LIGHT ATTENTION LIGHTS YOUR WORLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-light image enhancement remains challenging under spatially non-uniform illumination. Existing unified restoration pipelines often over-enhance bright areas or insufficiently restore severely underexposed regions, while processing high-resolution images incurs substantial computational overhead. To address these challenges, we propose L^4Net, a **L**ightweight **L**ight-Attention Network for **L**ow-**L**ight Image Enhancement. L^4Net performs restoration in the low-resolution feature space using two structurally identical illumination-aware branches that adaptively restore differently illuminated regions. Light-Attention (LA) blocks, the core components of the lightweight L^4 backbone, integrate global illumination modeling and local feature interactions. A light-guided learning strategy further encourages branch specialization. A super-resolution reconstruction module recovers fine details from compact feature representations. Extensive experiments on multiple low-light benchmarks demonstrate that L^4Net achieves state-of-the-art performance with substantially lower computational cost. Deployment on an embedded platform further demonstrates its practicality for resource-constrained low-light vision applications. Code will be released upon publication.

1 INTRODUCTION

Low-light image enhancement (LLIE) aims to recover perceptually natural and structurally faithful images from observations captured under insufficient illumination. LLIE is important for nighttime surveillance, autonomous driving, mobile photography, and robotic perception, where low contrast, amplified noise, color distortion, and missing structures degrade both visual quality and downstream perception tasks Huang et al. (2025); Liu et al. (2026b); Liao et al. (2025).

Existing methods commonly adopt end-to-end architectures Liao et al. (2025); Feijoo et al. (2025) to learn direct mappings from low-light inputs to normally exposed outputs. To improve illumination, some approaches Fu et al. (2023a); Fan et al. (2025) incorporate physical priors or imaging models to decouple illumination components from low-light observations, achieving considerable progress in restoring images affected by globally insufficient illumination. However, these approaches remain limited in addressing a fundamental conflict introduced by spatially non-uniform illumination: well-exposed and severely underexposed regions may coexist within a single image but require different restoration strengths. A unified restoration strategy must conservatively preserve visible content while aggressively recovering weak signals in dark regions Jin et al. (2025). This conflict often leads to local over-enhancement in bright areas or insufficient detail recovery in dark ones.

Furthermore, illumination variations in real-world low-light scenes cause both overall image degradation and detail loss in underexposed regions Cho et al. (2026); Xu et al. (2026). CNN-based LLIE methods Fu et al. (2023b); Feijoo et al. (2025); Yan et al. (2025) effectively capture local feature patterns but struggle to model global degradation characteristics. To overcome this limitation, recent studies have introduced Transformer-based architectures Cui et al. (2022); Wang et al. (2024); Yi et al. (2023); Balmez et al. (2026) to model global illumination variations by aggregating non-local context across the image. However, their considerable computational cost remains a major challenge, particularly when processing high-resolution (HR) low-light images Yi et al. (2023).

To address these issues, we propose L^4Net, a lightweight and efficient framework for HR LLIE that handles spatially varying illumination while reducing computational cost. Instead of treating all image regions uniformly, L^4Net learns complementary representations for bright and dark regions through dual branches built upon the same L^4 backbone. To enable illumination-aware and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

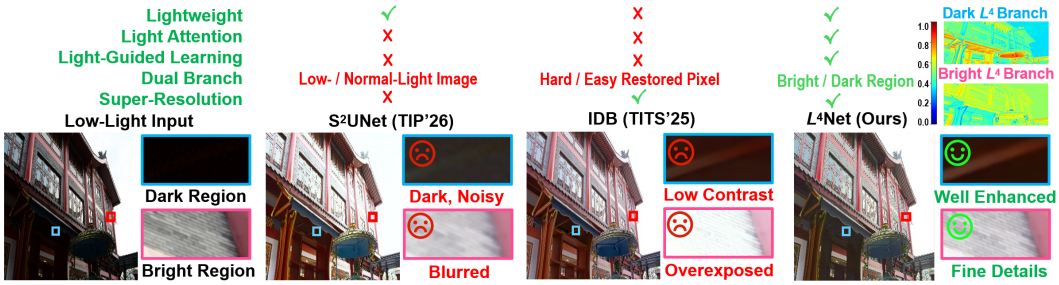


Figure 1: Highlights of the proposed L^4Net. Compared with state-of-the-art methods, L^4Net achieves high-fidelity restoration under non-uniform illumination by effectively enhancing underexposed regions while preserving structural details in well-exposed regions.

lightweight feature enhancement, we introduce Light-Attention (LA) blocks as the core components of the L^4 backbone. LA couples Transformer-based attention for modeling long-range illumination variations with convolutional layers for local feature interaction. The computational cost is further reduced via enhancement in the low-resolution (LR) space with HR reconstruction. As illustrated in Fig. 1, compared with two existing dual-branch methods, L^4Net more effectively restores severely dark regions while retaining fine details in relatively bright areas. Extensive experiments on multiple benchmarks validate the effectiveness of L^4Net, while edge-device evaluation confirms its practicality under computational resource constraints. Our main contributions are summarized as follows:

- We propose L^4Net, a **L**ightweight **L**ight-Attention Network for **L**ow-**L**ight Image Enhancement. L^4Net employs two structurally identical illumination-aware branches and a super-resolution (SR) reconstruction module for efficient and robust enhancement under spatially non-uniform illumination.
- We design lightweight Light-Attention (LA) blocks that integrate global illumination modeling with local feature enhancement, serving as the fundamental block of our L^4 backbone.
- We propose a light-guided learning strategy that provides branch-specific supervision through adaptive luminance masks and incorporates light-aware text prompts as multimodal semantic guidance.

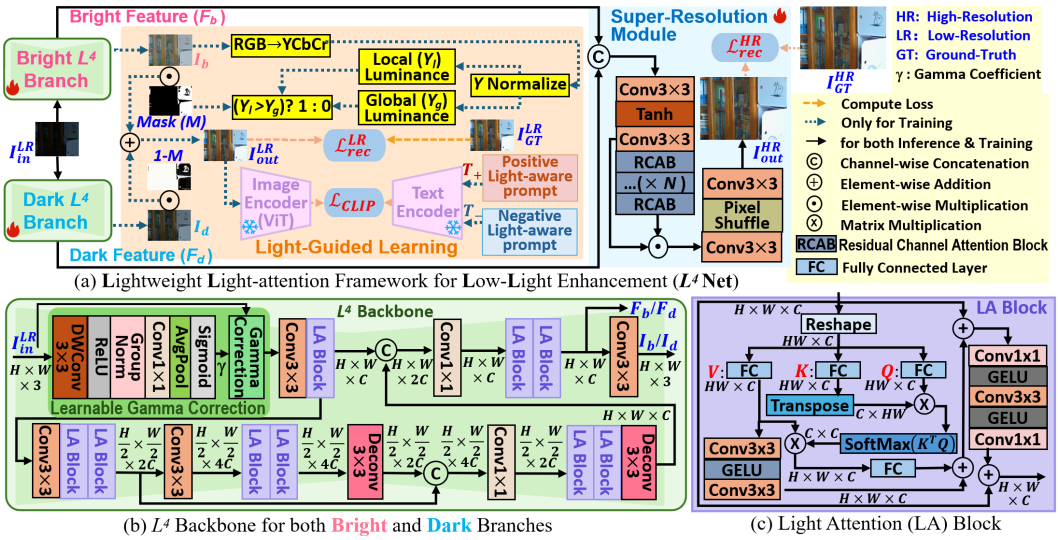
2 Related Work

2.1 Learning-Based Low-Light Image Enhancement

Existing learning-based methods for LLIE employ diverse illumination modeling strategies, including region-adaptive processing and Retinex-based decomposition. SNR Xu et al. (2022) combines Transformer and convolutional networks to perform spatially varying feature fusion guided by signal-to-noise information. Retinexformer Cai et al. (2023) and URetinex-Net++ Wu et al. (2025b) provide effective and interpretable illumination modeling using a unified framework for different luminance regions. SHAL-Net Xu et al. (2024) restores image structures and contrast before enhancement and decouples reflectance and illumination through a decomposition network. S^{2}UNet Liu et al. (2026a) decomposes shared reflectance and independent illumination components from an input image and its gamma-corrected version. IDB Li et al. (2025) introduces an isomorphic dual-branch architecture to process pixels with different restoration difficulties rather than varying illumination levels. Although these methods effectively model illumination variations, they rely on a single restoration pathway for the entire image. In contrast, L^4Net employs adaptive luminance masks to supervise two parameter-independent branches built upon identical light-attention backbones, enabling specialized restoration of well-exposed and underexposed regions.

2.2 Lightweight Models for Low-Light Enhancement

Efficiency in learning-based LLIE has been improved through lightweight architectures, efficient attention mechanisms, and alternative feature representations. Representative methods include the



(a) **L**ightweight **L**ight-attention Framework for **L**ow-**L**ight Enhancement (***L*⁴ Net**)

(b) L^4 Backbone for both **Bright** and **Dark** Branches

(c) Light Attention (LA) Block

Figure 2: Overall architecture of the proposed L^4Net.

compact illumination mappings of IAT Cui et al. (2022), the efficient attention of Restormer Zamir et al. (2022), Fourier-domain modeling of FourLLIE Wang et al. (2023a), and lightweight coarse-to-fine reconstruction framework of RFLLIE Jiang et al. (2024). Although these methods reduce model complexity, they still perform restoration in the original-resolution feature space, resulting in considerable computational cost for HR inputs. In contrast, L^4Net performs illumination enhancement in the LR feature space using the lightweight L^4 backbone and reconstructs HR outputs only at the final stage, substantially reducing computational cost.

3 Method

3.1 L^4Net Framework

As illustrated in Fig. 2, L^4Net takes a LR low-light image $\mathbf{I}_{\text{in}}^{\text{LR}}$ as input and reconstructs an enhanced HR image $\mathbf{I}_{\text{out}}^{\text{HR}}$. The input is processed by two branches, referred as bright and dark branches, each built upon an identical L^4 backbone with unshared parameters. Guided by an adaptive luminance mask, the two branches specialize in preserving reliable content from relatively well-exposed regions and recovering weak signals from severely underexposed regions, respectively. They produce branch-specific enhanced images $\mathbf{I}_b$ and $\mathbf{I}_d$, together with feature representations $\mathbf{F}_b$ and $\mathbf{F}_d$. During training, the light-guided learning strategy employs the luminance mask for region-specific supervision and light-aware text prompts for multimodal semantic guidance. The branch outputs are spatially combined to generate an enhanced LR result, while their complementary features are fused by the SR reconstruction module to recover fine structural details and produce the final HR output. This design performs enhancement in the compact LR space and reserves HR computation for reconstruction. The following subsections describe the L^4 backbone, light-guided learning strategy, and SR reconstruction module.

3.2 Lightweight Light Attention L^4 Backbone

We design a lightweight L^4 backbone based on light attention to restore region-specific features under spatially varying illumination, as shown in Fig. 2(b). L^4 backbone consists of a learnable gamma Wang et al. (2023b) correction block followed by a U-shaped architecture built primarily from LA blocks.

Learnable Gamma Correction. Given a low-light and low-resolution input image $\mathbf{I}_{\text{in}}^{\text{LR}}$, the block learns the gamma coefficient γ and applies gamma correction for pre-enhancement, providing illumination-adaptive input conditioning for subsequent feature restoration.

Light Attention Block. To restore features degraded by spatially varying illumination, the backbone adopts the LA block. Since non-uniform illumination exhibits global variation, capturing long-range dependencies is essential for modeling lighting patterns. However, local convolutions cannot adequately capture global context, whereas stacking multiple spatial attention layers incurs substantial computational overhead. Therefore, the LA block performs global interaction in the channel domain, enabling illumination-aware feature modeling with low computational cost. As illustrated in Fig. 2(c), the input features are reshaped and projected by lightweight fully connected (FC) layers to generate channel-wise query (Q), key (K), and value (V) embeddings. These FC projections reduce cost of feature transformation. Following the transposed attention formulation Zamir et al. (2022); Wu et al. (2025a), channel attention is computed as $K^T Q$, producing a $C \times C$ attention matrix instead of an $HW \times HW$ spatial attention map. This formulation captures long-range channel dependencies while reducing computational complexity. To complement global illumination modeling, the value branch (V) is refined by 3×3 convolutions with GELU activations before attention modulation, enhancing illumination-sensitive features and preserving fine structures in underexposed regions. The attended features are then fused with the locally refined representations, while residual connections facilitate gradient flow and stabilize training.

Overall, the L^4 backbone integrates illumination-adaptive pre-enhancement with efficient global and local feature modeling, enabling robust restoration under spatially varying illumination while maintaining high computational efficiency.

3.3 Light-Guided Learning Strategy

Although the bright and dark branches share the same L^4 backbone, simply aggregating their features does not explicitly enforce branch specialization. To address this issue, we propose a light-guided learning strategy that supervises each branch within its corresponding luminance range using adaptive masks, while providing semantic guidance through light-aware text prompts. As shown in Fig. 2(a), the two branches produce enhanced LR images $\mathbf{I}_b$ and $\mathbf{I}_d$, which are used to compute LR reconstruction and CLIP losses before HR reconstruction.

Adaptive Luminance Mask. Given the bright branch output $\mathbf{I}_b \in R^{H\times W\times 3}$, we convert it to the YCbCr color space Cheng et al. (2026) and normalize its luminance channel $\mathbf{Y}$ to $[0, 1]$. The global luminance Y_g is computed as the mean value of $\mathbf{Y}$:

$$Y_g = \frac{1}{HW} \sum_{x=1}^{H} \sum_{y=1}^{W} \mathbf{Y}(x, y). \tag{1}$$

Unlike a fixed threshold, Y_g reflects the relative illumination level of each input image. However, directly comparing pixel intensities with Y_g is sensitive to low-light noise and often produces fragmented masks. To improve spatial consistency, we partition $\mathbf{Y}$ into non-overlapping $k \times k$ regions. The local luminance Y_l is then computed as the regional mean:

$$\mathbf{Y}_l(x, y) = \frac{1}{k^2} \sum_{(u,v)\in\omega(x,y)} \mathbf{Y}(u, v), \tag{2}$$

where $\omega(x, y)$ denotes the $k \times k$ region containing pixel (x, y). This averaging yields a locally smoothed luminance map. The binary mask is then computed as

$$\mathbf{M}(x, y) = \mathbb{I}\left[\mathbf{Y}_l(x, y) > Y_g\right], \tag{3}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function, which equals 1 if the condition is satisfied and 0 otherwise. Pixels with $M = 1$ are assigned to the bright branch, while those with $M = 0$ are assigned to the dark branch. Unless otherwise stated, we set $k = 4$. The resulting mask serves as a spatial prior for branch-specific supervision.

Branch Differentiation via Luminance Masks. The outputs of the two branches are spatially gated to obtain a low-resolution prediction:

$$\mathbf{I}_{\text{out}}^{\text{LR}} = \mathbf{M} \odot \mathbf{I}_b + (1 - \mathbf{M}) \odot \mathbf{I}_d, \tag{4}$$

where $\odot$ denotes element-wise multiplication and the single-channel mask is broadcast to the RGB channels. The LR reconstruction objective incorporates an L_1 loss $\mathcal{L}_1^{\text{LR}}$ and an SSIM loss Wang

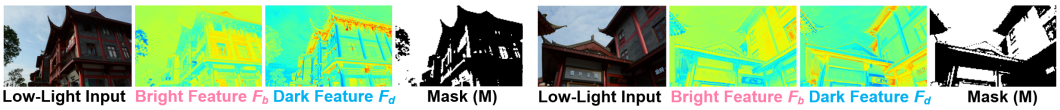


Figure 3: Visualization of branch feature maps under the guidance of the luminance mask. The two branches attend to complementary luminance regions.

et al. (2004) term, with the LR ground truth $\mathbf{I}_{\mathrm{GT}}^{\mathrm{LR}}$ obtained by nearest-neighbor downsampling of the HR reference image:

$$\mathcal{L}_{\mathrm{rec}}^{\mathrm{LR}} = \mathcal{L}_1^{\mathrm{LR}} + \left(1 - \mathrm{SSIM}\left(\mathbf{I}_{\mathrm{out}}^{\mathrm{LR}}, \mathbf{I}_{\mathrm{GT}}^{\mathrm{LR}}\right)\right), \quad \mathcal{L}_1^{\mathrm{LR}} = \left\|\mathbf{I}_{\mathrm{out}}^{\mathrm{LR}} - \mathbf{I}_{\mathrm{GT}}^{\mathrm{LR}}\right\|_1. \tag{5}$$

The binary mask in Eq. (3) is treated as a fixed spatial gate during each forward-backward pass. Accordingly, the gradients propagated to the two branches satisfy

$$\frac{\partial \mathcal{L}_{\mathrm{rec}}^{\mathrm{LR}}}{\partial \mathbf{I}_b} = \mathbf{M} \odot \frac{\partial \mathcal{L}_{\mathrm{rec}}^{\mathrm{LR}}}{\partial \mathbf{I}_{\mathrm{out}}^{\mathrm{LR}}}, \quad \frac{\partial \mathcal{L}_{\mathrm{rec}}^{\mathrm{LR}}}{\partial \mathbf{I}_d} = (1 - \mathbf{M}) \odot \frac{\partial \mathcal{L}_{\mathrm{rec}}^{\mathrm{LR}}}{\partial \mathbf{I}_{\mathrm{out}}^{\mathrm{LR}}}. \tag{6}$$

For the LR reconstruction L_1 loss $\mathcal{L}_1^{\mathrm{LR}}$ in Eq. (5), the gradients are given by

$$\frac{\partial \mathcal{L}_1^{\mathrm{LR}}}{\partial \mathbf{I}_b} = \mathbf{M} \odot \mathrm{sgn}\left(\mathbf{I}_{\mathrm{out}}^{\mathrm{LR}} - \mathbf{I}_{\mathrm{GT}}^{\mathrm{LR}}\right), \quad \frac{\partial \mathcal{L}_1^{\mathrm{LR}}}{\partial \mathbf{I}_d} = (1 - \mathbf{M}) \odot \mathrm{sgn}\left(\mathbf{I}_{\mathrm{out}}^{\mathrm{LR}} - \mathbf{I}_{\mathrm{GT}}^{\mathrm{LR}}\right), \tag{7}$$

where $\mathrm{sgn}(\cdot)$ denotes the element-wise sign function.

Consequently, the mask selectively routes gradients to the two branches. As shown in Fig. 3, the branch feature maps ($\mathbf{F}_{\mathrm{b}}$ and $\mathbf{F}_{\mathrm{d}}$) exhibit complementary responses to well-exposed and under-exposed regions defined by the dynamically updated luminance mask $\mathbf{M}$, validating the proposed branch specialization.

Light-Aware Multimodal Semantic Guidance. Pixel-level reconstruction alone cannot adequately constrain global exposure and perceptual appearance. Therefore, we introduce multimodal semantic guidance using light-aware text prompts for network optimization. Specifically, positive and negative prompts ($\mathbf{T}_+$/$\mathbf{T}_-$) include phrases such as "*a vibrant, well-exposed photo*" and "*a dark, underexposed photo*", respectively. During training, the enhanced LR image $\mathbf{I}_{\mathrm{out}}^{\mathrm{LR}}$ and text prompts are encoded by frozen pretrained contrastive language-image pretraining (CLIP) encoders Radford et al. (2021); Liang et al. (2023) to obtain the image embedding $\mathbf{E}_{\mathbf{I}_{\mathrm{out}}^{\mathrm{LR}}}$ and text embeddings $\mathbf{E}_{\mathrm{T}_\pm}$, respectively. A semantic loss is then defined as

$$\mathcal{L}_{\mathrm{CLIP}} = (s_- - s_+) + \frac{e^{s_-}}{e^{s_-} + e^{s_+}}, \quad s_\pm = s(\mathbf{E}_{\mathbf{I}_{\mathrm{out}}^{\mathrm{LR}}}, \mathbf{E}_{\mathrm{T}_\pm}), \tag{8}$$

where $s(\cdot, \cdot)$ denotes cosine similarity. The first term increases the relative similarity between the enhanced image and the positive prompt, while the second suppresses similarity to the negative prompt, encouraging semantically consistent restoration.

Total Loss. The final HR output $\mathbf{I}_{\mathrm{out}}^{\mathrm{HR}}$ is supervised by the HR reconstruction loss:

$$\mathcal{L}_{\mathrm{rec}}^{\mathrm{HR}} = \left\|\mathbf{I}_{\mathrm{out}}^{\mathrm{HR}} - \mathbf{I}_{\mathrm{GT}}^{\mathrm{HR}}\right\|_1 + \left(1 - \mathrm{SSIM}\left(\mathbf{I}_{\mathrm{out}}^{\mathrm{HR}}, \mathbf{I}_{\mathrm{GT}}^{\mathrm{HR}}\right)\right). \tag{9}$$

The overall training objective is defined as

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{rec}}^{\mathrm{LR}} + \mathcal{L}_{\mathrm{rec}}^{\mathrm{HR}} + \lambda \mathcal{L}_{\mathrm{CLIP}}, \tag{10}$$

where λ is empirically set to 0.1 to balance semantic guidance and reconstruction fidelity.

3.4 Super-Resolution Reconstruction

To efficiently generate enhanced HR images, we introduce an SR reconstruction module at the end of L^4Net. The bright and dark branch features ($\mathbf{F}_b$/$\mathbf{F}_d$) are fused and refined by a stack of Residual Channel Attention Blocks (RCABs) Zhang et al. (2018) to recover high-frequency details. This decoupled design enables the dual branches to focus on illumination enhancement in the LR space, while the SR reconstruction module restores fine structures for the final HR output.

Table 1: Quantitative comparison on LOLv1, LOLv2-Syn, and ACDC-Night. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	LOLv1						LOLv2-Syn		ACDC-Night		Params.
	Original		2×		4×		Original		4×		
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	(M)
SNR Xu et al. (2022)	24.60	0.841	22.24	0.768	19.93	0.694	24.14	0.928	24.87	0.794	39.121
IAT Cui et al. (2022)	23.38	0.808	18.78	0.727	17.25	0.629	20.23	0.870	23.14	0.766	0.09
Restormer Zamir et al. (2022)	22.67	0.784	22.18	0.702	21.32	0.641	22.28	0.902	22.14	0.759	26.127
FLW Zhang et al. (2023)	23.91	0.842	22.89	0.788	21.83	0.695	13.92	0.686	23.06	0.770	0.017
RetinexFormer Cai et al. (2023)	**25.16**	0.845	22.84	0.749	20.67	0.667	25.67	0.930	20.53	0.693	1.606
SHAL-Net Xu et al. (2024)	18.01	0.749	17.48	0.674	16.97	0.605	18.52	0.803	14.64	0.696	0.266
RFLLIE Jiang et al. (2024)	24.84	0.848	23.37	0.792	21.45	0.710	24.62	0.931	19.74	0.773	14.869
IDB Li et al. (2025)	22.96	0.835	22.61	0.796	22.30	0.741	25.20	0.934	27.59	0.866	100.335
URetinex++ Wu et al. (2025b)	23.83	0.841	20.72	0.764	19.92	0.672	11.11	0.578	20.64	0.737	0.372
S^2UNet Liu et al. (2026a)	20.43	0.788	20.01	0.734	19.32	0.642	18.75	0.794	17.11	0.692	0.149
L^4Net (Ours)	24.72	**0.872**	**23.49**	**0.836**	**22.69**	**0.786**	**26.14**	**0.947**	**28.68**	**0.882**	6.918

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Implementation Details. L^4Net is trained for 6000 epochs with a batch size of 16 using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) on a single NVIDIA RTX 3090 GPU. The learning rate is initialized to 1×10^{-4}, kept constant for the first 1500 epochs, and then linearly decayed to zero over the remaining epochs.

Datasets and Metrics. We evaluate L^4Net on LOLv1 Wei et al. (2018), LOLv2-Syn Yang et al. (2021), and the nighttime subset of ACDC (ACDC-Night) Sakaridis et al. (2021), covering both real and synthetic low-light scenarios. LOLv1 (485/15 training/testing pairs) contains paired real low-/normal-light images; LOLv2-Syn (900/100) provides synthetic image pairs with spatially varying illumination; and ACDC-Night (400/500) comprises HR nighttime driving scenes with severe underexposure and non-uniform illumination. For all methods, we evaluate three input scales: the original resolution and 2×/4× downsampled inputs on LOLv1 (600×400); the original resolution on LOLv2-Syn (384×384); and 4× downsampled inputs on ACDC-Night (1920×1080) due to its HR images. Nearest-neighbor interpolation is applied for all downsampling operations. We adopt peak signal-to-noise ratio (PSNR) He et al. (2016) and structural similarity index measure (SSIM) Wang et al. (2004) to evaluate reconstruction fidelity and structural preservation, respectively, where higher values indicate better restoration performance.

4.2 COMPARISONS WITH STATE-OF-THE-ART METHODS

Quantitative Results. As shown in Table 1, L^4Net generally outperforms state-of-the-art methods across different datasets and input scales. All competing methods except IDB are trained at the original resolution and tested without retraining on 2× and 4× downsampled inputs, with outputs restored to the original resolution via nearest-neighbor interpolation for evaluation. IDB is retrained at each scale to adapt its SR module to the corresponding input–output resolution mapping. L^4Net achieves the best results on most evaluation metrics across LOLv1 and LOLv2-Syn at the original resolution and maintains its advantage even under 4× downsampling on LOLv1, where most competing methods deteriorate noticeably. These results demonstrate effective illumination correction and detail recovery under spatially varying lighting conditions. On ACDC-Night, despite using only 1/16 of the original pixels as input (4× downsampling), L^4Net continues to achieve leading performance, demonstrating robust preservation of structural details under severe resolution degradation. Notably, L^4Net outperforms the second-best IDB on all evaluation metrics while using less than 7% of its parameters, highlighting a favorable quality–efficiency trade-off.



Figure 4: Visual comparisons on LOLv1 and LOLv2-Syn, and ACDC-Night. Local regions in red boxes are enlarged below.

Table 2: Quantitative ablation study of L^4Net components.

Component (Params.)	M1	M2	M3	M4	M5	L^4Net	(RCAB Number)		
Dual-Branch (2.68M)	✗	✓	✓	✓	✓	✓	✓	✓	✓
SR Module (4.24M)	✗	✗	✗	✓	✓	✓(5)	✓(0)	✓(1)	✓(10)
Light-Guided Learning	✗	✗	✓	✗	✓	✓	✓	✓	✓
Text Prompt	✗	✗	✗	✗	✗	✓	✓	✓	✓
PSNR	24.19	24.13	23.88	24.23	24.69	**24.72**	22.45	24.64	24.51
SSIM	0.855	0.858	0.858	0.862	0.866	**0.872**	0.860	0.868	0.869

Visual Comparisons. Fig. 4 presents visual comparisons under different input scales, including the original resolution and $2\times/4\times$ downsampling, across three datasets. Competing methods often suffer from color distortions, illumination imbalance, noise amplification, or structural detail loss, especially under severe downsampling. For example, SHAL-Net and S^{2}UNet introduce noticeable color shifts, IDB tends to over-enhance bright regions on LOLv1, while IAT, FLW, and RELLIE provide insufficient enhancement in dark regions with amplified noise. URetinex++ even fails on some LOLv2-Syn images due to extreme illumination variations. Most existing approaches exhibit blurred edges and degraded textures as the input resolution decreases. In contrast, L^4Net better preserves fine structures, achieves faithful color reproduction, and provides balanced enhancement with fewer artifacts across regions with different illumination levels. These results demonstrate the robustness of L^4Net to spatially varying illumination and resolution degradation.

4.3 Ablation Study

Ablation of Network Components. We evaluate five model variants on the LOLv1 dataset to assess the contribution of each component in L^4Net, as reported in Table 2. M1 denotes the single-branch L^4 backbone. M2 introduces the illumination-aware dual-branch architecture, where the features from the two branches are concatenated and mapped to the output through a 1×1 convolution. The dual-branch design enhances structural preservation under spatially non-uniform illumination. M3 and M4 incorporate light-guided learning and the SR reconstruction module, respectively. M3 preserves structural similarity despite a slight decrease in PSNR, indicating that light-guided learning mainly improves illumination-aware restoration rather than pixel-wise fidelity. M4 improves

Table 3: Effect of number of LA-block at three feature scales ($H \times W \times C$, $\frac{H}{2} \times \frac{W}{2} \times 2C$, and $\frac{H}{4} \times \frac{W}{4} \times 4C$) in the encoder of L^4 backbone. The adopted configuration is $[1, 2, 2]$.

LA Blocks	$[0, 0, 0]$	$[1, 1, 1]$	$[2, 1, 1]$	$[1, 2, 1]$	$[1, 1, 2]$	$[\mathbf{1}, \mathbf{2}, \mathbf{2}]$	$[1, 2, 3]$
PSNR	20.14	23.95	24.09	24.05	24.06	24.19	24.20
SSIM	0.789	0.848	0.852	0.858	0.853	0.855	0.857
FLOPs (G)	115	314	396	393	352	430	468
Params (M)	0.34	0.86	0.90	1.02	1.18	1.34	1.66

fine-detail recovery, and combining both modules in M5 yields further improvements, indicating their complementary roles. Finally, incorporating light-aware multimodal semantic guidance further improves perceptual consistency and semantic fidelity, leading to the best overall performance.

Effect of the Number of RCABs. The last four columns of Table 2 investigate the effect of the number of Residual Channel Attention Blocks (RCABs) Zhang et al. (2018) in the SR reconstruction module. We evaluate four configurations with 0, 1, 5, and 10 RCABs, while keeping all other settings unchanged from those of L^4Net. When no RCAB is used, only convolutional layers and PixelShuffle are retained for HR reconstruction. Removing all RCABs causes a noticeable performance degradation, indicating that convolutional upsampling alone is insufficient to recover high-frequency details lost during downsampling. Introducing RCABs progressively improves restoration quality by refining the fused bright and dark branch features. However, increasing the number from five to ten provides no further gain and slightly degrades the performance while increasing model complexity, suggesting redundant feature refinement and diminishing returns. Considering the quality–efficiency trade-off, we adopt five RCABs as the default configuration of the SR reconstruction module.

Effect of LA-Block Allocation in L^4 Backbone. Table 3 investigates the allocation of LA blocks in the encoder of L^4Net across three feature scales, i.e., $H \times W \times C$, $\frac{H}{2} \times \frac{W}{2} \times 2C$, and $\frac{H}{4} \times \frac{W}{4} \times 4C$. The decoder follows a symmetric LA-block configuration. Removing all the proposed LA blocks and retaining only convolutional layers causes a significant degradation in restoration quality, demonstrating the effectiveness of global–local feature interaction introduced by LA. Introducing LA blocks at all three scales substantially improves the results, while allocating additional blocks to different scales provides further but scale-dependent gains. This observation indicates that LA blocks at different feature scales capture complementary characteristics rather than providing redundant patterns. Distributing additional blocks over the middle- and low-resolution scales achieves a favorable balance between illumination modeling and structural detail preservation. In contrast, further deepening the lowest-resolution stage produces only marginal improvement while introducing considerable additional computational overhead, revealing diminishing returns. Based on this quality–efficiency trade-off, we adopt the $[1, 2, 2]$ LA-block allocation as the default settings for the L^4 backbone.

Visual Analysis of Key Modules. Fig. 5 presents a qualitative comparison of the light-guided learning strategy and the SR reconstruction module. The low-light LR input (4× downsampling) suffers from spatially varying underexposure, which obscures local structures. Without the light-guided learning strategy, the two branches fail to learn sufficiently distinct restoration behaviors, leading to imbalanced enhancement and local overexposure. Removing the SR module allows the LR output ($I_{\text{out}}^{\text{LR}}$) of L^4Net to recover the overall illumination but results in blurred object boundaries and indistinct textual structures due to the irreversible loss of high-frequency information during downsampling. In contrast, the complete L^4Net combines region-specific gradient routing with complementary feature reconstruction to generate HR output($I_{\text{out}}^{\text{HR}}$) with more balanced exposure and clearer structural details. Overall, light-guided learning enables spatially adaptive exposure control, while the SR module reconstructs fine details.

4.4 Efficiency Evaluation and Embedded Deployment

Computational Efficiency. To evaluate the efficiency of L^4Net for HR LLIE, we compare it with representative Transformer-based methods under the same output resolution and hardware config-



Figure 5: Qualitative ablation analysis of light-guided learning and SR reconstruction in L^4Net. Removing light-guided learning leads to exposure imbalance, while removing the SR module results in blurred high-frequency structures.

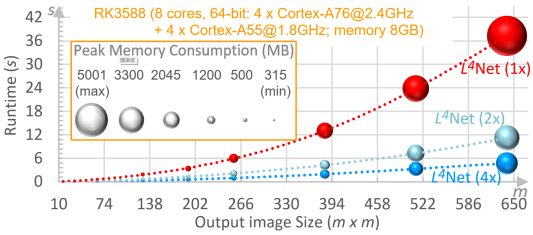


Figure 6: Efficiency evaluation on Orange Pi 5.

Table 4: Efficiency comparison of Transformer-based models and GPU memory reduction for 1600×1200 output images.

Model	Params (M)	FLOPs (G)	Time (s)	L^4Net Model	Input Image	GPU (G)	Save (%)
Restormer	26.13	4131	7.67	1×	1600 × 1200	5.95	0.0
RFLLIE	14.87	4467	1.25	2×	800 × 600	2.70	54.6
SNR	39.12	817	1.21	4×	400 × 300	2.13	64.2
Ours	**6.92**	**567**	**0.29**	Ours	**400×300**	**2.13**	**64.2**

uration. Since existing methods directly process HR inputs, they incur substantial computational overhead, whereas L^4Net performs enhancement in a compact feature space. As shown in the left half of Table 4, L^4Net reduces computational cost by 86.3%, 87.3%, and 30.6% and inference time by 96.2%, 76.8%, and 76.0% compared with Restormer, RFLLIE, and SNR, respectively. These results validate the effectiveness of the proposed LR-space enhancement strategy for HR LLIE. The right side of Table 4 further examines peak GPU memory consumption of L^4Net under different input scales. In the full-resolution setting, branch-wise enhancement and feature fusion are performed at the target resolution. In contrast, the proposed configuration conducts these operations in the LR domain and postpones spatial upsampling to the final SR reconstruction stage. For example, when generating a 1600×1200 enhanced image, L^4Net equipped with $4\times$ SR performs most operations on 400×300 LR features, reducing GPU memory consumption by 64.2%. This design effectively decreases the spatial memory footprint of branch outputs and fused features throughout most of the forward pass, thereby substantially alleviating the memory burden of HR enhancement.

Deployment on Embedded Platform. To evaluate deployment feasibility of L^4Net on edge devices, we conduct inference experiments on an Orange Pi 5 equipped with an RK3588 processor. Figure 6 reports the inference time and peak memory usage at different output resolutions. Both runtime and memory consumption increase with output resolution, making direct HR processing ($1\times$) increasingly costly. Under the same target resolution, the $2\times$ and $4\times$ settings reduce computational demands by operating on lower-resolution inputs, with the $4\times$ setting achieving the largest savings. Combined with the quantitative restoration results, these findings indicate that L^4Net achieves an effective balance between efficiency and restoration quality, highlighting its suitability for real-world low-light vision applications on resource-constrained devices.

5 Conclusion

In this paper, we presented L^4Net, a lightweight framework that addresses spatially non-uniform illumination through light-guided regional specialization and computational overhead in high-resolution LLIE through LR-space enhancement with HR reconstruction. Comprehensive experiments and ablation studies demonstrate that L^4Net achieves effective illumination correction and detail recovery with a favorable quality–efficiency trade-off, while edge-device evaluation verifies its practical applicability. A remaining limitation is that multimodal semantic guidance may occasionally introduce a discrepancy between perceptual enhancement and structural preservation. Future work will explore more adaptive strategies for incorporating semantic priors in real-world LLIE.

AI USE STATEMENT

We used generative AI tools for grammar correction, proofreading, and LaTeX formatting assistance during manuscript preparation. The research ideas, methodological motivation, proposed approach, and scientific contributions were independently developed by the authors. Generative AI tools were not used to formulate these elements or design the experiments. They were also not involved in the generation or modification of references cite, experimental data, results, and the scientific content of figures and tables. All analyses and interpretations of the experimental findings were conducted by the authors. The authors reviewed and verified all AI-assisted edits and take full responsibility for the accuracy, originality, and integrity of the final manuscript.

REFERENCES

Raul Balmez, Alexandru Brateanu, Ciprian Orhei, Codruta O. Ancuti, and Cosmin Ancuti. Isalux: Illumination and semantics-aware transformer employing mixture of experts for low light image enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7862–7872, March 2026.

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*, pp. 12504–12513, 2023.

Yang Cheng, Ziteng Cui, Lin Gu, Shenghan Su, and Zenghui Zhang. Perception-inspired color space design for photo white balance editing. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3741–3749. IEEE, 2026.

Joshua Cho, Sara Aghajanzadeh, Zhen Zhu, and David Forsyth. Mr. illuminate: Zero-shot low-light image enhancement with diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8460–8470, June 2026.

Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, Zhengkai Jiang, Yu Qiao, and Tatsuya Harada. You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In *Proceedings of the 33rd British Machine Vision Conference (BMVC)*, pp. 238. BMVA Press, 2022.

Guodong Fan, Zishu Yao, Guang-Yong Chen, Jian-Nan Su, and Min Gan. Iniretinex: Rethinking retinex-type low-light image enhancer via initialization perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2834–2842, 2025.

Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9716–9725, 2021.

Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10879–10889, 2025.

Huiyuan Fu, Wenkai Zheng, Xiangyu Meng, Xin Wang, Chuanming Wang, and Huadong Ma. You do not need additional priors or regularizers in retinex-based low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18125–18134, June 2023a.

Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22252–22261, June 2023b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yan Huang, Xiaoshan Liao, Jinxiu Liang, Yuhui Quan, Boxin Shi, and Yong Xu. Zero-shot low-light image enhancement via latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3815–3823, 2025.

Hai Jiang, Yang Ren, and Songchen Han. Revisiting coarse-to-fine strategy for low-light image enhancement with deep decomposition guided training. *Computer Vision and Image Understanding*, 241:103952, 2024.

Shuai Jin, Yuhua Qian, Feijiang Li, Guoqing Liu, and Xinyan Liang. Pasd: A pixel-adaptive swarm dynamics approach for unsupervised low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9070–9079, October 2025.

Zhan Li, Wenqing Kuang, Bir Bhanu, Yifan Deng, Yihang Chen, and Kangkang Xu. Low-visibility scene enhancement by isomorphic dual-branch framework with attention learning. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8094–8103, 2023.

Minwen Liao, Haobo Dong, Xinyi Wang, Kurban Ubul, Yihua Shao, and Ziyang Yan. Gm-moe: Low-light enhancement with gated-mechanism mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8766–8776, October 2025.

Jia Liu, Yu Luo, Guanghui Yue, Jie Ling, Liang Liao, Chia-Wen Lin, Guangtao Zhai, and Wei Zhou. Self-supervised unfolding network with shared reflectance learning for low-light image enhancement. *IEEE Transactions on Image Processing*, 35:800–815, 2026a.

Ruizhou Liu, Zhe Wu, Zimo Liu, Qingfang Zheng, and Qingming Huang. Dmgine: Day-memory guided nighttime image enhancement for dynamic traffic scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7251–7259, 2026b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10765–10775, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 7459–7469, 2023a.

Cong Wang, Jinshan Pan, Wei Wang, Gang Fu, Siyuan Liang, Mengzhu Wang, Xiao-Ming Wu, and Jun Liu. Correlation matching transformation transformers for uhd image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5336–5344, 2024.

Yinglong Wang, Zhen Liu, Jianzhuang Liu, Songcen Xu, and Shuaicheng Liu. Low-light image enhancement with illumination-aware gamma correction and complete image modelling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13128–13137, 2023b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018.

Chen Wu, Ling Wang, Zhuoran Zheng, Weidong Jiang, Yuning Cui, and Jingyuan Xia. Ultra-high-definition image restoration via high-frequency enhanced transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.

Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Interpretable optimization-inspired unfolding network for low-light image enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2545–2562, 2025b.

Heming Xu, Xintong Liu, Hanwen Zhang, Xiaohe Wu, and Wangmeng Zuo. Degraded structure and hue guided auxiliary learning for low-light image enhancement. *Knowledge-Based Systems*, 295:111779, 2024.

Senyan Xu, Zhijing Sun, Kean Liu, Xin Lu, Ruixuan Jiang, Xueyang Fu, and Zheng-Jun Zha. Event-illumination collaborative low-light image enhancement with a high-resolution real-world dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22270–22280, June 2026.

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*, pp. 17714–17724, June 2022.

Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. In *Proceedings of the Computer Vision and Pattern Recognition conference*, pp. 5678–5687, 2025.

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.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12302–12311, October 2023.

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*, pp. 5728–5739, 2022.

Yu Zhang, Xiaoguang Di, Junde Wu, Rao Fu, Yong Li, Yue Wang, Yanwu Xu, Guohui Yang, and Chunhui Wang. Simplifying low-light image enhancement networks with relative loss functions. *arXiv preprint arXiv:2304.02978*, 2023. doi: 10.48550/arXiv.2304.02978. URL `https://arxiv.org/abs/2304.02978`.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018.

A Appendix

A.1 Training Strategy

During training, the illumination-aware dual branches and the super-resolution (SR) reconstruction module are jointly optimized in an end-to-end manner. The bright and dark branches learn complementary restoration behaviors through a light-guided learning strategy. Regions with different illumination levels are distinguished by an adaptive luminance mask $\mathbf{M}$, which is dynamically updated based on the output image $\mathbf{I}_b$ of the bright branch during training. The branch outputs $\mathbf{I}_b$ and $\mathbf{I}_d$ are spatially combined using $\mathbf{M}$ to produce the enhanced low-resolution (LR) image $\mathbf{I}_{\text{out}}^{\text{LR}}$. This LR output is used to compute the LR reconstruction loss, while the CLIP loss is computed using the LR output together with light-aware text prompts $\mathbf{T}_+$ and $\mathbf{T}_-$. The SR reconstruction module recovers high-frequency details to generate the final enhanced high-resolution (HR) image $\mathbf{I}_{\text{out}}^{\text{HR}}$ by fusing the dual-branch features $\mathbf{F}_b$ and $\mathbf{F}_d$. This joint optimization enables LR-space illumination

Algorithm 1 Training Procedure of the Proposed L^4Net

Input: High-resolution low-light images $\mathbf{I}_{\text{in}}^{\text{HR}}$;
High-resolution normal-light images $\mathbf{I}_{\text{GT}}^{\text{HR}}$;
Positive text prompt $\mathbf{T}_+$; Negative text prompt $\mathbf{T}_-$
Output: Trained L^4Net

for each $\mathbf{I}_{\text{in}}^{\text{HR}}$ in training set **do**
  Downsample $\mathbf{I}_{\text{in}}^{\text{HR}}$ to obtain low-resolution image $\mathbf{I}_{\text{in}}^{\text{LR}}$
  Downsample $\mathbf{I}_{\text{GT}}^{\text{HR}}$ to obtain low-resolution image $\mathbf{I}_{\text{GT}}^{\text{LR}}$
  Feed $\mathbf{I}_{\text{in}}^{\text{LR}}$ into two L^4 branches to obtain bright/dark features $\mathbf{F}_b$/$\mathbf{F}_d$ and images $\mathbf{I}_b$/$\mathbf{I}_d$:
    $(\mathbf{F}_b, \mathbf{I}_b) \leftarrow$ Bright L^4 Branch $(\mathbf{I}_{\text{in}}^{\text{LR}})$
    $(\mathbf{F}_d, \mathbf{I}_d) \leftarrow$ Dark L^4 Branch $(\mathbf{I}_{\text{in}}^{\text{LR}})$
  Compute luminance mask $\mathbf{M}$ from $\mathbf{I}_b$ by Eq. (3)
  Obtain enhanced LR image $\mathbf{I}_{\text{out}}^{\text{LR}}$ from $\mathbf{I}_b$/ $\mathbf{I}_d$ by Eq. (4)
  Compute LR reconstruction loss $\mathcal{L}_{\text{rec}}^{\text{LR}}$ by Eq. (5)
  Compute loss $\mathcal{L}_{\text{CLIP}}$ from $\mathbf{I}_{\text{out}}^{\text{LR}}$, $\mathbf{T}_+$, $\mathbf{T}_-$ by Eq. (8)
  Feed $\mathbf{F}_b$ and $\mathbf{F}_d$ into SR module to obtain HR output:
    $\mathbf{I}_{\text{out}}^{\text{HR}} \leftarrow$ SR Module $(\mathbf{F}_b, \mathbf{F}_d)$
  Compute HR reconstruction loss $\mathcal{L}_{\text{rec}}^{\text{HR}}$ by Eq. (9)
  Compute total loss $\mathcal{L}_{\text{total}}$ by Eq. (10)
end for
Optimize the network by minimizing total loss $\mathcal{L}_{\text{total}}$
return trained L^4Net.

Table 5: Ablation studies on learnable gamma correction and attention mechanisms in LA modules. "w/" ("w/o") denotes with (without).

Setting	Variant	PSNR	SSIM	FLOPs	Params.
Gamma Correction	w/o	24.05	0.838	429.70G	1.34M
	w/	**24.19**	**0.855**	429.81G	1.34M
SA+MLP LA Block	SSA	21.56	0.838	**300.88G**	6.86M
	CSA	**23.95**	**0.848**	313.65G	**0.86M**

enhancement and HR detail reconstruction to be learned collaboratively, achieving better exposure recovery while preserving structural fidelity. The complete training procedure of the proposed L^4Net is summarized in Algorithm 1, where equations follow the numbering in the main paper.

A.2 Analysis of L^4 Backbone Design

To isolate the contribution of the L^4 backbone from the dual-branch architecture and the complete training objective, all experiments in this section are conducted using a single-branch network trained solely with the reconstruction loss. We investigate two key design choices of the backbone: the learnable gamma-correction block and the Light-Attention (LA) blocks. To avoid potential effects of downsampling, all ablation experiments are conducted on LOLv1 at its original resolution.

Gamma pre-enhancement. Low-light images commonly exhibit compressed intensity distributions and spatially non-uniform illumination, making it difficult for the restoration backbone to effectively extract informative features from the input. To alleviate this issue, the L^4 backbone employs a lightweight learnable gamma-correction block before the U-shaped restoration network. The block estimates the gamma coefficient and performs input pre-enhancement, providing illumination-adaptive conditioning for subsequent feature restoration. As shown in the first two rows of Table 5, learnable gamma correction consistently improves restoration performance while introducing negligible additional computational overhead. These results demonstrate that illumination-adaptive pre-enhancement provides a more favorable input distribution and aids subsequent feature restoration.

Choice of attention mechanism. We further compare spatial self-attention (SSA) Vaswani et al. (2017) and channel self-attention (CSA) by replacing LA blocks in the L^4 backbone with a standard

Table 6: Ablation study on learning strategy.

Learning Strategy	Applied on	PSNR↑	SSIM↑
Without guidance (M3)	-	24.23	0.862
Difficulty-guided (IDB)	Pixel-wise	24.10	0.862
Light-guided (Ours)	Patch-wise	**24.72**	**0.872**

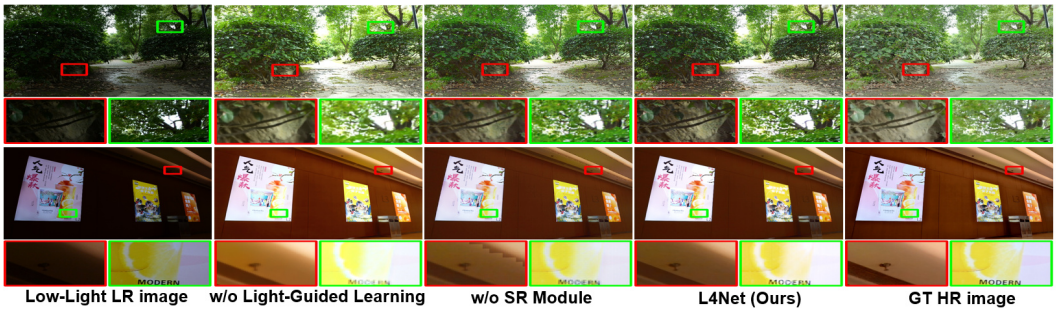


Figure 7: Qualitative ablation of light-guided learning and SR reconstruction in L^4Net. "w/o" denotes without.

attention module composed of SSA and a multilayer perceptron (MLP). Directly applying SSA to HR feature maps requires constructing an $HW \times HW$ attention map, resulting in substantial memory consumption and computational overhead. Therefore, the SSA-based variant adopts a $[0, 0, 2]$ configuration across the three feature scales, with two attention blocks placed only at the lowest resolution ($\frac{H}{4} \times \frac{W}{4} \times 4C$). In contrast, the CSA-based variant employs the proposed LA blocks with a $[1, 1, 1]$ configuration, where the CSA operation constructs a $C \times C$ attention map and enables efficient channel-wise dependency modeling across all feature scales. These configurations provide comparable computational and memory costs between the two variants for a fair comparison. The last two rows in Table 5 show that the CSA-based variant improves PSNR by 2.39 dB while using only 12.5% of the parameters of the SSA-based variant, demonstrating LA blocks achieve a better quality-efficiency trade-off for the L^4 backbone.

A.3 Ablation of the Light-Guided Learning

The effectiveness of the bright and dark branches, each built upon a structurally identical L^4 backbone with unshared parameters, depends on their ability to learn complementary restoration behaviors. To investigate the proposed light-guided learning strategy, we compare three branch specialization schemes in Table 6 on LOLv1. Without branch-specific guidance, the two branches are optimized with the same reconstruction objective and may converge to similar representations. Following IDB Li et al. (2025), we construct a pixel-wise difficulty map from the difference between the low-light input and its reference image, and route hard- and easy-to-restore pixels to different branches. However, this illumination-agnostic map is sensitive to noise, local misalignment, and high-frequency variations, often producing fragmented regions that encourage pixel-level fitting rather than spatially coherent illumination restoration.

In contrast, the proposed strategy employs an adaptive luminance mask to supervise the bright and dark branches. Since illumination typically varies smoothly within local regions, patch-level mask estimation provides spatially coherent guidance: the bright branch focuses on preserving reliable structures in well-exposed regions, whereas the dark branch specializes in recovering weak signals from severely underexposed regions. This illumination-aware specialization better aligns with spatially non-uniform illumination and promotes complementary restoration behaviors, while patch-wise estimation reduces sensitivity to noisy pixels.

Figure 7 further shows that the proposed light-guided learning strategy suppresses over-enhancement in relatively well-exposed regions while improving detail recovery in severely underexposed areas. Meanwhile, the SR reconstruction module restores high-frequency details lost during downsam-

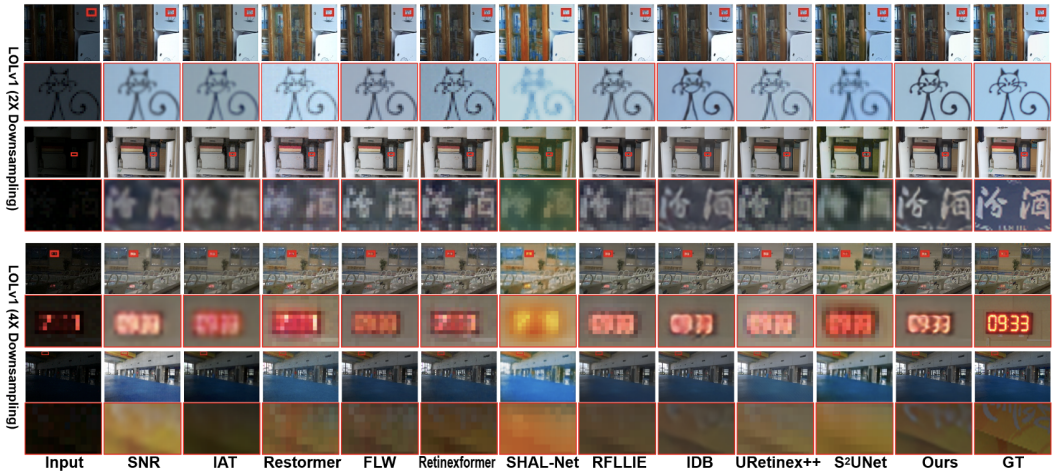


Figure 8: Visual comparisons on LOLv1 under $2\times$ and $4\times$ downsampling. Local regions in red boxes are enlarged below.

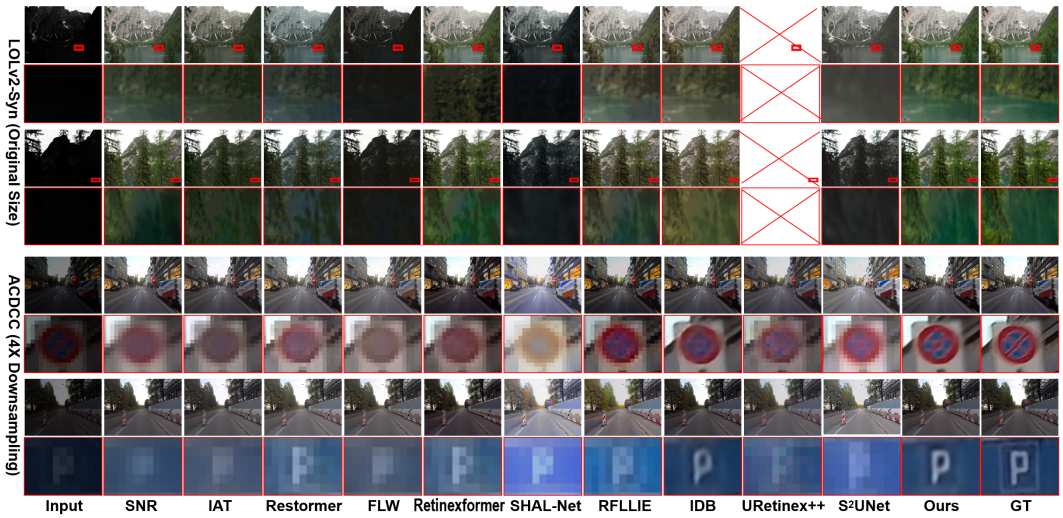


Figure 9: Visual comparisons on LOLv2-Syn and ACDC-Night. Local regions in red boxes are enlarged below.

pling, producing sharper boundaries and clearer textual structures. Consequently, the complete L^4Net achieves more balanced exposure and finer structural details.

A.4 More Visual Results

Additional results are presented in Figs. 8 and 9, where we compare L^4Net with several state-of-the-art methods, including SNR Xu et al. (2022), IAT Cui et al. (2022), Restormer Zamir et al. (2022), FLW Zhang et al. (2023), Retinexformer Cai et al. (2023), SHAL-Net Xu et al. (2024), RFLLIE Jiang et al. (2024), IDB Li et al. (2025), URetinex++ Wu et al. (2025b), S^{2}UNet Liu et al. (2026a). These comparisons provide further visual examples of the enhancement results produced by the different methods.

Fig. 8 presents visual comparisons on LOLv1 with $2\times$ and $4\times$ downsampling. SHAL-Net and S^{2}UNet introduce noticeable color distortions, while SNR, Restormer, and URetinex++ suffer from blurred edges and textures. IAT, Retinexformer, RFLLIE, and IDB provide insufficient illumination

enhancement in severely underexposed regions. In contrast, L^4Net achieves more balanced illumination adjustment and faithful color reproduction while preserving text details and fine structures.

Fig. 9 presents visual comparisons on LOLv2-Syn at the original resolution and ACDC-Night with $4\times$ downsampling. On LOLv2-Syn, competing methods commonly exhibit color shifts caused by over- or under-enhancement. URetinex-Net++ completely fails under extreme illumination variations, producing an almost entirely white output, while RFLLIE and IDB produce noticeable yellowish tones. FLW, SHAL-Net, and S^2UNet tend to generate overly dark outputs. In contrast, L^4Net produces more faithful colors and sharper edges, benefiting from light-aware multimodal semantic guidance and HR detail reconstruction. On ACDC-Night, existing methods often generate block artifacts and blurred boundaries when recovering geometric structures from severely downsampled low-light inputs. In contrast, L^4Net restores clearer sign contours and internal symbols with sharper edges. Although IDB also incorporates an SR module, its reconstructed boundaries remain blurred and unstable. These results demonstrate the effectiveness of jointly learning LR illumination enhancement and HR detail reconstruction rather than relying on SR reconstruction alone.

A.5 Application to Semantic Segmentation

Low-light conditions reduce scene visibility and degrade the performance of vision systems in autonomous driving and surveillance scenarios. To evaluate the effectiveness of L^4Net for downstream perception, we employ the STDC-Seg model Fan et al. (2021) with the same settings to perform semantic segmentation on ACDC-Night images enhanced by various low-light enhancement methods.

Figure 10 presents the semantic segmentation results on representative nighttime driving scenes. Competing methods produce inaccurate predictions and incomplete semantic regions due to insufficient enhancement, color distortion, or blurred details. In contrast, L^4Net achieves more accurate segmentation results with more complete road regions and clearer object boundaries, particularly in severely underexposed regions. This improvement can be attributed to the enhanced images produced by L^4Net, which provide more balanced illumination and richer structural information for downstream segmentation.



Figure 10: Visual comparison of semantic segmentation results on ACDC-Night. Red-boxed regions are enlarged on the right. GT denotes ground truth.