SeMaNet: Semantic-Guided Low-Light Image Enhancement with Hybrid Transformer-Mamba Architecture

Tianzhi Jia , Shikui Wei * , Yao Zhao

Posted Date: 23 April 2026

doi: 10.20944/preprints202604.1600.v1

Keywords: low-light image enhancement; vision-language model; attention; Mamba; Retinex decomposition

 doi:10.20944/preprints202604.1600.v1

Article

SeMaNet: Semantic-Guided Low-Light Image Enhancement with Hybrid Transformer-Mamba Architecture

Tianzhi Jia [1,2], Shikui Wei [1,2,*] and Yao Zhao [1,2]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[2] Visual Intelligence + X International Joint Laboratory of the Ministry of Education, Beijing 100044, China
* Correspondence: shkwei@bjtu.edu.cn

Abstract

Low-light image enhancement aims to recover high-quality visuals from poorly illuminated inputs, yet existing methods often suffer from over-enhancement, noise amplification, and semantic inconsistency in complex scenes. In this paper, we propose SeMaNet, a novel semantic-guided framework that integrates textual priors with a hybrid Transformer-Mamba architecture for controllable and efficient low-light enhancement. Our approach begins by leveraging pre-trained CLIP to generate semantically meaningful attention maps from natural language prompts, enabling interpretable region-aware enhancement without requiring pixel-level annotations. These semantic priors are then fused with illumination estimates and raw image features through a cross-attention mechanism, allowing dynamic interaction among multi-modal cues. To balance global context modeling and computational efficiency, we design a U-Net-based restoration network that interleaves Transformer blocks for long-range dependency capture and Mamba layers for linear-time sequence processing. Furthermore, our method explicitly models the image formation process via a perturbation-aware Retinex decomposition, enhancing physical plausibility. Extensive experiments on LOL v1, LOL-v2-real, LOL-v2-synthetic, SID, SMID, and SDSD-out datasets demonstrate that SeMaNet achieves state-of-the-art performance in both quantitative metrics (PSNR, SSIM) and qualitative quality, particularly excelling in preserving semantic coherence and fine details under challenging lighting conditions. The hybrid architecture also offers superior inference efficiency compared to pure Transformer-based models.

Keywords: low-light image enhancement; vision-language model; attention; Mamba; Retinex decomposition

1. Introduction

Low-light image enhancement (LLIE) is a fundamental problem in computer vision, aimed at recovering visually compelling and information-rich images from poorly illuminated inputs. The prevalence of low-light conditions in real-world scenarios—such as nighttime autonomous driving [1], surveillance in dim environments [2], indoor mobile photography [3], and medical imaging under exposure constraints [4]—has made LLIE increasingly critical for both human perception and machine vision systems. Images captured under inadequate illumination typically exhibit low visibility, elevated noise levels, color distortion, and loss of fine details, which severely degrade visual quality and impair the performance of downstream computer vision tasks including object detection, semantic segmentation, and scene understanding [5].

Despite decades of research, LLIE remains a challenging problem due to several inherent difficulties. First, the degradation process in low-light imaging is highly complex and often involves non-uniform illumination distributions, spatially varying noise patterns, and signal-dependent sensor characteristics. Second, enhancement algorithms must carefully balance multiple competing objectives:

 doi:10.20944/preprints202604.1600.v1

improving brightness while avoiding over-enhancement, suppressing noise without losing texture details, and restoring colors while maintaining naturalness. Third, many existing methods apply spatially uniform transformations that fail to account for semantic variations across different image regions—for instance, treating foreground objects (e.g., human faces) identically to background elements (e.g., sky or walls), which can lead to suboptimal perceptual quality.

Traditional enhancement techniques, including histogram equalization [6] and gamma correction [7], perform global intensity transformations that are computationally efficient but often produce over-enhancement, contrast distortion, and noise amplification. The Retinex theory [8,9], which models image formation as the product of reflectance and illumination components, has inspired numerous decomposition-based methods [10,11]. However, classical Retinex approaches frequently suffer from halo artifacts and struggle to handle real-world noise characteristics. The advent of deep learning has significantly advanced the field, with convolutional neural networks (CNNs) demonstrating impressive capabilities in learning complex mappings from low-light to well-exposed images [3,12–16]. More recently, attention mechanisms [17–19] and Vision Transformers [20–22] have enabled spatially adaptive enhancement and long-range dependency modeling, achieving state-of-the-art performance on standard benchmarks.

Despite these advances, current LLIE methods exhibit several critical limitations. First, most existing approaches lack high-level semantic awareness, treating all image regions uniformly regardless of their semantic importance. Human visual perception, however, naturally prioritizes semantically salient content—we tend to focus on faces, foreground objects, and regions of interest rather than processing the entire scene uniformly. Second, Transformer-based methods [21,22], while effective at capturing global context, suffer from quadratic computational complexity with respect to image resolution, limiting their applicability to high-resolution inputs and real-time scenarios. Third, recent State Space Models like Mamba [23–29] offer linear-time complexity but have not been fully explored in conjunction with semantic guidance for controllable enhancement.

To address these limitations, we propose SeMaNet (Semantic-Guided Mamba Network), a novel framework that integrates textual semantic priors with a hybrid Transformer-Mamba architecture for controllable and efficient low-light image enhancement. Our key insight is that leveraging pre-trained vision-language models such as CLIP [30] enables the extraction of semantically meaningful attention maps from natural language prompts (e.g., "brighten the person" or "enhance the sky"), providing interpretable, region-aware guidance without requiring pixel-level annotations. These semantic priors are dynamically fused with illumination estimates and raw image features through a cross-attention mechanism, allowing multi-modal cues to interact and mutually reinforce each other. To balance global context modeling with computational efficiency, we design a U-Net-based restoration network that strategically interleaves Transformer blocks—which excel at capturing long-range dependencies—with Mamba layers that process sequences in linear time. Furthermore, we explicitly model the image formation process through a perturbation-aware Retinex decomposition, which accounts for real-world noise and illumination estimation errors, thereby enhancing physical plausibility. The main contributions of this paper can be summarized as follows:

1. **Semantic-Guided Enhancement Framework:** We propose a novel LLIE method that integrates CLIP-based semantic priors from natural language prompts, enabling interpretable, region-aware enhancement without pixel-level supervision. This allows users to specify which regions or semantic content should be prioritized during enhancement.
2. **Hybrid Transformer-Mamba Architecture:** We design a novel U-Net backbone that synergistically combines Transformer blocks for global context modeling and Mamba layers for efficient sequential processing. This hybrid architecture achieves superior performance with significantly reduced computational overhead compared to pure Transformer-based models.
3. **Perturbation-Aware Retinex Decomposition:** We introduce a physically grounded decomposition framework that explicitly models noise and illumination estimation errors, improving the realism and robustness of the enhancement process under challenging real-world conditions.

 doi:10.20944/preprints202604.1600.v1

4. **Comprehensive Experimental Validation:** Extensive experiments on multiple benchmarks (LOL v1, LOL-v2-real, LOL-v2-synthetic, SID, SMID, and SDSD-out) demonstrate that SeMaNet achieves state-of-the-art performance in both quantitative metrics (PSNR, SSIM) and qualitative visual quality, particularly excelling in preserving semantic coherence and fine details under challenging lighting conditions.

The remainder of this paper is organized as follows: Section 2 reviews related work in low-light image enhancement, attention mechanisms, Transformers, State Space Models, and vision-language models. Section 3 presents the detailed methodology of SeMaNet, including the semantic prior extraction module, Transformer-Mamba architecture, and Retinex-based fusion. Section 4 describes the experimental setup and presents comprehensive results. Section 5 concludes the paper and discusses future directions.

2. Related Work

2.1. Traditional and CNN-Based Low-Light Enhancement

Early approaches to low-light image enhancement primarily relied on handcrafted techniques. Histogram equalization (HE) [6] redistributes pixel intensities to expand the dynamic range, while gamma correction [7] applies power-law transformations to modify luminance curves. Although computationally efficient, these methods operate globally and lack adaptability to local illumination variations, often resulting in over-enhancement, contrast distortion, and amplified noise.

The Retinex theory [8,9] provides a more principled framework by decomposing images into reflectance and illumination components. Jobson et al. [10] introduced Single-Scale Retinex (SSR) and Multi-Scale Retinex (MSR) to improve perceptual quality, while Multi-Scale Retinex with Color Restoration (MSRCR) [11] further enhanced color fidelity. Despite their theoretical elegance, classical Retinex methods frequently produce halo artifacts and struggle with real-world noise characteristics.

The emergence of deep learning revolutionized LLIE by enabling data-driven learning of complex mappings. Early CNN-based approaches such as DeepUPE [12] and RetinexNet [13] demonstrated the potential of end-to-end learning. RetinexNet [13] pioneered the integration of Retinex theory into deep networks by jointly optimizing decomposition and enhancement. Subsequent works refined this paradigm: KinD [14] and KinD++ [15] introduced dedicated subnetworks for decomposition and adjustment, while Zero-DCE [3] and Zero-DCE++ [16] proposed zero-reference learning based on curve estimation, eliminating the need for paired training data. EnlightenGAN [17] employed global-local discriminators with self-attention modules for selective enhancement. However, these methods typically apply spatially uniform or weakly adaptive processing, often overlooking semantic variations across different image regions and lacking explicit content-aware guidance.

2.2. Transformer-Based Architectures and Attention Mechanisms

Attention mechanisms have enabled spatially adaptive enhancement by allowing networks to dynamically focus on informative regions. Frequency-domain attention methods [18] decompose images into different frequency components for selective enhancement, while NAFNet [19] achieves an efficient balance through gated convolutions and simplified attention operations.

Vision Transformers (ViTs) [20] have emerged as powerful architectures for image restoration by leveraging self-attention mechanisms to capture long-range dependencies. Unlike CNNs with limited receptive fields, Transformers can model global context across the entire image, enabling better structural consistency. Restormer [21] combines multi-head self-attention with multi-scale hierarchical architectures, achieving remarkable performance in denoising, deblurring, and super-resolution. In low-light enhancement, RetinexFormer [22] integrates Retinex decomposition with Transformer blocks, leveraging global attention to improve illumination estimation and structural consistency. LLFormer [23] further explores global-local parallel attention for enhanced detail recovery.

However, the quadratic computational complexity of self-attention with respect to the number of tokens—poses significant challenges for processing high-resolution images. This computational

 doi:10.20944/preprints202604.1600.v1

burden limits the practical deployment of pure Transformer-based methods in real-time or resource-constrained scenarios, motivating the exploration of more efficient alternatives that can maintain modeling capacity while reducing computational costs.

2.3. Efficient State Space Models and Mamba Architecture

To address the computational limitations of Transformers, State Space Models (SSMs) have recently gained attention as efficient alternatives for sequence modeling. SSMs, rooted in control theory, model sequences as continuous dynamical systems and offer linear-time complexity O(n) through efficient parallel scan algorithms. Mamba [24], a prominent SSM-based architecture, introduces selective state-space modeling with input-dependent parameters, achieving impressive efficiency in long-sequence processing while maintaining competitive performance.

Mamba-based architectures have been successfully applied to various vision tasks, including image classification [25], medical image segmentation [26], video understanding [27], and general image restoration [28]. In the specific domain of low-light enhancement, RetinexMamba [29] pioneered the application of Mamba to LLIE, demonstrating that SSMs can reduce computational costs while maintaining or even improving enhancement quality compared to CNN-based and Transformer-based methods. The linear complexity of Mamba makes it particularly attractive for high-resolution image processing and real-time applications. However, existing Mamba-based approaches do not incorporate high-level semantic guidance, limiting their ability to perform content-aware, region-adaptive enhancement that aligns with human perceptual priorities.

2.4. Vision-Language Models

Vision-language models, exemplified by CLIP [30], have opened new avenues for incorporating high-level semantic understanding into low-level vision tasks. CLIP learns joint embeddings of images and text through contrastive learning on large-scale datasets, enabling zero-shot transfer and semantic alignment across modalities. CLIP has been successfully applied to various downstream tasks, including image editing [31], semantic segmentation [32], and text-to-image generation [33], demonstrating the power of vision-language priors for semantic-aware processing.

In the image enhancement domain, several recent works have explored unpaired learning strategies [34,35] and illumination-aware modeling [23,36] to improve realism and naturalness. However, the integration of vision-language models for semantic-guided low-light enhancement remains largely unexplored. Unlike existing methods that process all image regions uniformly, human enhancement strategies naturally prioritize semantically important content—for example, ensuring that faces are well-exposed and noise-free while allowing background regions to remain relatively dark. This semantic awareness is crucial for producing perceptually pleasing results that align with human expectations.

Despite significant progress, existing LLIE methods face key limitations: traditional and CNN-based approaches lack semantic awareness and apply largely uniform transformations; Transformer-based methods achieve strong performance but suffer from high computational costs; and recent Mamba-based architectures offer efficiency but do not leverage semantic priors for controllable enhancement. Our proposed SeMaNet addresses these limitations by integrating CLIP-generated semantic attention maps with a hybrid Transformer-Mamba architecture, enabling interpretable, region-aware enhancement guided by natural language prompts while maintaining computational efficiency.

3. Methodology

3.1. Overall Architecture of SeMaNet

As shown in Figure 1, the overall architecture of SeMaNet includes three core components: a text-guided semantic prior extraction module, a Retinex decomposition and semantic-guided fusion module, and a Transformer-Mamba enhanced network. The entire framework takes low-light images and natural language prompts input by users to achieve semantic-aware high-quality image enhancement through multi-stage collaborative processing.

 doi:10.20944/preprints202604.1600.v1

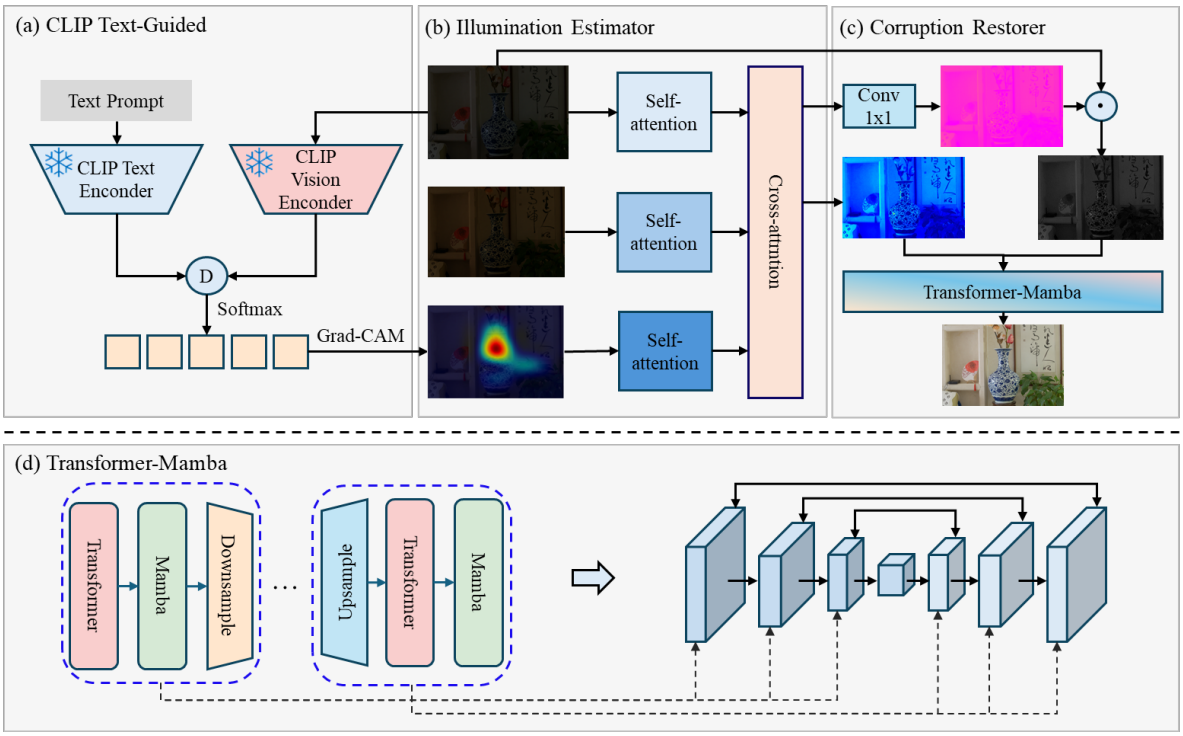

Figure 1. Overall framework of the proposed SeMaNet model.

In the CLIP Text-Guided branch, a pre-trained CLIP model is employed to extract semantic information from text prompts and perform cross-modal alignment with the input image. Semantic attention maps are generated using Gradient-weighted Class Activation Mapping (Grad-CAM) to mark key areas in the image that require enhancement, such as dark objects or noise-dense regions. This map serves as high-level semantic prior to guide the subsequent feature fusion process.

In the Illumination Estimator module, both the original low-light image and the illumination estimation map undergo feature enhancement through self-attention mechanisms, while incorporating the aforementioned semantic attention map as a third input stream. Dynamic interaction within a cross-attention module among the three inputs achieves deep integration between image content, illumination structure, and semantic priors. Features after fusion retain original information through residual connections and are passed on to the next stage.

In the Corruption Restorer module, fused features, after channel adjustment by a 1×1 convolution layer, are fed into a Transformer-Mamba hybrid network based on a U-Net architecture. The network combines the global modeling capability of Transformers with the efficient sequence processing characteristics of Mamba, alternately employing both along the encoder path to extract multi-scale features; the decoder path gradually restores spatial resolution through upsampling and merges information from both high and low levels via skip connections. The final output is an enhanced image with good visual quality and semantic consistency. The entire framework realizes an end-to-end semantic-driven enhancement process from text to image, improving the brightness and detail expressiveness of low-light images.

3.2. Text-Guided Semantic Prior Extraction Module

To introduce high-level semantic guidance in the process of low-light image enhancement, this paper proposes a text-guided semantic prior extraction module. The semantic prior module leverages the cross-modal understanding capability of vision-language models, combining natural language prompts to generate semantic-aware attention maps aimed at highlighting key areas in the image that require special processing, such as dark objects, distorted regions, or noise-dense areas.

Given an input low-light image I and text prompt T, a pre-trained CLIP vision-language model is used to calculate the image-text matching score, and Gradient-weighted Class Activation Mapping

 doi:10.20944/preprints202604.1600.v1

(Grad-CAM) is utilized to produce heatmap responses. The heatmap corresponding to the text prompt is defined as follows:

$$A = \mathrm{ReLU}\left(\sum_{c=1}^{C} \alpha_c F_c\right) \tag{1}$$

where F_c represents the feature map from the last layer of the visual encoder, the ReLU operation retains regions contributing positively to the current task, and α_c denotes the globally averaged gradients computed based on the text prompt:

$$\alpha_c = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \frac{\partial y}{\partial F_c(i,j)} \tag{2}$$

where y is the text matching score output by the model.

Subsequently, this heatmap A is bilinearly upsampled to the original resolution to generate the final semantic attention map. It further constitutes, along with the original low-light image and illumination prior map, the inputs for the enhancement network.

The text-guided semantic prior extraction module can adaptively focus on key areas described by text during the enhancement process, achieving semantic-aware local brightness adjustment and detail recovery. This module does not require end-to-end training of the vision-language model but utilizes it solely for generating prior attention, offering advantages of computational efficiency and ease of integration while providing interpretable external semantic guidance for the low-light enhancement task.

3.3. Transformer-Mamba Attention Model Architecture

In low-light environments, images commonly suffer from insufficient brightness, low contrast, noticeable noise, and loss of details. Traditional methods are limited by inadequate attention mechanisms, while existing deep learning models incur significant computational overhead for long-sequence modeling. This paper proposes a Transformer-Mamba hybrid architecture that integrates global modeling capabilities with efficient sequence processing advantages to achieve high-quality low-light image enhancement.

As shown in Figure 2, the Transformer architecture achieves powerful global dependency modeling through self-attention mechanisms, achieving remarkable results in fields such as computer vision and natural language processing. However, its quadratic computational complexity $O(n^2)$ poses challenges. Recently, the Mamba architecture based on State Space Models (SSM) has demonstrated excellent efficiency in long-sequence modeling with linear complexity but is less effective at capturing local detail features. This paper designs a Transformer-Mamba hybrid architecture to synergistically optimize global semantic understanding and efficient local feature extraction.

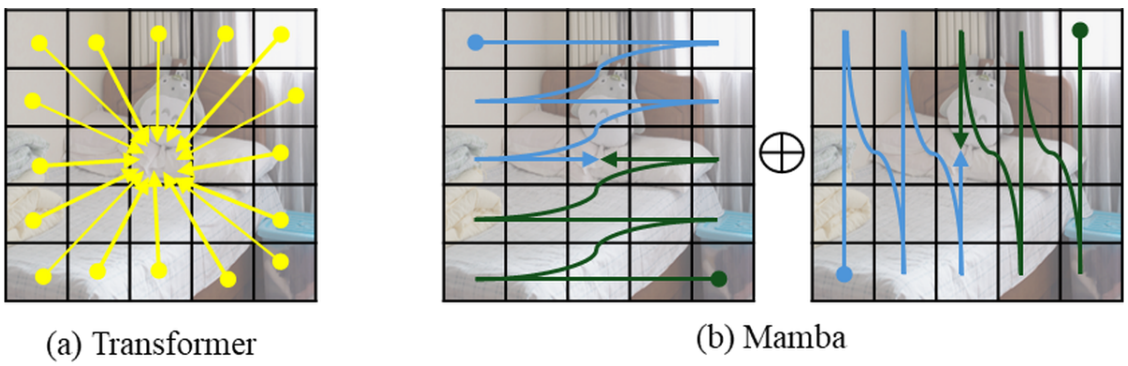

Figure 2. Computational flowcharts of Transformer and Mamba.

The Transformer-Mamba hybrid architecture adopts U-Net as the basic encoder-decoder structure. In the encoder path, each encoding block consists of a Transformer module serially connected with

 doi:10.20944/preprints202604.1600.v1

multiple layers of Mamba modules, followed by downsampling to enter the next level, repeating this process several times to extract multi-scale features. The decoder path gradually restores spatial resolution through upsampling, alternately using Transformer and Mamba modules for feature reconstruction. Multi-scale information is fused between corresponding levels of the encoder and decoder via skip connections, effectively alleviating gradient vanishing issues in deep networks and preserving rich spatial details. Transformer layers capture global context dependencies, enhancing understanding of the overall scene structure, whereas Mamba modules efficiently handle long-sequence features using selective state space models, significantly reducing computational burdens. Their coordinated design enables the model to maintain high performance with good real-time capability.

As illustrated in Figure 3, the core of the Transformer is the Multi-Head Self-Attention mechanism. Given an input sequence,

$$\mathbf{X} \in \mathbb{R}^{n \times d} \tag{3}$$

where n is the sequence length and d is the feature dimension.

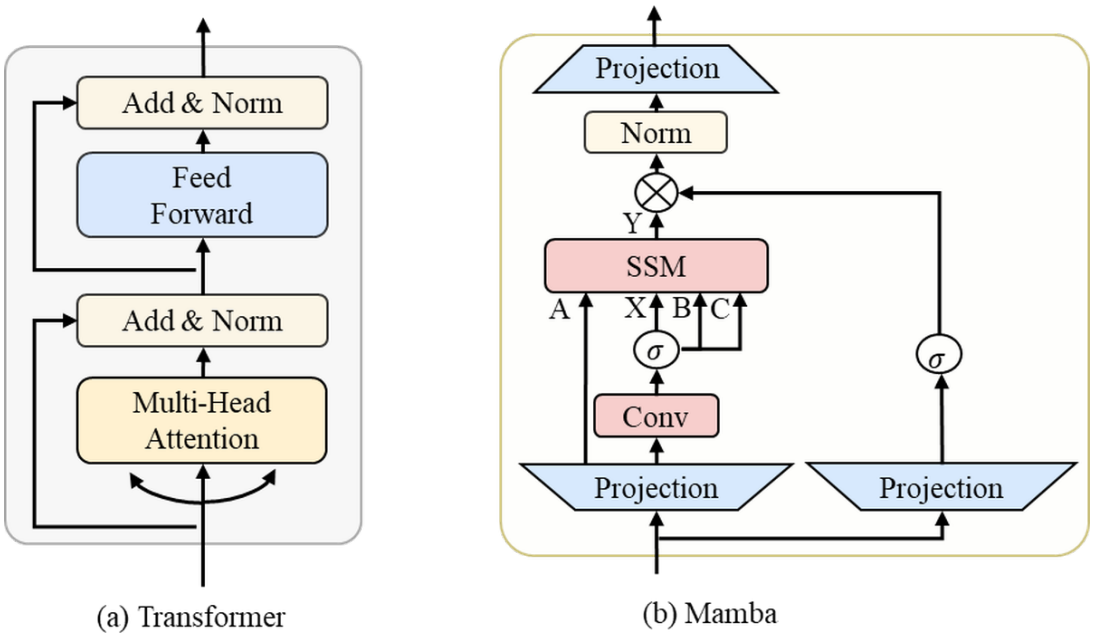

Figure 3. Detailed module structures of Transformer and Mamba.

Learnable linear projections generate Query (Q), Key (K), and Value (V) matrices:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q} \tag{4}$$

$$\mathbf{K} = \mathbf{X}\mathbf{W}^{K} \tag{5}$$

$$\mathbf{V} = \mathbf{X}\mathbf{W}^{V} \tag{6}$$

where $\mathbf{W}^{Q}$, $\mathbf{W}^{K}$, and $\mathbf{W}^{V}$ are projection parameters, and d_k denotes the dimension of each attention head. The scaled dot-product attention is computed as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \tag{7}$$

To enhance representation ability, a multi-head mechanism computes h attention heads in parallel, with the final output concatenated from each head's output:

 doi:10.20944/preprints202604.1600.v1

$$\text{head}_i = \text{Attention}\left(\mathbf{X}\mathbf{W}_i^Q, \mathbf{X}\mathbf{W}_i^K, \mathbf{X}\mathbf{W}_i^V\right), \quad i = 1, \ldots, h \tag{8}$$

After computing attention, features undergo further nonlinear transformation through a feed-forward neural network (FFN), combined with residual connections and layer normalization, enhancing model expressiveness and mitigating gradient vanishing problems.

The State Space Model (SSM) in the Mamba model originates from control theory, treating sequence modeling as a state evolution process of a continuous dynamic system. Its continuous form is defined as

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{A}\mathbf{h}(t) + \mathbf{B}x(t) \tag{9}$$

$$y(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}x(t) \tag{10}$$

where $\mathbf{h}(t)$ is the hidden state, $x(t)$ is the input, $y(t)$ is the output, and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are system parameter matrices.

To adapt to discrete sequence data, the zero-order hold method is adopted for discretization. The discrete state update equations are

$$\mathbf{h}_t = \overline{\mathbf{A}}\mathbf{h}_{t-1} + \overline{\mathbf{B}}x_t \tag{11}$$

$$y_t = \mathbf{C}\mathbf{h}_t + \mathbf{D}x_t \tag{12}$$

Traditional SSMs have fixed parameters and lack adaptability to input context. Mamba introduces a selective mechanism, making parameters $\mathbf{B}$, $\mathbf{C}$, and the sampling interval Δ_t dynamically dependent on the current input x_t:

$$\Delta_t = \text{softplus}(\text{Linear}_\delta(x_t)) \tag{13}$$

$$\mathbf{B}_t = \text{Linear}_B(x_t) \tag{14}$$

$$\mathbf{C}_t = \text{Linear}_C(x_t) \tag{15}$$

This input-aware parameterization enables the model to dynamically adjust state transition behavior based on context, achieving gating and selection functionality similar to LSTM or GRU, thereby better capturing critical information. Mamba can achieve efficient inference with near-linear time complexity $O(n)$ through a parallel scan algorithm.

Building upon illumination guidance, the Transformer-Mamba hybrid architecture focuses on detail reconstruction and noise suppression in low-light images. Using enhanced illumination features as guidance, it performs deep modeling of low-light images, effectively restoring texture structures and removing noise introduced by low-light conditions. By leveraging Transformer to capture long-range spatial dependencies for global content consistency, combined with Mamba's efficient modeling of local sequential features, the architecture improves detail restoration quality in complex regions, achieving high-quality image reconstruction and denoising while preserving color fidelity.

3.4. Retinex Decomposition and Semantic-Guided Fusion Module

The core objective of image enhancement is to improve visual readability under low-light conditions while preserving structural and color fidelity. Retinex theory provides a physically meaningful way to model image formation, positing that human visual perception of color exhibits illumination invariance—that is, an object's color (reflectance property) remains unchanged regardless of environ-

 doi:10.20944/preprints202604.1600.v1

mental lighting variations. Based on this perceptual mechanism, a natural image can be decomposed into the element-wise product of two intrinsic components:

$$I(x,y) = R(x,y) \cdot L(x,y), \quad \forall (x,y) \in \Omega \tag{16}$$

where Ω denotes the image spatial domain, $R(x,y)$ is the reflectance map representing inherent surface properties of objects such as material, texture, and color, and $L(x,y)$ is the illumination map reflecting the intensity distribution of incident light in the scene and governing overall brightness variations. Ideally, by separating and appropriately adjusting L, image brightness can be globally or locally enhanced without altering the essential characteristics of objects.

To establish a model more closely aligned with real-world low-light imaging processes, this paper introduces a perturbation-aware Retinex decomposition framework. It assumes that the observed image I is a nonlinear product of ideal reflectance R and true illumination L under noise interference, whose generation process can be modeled as

$$I = (R + \varepsilon_r) \odot (L + \varepsilon_l) \tag{17}$$

where ε_r is the reflectance perturbation term, encompassing sensor noise, compression artifacts, and local reflectance abrupt changes, and ε_l is the illumination estimation error characterizing local non-uniformities in illumination distribution (e.g., spotlight edges and shadow transition zones). Expanding the above equation yields

$$I = R \odot L + R \odot \varepsilon_l + \varepsilon_r \odot L + \varepsilon_r \odot \varepsilon_l \tag{18}$$

In practical modeling, the illumination map L is typically assigned a smoothness prior due to its slowly varying nature in space. Therefore, the enhanced illumination can be estimated through the following transformation:

$$\begin{aligned} I_{\text{enh}} &= (R + \varepsilon_r) \odot (L + \varepsilon_l) \odot \overline{L} \\ &= (R + \varepsilon_r) \odot (L + \varepsilon_l) \cdot \frac{1}{\alpha L + \beta} \\ &= (R + \varepsilon_r) \odot L^*, \end{aligned} \tag{19}$$

where α is a stabilization hyperparameter, β prevents division by zero, and $\overline{L}$ can be regarded as a robust approximation of $1/L$.

To introduce high-level semantic guidance in the low-light image enhancement process and achieve dynamic fusion of multi-source information, this paper proposes a semantic prior fusion module based on a cross-attention mechanism, taking the original low-light image, illumination estimation map, and text-guided semantic attention map as inputs, and enabling cross-modal interaction and semantic alignment among the three through a cross-attention mechanism.

As shown in Figure 4, each input is first enhanced via an independent self-attention branch to capture long-range spatial dependencies within its respective modality. Subsequently, the features from the three branches are fed into a cross-attention module to enable cross-modal information interaction. During this process, different inputs can mutually "perceive" each other: for instance, image features can enhance the recovery of details in dark regions guided by the semantic attention map, while the illumination map can perform locally adaptive adjustments based on semantic cues. The fused features are combined with the original branch outputs via residual connections, preserving base information while incorporating high-level semantic guidance. This structure fully leverages the dynamic modeling capability of attention mechanisms, achieving effective alignment and fusion of multi-source information without introducing significant computational overhead, thereby providing a richer and more semantically consistent feature representation for the subsequent enhancement network.

 doi:10.20944/preprints202604.1600.v1

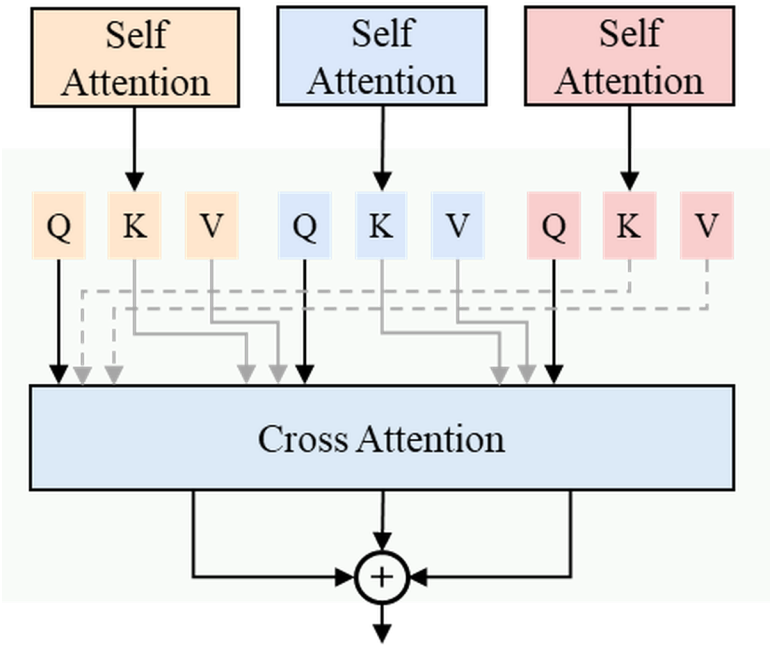

Figure 4. Structure of the Cross-Attention module.

3.5. Loss Function

To achieve high-quality enhancement of low-light images, this paper designs a composite loss function that integrates pixel fidelity, structural clarity, and color naturalness, consisting of reconstruction loss, edge-aware loss, and color consistency loss.

Reconstruction loss ensures high pixel-level consistency between the enhanced image I_{enh} and the ground-truth image I_{gt}:

$$\mathcal{L}_{\text{recon}} = \left\| I_{\text{enh}} - I_{\text{gt}} \right\|_1 + \left\| I_{\text{enh}} - I_{\text{gt}} \right\|_2^2 \tag{20}$$

Edge-aware loss is designed to address the issue of blurred edges and detail loss in low-light images. During the enhancement of dark areas, traditional methods often lead to structural degradation due to noise amplification or over-smoothing. To this end, we introduce an edge modeling mechanism based on the Sobel operator to extract image gradients:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{21}$$

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{22}$$

The edge intensity map is calculated as

$$E(I) = \sqrt{(I * G_x)^2 + (I * G_y)^2} \tag{23}$$

Edge-aware loss not only requires matching edge intensity but also further constrains the spatial gradient changes to maintain the sharpness and geometric consistency of edges:

$$\mathcal{L}_{\text{edge}} = \left\| E(I_{\text{enh}}) - E(I_{\text{gt}}) \right\|_2^2 + \left\| \nabla E(I_{\text{enh}}) - \nabla E(I_{\text{gt}}) \right\|_1 \tag{24}$$

 doi:10.20944/preprints202604.1600.v1

Color consistency loss is used to preserve the color statistical properties of images, preventing color bias or saturation distortion during enhancement. By constraining the consistency of mean μ and covariance matrix Σ in RGB space, it enhances visual naturalness:

$$\mathcal{L}_{\text{color}} = \left\|\mu(I_{\text{enh}}) - \mu(I_{\text{gt}})\right\|_2^2 + \left\|\Sigma(I_{\text{enh}}) - \Sigma(I_{\text{gt}})\right\|_F^2 \tag{25}$$

The loss function effectively takes into account structural integrity and color authenticity while ensuring brightness restoration and detail enhancement, providing a reliable optimization direction for the model to produce clear, natural, and visually superior enhancement results.

4. Experiments

4.1. Experimental Setup

This experiment is evaluated on several public low-light image enhancement datasets, including LOL dataset v1, LOL-v2-real, LOL-v2-synthetic, SID, SMID, and SDSD-out, which are widely used in the field for benchmarking low-light enhancement methods [37]. Specifically, the LOL dataset v1 contains 485 training image pairs and 15 testing image pairs, all resized to 400×600 resolution; LOL-v2 further provides two subsets: real-world scenes (LOL-v2-real) and synthetic scenes (LOL-v2-synthetic), used to evaluate model performance under real noise and ideal conditions, respectively, with 100 image pairs in each test set; the SID, SMID, and SDSD-out datasets are primarily used for further evaluating enhancement results under complex noise patterns and real-world degradation.

The model is implemented in the PyTorch framework and trained using the Adam optimizer with an initial learning rate of 2×10^{-4}, gradually decayed to 1×10^{-6} using a cosine annealing strategy. The batch size is set to 8, with 2.5×10^5 training iterations. During training, input images are randomly cropped to 128×128 to increase sample diversity, along with data augmentation via horizontal flipping and random rotation; during inference, full-resolution images are processed directly. The model is trained to convergence on an NVIDIA A100 40GB GPU.

The input low-light image is accompanied by a text prompt T, which is set as

$$T = \{\text{“dark objects”, “not natural appearance”, “noise suppression”}\} \tag{26}$$

PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) are adopted as quantitative evaluation metrics, measuring pixel-level reconstruction accuracy and structural information preservation, respectively.

PSNR is defined as

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{L^2}{\text{MSE}}\right) \tag{27}$$

where L is the maximum possible pixel value of the image, and

$$\text{MSE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(I(i,j) - \widehat{I}(i,j)\right)^2 \tag{28}$$

where MSE is the mean squared error between the ground-truth image I and the enhanced image $\widehat{I}$, with $M \times N$ being the image dimensions.

SSIM between two images I and $\widehat{I}$ is computed as

$$\text{SSIM}(I, \widehat{I}) = \frac{\left(2\mu_I\mu_{\widehat{I}} + C_1\right)\left(2\sigma_{I\widehat{I}} + C_2\right)}{\left(\mu_I^2 + \mu_{\widehat{I}}^2 + C_1\right)\left(\sigma_I^2 + \sigma_{\widehat{I}}^2 + C_2\right)} \tag{29}$$

where μ_I and $\mu_{\widehat{I}}$ are the local means, σ_I^2 and $\sigma_{\widehat{I}}^2$ are the variances, $\sigma_{I\widehat{I}}$ is the covariance, and C_1 and C_2 are stabilization constants. The SSIM value ranges from -1 to 1, with higher values indicating better structural fidelity.

 doi:10.20944/preprints202604.1600.v1

4.2. Experimental Results

4.2.1. Quantitative Comparison

Table 1 presents the quantitative performance comparison on the LOL-v1, LOL-v2-real, and LOL-v2-syn datasets. Our proposed SeMaNet achieves the highest scores across all metrics and datasets, demonstrating its superior restoration capability. On the LOL-v1 benchmark, SeMaNet attains a PSNR of 25.46 dB and an SSIM of 0.850, surpassing the previous state-of-the-art method RetinexFormer (25.16 dB, 0.845) and significantly outperforming CNN-based models such as MIRNet (24.14 dB, 0.830). Notably, SeMaNet also shows strong generalization on the more challenging LOL-v2 datasets: it achieves 22.89 dB (PSNR) and 0.852 (SSIM) on LOL-v2-real, and 25.71 dB (PSNR) and 0.938 (SSIM) on LOL-v2-syn, setting new benchmarks. The consistent improvement over RetinexFormer and SNR-Net highlights the effectiveness of our hybrid Transformer-Mamba architecture and semantic guidance in preserving structural fidelity and enhancing perceptual quality under diverse lighting conditions.

Table 1. Quantitative comparisons on LOL v1, LOL-v2-real and LOL-v2-syn.

Methods	LOL-v1		LOL-v2-real		LOL-v2-syn	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SID [38]	14.35	0.436	13.24	0.442	15.04	0.610
DeepUPE [39]	14.38	0.446	13.27	0.452	15.08	0.623
DeepLPF [40]	15.28	0.473	14.10	0.480	16.02	0.587
UFormer [41]	16.36	0.771	18.82	0.771	19.66	0.871
RetinexNet [13]	16.77	0.560	15.47	0.567	17.13	0.798
EnGAN [17]	17.48	0.650	18.23	0.617	16.57	0.734
RUAS [42]	18.23	0.720	18.37	0.723	16.55	0.652
FIDE [18]	18.27	0.665	16.85	0.678	15.20	0.612
KinD [14]	20.86	0.790	14.74	0.641	13.29	0.578
Restormer [21]	22.43	0.823	19.94	0.827	21.41	0.830
MIRNet [43]	24.14	0.830	20.02	0.820	21.94	0.876
SNR-Net [44]	24.61	0.842	21.48	0.849	24.14	0.928
RetinexFormer [22]	25.16	0.845	22.80	0.840	25.67	0.930
SeMaNet	**25.46**	**0.850**	**22.89**	**0.852**	**25.71**	**0.938**

Further evaluation on the SID, SMID, and SDSD-out datasets (Table 2) reinforces the robustness of SeMaNet. On the SID dataset, our method achieves 24.55 dB in PSNR and 0.687 in SSIM, outperforming all compared methods including the strong baseline Restormer (22.27 dB, 0.649) and RetinexFormer (24.44 dB, 0.680). On SMID and SDSD-out—datasets characterized by complex noise patterns and real-world degradation—SeMaNet achieves 29.35 dB (PSNR, 0.826 SSIM) and 29.88 dB (PSNR, 0.883 SSIM), respectively, ranking first in all metrics. These results validate that the perturbation-aware Retinex decomposition and CLIP-guided semantic attention enable SeMaNet to effectively handle real-world noise while maintaining high structural consistency and color accuracy. The marginal but consistent gains over RetinexFormer further confirm the benefits of integrating long-range modeling with efficient sequential processing under semantic guidance.

Table 2. Quantitative comparisons on SID, SMID and SDSD-out.

Methods	SID		SMID		SDSD-out	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SID [38]	16.97	0.591	24.78	0.718	24.90	0.693
DeepUPE [39]	17.01	0.604	23.91	0.690	21.94	0.698
DeepLPF [40]	18.07	0.600	24.36	0.688	22.76	0.658
UFormer [41]	18.54	0.577	27.20	0.792	23.85	0.748
RetinexNet [13]	16.48	0.578	22.83	0.684	20.96	0.629
EnGAN [17]	17.23	0.543	22.62	0.674	20.10	0.616
RUAS [42]	18.44	0.581	25.88	0.744	23.84	0.743
FIDE [18]	18.34	0.578	24.42	0.692	22.20	0.629

 doi:10.20944/preprints202604.1600.v1

Table 2. *Cont.*

Methods	SID		SMID		SDSD-out	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
KinD [14]	18.02	0.583	22.18	0.634	21.97	0.654
Restormer [21]	22.27	0.649	26.97	0.758	24.79	0.802
MIRNet [43]	20.84	0.605	25.66	0.762	27.13	0.837
SNR-Net [44]	22.87	0.625	28.49	0.805	28.66	0.866
RetinexFormer [22]	24.44	0.680	29.15	0.815	29.84	0.877
SeMaNet	**24.55**	**0.687**	**29.35**	**0.826**	**29.88**	**0.883**

4.2.2. Qualitative Comparison

Figures 5–7 present qualitative comparisons of low-light enhancement results from different methods on the LOL-v1, LOL-v2-real, and LOL-v2-syn datasets, respectively. In the indoor swimming pool scene from LOL-v1, most methods improve visibility to some extent, yet several suffer from over-enhancement or halo artifacts around bright regions. SeMaNet, by contrast, preserves the clarity of fine details such as the digital clock and glass surfaces while achieving smooth and natural lighting transitions, demonstrating its ability to recover brightness without compromising structural fidelity. In the kitchen scene from LOL-v2-real, which features smooth surfaces and small objects under uneven illumination, many competing methods either amplify noise on the tabletop or introduce color distortions. SeMaNet effectively suppresses noise in homogeneous regions while maintaining sharp edges and color consistency, resulting in a more realistic and visually clean restoration. Finally, in the outdoor backlit scene from LOL-v2-syn, many approaches struggle to recover details in shadowed regions—particularly on the person—or overexpose the sky. SeMaNet successfully enhances the subject's silhouette and fabric texture without sacrificing background details, preserving both contrast and semantic coherence. This underscores its superior performance in handling complex, real-world lighting conditions.

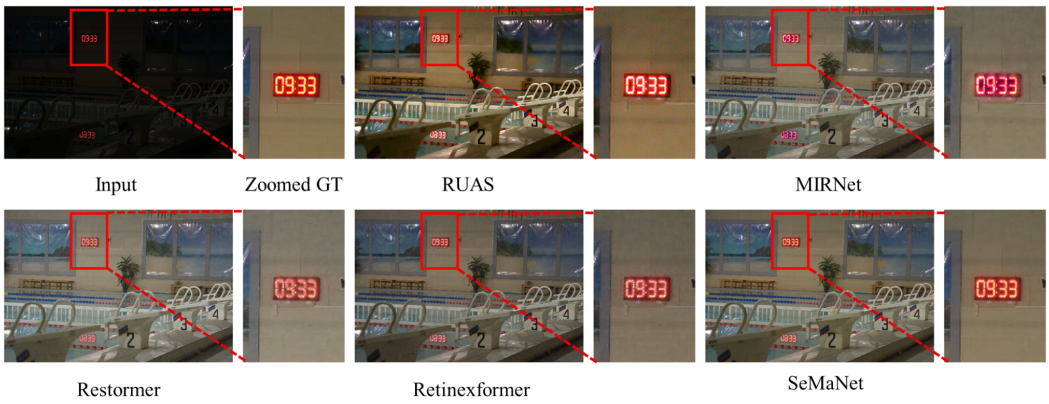

Figure 5. Comparison of low-light enhancement results using different methods on the LOL-v1 dataset.

Figure 8 shows the semantic attention heatmaps generated by our CLIP-guided approach, clearly revealing how the model dynamically focuses on semantically meaningful key regions during the low-light image enhancement process. The heatmaps indicate strong attention responses from the network in areas such as cups, washing machines, cabinets, and human faces—objects that typically hold high visual importance in indoor scenes. This demonstrates that the CLIP-based semantic prior effectively helps the model distinguish between foreground objects and background regions, enabling adaptive enhancement strategies tailored to specific semantic areas. For instance, the strong attention activation around the cup and washing machine suggests that the model prioritizes preserving texture details and structural integrity within these object-centric regions. By leveraging the cross-modal image-text alignment capability provided by CLIP, our method avoids the over-saturation and loss of fine details commonly caused by uniform global enhancement in traditional approaches. Instead,

 doi:10.20944/preprints202604.1600.v1

it achieves scene-aware local illumination adjustment, ensuring that visually salient regions receive appropriate exposure while less critical areas are processed more conservatively. This semantic-aware mechanism significantly improves visual quality in complex scenes, effectively preventing common artifacts such as over-brightening and color distortion seen in non-semantic methods like RetinexNet.

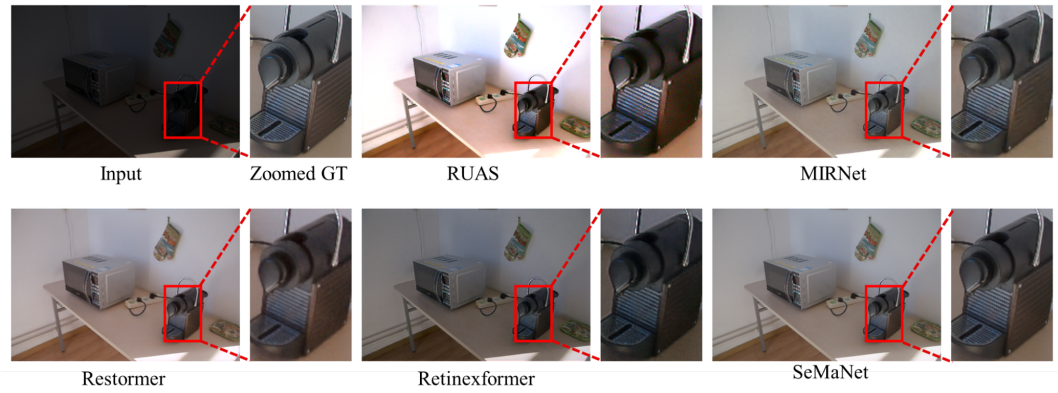

Figure 6. Comparison of low-light enhancement results using different methods on the LOL-v2-real dataset.

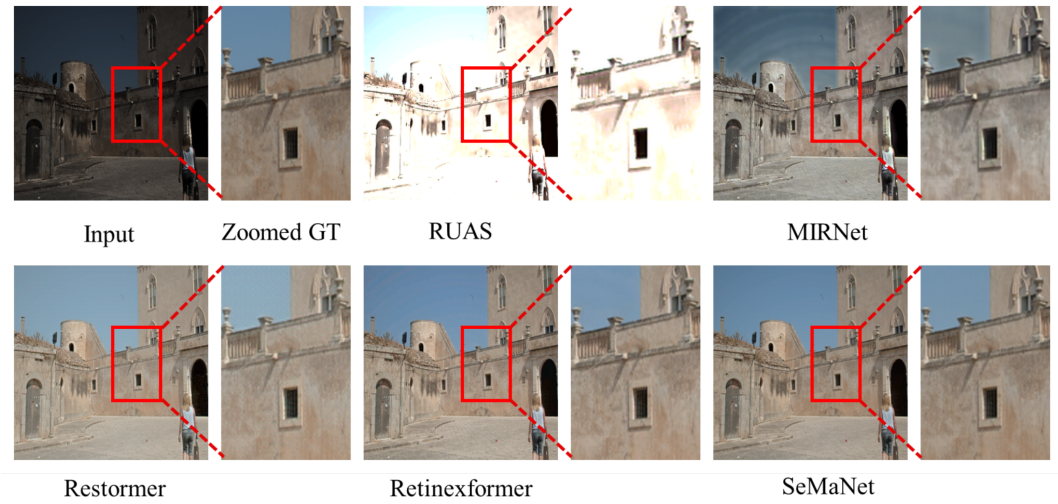

Figure 7. Comparison of low-light enhancement results using different methods on the LOL-v2-syn dataset.

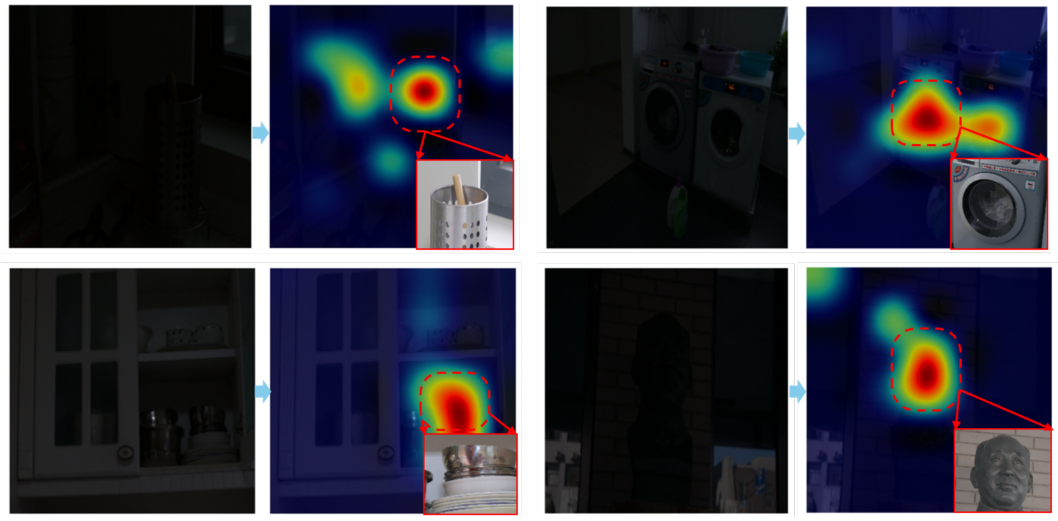

Figure 8. Visualization of CLIP-based semantic-guided attention regions.

 doi:10.20944/preprints202604.1600.v1

4.2.3. Ablation Study

To systematically evaluate the contribution of each proposed component in SeMaNet, comprehensive ablation experiments are conducted on both LOL-v2-real and LOL-v2-syn datasets. A strong baseline model is established following a typical Retinex-based decomposition-enhancement paradigm, comprising a U-Net backbone with Transformer blocks integrated into the encoder-decoder for global context modeling. Subsequently, the key components are incrementally incorporated: (1) the hybrid Transformer-Mamba (H-TM) backbone, (2) the edge-aware loss (EAL), and (3) the CLIP-guided semantic prior (CLIP-SP). All variants are trained under identical experimental configurations to ensure fair comparison.

The quantitative results are summarized in Table 3. The baseline model achieves 21.42 dB PSNR and 0.802 SSIM on LOL-v2-real, and 24.03 dB PSNR and 0.901 SSIM on LOL-v2-syn. Replacing the decoder's Transformer blocks with Mamba layers to form the hybrid Transformer-Mamba (H-TM) architecture brings significant improvements, increasing PSNR to 22.15 dB and SSIM to 0.826 on real data, and to 24.88 dB and 0.920 on synthetic data. This demonstrates that Mamba enhances long-range feature propagation and detail recovery, particularly under real-world noise.Further gains are achieved by introducing the edge-aware loss (EAL), which improves structural fidelity. PSNR rises to 22.56 dB and SSIM to 0.841 on LOL-v2-real, and to 25.32 dB and 0.931 on LOL-v2-syn, confirming the effectiveness of EAL in preserving edges and fine textures. The full SeMaNet model integrates the CLIP-guided semantic prior (CLIP-SP), enabling semantically aware enhancement. It achieves the highest performance: 22.89 dB PSNR and 0.852 SSIM on LOL-v2-real, and 25.71 dB PSNR and 0.938 SSIM on LOL-v2-syn—matching the results in Table 1. The consistent improvements validate the complementary roles of each component. The integration of CLIP-derived semantics allows the model to adaptively modulate enhancement strength based on scene content, such as preserving natural skin tones and suppressing noise in uniform regions. This semantic awareness, combined with efficient long-range modeling (H-TM) and structural regularization (EAL), is essential to SeMaNet's state-of-the-art performance.

Table 3. Ablation study on LOL-v2-real and LOL-v2-syn.

Components	LOL-v2-real		LOL-v2-syn	
	PSNR	SSIM	PSNR	SSIM
Baseline	21.42	0.802	24.03	0.901
+ H-TM	22.15	0.826	24.88	0.920
+ H-TM + EAL	22.56	0.841	25.32	0.931
+ H-TM + EAL + CLIP-SP (SeMaNet)	22.89	0.852	25.71	0.938

5. Conclusion

This paper presents SeMaNet, a semantic-guided framework for low-light image enhancement that unifies vision-language understanding, an efficient hybrid architecture, and a physically informed Retinex model. By harnessing pre-trained CLIP, SeMaNet interprets natural language prompts—such as "brighten the face" or "reduce sky noise"—to generate spatially adaptive attention maps. This enables intuitive, user-controllable enhancement without relying on pixel-level labels. These semantic cues are seamlessly integrated with illumination estimates and image features through cross-attention, ensuring that restoration aligns with both scene semantics and lighting physics.

The backbone combines Transformers for global context and Mamba blocks for efficient sequential modeling within a U-Net structure. This design preserves fine details while significantly reducing computational overhead compared to pure Transformer-based methods. A perturbation-aware Retinex component further accounts for real-world imperfections like sensor noise and uneven lighting, yielding more realistic results. Experiments across multiple low-light benchmarks show that SeMaNet consistently outperforms existing methods in both visual quality and quantitative evaluation. It achieves leading performance on both synthetic and real-world datasets, with clear gains attributed to

 doi:10.20944/preprints202604.1600.v1

each key component—semantic guidance, hybrid architecture, and edge-aware optimization. Visually, the method avoids common pitfalls such as over-enhancement or texture loss, particularly in sensitive regions like faces or smooth surfaces. Beyond low-light enhancement, SeMaNet demonstrates how high-level semantic reasoning can be effectively embedded into low-level vision pipelines. The framework opens avenues for interpretable, controllable restoration and can be extended to video processing, edge deployment, and other tasks where efficiency and semantic awareness matter.

Author Contributions: Conceptualization, T.J. and S.W.; methodology, T.J. and H.P.; software, T.J.; validation, T.J. and H.P.; formal analysis, T.J.; writing—original draft preparation, T.J.; writing—review and editing, S.W. and Y.Z.; supervision, S.W. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available. LOL v1, LOL-v2-real, LOL-v2-synthetic, SID, SMID, and SDSD-out can be accessed from the corresponding public sources cited in the manuscript. The source code and additional evaluation results will be made publicly available upon acceptance of the manuscript.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLIE	Low-Light Image Enhancement
CLIP	Contrastive Language–Image Pre-training
Grad-CAM	Gradient-weighted Class Activation Mapping
SSM	State Space Model
FFN	Feed-Forward Neural Network
MSE	Mean Squared Error
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure

References

1. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223. [CrossRef]
2. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
3. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; Seattle, WA, USA, 13–19 June 2020; pp. 1777–1786. [CrossRef]
4. Zhao, L.; Wang, K.; Zhang, J.; Wang, A.; Bai, H. Learning Deep Texture-Structure Decomposition for Low-Light Image Restoration and Enhancement. *Neurocomputing* **2023**, *524*, 126–141. [CrossRef]
5. Wang, W.; Wu, X.; Yuan, X.; Gao, Z. An Experiment-Based Review of Low-Light Image Enhancement Methods. *IEEE Access* **2020**, *8*, 87884–87917. [CrossRef]
6. Pizer, S.M.; Amburn, E.P.; Austin, J.D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J.B.; Zuiderveld, K. Adaptive Histogram Equalization and Its Variations. *Comput. Vis. Graph. Image Process.* **1987**, *39*, 355–368. [CrossRef]
7. Rahman, S.; Rahman, M.M.; Abdullah-Al-Wadud, M.; Al-Quaderi, G.D.; Shoyaib, M. An Adaptive Gamma Correction for Image Enhancement. *EURASIP J. Image Video Process.* **2016**, *2016*, 35. [CrossRef]
8. Land, E.H.; McCann, J.J. Lightness and Retinex Theory. *J. Opt. Soc. Am.* **1971**, *61*, 1–11. [CrossRef]

 doi:10.20944/preprints202604.1600.v1

9. Land, E.H. The Retinex Theory of Color Vision. *Sci. Am.* **1977**, *237*, 108–128. [CrossRef]
10. Jobson, D.J.; Rahman, Z.-u.; Woodell, G.A. A Multiscale Retinex for Bridging the Gap between Color Images and the Human Observation of Scenes. *IEEE Trans. Image Process.* **1997**, *6*, 965–976. [CrossRef]
11. Jobson, D.J.; Rahman, Z.; Woodell, G.A. Properties and Performance of a Center/Surround Retinex. *IEEE Trans. Image Process.* **1997**, *6*, 451–462. [CrossRef]
12. Cai, J.; Gu, S.; Zhang, L. Learning a Deep Single Image Contrast Enhancer from Multi-Exposure Images. *IEEE Trans. Image Process.* **2018**, *27*, 2049–2062. [CrossRef]
13. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep Retinex Decomposition for Low-Light Enhancement. *arXiv* **2018**, arXiv:1808.04560. [arXiv]
14. Zhang, Y.; Zhang, J.; Guo, J.; Chen, X.; Chen, X.; Zhu, X. Kindling the Darkness: A Practical Low-Light Image Enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM)*; Nice, France, 21–25 October 2019; pp. 1635–1643. [CrossRef]
15. Liu, J.; Xu, D.; Yang, W.; Fan, M.; Huang, H. Benchmarking Low-Light Image Enhancement and Beyond. *Int. J. Comput. Vis.* **2021**, *129*, 1153–1184. [CrossRef]
16. Li, C.; Guo, C.; Loy, C.C. Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation. *arXiv* **2021**, arXiv:2103.00860. [arXiv]
17. Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; Wang, Z.; Lin, L. EnlightenGAN: Deep Light Enhancement without Paired Supervision. *IEEE Trans. Image Process.* **2021**, *30*, 2340–2349. [CrossRef]
18. He, Z.; Ran, W.; Liu, S.; Li, K.; Lu, J.; Xie, C.; Liu, Y.; Lu, H. Low-Light Image Enhancement with Multi-Scale Attention and Frequency-Domain Optimization. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 2861–2873. [CrossRef]
19. Chen, L.; Chu, X.; Zhang, X.; Sun, J. Simple Baselines for Image Restoration. In *Computer Vision – ECCV 2022*; Tel Aviv, Israel, 23–27 October 2022; pp. 17–33. [CrossRef]
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [arXiv]
21. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.-H. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; New Orleans, LA, USA, 19–24 June 2022; pp. 5718–5729. [CrossRef]
22. Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; Zhang, Y. Retinexformer: One-Stage Retinex-Based Transformer for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; Paris, France, 2–6 October 2023; pp. 12470–12479. [CrossRef]
23. Wang, T.; Zhang, K.; Shen, T.; Luo, W.; Stenger, B.; Lu, T. Ultra-High-Definition Low-Light Image Enhancement: A Benchmark and Transformer-Based Method. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 2974–2982. [CrossRef]
24. Gu, A.; Goel, K.; Kumar, A.; et al. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752. [arXiv]
25. Zhu, L.; Hou, M.R.; Wang, Y.; et al. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv* **2024**, arXiv:2401.09417. [arXiv]
26. Archit, A.; Pape, C. ViM-UNet: Vision Mamba for Biomedical Segmentation. *arXiv* **2024**, arXiv:2404.07705. [arXiv]
27. Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; Qiao, Y. VideoMamba: State Space Model for Efficient Video Understanding. In *Computer Vision – ECCV 2024*; Milan, Italy, 29 September–4 October 2024; pp. 237–255. [CrossRef]
28. Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; Xia, S.-T. MambaIR: A Simple Baseline for Image Restoration with State-Space Model. In *Computer Vision – ECCV 2024*; Milan, Italy, 29 September–4 October 2024; pp. 222–241. [CrossRef]
29. Bai, J.; Yin, Y.; He, Q.; Li, Y.; Zhang, X. Retinexmamba: Retinex-Based Mamba for Low-Light Image Enhancement. In *Neural Information Processing*; 31st International Conference, ICONIP 2024, Auckland, New Zealand, 2–6 December 2024, Proceedings, Part VIII; Springer: Singapore, 2025; pp. 427–442. [CrossRef]
30. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020. [arXiv]

 doi:10.20944/preprints202604.1600.v1

31. Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; Lischinski, D. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; Montreal, QC, Canada, 11–17 October 2021; pp. 2065–2074. [CrossRef]
32. Xu, D.; Zhu, Y.; Choy, C.B.; Fei-Fei, L. Scene Graph Generation by Iterative Message Passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; Honolulu, HI, USA, 21–26 July 2017; pp. 5410–5419. [CrossRef]
33. Ramesh, A.; Dhariwal, P.; Nichol, A.; et al. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* **2022**, arXiv:2204.06125. [arXiv]
34. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef]
35. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation with Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; Venice, Italy, 22–29 October 2017; pp. 2242–2251. [CrossRef]
36. Anaya, J.; Barbu, A. RENOIR: A Dataset for Real Low-Light Image Noise Reduction. *J. Vis. Commun. Image Represent.* **2018**, *51*, 144–154. [CrossRef]
37. Zheng, S.; Ma, Y.; Pan, J.; Lu, C.; Gupta, G. Low-Light Image and Video Enhancement: A Comprehensive Survey and Beyond. *arXiv* **2022**, arXiv:2212.10772. [arXiv]
38. Chen, C.; Chen, Q.; Xu, J.; Koltun, V. Learning to See in the Dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; Salt Lake City, UT, USA, 18–22 June 2018; pp. 3291–3300. [CrossRef]
39. Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; Jia, J. Underexposed Photo Enhancement Using Deep Illumination Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; Long Beach, CA, USA, 16–20 June 2019; pp. 6849–6857. [CrossRef]
40. Moran, S.; Marza, P.; McDonagh, S.; Parisot, S.; Slabaugh, G. DeepLPF: Deep Local Parametric Filters for Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; Seattle, WA, USA, 13–19 June 2020; pp. 12823–12832. [CrossRef]
41. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A General U-Shaped Transformer for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; New Orleans, LA, USA, 19–24 June 2022; pp. 17662–17672. [CrossRef]
42. Liu, R.; Ma, L.; Zhang, J.; Fan, X.; Luo, Z. Retinex-Inspired Unrolling with Cooperative Prior Architecture Search for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; Virtual Event, 19–25 June 2021; pp. 10561–10570. [CrossRef]
43. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.-H.; Shao, L. Learning Enriched Features for Real Image Restoration and Enhancement. In *Computer Vision – ECCV 2020*; Glasgow, UK, 23–28 August 2020; pp. 492–511. [CrossRef]
44. Xu, X.; Wang, R.; Fu, C.-W.; Jia, J. SNR-Aware Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; New Orleans, LA, USA, 19–24 June 2022; pp. 17693–17703. [CrossRef]