# TOWARDS SCALE-AWARE LOW-LIGHT ENHANCEMENT VIA STRUCTURE-GUIDED TRANSFORMER DESIGN

# TOWARDS SCALE-AWARE LOW-LIGHT ENHANCEMENT VIA STRUCTURE-GUIDED TRANSFORMER DESIGN

By YAN MIN, M.A.Sc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Applied Science

McMaster University

MASTER OF APPLIED SCIENCE (2025)

Hamilton, Ontario, Canada (Electrical and Computer Engineering)

| | |
|---|---|
| TITLE: | Towards Scale-Aware Low-Light Enhancement via Structure-Guided Transformer Design |
| AUTHOR: | Yan Min<br>M.A.Sc (Electrical and Computer Engineering),<br>McMaster University, Hamilton, Canada |
| SUPERVISOR: | Dr. Jun Chen |
| NUMBER OF PAGES: | xiii, 44 |

# Abstract

We present a novel Transformer-based framework for low-light image enhancement, Towards Scale-Aware Low-Light Enhancement via Structure-Guided Transformer Design. Our model is built upon a U-Net-style encoder–decoder architecture, where we introduce a customized Hybrid Structure-Guided Feature Extractor (HSGFE) at each scale. The HSGFE integrates three key components: (1) a Dilated Residual Dense Block (DRDB) for effective feature refinement, (2) a Structure-Guided Transformer Block (SGTB) that incorporates structural priors to preserve edges and suppress noise, and (3) a Semantic-Aligned Scale-Aware Module (SAM) to handle multi-scale variations. This design enables our network to enhance low-light images while maintaining structural integrity and reducing color distortion. Extensive experiments show that our method achieves state-of-the-art performance in both quantitative metrics and visual quality. Our approach also achieve top-tier results on standard LLIE benchmarks and ranked second in the NTIRE 2025 Low-Light Image Enhancement Challenge.

*To my dear parents, with love and gratitude.*

# Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Jun Chen, for providing me with valuable research opportunities and the chance to study in Canada. His guidance and support have been instrumental throughout my academic journey.

I'm also especially grateful to Han Zhou and Wei Dong for their thoughtful guidance and generous support. Their willingness to share their knowledge and experience not only introduced me to the world of deep learning and computer vision, but also gave me the confidence to begin my own journey in research. Working with them has been both a privilege and a joy, and I've learned so much from their mentorship and collaboration.

# Table of Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **LLIE** | Low Light Image Enhancement |
| **SGCA** | Structure-Guided Cross Attention |
| **HSGFE** | Hybrid Structure-Guided Feature Extractor |
| **NTIRE** | New Trend in Image Restoration Competition |
| **DRDB** | Dilated Residual Dense Block |
| **SAM** | Semantic-Aligned Scale-Aware Module |
| **PSNR** | Peak Signal-to-Noise Ratio |
| **HE** | Histogram Equalization |
| **CNN** | Convolutional Neural Network |
| **ViT** | Vision Transformer |
| **IPT** | Image Processing Transformer |
| **CIConv** | Color Invariant Convolution |
| **KM** | Kubelka-Munk reflection model |

| | |
|---|---|
| **CSA** | Channel-wise Self-Attention |
| **FFN** | Feed-Forward Network |
| **MS-SSIM** | Multi-Scale SSIM Loss |
| **LPIPS** | Learned Perceptual Image Patch Similarity |
| **NIQE** | Natural Image Quality Evaluator |
| **LN** | Layer Normalization |
| **Adam** | Adam optimizer |
| **LOL-v1** | Low-Light Dataset v1 |
| **LOL-v2** | Low-Light Dataset v2 |
| **DSLR** | Digital Single-Lens Reflex |

# Chapter 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 issues not only lead to the degradation of image qualities, but also hinder the performance of high-level vision tasks such as object detection, recognition, and tracking. Therefore, extensive research has been conducted on this task, aiming to enhance low-light images and make them appear as if captured under better lighting conditions.

Although many methods have been proposed in recent years, they still suffer from several limitations. Most popular traditional methods for LLIE can be categorized into two types. One is histogram equalization-based methods [1, 27] , 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 [32, 15, 55, 22,

47, 34] represent another popular class of traditional methods for LLIE. According to Retinex theory, 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 [23, 19, 8, 11, 61, 60, 6] 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 [14] 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. 1.1, which motivates us to explore and integrate other more robust physical priors into deep learning models for LLIE.

Figure 1.1: The enhancement results of Retinexformer (pre-trained on LOLv2-real dataset) for real-world data.

In this paper, we propose a model titled *Towards Scale-Aware Low-Light Enhancement via Structure-Guided Transformer Design*. Our network follows a U-Net-like encoder–decoder architecture. At each level, we employ our customized Hybrid Structure-Guided Feature Extractor (HSGFE), which leverages structural priors from the input images to better preserve original details. Within the HSGFE, we first use a Dilated Residual Dense Block (DRDB) [65] to refine the input features. Then, the Structure-Guided Transformer Block (SGTB) is developed to incorporate structural priors into the feature maps, enabling the use of invariant edge detectors [17]. After the SGTB, a Semantic-Aligned Scale-Aware Module (SAM) [65] is integrated to address scale variations. Additionally, skip connections are employed to preserve the spatial information of the images.

To summarize, we highlight three key contributions of this work:

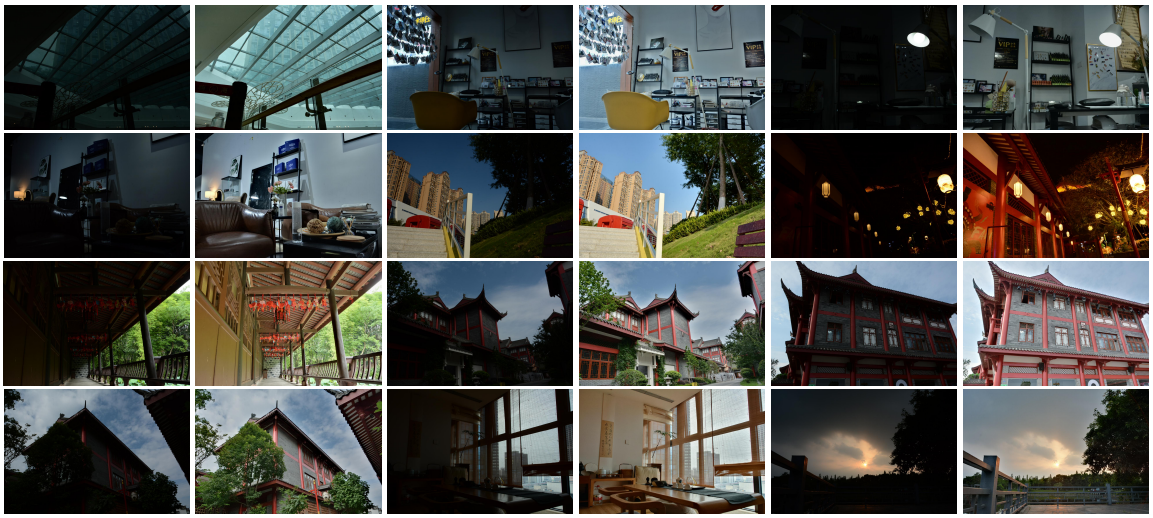


Figure 1.2: Our enhanced results on *NTIRE 2025 Low Light Enhancement Challenge*. Our method secures the highest PSNR value, achieves the second-best overall performance, and effectively enhance low light inputs without over-exposed artifacts.

- **First,** we build an encoder–decoder-style network, similar to U-Net, using customized Transformers.
- **Second,** 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.
- **Third,** 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.

# Chapter 2

# Related Work

## 2.1 Methods in LLIE

Histogram Equalization (HE) [1, 27] and gamma correction [24, 48] 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 low-light images is typically dynamic and diverse. Convolution Neural Networks (CNNs) [23, 19, 8, 11, 61] really change how people would deal with low-light images. LLNet [43] is the first to propose an autoencoder network, which enhances low-light images without oversaturating the bright regions. Retinex theory are also widely applied in the deep learning-based methods. Yang *et* al. [63] introduces SGM-Net to combine priors and data-driven learning for LLIE, and effectively suppresses noise and improve contrast. Guo *et* al. [21] propose Zero-DCE, a lightweight deep learning method that enhances low-light images by estimating pixel-wise adjustment curves without relying on paired or unpaired training data. Their approach uses non-reference loss functions to guide learning and shows strong generalization across

diverse lighting conditions. A brightness-aware network is proposed by Liu *et* al. [42], which leverages normal-light priors and attention mechanisms to enhance low-light images more naturally. By integrating a residual-quantized codebook and a fusion module, their method effectively combines low-light and normal-light features, outperforming prior approaches in both synthetic and real-world scenarios. Zhang *et* al. [70] propose a highly efficient single convolutional layer model (SCLM) for low-light image enhancement, utilizing structural re-parameterization for global enhancement. To address uneven illumination, they incorporate a local adaptation module, achieving competitive results with minimal parameters and low computational complexity. Makwana [44] *et* al. introduce LIVENet, a two-stage deep learning framework that jointly performs noise reduction and low-light enhancement. By leveraging a denoising block and an atmospheric scattering model, LIVENet enhances illumination and texture while preserving natural appearance, achieving superior results across multiple benchmarks. Wang *et* al. [54] propose BCNet, a "brighten-and-colorize" network that treats LLIE as a multi-task problem by enhancing both lightness and chrominance. Unlike prior methods, it enables user-customized color enhancements without affecting image structure, achieving state-of-the-art results and offering flexible visual outputs. Shakibania *et* al. [52] present CDAN, an attention-guided autoencoder network that enhances low-light images by combining convolutional and dense blocks with skip connections. Through a composite loss and post-processing, CDAN effectively restores brightness, texture, and color, achieving strong performance across challenging low-light scenarios.

Due to the limitations of CNNs, these deep learning-based methods struggle to capture non-local information in images.

## 2.2 Transformers in Image Restoration

Vision Transformers (ViTs) [14] are widely adopted in modern computer vision tasks, and the self-attention mechanism is utilized to capture global dependency within images. Many researches have integrated or customized vision transformers in their networks for low-level tasks [5, 7, 68, 25]. For example, Chen *et* al. [7] introduces a vision transformer named image processing transformer (IPT), a pre-trained transformer model for low-level vision tasks like denoising, super-resolution, and deraining. Trained on corrupted image pairs from ImageNet with a multi-head, multi-tail structure and contrastive learning, IPT can be fine-tuned for various tasks, not limited to image restoration. Conde *et* al. [10] proposes a network called Swin2SR, which builds upon SwinIR [38] by incorporating Swin Transformer v2 to enhance performance on compressed image super-resolution. This approach addresses key challenges in training vision transformers—such as instability and resolution gaps, and achieves impressive results on JPEG artifact removal and lightweight super-resolution tasks. Zhang *et* al. [71] propose an efficient Transformer-based approach for image restoration that captures superpixel-wise global dependencies through a coarse-to-fine design. Their method uses condensed attention and dual adaptive blocks to transfer global context to the pixel level, achieving competitive results with significantly lower computational cost compared to SwinIR. Ren *et* al. [49] introduce the Key-Graph Transformer (KGT), which reduces the computational burden of global attention by constructing a sparse key-graph that connects only essential nodes. This selective attention mechanism enables efficient image restoration with strong performance across multiple tasks. Yang *et* al. [64] propose the Region Attention Transformer (RAT), which performs self-attention within semantically segmented regions rather than fixed

patches, reducing interference from unrelated areas. By leveraging dynamic region partitioning and a focal region loss, RAT achieves strong results across various medical image restoration tasks. Jiang *et* al. [28] present SFHformer, a dual-domain Transformer that integrates Fast Fourier Transform for global frequency modeling alongside spatial domain processing for local features. This hybrid design enables robust performance across diverse image restoration tasks, including low-light enhancement, while maintaining efficiency in terms of parameters and computation. Wang *et* al. [57] propose Uformer, a Transformer-based encoder-decoder architecture for image restoration that introduces a locally-enhanced window attention mechanism to balance efficiency and context modeling. A learnable multi-scale spatial bias further enhances detail restoration, enabling strong performance across multiple restoration tasks with minimal computational overhead. Chen *et* al. [9] propose the Cross Aggregation Transformer (CAT), which enhances long-range dependency modeling through a novel Rectangle-Window Self-Attention mechanism and axial-shift operations. By combining global attention with local inductive biases via a Locality Complementary Module, CAT achieves state-of-the-art performance across various image restoration tasks. Xiao *et* al. [59] propose Stoformer, a Transformer-based model that introduces a stochastic window shifting strategy to enhance translation invariance and better preserve local relationships in image restoration. By combining random window partitioning with a layer expectation propagation algorithm, their method achieves notable improvements across tasks like deraining, denoising, and deblurring.

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 dependency.

## 2.3 Structure Prior

In recent years, image structure priors have been increasingly utilized in various low-level vision tasks, including image inpainting [13, 45, 50], depth enhancement [20, 26, 37], and image restoration [12, 36, 35, 46]. In a different line of work, Lengyel *et al.* [33] proposes the integration of trainable, color-invariant edge detection layers into neural architectures to increase resilience to illumination variations. Their approach demonstrated that leveraging such structure-based priors can mitigate the distribution gap between day and night scenes, leading to improved generalization across multiple tasks (*e.g.*, classification, segmentation, and place recognition) without relying on any target domain data. 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, the approach improves the robustness of deep networks without altering their architecture, highlighting the value of pre-processing for handling illumination variation. Ulyanov *et* al. [53] demonstrate that the architecture of a convolutional generator network itself—without any learning—can serve as a powerful image prior. Their method leverages randomly initialized networks to perform tasks like denoising and super-resolution, revealing the strong inductive bias inherent in CNN structures and bridging the gap between learned and handcrafted priors. Arjomand *et* al. [3] introduce a restoration method that builds on a smoothed approximation of natural image statistics, enabling image enhancement without explicit noise level input. Their approach leverages denoising autoencoders to estimate gradients of this prior, guiding a gradient-based optimization process that restores images

effectively across various tasks such as super-resolution, deblurring, and demosaicing. To improve generalization in image restoration, Liu *et* al. [39] propose TAPE, a transformer-based framework that first learns a task-independent representation of natural image statistics through pretraining. This prior is later adapted to specific restoration tasks via fine-tuning, allowing the model to handle diverse degradations effectively and even outperform task-specific methods in certain cases.

In this work, we first extract robust structure priors from low light inputs and then integrate these priors to modulate the LLIE process.

# Chapter 3

# Method

The contribution of our work mainly lies in the integration of illumination-invariant structure priors and the design of multi-scale CNN-Transformer hybrid network. The framework of our method is illustrated in Figure 3.1, 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 [33], which serves as a task-adaptive edge detector. CIConv applies

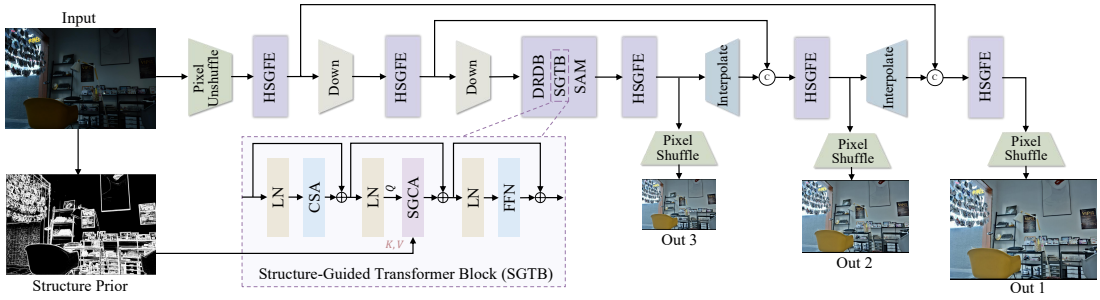


Figure 3.1: The overall framework of our propose method. We develop our method based on ESDNet [65] and adopt a similar UNet architecture. At each level of the encoder and decoder, our customized Hybrid Structure-Guided Feature Extractor (HSGFE) module is employed. Within each HSGFE, besides the Dilated Residual Dense Block (DRDB) and Semantic-Aligned Scale-Aware Module (SAM) proposed in ESDNet [65], we first extract structure priors based on color invariant layers [33] and then develop the Structure-Guided Transformer Block (SGTB) to integrate these priors as guidance. With the integration of illumination-invariant 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.

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—E, W, C, N, and H—based on the Kubelka-Munk (KM) reflection model [51, 18, 16], we adopt $W$ as it provides robust edge detection under varying illumination, shading, and reflectance conditions:

$$W_{out} = CIConv(I_{in}) \tag{3.1.1}$$

Here, $I_{in}$ denotes the input image, specifically the low-light image to be enhanced. $W_{\text{out}}$ represents the structural prior, which is later integrated into the Structure-Guided Transformer Block (SGTB) for guidance. Regarding to *CIConv*, the fomular is like following:

$$\text{CIConv}(I_{in}) = \frac{\log\left(\text{W}^2(I_{in}) + \epsilon\right) - \mu_{\mathcal{S}}}{\sigma_{\mathcal{S}}} \tag{3.1.2}$$


Figure 3.2: Example visualization of the extracted structure priors.

Here, $\mu_{\mathcal{S}}$, $\sigma_{\mathcal{S}}$ and $\epsilon$ refer to the sample mean, standard deviation, and small perbutation. To compute $\mathrm{W}(I_{in})$, we first use the Gaussian Color Model [17] to obtain the initial edge detectors, denoted as $E$. Then, we use $E$ to derive the second phase of edge detectors, denoted as $W$, as follows:

$$W = \sqrt{W_x^2 + W_{\lambda x}^2 + W_{\lambda\lambda x}^2 + W_y^2 + W_{\lambda y}^2 + W_{\lambda\lambda y}^2}, \tag{3.1.3}$$

$$W_x = \frac{E_x}{E}, \quad W_{\lambda x} = \frac{E_{\lambda x}}{E}, \quad W_{\lambda\lambda x} = \frac{E_{\lambda\lambda x}}{E} \tag{3.1.4}$$

Finally, we obtain the structure prior map based on the above formulas and the input image $I_{in}$. This structure prior is then used to guide our customized attention block in Section 3.2.

**Physical Explanation** These equations mathematically indicates that $W$ characterizes the spatial derivatives of spectral intensity. We visualize our extracted structure prior $W$ for low-light and normal-light in Fig. 3.2. It is clear that $W$ represents the stable edge and structure map across images with differing illumination conditions, highlighting its great potential to guide the enhancement process.

## 3.2 Model Architecture

As illustrated in Fig. 3.1, our model adopts the encoder-decoder architecture. The PixelShuffule and several convolutional layers (denoted as "Down" in Fig. 3.1) are utilized for down-sampling, and the PixelShuffle or Interplation 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.

### 3.2.1 HSGFE Module

HSGFE Module is designed to exploit structural cues inherent in 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 in Section 3.1 are explicitly injected into the feature flow, guided by invariant edge descriptors [17]. Moreover, the Semantic-Aligned Scale-Aware Module (SAM) is incorporated to further accommodate scale diversity across scenes.

### 3.2.2 Structure-Guided Transformer Block (SGTB)

As shown in Figure 3.1, 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:

$$F_{out} = CSA(LN(F_{in})) + F_{in}, \tag{3.2.1}$$

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 dependence while ensuring model efficiency. Then, the processed representations are passed into the SGCA for further refinement.

### 3.2.3 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 $\mathbf{X_F}$ is linearly projected to generate the Query ($\mathbf{Q}$) representation:

$$\mathbf{Q} = \mathbf{X_F}\mathbf{W}_Q^{\top}. \tag{3.2.2}$$

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:

$$\mathbf{K_p} = \mathbf{X_s}\mathbf{W}_{K_p}^{\top}, \quad \mathbf{V_p} = \mathbf{X_s}\mathbf{W}_{V_p}^{\top}, \tag{3.2.3}$$

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, and $\mathbf{X_s}$ stands for the structure priors at corresponding levels. We then formulate our structure-guided attention mechanism as follows:

$$Attention(Q, K_p, V_p) = softmax(\frac{Q \cdot K_p}{\lambda}) \cdot V_p, \tag{3.2.4}$$

where $\lambda$ 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.

### 3.2.4 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: $Y_{in,0}$, $Y_{in,1}$, and $Y_{in,2}$.

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 $\alpha_i$, 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:

$$[\beta_0, \beta_1, \beta_2 = MLP(\alpha_0, \alpha_1, \alpha_2)]. \tag{3.2.5}$$

Finally, the fused feature map $F_{\text{fused}}$ is obtained as follows:

$$F_{fused} = F_{in,0} + \beta_1 \odot Y_{in,1} + \beta_2 \odot Y_{in,2}. \tag{3.2.6}$$

Additionally, skip connections are incorporated within this module to retain more information from the input feature maps.

## 3.3 Loss Function

As Figure 3.1 shows, for network training, we design the loss function based on three levels of output images—$\hat{I}_1$, $\hat{I}_2$, and $\hat{I}_3$—which correspond to the output images at

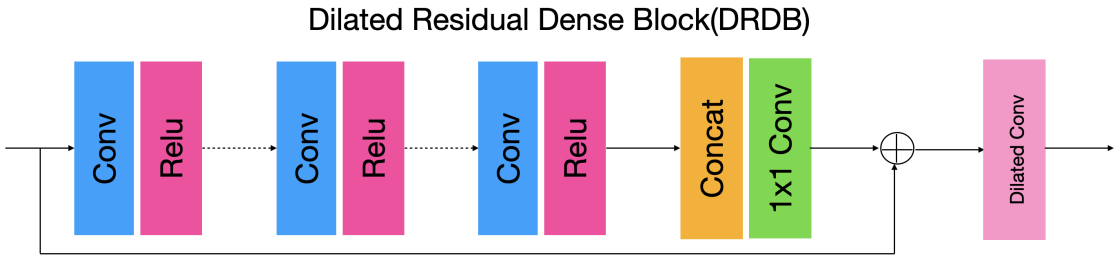


Figure 3.3: The architecture of the Dilated Residual Dense Block, originally proposed by [65], is modified for use in our network model.

three different resolutions from the respective levels of our decoder:

$$L_{\text{total}} = \sum_{i=1}^{3} L_C\left(I_i, \hat{I}_i\right) + \lambda \cdot \sum_{i=1}^{3} L_P\left(I_i, \hat{I}_i\right) + \gamma \cdot \sum_{i=1}^{3} L_{\text{MS-SSIM}}\left(I_i, \hat{I}_i\right) \tag{3.3.1}$$

Here, $L_1$, $L_{\text{P}}$, and $L_{\text{MS-SSIM}}$ represent the Charbonnier loss [31], perceptual loss [29], and Multi-Scale SSIM (MS-SSIM) loss, respectively. The weighting factors are set as $\lambda = 0.01$ and $\gamma = 0.4$.

# Chapter 4

# Experiments

## 4.1 Performance on NTIRE 2025 LLIE Challenge

### 4.1.1 Chanllenge Details

This project was undertaken as part of the Low-Light Image Enhancement (LLIE) challenge [41], a globally recognized competition track hosted by the NTIRE workshop alongside CVPR. Widely respected in the computer vision community, the challenge draws participation from diverse research teams worldwide, all aiming to push the limits of visual enhancement in dim lighting scenarios. The track offers a shared testing ground that promotes creativity while enabling fair performance comparisons among different approaches.

### 4.1.2 Dataset Details

We use the dataset provided by the *NTIRE 2025 Low Light Image Enhancement Challenge* [41]. This dataset consists of 219 training images, 46 validation images, and

Figure 4.1: **Example training pairs from the subset of the *NTIRE* 2025 dataset.** Each pair includes a low-light input and its corresponding high-quality ground truth image.

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 FiveK [4], 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 Image Enhancement Challenge* [40] for fine-tuning.

### 4.1.3 Implementation Details

We develop our model using PyTorch and train it on a single A100 GPU. Regarding the competition [41], the model is trained on paired data from the *NTIRE 2025 LLIE Challenge* [41] dataset, without any pretrained weights. The model is optimized with the Adam optimizer [30] ($lr = 2 \times 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$), 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 per GPU. We adopt a two-stage learning rate schedule using *CosineAnnealingRestartCyclicLR*: an initial phase of 46,000 iterations at a constant

rate of $3 \times 10^{-4}$, followed by cosine annealing down to $1 \times 10^{-6}$ over 104,000 iterations. The loss function is described in Sec. 3.3. Data augmentation includes geometric transforms and mixup with $\beta = 1.2$ and identity mapping enabled.

### 4.1.4 Quantitative and Qualitative on NTIRE 2025 LLIE Challenge

As presented in Tab. 4.4, our method achieves the highest PNSR score, 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 performance of our proposed approach. The enhancement outputs of our method are presented in Fig. 1.2, which demonstrate 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

### 4.2.1 Datasets

**LOL-v1** This dataset includes a total of 500 image pairs—485 for training and 15 for testing. All images were taken with DSLR cameras in genuinely low-light conditions, mostly capturing indoor scenes. The average resolution is around 400 $\times$ 600 pixels, which makes the dataset suitable for GPU-based training and testing without needing to crop or resize the input. LOL-v1 [58] has become a widely used benchmark in the field of low-light image enhancement. Training pairs of examples are given in Figure 4.3.

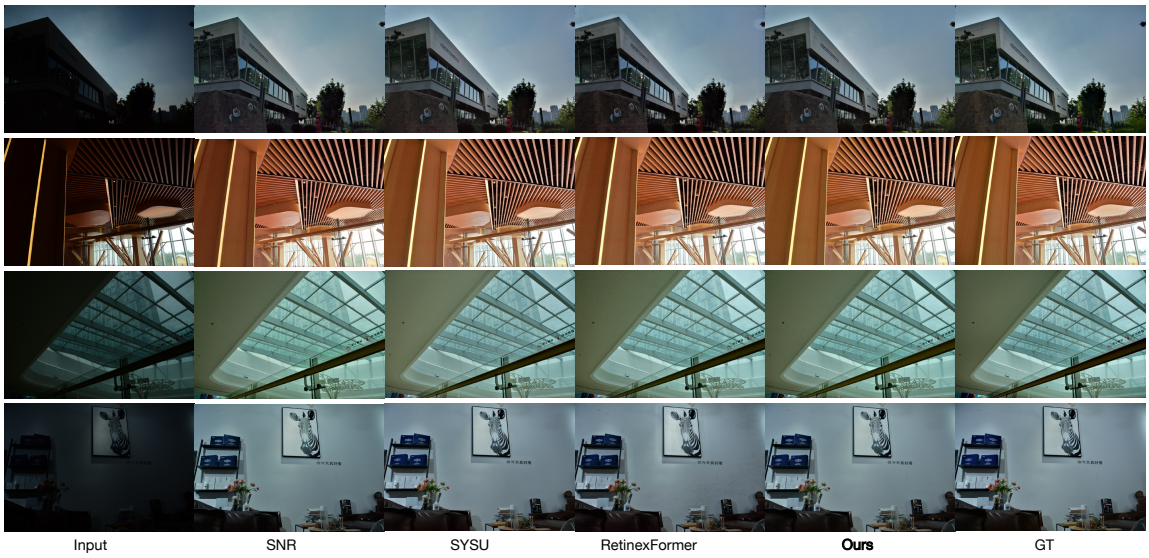


Figure 4.2: Qualitative comparisons on the *NTIRE 2025 LLIE Challenge* dataset. We compare our model with SNR [60], UHDM (SYSU) [40], and Retinexformer [6]. Our model consistently performs best or comparably well. Please zoom in for a better view.

Figure 4.3: **Example training pairs from the LOL-v1 dataset**. Each pair consists of a low-light input and its corresponding well-lit ground truth image.

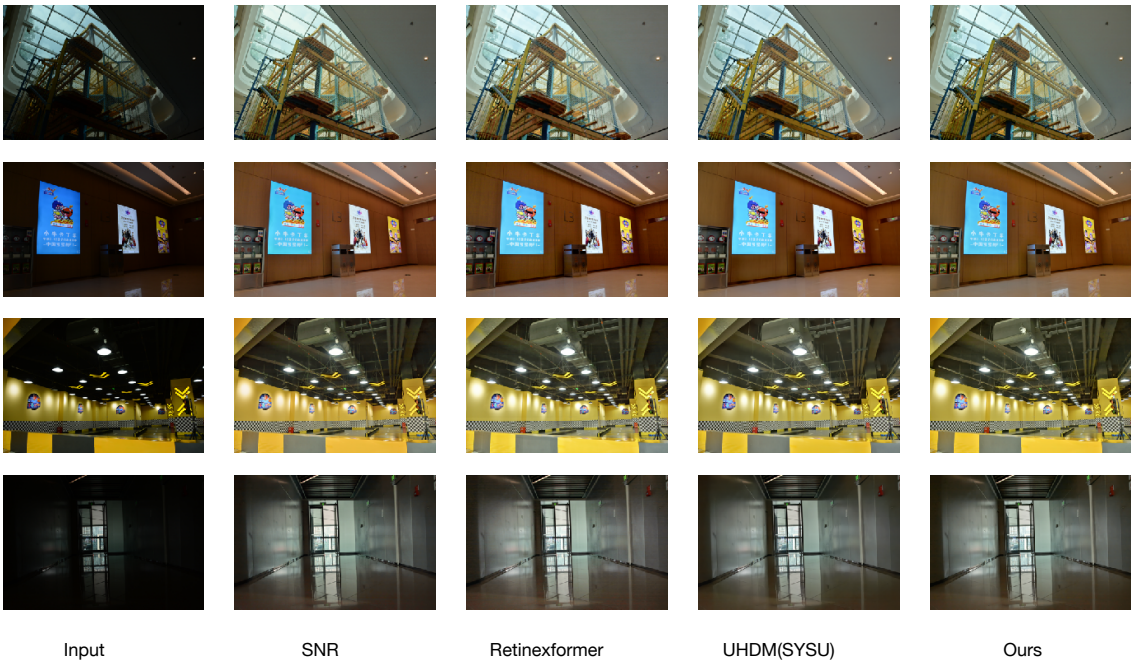


Figure 4.4: Additional examples of our results compared to other state-of-the-art (SOTA) methods.(1)

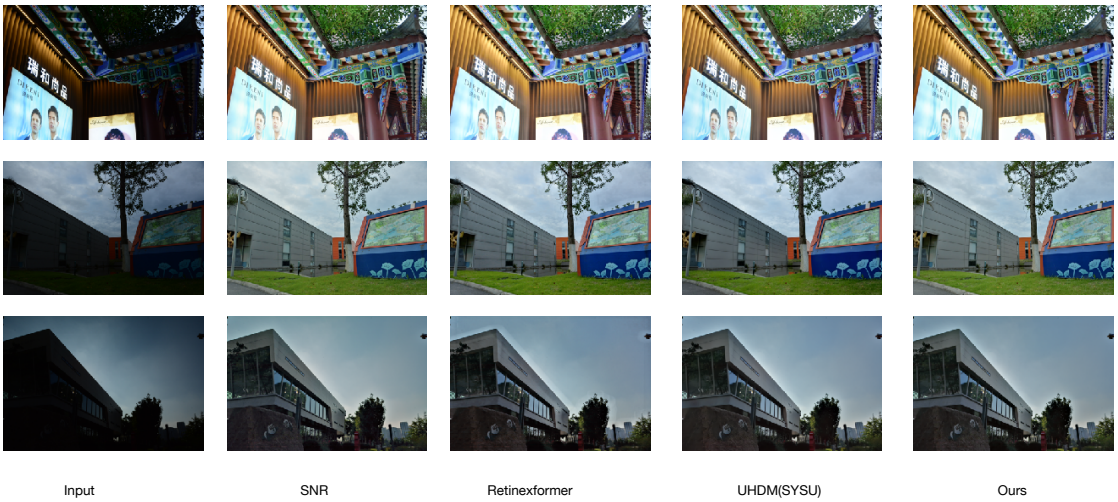


Figure 4.5: Additional examples of our results compared to other state-of-the-art (SOTA) methods.(2)

Figure 4.6: **Example training pairs from the real subset of the LOL-v2 dataset.** Each pair includes a low-light input and its corresponding high-quality ground truth image.

Figure 4.7: **Example training pairs from the synthetic subset of the LOL-v2 dataset.** Each pair includes a low-light input and its corresponding high-quality ground truth image.

**LOL-v2** This dataset is another well-known benchmark for low-light image enhancement, and it serves as an extension of LOL-v1. LOL-v2 introduces both real and synthetic scenes, covering a mix of indoor and outdoor environments. Compared to LOL-v1, it offers a larger number of high-resolution image pairs—689 for training and 100 for testing. First introduced by Yang *et* al. [62] in 2020, it has since become one of the most widely used benchmarks in the low-light enhancement field. Both real and synthetic subsets of LOL-v2 are shown in Figure 4.6 and Figure 4.7, respectively.

### 4.2.2 Experiment Settings

To comprehensively evaluate the performance of our method, we compare our method with KinD [69], MIRNet [66], Restormer [68], LLFlow [56], Retinexformer [6], SNR [60], UHDM (refined by Team SYSU-FVL-T2 [40] in *NTIRE 2024 LLIE Challenge*) on the *NTIRE 2025 LLIE* dataset. This dataset contains 219 training pairs, 46 and 30 low light images for validation and testing. 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 $512 \times 512$ 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 [63], where the patch size and batch size are set to $384 \times 384$ and 4, respectively.

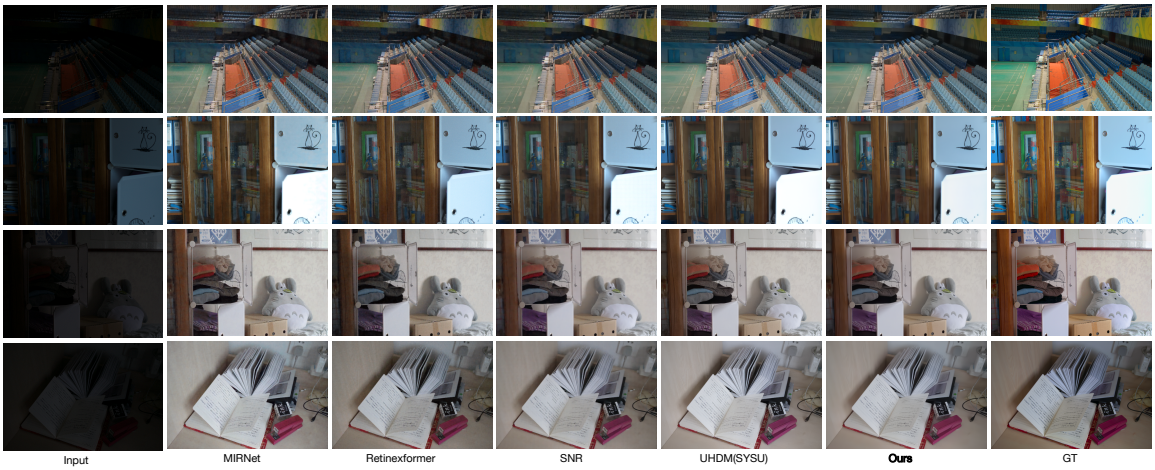


Figure 4.8: Visual comparisons on the LOL-v1 dataset, with MIRNet [66], Retinex-Former [6], SNR [60], and UHDM (SYSU) [40].

### 4.2.3 Quantitative Results

The quantitative comparisons on the *NTIRE 2025 LLIE*, LOL-v1, and LOL-v2-real datasets are shown in Tab 4.1. On LOL-v1, we obtain the highest SSIM (0.873) and the lowest LPIPS (0.092), 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 UHDM (0.111).

### 4.2.4 Qualitative Results

We present visual comparisons of our model with the methods listed in Table 4.1. As shown in Figure 4.8, Figure 4.9, and Figure 4.2, 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

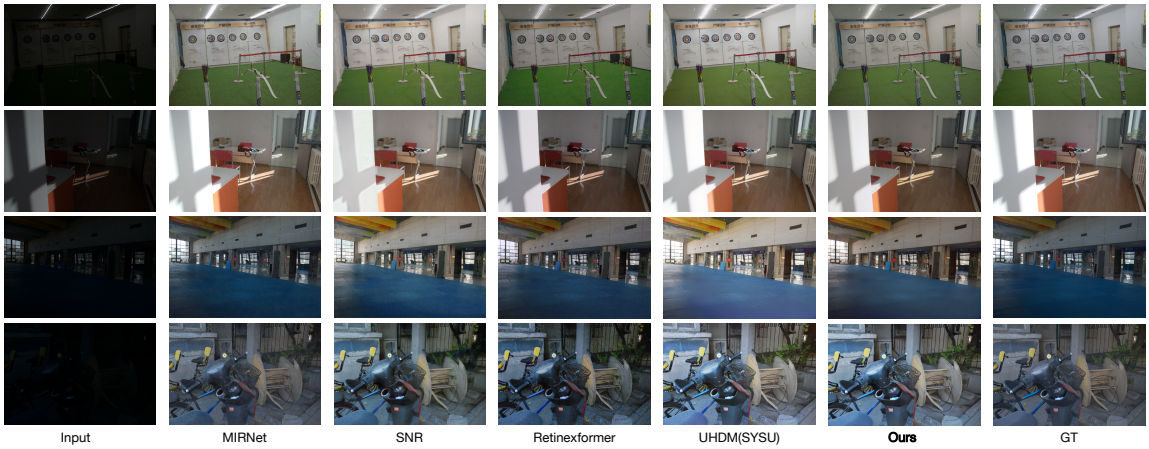


Figure 4.9: Visual comparisons on LOL-v2-real dataset with MIRNet [66], Retinex-Former [6], SNR [60], and UHDM (SYSU) [40].

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.

| **Methods** | **LOL-v1** | | | **LOL-v2-real** | | | **NTIRE 2025 LLIE** | | |
|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| KinD [69] | 20.87 | 0.802 | 0.207 | 17.54 | 0.669 | 0.365 | — | — | — |
| MIRNet [66] | 24.14 | 0.832 | 0.131 | 19.35 | 0.708 | 0.138 | — | — | — |
| SNR [60] | 24.61 | 0.842 | 0.151 | 20.82 | 0.811 | 0.161 | 24.39 | 0.878 | 0.149 |
| Restormer [67] | 22.43 | 0.823 | 0.147 | 18.69 | 0.834 | 0.232 | — | — | — |
| Retinexformer [6] | [1]**25.16** | 0.845 | 0.131 | 22.79 | 0.840 | 0.171 | 25.06 | 0.872 | 0.183 |
| LLFlow [56] | 21.09 | [2]0.861 | 0.116 | 17.43 | 0.831 | 0.129 | — | — | — |
| UHDM (SYSU) [40] | 23.10 | 0.846 | [2]0.094 | 20.48 | 0.841 | 0.134 | 26.26 | 0.892 | [1]**0.111** |
| **Ours** | [2]24.63 | [1]**0.873** | [1]**0.092** | [1]**22.84** | [1]**0.859** | [1]**0.126** | [1]**26.75** | [1]**0.899** | [2]0.113 |

Table 4.1: Quantitative comparison on LOL-v1, LOL-v2-real, and *NTIRE 2025 LLIE* datasets. [1] and [2] indicate the best and second-best performances, respectively. "——" denotes ongoing experiments.

| **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 4.2: 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.

| Input | Training Time | Epochs | Extra Data | Diffusion | Attention | Quantization | # Params. (M) | Runtime | GPU |
|---|---|---|---|---|---|---|---|---|---|
| (1600, 1600, 3) | 36h | 260 | No | No | Yes | No | 12.7 | 0.7s on GPU | A100 |

Table 4.3: Details of our model for the *NTIRE 2025 LLIE Challenge*

## 4.3 Ablation Study

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

### 4.3.1 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. 4.2 reports the quantitative performance of this modification on LOL-v1 dataset, which still achieve competitive enhancement performance on all measured metrics compared to UHDM(SYSU), Restormer, and MIRNet in Tab. 4.1. 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.

| 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 4.4: 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.

### 4.3.2 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. 4.2. 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.

### 4.3.3 Loss Function

As introduced in Sec. 4.2, we add the muti-scale MS-SSIM loss into our complete optimization objective, which is not included in UHDM(SYSU). To verify the effectiveness of this new introduced loss function, we remove it and adopt the same optimization strategy in UHDM(SYSU). The quantitative results are reported in Tab. 4.2, which demonstrate that the integration of multi-scale MS-SSIM loss helps achieve higher SSIM performance. The loss function of UHDM (SYSU) is defined as follows:

$$L_{\text{total}} = \sum_{i=1}^{3} L_C\left(I_i, \hat{I}_i\right) + \lambda \cdot L_{Percep}\left(I_i, \hat{I}_i\right), \tag{4.3.1}$$

where $L_1$ and $L_{Percep}$ represent Charbonnier loss [31], and perceptual loss[29], respectively. The weighting factor is set as $\lambda = 0.04$.

# Chapter 5

# Future Improvements

**Evaluation and Metrics** As stated in Section 3.3, we have incorporated PSNR, LPIPS, and MS-SSIM into our loss function. However, we plan to include additional metrics in future versions of the loss, such as NIQE. Furthermore, we aim to introduce a User Study Score in our comparison with other models. A group of participants will be invited to rate the enhanced results produced by our model and others. The scoring will range from 1 to 5, allowing participants to evaluate image quality based on their visual perception.

**Structure Priors** As introduced in Section 3.2, we have incorporated structure priors into our model. Lengyel *et al.* proposed five color invariants—E, W, C, N, and H. In our current approach, we use W as the source of structure priors. However, the remaining four invariants have not been thoroughly explored. In future work, we plan to substitute the current structure prior with each of the other invariants (E, C, N, and H) to fully investigate their potential and better understand the role of structure priors in our framework.

**Model Architecture** As shown in Figure 3.1, each SGTB block in our model includes only one attention module. As the patch size increases, the computational cost of the transformer-like architecture also grows. To address this, we plan to introduce multi-head attention in future versions of our model. This would allow the network to better capture both local and non-local information within the input patches.

# Chapter 6

# Conclusion

We present an encoder–decoder architecture augmented with Hybrid Structure-Guided Feature Extractors (HSGFEs) 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, a semantic-aligned, scale-aware module is incorporated, and skip connections are maintained 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, which achieveds the **best PSNR score** and **second-best overall performance** among 28 teams globally in the *NTIRE 2025 Low Light Enhancement Challenge* [41].

# Bibliography

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

[2] N. Alshammari, S. Akcay, and T. 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] S. Arjomand Bigdeli, M. Zwicker, P. Favaro, and M. Jin. Deep mean-shift priors for image restoration. *Advances in neural information processing systems*, 30, 2017.

[4] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011.

[5] Y. Cai, J. Lin, H. Wang, X. Yuan, H. Ding, Y. Zhang, R. Timofte, and L. V. Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive

imaging. *Advances in Neural Information Processing Systems*, 35:37749–37761, 2022.

[6] 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*, pages 12504–12513, 2023.

[7] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021.

[8] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. 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.

[9] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yuan, et al. Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35: 25478–25490, 2022.

[10] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *European Conference on Computer Vision*, pages 669–687. Springer, 2022.

[11] Y. Deng, C. C. Loy, and X. Tang. Aesthetic-driven image enhancement by

adversarial learning. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 870–878, 2018.

[12] B. Dogan, S. Gu, and R. Timofte. Exemplar guided face image super-resolution without facial landmarks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[13] B. Dolhansky and C. C. 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.

[14] 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.

[15] X. Fu, Y. Liao, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding. A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation. *IEEE Transactions on Image Processing*, 24(12):4965–4977, 2015.

[16] J.-M. Geusebroek. Color and geometrical structure in images. *Appl. Microsc*, 2000.

[17] J.-M. Geusebroek, R. Van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern analysis and machine intelligence*, 23(12):1338–1350, 2001.

[18] T. Gevers, A. Gijsenij, J. Van de Weijer, and J.-M. Geusebroek. *Color in computer vision: Fundamentals and applications*. John Wiley & Sons, 2012.

[19] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[20] S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang. Learning dynamic guidance for depth image enhancement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3769–3778, 2017.

[21] 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*, pages 1780–1789, 2020.

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

[23] Y. Hu, H. He, C. Xu, B. Wang, and S. Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2):1–17, 2018.

[24] 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*, 22(3):1032–1041, 2012.

[25] D. A. Hudson and L. Zitnick. Generative adversarial transformers. In *International conference on machine learning*, pages 4487–4499. PMLR, 2021.

[26] T.-W. Hui, C. C. Loy, and X. 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.

[27] H. Ibrahim and N. S. P. Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007.

[28] X. Jiang, X. Zhang, N. Gao, and Y. Deng. When fast fourier transform meets transformer for image restoration. In *European Conference on Computer Vision*, pages 381–402. Springer, 2024.

[29] J. Johnson, A. Alahi, and L. 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.

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

[31] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. 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.

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

[33] A. Lengyel, S. Garg, M. Milford, and J. 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.

[34] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE transactions on image processing*, 27(6):2828–2841, 2018.

[35] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 272–289, 2018.

[36] X. Li, W. Li, D. Ren, H. Zhang, M. Wang, and W. 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.

[37] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. 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.

[38] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[39] L. Liu, L. Xie, X. Zhang, S. Yuan, X. Chen, W. Zhou, H. Li, and Q. Tian. Tape: Task-agnostic prior embedding for image restoration. In *European Conference on Computer Vision*, pages 447–464. Springer, 2022.

[40] X. Liu, Z. Wu, A. Li, F.-A. Vasluianu, Y. Zhang, S. Gu, L. Zhang, C. Zhu, R. 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.

[41] X. Liu, Z. Wu, F.-A. Vasluianu, H. Yan, R. Bin, Y. Zhang, S. Gu, L. Zhang, C. Zhu, R. Timofte, et al. Ntire 2025 challenge on low light image enhancement: Methods and results. In *CVPRW*, 2025.

[42] Y. Liu, T. Huang, W. Dong, F. Wu, X. Li, and G. Shi. Low-light image enhancement with multi-stage residue quantization and brightness-aware attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12140–12149, 2023.

[43] K. G. Lore, A. Akintayo, and S. Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

[44] D. Makwana, G. Deshmukh, O. Susladkar, S. Mittal, et al. Livenet: A novel network for real-world low-light image denoising and enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5856–5865, 2024.

[45] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. 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.

[46] J. Pan, Z. Hu, Z. Su, and M.-H. 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.

[47] S. Park, S. Yu, B. Moon, S. Ko, and J. Paik. Low-light image enhancement using variational optimization-based retinex model. *IEEE Transactions on Consumer Electronics*, 63(2):178–184, 2017.

[48] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, G. D. Al-Quaderi, and M. Shoyaib. An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing*, 2016:1–13, 2016.

[49] B. Ren, Y. Li, J. Liang, R. Ranjan, M. Liu, R. Cucchiara, L. Van Gool, and N. Sebe. Key-graph transformer for image restoration. *arXiv preprint arXiv:2402.02634*, 2024.

[50] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 181–190, 2019.

[51] S. A. Shafer. Using color to separate reflection components. *Color Research & Application*, 1985.

[52] H. Shakibania, S. Raoufi, and H. Khotanlou. Cdan: convolutional dense attention-guided network for low-light image enhancement. *Digital Signal Processing*, 156:104802, 2025.

[53] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[54] C. Wang and Z. Jin. Brighten-and-colorize: A decoupled network for customized

low-light image enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8356–8366, 2023.

[55] S. Wang, J. Zheng, H.-M. Hu, and B. Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing*, 22(9):3538–3548, 2013.

[56] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. C. Kot. Low-light image enhancement with normalizing flow. *arXiv preprint arXiv:2109.05923*, 2021.

[57] 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*, pages 17683–17693, 2022.

[58] C. Wei, W. Wang, W. Yang, and J. Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[59] J. Xiao, X. Fu, F. Wu, and Z.-J. Zha. Stochastic window transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:9315–9329, 2022.

[60] 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*, pages 17714–17724, 2022.

[61] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, 35(2):1–15, 2016.

[62] 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*, pages 3063–3072, 2020.

[63] 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*, 30:2072–2086, 2021.

[64] Z. Yang, H. Chen, Z. Qian, Y. Zhou, H. Zhang, D. Zhao, B. Wei, and Y. Xu. Region attention transformer for medical image restoration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–613. Springer, 2024.

[65] X. Yu, P. Dai, W. Li, L. Ma, J. Shen, J. Li, and X. Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *European Conference on Computer Vision*, pages 646–662. Springer, 2022.

[66] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020.

[67] 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 *CVPR*, 2022.

[68] 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*, pages 5728–5739, 2022.

[69] 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*, pages 1632–1640, 2019.

[70] Y. Zhang, B. Teng, D. Yang, Z. Chen, H. Ma, G. Li, and W. Ding. Learning a single convolutional layer model for low light image enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5995–6008, 2023.

[71] H. Zhao, Y. Gou, B. Li, D. Peng, J. Lv, and X. Peng. Comprehensive and delicate: An efficient transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14122–14132, 2023.