IA2former: Illumination-Aware Attention-based Transformer for Low-light Image Enhancement

Tianqi Jiang[1,2,*] [0009-0003-3976-3101], Danqing Ju[1,2,*][0009-0000-3536-2253],

Han Wu[1,2] [0009-0003-9912-714X], and Ping Liang[1,2,†]

1 School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
2 Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China

lpnjh@wust.edu.cn

* *Tianqi Jiang and Danqing Ju contributed equally; † Corresponding author: Ping Liang.*

Abstract. Enhancing images acquired under weak illumination is challenging because exposure correction, noise control, color fidelity, and detail preservation must be balanced at the same time. Classical Retinex algorithms describe an image through illumination and reflectance, but their hand-crafted assumptions are often fragile in extremely dark scenes. Recent Transformer restorers, including RetinexFormer and Restormer, improve long-range dependency modeling; however, the interaction between illumination variation and spatial-semantic content is still not sufficiently exploited. This paper presents IA2former, an Illumination-Aware Attention-based Transformer for low-light image enhancement. The model organizes illumination-guided tokens and reflectance-related content features through a two-stage attention interaction, and introduces an illumination-aware loss to regularize feature changes in illumination-sensitive regions. Experiments on LOL-v1 and LOL-v2 show that IA2former achieves competitive or superior PSNR, SSIM, and LPIPS compared with representative enhancement methods. Visual comparisons and error-map analysis further indicate improved color consistency, noise suppression, and structure preservation.

Keywords: Low-light Enhancement , Retinex , Transformer.

1 Introduction

Low-light image enhancement aims to convert poorly illuminated images into visually clear and task-friendly results. Images captured in dim indoor scenes or complex night environments usually contain weak contrast, blurred textures, color bias, and amplified noise, which not only affect human viewing but also degrade downstream vision tasks such as detection and recognition. A practical enhancement method should therefore brighten under-exposed regions while avoiding over-smoothing, color distortion, and

artificial artifacts. Fig. 1 illustrates typical low-light degradation and the corresponding improvement after enhancement in indoor and outdoor scenes.

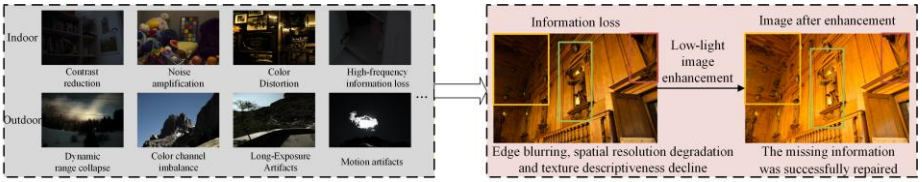

Fig. 1. Examples of low-light effects in two typical scenarios: indoors under low-light environments and outdoors under dynamic lighting conditions. The figure also compares the visual effects before and after low-light image enhancement.

Earlier enhancement methods, including histogram equalization [1,2,3] and gamma correction [4,5,6], mainly remap pixel intensities to make dark areas brighter. Although these operations are simple and efficient, they do not explicitly model scene illumination and may produce unnatural colors or visible artifacts. Retinex-based approaches [7,8,9,10] introduce an illumination-reflectance view and provide a more interpretable framework, but they remain sensitive to noise and imperfect decomposition. Deep models such as the Retinex-CNN method of Wei et al. [11] improve representation learning, while generative models [12,13] and flow-based models [14] further enrich the solution space. Nevertheless, robustly modeling global context, illumination variation, and fine detail under severe darkness remains difficult.

Convolutional networks are effective at extracting local patterns, yet their limited receptive field makes it difficult to capture non-local similarity and scene-level illumination structure. Transformer-based restoration networks address this limitation by using attention to relate distant regions, and models such as SNR-Net [15] and Restormer [16] have demonstrated the value of combining convolutional features with Transformer blocks. However, general restoration backbones usually treat illumination only implicitly. RetinexFormer [17] and RetinexMamba [18] make stronger use of Retinex-inspired priors, but their illumination cues are still mainly injected as auxiliary modulation rather than being used to build explicit feature interactions. This motivates an architecture that places illumination-aware modeling inside the attention process itself.

To address these issues, we develop IA2former, an Illumination-Aware Attention-based Transformer designed for low-light restoration. Instead of using illumination merely as an element-wise guidance map, IA2former learns interactions between illumination variation and spatial-semantic content. Its IA2-MSA module first forms illumination-aware token representations and then relates them to reflectance-related content features through a second cross-attention stage. This design allows the network to recover brightness while retaining texture, color, and structural information in complex illumination conditions.

We also introduce an Illumination-Aware Loss as an auxiliary regularizer. Rather than constraining the output image to remain close to the low-light input, the loss penalizes illumination-weighted feature drift in regions where aggressive brightening is likely to cause color shift, noise amplification, or local artifacts. In this way, the

reconstruction objective remains responsible for brightness recovery, while the additional regularization encourages more stable chromatic and structural changes.

The effectiveness of IA2former is evaluated against representative low-light enhancement methods, including EnlightenGAN [13], KinD [19], Restormer [16], SNR-Net [15], LLFlow [14], RetinexMamba [18], and RetinexFormer [17]. On LOL-v1 and LOL-v2, IA2former provides a strong balance among reconstruction fidelity, structural similarity, and perceptual quality. It obtains the best PSNR and SSIM on LOL-v1, the highest PSNR on LOL-v2-Synthetic, and the best SSIM and LPIPS on LOL-v2-Real. For instance, the model reaches 25.89 dB PSNR and 0.878 SSIM on LOL-v1, and 26.71 dB PSNR on LOL-v2-Synthetic. These results indicate that illumination-aware feature interaction can improve both quantitative scores and visual restoration quality.

Our contributions can be summarized as follows:

IA2-MSA: We propose the Illumination-Aware Attention-based Multi-Head Self-Attention (IA2-MSA) mechanism, which enhances the low-light image by explicitly modeling the interaction between illumination variations and spatial-semantic features. Unlike traditional methods, our mechanism does not modify the reflectance-related content representation directly but rather establishes a relationship between the reflectance-related content representation and the illumination-guided low-light image for more effective image enhancement.

IA Loss: We introduce an Illumination-Aware Loss function as an auxiliary regularization term to constrain illumination-inconsistent feature drift during enhancement. Instead of forcing the enhanced image to remain unchanged, this loss preserves illumination-weighted structural and chromatic consistency, thereby reducing color shift, noise amplification, and local artifacts.

Experimental Validation: We conduct comprehensive experiments comparing IA2former with state-of-the-art models, demonstrating its favorable overall performance across reconstruction fidelity, structural similarity, and perceptual quality on several low-light datasets.

2 Related work

2.1 Low-light Image Enhancement

Existing low-light enhancement studies can be roughly grouped into distribution remapping methods, model-driven Retinex methods, and learning-based methods.

Distribution Mapping Methods. Histogram equalization and gamma correction enhance dark images by reshaping the intensity distribution of the input. Their low computational cost makes them easy to deploy, but the absence of semantic and illumination awareness often causes local over-enhancement, color inconsistency, and visible artifacts [1,2,3,4,5,6].

Traditional Model Methods. Retinex-based methods attempt to explain the observed image through illumination and reflectance, which gives them a clearer physical motivation than direct intensity remapping. Early work by Horn, Blake, and related studies established important Retinex formulations [7,8,9,10]. However, hand-crafted assumptions about illumination smoothness, noise, and degradation are easily violated in real

low-light images. Later Retinex variants improved illumination estimation and decomposition optimization [20,21], but robust restoration in noisy and spatially uneven lighting remains difficult.

Deep Learning Methods. Learning-based approaches replace hand-crafted priors with data-driven representations. Wei et al. [11] combined Retinex decomposition with CNNs, while later multi-level feature fusion methods [22,23] improved detail recovery and noise suppression. GAN- and VAE-based models [12,13] introduce generative priors for complex lighting, diffusion models [24] recover details through iterative denoising, and LLFlow [14] models the mapping between low- and normal-light images with normalizing flows. Recent lightweight and prompt-based restoration networks [25,26,27] further improve efficiency and degradation awareness. Even so, many methods still struggle to coordinate illumination correction with structure preservation when scenes contain strong darkness, colored light, or non-uniform exposure.

2.2 Vision Transformer

Vision Transformers have become a powerful tool for both semantic vision tasks and low-level restoration because attention can connect distant image regions. In low-light enhancement, SNR-Net [15] combines convolutional layers with Transformer components to handle varying illumination, whereas Restormer [16] offers a general restoration backbone with strong global modeling ability. RetinexFormer [17] introduces Retinex-inspired decomposition into a Transformer framework, and RetinexMamba [18] explores state-space modeling for low-light restoration. These studies show clear progress, but illumination is still often used as side information rather than as a feature stream that actively interacts with content features. Beyond image enhancement, deep learning models have also been applied to intelligent prediction and visual inspection tasks, including PCB defect detection, robotic grasping, and wire defect detection and segmentation [29-32].

Although these methods demonstrate the effectiveness of combining Retinex decomposition with advanced backbone architectures, illumination cues are not explicitly involved in feature interaction modeling, leaving the relationship between illumination variations and spatial–semantic structures insufficiently explored and motivating more principled illumination-aware feature modeling strategies.

3 Method

Fig. 2 illustrates the overall architecture of the proposed low-light image enhancement framework. Given a low-light input image I_l, the model first extracts an illumination map through a lightweight illumination estimation module $\mathcal{E}(\cdot)$:

$$L_c = \text{clamp}(\mathcal{E}(I_l), 0, 255), \tag{1}$$

where $\mathcal{E}(\cdot)$ comprises two 3 × 3 convolutional layers with ReLU activation, and clamp(·, 0, 255) ensures that Lc shares the same value range as Il for subsequent

channel-wise concatenation. The illumination map is then concatenated with the original input to form the illumination-guided feature:

$$F_c = [L_c, I_l], \tag{2}$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation. F_c encapsulates both the spatial brightness distribution and the raw image content, and serves as the primary input to the proposed IA2-MSA module. In parallel, a reflectance-related feature is extracted from F_c through a shallow content extraction module $\mathcal{R}(\cdot)$:

$$F_R = \mathcal{R}(F_c), \tag{3}$$

where $\mathcal{R}(\cdot)$ consists of a 1 × 1 convolution for channel alignment followed by a 3 × 3 convolution for spatial feature extraction. Note that F_R is not intended as a strict reflectance decomposition in the Retinex sense; rather, $\mathcal{R}(\cdot)$ learns to extract content and structural features that are complementary to the illumination information encoded in L_c. The subsequent two-stage attention mechanism further disentangles the roles of illumination and content, progressively separating illumination-dominated representations from reflectance-related structures.

Accordingly, the term reflectance-related feature is used throughout this paper to denote a learned content representation inspired by Retinex separation, rather than a physically interpretable reflectance map obtained by explicit decomposition.

The network follows a shallow-to-deep illumination modeling strategy. In early layers, the dominant degradation is usually insufficient or uneven exposure, so illumination cues provide strong guidance. In deeper layers, however, brightness information must be coordinated with object boundaries, textures, and semantic layout. IA2former therefore uses two stages: the first stage builds illumination-aware tokens from fused shallow features, and the second stage performs cross-attention between these tokens and reflectance-related features. This progression helps the model move from exposure correction to joint illumination-content reasoning.

The framework is inspired by Retinex separation but avoids explicit physical decomposition. Classical Retinex assumes I = R x L, an assumption that can become unstable when illumination is spatially complex or noise is severe. IA2former instead uses attention to learn a functional separation: illumination-aware dependencies are captured first, and content features are then refined under illumination guidance. Compared with the standard Restormer blocks [16], the IA2former block embeds this illumination-aware interaction directly into the restoration backbone.

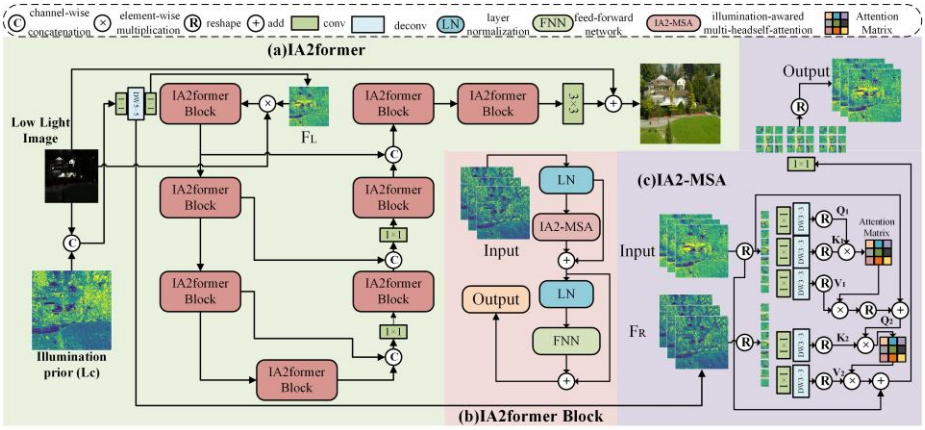

Fig. 2. Overview of the proposed framework, including (a) the Illumination-Aware Attention-based Transformer (IA2former), (b) the IA2former block, and (c) the Illumination-Aware Attention-based Multi-Head Self-Attention (IA2-MSA). For compactness, the IA2former block is shown with a fused input in the overall framework, while the IA2-MSA module explicitly illustrates the illumination-guided feature and reflectance-related feature streams.

Each IA2former block jointly uses the illumination-guided feature Fc and the reflectance-related feature FR. These streams are first projected into a fused token representation, and IA2-MSA then separates their roles through the two-stage attention process. The first stage emphasizes global illumination distribution and exposure correction, while the second stage encourages interaction between illumination-aware tokens and structural content features. Unlike Retinex-based Transformers that mainly modulate features with illumination maps, IA2former treats illumination and content as mutually related streams and fuses them through cross-attention.

In the following, we detail the two core components of the proposed framework. Section 3.1 introduces the Illumination-Aware Attention-based Multi-Head Self-Attention (IA2-MSA), and Section 3.2 presents the Illumination-Aware Loss designed to further constrain the enhancement process from an illumination perspective.

3.1 Illumination-Aware Attention-based Multi-Head Self-Attention

The IA2-MSA module is designed to make illumination participate in attention rather than act only as a scalar guidance map. It first derives illumination-aware token representations from fused shallow features, and then uses these tokens to interact with reflectance-related content features. This organization enables a gradual transition from brightness-oriented modeling to deeper illumination-content reasoning, which benefits exposure recovery and structure preservation under complex lighting.

Compared with IG-MSA and RetinexFormer-like illumination modulation, IA2-MSA differs mainly in feature flow. It still uses the standard query-key-value attention formulation, but illumination-aware tokens and reflectance-related tokens are arranged

into a two-stage interaction. Thus, illumination variation and scene content are modeled as coupled feature streams instead of a single modulated representation.
Before entering IA2-MSA, the illumination-guided feature F_c and the reflectance-related feature F_R are concatenated and projected into a unified input token sequence:

$$X = Z_{\text{in}} = \text{Tok}(\text{Proj}([F_R, F_c])), \tag{4}$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation, Proj(·) aligns the fused features to the embedding dimension, Tok(·) converts spatial features into tokens, and Z_{in} denotes the input token sequence used for residual connections.

The fused token sequence is then reshaped back to the spatial domain and passed through three lightweight embedding branches:

$$\overline{F_{\text{in}}} = \text{Tok}^{-1}(X) \in \mathbb{R}^{H \times W \times C}, F^{(i)} = \phi_i(\overline{F_{\text{in}}}), i \in \{1,2,3\}, \tag{5}$$

where $\overline{F_{\text{in}}}$ denotes the spatial feature map reconstructed from the fused input tokens X, and $F^{(i)}$ represents the feature map generated by the i-th lightweight embedding branch ϕ_i. The embedded features are then tokenized as $Z^{(i)}$.

Stage 1: Illumination Self-Awareness. In the first stage, cross-layer attention is performed among the embedded input features to capture global illumination-aware dependencies. The query, key, and value are constructed as:

$$Q^{(1)} = Z^{(1)}W_Q^{(1)}, K^{(1)} = Z^{(2)}W_K^{(1)}, V^{(1)} = Z^{(3)}W_V^{(1)}, \tag{6}$$

where the three token sequences $Z^{(1)}$, $Z^{(2)}$, $Z^{(3)}$ are generated by parallel lightweight embedding branches and capture complementary contextual cues. The attention output is computed and followed by a residual connection in the token space:

$$\tilde{Z}^{(1)} = \text{softmax}\left(\frac{Q^{(1)}(K^{(1)})^T}{\sqrt{d}}\right)V^{(1)}, Z_{\text{out}}^{(1)} = \tilde{Z}^{(1)} + Z_{\text{in}}. \tag{7}$$

This operation produces an illumination-aware token representation that captures global cross-layer dependencies while preserving the original input information through the residual connection.

Stage 2: Illumination-Guided Enhancement. The second stage explicitly models the interaction between illumination-aware representations and reflectance-related features. The illumination-aware tokens obtained from the first stage are first mapped back to the spatial domain and re-embedded:

$$Z_{\text{IA}} = \text{Tok}(\phi(\text{Tok}^{-1}(Z_{\text{out}}^{(1)}))), \tag{8}$$

where $Z_{\text{out}}^{(1)}$ is reshaped to the spatial domain via $\text{Tok}^{-1}(\cdot)$, refined by a lightweight embedding function $\phi(\cdot)$, and then re-tokenized to obtain the illumination-aware token representation Z_{IA}.

Using the illumination-aware tokens as queries and the tokenized reflectance-related features as keys and values, the second-stage cross-attention is formulated as:

$$Z_R = \mathrm{Tok}(F_R), Q^{(2)} = Z_{\mathrm{IA}} W_Q^{(2)}, K^{(2)} = Z_R W_K^{(2)}, V^{(2)} = Z_R W_V^{(2)}. \quad (9)$$

The second-stage attention output is computed as:

$$\tilde{Z}^{(2)} = \mathrm{softmax}\left(\frac{Q^{(2)}(K^{(2)^T})}{\sqrt{d}}\right) V^{(2)}. \quad (10)$$

The final token representation is obtained through residual addition between the second-stage attention output and the original input tokens:

$$Z_{\mathrm{out}}^{(2)} = \tilde{Z}^{(2)} + Z_{\mathrm{in}}, \hat{F} = \mathrm{Tok}^{-1}(Z_{\mathrm{out}}^{(2)}), \quad (11)$$

where + denotes element-wise residual addition between token representations with matched dimensions, and F is the final enhanced feature map.

3.2 Illumination-Aware Loss

Although the proposed IA2-MSA module enhances low-light images through illumination-guided feature interaction, directly amplifying low-light regions may also introduce undesirable feature drift, color deviation, and noise amplification. To mitigate this issue, we introduce an Illumination-Aware Loss as an auxiliary regularization term. Specifically, the estimated illumination map L_c is normalized to [0, 1] and serves as an illumination-sensitive weighting factor in the feature regularization. Let $L_c(i)$ denote the illumination map value at spatial location i, and let $F_R(i)$ and $F_R^{enhanced}(i)$ denote the reflectance-related features before and after enhancement, respectively. The proposed loss is defined as:

$$\mathcal{L}_{\mathrm{IA}} = \sum_i \| \left(\frac{L_c(i)}{255} \odot F_R^{\mathrm{enhanced}}(i)\right) - \left(\frac{L_c(i)}{255} \odot F_R(i)\right) \|^2. \quad (12)$$

This loss does not aim to enforce identity between the enhanced image and the low-light input. Instead, it regularizes the illumination-weighted reflectance-related representation so that the enhancement process avoids illumination-inconsistent changes in structural and chromatic features. The actual enhancement is driven by the reconstruction objective against the normal-light reference, while $\mathcal{L}_{\mathrm{IA}}$ acts as a constraint that suppresses excessive reflectance drift in illumination-sensitive regions. In this manner, the proposed loss helps reduce color shift, noise amplification, and local artifacts without impeding brightness recovery or detail enhancement. The overall training objective is formulated as:

Accordingly, IA Loss should be viewed as an empirical regularizer. Its purpose is to stabilize feature changes in dark or spatially uneven regions, reducing color shift, noise amplification, and detail distortion while leaving the reconstruction objective to learn the desired normal-light appearance.

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{IA}} \mathcal{L}_{\mathrm{IA}}, \quad (13)$$

where $\mathcal{L}_{\text{rec}}$ denotes the reconstruction loss between the enhanced image and the normal-light ground truth, and λ_{IA} controls the strength of the illumination-aware regularization.

4 Experiment

4.1 Datasets and Implementation Details

Experiments are conducted on LOL-v1 [11] and LOL-v2 [28]. LOL-v2 includes both Real and Synthetic subsets. Following the commonly used splits [11,17,28], LOL-v1 uses 485 training and 15 testing pairs, LOL-v2-Real uses 689 training and 100 testing pairs, and LOL-v2-Synthetic uses 900 training and 100 testing pairs.

Our method is implemented in PyTorch and trained with DDP on 4× NVIDIA RTX 3090 GPUs. We use the Adam optimizer with _1=0.9 and _2=0.999 for 2.5×10^5 iterations. The learning rate is initialized as 2×10^-4 and decayed to 1×10^-6 by cosine annealing. Paired patches of size 128×128 are randomly cropped for training, with random flipping and rotation for data augmentation. The batch size is 16 per GPU. The training objective combines L1, SSIM, and perceptual losses as the reconstruction objective, together with the proposed IA Loss. We report PSNR, SSIM, and LPIPS for quantitative evaluation.

4.2 Low-light Image Enhancement

Quantitative Results. Table 1 compares full-reference metrics on LOL-v1 and LOL-v2, where higher PSNR/SSIM and lower LPIPS indicate better quality. IA2former provides a balanced result across the three benchmarks. It ranks first on LOL-v1 with 25.89 dB PSNR and 0.878 SSIM, achieves the best PSNR on LOL-v2-Synthetic with 26.71 dB, and obtains the strongest SSIM and LPIPS on LOL-v2-Real. LLFlow reports the highest PSNR on LOL-v2-Real, but IA2former shows better structural and perceptual scores on this subset, suggesting that the proposed illumination-aware interaction is beneficial for preserving visible details and reducing perceptual distortion.

Table 1. Full-reference evaluation on LOL-v1 and LOL-v2. The best results are highlighted in bold, and the second-best results are underlined.

Method	LOL-v1			LOL-v2-Real			LOL-v2-Syn		
	PSNR ↑	SSIM↑	LPIPS ↓	PSNR ↑	SSIM↑	LPIPS ↓	PSNR ↑	SSIM↑	LPIPS ↓
EnlightenGAN	17.48	0.652	0.322	18.68	0.678	0.364	15.91	0.637	0.371
KinD	19.55	0.813	0.168	18.19	0.836	0.166	17.99	0.812	0.259
Restormer	22.33	0.819	0.153	19.02	0.791	0.236	21.83	0.861	0.148
SNR-Net	23.89	0.842	0.151	21.48	0.848	<u>0.137</u>	24.01	0.927	<u>0.057</u>
LLFlow	25.13	<u>0.872</u>	**0.117**	**26.20**	<u>0.888</u>	<u>0.137</u>	24.81	0.919	0.067
RetinexMamba	24.09	0.831	0.148	23.07	0.848	0.178	<u>25.93</u>	**0.934**	**0.042**
RetinexFormer	<u>25.16</u>	0.845	0.131	22.80	0.840	0.171	25.67	0.930	0.059
IA2former	**25.89**	**0.878**	<u>0.129</u>	<u>25.14</u>	**0.891**	**0.132**	**26.71**	<u>0.933</u>	0.061

Qualitative Results. Fig. 3 presents a visual comparison on an image captured under extremely weak illumination. The input image in (a) is dark and color-biased, whereas the normal-light reference in (b) contains clearer brightness, texture, and chromatic information. The zoomed regions on the folded clothes and plush toy reveal different failure patterns. EnlightenGAN (c) increases brightness only partially and leaves strong under-exposure on the clothes. KinD (d) improves global illumination but introduces blotchy transitions in shadowed areas.

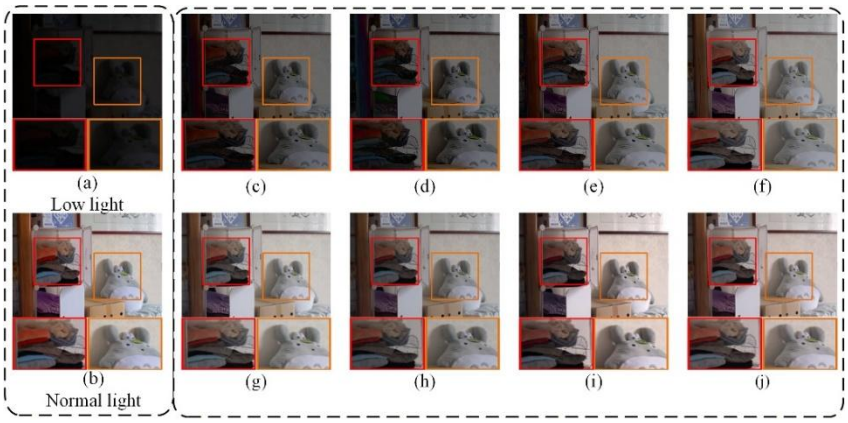

Fig. 3. Qualitative comparison on a low-light image enhancement example. (a) Low-light input; (b) normal-light reference ground truth; (c) EnlightenGAN [13]; (d) KinD [19]; (e) Restormer [16]; (f) SNR-Net [15]; (g) LLFlow [14]; (h) RetinexMamba [18]; (i) RetinexFormer [17]; and (j) IA2former (ours). The red and orange boxes indicate zoom-in regions for clearer comparison of detail restoration, noise suppression, and color fidelity across methods.

Restormer (e) brightens the scene but smooths fabric textures and weakens fine details. SNR-Net (f) controls noise effectively, yet its colors appear desaturated and less natural. LLFlow (g) enhances the overall exposure but leaves residual noise in darker folds. RetinexMamba (h) keeps contrast but introduces over-enhancement on brighter regions of the toy. RetinexFormer (i) produces a more balanced result, although some edge sharpness and color fidelity are still limited. IA2former (j) better recovers cloth color and toy texture, which supports the benefit of explicit illumination-content interaction.

Fig. 4 reports pixel-wise error maps on LOL-v1, LOL-v2-Real, and LOL-v2-Synthetic. Brighter colors correspond to larger differences from the normal-light target. On LOL-v1, Restormer and RetinexFormer show clear errors around boundaries such as the stadium seats and railings, while IA2former produces a darker and more spatially consistent map. On LOL-v2-Real, IA2former reduces error responses over water reflections and mountain textures, indicating better exposure and color consistency. On LOL-v2-Synthetic, its error map contains fewer high-frequency speckles around foliage, suggesting stronger robustness to synthetic degradation and detailed structures.

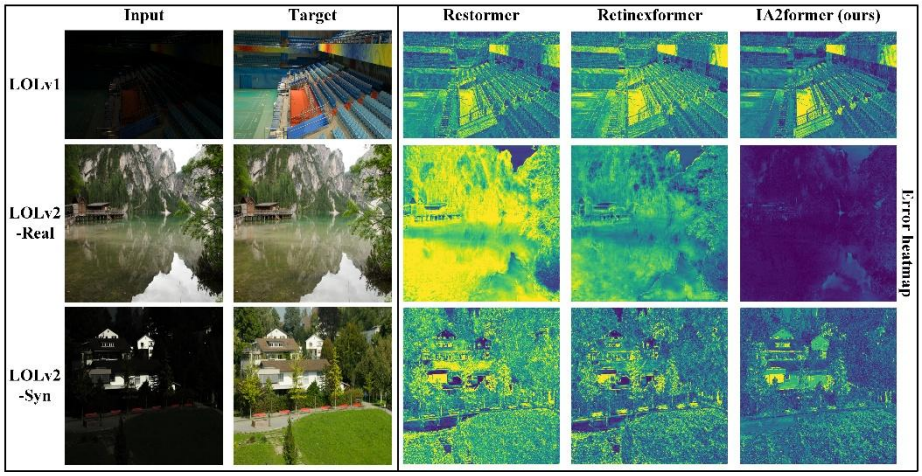

Fig. 4. Error heatmap visualization on three benchmarks: LOL-v1, LOL-v2-Real, and LOL-v2-Synthetic. From left to right are the input, target ground truth, and the corresponding error maps of Restormer [16], RetinexFormer [17], and IA2former (ours). Brighter colors indicate larger reconstruction errors, highlighting regions where the enhancement results deviate more from the target.

4.3 Ablation Study

We further conduct an ablation study to isolate the contribution of each component in IA2former. Six variants are evaluated:

IG-MSA [17]: Baseline using illumination-guided multi-head self-attention. **IG-MSA + IA Loss**: IG-MSA equipped with illumination-aware loss. **IA2-MSA**: Proposed illumination-aware attention-based MSA module. **IA2-MSA + IA Loss**: Full IA2former model. **Stage-1 only**: Only the first stage of IA2-MSA is used. **Stage-2 only**: Only the second stage of IA2-MSA is used.

Table 2. Ablation results on LOL-v1.

Method	**LOL-v1**			**Params (M)↓**	**FLOPs (G)↓**
	PSNR↑	SSIM↑	LPIPS↓		
IG-MSA	24.99	0.844	0.134	**3.84**	**24.71**
IG-MSA + IA Loss	25.18	0.848	<u>0.130</u>	**3.84**	<u>24.83</u>
Stage-1 only	25.12	0.855	0.132	4.05	26.10
Stage-2 only	25.17	0.857	0.131	4.05	26.10
IA2-MSA	<u>25.53</u>	<u>0.861</u>	0.131	4.31	29.37
IA2-MSA + IA Loss	**25.89**	**0.878**	**0.129**	4.31	29.50

Table 2 shows that replacing IG-MSA with the proposed IA2-MSA consistently improves all metrics. IA2-MSA increases PSNR from 24.99 to 25.53, raises SSIM from 0.844 to 0.861, and reduces LPIPS from 0.134 to 0.131. The single-stage versions also improve over the baseline, with Stage-1 only and Stage-2 only reaching 25.12 and 25.17 dB PSNR, respectively, but neither matches the complete two-stage design. Adding IA

Loss further improves the full model to 25.89 dB PSNR, 0.878 SSIM, and 0.129 LPIPS. These results indicate that the two-stage attention structure and illumination-aware regularization contribute complementary gains.

The performance improvement comes with a moderate increase in computation. Moving from IG-MSA to the full two-stage IA2-MSA raises FLOPs from 24.71G to 29.50G, showing a trade-off between quality and efficiency. Future work will explore lightweight attention, parameter sharing, and systematic runtime and memory evaluation.

5 Conclusion

This paper presents IA2former, a low-light enhancement framework that incorporates illumination-aware interaction into a Transformer-style restoration model. The proposed IA2-MSA module models the relationship between illumination-aware tokens and reflectance-related content features, while IA Loss regularizes illumination-weighted feature changes during enhancement. Experiments on LOL-v1 and LOL-v2 demonstrate strong quantitative and qualitative performance compared with representative low-light enhancement methods. The results suggest that explicitly coupling illumination and content features is effective for improving brightness recovery, color consistency, and structural detail.

6 Limitations

IA2former still has several limitations. Extremely dark scenes or highly dynamic lighting may lead to local under-enhancement or slight over-exposure. In addition, the two-stage cross-attention design increases computational cost, which may limit deployment on resource-constrained devices.

The current evaluation is also limited to LOL-v1 and LOL-v2, which are useful standardized benchmarks but cannot fully cover real-world low-light diversity. Future studies should include more challenging paired, unpaired, and real-captured datasets such as SID, SMID, VE-LOL, ExDark, and naturally collected low-light images. Broader comparisons with diffusion-based, prompt-based, all-in-one restoration, and newer Mamba-based methods would further clarify the generalization and efficiency of IA2former.

References

1. Arici, Tarik, Salih Dikbas, and Yucel Altunbasak. "A histogram modification framework and its application for image contrast enhancement." IEEE Transactions on image processing 18.9 (2009): 1921-1935.

2. Wang, Xuewen, and Lixia Chen. "An effective histogram modification scheme for image contrast enhancement." Signal Processing: Image Communication 58 (2017): 187-198.

3. Qiuqi R, Yuzhi R (2013) Digital Image Processing, 3rd edn. Publishing House of Electronics Industry.

4. Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. TIP, 2012.

5. Shanto Rahman, Md Mostafijur Rahman, Mohammad Abdullah-Al-Wadud, Golam Dastegir Al-Quaderi, and Mohammad Shoyaib. An adaptive gamma correction for image enhancement. EURASIP Journal on Image and Video Processing, 2016.

6. Zhi-Guo Wang, Zhi-Hu Liang, and Chun-Liang Liu. A real time image processor with combining dynamic contrast ratio enhancement and inverse gamma correction for PDP. Displays, 2009.

7. Zosso, Dominique, Giang Tran, and Stanley J. Osher. "Non-local Retinex–-A unifying framework and beyond." SIAM Journal on Imaging Sciences 8.2 (2015): 787-826.

8. Blake, Andrew. "Boundary conditions for lightness computation in Mondrian world." Computer vision, graphics, and image processing 32.3 (1985): 314-327.

9. Blake, Andrew. "On lightness computation in Mondrian world." Central and Peripheral Mechanisms of Colour Vision. 1985.

10. Brelstaff, Gavin, and Andrew Blake. "Computing lightness." Pattern Recognition Letters 5.2 (1987): 129-138.

11. Wei, Chen, et al. "Deep retinex decomposition for low-light enhancement." arXiv preprint arXiv:1808.04560 (2018).

12. Koohestani, Farzaneh, et al. "BrightVAE: Luminosity Enhancement in Underexposed Endoscopic Images." arXiv preprint arXiv:2411.14663 (2024).

13. Jiang, Yifan, et al. "EnlightenGAN: Deep light enhancement without paired supervision." IEEE transactions on image processing 30 (2021): 2340-2349.

14. Wang, Yufei, et al. "Low-light image enhancement with normalizing flow." Proceedings of the AAAI conference on artificial intelligence. 2022.

15. Xu, Xiaogang, et al. "SNR-aware low-light image enhancement." In CVPR, 2022.

16. Zamir, Syed Waqas, et al. "Restormer: Efficient transformer for high-resolution image restoration." CVPR. 2022.

17. Cai, Yuanhao, et al. "RetinexFormer: One-stage retinex-based transformer for low-light image enhancement." ICCV. 2023.

18. Bai, Jiesong, et al. "RetinexMamba: Retinex-based mamba for low-light image enhancement." arXiv preprint arXiv:2405.03349 (2024).

19. Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. ACM MM (2019).

20. Rahman, Z.u., Jobson, D.J., Woodell, G.A.: Retinex processing for automatic image enhancement. Journal of Electronic imaging (2004).

21. Jobson, D.J., Rahman, Z.u., Woodell, G.A.: Properties and performance of a center/surround retinex. IEEE transactions on image processing (1997).

22. Gharbi, Michaël, et al. "Deep bilateral learning for real-time image enhancement." ACM Transactions on Graphics (TOG) (2017).

23. Zamir, Syed Waqas, et al. "Learning enriched features for real image restoration and enhancement." ECCV. 2020.

24. Zhou, Dewei, Zongxin Yang, and Yi Yang. "Pyramid diffusion models for low-light image enhancement." arXiv preprint arXiv:2305.10028 (2023).

25. Zhang, Ling, et al. "DLIENet: A lightweight low-light image enhancement network via knowledge distillation." Pattern Recognition 169 (2026): 111777.

26. Sun, Wei, et al. "AdaPrompt-IR: Adaptive learning to perceive degradation semantic and prompting for all-in-one image restoration." Pattern Recognition (2026).

27. Zhang, Xiaofeng, et al. "Memory augment is all you need for image restoration." IEEE Transactions on Consumer Electronics (2026).

28. Yang, Wenhan, et al. "Sparse gradient regularized deep Retinex network for robust low-light image enhancement." IEEE Transactions on Image Processing 30 (2021): 2072-2086.

29. Wu, H., Lin, Y.: A high-performance and enhanced generalization small target defect detection method for PCB boards based on YOLO-EMAC. IEEE Transactions on Instrumentation and Measurement 74, Art. no. 5042513, 1-13 (2025). https://doi.org/10.1109/TIM.2025.3602542

30. Xiong, S., Wu, H., Lin, Y.: A Robotic Grasping Framework for Railway Cargo Based on an Improved YOLOv8 Model. In: 2025 8th International Conference on Robotic Systems and Applications (ICRSA) (2025). https://doi.org/10.1109/ICRSA66467.2025.11383796

31. Wu, H., Xiong, S., Lin, Y.: A Novel 24 h × 7 Days Broken Wire Detection and Segmentation Framework Based on Dynamic Multi-Window Attention and Meta-Transfer Learning. Sensors 25(12), 3718 (2025). https://doi.org/10.3390/s25123718

32. Wu, H., Lin, Y., Chen, Y., Li, J.: Damaged Wire Detection and Segmentation Method Based on Lightweight Channel Attention Mechanism and Transfer Learning. In: 2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 2569-2574 (2025). https://doi.org/10.1109/CSCWD64889.2025.11033272