This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Scale-Aware Low-Light Enhancement via Structure-Guided Transformer Design

Wei Dong* Yan Min* Han Zhou† Jun Chen
McMaster University *Equal Contribution †Corresponding Author
{dongw22, miny13, zhouh115, chenjun}@mcmaster.ca

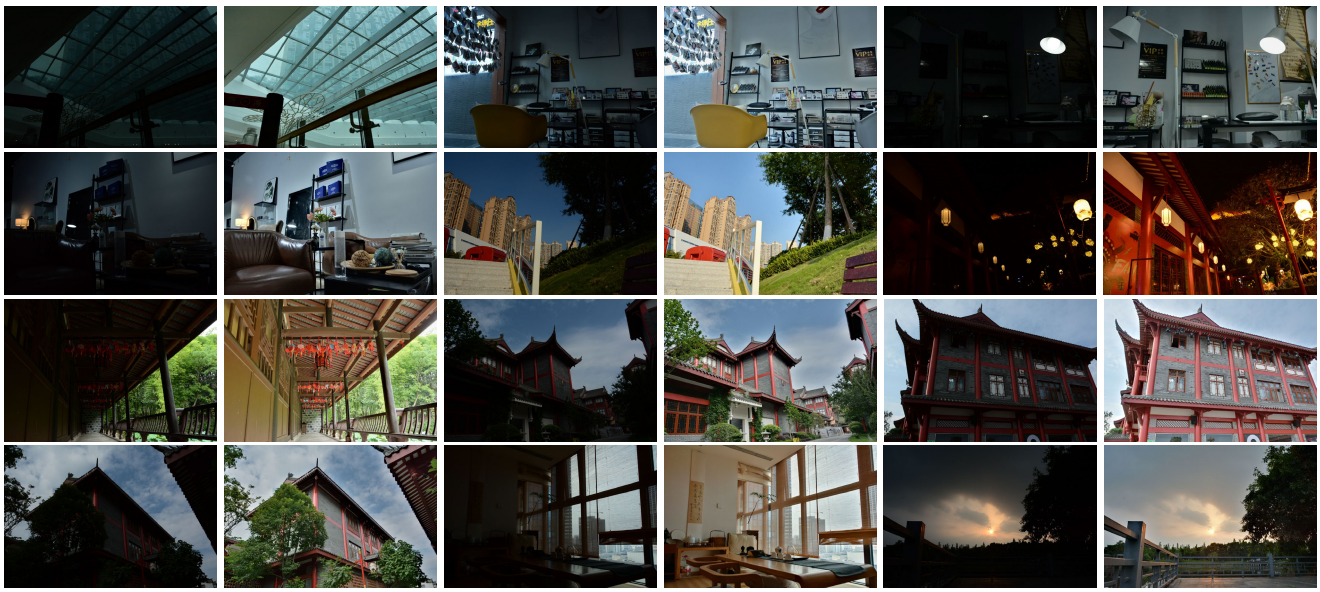


Figure 1. Our enhanced results on *NTIRE 2025 Low Light Enhancement Challenge*. Our method secures the **best PSNR**, achieves the *second-best* overall performance, and effectively enhance low light inputs without over-exposed artifacts.

Abstract

Current Low-light Image Enhancement (LLIE) techniques predominantly rely on either direct Low-Light (LL) to Normal-Light (NL) mappings or guidance from semantic features or illumination maps. Nonetheless, the intrinsic ill-posedness of LLIE and the difficulty in retrieving robust semantics from heavily corrupted images hinder their effectiveness in extremely low-light environments. To tackle this challenge, we present ***SG-LLIE****, a new multi-scale CNN-Transformer hybrid framework guided by structure priors. Different from employing pre-trained models for the extraction of semantics or illumination maps, we choose to extract robust structure priors based on illumination-invariant edge detectors. Moreover, we develop a CNN-Transformer Hybrid Structure-Guided Feature Extractor (HSGFE) module at each scale with in the UNet encoder-decoder architecture. Besides the CNN blocks which excels in multi-scale feature extraction and fusion, we introduce a Structure-Guided Transformer Block (SGTB) in each HSGFE that incorporates structural priors to modulate the enhancement process. Extensive experiments show that our method achieves state-of-the-art performance on several LLIE benchmarks in both quantitative metrics and visual quality. Our solution ranks second in the NTIRE 2025 Low-Light Enhancement Challenge. Code is released at* https://github.com/minyan8/imagine.

1. Introduction

Low Light Image Enhancement (LLIE) is an important task in computer vision. Normally captured images in dark environments often suffer from poor visibility, low contrast, and significant noise due to limited lighting conditions. These



Figure 2. The enhancement results of Retinexformer [6] (pretrained on LOL-v2-real dataset [63]) for real-world data.

issues not only lead to the degradation of image qualities, but also hinder the performance of high-level vision tasks such as object segmentation, recognition, and tracking [13, 27, 59]. Therefore, extensive research has been conducted to enhance low-light images and make them appear as if captured under better lighting conditions.

Most popular traditional methods for LLIE can be categorized into two types. One is histogram equalization-based methods [1, 31] , which adopt a straightforward strategy to perform gray-level remapping to enhance images with low visibility and contrast. However, these methods tends to introduce artifacts into the enhanced outputs. For example, in certain regions with uniform pixel values, the remapping process may produce excessively high or low pixel intensities, resulting in undesirably extremely bright or dark areas. Retinex-based approaches [19, 25, 37, 50, 56] represent another popular class of traditional methods for LLIE. According to Retinex theory [35], a color image can be decomposed into two components: reflectance and illumination. Thus, the estimated illumination is essentially utilized as a physical prior to guide the enhancement process of the reflectance component. However, this strategy assumes that the reflectance is noise-free, which is not realistic in practical scenarios. Moreover, inaccurate illumination priors can lead to visual distortions and color inconsistencies.

Recently, many deep learning models [6, 8, 10, 23, 26, 60, 61, 72] based on Convolutional Neural Network (CNN) and Transformers have been introduced for LLIE. However, end-to-end CNN methods often come with limited transparency and lacks solid theoretical backing, which can occasionally lead to unexpected or inconsistent results. While Transformer-based models excel at modeling non-local dependencies, directly applying Transformers [18] to LLIE does not necessarily yield satisfactory results. Therefore, dedicated architectural designs are essential. To this end, Retinexformer [6] proposes to introduce the illumination prior into Transformers to guide the reflectance enhancement process and ultimately produce impressive well-lit images, which demonstrates the significant potential of employing physical priors in deep learning models. However, we observe that the illumination prior utilized in Retinexformer is also learned from low light inputs via a lightweight neural network and there are no ground truth of illumination component to supervise such learning. Therefore, Retinexformer often struggles to generalize well in real-world scenarios and tends to produce unnatural color intensity and contrast, as presented in Fig. 2, which motivates us to explore and integrate other more robust physical priors into deep learning models for LLIE.

In this paper, we propose **SG-LLIE**, a scale-aware CNN-Transformer framework for **LLIE** with the **G**uidance of **S**tructure priors. Overall, our method follows a UNet encoder–decoder architecture, and we develop the Hybrid Structure-Guided Feature Extractor (HSGFE) module at each level. Within each HSGFE module, we first employ illumination-invariant edge detectors [21] to extract stable structure priors, and then we introduce the Structure-Guided Transformer Block (SGTB) to incorporate structural priors into the restoration process. Besides, the CNN-based Dilated Residual Dense Block (DRDB) and Semantic-Aligned Scale-Aware Module (SAM) proposed in [65] are employed for multi-scale feature fusion. To summarize, we highlight three key contributions of this work:

- We introduce **SG-LLIE**, a multi-scale CNN-Transformer hybrid framework guided by structure priors.
- We first extract robust structure priors from low light images and then integrate these priors into our customized Transformer via structure-guided cross-attention, providing more effective guidance for LLIE.
- Our model shows solid performance in both quantitative metrics and visual results, taking the lead in the *NTIRE 2025 Low-Light Image Enhancement Challenge* and staying competitive with other top solutions.

2. Related Work

Methods in LLIE Histogram Equalization (HE) [1, 31] and gamma correction [28, 51] are widely used in traditional techniques. These methods perform well under relatively uniform lighting conditions. However, their effectiveness often degrades in real-world scenarios, where illumination in LL images is typically dynamic and diverse. CNN-based models [3, 8, 10, 23, 26, 61, 70] has fundamentally altered the paradigm of image restoration and enhancement. LLNet [46] is the first to propose an autoencoder network, which enhances LL images without oversaturating the bright regions. Retinex theory are also widely applied in the deep learning-based methods. Yang *et* al. [64] introduces SGM-Net to combine priors and data-driven learning for LLIE, and effectively suppresses noise and improve contrast. However, due to the limitations of CNNs, these methods struggle to capture non-local information in images.

Transformers in Image Restoration Vision Transformers (ViTs) [18] are widely adopted in diverse low-level vi-

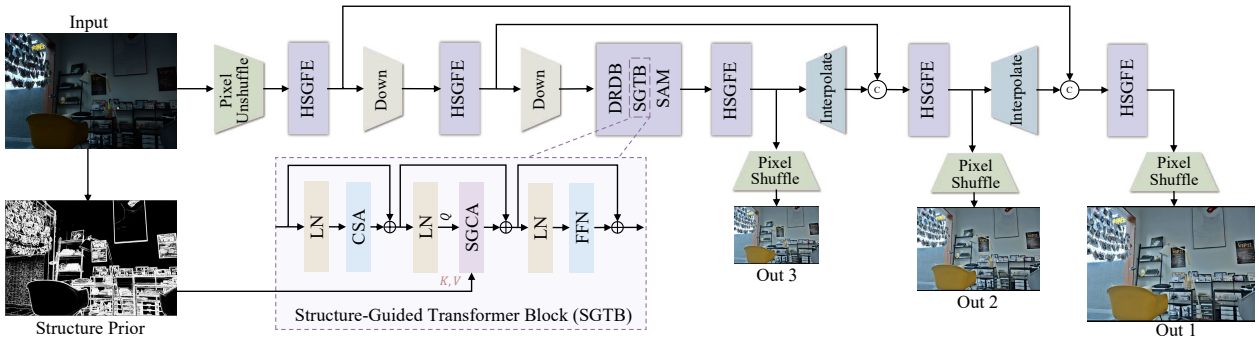


Figure 3. The overall framework of **SG-LLIE**. We develop our method based on ESDNet [65] and adopt a similar UNet architecture. At each level of the encoder and decoder, our customized Hybrid Structure-Guided Feature Extractor (HSGFE) module is employed. Within each HSGFE, besides the Dilated Residual Dense Block (DRDB) and Semantic-Aligned Scale-Aware Module (SAM) proposed in ESDNet [65], we first extract structure priors based on color-invariant edge detectors [36] and then develop the Structure-Guided Transformer Block (SGTB) to integrate these priors as guidance. With the integration of structure priors and our designed CNN-Transformer hybrid network, our method effectively improve the visibility and contrast with good noise suppression for diverse low light images.

sion tasks [4, 5, 7, 14–17, 29, 40, 54, 55, 62, 68]. For example, Chen *et* al. [7] introduces IPT, a pre-trained transformer model for low-level vision tasks like denoising, super-resolution, and deraining. Trained on corrupted image pairs from ImageNet, IPT can be fine-tuned for various tasks, not limited to image restoration. Conde *et* al. [9] proposes a network called Swin2SR, which builds upon SwinIR [42] by incorporating Swin Transformer v2 [45] to enhance performance on compressed image super-resolution. In this work, we aim to combine CNN and Transformer to develop a hybrid network, which is capable of effectively capturing local and long range dependencies.

Structure Prior In recent years, image structure priors have been increasingly utilized in various low-level vision tasks, including image inpainting [12, 48, 52], depth enhancement [24, 30, 41], and image restoration [11, 38, 39, 49]. In a different line of work, Lengyel *et al.* [36] proposes the integration of trainable, color-invariant edge detection layers into neural architectures to increase resilience to illumination variations. Alshammari *et* al. [2] introduces the use of illumination-invariant image transforms to enhance scene understanding and segmentation in challenging lighting conditions. By combining invariant representations with chromatic cues, this approach improves the robustness of deep networks without altering their architecture, highlighting the value of pre-processing for handling illumination variation. In this work, we aim to extract robust structure priors from low light inputs and then integrate these priors to modulate the LLIE process.

3. Method

Our contribution mainly lies in the design of multi-scale CNN-Transformer hybrid network and the integration of illumination invariant structure priors. The framework of our method is illustrated in Figure 3, where an encoder–decoder UNet architecture is designed. At each hierarchical level, the proposed Hybrid Structure-Guided Feature Extractor (HSGFE) functions as the core module, leveraging structural cues from the input image to facilitate the preservation of fine-grained details. In this section, We first discuss the structure prior extraction in Sec. 3.1. Then, we introduce the details of our developed HSGFE module including the Structure-Guided Transformer Block (SGTB) and Structure-Guided Cross Attention (SGCA) in Sec. 3.2. Finally, the multi-scale loss functions are specified in Sec. 3.3.

3.1. Structure Prior Extraction

To extract the structure prior, we adopt the Color Invariant Convolution (CIConv) proposed in [36, 73], which serves as a task-adaptive edge detector. CIConv applies a learnable scale-aware transformation to the color-invariant representation of the input, producing a normalized edge response map that reflects task-relevant structure. Among the derived color-invariant representations based on the Kubelka-Munk (KM) reflection model [20, 22, 53], we adopt $\mathbf{W}$ as it provides robust edge detection under varying illumination, shading, and reflectance conditions:

$$\mathbf{W}_{out} = \text{CIConv}(\mathbf{I}_{in}), \tag{1}$$

where $\mathbf{I}_{in}$ denotes the input LL image. $\mathbf{W}_{out}$ represents the structural prior, which is later integrated into the Structure-Guided Transformer Block (SGTB) for guidance. The formula of CIConv can be expressed as:

$$\text{CIConv}(\mathbf{I}_{in}) = \frac{\log\left(\mathbf{W}^2(\mathbf{I}_{in}) + \epsilon\right) - \mu_{\mathcal{S}}}{\sigma_{\mathcal{S}}} \tag{2}$$

where $\mu_{\mathcal{S}}$, $\sigma_{\mathcal{S}}$ and ϵ refer to the sample mean, standard deviation, and small perbutation. To compute $\mathbf{W}(\mathbf{I}_{in})$, we



Figure 4. Example visualization of the extracted structure priors.

first use the Gaussian Color Model [21] to obtain the initial edge detectors, denoted as $\mathbf{E}$. Then, we use $\mathbf{E}$ to derive the second phase of edge detectors, denoted as $\mathbf{W}$, as follows:

$$\mathbf{W} = \sqrt{\mathbf{W}_x^2 + \mathbf{W}_{\lambda x}^2 + \mathbf{W}_{\lambda\lambda x}^2 + \mathbf{W}_y^2 + \mathbf{W}_{\lambda y}^2 + \mathbf{W}_{\lambda\lambda y}^2}, \tag{3}$$

$$\mathbf{W}_x = \frac{\mathbf{E}_x}{\mathbf{E}}, \quad \mathbf{W}_{\lambda x} = \frac{\mathbf{E}_{\lambda x}}{\mathbf{E}}, \quad \mathbf{W}_{\lambda\lambda x} = \frac{\mathbf{E}_{\lambda\lambda x}}{\mathbf{E}} \tag{4}$$

Physical Explanation These equations mathematically indicates that $\mathbf{W}$ characterizes the spatial derivatives of spectral intensity. We visualize structure priors $\mathbf{W}$ for low-light and normal-light images in Fig. 4. It is clear that W represents the stable edge and structure map across images with varying illumination conditions, highlighting its great potential to guide the enhancement process.

3.2. Model Architecture

As illustrated in Fig. 3, our model adopts the encoder-decoder architecture. The "PixelUnShuffule" and several convolutional layers (denoted as "Down" in Fig. 3) are utilized for down-sampling, and the "PixelShuffle" or "Interpolation" process are used for up-sampling. Throughout the network, skip connections bridge corresponding encoder and decoder stages to maintain spatial coherence and support feature fusion. In each level in the encoding and decoding process, the Hybrid Structure-Guided Feature Extractor (HSGFE) module is proposed for representation learning and structure prior integration.

HSGFE Module HSGFE Module is designed to exploit structural cues inherent extracted from the input for enhanced detail retention. The HSGFE begins by processing features through a Dilated Residual Dense Block (DRDB) [65], which enhances local representations. This is followed by a Structure-Guided Transformer Block (SGTB), where structural priors extracted in Sec. 3.1 are explicitly injected into the feature flow, guided by invariant edge descriptors [21]. Moreover, the Semantic-Aligned Scale-Aware Module (SAM) is incorporated to further accommodate scale diversity and adaptively fuse representations from different scales. Within SAM, residual connections are retained to maintain high-resolution spatial fidelity.

Structure-Guided Transformer Block (SGTB) As shown in Fig. 3, our SGTB is composed of three main components: a Channel-wise Self-Attention (CSA), a Structure-Guided Cross Attention (SGCA), and a Feed-Forward Network (FFN). In addition, Layer Normalization is applied before each of these mechanisms, and three residual connections are applied to preserve residual information and support stable feature learning. Therefore, the feature flow for CSA can be represented as:

$$\mathbf{F}_{out} = \text{CSA}(\text{LN}(\mathbf{F}_{in})) + \mathbf{F}_{in}, \tag{5}$$

where $\mathbf{F}_{\text{in}} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{F}_{\text{out}} \in \mathbb{R}^{H\times W\times C}$ represent the input and output feature maps. We apply channel-wise attention here to let the model exchange information across channels, which facilitates capture long-range dependencies while ensuring model efficiency. Then, the processed representations are passed into the SGCA for further refinement.

Structure-Guided Cross Attention (SGCA) This mechanism firstly reshape the input features $\mathbf{F}_{\text{in}} \in \mathbb{R}^{H\times W\times C}$ into $\mathbf{X} \in \mathbb{R}^{HW\times C}$. Then, the resulting sequence $\mathbf{X}$ is linearly projected to generate the Query ($\mathbf{Q}$) representation:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q^\top. \tag{6}$$

We observe that conventional low-light image enhancement models often distort the original structural details of input images. To address this issue, we incorporate structural priors into our attention mechanism to better preserve spatial consistency. Specifically, structural priors are introduced into the Structure-Guided Cross Attention (SGCA) block to provide the **K**ey and **V**alue elements, denoted as $\mathbf{K}_p$ and $\mathbf{V}_p$, where the subscript p refers to structure **p**riors. These two representations are obtained by:

$$\mathbf{K_p} = \mathbf{X}\mathbf{W}_{K_p}^\top, \quad \mathbf{V_p} = \mathbf{X}\mathbf{W}_{V_p}^\top, \tag{7}$$

where $\mathbf{W}_{K_p} \in \mathbb{R}^{C\times C}$ and $\mathbf{W}_{V_p} \in \mathbb{R}^{C\times C}$ are learnable parameter matrice. We then formulate our structure-guided attention mechanism as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}_p, \mathbf{V}_p) = \text{softmax}(\frac{\mathbf{Q}\cdot\mathbf{K}_p}{\lambda})\cdot\mathbf{V}_p, \tag{8}$$

where λ is a learnable parameter that adaptively adjusts the scale of the matrix multiplication. To this end, we design our cross-attention mechanism to not only gather long-range dependencies but also blend structural cues directly into the current feature representations.

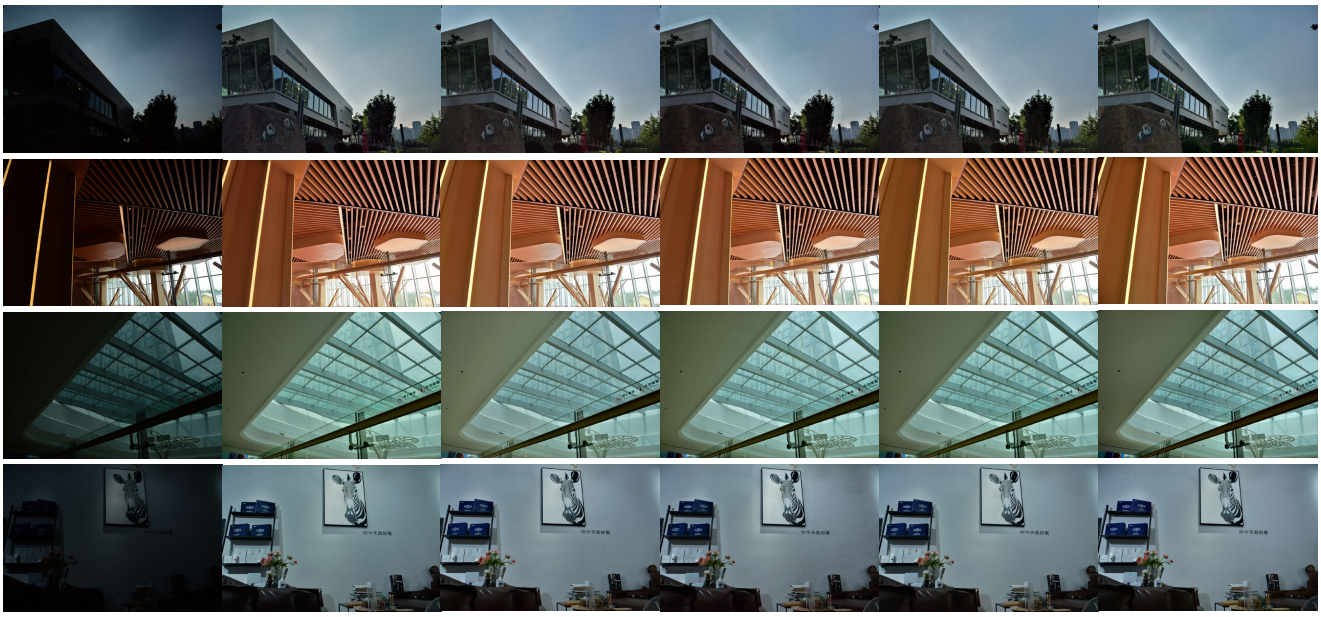


Figure 5. Qualitative comparisons on the *NTIRE 2025 LLIE Challenge* dataset. We compare our model with SNR [60], UHDM (SYSU) [43], and Retinexformer [6]. Our model consistently performs best or comparably well. Please zoom in for a better view.

Team	Rank	PSNR↑	SSIM↑	LPIPS↓	NIQE↓
NWPU-HVI	1	26.24	0.861	0.128	10.9539
Imagine(Ours)	2	26.345	0.858	0.133	11.8073
pengpeng-yu	3	25.849	0.858	0.134	11.2933
DAVIS-K	4	25.138	0.863	0.127	10.5814
SoloMan	5	25.801	0.856	0.130	11.4979
Smartdsp	6	25.47	0.848	0.120	10.5387
Smart210	7	26.148	0.855	0.137	11.5165
WHU-MVP	8	25.755	0.855	0.138	11.2140
BUPTMM	9	25.673	0.855	0.137	11.2831
NJUPTIPR	10	25.011	0.848	0.122	10.1485
SYSU-FVL-T2	11	25.652	0.857	0.135	11.5897

Table 1. Performance comparison of the top 11 teams in the *NTIRE 2025 LLIE Challenge*. Metrics include PSNR, SSIM, LPIPS (lower is better), and NIQE (lower is better). Our team (**Imagine**) achieved the best *PSNR* and the second-best overall score among 28 teams.

Scale-Adaptive Neural Architecture Inspired by the approach in Yu *et* al. [65], we integrate a Semantic-Aligned Scale-Aware Module (SAM) following the SGTB within our HSGFE module to effectively extract features across multiple scales. In real-world scenarios, images are often captured at varying resolutions (*e.g.*, 6000×4000 or 2992×2000), which poses challenges for consistent feature representation. To address this, SAM leverages a combination of pyramid-based feature extraction and cross-scale dynamic fusion.

Initially, the input feature map $\mathbf{F}_{in,0} \in \mathbb{R}^{H\times W\times C}$ undergoes bilinear interpolation to produce two additional versions at coarser scales: $\mathbf{F}_{in,1} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ and $\mathbf{F}_{in,2} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$. These multi-resolution feature maps are then processed independently through convolutional layers to yield the corresponding pyramid representations: $\mathbf{Y}_{in,0}$, $\mathbf{Y}_{in,1}$, and $\mathbf{Y}_{in,2}$.

Subsequently, cross-scale fusion is performed to integrate the multi-resolution feature maps. To achieve this, each scale-specific feature map is assigned a learnable weight matrix α_i, where $i = 0, 1, 2$. These weights are derived by applying global average pooling to each of the three feature maps independently. The resulting pooled features are then passed through a multi-layer perceptron (MLP), which facilitates adaptive and effective cross-scale interaction. This process can be formulated as follows:

$$[\beta_0, \beta_1, \beta_2 = \text{MLP}(\alpha_0, \alpha_1, \alpha_2)]. \tag{9}$$

Finally, the fused feature map $\mathbf{F}_f$ is obtained as follows,

Methods	LOL-v1			LOL-v2-real			NTIRE 2025 LLIE			Complexity	
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Params (M)	FLOPs
KinD [69]	20.87	0.802	0.207	17.54	0.669	0.365	21.29	0.829	0.186	8.02	34.09 G
MIRNet [66]	24.14	0.832	0.131	19.35	0.708	0.138	24.21	0.857	0.141	31.76	785 G
SNR [60]	24.61	0.842	0.151	20.82	0.811	0.161	24.39	0.878	0.149	4.01	26.35 G
Restormer [67]	22.43	0.823	0.147	18.69	0.834	0.232	22.47	0.846	0.145	26.13	144.25 G
Retinexformer [6]	25.16	0.845	0.131	22.79	0.840	0.171	25.06	0.872	0.183	1.61	17.02 G
LLFlow [57]	21.09	0.861	0.116	17.43	0.831	0.129	23.49	0.880	0.131	17.42	1.05 T
ESDNet [65]	23.10	0.846	0.094	20.48	0.841	0.134	26.26	0.892	0.111	10.62	28.82 G
GLARE [71]	23.55	0.863	0.086	22.75	0.856	0.105	25.35	0.887	0.122	24.74	223 G
Ours	24.63	0.873	0.092	22.84	0.859	0.126	26.75	0.899	0.113	12.67	35.80 G

Table 2. Quantitative comparison on LOL-v1, LOL-v2-real, and *NTIRE 2025* datasets. [Key: Best Performance; Second-Best Performance]

and $\mathbf{F}_{in,0}$ is added to retain more input information.

$$\mathbf{F}_f = \mathbf{F}_{in,0} + \beta_0 \odot \mathbf{Y}_{in,0} + \beta_1 \odot \mathbf{Y}_{in,1} + \beta_2 \odot \mathbf{Y}_{in,2}. \quad (10)$$

3.3. Loss Function and Adaptive Adjustment Layer

Loss Function As Fig. 3 shows, for network training, we design the loss function based on three levels of output images—$\hat{\mathbf{I}}_1$, $\hat{\mathbf{I}}_2$, and $\hat{\mathbf{I}}_3$—which correspond to the output images at three different resolutions from the respective levels of our decoder:

$$L_{\text{total}} = \sum_{i=1}^{3} L_C\left(\mathbf{I}_i, \hat{\mathbf{I}}_i\right) + \lambda \cdot \sum_{i=1}^{3} L_P\left(\mathbf{I}_i, \hat{\mathbf{I}}_i\right) + \gamma \cdot \sum_{i=1}^{3} L_{\text{MS-SSIM}}\left(\mathbf{I}_i, \hat{\mathbf{I}}_i\right), \quad (11)$$

where L_1, L_{P}, and $L_{\text{MS-SSIM}}$ represents the Charbonnier loss [34], perceptual loss [32], and Multi-Scale SSIM (MS-SSIM) loss, respectively. The weighting factors are set as $\lambda = 0.01$ and $\gamma = 0.4$.

Adaptive Adjustment Layer When evaluating our method on the dataset provided by NTIRE 2025 Low-Light Enhancement Challenge, we observe the illumination of enhanced results differs a lot even for the training data. We attribute this to the diverse light condition within the training set and we aim to develop an additional layer for slight illumination adjustment. Specifically, by analyzing the illumination level of the training set (including the input and enhanced results of our proposed SG-LLIE), we classify all training samples into five classes, and employ a specific adjustment layer between 0.96 and 1.05 to each class. Moreover, three linear layers connected by ReLU is employed to learn this classification pattern and the cross-entropy is utilized for the optimization of this adaptive adjustment layer. Finally, we add this adjustment layer into SG-LLIE and optimize SG-LLIE using a low learning rate ($2 \times 10^{-}7$) without updating the adaptive adjustment layer. Besides, only $\hat{\mathbf{I}}_1$ is leveraged as the loss function in this stage.

4. Experiments

4.1. Performance on NTIRE 2025 LLIE Challenge

Dataset We use the dataset provided by the *NTIRE 2025 Low Light Enhancement Challenge* [44]. This dataset consists of 219 training images, 46 validation images, and 30 test images. Most images have a resolution of 2992×2000, with several reaching up to 6000×4000. Compared to other low-light image datasets, such as LOL [58] and LOL-v2-real [63], the *NTIRE* dataset offers higher-resolution images and more diverse content, better reflecting real-world captures from modern smartphones. Both indoor and outdoor scenes are included. Besides, we also use the training set from the *NTIRE 2024 Low-Light Enhancement Challenge* [43] for fine-tuning.

Implementation Details We develop and train our model using PyTorch on a single A100 GPU, without any pre-trained weights. The model is optimized with the Adam optimizer [33] ($lr = 2 \times 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$), and gradient clipping is applied. Training runs for 15,000 iterations with a fixed patch size of 1600×1600 and a batch size of 1. The initial learning rate is set as 2×10^{-4} and the cyclic cosine annealing schedule [47] is employed. Data augmentation includes geometric transforms and mixup with $\beta = 1.2$ and identity mapping enabled. The training of the adjustment layer and the fine-tuning of SG-LLIE is provided in Sec. 3.3.

Quantitative and Qualitative on NTIRE 2025 LLIE Challenge As presented in Tab. 1, our method achieves the best PNSR, the third best SSIM in *NTIRE 2025 LLIE Challenge*. Overall, our method ranks 2nd out of 28 teams from around the world, highlighting the outstanding perfor-



Figure 6. Visual comparisons on the LOL-v1 dataset, with MIRNet [66], RetinexFormer [6], SNR [60], and ESDNet [65].

mance of our proposed approach. The enhancement outputs of our method are presented in Fig. 1, which demonstrates that our method can handle diverse low light images captured with varying illuminations and indoor or outdoor scenarios. Both quantitative and qualitative results demonstrate the effectiveness of our proposed method.

4.2. Performance on LLIE Benchmark Datasets

Experiment Settings To comprehensively evaluate the performance of our method, we compare our method with KinD [69], MIRNet [66], Restormer [68], LLFlow [57], Retinexformer [6], SNR [60], ESDNet (refined by Team SYSU-FVL-T2 [43] in *NTIRE 2024 LLIE Challenge*), and GLARE [71] on the *NTIRE 2025 LLIE* dataset. Due to the unavailability of the ground truth for validation and test set, we build a test set includes 11 images with scenes that are similar but not identical to those in the training set. We construct this test set by analyzing the differences between the *NTIRE 2024 LLIE* and *NTIRE 2025 LLIE* training sets. For a fair comparison, we use a patch size of 512×512 and a batch size of 2 during training on the NTIRE 2025 LLIE dataset for all methods being compared. Besides the NTIRE dataset, we also evaluate our method on two additional datasets: LOL-v1 [58] and LOL-v2-real [64], where the patch size and batch size are set to 384×384 and 4, respectively. Notably, the adjustment layer is not included in this setting for a fair comparison.

Quantitative Results The quantitative comparisons on the *NTIRE 2025 LLIE*, LOL-v1, and LOL-v2-real datasets are shown in Tab 2. On LOL-v1, we obtain the highest SSIM (0.873) and achieve the second-bset performance on PSNR and LPIPS, indicating superior structural fidelity and perceptual quality, while maintaining a competitive PSNR (24.63). For LOL-v2-real, our approach continues to lead with a PSNR of 22.84, SSIM of 0.859, and LPIPS of 0.126, surpassing all other methods. On the *NTIRE 2025 LLIE* benchmark, our method again outperforms all baselines with the highest PSNR (26.75), SSIM (0.899), and second-best LPIPS (0.113), closely following ESDNet (0.111). Although Retinexformer [6] achieves higher PSNR scores with fewer parameters on the LOL-v1 and LOL-v2-real datasets, it falls short in recovering structural details compared to our method. In addition, its performance on high-resolution images degrades noticeably, and it incurs higher GPU memory consumption owing to its fully Transformer-based design. In contrast, our method outperforms both GLARE [71] and LLFlow [57] in terms of computational efficiency and enhancement performance. Furthermore, our SG-LLIE introduces only approximately 2M additional parameters compared to ESDNet [65], yet achieves an average gain of 1.46 in PSNR and 0.017 in SSIM.

Qualitative Results We present visual comparisons of our model with the methods listed in Table 2. As shown in Figure 6, Figure 7, and Figure 5, our results are perceptually better or comparable to those of other methods. Our method more effectively captures underlying lighting conditions and preserves the original content of the input images. While MIRNet [66] and SNR [60] occasionally produce slightly brighter outputs, our model delivers a more perceptually pleasing balance of contrast and natural appearance, particularly in terms of retaining underlying image information.

4.3. Ablation Study

To verify the contribution of each component in our method, we conduct ablation studies on the LOL-v1 dataset [8].



Figure 7. Visual comparisons on LOL-v2-real dataset with MIRNet [66], RetinexFormer [6], SNR [60], and ESDNet [65].

Configuration	PSNR↑	SSIM↑	LPIPS↓
w/o SGCA	23.67	0.855	0.095
w/o SGTB	23.38	0.849	0.096
w/o MS-SSIM	24.48	0.868	0.093
Full Model	**24.63**	**0.873**	**0.092**

Table 3. Ablation results on LOL-v1 dataset. Our full model achieves the best performance in terms of all metrics and removing any component our complete model leads to obvious performance drop, highlighting the rationality of the design of our method.

Structure-Guided Cross Attention (SGCA) To study the importance of the extracted structure prior and our proposed SGCA, we remove these two parts from our method. Tab. 3 reports the quantitative performance of this modification on LOL-v1 dataset, which still achieve competitive enhancement performance on all measured metrics compared to ESDNet, Restormer, and MIRNet in Tab. 2. However, compared to this modified version, our full model shows significantly enhanced SSIM and PSNR scores by integrating structure prior using our proposed SGCA. This comparisons manifest the importance of our proposed SGCA and the structure prior generated by the extraction pipeline discussed in Sec. 3.1.

Structure-Guided Transformer Block (SGTB) Similarly, we implement an adaptation to illustrate the effectiveness of the developed SGTB. Specifically, we remove the SGTB from our customized Hybrid Structure-Guided Feature Extractor (HSGFE) module. The quantitative results of the remaining network are reported in Tab. 3. The discernible performance gap between this configuration and the full model (*i.e.*, a 1.25 dB drop in PSNR and a 0.024 drop in SSIM) demonstrates the superiority of our proposed Structure-Guided Transformer Block (SGTB). It is worth noting that the remaining network is entirely CNN-based, and its relatively poor performance underscores the importance of our hybrid CNN-Transformer architecture.

Loss Function As introduced in Sec. 3.3, we add the muti-scale MS-SSIM loss into our complete optimization objective, which is not included in ESDNet. To verify the effectiveness of this new introduced loss function, we remove it and adopt the same optimization strategy in ESDNet. The quantitative results are reported in Tab. 3, which demonstrate that the integration of multi-scale MS-SSIM loss helps achieve higher SSIM performance.

5. Conclusion

We present SG-LLIE, a CNN-Transformer hybrid encoder-decoder architecture augmented with Hybrid Structure-Guided Feature Extractor (HSGFE) modules to effectively enhance low-light images. These modules leverage structural cues, extracted from the input by color-invariant edge detector, to preserve fine-grained details, combining dilated residual dense blocks with transformer layers guided by structural priors. To further address scale variation, the scale-aware semantic-aligned module is incorporated, and skip connections are employed to preserve spatial consistency throughout the network. Our design enables robust and detail-aware low-light enhancement across diverse lighting conditions. Extensive experiments validate the effectiveness of our method, achieving the **best PSNR score** and **second-best overall performance** among 28 teams in *NTIRE 2025 Low Light Enhancement Challenge* [44].

References

[1] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, and Oksam Chae. A dynamic histogram equalization for image contrast enhancement. *IEEE transactions on consumer electronics*, 53(2):593–600, 2007. 2

[2] Naif Alshammari, Samet Akcay, and Toby P Breckon. On the impact of illumination-invariant image pre-transformation for contemporary automotive semantic scene understanding. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1027–1032. IEEE, 2018. 3

[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Han Zhou, Wei Dong, Yangyi Liu, Jun Chen, Huan Liu, Liangyan Li, et al. Ntire 2023 hr non-homogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1808–1825, 2023. 2

[4] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Yidi Liu, Xingbo Wang, Yurui Zhu, Gege Shi, Xin Lu, Xueyang Fu, et al. Ntire 2024 dense and non-homogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6453–6468, 2024. 3

[5] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc V Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. *Advances in Neural Information Processing Systems*, 35:37749–37761, 2022. 3

[6] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12504–12513, 2023. 2, 5, 6, 7, 8

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 3

[8] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6306–6314, 2018. 2, 7

[9] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *European Conference on Computer Vision*, pages 669–687. Springer, 2022. 3

[10] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 870–878, 2018. 2

[11] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019. 3

[12] Brian Dolhansky and Cristian Canton Ferrer. Eye in-painting with exemplar generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7902–7911, 2018. 3

[13] Wei Dong, Han Zhou, and Dong Xu. A new sclera segmentation and vessels extraction method for sclera recognition. In *2018 10th International Conference on Communication Software and Networks (ICCSN)*, 2018. 2

[14] Wei Dong, Han Zhou, Yuqiong Tian, Jingke Sun, Xiaohong Liu, Guangtao Zhai, and Jun Chen. Shadowrefiner: Towards mask-free shadow removal via fast fourier transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6208–6217, 2024. 3

[15] Wei Dong, Han Zhou, Ruiyi Wang, Xiaohong Liu, Guangtao Zhai, and Jun Chen. Dehazedct: Towards effective non-homogeneous dehazing via deformable convolutional transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6405–6414, 2024.

[16] Wei Dong, Han Zhou, Yulun Zhang, Xiaohong Liu, and Jun Chen. Ecmamba: Consolidating selective state space model with retinex guidance for efficient multiple exposure correction. *Advances in Neural Information Processing Systems*, 37:53438–53457, 2024.

[17] Wei Dong, Han Zhou, Seyed Amirreza Mousavi, and Jun Chen. Retinex-guided histogram transformer for mask-free shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 3

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[19] Xueyang Fu, Yinghao Liao, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation. *IEEE Transactions on Image Processing*, 24(12):4965–4977, 2015. 2

[20] Jan-Mark Geusebroek. Color and geometrical structure in images. *Appl. Microsc*, 2000. 3

[21] J-M Geusebroek, Rein Van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. Color invariance. *IEEE Transactions on Pattern analysis and machine intelligence*, 23 (12):1338–1350, 2001. 2, 4

[22] Theo Gevers, Arjan Gijsenij, Joost Van de Weijer, and Jan-Mark Geusebroek. *Color in computer vision: Fundamentals and applications*. John Wiley & Sons, 2012. 3

[23] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 2

[24] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang. Learning dynamic guidance for depth image enhancement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3769–3778, 2017. 3

[25] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016. 2

[26] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2): 1–17, 2018. 2

[27] Guan Huang, Xingang Wang, Wenqi Wu, Han Zhou, and Yuanyuan Wu. Real-time lane-vehicle detection and tracking system. In *Chinese Control and Decision Conference (CCDC)*, 2016. 2

[28] Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE transactions on image processing*, 22(3):1032–1041, 2012. 2

[29] Drew A Hudson and Larry Zitnick. Generative adversarial transformers. In *International conference on machine learning*, pages 4487–4499. PMLR, 2021. 3

[30] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 353–369. Springer, 2016. 3

[31] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007. 2

[32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 6

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[34] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018. 6

[35] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 2

[36] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021. 3

[37] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE transactions on image processing*, 27(6):2828–2841, 2018. 2

[38] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 272–289, 2018. 3

[39] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020. 3

[40] Xin Li, Yeying Jin, Xin Jin, Zongwei Wu, Bingchen Li, Yufei Wang, Wenhan Yang, Yu Li, Zhibo Chen, Bihan Wen, Robby Tan, Radu Timofte, et al. NTIRE 2025 challenge on day and night raindrop removal for dual-focused images: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 3

[41] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 154–169. Springer, 2016. 3

[42] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 3

[43] Xiaoning Liu, Zongwei Wu, Ao Li, Florin-Alexandru Vasluianu, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, et al. Ntire 2024 challenge on low light image enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6571–6594, 2024. 5, 6, 7

[44] Xiaoning Liu, Zongwei Wu, Florin-Alexandru Vasluianu, Hailong Yan, Ren Bin, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, et al. Ntire 2025 challenge on low light image enhancement: Methods and results. In *CVPRW*, 2025. 6, 8

[45] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 3

[46] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017. 2

[47] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[48] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3

[49] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring face images with exemplars. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 47–62. Springer, 2014. 3

[50] Seonhee Park, Soohwan Yu, Byeongho Moon, Seungyong Ko, and Joonki Paik. Low-light image enhancement using variational optimization-based retinex model. *IEEE Transactions on Consumer Electronics*, 63(2):178–184, 2017. 2

[51] Shanto Rahman, Md Mostafijur Rahman, Mohammad Abdullah-Al-Wadud, Golam Dastegir Al-Quaderi, and Mo-

hammad Shoyaib. An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing*, 2016:1–13, 2016. 2

[52] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 181–190, 2019. 3

[53] Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 1985. 3

[54] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, Wei Dong, Han Zhou, Yuqiong Tian, Jun Chen, et al. Ntire 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6547–6570, 2024. 3

[55] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, et al. NTIRE 2025 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 3

[56] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing*, 22(9):3538–3548, 2013. 2

[57] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex C Kot. Low-light image enhancement with normalizing flow. *arXiv preprint arXiv:2109.05923*, 2021. 6, 7

[58] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 6, 7

[59] Dong Xu, Wei Dong, and Han Zhou. Sclera recognition based on efficient sclera segmentation and significant vessel matching. In *The Computer Journal*, 2022. 2

[60] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17714–17724, 2022. 2, 5, 6, 7, 8

[61] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, 35(2): 1–15, 2016. 2

[62] Kangning Yang, Jie Cai, Ling Ouyang, Florin-Alexandru Vasluianu, Radu Timofte, Jiaming Ding, Huiming Sun, Lan Fu, Jinlong Li, Chiu Man Ho, Zibo Meng, et al. NTIRE 2025 challenge on single image reflection removal in the wild: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 3

[63] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. 2, 6

[64] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. 2, 7

[65] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *European Conference on Computer Vision*, pages 646–662. Springer, 2022. 2, 3, 4, 5, 6, 7, 8

[66] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020. 6, 7, 8

[67] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 6

[68] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 3, 7

[69] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1632–1640, 2019. 6, 7

[70] Han Zhou, Wei Dong, Yangyi Liu, and Jun Chen. Breaking through the haze: An advanced non-homogeneous dehazing method based on fast fourier convolution and convnext. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1895–1904, 2023. 2

[71] Han Zhou, Wei Dong, Xiaohong Liu, Shuaicheng Liu, Xiongkuo Min, Guangtao Zhai, and Jun Chen. Glare: Low light image enhancement via generative latent feature based codebook retrieval. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024. 6, 7

[72] Han Zhou, Wei Dong, Xiaohong Liu, Yulun Zhang, Guangtao Zhai, and Jun Chen. Low-light image enhancement via generative perceptual priors. *arXiv preprint arXiv:2412.20916*, 2024. 2

[73] Han Zhou, Wei Dong, and Jun Chen. Lita-gs: Illumination-agnostic novel view synthesis via reference-free 3d gaussian splatting and physical priors. *arXiv preprint arXiv:2504.00219*, 2025. 3