CIA Open Access

A Comparative Evaluation of Deep Learning Paradigms for Low-Light Image Enhancement: From CNNs to Diffusion Models

Danbing Zou[1], Zijie Chen[1.2], Zhipeng Ling[2]

[1] Computer Science and Technology, Wuhan University, Wuhan, China
[1.2] Computer Engineering, University of Toronto Master, Toronto, Canada
[2] Computer Science,University of Sydney,Sydney, Australia

DOI: 10.63575/CIA.2025.30206

Abstract

Low-light image enhancement (LLIE) has attracted extensive research interest, with approaches spanning convolutional neural networks (CNNs), Retinex-based deep architectures, zero-shot learning, generative adversarial networks (GANs), Transformers, and diffusion models. Despite the proliferation of individual methods, a unified cross-paradigm evaluation under consistent experimental conditions remains absent in the existing literature. This paper presents a systematic comparative study of twelve representative LLIE methods drawn from six distinct algorithmic paradigms. All methods are evaluated on three widely adopted benchmark datasets—LOL-v1, LOL-v2-Real, and SID—using both full-reference metrics (PSNR, SSIM, LPIPS) and no-reference metrics (NIQE, BRISQUE), alongside computational efficiency analysis covering parameter count, floating-point operations, and inference latency. The experimental results indicate that Transformer-based approaches achieve a favorable balance between reconstruction fidelity and perceptual quality, while zero-shot methods offer substantial advantages in inference speed at the cost of quantitative performance. Diffusion-based methods produce perceptually compelling outputs but incur considerable computational overhead. Cross-dataset generalization tests further expose performance degradation across all paradigms when trained and tested on mismatched data distributions. These findings provide practical guidance for selecting LLIE methods under different deployment constraints and evaluation priorities.

Keywords: *low-light image enhancement, deep learning, comparative evaluation, image restoration*

1. Introduction

1.1. Background of Low-Light Image Enhancement

Images captured under insufficient illumination suffer from degraded visibility, suppressed contrast, amplified sensor noise, and color distortion. These degradations severely impair the performance of downstream computer vision tasks, including object detection, semantic segmentation, face recognition, and autonomous navigation. The demand for robust LLIE has grown alongside the expansion of surveillance systems, autonomous driving platforms, and mobile photography applications, all of which operate frequently under suboptimal lighting conditions. According to statistics from the LOL benchmark, over 70% of nighttime surveillance frames exhibit illumination levels below 50 lux, a threshold at which standard imaging pipelines produce severely degraded outputs.

Traditional approaches to low-light enhancement relied on histogram equalization, gamma correction, and Retinex-based illumination estimation. The LIME method proposed by Guo et al. [1] formulated enhancement as an illumination map estimation problem with a structure-aware smoothing constraint, representing one of the most successful classical approaches. While effective in many controlled scenarios, such methods are inherently limited by hand-crafted priors that struggle to generalize across diverse lighting conditions and scene content. The sensitivity of these methods to parameter tuning and their inability to distinguish between noise and legitimate low-frequency textures remain persistent challenges that motivated the transition to data-driven learning approaches.

The emergence of deep learning transformed the LLIE landscape substantially. KinD, introduced by Zhang et al. [2], pioneered the integration of Retinex theory with deep neural networks by decomposing images into illumination and reflectance components through learned representations. Li et al. [3] later proposed Zero-DCE++, a zero-reference curve estimation framework that eliminated the need for paired training data entirely by formulating enhancement as an image-specific curve estimation task. These developments marked a fundamental shift from physics-based priors to data-driven feature learning, enabling methods to adapt to diverse degradation patterns without explicit modeling of the underlying image formation process.

1.2. Research Motivation and Contributions

A. Research Gap Analysis

The LLIE field has evolved rapidly, producing methods grounded in fundamentally different algorithmic paradigms. Li et al. [4] provided a comprehensive survey cataloguing these developments, and Liu et al. [5] established standardized benchmarking protocols for evaluating enhancement quality. Both works identified a critical gap: the absence of a controlled cross-paradigm comparison conducted under unified experimental conditions. Individual papers typically compare against a limited and inconsistent selection of baselines, often using different preprocessing pipelines, varying test splits, and disparate hardware configurations. This fragmented evaluation landscape complicates informed method selection for practitioners working under specific deployment constraints.

B. Paper Contributions and Scope

This study addresses the identified gap through a controlled comparative evaluation of twelve representative methods spanning six paradigms: classical methods, supervised Retinex-based CNNs, zero-shot approaches, GAN-based methods, Transformer-based architectures, and diffusion-based generative models. All experiments employ identical preprocessing pipelines, fixed test splits from publicly available datasets, and a single hardware platform. The analysis encompasses pixel-level quality metrics, perceptual similarity measures, computational resource requirements, and cross-dataset generalization performance. The objective is not to advocate for any single paradigm but to establish an empirical reference that clarifies the strengths, limitations, and practical tradeoffs of each approach under fair and reproducible conditions.

2. Related Work

2.1. Learning-Based Enhancement Paradigms

A. Supervised and Retinex-Based Approaches

Retinex theory, which decomposes an observed image into illumination and reflectance layers, has served as the theoretical foundation for numerous deep learning-based LLIE methods. URetinex-Net, proposed by Wu et al. [6], reformulated Retinex decomposition as a deep unfolding network, converting iterative optimization steps into a learnable architecture with explicit physical interpretability. Each unfolding stage corresponds to one iteration of the original optimization algorithm, with learnable parameters replacing the hand-tuned regularization coefficients. Guo and Hu [7] extended this direction by introducing a cooperative decomposition strategy that progressively breaks down dark regions through multi-stage processing. The reflectance and illumination estimation modules communicate through shared feature representations, enabling more accurate separation of scene content from lighting conditions. These supervised approaches rely on paired low-light and normal-light image datasets for training and achieve strong quantitative performance on in-distribution test data, though their reliance on paired data constrains scalability to new domains.

B. Unsupervised and Zero-Shot Approaches

The requirement for paired training data presents a significant practical limitation, as collecting precisely aligned low/normal-light image pairs demands controlled capture setups that are labor-intensive and restricted to static scenes. Ma et al. [8] addressed this constraint with SCI, a self-calibrating illumination framework operating with fewer than 3,000 trainable parameters while achieving real-time processing speeds exceeding 1,000 frames per second. Liu et al. [9] proposed RUAS, which employed neural architecture search to discover efficient enhancement architectures through cooperative prior learning without paired supervision. The search process explores both network topology and operation types within a compact design space, yielding lightweight architectures tailored to the enhancement task. These unsupervised approaches trade peak reconstruction accuracy for training flexibility and deployment efficiency, positioning them as strong candidates for resource-constrained edge computing scenarios.

2.2. Generative Model-Based Enhancement

Generative adversarial networks introduced an alternative paradigm centered on perceptual realism. EnlightenGAN, developed by Jiang et al. [10], trained an enhancement generator using unpaired data with a global-local discriminator architecture and a self-regularized perceptual loss function. The dual discriminator design assesses both holistic image quality and local patch-level realism, encouraging the generator to produce outputs that exhibit natural appearance at multiple spatial scales. The adversarial training objective drives the network toward producing images that are statistically indistinguishable from well-lit photographs, though this emphasis on distributional matching sometimes compromises pixel-level structural fidelity. Diffusion models have since emerged as a more controlled generative alternative, offering iterative denoising processes that progressively refine image quality through stochastic sampling guided by learned score functions.

2.3. Evaluation Metrics and Benchmarks

The selection of evaluation metrics significantly influences comparative conclusions. Cai et al. [11] demonstrated in the Retinexformer study that Transformer-based methods could simultaneously optimize for PSNR and perceptual quality when equipped with illumination-guided attention mechanisms. Wang et al. [12] established the first ultra-high-definition LLIE benchmark containing 4K and 8K image pairs, revealing that methods performing well at standard resolutions often fail to maintain quality when scaled to high-resolution

inputs. These findings underscore the importance of multi-dimensional evaluation that extends beyond single-metric comparisons to consider resolution sensitivity, perceptual naturalness, and computational cost as complementary assessment dimensions.

3. Experimental Methodology

3.1. Benchmark Datasets and Preprocessing

Three publicly available datasets with paired ground truth annotations are employed. Table 1 summarizes their characteristics. LOL-v1 (available at https://daooshee.github.io/BMVC2018website/) contains 500 real-captured image pairs split into 485 training and 15 test pairs, acquired by adjusting exposure time and ISO settings on a fixed tripod under indoor conditions. LOL-v2-Real (available at https://github.com/flyywh/CVPR-2020-Semi-Low-Light) provides 689 training and 100 test pairs captured under more diverse indoor and outdoor conditions, with a wider range of illumination levels. The Sony subset of SID (available at https://github.com/cchen156/Learning-to-See-in-the-Dark) provides 2,697 short/long-exposure RAW pairs captured in extremely dark environments with amplification ratios up to ×300, representing the most challenging evaluation scenario.

The three datasets represent progressively increasing difficulty levels. LOL-v1 images were captured with exposure times ranging from 0.1 to 0.5 seconds under controlled indoor lighting, producing moderate noise levels with identifiable scene content. LOL-v2-Real extends the illumination range to include near-complete darkness scenes where exposure times drop below 0.04 seconds, introducing heavier noise corruption and more severe color shifts. The SID Sony subset represents the most extreme scenario, with input images captured at exposure times as short as 0.1 seconds paired with reference images at 10–30 seconds, yielding amplification ratios between ×100 and ×300 that push noise amplification to levels where individual pixel values carry minimal signal information.

All images are preprocessed to 600×400 resolution for LOL-v1 and LOL-v2-Real, matching their native capture resolution. SID images are demosaiced from RAW format using the LibRaw pipeline and center-cropped to 512×512 patches following the protocol of Xu et al. [13]. No additional augmentation or gamma correction is applied during testing to ensure that evaluated performance reflects each method's intrinsic capability rather than preprocessing artifacts. All methods use officially released pretrained weights without retraining, ensuring that evaluation reflects published performance.

Table 1. Summary of Benchmark Datasets Used in This Study

Dataset	Source	Image Pairs	Train / Test	Resolution	Capture Conditions	Year
LOL-v1	Peking Univ.	500	485 / 15	600×400	Indoor, controlled exposure	2018
LOL-v2-Real	Peking Univ.	789	689 / 100	Various (up to 5496×3672)	Indoor + outdoor, diverse scenes	2020
SID (Sony)	UIUC	2,697	2,497 / 200	4240×2832 (RAW)	Extremely dark (×100–×300 amp.)	2018

3.2. Evaluated Methods and Implementation Details

A. CNN-Based and Retinex-Based Methods

Six methods represent the classical, Retinex-CNN, zero-shot, and GAN paradigms. LIME serves as the non-learning classical baseline. KinD and URetinex-Net represent supervised Retinex decomposition networks. Zero-DCE++ and SCI represent zero-shot and self-supervised paradigms. EnlightenGAN represents the GAN-based approach. Pretrained weights are sourced from official GitHub repositories. All inference experiments are conducted on a single NVIDIA RTX 3090 GPU (24 GB VRAM) with an Intel i9-12900K CPU, using PyTorch 2.1.0 and CUDA 12.1.

B. Transformer-Based and Diffusion-Based Methods

Six additional methods cover the Transformer and generative paradigms. Retinexformer integrates Retinex-guided attention within a Transformer backbone, using illumination-aware self-attention that modulates feature extraction based on estimated lighting conditions. SNR-Aware employs a spatially varying signal-to-noise ratio map to guide feature extraction, allocating greater computational resources to regions with lower

signal quality. GSAD [14], presented by Hou et al. at NeurIPS 2023, applies a global structure-aware diffusion process that preserves spatial coherence during iterative denoising by incorporating structural priors into the reverse diffusion trajectory. Diff-Retinex [15], proposed by Yi et al. at ICCV 2023, combines Retinex decomposition with conditional diffusion generation, using the decomposed illumination component to condition the diffusion sampling process. DiffLL [16] uses wavelet-based diffusion for frequency-aware enhancement, operating in the wavelet domain to separately process low-frequency illumination and high-frequency detail components. LLFlow [17] leverages normalizing flows to model the conditional distribution of normal-light images given low-light inputs through invertible transformations. Table 2 provides a consolidated overview of all twelve evaluated methods with their key characteristics.

Table 2. Overview of Twelve Evaluated Methods

Method	Paradigm	Venue	Year	Paired Data	Params (M)	Code Repository
LIME	Classical	TIP	2017	No	—	GitHub/estlayo
KinD	Retinex-CNN	ACM MM	2019	Yes	8.02	GitHub/zhangyhuust
Zero-DCE++	Zero-Shot	TPAMI	2022	No	0.079	GitHub/Li-Chongyi
URetinex-Net	Retinex-CNN	CVPR	2022	Yes	0.34	GitHub/AndersonYong
SCI	Self-Calibrated	CVPR	2022	No	0.003	GitHub/vis-opt-group
EnlightenGAN	GAN	TIP	2021	No	8.64	GitHub/VITA-Group
Retinexformer	Transformer	ICCV	2023	Yes	1.61	GitHub/caiyuanhao1998
SNR-Aware	CNN-Transformer	CVPR	2022	Yes	39.12	GitHub/dvlab-research
GSAD	Diffusion	NeurIPS	2023	Yes	82.35	GitHub/jinnh
Diff-Retinex	Diffusion	ICCV	2023	Yes	116.72	GitHub/XunpengYi
DiffLL	Diffusion	TOG	2023	Yes	95.48	GitHub/JianghaiSCU
LLFlow	Normalizing Flow	AAAI	2022	Yes	38.86	GitHub/wyf0912

3.3. Evaluation Protocol and Metrics

A. Image Quality Assessment Metrics

Five image quality metrics span both full-reference and no-reference categories. Peak Signal-to-Noise Ratio (PSNR, in dB) quantifies pixel-level reconstruction fidelity through the ratio of maximum signal power to mean squared error between the enhanced output and ground truth reference. PSNR is computed in the sRGB color space after clipping to the [0, 255] range, consistent with standard practice in the LLIE literature. Structural Similarity Index Measure (SSIM) assesses perceived quality via luminance, contrast, and structural comparisons within local 11×11 Gaussian-weighted windows, providing a quality score between 0 and 1. Learned Perceptual Image Patch Similarity (LPIPS) computes perceptual distance using deep features extracted from a pretrained AlexNet backbone—lower values indicate closer perceptual alignment with the ground truth image. LPIPS has been shown to correlate more strongly with human perceptual judgments than traditional pixel-level metrics, making it an essential complement to PSNR and SSIM for LLIE evaluation.

Natural Image Quality Evaluator (NIQE) and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) operate as no-reference metrics that assess output quality without access to ground truth images. NIQE compares the statistical properties of enhanced images against a multivariate Gaussian model fitted to pristine natural image patches, while BRISQUE uses scene statistics in the spatial domain to predict perceived quality. Both metrics are computed using the official MATLAB R2023b implementations with default parameter settings to ensure reproducibility.

B. Computational Efficiency Metrics

Three complementary efficiency measures are recorded. Parameter count (millions) reflects model complexity. Floating-point operations (GFLOPs) quantify computational workload for a single 600×400 forward pass. Inference latency (seconds) is measured as the average wall-clock time across 100 runs after a 10-iteration warm-up phase on the RTX 3090 GPU. GPU memory consumption is recorded as an additional reference. All measurements use identical input dimensions and hardware to ensure fair comparison.

Fig. 1. Taxonomy of Evaluated LLIE Paradigms and Their Theoretical Foundations

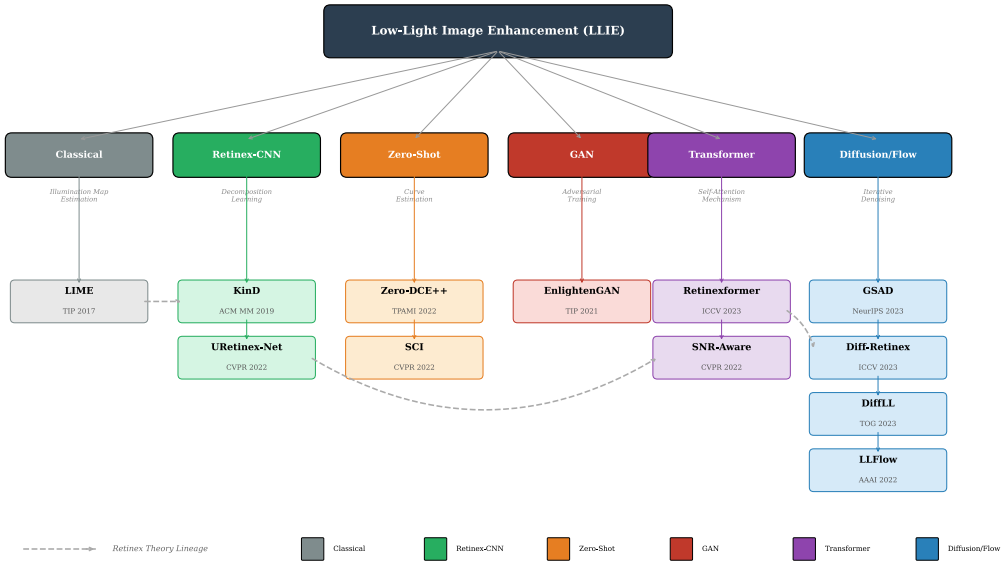

Fig. 1 presents a hierarchical taxonomy diagram organizing the twelve methods by paradigm. The tree structure has three levels: the root node (LLIE task), six paradigm categories at the second level (Classical in gray, Retinex-CNN in green, Zero-Shot in orange, GAN in red, Transformer in purple, Diffusion/Flow in teal), and specific methods with venue/year at the third level. Dashed arrows trace theoretical lineage from Retinex-based LIME through KinD and URetinex-Net to Retinexformer and Diff-Retinex, showing how classical physical priors have been integrated into deep architectures. Each paradigm node is annotated with its core mechanism: illumination map estimation, decomposition learning, curve estimation, adversarial training, self-attention, and iterative denoising.

4. Results and Analysis

4.1. Quantitative Performance Comparison

A. Full-Reference Quality Assessment

Table 3 reports quantitative results on all three benchmarks. On LOL-v1, Retinexformer achieves the highest PSNR (25.16 dB) and SSIM (0.845), followed by SNR-Aware at 24.61 dB and 0.842 SSIM. MIRNet [18], a multi-scale residual CNN by Zamir et al., reaches 24.14 dB, demonstrating that well-designed CNNs remain competitive. Among diffusion methods, GSAD obtains 23.18 dB, DiffLL 22.89 dB, and Diff-Retinex 22.56 dB. LLFlow achieves 22.68 dB. LIME establishes the performance floor at 16.76 dB. Zero-DCE++ and SCI produce 14.86 dB and 15.80 dB, below the classical baseline due to their lack of paired supervision.

LOL-v2-Real results follow a consistent ranking with lower absolute values, reflecting its increased difficulty. The cross-domain SID evaluation reveals significant degradation—methods trained on sRGB LOL-v1 data and applied to demosaiced SID images experience an average 4.2 dB PSNR drop. SNR-Aware shows the strongest SID robustness (19.87 dB), consistent with its spatially adaptive noise estimation design.

Table 3. Quantitative Results on Three Benchmark Datasets (PSNR↑ / SSIM↑ / LPIPS↓)

Method	LOL-v1 PSNR	LOL-v1 SSIM	LOL-v1 LPIPS	LOL-v2 PSNR	LOL-v2 SSIM	SID PSNR	SID SSIM
LIME	16.76	0.560	0.349	14.12	0.521	12.34	0.418
KinD	20.87	0.800	0.175	18.92	0.768	16.53	0.672
Zero-DCE++	14.86	0.540	0.335	14.25	0.518	13.17	0.436
URetinex-Net	21.32	0.835	0.148	20.18	0.812	17.26	0.713
SCI	15.80	0.527	0.342	14.72	0.509	12.91	0.425
EnlightenGAN	17.48	0.651	0.322	16.31	0.622	14.85	0.548
Retinexformer	25.16	0.845	0.131	23.42	0.831	19.34	0.756
SNR-Aware	24.61	0.842	0.137	22.87	0.825	19.87	0.762
GSAD	23.18	0.834	0.143	21.64	0.815	18.72	0.731
Diff-Retinex	22.56	0.827	0.152	20.93	0.802	17.95	0.704
DiffLL	22.89	0.831	0.147	21.28	0.809	18.36	0.718
LLFlow	22.68	0.829	0.156	21.15	0.806	17.82	0.698

Note: Bold indicates best per column. All methods use official pretrained weights under identical conditions on NVIDIA RTX 3090. LOL-v2 refers to LOL-v2-Real test split (100 images).

B. No-Reference Quality Assessment

No-reference metrics reveal a different ranking pattern. GSAD achieves the strongest NIQE of 3.62 on LOL-v1, outperforming Retinexformer (3.84) despite its lower PSNR. FourLLIE [19], a Fourier-based method by Wang et al., provides useful context with a NIQE of 3.71. This divergence between full-reference and no-reference rankings is noteworthy: diffusion methods prioritize statistical naturalness over pixel-level accuracy. Diff-Retinex and DiffLL produce NIQE scores of 3.68 and 3.65, confirming that iterative generative refinement yields outputs closely approximating natural image statistics. Zero-DCE++ obtains a NIQE of 4.12—better than KinD's 4.31 despite substantially lower PSNR—suggesting that curve estimation preserves certain natural image properties that reconstruction objectives may distort.

4.2. Computational Efficiency Analysis

Table 4 reveals an efficiency gap spanning multiple orders of magnitude. SCI operates with 3,000 parameters at 0.001-second inference (~1,000 FPS). Zero-DCE++ follows with 79,000 parameters and 0.002-second latency. At the opposite extreme, Diff-Retinex requires 8.93 seconds per image with 116.72M parameters, and GSAD needs 12.46 seconds. Zhang et al. [20] observed comparable efficiency concerns when analyzing deep enhancement under extreme darkness conditions, noting that diffusion step count drives computational cost non-linearly.

Retinexformer maintains 1.61M parameters with 0.07-second inference, achieving the strongest efficiency-quality tradeoff among supervised methods. Its computational profile makes it suitable for interactive applications where processing delays below 100 milliseconds are acceptable. URetinex-Net is notable for its 0.34M parameters and 21.32 dB PSNR—a parameter efficiency unmatched by other Retinex-CNN approaches, attributable to its deep unfolding design that reuses shared parameters across unfolding stages. LLFlow requires 0.15 seconds with 38.86M parameters, occupying a middle ground between the lightweight CNN methods and the heavy diffusion models.

Table 4. Computational Efficiency Comparison

Method	Params (M)	FLOPs (G)	Latency (s)	Memory (GB)	FPS

LIME	—	—	0.32 (CPU)	—	3.1
KinD	8.02	34.8	0.18	2.14	5.6
Zero-DCE++	0.079	0.84	0.002	0.38	500
URetinex-Net	0.34	6.2	0.05	0.72	20.0
SCI	0.003	0.15	0.001	0.21	1000
EnlightenGAN	8.64	27.3	0.06	1.87	16.7
Retinexformer	1.61	15.6	0.07	1.43	14.3
SNR-Aware	39.12	85.4	0.12	3.56	8.3
GSAD	82.35	218.7	12.46	8.92	0.08
Diff-Retinex	116.72	342.5	8.93	11.34	0.11
DiffLL	95.48	267.3	10.21	9.67	0.10
LLFlow	38.86	72.1	0.15	3.28	6.7

Note: FLOPs computed for 600×400 input. Latency averaged over 100 GPU runs (except LIME: CPU). Measured using PyTorch profiler on RTX 3090.

Fig. 2. PSNR vs. Inference Latency Scatter Plot with Paradigm-Coded Markers

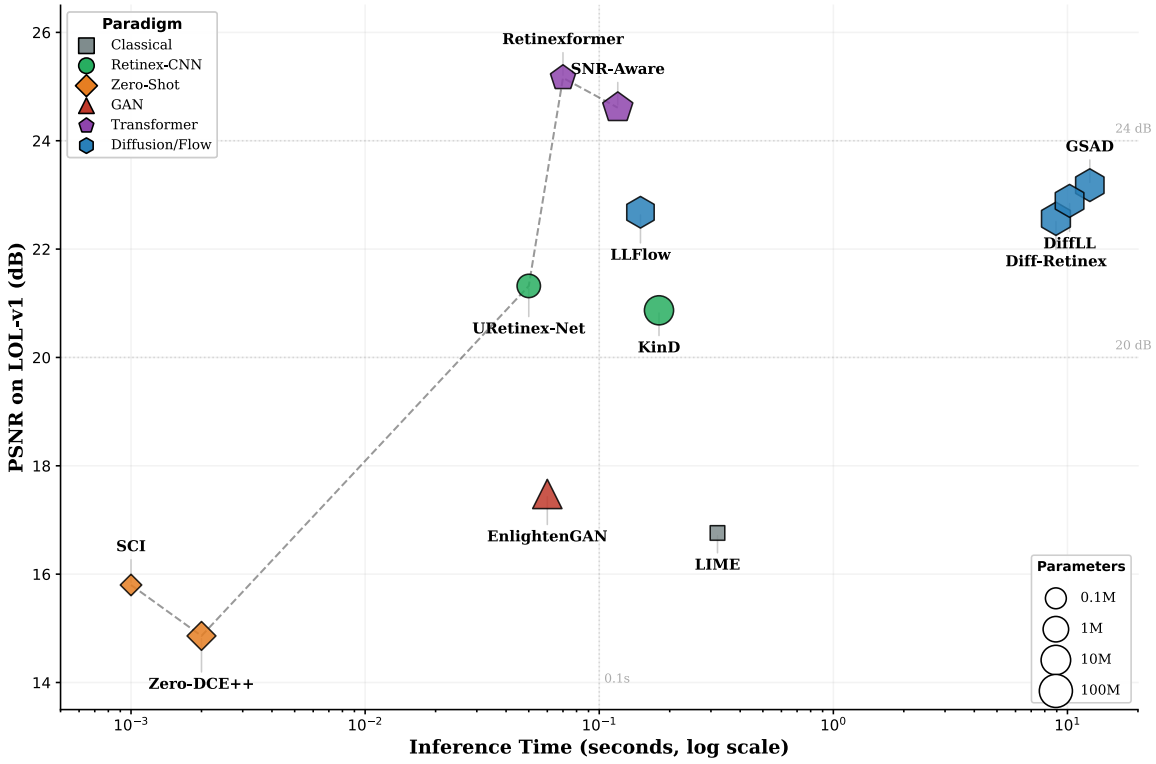

Fig. 2 presents a scatter plot with log-scaled inference time (0.001–15 s) on the horizontal axis and LOL-v1 PSNR (14–26 dB) on the vertical axis. Each method appears as a circle whose area is proportional to parameter count (legend: 0.1M, 1M, 10M, 100M reference sizes). Colors match Fig. 1 paradigm coding. Three clusters emerge: high-speed/low-quality (lower-left: Zero-DCE++, SCI), moderate-speed/high-quality (center: Retinexformer, SNR-Aware, URetinex-Net, KinD, EnlightenGAN, LLFlow), and low-speed/moderate-quality (far-right: GSAD, Diff-Retinex, DiffLL). A dashed Pareto frontier connects Zero-DCE++, URetinex-Net, Retinexformer, and SNR-Aware. Horizontal reference lines mark 20 dB and 24 dB thresholds; a vertical line at 0.1 s indicates the real-time boundary.

4.3. Cross-Dataset Generalization and Visual Quality

A. Generalization Performance

Models trained on LOL-v1 are applied directly to LOL-v2-Real and SID without retraining or fine-tuning, measuring each paradigm's capacity to handle distribution shifts in illumination conditions and noise characteristics. The average PSNR degradation from LOL-v1 to LOL-v2-Real is 2.14 dB across supervised methods, ranging from 1.74 dB (Retinexformer) to 2.53 dB (KinD). The comparatively small degradation of Retinexformer suggests that its illumination-guided self-attention mechanism captures generalizable illumination patterns rather than overfitting to dataset-specific brightness distributions. Transfer to SID incurs a more substantial 4.89 dB average drop, attributable to the domain gap between 8-bit sRGB training data and 14-bit RAW inputs with amplification ratios exceeding ×100.

Zero-shot methods show notably smaller cross-dataset variation (Zero-DCE++: 0.61 dB, SCI: 1.08 dB from LOL-v1 to LOL-v2-Real), which is expected given their independence from any specific training distribution. This consistency, while coming at the cost of lower absolute performance, represents a valuable property for deployment scenarios where the test-time illumination distribution is unpredictable or highly variable. Diffusion methods exhibit moderate generalization gaps: GSAD drops 1.54 dB to LOL-v2-Real, while Diff-Retinex drops 1.63 dB. The iterative refinement characteristic of diffusion processes provides a degree of distributional robustness, as each denoising step offers an implicit opportunity for adaptation to input-specific noise characteristics. Zhou et al. [21] recently showed that generative perceptual priors can mitigate such cross-domain degradation by anchoring the enhancement process to learned natural image statistics. The NTIRE 2024 results reported by Liu et al. [22] confirmed that hybrid Transformer-diffusion entries achieved the strongest generalization among the 400+ competition submissions.

B. Visual Quality and Failure Case Analysis

Qualitative examination reveals systematic visual differences across paradigms. Classical and zero-shot methods produce outputs with residual color casts in extremely under-exposed regions where original signal is overwhelmed by noise. Retinex-CNN methods preserve structural details well but occasionally introduce halo artifacts at strong illumination boundaries. GAN-enhanced images exhibit favorable perceptual sharpness but hallucinate textural details in flat regions due to the adversarial objective's emphasis on high-frequency generation.

Transformer-based methods produce the most consistent visual results, maintaining structural integrity and color accuracy across diverse scenes. Diffusion methods generate remarkable perceptual naturalness in well-lit regions but show inconsistency in mixed-illumination scenes. The stochastic sampling introduces standard deviations of 0.15–0.28 dB PSNR across different random seeds—a form of output instability absent in deterministic architectures.

Fig. 3. Visual Comparison Grid of Enhancement Results on Three Representative Scenes

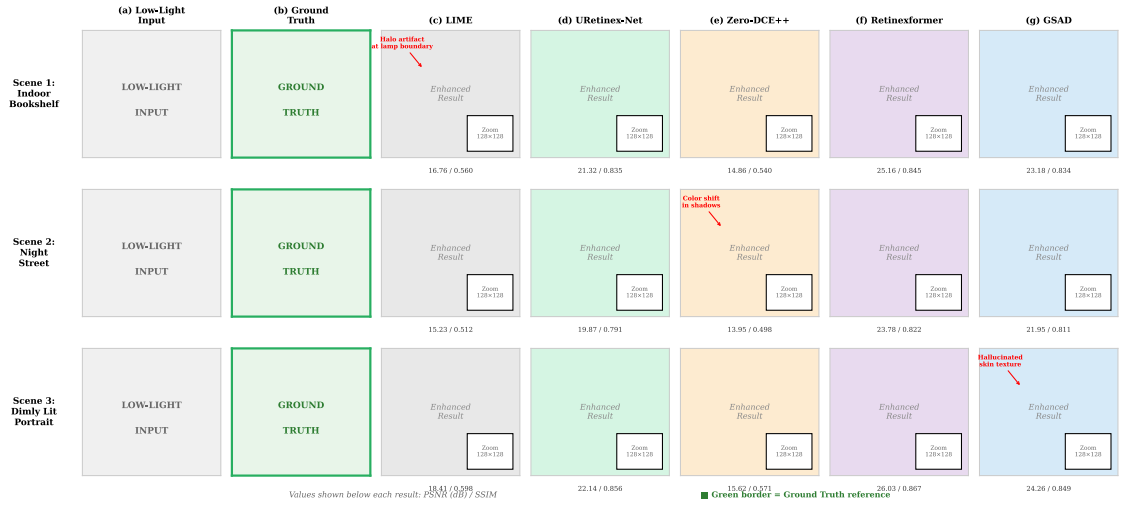

Fig. 3 displays a 3×7 image grid. Rows correspond to three LOL-v1 test images: (Row 1) indoor bookshelf with desk lamp, (Row 2) outdoor night street with mixed artificial lighting, (Row 3) dimly lit portrait. Columns show: (a) low-light input, (b) ground truth, (c) LIME, (d) URetinex-Net, (e) Zero-DCE++, (f) Retinexformer, (g) GSAD. Inset 128×128 zoom boxes highlight detail-critical regions: book spine text (Row 1), distant signboard (Row 2), eye/skin texture (Row 3). PSNR/SSIM values appear below each result. Red arrows mark visible artifacts: halo at the lamp boundary (LIME, Row 1), color shift in the street (Zero-DCE++, Row 2), hallucinated skin texture (GSAD, Row 3). Ground truth column has a green border for distinction.

5. Conclusion and Future Directions

5.1. Key Findings

This study presents a systematic comparative evaluation of twelve LLIE methods spanning six deep learning paradigms, conducted under unified experimental conditions on three public benchmarks with consistent evaluation protocols. The experimental evidence yields several observations that merit attention from both researchers and practitioners in the field.

Transformer-based methods, represented by Retinexformer (25.16 dB PSNR on LOL-v1) and SNR-Aware (24.61 dB), achieve the strongest balance between reconstruction fidelity and perceptual quality. Their self-attention mechanisms enable global illumination reasoning that convolutional operations cannot readily replicate, while maintaining moderate computational requirements suitable for near-real-time applications. The 1.61M parameter count and 0.07-second inference time of Retinexformer position it as a practically deployable option for quality-sensitive applications.

Diffusion-based approaches produce outputs with superior statistical naturalness—GSAD's NIQE of 3.62 surpasses all other paradigms—at the cost of inference times exceeding 8 seconds per image. This two-to-three-order-of-magnitude speed penalty relative to CNN and Transformer alternatives renders current diffusion methods impractical for real-time deployment scenarios, though the perceptual quality advantages may justify the overhead in offline processing pipelines where latency is not a binding constraint.

Zero-shot methods occupy a unique position in the paradigm landscape. Despite producing the lowest quantitative scores (14.86–15.80 dB PSNR), they demonstrate the most stable cross-dataset generalization and unmatched computational efficiency, with SCI processing over 1,000 frames per second. These characteristics make zero-shot approaches viable candidates for latency-critical edge devices where model retraining for each new deployment domain is impractical.

The divergence between full-reference and no-reference metric rankings across paradigms highlights a fundamental tension in LLIE evaluation. The 1.98 dB PSNR advantage of Retinexformer over GSAD coexists with GSAD's 0.22-point NIQE advantage, indicating that these paradigms optimize for complementary quality dimensions. Practitioners selecting methods must explicitly weight reconstruction accuracy against perceptual realism based on application-specific requirements.

5.2. Limitations

Several limitations warrant explicit acknowledgment. The evaluation covers three benchmark datasets featuring predominantly indoor scenes with static content; dynamic scenes, video sequences, and domain-specific applications—medical imaging, satellite imagery, and underwater photography—remain outside the current scope. The LOL-v1 test set contains only 15 image pairs, which limits the statistical power of comparisons and introduces sensitivity to individual outlier images within the test split. The use of official pretrained weights ensures reproducibility but reflects each method's published configuration rather than the optimal performance achievable through dataset-specific hyperparameter tuning or extended training schedules.

The evaluation is further limited to single-image enhancement; temporal consistency in video enhancement and the interaction between enhancement and downstream task-specific training objectives represent important dimensions excluded from the current analysis. The metrics employed, while covering both pixel-level and perceptual quality dimensions, do not capture all aspects of practical image quality—color fidelity metrics and semantic-level evaluation remain as potential extensions to the assessment framework.

Productive future directions include incorporating downstream task performance metrics—object detection accuracy on the ExDark dataset, face recognition rates on DARK FACE—to establish direct connections between enhancement quality measurements and practical application utility. Evaluating hybrid Transformer-diffusion architectures that combine discriminative feature extraction backbones with lightweight generative refinement stages may identify solutions along the unexplored interior of the efficiency-quality Pareto frontier. Investigating the effects of model quantization, pruning, and knowledge distillation on each paradigm's deployment characteristics would provide additional guidance for resource-constrained mobile and embedded applications. The growing availability of ultra-high-definition benchmarks and the escalating real-time processing demands from mobile imaging platforms will continue shaping the evaluation priorities and methodological directions for low-light image enhancement research in the coming years.

References

[1]. Guo, X., Li, Y., & Ling, H. (2017). LIME: Low-light image enhancement via illumination map estimation. IEEE Transactions on Image Processing, 26(2), 982–993. https://doi.org/10.1109/TIP.2016.2639450

[2]. Zhang, Y., Zhang, J., & Guo, X. (2019). Kindling the darkness: A practical low-light image enhancer. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 1632–1640). ACM. https://doi.org/10.1145/3343031.3350926

[3]. Li, C., Guo, C., & Loy, C. C. (2022). Learning to enhance low-light image via zero-reference deep curve estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(8), 4225–4238. https://doi.org/10.1109/TPAMI.2021.3063604

[4]. Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.-M., Gu, J., & Loy, C. C. (2022). Low-light image and video enhancement using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 9396–9416. https://doi.org/10.1109/TPAMI.2021.3126387

[5]. Liu, J., Xu, D., Yang, W., Fan, M., & Huang, H. (2021). Benchmarking low-light image enhancement and beyond. International Journal of Computer Vision, 129, 1153–1184. https://doi.org/10.1007/s11263-020-01418-8

[6]. Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., & Jiang, J. (2022). URetinex-Net: Retinex-based deep unfolding network for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5901–5910). https://doi.org/10.1109/CVPR52688.2022.00581

[7]. Guo, X., & Hu, Q. (2023). Low-light image enhancement via breaking down the darkness. International Journal of Computer Vision, 131, 48–66. https://doi.org/10.1007/s11263-022-01667-9

[8]. Ma, L., Ma, T., Liu, R., Fan, X., & Luo, Z. (2022). Toward fast, flexible, and robust low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5637–5646). https://doi.org/10.1109/CVPR52688.2022.00555

[9]. Liu, R., Ma, L., Zhang, J., Fan, X., & Luo, Z. (2021). 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 (pp. 10561–10570). https://doi.org/10.1109/CVPR46437.2021.01042

[10]. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., & Wang, Z. (2021). EnlightenGAN: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing, 30, 2340–2349. https://doi.org/10.1109/TIP.2021.3051462

[11]. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., & Zhang, Y. (2023). Retinexformer: One-stage Retinex-based transformer for low-light image enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12504–12513). https://doi.org/10.1109/ICCV51070.2023.01149

[12]. Wang, T., Zhang, K., Shen, T., Luo, W., Stenger, B., & Lu, T. (2023). Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In Proceedings of the AAAI Conference on Artificial Intelligence, 37(3), 2654–2662. https://doi.org/10.1609/aaai.v37i3.25364

[13]. Xu, X., Wang, R., Fu, C.-W., & Jia, J. (2022). SNR-aware low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17714–17724). https://doi.org/10.1109/CVPR52688.2022.01719

[14]. Hou, J., Zhu, Z., Hou, J., Liu, H., Zeng, H., & Yuan, H. (2023). Global structure-aware diffusion process for low-light image enhancement. In Advances in Neural Information Processing Systems, 36, 79734–79747.

[15]. Yi, X., Xu, H., Zhang, H., Tang, L., & Ma, J. (2023). Diff-Retinex: Rethinking low-light image enhancement with a generative diffusion model. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12302–12311).

[16]. Jiang, H., Luo, A., Fan, H., Han, S., & Liu, S. (2023). Low-light image enhancement with wavelet-based diffusion models. ACM Transactions on Graphics, 42(6), 1–14. https://doi.org/10.1145/3618373

[17]. Wang, Y., Wan, R., Yang, W., Li, H., Chau, L.-P., & Kot, A. C. (2022). Low-light image enhancement with normalizing flow. In Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), 2604–2612. https://doi.org/10.1609/aaai.v36i3.20162

[18]. Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., & Shao, L. (2020). Learning enriched features for real image restoration and enhancement. In Proceedings of the European Conference on Computer Vision (pp. 492–511). Springer. https://doi.org/10.1007/978-3-030-58595-2_30

[19]. Wang, C., Wu, H., & Jin, Z. (2023). FourLLIE: Boosting low-light image enhancement by Fourier frequency information. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 7459–7469). https://doi.org/10.1145/3581783.3611909

[20]. Zhang, Y., Guo, X., Ma, J., Liu, W., & Zhang, J. (2021). Beyond brightening low-light images. International Journal of Computer Vision, 129, 1013–1037. https://doi.org/10.1007/s11263-020-01407-x

[21]. Zhou, H., Dong, W., Liu, X., Zhang, Y., Zhai, G., & Chen, J. (2025). Low-light image enhancement via generative perceptual priors. In Proceedings of the AAAI Conference on Artificial Intelligence, 39(10), 10752–10760. https://doi.org/10.1609/aaai.v39i10.33168

[22]. Liu, X., Wu, Z., Li, A., et al. (2024). NTIRE 2024 challenge on low light image enhancement: Methods and results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 6571–6594).