. 22, no. 12, pp. 3025–3038, 2020.

[12] Q. Ma, Y. Wang, and T. Zeng, "Retinex-based variational framework for low-light image enhancement and denoising," *IEEE Transactions on Multimedia*, 2022.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[14] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–14, 2023.

[15] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[16] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in*

• **Comment (2):** For technical details, according to section "section III. METHODS", "section A. Conditional Diffusion Model", "section B. Structure Enhancement Module", "section (1) Enhancement", "section (2) Fusion", "section C. Denoising of Diffusion Transformer", "Fig. 2: Overview of SDTL", "Fig. 3: Visualization of hot feature for SEM" and "Fig. 4: Details of Denoising Diffusion Transformer", the proposed method combines multiple components and/or modules and/or steps. Thus, it would be better to have more discussions of time and space complexity for the proposed approach with theory analysis and/or with experiment analysis. The model complexity has great impact on the evaluation. The heavyweight models are generally better the lightweight ones. This could be further clarified.

• **Response:** We have conducted experiments to verify it in section "EXPERIMENTS". We discuss the parameters of the model in "TABLE IV" of the experimental section. It can be seen that our model has PSNR=26.67 when depth=2, and PSNR=28.85 when depth=6. This showed that our model achieves advanced performance with a smaller number of parameters. On the other hand, our time complexity is not an advantage. We test the model in a single image, it need 0.3 second to predict the clear image. This is because our diffusion model uses normal Diffusion Transformer (DiT) and the subsequent Masked Diffusion Transformer (MDT) [ICCV2023] significantly reduces the inference speed through mask modeling, which is also the focus of our later research.

Changes in page 7:

TABLE IV: Effect of the depth (*i.e.*, the number of SDT block) of Denoising DiT on the LOLv2-real dataset.

Depth	PSNR↑	SSIM↑	LPIPS↓	Parameters (M)
2	26.67	0.822	0.215	11.08
4	27.66	0.866	0.148	18.77
6	28.85	0.875	0.131	24.84

• **Comment (3):**For experimental details, the authors evaluate the performance of the system, which are interesting and convincing. However, according to section "IV. EXPERIMENTS" and "TABLE I: Quantitative comparisons of different methods", the authors compared the proposed approach with a number of approaches from 2018 to 2023 (such as [12] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, Low-light image enhancement with wavelet-based diffusion models, ACM Transactions on Graphics (TOG), vol. 42, no. 6, pp. 1 – 14, 2023. [44] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, Retinexformer: One-stage retinex-based transformer for low-light image enhancement, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12 504 – 12 513. [43] X. Xu, R. Wang, and J. Lu, "Low-light image enhancement via structure modeling and guidance," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9893 – 9903) It would be better to compare with the latest works from 20024 in the same task. This comment is optional but not necessary.

• **Response:** Thanks for your suggestion. We have compared two latest works from 2024 (RSFNet, LightenDiffusion, ExpoMamba) with our model to improve the advanced peroformance of SDTL, experiment results are shown in section IV.A. In addition, we complement the performance of other methods on the LSRW dataset, which further demonstrates the advancedness of our method.

Changes in page 5:

TABLE I: Quantitative comparisons of different methods on the LOLv1 [8], LOLv2 [35] and LSRW [36] datasets. The best results and second-best results are highlighted in blue and green, respectively. Note that we obtained these results either from the corresponding papers, or by running the officially released models, and the missing results on LSRW are marked as "-".

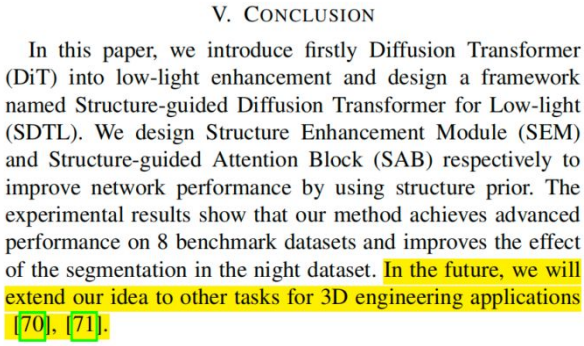
Methods	References	LSRW				LOLv1				LOLv2-real			
		PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
RetinexNet [8]	BMVC'18	15.61	0.414	0.454	108.350	16.77	0.462	0.474	126.266	18.37	0.723	0.365	133.905
KinD [37]	MM'19	-	-	-	-	20.87	0.799	0.207	104.632	17.54	0.669	0.375	137.346
DRBN [38]	CVPR'20	16.73	0.507	0.457	80.727	19.86	0.834	0.155	98.732	20.13	0.830	0.147	89.085
EnlightenGAN [9]	TIP'21	17.11	0.463	0.406	69.033	17.48	0.652	0.322	94.704	18.64	0.677	0.309	84.044
Restormer [39]	CVPR'22	16.30	0.453	0.427	69.219	20.61	0.797	0.288	72.998	24.91	0.851	0.264	58.649
URetinex-Net [40]	CVPR'22	18.27	0.518	0.419	66.871	19.84	0.824	0.237	52.383	21.09	0.858	0.208	49.836
Uformer [41]	CVPR'22	16.59	0.494	0.435	82.299	19.00	0.741	0.354	109.351	18.44	0.759	0.347	98.138
MIRNet [42]	TPAMI'22	16.47	0.477	0.430	93.811	24.14	0.842	0.131	69.179	20.36	0.782	0.317	49.108
SNR-Aware [25]	CVPR'22	16.49	0.505	0.419	65.807	24.61	0.842	0.152	55.121	21.48	0.849	0.157	54.532
IAT [43]	BMVC'22	20.81	0.565	0.467	80.499	23.38	0.809	0.134	67.412	23.50	0.824	0.191	62.153
LLFormer [26]	AAAI'23	20.69	0.560	0.518	96.782	25.76	0.823	0.167	65.271	26.20	0.819	0.209	63.432
FourLLIE [27]	MM'23	18.61	0.505	0.316	73.55	20.22	0.766	0.250	91.793	22.34	0.847	0.051	89.334
UHDFour [44]	ICLR'23	17.30	0.529	0.443	62.032	23.09	0.821	0.259	56.912	21.79	0.854	0.292	60.837
SMG [45]	CVPR'23	19.04	0.568	0.392	101.56	23.68	0.826	0.118	58.846	24.62	0.867	0.148	78.582
Retinexformer [46]	ICCV'23	20.15	0.534	0.336	70.36	25.15	0.845	0.131	71.148	22.80	0.840	0.171	62.439
DiffLL [14]	TOG'23	19.28	0.552	0.350	45.294	26.33	0.845	0.217	48.114	28.85	0.876	0.207	45.359
RSFNet [47]	CVPR'24	-	-	-	-	22.15	0.860	0.265	-	21.59	0.843	0.278	-
LightenDiffusion [48]	ECCV'24	18.55	0.539	0.311	-	20.45	0.803	0.192	-	-	-	-	-
ExpoMamba [49]	ICML'24	-	-	-	-	25.77	0.860	0.212	89.210	28.04	0.885	0.232	85.920
SDTL (Ours)	-	21.23	0.576	0.392	42.569	27.34	0.862	0.118	50.432	28.85	0.875	0.131	43.237

• **Comment (4):** For reproducibility, it would be better to open the source code (algorithm source code and/or data preparing code) on GitHub. The open source will greatly contribute to the community of related research areas. This comment is optional but not necessary.

• **Response:** The code is available at https://github.com/XiangchenYin/SDTL.

• **Comment (5):** For generality exploration, besides the 2D task, in order to show the beauty of the proposed idea, it would be better to extend the proposed idea to 3D tasks in section of "Conclusion and Future Work", such as "In the future, we will extend our idea to other tasks for 3D engineering applications [3D Reconstruction based on Hierarchical Reinforcement Learning with Transferability. http://dx.doi.org/10.3233/ICA-230710] [TPNet: A novel mesh analysis method via topology preservation and perception enhancement, https://doi.org/10.1016/j.cagd.2023.102219.]". The above generality exploration will be interesting to the readers.

• **Response:** We have further discussed about application scenarios and scalability of SDTL in section "CONCLUSION" such as 3D engineering applications. We will cite the inference in the final version. Changes in page 9:



V. CONCLUSION

In this paper, we introduce firstly Diffusion Transformer (DiT) into low-light enhancement and design a framework named Structure-guided Diffusion Transformer for Low-light (SDTL). We design Structure Enhancement Module (SEM) and Structure-guided Attention Block (SAB) respectively to improve network performance by using structure prior. The experimental results show that our method achieves advanced performance on 8 benchmark datasets and improves the effect of the segmentation in the night dataset. In the future, we will extend our idea to other tasks for 3D engineering applications [70], [71].

• **Comment (6):** The readability and presentation of this manuscript can be improved. Visualization of some figures and tables can be enhanced. Please carefully read and check the language, coefficients, functional, notations throughout the manuscript.

• **Response:** We have modified the description for better understanding and we will thoroughly check and fix grammatical errors in the new submission and the modified parts are highlighted. Besides, we have modified the design of figure 1,2 and 4 to make them more clear to read.

Sincerely,

Xiangchen Yin
University of Science and Technology of China