SWIFTLIGHT: ONE-STEP LOW-LIGHT IMAGE ENHANCEMENT VIA TRAJECTORY-AWARE DISTILLATION AND GLOBAL-LOCAL CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have shown promising performance in low-light image enhancement. However, iterative sampling incurs substantial computation, especially for high-resolution images. Aggressive acceleration can degrade restoration quality, while independent patch processing may produce inconsistent illumination. To address these challenges, we propose **SwiftLight**, a trajectory-aware distillation framework for one-step low-light image enhancement. Specifically, we introduce a three-stage training strategy. In the first stage, we distill a multi-step teacher into a few-step model through segment-wise consistency learning. In the second stage, the few-step model supervises a one-step student that starts from the low-light input. Teacher trajectory cues and input darkness guide spatial supervision in both stages to retain important restoration corrections. In the third stage, adversarial refinement is incorporated into one-step training to improve texture realism. Furthermore, we design a lightweight consistency head to calibrate independently enhanced patches using a full-image prediction. Adaptive low-frequency correction and center-weighted fusion align illumination and color while retaining local details. Experiments demonstrate strong perceptual quality with one-step patch inference. Code will be made publicly available.

1 INTRODUCTION

Low-light image enhancement (LLIE) aims to recover clear details and natural colors from poorly illuminated images. Such images often suffer from noise, low contrast, and color distortion. Early approaches adjust contrast or estimate illumination through handcrafted priors (Land, 1977). Deep learning methods improve restoration through specialized architectures and illumination-aware guidance (Cai et al., 2023; Xu et al., 2022). However, recovering fine details from severely underexposed regions remains challenging due to weak signals and noise.

Recently, diffusion models have advanced LLIE by learning rich image priors (Zhou et al., 2023; Yi et al., 2023). Conditioned on the low-light image, they refine intermediate predictions through successive denoising steps. These priors help recover structures and textures obscured by noise and underexposure. Illumination cues, structural guidance, and frequency information further improve restoration quality (Yi et al., 2023; Hou et al., 2023; Shang et al., 2024). Patch-based diffusion extends this paradigm to high-resolution images (Özdenizci & Legenstein, 2023; Shang et al., 2024). Local processing reduces memory demands and accommodates images of varying sizes. It also retains fine spatial information by avoiding full-image resizing to a fixed input resolution.

Despite this progress, patch-based diffusion enhancement faces two challenges. **(1) Quality loss with fewer sampling steps.** Iterative sampling is costly, but aggressive acceleration can discard useful restoration refinements (Song et al., 2023; Lan et al., 2025). One-step inference must reproduce these accumulated corrections in a single evaluation. Dark regions are particularly affected because weak signals provide limited evidence for reconstructing local structures. **(2) Inconsistent illumination across patches.** Local crops lack full-image context, so their brightness and color predictions may disagree (Guo et al., 2026). Averaging overlapping outputs softens boundaries, but low-frequency brightness and color offsets can remain in the assembled image. Efficient enhancement therefore requires both effective trajectory compression and coherent patch restoration.

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



Figure 1: Visual comparisons on LOL-v2-real. SwiftLight produces globally coherent enhancements while preserving local structures and avoiding patch-related illumination discontinuities.

To address these challenges, we propose SwiftLight, a trajectory-aware distillation framework for one-step LLIE. SwiftLight follows a three-stage training strategy. In the first stage, a multi-step teacher is distilled into a few-step model through segment-wise consistency learning (Xie et al., 2026). In the second stage, the few-step model supervises a one-step student that starts from the low-light input. Teacher trajectory cues and input darkness guide spatial supervision in both stages. The third stage incorporates adversarial refinement into one-step training to improve texture realism. Together, these stages preserve key restoration corrections while enabling efficient one-step patch enhancement.

Building on the distilled one-step enhancer, we design a lightweight consistency head to calibrate independently restored patches. Guided by a full-image prediction, the head adaptively corrects low-frequency illumination and color differences across patches. Center-weighted fusion then combines the calibrated outputs. This design promotes a coherent global appearance while retaining fine local structures and textures. Experiments on LOL-v1, LOL-v2-real, and LOL-v2-synthetic demonstrate strong perceptual quality with one-step patch inference.

Our main contributions are summarized as follows:

- We introduce SwiftLight with trajectory-aware staged distillation for one-step LLIE. Teacher trajectory cues and input darkness guide the transfer of restoration knowledge from a multi-step model.
- We design global-local adaptive consistency enhancement to align patch illumination and color through output calibration and center-weighted fusion.
- SwiftLight achieves the best LPIPS and DISTS scores among compared methods on three low-light benchmarks with one-step patch inference.

2 Related Works

2.1 Low-light Image Enhancement

LLIE maps low-light observations to normally exposed images using specialized architectures, auxiliary guidance, image decomposition, unsupervised objectives, or alternative color spaces (Cai et al., 2023; Xu et al., 2022; Zhang et al., 2022; Guo et al., 2020; Yan et al., 2025). Diffusion-based methods further exploit generative priors: PyDiff uses pyramidal conditioning (Zhou et al., 2023), Diff-Retinex incorporates Retinex decomposition (Yi et al., 2023), and WCDM restores wavelet-domain high-frequency components (Jiang et al., 2023). Frequency, structural, brightness, degradation, and pretrained priors have also been explored (Lv et al., 2024; Hou et al., 2023; Yin et al., 2023; Wang et al., 2025; Xu et al., 2024; Wang et al., 2024b). However, most diffusion LLIE methods still require iterative sampling, and severely degraded regions remain difficult to restore perceptually.

Patch-based diffusion scales restoration to high-resolution inputs through local crops (Wang et al., 2023b; Hu et al., 2024). Existing approaches merge overlapping features, use hierarchical resolutions, average independently denoised patches, aggregate multiple patch scales, or process semantic subregions (Ding et al., 2024; Skorokhodov et al., 2024; Hur et al., 2025; Özdenizci & Legenstein, 2023; Shang et al., 2024; Aversa et al., 2023). These strategies either increase computation or leave low-frequency illumination and color discrepancies unresolved. SwiftLight instead combines one-step patch restoration with global low-frequency guidance and content-adaptive calibration.

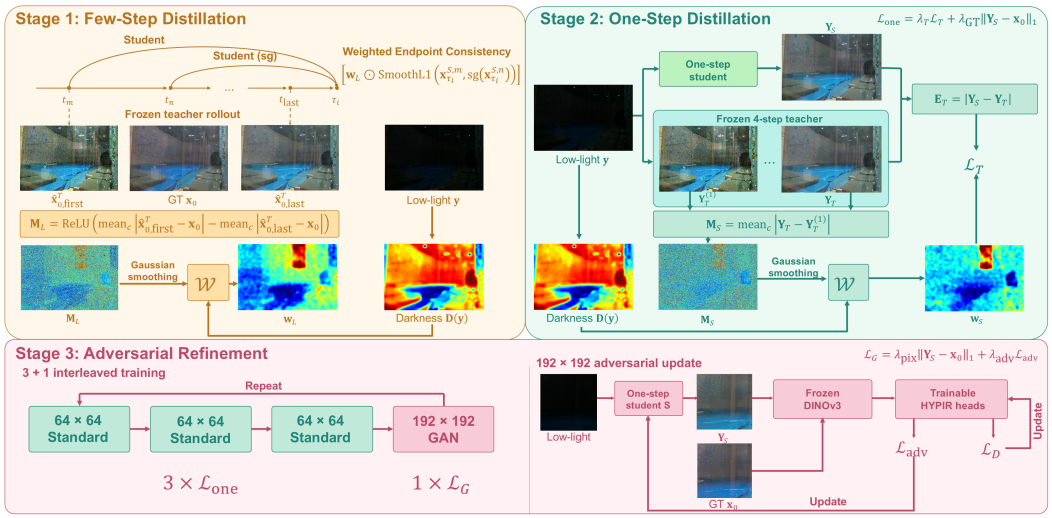

Figure 2: Overview of the trajectory-aware staged distillation in SwiftLight. Stage 1 distills the multi-step model into a few-step student through weighted endpoint consistency, where the frozen teacher's improvement map and the input darkness map determine the spatial weight $\mathbf{w}_L$. Stage 2 trains a low-light-initialized one-step student using a frozen four-step teacher, with the teacher's accumulated trajectory correction and input darkness forming $\mathbf{w}_S$. Stage 3 interleaves three standard 64×64 one-step updates with one 192×192 adversarial update, using frozen DINOv3 features and trainable HYPIR heads to improve perceptual detail.

2.2 Diffusion Model Acceleration

Diffusion acceleration relies on fast samplers or training-based distillation. Training-free samplers reduce the number of evaluations but often degrade below ten steps (Lu et al., 2022; 2025; Song et al., 2020). PD and CAD progressively compress sampling, while CD, CM, and CTM learn consistency across timesteps or trajectories (Salimans & Ho, 2022; Meng et al., 2023; Song et al., 2023; Kim et al., 2024). Latent approaches extend this idea to few- or one-step generation (Luo et al., 2023; Liu et al., 2023; Yin et al., 2024; Kang et al., 2024; Sauer et al., 2024; Wang et al., 2024a; Ren et al., 2024), but mainly target text-to-image synthesis. LLIE instead requires faithful conditional restoration. SwiftLight therefore transfers a 25-step LLIE model into a few-step teacher and then a low-light-initialized one-step student, followed by perceptual adversarial refinement.

3 Methods

Given a pretrained multistep diffusion enhancer, our goal is to obtain efficient one-step restoration without discarding the illumination and structural corrections accumulated along its denoising trajectory. As illustrated in Figure 2, our framework contains two complementary components. We first compress the multistep model through trajectory-aware staged distillation. We then introduce global-local adaptive consistency enhancement to correct the local brightness deviations and inter-patch inconsistencies that may arise during patch-wise one-step inference.

3.1 Preliminary

Diffusion Models. Diffusion models define a forward process that gradually transforms clean data into Gaussian noise and learn to reverse this corruption through iterative denoising. Given a clean image $\mathbf{x}_0$, its noisy state at timestep $t \in [0, T]$ can be directly sampled as

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \alpha_t^2 + \sigma_t^2 = 1, \tag{1}$$

where α_t and σ_t are deterministic coefficients specified by a fixed noise schedule, controlling the signal and noise levels at timestep t. The variable $\mathbf{c}$ collectively denotes the conditioning information available to the denoiser during restoration. The conditional denoiser $\boldsymbol{\epsilon}_\phi(\mathbf{x}_t, \mathbf{c}, t)$ is trained to

recover the injected noise using

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0,t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, \mathbf{c}, t)\|_2^2\right]. \tag{2}$$

The corresponding clean-image estimate is

$$\widehat{\mathbf{x}}_{0,\phi}(\mathbf{x}_t, \mathbf{c}, t) = \frac{\mathbf{x}_t - \sigma_t \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, \mathbf{c}, t)}{\alpha_t}. \tag{3}$$

Following DDIM (Song et al., 2020), a deterministic transition from t to any earlier timestep $s < t$ can then be written as

$$\Psi_\phi(\mathbf{x}_t, \mathbf{c}, t, s) = \alpha_s \widehat{\mathbf{x}}_{0,\phi}(\mathbf{x}_t, \mathbf{c}, t) + \sigma_s \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, \mathbf{c}, t), \tag{4}$$

which defines a deterministic teacher trajectory for subsequent distillation.

Consistency Distillation. CD (Song et al., 2023; Luo et al., 2023) trains a model $f_\theta(\mathbf{x}_t, \mathbf{c}, t)$ to map different states along the same denoising trajectory to a common endpoint. Given two timesteps $t_n < t_m$, an online model with trainable parameters θ and a target model with parameters θ^- are constrained to produce consistent predictions from two states on the teacher trajectory:

$$\mathcal{L}_{\text{CD}} = d(f_\theta(\mathbf{x}_{t_m}, \mathbf{c}, t_m), f_{\theta^-}(\mathbf{x}_{t_n}, \mathbf{c}, t_n)), \tag{5}$$

where $\mathbf{x}_{t_n} = \Psi_\phi(\mathbf{x}_{t_m}, \mathbf{c}, t_m, t_n)$ is obtained using a frozen diffusion teacher with parameters ϕ, $d(\cdot,\cdot)$ is a distance measure, and θ^- is updated as an exponential moving average of θ and receives no gradient. CTM (Kim et al., 2024) further introduces a target timestep s, allowing the model to map a state at t to a specified earlier point s on the same trajectory.

To alleviate the difficulty of fitting a single consistency mapping over the entire trajectory, segment-wise CD enforces consistency within shorter temporal intervals (Wang et al., 2024a; Ren et al., 2024; Xie et al., 2026). Let $0 = \tau_0 < \tau_1 < \cdots < \tau_K = T$ denote the interval boundaries, with τ_i being the left boundary of a randomly selected interval. The timestep t_m is sampled from this interval, $t_n = t_m - \Delta$ is set with $t_n < t_m$, and $\mathbf{x}_{t_m}$ is constructed by forward diffusing the clean image $\mathbf{x}_0$. The frozen teacher and two evaluations of the same student then produce

$$\mathbf{x}_{t_n}^T = \Psi_\phi(\mathbf{x}_{t_m}, \mathbf{c}, t_m, t_n), \quad \mathbf{x}_{\tau_i}^{S,m} = \Psi_\theta(\mathbf{x}_{t_m}, \mathbf{c}, t_m, \tau_i), \quad \mathbf{x}_{\tau_i}^{S,n} = \Psi_\theta\big(\mathbf{x}_{t_n}^T, \mathbf{c}, t_n, \tau_i\big). \tag{6}$$

Here, the superscripts T and S distinguish the teacher-generated state from the student predictions, respectively. The teacher parameters ϕ remain frozen, whereas θ denotes the trainable student parameters shared by the two student evaluations. The low-timestep prediction $\mathbf{x}_{\tau_i}^{S,n}$ is treated as a stop-gradient target for $\mathbf{x}_{\tau_i}^{S,m}$. Minimizing their distance encourages both predictions to represent the common interval endpoint τ_i, thereby limiting each consistency mapping to a shorter trajectory span.

3.2 Trajectory-Aware Staged Distillation

Starting from a pretrained multi-step conditional diffusion model (Jiang et al., 2023; Yi et al., 2023; Shang et al., 2024), we progressively reduce the number of denoising steps in two stages: the first stage uses segment-wise consistency distillation to obtain a few-step model, while the second uses the resulting model as a frozen teacher to train a one-step student.

Stage 1: Few-Step Distillation. Low-light degradation is spatially non-uniform: severely underexposed regions generally contain weaker observable information and are more difficult to restore (Xu et al., 2022; Cai et al., 2023). Compressing the teacher trajectory into fewer student transitions can also discard useful iterative refinements unevenly across the image, but a uniform consistency objective treats all spatial locations equally.

We therefore construct the distillation weight from two complementary cues. A GT-referenced teacher-improvement map identifies where the remaining teacher trajectory reduces reconstruction error, and an input-derived darkness map reflects the degree of local underexposure. Together, they assign greater emphasis to beneficial teacher corrections in darker regions.

Specifically, the frozen-teacher transition from t_m to t_n also yields the first clean prediction $\widehat{\mathbf{x}}_{0,\text{first}}^T = \widehat{\mathbf{x}}_{0,\phi}(\mathbf{x}_{t_m}, \mathbf{c}, t_m)$. Starting from $\mathbf{x}_{t_n}^T$, we continue the frozen teacher with stride Δ toward

τ_i and retain the clean prediction from its final evaluation as $\widehat{\mathbf{x}}^T_{0,\text{last}}$. This final evaluation occurs at $\tau_i + \Delta$ for a regular boundary and at $t = 0$ when $\tau_i = -1$. The resulting teacher-improvement map is

$$\mathbf{M}_L = \mathrm{ReLU}\big(\mathrm{mean}_c \left|\widehat{\mathbf{x}}^T_{0,\text{first}} - \mathbf{x}_0\right| - \mathrm{mean}_c \left|\widehat{\mathbf{x}}^T_{0,\text{last}} - \mathbf{x}_0\right|\big)\,. \tag{7}$$

Since the auxiliary rollout starts from the sampled $\mathbf{x}_{t_m}$, $\mathbf{M}_L$ provides a spatial estimate of the utility of the specific remaining trajectory and accordingly reweights the consistency error for the current training pair. When $t_n = \tau_i$ at a regular boundary, no additional teacher evaluation remains, the two clean predictions are identical, and the subsequent weighting operator returns a neutral weight.

We obtain the low-light-specific modulation from a darkness map that characterizes the relative underexposure at each spatial location. We derive this map using Max-RGB (Land, 1977), a classical strategy that estimates the initial illumination at each pixel. For a normalized low-light observation $\mathbf{y}$, we apply Gaussian smoothing to suppress local intensity fluctuations and take the complement of the smoothed estimate as the darkness map:

$$\widetilde{\mathbf{L}}(x) = \mathbf{y}^{\max}(x) = \max_{c\in\{R,G,B\}} \mathbf{y}^c(x), \qquad \mathbf{L}(\mathbf{y}) = \mathrm{Gauss}\Big(\widetilde{\mathbf{L}}\Big)\,, \qquad \mathbf{D}(\mathbf{y}) = 1 - \mathbf{L}(\mathbf{y}), \tag{8}$$

where $\widetilde{\mathbf{L}}$ is the initial Max-RGB estimate, $\mathbf{L}$ is its smoothed illumination map, and $\mathbf{D}$ is the corresponding darkness map. To combine a teacher-derived trajectory map $\mathbf{M}$ with $\mathbf{D}$, we define the following weighting operator:

$$\mathcal{W}(\mathbf{M}; \mathbf{y}, \gamma) = \mathcal{N}(\mathrm{Gauss}(\mathbf{M}) \odot (1 + \gamma \mathbf{D}(\mathbf{y})))\,, \tag{9}$$

Within the weighting operator, Gauss spatially smooths the teacher signal, γ controls the contribution of darkness, and $\mathcal{N}$ normalizes and bounds the resulting weights to maintain a stable loss scale. Further implementation details are provided in Appendix A.1. Applying the weighting operator to $\mathbf{M}_L$ gives the first-stage weight $\mathbf{w}_L = \mathcal{W}(\mathbf{M}_L; \mathbf{y}, \gamma_L)$.

For the first stage, we adopt the fixed-boundary segment-wise CD construction in Equation 6, match the two student predictions of the same interval endpoint, and use $\mathbf{w}_L$ to reweight their pixel-wise consistency error. Together with direct ground-truth supervision, the first-stage objective is

$$\mathcal{L}_{\text{few}} = \lambda_c \,\mathrm{mean}\big[\mathbf{w}_L \odot \mathrm{SmoothL1}\big(\mathbf{x}^{S,m}_{\tau_i}, \mathrm{sg}(\mathbf{x}^{S,n}_{\tau_i})\big)\big] + \lambda_{\text{sup}} \left\|\widehat{\mathbf{x}}^S_0 - \mathbf{x}_0\right\|_1\,, \tag{10}$$

where $\mathrm{sg}(\cdot)$ stops gradients through the low-timestep branch, and $\widehat{\mathbf{x}}^S_0$ denotes the clean prediction of the active branch. The resulting few-step model is then configured with a short sampling schedule and frozen as the teacher for the second stage.

Stage 2: One-Step Distillation. The few-step teacher still accumulates restoration corrections over several denoising evaluations, whereas the one-step student must reproduce its final result in a single evaluation. Regions that change substantially after the first teacher evaluation therefore impose a greater compression burden on the student. We use the accumulated trajectory correction of the teacher to identify these regions, combine it with the shared darkness map, and use the resulting spatial weight to guide teacher–student matching.

Both the frozen teacher and the one-step student are initialized from the few-step model obtained in Stage 1. Starting from Gaussian noise, the first teacher evaluation produces a provisional clean-image estimate $\mathbf{Y}^{(1)}_T$, while completing its short denoising schedule produces the final output $\mathbf{Y}_T$. The one-step student starts denoising from the low-light image $\mathbf{y}$ and produces the restored output $\mathbf{Y}_S$ in a single network evaluation. We measure the teacher correction accumulated after its first evaluation as

$$\mathbf{M}_S = \mathrm{mean}_c \left|\mathbf{Y}_T - \mathbf{Y}^{(1)}_T\right|. \tag{11}$$

Here, mean_c averages the absolute correction over the color channels, so larger values of $\mathbf{M}_S$ indicate a stronger dependence on subsequent teacher denoising evaluations. Applying the weighting operator in Equation 9 to $\mathbf{M}_S$ gives the second-stage spatial weight

$$\mathbf{w}_S = \mathcal{W}(\mathbf{M}_S; \mathbf{y}, \gamma_S)\,. \tag{12}$$

We apply $\mathbf{w}_S$ to the teacher–student error and retain direct ground-truth reconstruction supervision. The second-stage objective is

$$\mathcal{L}_{\text{one}} = \lambda_T \,\mathrm{mean}[\mathbf{w}_S \odot |\mathbf{Y}_S - \mathbf{Y}_T|] + \lambda_{\text{GT}} \left\|\mathbf{Y}_S - \mathbf{x}_0\right\|_1\,. \tag{13}$$

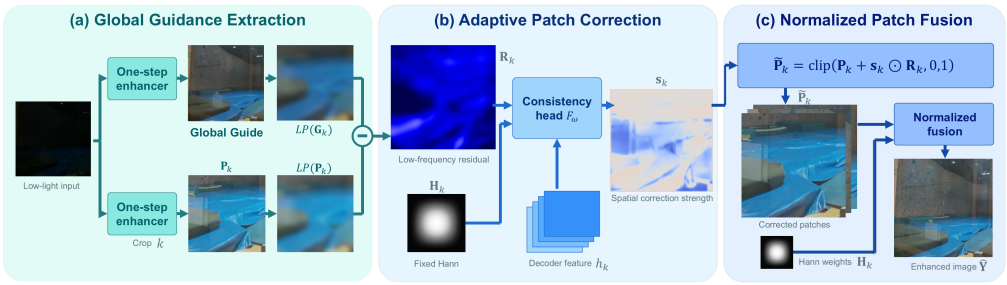

Figure 3: Overview of the consistency head and normalized patch fusion.

The two trajectory maps serve different roles. In the first stage, $\mathbf{M}_L$ identifies local teacher refinements that reduce the error relative to the ground truth. In the second stage, $\mathbf{M}_S$ identifies regions where the one-step student must absorb larger corrections accumulated by the few-step teacher. The weighting operator combines both maps with $\mathbf{D}(\mathbf{y})$ to adapt their spatial emphasis to the low-light input.

Stage 3: Adversarial Refinement. Although the first two stages establish basic restoration capability in the one-step student, the resulting images may still exhibit over-smoothed textures and insufficient perceptual detail. We therefore apply adversarial training to improve perceptual quality. However, applying large-patch adversarial supervision at every iteration would be costly and may destabilize reconstruction training. To balance perceptual refinement with efficiency and stability, we insert one adversarial update on a 192×192 patch into the ongoing Stage 2 training schedule after every three standard updates on 64×64 patches.

During the inserted adversarial update, the student generates $\mathbf{Y}_S$ from the low-light patch, and the discriminator combines a frozen DINOv3 feature extractor with trainable HYPIR discrimination heads. Using r as the real-image label, the discriminator and generator objectives are

$$\mathcal{L}_D = \sum_l \left[\mathrm{BCE}(D_l(\mathbf{x}_0), r) + \mathrm{BCE}(D_l(\mathrm{sg}(\mathbf{Y}_S)), 0)\right], \tag{14}$$

$$\mathcal{L}_G = \lambda_{\mathrm{pix}} \left\|\mathbf{Y}_S - \mathbf{x}_0\right\|_1 + \lambda_{\mathrm{adv}} \sum_l \mathrm{BCE}(D_l(\mathbf{Y}_S), r). \tag{15}$$

Here, D_l denotes the l-th discrimination head, BCE is the binary cross-entropy loss, sg is the stop-gradient operator, and r and 0 are the real and fake labels, respectively; λ_{pix} and λ_{adv} balance the pixel and adversarial terms.

3.3 Global-Local Adaptive Consistency Enhancement

High-resolution image restoration is computationally demanding, and two common strategies for improving efficiency are reducing diffusion sampling to a single step and dividing the input image into smaller patches. However, both strategies can introduce grid-like artifacts: patch-wise processing produces independent predictions across neighboring crops, while one-step inference lacks subsequent denoising iterations that could correct their illumination and color discrepancies. To suppress these artifacts, SwiftLight first performs **global guidance extraction** to obtain a low-frequency reference from a low-resolution full-image prediction. It then uses **adaptive patch correction** to calibrate each locally restored patch with this guidance. Finally, it applies **normalized patch fusion** to combine the corrected overlapping patches with Hann weights into the final output.

Global Guidance Extraction. We apply the frozen one-step enhancer once to a downsampled full image and resize its output to the original resolution, obtaining a global guide. For the k-th

local crop, we take the spatially corresponding guide patch $\mathbf{G}_k$, while local inference produces the enhanced patch $\mathbf{P}_k$ and its final decoder feature $\mathbf{h}_k$. We extract their low-frequency discrepancy as

$$\mathbf{R}_k = \mathrm{LP}(\mathbf{G}_k) - \mathrm{LP}(\mathbf{P}_k)\,, \tag{16}$$

where LP is a low-pass operator. Thus, $\mathbf{R}_k$ retains the illumination and color difference provided by the global guide while excluding most texture differences. We specify the low-pass implementation in the experimental section.

Adaptive Patch Correction. A lightweight consistency head uses the decoder feature $\mathbf{h}_k$, the global-local residual $\mathbf{R}_k$, and the fixed Hann window $\mathbf{H}_k$ to construct a bounded, single-channel correction-strength map $\mathbf{s}_k$ and inject the global residual into the local patch:

$$\mathbf{s}_k = s_0 + \rho \tanh(F_\omega(\mathbf{h}_k, \mathbf{R}_k, \mathbf{H}_k))\,, \qquad \widetilde{\mathbf{P}}_k = \mathrm{clip}(\mathbf{P}_k + \mathbf{s}_k \odot \mathbf{R}_k, 0, 1)\,. \tag{17}$$

Here, F_ω denotes the consistency head, and $\mathbf{s}_k$ controls how much of the global-local residual is injected at each spatial location, with the same strength shared across RGB channels. The form above centers the correction at $s_0 = 0.5$ and limits its adaptive deviation with $\rho = 0.25$, yielding $\mathbf{s}_k \in (0.25, 0.75)$. This design allows the head to strengthen or attenuate the global guidance while preventing reversed or excessive corrections. The corrected patch $\widetilde{\mathbf{P}}_k$ is finally clipped to the valid image range.

Normalized Patch Fusion. Finally, we combine all corrected patches by normalized overlap-add:

$$\widehat{\mathbf{Y}}(p) = \frac{\sum_{k:p\in\Omega_k} \mathbf{H}_k(p)\widetilde{\mathbf{P}}_k(p)}{\sum_{k:p\in\Omega_k} \mathbf{H}_k(p) + \varepsilon}, \tag{18}$$

where Ω_k is the spatial support of patch k. For each output location p, the numerator accumulates predictions from all corrected patches covering that location, weighted by their corresponding Hann values, while the denominator normalizes the accumulated weight. The Hann window assigns larger weights near patch centers and smaller weights near patch boundaries, reducing the influence of less reliable boundary predictions; ε prevents division by zero.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on three paired low-light image enhancement benchmarks: LOL-v1 (Wei et al., 2018), LOL-v2-real, and LOL-v2-synthetic (Yang et al., 2021). LOL-v1 contains 485 low-/normal-light image pairs for training and 15 pairs for testing. LOL-v2-real contains 689 training pairs and 100 test pairs captured from real scenes, while LOL-v2-synthetic contains 900 training pairs and 100 test pairs generated with synthetic low-light degradation. For each benchmark, we train a separate model using its official training split and evaluate it on the corresponding test split.

Implementation Details. SwiftLight is implemented in PyTorch. Stages 1 and 2 use Adam on 64×64 patches; the consistency head uses AdamW with the enhancer frozen. Inference uses 192×192 patches, stride 144, a global-guide long edge of 192, and Gaussian filtering with kernel size 37 and standard deviation 6. Dataset-wise optimization settings and training schedules are detailed in Appendix A.3.

Metrics. We use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) (Wang et al., 2004) to evaluate pixel and structural fidelity, with SSIM computed over the RGB channels. Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) and Deep Image Structure and Texture Similarity (DISTS) (Ding et al., 2020) measure full-reference perceptual distance. To complement these reference-based measures, we further assess the perceptual quality and naturalness of the enhanced results without ground-truth images using four no-reference metrics: CLIP-IQA (Wang et al., 2023a), QualiCLIP+ (Agnolucci et al., 2024), TOPIQ-NR (Chen et al., 2024), and LIQE (Zhang et al., 2023).

Table 1: Quantitative comparison with state-of-the-art LLIE methods. Higher is better for PSNR, SSIM, CLIP-IQA, QualiCLIP+, TOPIQ-NR, and LIQE; lower is better for LPIPS and DISTS. The best and second-best results are marked in red and blue, respectively.

		Non-diffusion					Diffusion-based					
Dataset	Metric	KinD	EnlightenGAN	LLFlow	HVI-CIDNet	CWNet	Diff-Retinex	LightenDiff.	MDMS	GPP-LLIE	P-PatchDiff	SwiftLight
LOL-v1	PSNR↑	20.3717	17.4828	21.1491	23.8083	23.6011	22.6903	20.1652	22.7464	23.2701	22.6907	23.3623
	SSIM↑	0.8803	0.7156	0.9053	0.8894	0.9023	0.8991	0.8693	0.9060	0.8999	0.9091	0.9012
	LPIPS↓	0.1593	0.3223	0.1189	0.0856	0.1230	0.0980	0.1777	0.0835	0.0801	0.1052	0.0791
	DISTS↓	0.1204	0.1959	0.1185	0.0783	0.1073	0.0962	0.1654	0.0796	0.0722	0.1060	0.0686
	CLIP-IQA↑	0.3703	0.4301	0.3968	0.4879	0.3905	0.5027	0.2938	0.5135	0.4778	0.4147	0.6344
	QualiCLIP+↑	0.5833	0.4697	0.6597	0.6376	0.6125	0.6886	0.5318	0.6059	0.6415	0.6654	0.6617
	TOPIQ-NR↑	0.6304	0.4832	0.6904	0.6764	0.6477	0.7014	0.5074	0.6607	0.6562	0.6853	0.7065
	LIQE↑	3.5996	2.3222	4.4120	4.2335	3.7528	4.5690	2.4668	4.1189	4.4129	4.4558	4.5137
LOL-v2-real	PSNR↑	23.7729	18.6397	17.4331	23.9045	21.6474	21.5680	22.9645	18.9227	–	20.1033	20.6938
	SSIM↑	0.8944	0.7277	0.8463	0.8953	0.8815	0.8559	0.8788	0.8304	–	0.8647	0.8736
	LPIPS↓	0.1213	0.3087	0.1758	0.1219	0.1357	0.1615	0.1656	0.1213	–	0.1503	0.1080
	DISTS↓	0.1166	0.1923	0.1464	0.1030	0.1144	0.1427	0.1660	0.1009	–	0.1353	0.0953
	CLIP-IQA↑	0.3997	0.4260	0.3310	0.5031	0.3645	0.3918	0.2775	0.5413	–	0.3685	0.5202
	QualiCLIP+↑	0.5727	0.4672	0.6103	0.5799	0.5908	0.6182	0.5282	0.6067	0.4693	0.6270	0.6292
	TOPIQ-NR↑	0.6508	0.5100	0.6583	0.6428	0.6612	0.6592	0.5072	0.6688	0.4698	0.6750	0.7015
	LIQE↑	3.6730	2.4119	3.9485	3.7631	3.9505	4.0319	2.4578	3.8904	2.1119	4.1389	4.1447
LOL-v2-syn	PSNR↑	18.3159	16.5727	23.4258	25.7045	25.5012	22.5197	21.6075	23.9942	26.8717	23.8745	25.0213
	SSIM↑	0.8327	0.8046	0.9506	0.9596	0.9536	0.9151	0.8984	0.9497	0.9604	0.9434	0.9566
	LPIPS↓	0.2521	0.2117	0.0498	0.0471	0.0436	0.0887	0.1552	0.0478	0.0421	0.0568	0.0413
	DISTS↓	0.1920	0.1563	0.0526	0.0570	0.0534	0.0839	0.1408	0.0509	0.0551	0.0600	0.0467
	CLIP-IQA↑	0.3759	0.4884	0.4988	0.4924	0.5261	0.5299	0.4046	0.5204	0.5190	0.5037	0.5306
	QualiCLIP+↑	0.5486	0.5646	0.6452	0.6269	0.6424	0.6484	0.5731	0.6523	0.6400	0.6476	0.6562
	TOPIQ-NR↑	0.5762	0.6084	0.6784	0.6629	0.6796	0.6888	0.5228	0.6736	0.6680	0.6790	0.6874
	LIQE↑	2.9499	3.1226	4.1848	4.0982	4.2579	4.1669	3.1484	4.1766	4.1653	4.1938	4.2234

4.2 Comparison with State-of-the-Art Methods

We compare SwiftLight with representative non-diffusion approaches, including KinD (Zhang et al., 2019), EnlightenGAN (Jiang et al., 2021), LLFlow (Wang et al., 2022), HVI-CIDNet (Yan et al., 2025), and CWNet (Zhang et al., 2025), as well as diffusion-based approaches, including Diff-Retinex (Yi et al., 2023), LightenDiffusion (Jiang et al., 2024), MDMS (Shang et al., 2024), GPP-LLIE (Zhou et al., 2025), and P-PatchDiff (Guo et al., 2026).

Quantitative Results. Table 1 reports results on LOL-v1, LOL-v2-real, and LOL-v2-synthetic. SwiftLight achieves the best LPIPS and DISTS on all three datasets, indicating consistently strong perceptual fidelity. It also ranks first or second on CLIP-IQA, TOPIQ-NR, and LIQE across all datasets, while obtaining the best QualiCLIP+ on LOL-v2-real and LOL-v2-synthetic. Although the highest PSNR and SSIM are achieved by different competing methods, the consistent gains across both full-reference perceptual distances and no-reference quality metrics demonstrate the effectiveness of SwiftLight in preserving perceptual quality under aggressive one-step acceleration.

Qualitative Results. Figure 4 presents representative visual comparisons across the three benchmarks, with additional results provided in Appendix A.5. SwiftLight restores clearer local structures and fine details while maintaining a natural, spatially coherent global color tone under diverse illumination conditions.

Inference Efficiency. Table 2 reports the inference time for a single 1920×1080 image with batch size 1. SwiftLight completes full-image enhancement in 5.895 s, including low-resolution global-guidance prediction, 104 local patch evaluations, adaptive consistency correction, and center-weighted fusion. It is the fastest among the compared approaches and reduces latency by 31.7% relative to the second-fastest method, CWNet. The original 25-step MDMS is particularly expensive at this resolution because its dense three-scale sampling with a stride of 8 requires 85,220 patch evaluations. These results show that the proposed one-step formulation substantially improves high-resolution inference efficiency while retaining global guidance and patch-consistency calibration.

4.3 Ablation Studies

We conduct ablation studies on LOL-v1 to examine the staged training strategy, the consistency head, and the low-frequency extraction operator.

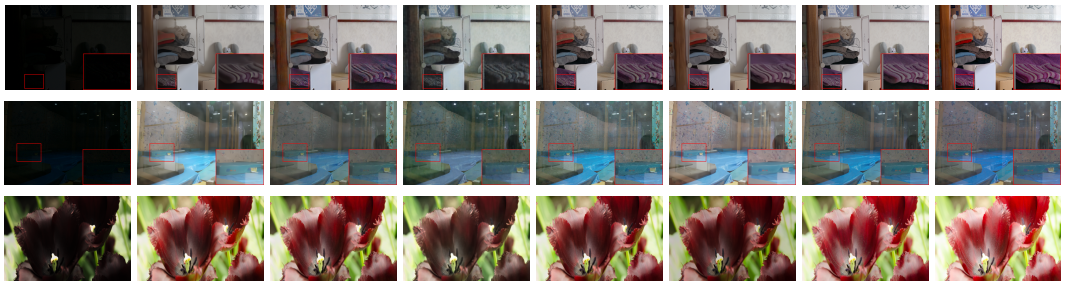

Figure 4: Visual comparisons across the three benchmarks. From top to bottom, the rows show results on LOL-v1, LOL-v2-real, and LOL-v2-synthetic, respectively.

Table 2: Inference time for a single 1920×1080 image with batch size 1. The best and second-best results are shown in **bold** and underlined, respectively.

Method	LLFlow	CWNet	Diff-Retinex	LightenDiff.	MDMS	GPP-LLIE	P-PatchDiff	SwiftLight
Runtime (s)	50.179	8.636	35.986	40.808	70,761.88	45.819	16.681	**5.895**

Staged Training. Table 3 compares five training configurations. *Pre-25Steps* is the initial multi-step model. *GT-1step* directly trains it as a one-step predictor with GT supervision until convergence. *S1-only* first distills it into a four-step model, then performs GT-only one-step training for 120K iterations. From the same four-step model, *S1+S2* applies one-step distillation for 60K iterations, while *S1+S2+S3* uses Stage 2 for 30K and Stage 3 for 90K iterations. The full schedule achieves the best PSNR, LPIPS, and CLIP-IQA with slightly lower SSIM than the 25-step model, while requiring only one denoising step.

Train Config.	PSNR↑	SSIM↑	LPIPS↓	CLIP-IQA↑
Pre-25Steps	22.7464	**0.9060**	0.0835	0.5135
GT-1step	23.1956	0.9056	0.0898	0.5178
S1-only	23.1679	0.9058	0.0889	0.5139
S1+S2	22.8178	0.9059	0.0836	0.5139
S1+S2+S3	**23.3869**	0.9026	**0.0800**	**0.6239**

Table 3: Ablation on staged training.

Consistency Head. Table 4 compares uniform averaging (*w/o Align*), fixed low-frequency correction with $s = 0.5$ (*Fixed LP*), added Hann fusion (*Fixed LP+H*), and adaptive correction (*SwiftLight*). Despite slightly higher PSNR and SSIM, *w/o Align* produces severe grid artifacts. SwiftLight suppresses these artifacts and achieves the best LPIPS and CLIP-IQA, improving perceptual quality and spatial coherence. Further visual comparisons appear in Appendix A.5.

Table 4: Ablation on the consistency head on LOL-v1. Quantitative results are shown on the left, with representative visual comparisons on the right. The best and second-best results are shown in **bold** and underlined, respectively.

Metric	w/o Align	Fixed LP	Fixed LP+H	SwiftLight
PSNR↑	**23.3869**	23.3092	23.3055	23.3623
SSIM↑	**0.9026**	0.8996	0.8996	0.9012
LPIPS↓	0.0800	0.0795	0.0797	**0.0791**
CLIP-IQA↑	0.6239	0.6272	0.6317	**0.6344**



5 Conclusion

We presented SwiftLight, a one-step diffusion framework for efficient low-light image enhancement. Its trajectory-aware distillation and adversarial refinement preserve restoration quality, while the global–local consistency head calibrates high-resolution patches using full-image guidance. Experiments on three LOL benchmarks demonstrate strong perceptual quality and substantially faster inference than iterative diffusion baselines.

AI USE STATEMENT

Generative AI tools were used to assist with translation, language editing, code debugging, and the formatting of figures and tables. They were not used to fabricate training data, experimental results, evaluation scores, or scientific claims. All AI-assisted text and code were reviewed and revised by the authors, and executable code was tested before use. The authors take full responsibility for the methodology, experiments, analysis, and final presentation of this work.

ETHICS STATEMENT

This work uses established low-light image enhancement benchmarks and does not involve human-subject studies or the collection of new personal data. We follow the usage conditions of the employed datasets and pretrained models. Because enhancement algorithms can modify the visible appearance of a scene, outputs from SwiftLight should not be treated as faithful forensic evidence without independent verification. We are not aware of other material ethical risks specific to the experiments reported in this paper.

REPRODUCIBILITY STATEMENT

The model formulation, training objectives, and inference procedure are described in Section 3. Dataset splits, evaluation metrics, and inference settings are reported in Section 4.1, while detailed weighting, consistency-head training, and dataset-wise optimization settings are provided in Appendices A.1–A.3. The high-resolution runtime protocol is specified in Table 2, and the principal component analyses are reported in Section 4.3 and Appendix A.4. We plan to release the training and evaluation code, configuration files, and model checkpoints upon publication.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for real-world image quality assessment. *arXiv e-prints*, pp. arXiv–2403, 2024.

Marco Aversa, Gabriel Nobis, Miriam Hägele, Kai Standvoss, Mihaela Chirica, Roderick Murray-Smith, Ahmed M Alaa, Lukas Ruff, Daniela Ivanova, Wojciech Samek, et al. Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology. *Advances in Neural Information Processing Systems*, 36:78126–78141, 2023.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 12470–12479. IEEE, 2023.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Zheng Ding, Mengqi Zhang, Jiajun Wu, and Zhuowen Tu. Patched denoising diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 46932–46954, 2024.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1777–1786. Ieee, 2020.

Ruoyu Guo, Haonan Zhong, Maurice Pagnucco, and Yang Song. P-patchdiff: Progressive patch diffusion models for low-light image enhancement. *International Journal of Computer Vision*, 134(9):404, 2026.

Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36:79734–79747, 2023.

Jason Hu, Bowen Song, Xiaojian Xu, Liyue Shen, and Jeffrey Fessler. Learning image priors through patch-based diffusion models for solving inverse problems. *Advances in Neural Information Processing Systems*, 37:1625–1660, 2024.

Junhwa Hur, Charles Herrmann, Saurabh Saxena, Janne Kontkanen, Wei-Sheng Lai, Yichang Shih, Michael Rubinstein, David J Fleet, and Deqing Sun. High-resolution frame interpolation with patch-based cascaded diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3868–3876, 2025.

Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (ToG)*, 42(6):1–14, 2023.

Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In *European conference on computer vision*, pp. 161–179. Springer, 2024.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.

Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. In *European Conference on Computer Vision*, pp. 428–447. Springer, 2024.

Dongjun Kim, Chieh-Hsin Lai, WeiHsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *International Conference on Learning Representations*, volume 2024, pp. 44493–44525, 2024.

Guanzhou Lan, Qianli Ma, Yuqi Yang, Zhigang Wang, Dong Wang, Xuelong Li, and Bin Zhao. Efficient diffusion as low light enhancer. In *CVPR*, 2025.

Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The twelfth international conference on learning representations*, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Xiaoqian Lv, Shengping Zhang, Chenyang Wang, Yichen Zheng, Bineng Zhong, Chongyi Li, and Liqiang Nie. Fourier priors-guided diffusion for zero-shot joint low-light enhancement and deblurring. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25378–25388. IEEE, 2024.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306. IEEE, 2023.

Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE transactions on pattern analysis and machine intelligence*, 45(8):10346–10357, 2023.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. *Advances in neural information processing systems*, 37:117340–117362, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Kai Shang, Mingwen Shao, Chao Wang, Yuanshuo Cheng, and Shuigen Wang. Multi-domain multi-scale diffusion model for low-light image enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4722–4730, 2024.

Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, and Sergey Tulyakov. Hierarchical patch diffusion models for high-resolution video generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7569–7579. IEEE, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Fu-Yun Wang, Zhaoyang Huang, Alexander W Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024a.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023a.

Tao Wang, Kaihao Zhang, Yong Zhang, Wenhan Luo, Björn Stenger, Tong Lu, Tae-Kyun Kim, and Wei Liu. Lldiffusion: Learning degradation representations in diffusion models for low-light image enhancement. *Pattern Recognition*, 166:111628, 2025.

Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26057–26066. IEEE, 2024b.

Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2604–2612, 2022.

Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in neural information processing systems*, 36:72137–72154, 2023b.

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

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

Qingsong Xie, Zhenyi Liao, Chen Chen, Zhijie Deng, and Haonan Lu. Logcd: Local-to-global consistency distillation for few-step image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8954–8964, 2026.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 17693–17703. IEEE, 2022.

Xiaogang Xu, Shu Kong, Tao Hu, Zhe Liu, and Hujun Bao. Boosting image restoration via priors from pre-trained models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2900–2909. IEEE, 2024.

Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5678–5687. IEEE, 2025.

Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12268–12277. IEEE, 2023.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024.

Yuyang Yin, Dejia Xu, Chuangchuang Tan, Ping Liu, Yao Zhao, and Yunchao Wei. Cle diffusion: Controllable light enhancement diffusion model. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 8145–8156, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Tongshun Zhang, Pingping Liu, Yubing Lu, Mengen Cai, Zijian Zhang, Zhe Zhang, and Qiuzhan Zhou. Cwnet: Causal wavelet network for low-light image enhancement. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8789–8799. IEEE, 2025.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081. IEEE, 2023.

Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 1632–1640, 2019.

Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. Deep color consistent network for low-light image enhancement. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 1889–1898. IEEE, 2022.

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

Han Zhou, Wei Dong, Xiaohong Liu, Yulun Zhang, Guangtao Zhai, and Jun Chen. Low-light image enhancement via generative perceptual priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10752–10760, 2025.

A Appendix

A.1 Weighting Operator

The weighting operator in Equation 9 converts a single-channel trajectory map into detached spatial weights with a controlled range and a unit spatial mean. Let Ω denote the spatial domain of one training sample. Given a trajectory map $\mathbf{M}$ and a low-light input $\mathbf{y}$, we first compute the unnormalized importance and its relative magnitude as

$$\mathbf{I} = \mathrm{Gauss}(\mathbf{M}) \odot (1 + \gamma \mathbf{D}(\mathbf{y})), \qquad \mathbf{Q} = \frac{\mathbf{I}}{\langle \mathbf{I} \rangle_\Omega + \epsilon}, \tag{19}$$

where $\langle \cdot \rangle_\Omega$ denotes spatial averaging. Gaussian smoothing makes the emphasis depend on spatially coherent teacher corrections rather than isolated pixels, while the multiplicative darkness term increases their relative importance in underexposed regions. We then obtain a bounded zero-mean modulation map

$$\mathbf{A} = \tanh(\mathbf{Q} - 1) - \langle \tanh(\mathbf{Q} - 1) \rangle_\Omega . \tag{20}$$

The hyperbolic tangent limits extreme relative responses, and subtracting the spatial mean ensures that increasing the weights at selected locations is balanced by decreasing them elsewhere.

To bound the final weights, let $a_+ = \max_{p \in \Omega} \mathbf{A}(p)$ and $a_- = -\min_{p \in \Omega} \mathbf{A}(p)$. The admissible scales for the positive and negative sides are

$$\begin{aligned} \beta_+ &= \begin{cases} (w_{\max} - 1)/a_+, & a_+ > \epsilon, \\ +\infty, & \text{otherwise}, \end{cases} \\ \beta_- &= \begin{cases} (1 - w_{\min})/a_-, & a_- > \epsilon, \\ +\infty, & \text{otherwise}, \end{cases} \\ \beta &= \min(\beta_+, \beta_-). \end{aligned} \tag{21}$$

If both candidates are infinite, which occurs when the modulation has no spatial variation, we set $\beta = 0$. At training iteration r, the normalization $\mathcal{N}$ in Equation 9 is implemented as

$$\eta_r = \min\left(1, \frac{r}{R}\right), \qquad \mathcal{W}(\mathbf{M}; \mathbf{y}, \gamma) = \mathbf{1} + \eta_r \beta \mathbf{A}, \tag{22}$$

where R is the warm-up duration. Consequently, the weights begin near a uniform map, remain within $[w_{\min}, w_{\max}]$, and preserve a spatial mean of one. A trajectory map without spatial variation therefore produces the neutral weight $\mathbf{1}$. In both distillation stages, we use $w_{\min} = 0.5$, $w_{\max} = 2.0$, $R = 5000$, $\gamma_L = \gamma_S = 1$, and $\epsilon = 10^{-6}$. The Gaussian operator uses a 37×37 kernel with standard deviation 6. The complete weighting computation is detached from backpropagation so that it only redistributes the corresponding distillation error.

A.2 Consistency Head Training

After the three-stage training, the one-step enhancer is frozen and only the consistency head is optimized. Each training sample is constructed from one paired low-light and ground-truth image. We randomly select a 336×336 region and extract four 192×192 patches whose top-left coordinates within this region are $(0, 0)$, $(0, 144)$, $(144, 0)$, and $(144, 144)$. The stride of 144 produces a 48-pixel overlap between horizontally or vertically adjacent patches, while the central 48×48 area is shared by all four patches, as illustrated in Figure 5.

The frozen enhancer processes the four low-light patches to obtain the local predictions $\mathbf{P}_k$ and decoder features $\mathbf{h}_k$. To construct the global guide, the complete low-light image is resized such that its longer edge is at most 192 pixels, restored once by the same frozen enhancer, and resized back to the original resolution. The corresponding crops form $\mathbf{G}_k$. The low-frequency residual $\mathbf{R}_k$ is then computed by Equation 16 using a Gaussian filter with kernel size 37 and standard deviation 6.

The consistency head first projects $\mathbf{h}_k$ to 32 channels with a 1×1 convolution. The projected feature is concatenated with the three-channel residual $\mathbf{R}_k$ and the single-channel Hann window $\mathbf{H}_k$, followed by two 3×3 convolution–SiLU blocks and a final 3×3 convolution that produces one

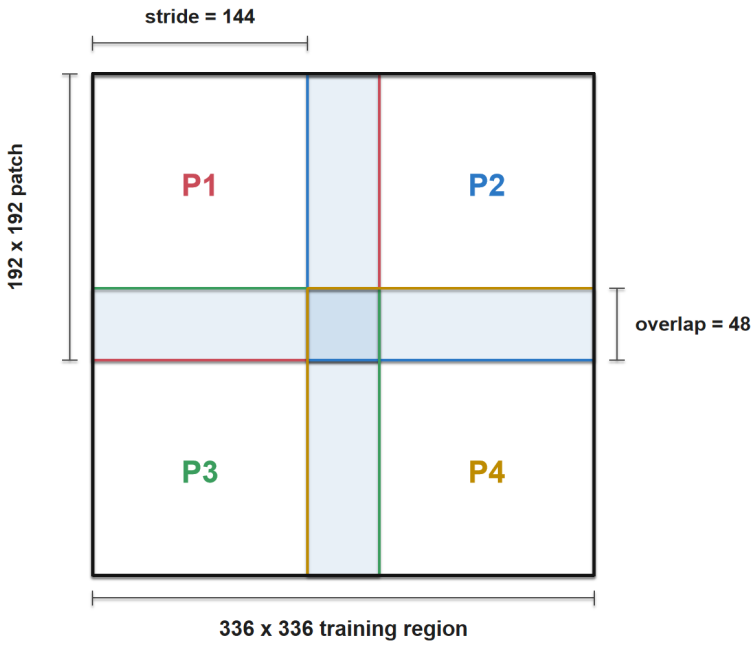

Figure 5: Construction of one training group for the consistency head. Four 192×192 patches are sampled on a 2×2 grid with a stride of 144, jointly covering a 336×336 region. Adjacent patches overlap by 48 pixels in each direction.

channel. Equation 17 transforms this output into the correction strength $\mathbf{s}_k$. The final convolution is initialized to zero, so the initial strength is $s_0 = 0.5$ before learning content-dependent deviations. We use $\rho = 0.25$, and construct $\mathbf{H}_k$ as a normalized two-dimensional Hann window with a minimum value of 0.02.

The corrected patches are fused by Equation 18 to form $\widehat{\mathbf{Y}}$ over the sampled 336×336 region. In addition, a detached fixed-strength reference is constructed using

$$\mathbf{P}_k^{\text{fix}} = \text{clip}\left(\mathbf{P}_k + s_0 \mathbf{R}_k, 0, 1\right), \qquad \mathbf{Y}_{\text{fix}}(p) = \frac{\sum_{k:p\in\Omega_k} \mathbf{H}_k(p)\mathbf{P}_k^{\text{fix}}(p)}{\sum_{k:p\in\Omega_k} \mathbf{H}_k(p) + \varepsilon}. \tag{23}$$

This reference retains the original fixed correction strength and the same normalized Hann fusion. It acts as an anchor rather than a trainable target branch. Let $\mathbf{Y}$ denote the ground-truth region corresponding to $\widehat{\mathbf{Y}}$. The consistency head is optimized by

$$\begin{aligned}\mathcal{L}_{\text{head}} = \lambda_{\text{img}} \,\text{Charb}(\widehat{\mathbf{Y}}, \mathbf{Y}) + \lambda_{\text{low}} \left\| \text{LP}(\widehat{\mathbf{Y}}) - \text{LP}(\mathbf{Y}) \right\|_1 \\ + \lambda_{\text{anchor}} \left\| \widehat{\mathbf{Y}} - \mathbf{Y}_{\text{fix}} \right\|_1 + \lambda_{\text{reg}} \,\text{mean}_k \left\| F_\omega(\mathbf{h}_k, \mathbf{R}_k, \mathbf{H}_k) \right\|_1,\end{aligned} \tag{24}$$

where

$$\text{Charb}(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\left(\widehat{\mathbf{Y}}_i - \mathbf{Y}_i\right)^2 + \epsilon_c^2}, \qquad \epsilon_c = 10^{-3}, \tag{25}$$

and N counts all pixels and color channels. The image term supervises the final fused reconstruction, and the low-frequency term emphasizes illumination and color consistency. The anchor term limits excessive deviation from the established fixed-strength result, while the final term regularizes the raw head output before the bounded transformation in Equation 17. We set $\lambda_{\text{img}} = 1$, $\lambda_{\text{low}} = 0.5$, $\lambda_{\text{anchor}} = 0.2$, and $\lambda_{\text{reg}} = 10^{-3}$.

For each dataset-specific model, the head is optimized using AdamW with a learning rate of 10^{-4} and zero weight decay; the 5K checkpoint is selected for evaluation. One full-image pair, corresponding to one four-patch group, is used per iteration. Mixed-precision training is enabled, and

Table 5: Dataset-wise training configuration. For Stage 3, paired entries separated by "/" correspond to the standard one-step update and the adversarial update, respectively.

Dataset	Training phase	Pairs	Iterations	Patch	Batch	Optimizer	Learning rate
LOL-v1	Stage 1	485	45K	64^2	16	Adam	2×10^{-5}
	Stage 2		30K	64^2	8	Adam	5×10^{-6}
	Stage 3 (standard/GAN)		90K	$64^2/192^2$	$8/1$	Adam	$5 \times 10^{-6}/10^{-5}$
	Consistency head		5K	4×192^2	1	AdamW	1×10^{-4}
LOL-v2-real	Stage 1	689	85K	64^2	16	Adam	2×10^{-5}
	Stage 2		20K	64^2	8	Adam	5×10^{-6}
	Stage 3 (standard/GAN)		10K	$64^2/192^2$	$8/1$	Adam	$5 \times 10^{-6}/10^{-5}$
	Consistency head		5K	4×192^2	1	AdamW	1×10^{-4}
LOL-v2-synthetic	Stage 1	900	110K	64^2	16	Adam	2×10^{-5}
	Stage 2		20K	64^2	8	Adam	5×10^{-6}
	Stage 3 (standard/GAN)		150K	$64^2/192^2$	$8/1$	Adam	$5 \times 10^{-6}/10^{-5}$
	Consistency head		5K	4×192^2	1	AdamW	1×10^{-4}

the gradient norm is clipped to 1. The enhancer remains in evaluation mode throughout training, so gradients are propagated only through the consistency head.

A.3 Training Details

We train an independent model for each benchmark using only its official paired training split. All runs start from the corresponding dataset-specific 25-step enhancer, and Table 5 summarizes the optimization used by our three-stage distillation and consistency head. The listed batch size is the number of paired images per update. In Stage 3, "standard/GAN" denotes the interleaved schedule of three standard one-step updates followed by one adversarial update. The optimization hyperparameters are shared across datasets, while the number of training iterations is selected separately for each benchmark.

The diffusion process uses 1000 linear noise levels with $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$. Stage 1 starts at $t = 960$ and divides the trajectory into intervals of 120; the frozen teacher advances with a stride of 40. We use equal consistency and supervised loss weights, Smooth L1 with $\beta = 1$, mixed-precision training, and a gradient-norm limit of 1.

For Stages 2 and 3, the student starts from the low-light image at $t = 333$ and is supervised by the frozen four-step teacher. We use an ℓ_1 trajectory loss with $\lambda_{\text{teacher}} = 1$ and a ground-truth reconstruction weight of $\lambda_{\text{GT}} = 0.1$. The adversarial update operates on a single 192×192 patch after every three standard 64×64 updates. Its reconstruction and adversarial weights are 0.1 and 0.01, respectively; the discriminator uses a learning rate of 10^{-5}, a real-label target of 0.8, and mixed-precision training. Adam uses zero weight decay in all distillation stages, and the student EMA decay is 0.9999. Validation is performed every 500 iterations and checkpoints are stored every 5K iterations.

A.4 Low-Frequency Extraction Ablation

Table 6 examines the low-pass operator used to construct the global–local residual. *Bilinear (1-ch.)* obtains a single-channel low-frequency estimate through bilinear resizing, whereas *Area* uses area downsampling followed by bilinear upsampling. *Gaussian* instead applies channel-wise Gaussian filtering at the original resolution, with $k = 37$ and $\sigma = 6$. Gaussian filtering achieves the best PSNR and CLIP-IQA and ties Area for the best SSIM and LPIPS, providing the most balanced performance and therefore serving as the default low-frequency extractor.

Table 6: Ablation on low-frequency extraction on LOL-v1. The best and second-best results are shown in **bold** and underlined, respectively.

Metric	Bilinear (1-ch.)	Area	Gaussian ($k = 37,\ \sigma = 6$)
PSNR↑	23.3051	23.3602	**23.3623**
SSIM↑	0.9005	**0.9012**	**0.9012**
LPIPS↓	0.0799	**0.0791**	**0.0791**
CLIP-IQA↑	0.6328	0.6326	**0.6344**

A.5 Additional Visual Comparisons

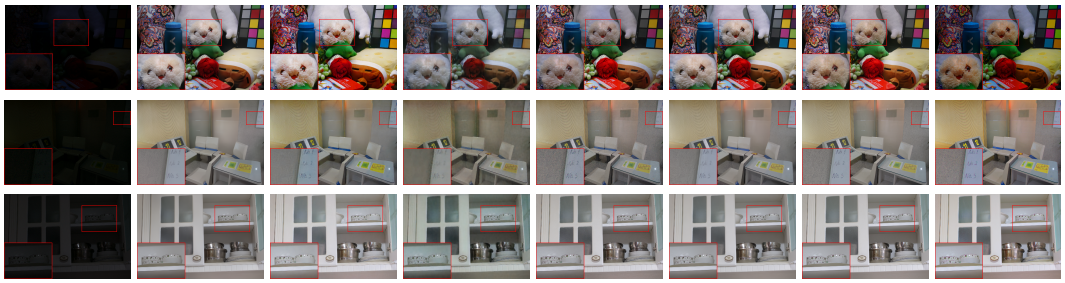

Figure 6: Additional visual comparisons on LOL-v1.

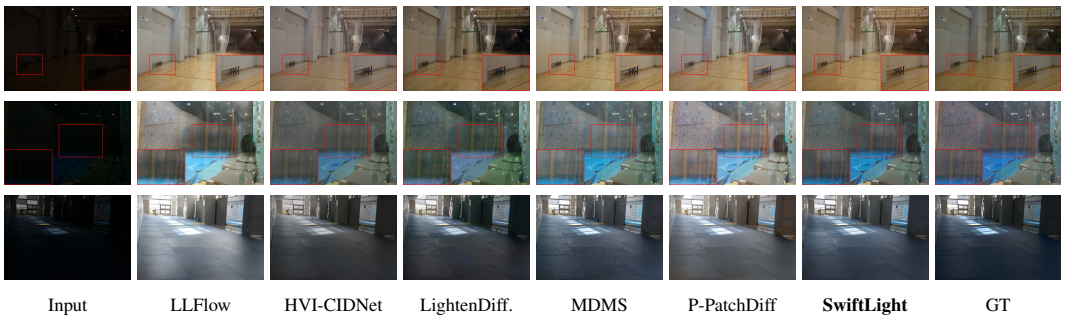

Figure 7: Additional visual comparisons on LOL-v2-real.

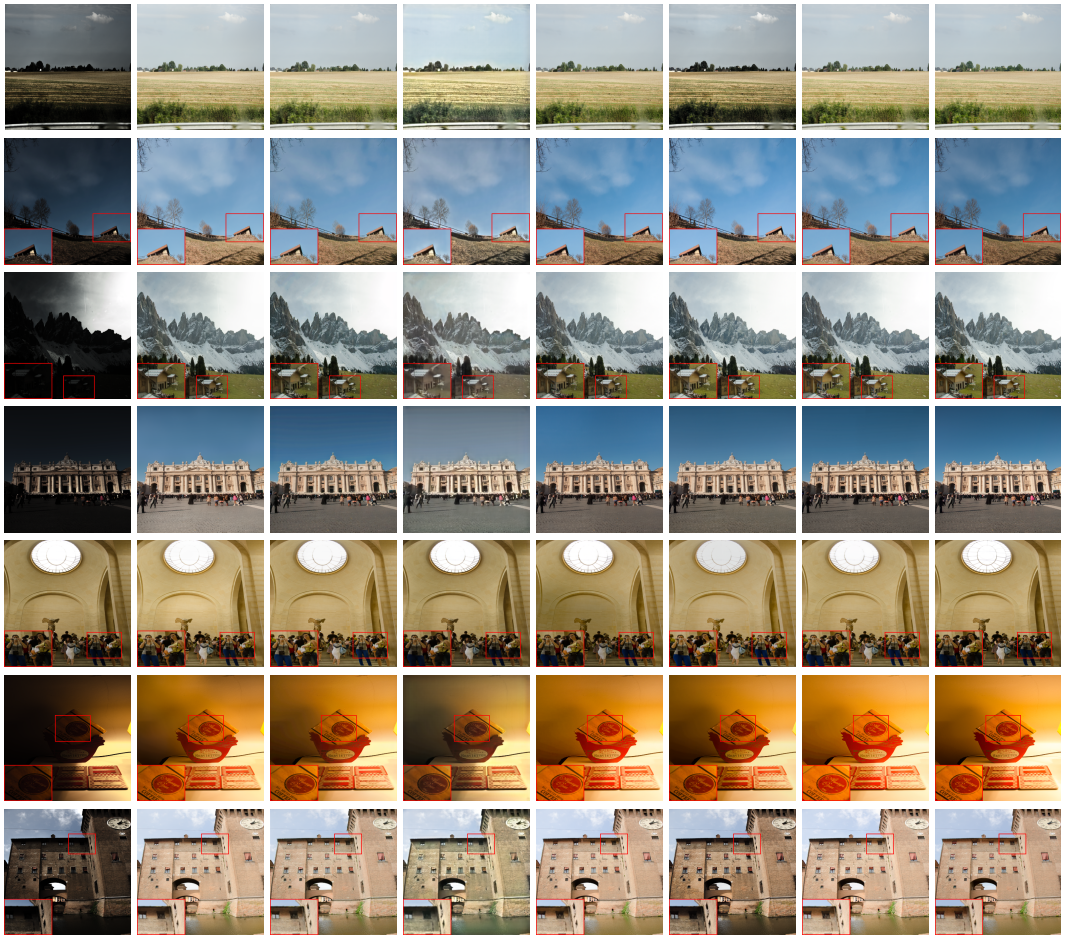

Figure 8: Additional visual comparisons on LOL-v2-synthetic.