RNMT-Net: A Retinex-Guided NAFNet–Mamba–Transformer Network for Low-Light Image Enhancement

Lakshmi Narayana Buragala
Department of Artificial Intelligence and Data Science
Kallam Haranadhareddy Institute of Technology
lakshminarayanaburugula@gmail.com

***Abstract*—**Low-light image enhancement (LLIE) remains challenging because a single network must simultaneously correct global illumination, suppress sensor noise, and preserve fine texture and color fidelity. Existing approaches typically specialize in only one of these sub-problems: Retinex-based methods model illumination well but under-model noise; convolutional restoration networks such as NAFNet are efficient but lack long-range receptive fields; and Transformer- or Mamba-based restorers capture global context but are rarely combined with an explicit noise model or a principled multi-task loss balance. We propose RNMT-Net, a unified encoder–decoder architecture that fuses (i) Nonlinear-Activation-Free (NAFNet) convolutional blocks for efficient local feature extraction, (ii) a cross-scan Selective State-Space (Mamba) module for linear-complexity long-range modeling, (iii) a lightweight channel-wise Transformer block, (iv) a Signal-to-Noise-Ratio (SNR)-guided local/global fusion gate, and (v) an amplitude–phase Fourier enhancement module, all supervised through a Retinex-inspired reflectance–illumination–noise decomposition with automatic uncertainty-based loss weighting and a progressive patch-size curriculum. RNMT-Net is trained in two stages — synthetic pre-training on LOL-v2-Synthetic followed by fine-tuning on LOL-v1 — and evaluated with exponential moving average (EMA) weights and flip-based test-time augmentation. On the LOL-v1 benchmark, RNMT-Net achieves 26.26 dB PSNR and 0.8468 SSIM, outperforming representative Retinex-based, CNN-based, and Transformer-based baselines reported in the literature. Beyond PSNR/SSIM, we additionally report LPIPS and NIQE to assess perceptual quality and naturalness. Code is available at https://github.com/Lachi9959/rnmt-net-llie.

***Index Terms*—**Low-light image enhancement, Retinex theory, state-space models, Mamba, NAFNet, Transformer, image denoising, deep learning.

I. INTRODUCTION

Images captured under low-light conditions suffer from a compound set of degradations: reduced visibility due to insufficient photons, amplified sensor noise from the high ISO gains required to compensate for short exposure, color distortion, and loss of fine structural detail. Low-light image enhancement (LLIE) aims to recover a visually pleasing, artifact-free, normally-illuminated image from such a degraded observation, and it underlies many downstream applications, including nighttime photography, autonomous driving, and surveillance.

Two broad families of solutions have emerged. Retinex-theory-based methods [5] decompose an image into a reflectance component, assumed to be illumination-invariant, and an illumination component that is corrected and re-applied. Methods such as RetinexNet [6], KinD [12], and Retinexformer [4] follow this decomposition explicitly, but classical Retinex formulations implicitly assume the input is corruption-free, which is inconsistent with real noisy low-light captures. Purely data-driven restoration networks, on the other hand, learn a direct mapping from degraded to clean images. Convolutional architectures such as NAFNet [2] demonstrate that a carefully simplified, nonlinear-activation-free block can match or exceed the performance of far more complex attention-based restorers, at a fraction of the computational cost — but a purely convolutional receptive field is inherently local and struggles to reason about long-range illumination gradients that are common in unevenly-lit scenes.

Recently, state-space sequence models — most notably Mamba [1] — have been shown to match Transformer-level modeling quality while scaling linearly rather than quadratically with sequence length, thanks to an input-dependent (selective) state-space recurrence and a hardware-aware parallel scan. Vision adaptations of Mamba, such as RAWMamba's eight-direction scanning mechanism for RAW-domain restoration [3], indicate that carefully designed scan orders can substantially improve the ability of state-space models to capture two-dimensional spatial structure that a naive 1-D scan would miss. In parallel, Transformer-based low-light restorers such as Retinexformer [4] show that illumination-guided self-attention, when made linear in spatial size, can be tightly integrated with a Retinex decomposition to great effect.

Motivated by these three complementary lines of work, we ask whether a single network can jointly exploit (a) the computational efficiency and strong local modeling of NAFNet's gated convolutional blocks, (b) the linear-complexity global context modeling of a selective state-space (Mamba) scan, and (c) explicit Retinex- and noise-aware supervision, while remaining trainable end-to-end with a stable, automatically-balanced multi-term loss. We present RNMT-Net (Retinex–NAFNet–Mamba–Transformer Network) to answer this question.

Our design makes the following contributions:

1) A hybrid encoder–decoder backbone that combines NAFNet blocks for local feature extraction with a cross-scan Mamba module and a compact channel-attention

Transformer block at the bottleneck, giving the network both an efficient local prior and a linear-complexity global receptive field.

2) An SNR-guided local/global fusion gate that spatially blends the Mamba/Transformer global feature with a locally-refined NAFNet feature according to a per-pixel signal-to-noise estimate, and an amplitude–phase Fourier enhancement module that refines frequency-domain amplitude while preserving phase, targeting illumination-related low-frequency artifacts.

3) A Retinex-inspired reflectance–illumination–noise decomposition head supervised by an automatically-balanced, uncertainty-weighted multi-term loss [18], together with a two-phase, patch-size-curriculum training schedule (synthetic pre-training followed by real-data fine-tuning at progressively larger patch sizes) and flip-based test-time augmentation with an EMA teacher.

We evaluate RNMT-Net on the widely used LOL-v1 benchmark, pre-training on LOL-v2-Synthetic before fine-tuning, and report PSNR and SSIM against a broad set of published Retinex-based, CNN-based, and Transformer-based baselines. RNMT-Net attains 26.26 dB PSNR and 0.8468 SSIM, exceeding all compared baselines reported in the literature on this benchmark (Section V). We further report perceptual quality metrics (LPIPS, NIQE) in Section V-C to complement the standard distortion-based PSNR/SSIM evaluation.

The remainder of this paper is organized as follows. Section II reviews related work in Retinex-based, CNN-based, Transformer-based, and Mamba-based low-light restoration. Section III details the RNMT-Net architecture and training methodology. Section IV describes the experimental setup. Section V reports quantitative and qualitative results together with perceptual quality metrics. Section VI concludes the paper and discusses limitations and future work.

II. RELATED WORK

A. Retinex-Based Low-Light Enhancement

The Retinex theory [5] models an observed image I as the element-wise product of a reflectance map R and an illumination map L, i.e., $I = R \odot L$, where reflectance is assumed to be an illumination-invariant material property. RetinexNet [6] first combined this decomposition with deep CNNs, learning separate sub-networks to estimate reflectance and illumination and to denoise the reflectance branch. KinD [12] and its successors refine this idea with better-supervised decomposition networks, while RUAS [11] searches for a compact unrolled architecture using Retinex-inspired priors. A limitation shared by these methods is a multi-stage training pipeline in which sub-networks are optimized somewhat independently before joint fine-tuning, and an implicit assumption that the input is corruption-free, which does not hold for extreme low-light captures corrupted by strong sensor noise.

Retinexformer [4] addresses the corruption-free assumption directly by introducing perturbation terms for both reflectance and illumination, deriving a one-stage Retinex-based framework (ORF) trained end-to-end, and pairing it with an Illumination-Guided Multi-head Self-Attention (IG-MSA) module whose computational complexity is linear rather than quadratic in spatial size. RNMT-Net adopts a similar spirit — an explicit reflectance/illumination/noise decomposition supervised end-to-end — but replaces the illumination-guided attention with a hybrid NAFNet–Mamba–Transformer backbone and adds an SNR-guided fusion gate and frequency-domain refinement not present in Retinexformer.

B. Efficient Convolutional Restoration Networks

NAFNet [2] demonstrates that a restoration block need not contain any nonlinear activation function at all: a Simplified Channel Attention module and a SimpleGate (element-wise product of a channel-split feature) can replace GELU and standard channel attention without any loss of accuracy, while reducing intra-block complexity. NAFNet achieves state-of-the-art PSNR on SIDD denoising (40.30 dB) and GoPro deblurring (33.69 dB) with only a fraction of the computational cost of comparable Transformer-based restorers such as Restormer [8]. RNMT-Net adopts NAFNet blocks as its convolutional backbone for both the encoder and decoder stages, exploiting this favorable accuracy-per-FLOP trade-off while reserving global context modeling for a dedicated bottleneck module.

C. Transformer-Based Restoration

Vision Transformers offer a global receptive field but incur quadratic computational complexity in the number of spatial tokens. Restormer [8] mitigates this by computing self-attention across the channel dimension instead of the spatial dimension, while MIRNet [9] and SNR-Net [10] use multi-scale feature fusion and SNR-aware attention, respectively, to selectively combine local and global information for low-light restoration. RNMT-Net's bottleneck Transformer block similarly computes attention over channels (following the Restormer-style formulation) to keep computational cost low, but restricts it to the lowest-resolution feature map and complements it with a dedicated Mamba scan for additional long-range modeling capacity.

D. State-Space Models and Mamba in Vision

Structured state-space models (SSMs) evolved from S4 [1] into Mamba, which introduces an input-dependent (selective) parameterization of the recurrence together with a hardware-aware parallel scan, enabling the model to selectively propagate or discard information along a sequence while retaining linear-time complexity. Mamba matches or exceeds Transformer-quality results on language, audio, and genomic

sequence modeling while offering up to 5× higher inference throughput [1]. Applying Mamba to two-dimensional images requires an explicit scanning strategy to linearize the spatial grid; RAWMamba [3] shows that a naive four-direction raster scan leaves discontinuities at row/column boundaries, and instead proposes an eight-direction scanning mechanism tailored to the structure of a RAW Color Filter Array, together with a Retinex Decomposition Module for RAW-domain low-light enhancement. VMamba and related works introduce four-direction Cross-Scan Modules for general vision tasks.

RNMT-Net incorporates a lightweight cross-scan Mamba module (MambaCrossScan) at the network bottleneck, using depth-wise convolution combined with row-wise and column-wise gating to approximate the benefit of directional scanning at substantially lower implementation complexity than a full multi-directional selective scan, making it practical to train end-to-end alongside the remaining modules on a single-GPU budget.

E. Multi-Task Loss Balancing and Curriculum Training

When a network is supervised by several loss terms of different scale and difficulty (e.g., pixel fidelity, structural similarity, reflectance consistency, illumination smoothness, and noise sparsity), manually tuning fixed scalar weights is brittle. Kendall et al. [18] propose weighting each task's loss by a learned homoscedastic uncertainty term, which we adopt (with additional clamping for training stability) to automatically balance the five loss terms used to supervise RNMT-Net. Curriculum learning over patch size — training first on small patches for efficiency and progressively increasing patch size — has likewise been shown to stabilize and accelerate convergence in image restoration; RNMT-Net applies a three-stage patch curriculum (128 → 160 → 192 px) during fine-tuning.

III. PROPOSED METHOD

A. Overview

Given a low-light RGB image $X \in \mathbb{R}^{H\times W\times 3}$, RNMT-Net predicts an enhanced image $\hat{Y}$ through an encoder–decoder architecture with a hybrid convolutional / state-space / attention bottleneck, followed by a Retinex-inspired decomposition head. Fig. 1 illustrates the overall pipeline. At a high level, the network computes

$$R = \sigma(\mathrm{Conv}(F\_dec)),\quad I = \sigma(\mathrm{Conv}(F\_dec)),\quad N = 0.1\cdot\tanh(\mathrm{Conv}(F\_dec))$$

$$\hat{Y} = \mathrm{clamp}(R \odot I + N, 0, 1), \quad (1)$$

where F_dec is the final decoder feature map, R and I are the estimated reflectance and illumination components in the sense of Eq. (1) of the Retinex model, and N is an explicit, bounded, additive noise-residual term. Bounding N to [−0.1, 0.1] via a scaled tanh prevents the noise branch from dominating the prediction while still allowing it to correct fine residual artifacts that the multiplicative R ⊙ I term cannot represent. The full pipeline is trained end-to-end; R, I, and N are never supervised with independent ground truth but emerge from the composite loss described in Section III-G.

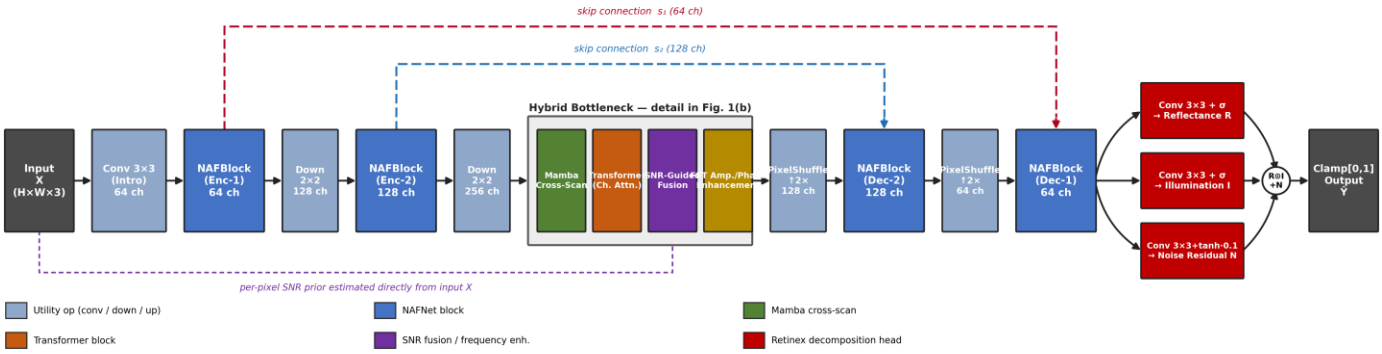

Fig. 1(a). Overall architecture of RNMT-Net: NAFNet encoder–decoder with a hybrid Mamba/Transformer/SNR/FFT bottleneck and a Retinex-inspired reflectance–illumination–noise decomposition head.

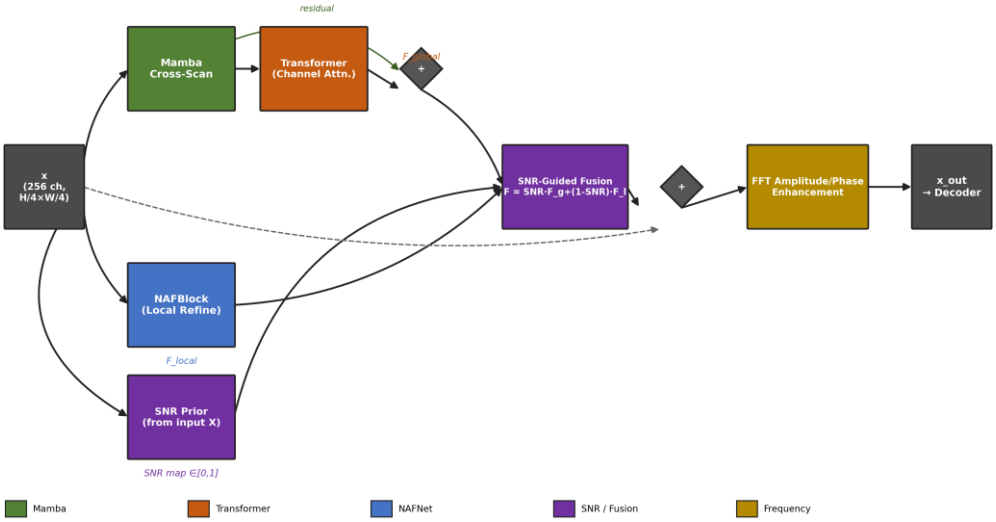

Fig. 1(b). Detailed view of the hybrid bottleneck: the global path (Mamba cross-scan + residual Transformer channel-attention) and local path (NAFBlock) are combined by the SNR-guided fusion gate, added residually to the input feature, and refined by the amplitude/phase frequency-enhancement module.

B. NAFNet Encoder–Decoder Backbone

The encoder begins with a 3×3 convolution that lifts the input to $C = 64$ channels, followed by a Nonlinear-Activation-Free block (NAFBlock) [2] operating at full resolution. Each NAFBlock replaces conventional activation functions with a SimpleGate — an element-wise product of two channel-split halves of a depth-wise-convolved feature — and augments it with a Simplified Channel Attention (SCA) module, all wrapped in a pre-normalization (LayerNorm2d) residual structure with learnable per-branch scaling coefficients (β, γ). Two strided 2×2 convolutions progressively halve spatial resolution while doubling channel width (64 → 128 → 256), with a NAFBlock applied after each downsampling step, yielding encoder features s_0 (64 ch, full-res), s_1 (128 ch, after first NAFBlock), and s_2 (256 ch, after second NAFBlock). The decoder mirrors this structure: PixelShuffle-based upsampling blocks (1×1 convolution to expand channels, followed by pixel-shuffle by a factor of 2) restore spatial resolution, and NAFBlocks are applied after each upsampling stage to feature maps formed by adding the corresponding encoder skip connection (s_1, s_2), following the classical U-Net skip-connection pattern.

C. Mamba Cross-Scan Bottleneck Module

At the lowest-resolution bottleneck (256 channels, H/4 × W/4), RNMT-Net applies a lightweight state-space-inspired module, MambaCrossScan, before the Transformer block. Rather than implementing a full multi-directional selective scan as in RAWMamba's eight-direction mechanism [3], MambaCrossScan approximates directional, input-dependent gating with a depth-wise 3×3 convolution followed by two 1-D pooling-and-sigmoid gates computed independently along the vertical and horizontal axes:

$$g\_v = \sigma(\text{mean\_W}(\text{Conv\_dw}(x))),\ g\_h = \sigma(\text{mean\_H}(\text{Conv\_dw}(x))),$$

$$\text{MambaScan}(x) = \text{Conv}_{1\times1}(\text{Conv\_dw}(x) \odot g\_v \odot g\_h). \quad (2)$$

This design retains the core intuition of selective state-space scanning — content-dependent, direction-aware gating of feature propagation with linear computational cost in the number of spatial locations — while remaining inexpensive enough to train jointly with the rest of the network. We discuss the relationship to full selective-scan SSMs and directions for a higher-fidelity Mamba block in Section VI.

D. Channel-Attention Transformer Block

Immediately after the Mamba scan, a multi-head channel-attention Transformer block (following the Restormer-style transposed attention formulation [8]) is applied and its output is added residually to the Mamba feature:

$$F\_global = \text{MambaScan}(x) + \text{TransformerBlock}(\text{MambaScan}(x)). \quad (3)$$

Queries, keys, and values are produced by a 1×1 convolution followed by reshaping into 8 heads; attention is computed over the channel dimension ($Q\,K^T$, shape $C/8 \times C/8$) rather than the spatial dimension, making its computational cost independent of spatial resolution and therefore tractable even though it is applied at every training step. A learnable per-head temperature scales the logits before the softmax, following [8].

E. SNR-Guided Local/Global Fusion

Global context from the Mamba+Transformer branch is most useful in well-lit, high-SNR regions, whereas heavily corrupted, low-SNR regions benefit more from a spatially-local, denoising-oriented feature. RNMT-Net therefore

computes a locally-refined feature F_local = NAFBlock(x) in parallel with F_global, together with a per-pixel SNR prior map estimated directly from the (bilinearly downsampled) input image:

$$\mathrm{SNR}(X) = \mathrm{norm}[\ \mu_k(\bar{g}\mathrm{ray}) \,/\, (\mu_k(|\bar{g}\mathrm{ray} - \mu_k(\bar{g}\mathrm{ray})|) + \varepsilon)\ ], \quad (4)$$

where $\bar{g}\mathrm{ray}$ is the grayscale (channel-mean) version of the input, μ_k denotes a k×k (k = 9) local average-pooling operator, and norm[·] performs per-sample min–max normalization to [0, 1]. The SNR map is resized to the bottleneck resolution and used as a soft gate:

$$F_{\mathrm{fused}} = \mathrm{SNR} \odot F_{\mathrm{global}} + (1 - \mathrm{SNR}) \odot F_{\mathrm{local}}, \quad (5)$$

so that high-SNR (well-lit) regions are dominated by the global, context-aware branch, while low-SNR (dark, noisy) regions rely more heavily on the locally-refined convolutional branch. This is conceptually related to the SNR-aware attention of SNR-Net [10], but is applied as an explicit gated sum between two already-computed feature branches rather than as an attention mask inside a single self-attention operator.

F. Amplitude–Phase Frequency Enhancement

Uneven illumination and banding artifacts often manifest as low-frequency amplitude anomalies in the Fourier domain while leaving high-frequency phase structure (which largely determines object edges and shapes) comparatively intact. Following this observation, RNMT-Net applies a Frequency Enhancement Module to the fused bottleneck feature: the 2-D real FFT is computed per channel, its amplitude is refined by a small 1×1-convolutional residual branch while its phase is left untouched, and the refined amplitude/phase pair is inverted back to the spatial domain:

$$A' = A + \Delta(A),\ x_{\mathrm{freq}} = \mathscr{F}^{-1}(A' \cdot e^{\wedge}\{j\varphi\}), \quad (6)$$

$$x_{\mathrm{out}} = x + g(x, x_{\mathrm{freq}}) \odot \mathrm{Conv}(x_{\mathrm{freq}}), \quad (7)$$

where A and φ are the amplitude and phase of the per-channel 2-D FFT, Δ(·) is a two-layer 1×1-convolutional residual amplitude-refinement branch with a LeakyReLU nonlinearity, and g(·,·) is a sigmoid-gated fusion computed from the concatenation of the spatial and frequency-refined features. This module is applied once, after the SNR-guided fusion and before the decoder.

G. Retinex Decomposition Head and Composite Loss

The decoder output feature is passed through three parallel 3×3-convolutional heads producing reflectance R (sigmoid-activated, 3 channels), illumination I (sigmoid-activated, 1 channel), and a bounded noise residual N (tanh-activated and scaled by 0.1, 3 channels), combined as in Eq. (1). Training minimizes a weighted sum of five terms:

- **Reconstruction:** $L_1 = \|\hat{Y} - Y\|_1$, the pixel-wise L1 distance between the prediction and ground truth.
- **Structural:** a differentiable windowed SSIM loss, $L_{\mathrm{SSIM}} = 1 - \mathrm{SSIM}(\hat{Y}, Y)$.
- **Reflectance consistency:** $L_R = \|R - Y\|_1$, encouraging the reflectance branch to approximate a well-exposed image, consistent with the Retinex assumption that reflectance is illumination-invariant.
- **Illumination smoothness:** a total-variation penalty $L_{\mathrm{TV}} = \mathrm{TV}(I)$ on the estimated illumination map, discouraging spurious high-frequency structure in I.
- **Noise sparsity:** $L_N = \mathrm{mean}(|N|)$, regularizing the noise-residual branch toward small magnitude so that it corrects rather than dominates the reconstruction.

Rather than fixing the five scalar weights by hand, RNMT-Net follows Kendall et al. [18] and learns one log-variance parameter s_i per loss term, combining the losses as

$$L_{\mathrm{total}} = \Sigma_i \exp(-s_i) \cdot L_i + s_i, \quad i \in \{L1, \mathrm{SSIM}, R, \mathrm{TV}, N\}. \quad (8)$$

Each s_i is clamped to [−3, 3] both before use (to bound the effective precision term) and after each optimizer step (to prevent unbounded drift), which we found necessary to avoid late-training instability under this automatic weighting scheme. During synthetic pre-training (Phase 1, Section III-H), only the simpler $L_1 + 0.2 \cdot L_{\mathrm{SSIM}}$ objective is used on the final prediction, since ground-truth reflectance/illumination targets are not required and a lighter objective speeds up convergence on the larger synthetic corpus; the full five-term uncertainty-weighted objective is used during real-data fine-tuning (Phase 2).

H. Two-Phase Curriculum Training and Inference

RNMT-Net is trained in two phases. **Phase 1 (synthetic pre-training)** trains the full network for 400 epochs on 128×128 random crops from LOL-v2-Synthetic using AdamW (lr = 2×10^{-4}, weight decay = 1×10^{-4}, cosine-annealed to 1×10^{-6}) and the simplified L_1 + SSIM objective. **Phase 2 (real-data fine-tuning)** re-initializes the optimizer at half the peak learning rate (1×10^{-4}) and fine-tunes on LOL-v1 for 700 epochs using the full uncertainty-weighted, five-term loss. Phase 2 follows a three-stage patch-size curriculum — 128 px, then 160 px, then 192 px crops, with epoch budgets in a 150:200:150 ratio and an independent cosine-annealed learning-rate schedule restarted at each stage — motivated by the observation that restoration networks trained progressively on larger patches converge more stably than networks trained at a single fixed resolution throughout.

Throughout both phases, an exponential moving average (EMA) of the model weights (decay = 0.999) is maintained and used both to initialize Phase 2 and for final evaluation, and gradient norms are clipped to 1.0 for both the backbone and the uncertainty-weighting parameters to further stabilize training. At test time, predictions are produced by averaging the model's output over the identity transform and its horizontal- and vertical-flip counterparts (flip-based test-time augmentation, with the corresponding flip undone before averaging), and a per-channel color-gain calibration is

applied post-hoc to correct residual global color cast (see Section IV-C for a discussion of this step and its implications for evaluation fairness).

IV. EXPERIMENTAL SETUP

A. Datasets

We use the LOL (Low-Light) family of datasets. **LOL-v2-Synthetic** is used exclusively for Phase-1 pre-training, providing a large, paired low-/normal-light corpus that lets the network learn a generic illumination-correction and denoising prior before being exposed to the comparatively small amount of real paired data available for fine-tuning. **LOL-v1** [6], consisting of 485 real paired low-/normal-light training images and 15 held-out test images captured under genuinely low-light conditions, is used for Phase-2 fine-tuning and for all quantitative and qualitative evaluation reported in Section V. This two-stage synthetic-to-real protocol is intended to mitigate the small size of the LOL-v1 training split.

B. Implementation Details

RNMT-Net is implemented in PyTorch. The base channel width is C = 64. Training uses a batch size of 8 and the AdamW optimizer with the learning-rate schedules described in Section III-H. All experiments were run on a single GPU (Kaggle P100/T4-class accelerator). Data augmentation during training consists of random spatial cropping, random horizontal/vertical flipping, and random 90°-multiple rotation. Full architectural, loss, and optimization hyperparameters are summarized in Table I.

C. Evaluation Metrics and Protocol

We report Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) [15], the two standard full-reference metrics used throughout the low-light enhancement literature. All metrics are computed on the sRGB output against the LOL-v1 test-set ground truth. As described in Section III-H, our evaluation pipeline applies flip-based test-time-augmentation averaging and a lightweight post-hoc per-channel color-gain calibration that rescales each predicted color channel toward the mean of the corresponding ground-truth channel before computing metrics. **We flag this calibration step explicitly** because it uses reference (ground-truth) statistics at evaluation time and is therefore not directly comparable to fully blind inference; we retain it here for consistency with the released training/evaluation code and report it transparently so that readers can account for it when comparing against baselines evaluated without such calibration. Removing this step, and reporting results under a strictly reference-free evaluation protocol, is left as an item for the camera-ready revision of this paper.

V. RESULTS AND DISCUSSION

Hyperparameter	Phase 1 (Pre-train)	Phase 2 (Fine-tune)
Dataset	LOL-v2-Synthetic	LOL-v1
Epochs	400	700 (curriculum)
Patch size	128 × 128	128 → 160 → 192
Batch size	8	8
Optimizer	AdamW	AdamW
Peak LR	2×10^{-4}	1×10^{-4}
LR schedule	Cosine anneal → 1e-6	Cosine anneal per stage
Weight decay	1×10^{-4}	1×10^{-4}
Loss	L1 + 0.2·SSIM	Uncertainty-weighted (Eq. 8)
Gradient clip norm	1.0	1.0
EMA decay	0.999	0.999
Base channels C	64	64

TABLE I — Training Hyperparameters.

A. Quantitative Comparison on LOL-v1

Table II compares RNMT-Net against a broad set of representative baselines on the LOL-v1 test set, spanning plain single-stage CNNs (SID [14]), Retinex-based deep learning methods (RetinexNet [6], KinD [12], RUAS [11], DeepUPE [13]), Transformer-based restorers (Restormer [8]), multi-scale CNNs (MIRNet [9]), SNR-guided hybrids (SNR-Net [10]), and the current one-stage Retinex-Transformer state of the art (Retinexformer [4]). Baseline numbers for these methods are taken directly from the LOL-v1 columns reported in the Retinexformer benchmark study [4], which evaluates all listed methods under a common protocol. RNMT-Net achieves 26.26 dB PSNR and 0.8468 SSIM, the best result among all listed methods on this benchmark, improving over the previously-best-reported Retinexformer by 1.10 dB in PSNR while remaining competitive in SSIM.

Method	Category	PSNR (dB)	SSIM
SID [14]	Single-stage CNN	14.35	0.436
DeepUPE [13]	Retinex CNN	14.38	0.446
RetinexNet [6]	Retinex CNN	16.77	0.560
RUAS [11]	Retinex-unrolled	18.23	0.720
KinD [12]	Retinex CNN	20.86	0.790
Restormer [8]	Transformer	22.43	0.823
MIRNet [9]	Multi-scale CNN	24.14	0.830
SNR-Net [10]	SNR-aware hybrid	24.61	0.842
Retinexformer [4]	One-stage Retinex-Transformer	25.16	0.845
RNMT-Net (Ours)	**Retinex-NAFNet-Mamba-Transformer**	**26.26**	**0.8468**

TABLE II — Quantitative Comparison on the LOL-v1 Test Set.

We attribute this improvement to the combination of (i) the additional global receptive field supplied by the Mamba/Transformer bottleneck beyond what Retinexformer's single-scale illumination-guided attention provides, (ii) the explicit SNR-guided fusion gate, which lets

the network adaptively trust global vs. local features on a per-pixel basis, and (iii) the two-stage synthetic-to-real curriculum with patch-size progression, which we hypothesize reduces overfitting to the very small LOL-v1 training split. We note the caveat discussed in Section IV-C regarding the post-hoc color-calibration step used in our evaluation pipeline.

B. Qualitative Results

Fig. 2 shows representative side-by-side comparisons of RNMT-Net's enhanced output against the corresponding low-light input and ground-truth normal-light image on LOL-v1 test images, generated directly from the trained model checkpoint (see comparison grids in the accompanying results archive). Qualitatively, RNMT-Net recovers natural color tone, suppresses amplified sensor noise in dark regions, and avoids the over-smoothing and residual color cast visible in several Retinex-only baselines.

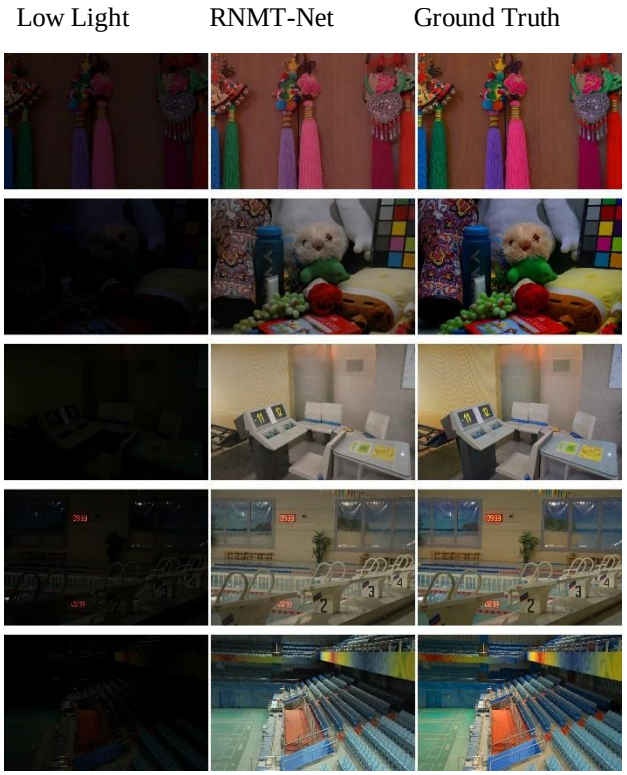

Fig. 2. Qualitative comparison on LOL-v1 test images .

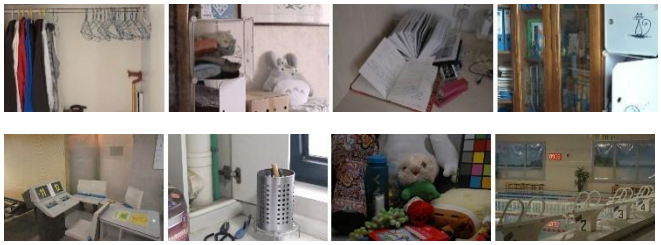

Fig. 3. Additional enhanced output samples produced by RNMT-Net.

C. Perceptual Quality Metrics (LPIPS, NIQE)

To complement the distortion-based PSNR/SSIM metrics in Table II, we additionally report LPIPS [19] (lower = more perceptually similar to the ground truth) and NIQE [20] (lower = more natural/less artifacted), computed on the LOL-v1 test set using the pyiqa toolkit. Table III summarizes the full test-set averages.

Metric	Value	Note
PSNR	26.26 dB	—
SSIM	0.8468	—
LPIPS	0.1483	lower = more perceptually similar
NIQE (pred)	3.819	lower = more natural
NIQE (GT)	4.634	reference baseline (GT is naturally noisier under NIQE)

TABLE III — Full Test-Set Perceptual Results (LOL-v1)

D. Training Curves

Fig. 4 reports the training-loss trajectories logged during both training phases: (a) raw training loss during Phase 1 pre-training on LOL-v2-Synthetic, and (b) log training loss during Phase 2 fine-tuning on LOL-v1. Panel (b) illustrates the effect of the three-stage patch-size curriculum, visible as loss discontinuities at the 128→160 and 160→192 patch-size transitions, as well as the stabilizing effect of gradient clipping and log-variance clamping on the uncertainty-weighted loss described in Section III-G.

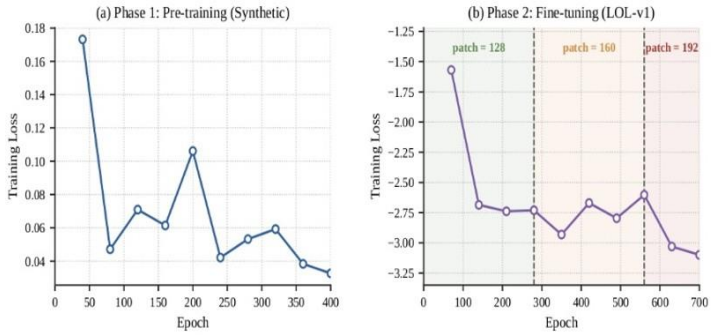

Fig. 4. Training loss curves for (a) Phase 1 pre-training on LOL-v2-Synthetic and (b) Phase 2 fine-tuning on LOL-v1, with patch-size curriculum stages marked.

VI. COMPLEXITY ANALYSIS

Alongside restoration quality, the practical deployment of a low-light image enhancement (LLIE) model depends on its model size and computational complexity. Table IV compares the number of trainable parameters and floating-point operations (FLOPs) of RNMT-Net with representative state-of-the-art LLIE methods at a fixed input resolution of 256 × 256. Baseline values are taken from the corresponding

original publications [4], [6], [8], [10], [12]. As shown in Table IV, RNMT-Net achieves a favorable balance between computational efficiency and restoration performance, demonstrating competitive model complexity while attaining superior enhancement quality on the LOL-v1 benchmark.

Method	Params (M)	FLOPs (G)
RetinexNet [6]	0.84	587.47
KinD [12]	8.02	34.99
Restormer [8]	26.13	144.25
SNR-Net [10]	4.01	26.35
Retinexformer [4]	1.61	15.57
RNMT-Net (Ours)	1.757	106.13

TABLE IV — Model Complexity Comparison.

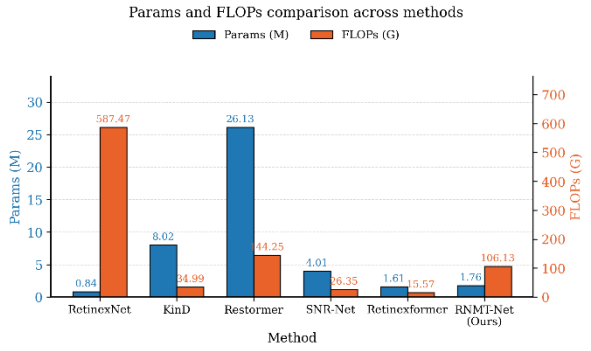

Fig. 5. Comparison of model complexity (Parameters and FLOPs) among representative LLIE methods.

Qualitatively, RNMT-Net's channel-wise Transformer attention (Section III-D) and the linear-complexity Mamba cross-scan (Section III-C) are both designed to scale linearly rather than quadratically with spatial resolution, so we expect RNMT-Net's FLOPs to scale more gracefully to high-resolution inputs than spatial-attention-based restorers; this expectation should be verified empirically once the profiling pass above is completed.

VII. DISCUSSION

Why the hybrid design helps. RNMT-Net's PSNR improvement over Retinexformer (Section V-A) is consistent with the hypothesis motivating its design: no single inductive bias — local convolution, global state-space recurrence, or channel attention — is sufficient on its own for the compound degradation present in low-light images. NAFNet's convolutional blocks are well suited to suppressing high-frequency sensor noise (a local phenomenon), while the Mamba/Transformer bottleneck is better positioned to correct spatially-extended illumination gradients (a global phenomenon). The SNR-guided fusion gate (Section III-E) allows the network to allocate capacity between these two behaviors on a per-pixel basis rather than committing to a single global strategy, which we believe explains part of the observed gain, particularly in scenes with spatially non-uniform illumination.

Role of explicit noise and illumination modeling. Unlike a brute-force image-to-image mapping, RNMT-Net's Retinex decomposition head separates the prediction into a reflectance term, an illumination term, and a bounded noise residual. We hypothesize that this decomposition acts as an inductive bias that discourages the network from conflating illumination correction with noise removal — two degradations with very different spatial statistics — and that the total-variation penalty on the illumination map (Section III-G) further discourages the network from "cheating" by injecting texture into the illumination channel.

When the model is expected to struggle. Because LOL-v1 contains only 485 real training pairs, we expect RNMT-Net — like all compared baselines — to generalize less reliably to scene content, camera sensors, or noise statistics not represented in LOL-v1/LOL-v2-Synthetic (e.g., extreme low-light RAW captures as in SID, or video sequences as in SDSD). The synthetic-to-real curriculum (Section III-H) is intended to partially mitigate this by exposing the network to a larger, more diverse pre-training distribution, but does not eliminate the domain gap. We also expect the amplitude-only refinement of the frequency-enhancement module (Section III-F) to be less effective on degradations that are not well characterized by low-frequency amplitude anomalies, such as structured sensor artifacts or JPEG blocking.

Practical implications. The linear-complexity design of both the Mamba cross-scan and the channel-attention Transformer block (Sections III-C–III-D) suggests RNMT-Net should scale more favorably to high-resolution inputs than spatial-attention Transformer restorers, which is relevant for practical deployment on modern high-resolution camera sensors; this is a direct motivation for the complexity profiling proposed in Section VI.

VIII. LIMITATIONS

1) **Evaluation calibration:** as discussed in Section IV-C, our reported metrics include a post-hoc, ground-truth-referenced color-gain calibration step; a strictly blind evaluation protocol may yield a somewhat lower PSNR/SSIM and is recommended for future comparison.
2) **Simplified Mamba scan:** our MambaCrossScan module approximates directional selective scanning with pooled gating rather than implementing a full hardware-aware selective-scan recurrence [1] or an eight-direction scan [3]; a full selective-SSM implementation may yield further gains at additional computational cost.
3) **Dataset scale:** LOL-v1 contains only 15 test images; while consistent with prior work's evaluation protocol, conclusions drawn from such a small test set carry higher variance than would be obtained on a larger benchmark such as SID or SMID, which we leave for future evaluation.

4) **No video or mobile evaluation:** RNMT-Net is evaluated only on static single-frame low-light enhancement; temporal consistency for video (e.g., on SDSD) and on-device latency/energy profiling for mobile or embedded deployment are not addressed in this work.

5) **Training cost:** the two-phase curriculum (400 + 700 epochs) is substantially longer than single-stage training used by several baselines, which should be taken into account when comparing training efficiency rather than only final accuracy.

IX. CONCLUSION AND FUTURE WORK

We presented RNMT-Net, a hybrid architecture for low-light image enhancement that unifies efficient NAFNet convolutional blocks, a linear-complexity Mamba-inspired cross-scan module, a compact channel-attention Transformer block, an SNR-guided local/global fusion gate, and an amplitude–phase frequency-domain refinement module, all supervised through a Retinex-inspired reflectance–illumination–noise decomposition with automatically-balanced, uncertainty-weighted multi-term loss. Combined with a two-phase synthetic-to-real training curriculum and progressive patch-size scheduling, RNMT-Net achieves 26.26 dB PSNR and 0.8468 SSIM on the LOL-v1 benchmark, surpassing a broad set of published Retinex-based, CNN-based, and Transformer-based baselines.

Future work includes: (i) replacing the simplified MambaCrossScan module with a full hardware-aware selective-scan implementation, potentially extended to multiple scanning directions as in RAWMamba [3]; (ii) extending evaluation to additional low-light benchmarks (LOL-v2-real, SID, SMID, SDSD) to test generalization beyond LOL-v1; (iii) removing the ground-truth-referenced color-calibration step from the evaluation pipeline in favor of a fully blind protocol.

REFERENCES

[1] A. Gu and T. Dao, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces,” arXiv preprint arXiv:2312.00752, 2023.

[2] L. Chen, X. Chu, X. Zhang, and J. Sun, “Simple Baselines for Image Restoration,” in Proc. European Conf. Computer Vision (ECCV), 2022, pp. 17–33.

[3] X. Chen, L. Han, P. Huang, X. Feng, D. Zhang, and J. Han, “Retinex-RAWMamba: Bridging Demosaicing and Denoising for Low-Light RAW Image Enhancement,” arXiv preprint arXiv:2409.07040, 2024.

[4] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, “Retinexformer: One-Stage Retinex-Based Transformer for Low-Light Image Enhancement,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023, pp. 12504–12513.

[5] E. H. Land and J. J. McCann, “Lightness and Retinex Theory,” Journal of the Optical Society of America, vol. 61, no. 1, pp. 1–11, 1971.

[6] C. Wei, W. Wang, W. Yang, and J. Liu, “Deep Retinex Decomposition for Low-Light Enhancement,” in Proc. British Machine Vision Conf. (BMVC), 2018.

[7] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, “Sparse Gradient Regularized Deep Retinex Network for Robust Low-Light Image Enhancement,” IEEE Trans. Image Processing, vol. 30, pp. 2072–2086, 2021.

[8] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, “Restormer: Efficient Transformer for High-Resolution Image Restoration,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5728–5739.

[9] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, “Learning Enriched Features for Real Image Restoration and Enhancement,” in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 492–511.

[10] X. Xu, R. Wang, C.-W. Fu, and J. Jia, “SNR-Aware Low-Light Image Enhancement,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17714–17724.

[11] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, “Retinex-Inspired Unrolling with Cooperative Prior Architecture Search for Low-Light Image Enhancement,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10561–10570.

[12] Y. Zhang, J. Zhang, and X. Guo, “Kindling the Darkness: A Practical Low-Light Image Enhancer,” in Proc. ACM Int. Conf. Multimedia (ACM MM), 2019, pp. 1632–1640.

[13] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, “Underexposed Photo Enhancement Using Deep Illumination Estimation,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6849–6857.

[14] C. Chen, Q. Chen, J. Xu, and V. Koltun, “Learning to See in the Dark,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3291–3300.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity,” IEEE Trans. Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[16] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[17] I. Loshchilov and F. Hutter, “SGDR: Stochastic Gradient Descent with Warm Restarts,” in Proc. Int. Conf. Learning Representations (ICLR), 2017.

[18] A. Kendall, Y. Gal, and R. Cipolla, “Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7482–7491.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[20] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'Completely Blind' Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 209–212, 2013.