Learning to Enhance in One Step: Average Residual Flow for Efficient Low-light Image Enhancement

1st Weixuan Huang
Tongji University
Shanghai, China
hwx_xuan@tongji.edu.cn

2nd Zhenghan Zhang
Tongji University
Shanghai, China
zzh1030@tongji.edu.cn

3rd Yiyang Li*
Tongji University
Shanghai, China
2252078@tongji.edu.cn

***Abstract*—While diffusion and flow models exhibit strong generative ability for low-light image enhancement, they commonly suffer from trajectory drift accumulation: color bias builds up along curved sampling trajectories from Gaussian noise, and iterative sampling further amplifies this drift, causing color distortion and slow inference that hinder real-time perception in nighttime autonomous driving. Experiments show that the recursive updates in conventional flow models introduce redundant computation and, more critically, break the direct mapping from low-light to normal-light domains. To address these problems, we propose the Average Residual Flow (ARF), which redefines the generative process as a direct mapping in the residual space. By averaging multi-step residual fields, ARF eliminates the Markov chain dependency and compresses the original multi-step iteration into a one-step transformation. To better approximate average residuals, we design a lightweight Conditioned Illumination-Temporal Enhancement Network (CITE-Net), which explicitly models the joint representation of temporal evolution and illumination disparity. Experiments show that the joint design of ARF and CITE-Net achieves state-of-the-art performance across six benchmarks from general and real-world driving datasets, delivering 2–5× speedup over comparable methods, validating its effectiveness and practicality.**

***Index Terms*—Low-light Image Enhancement, Diffusion models, Flow-based models**

I. Introduction

The safety of autonomous driving systems relies heavily on the accuracy of visual perception, yet image degradation under low-light conditions poses a serious threat to their reliable operation. In scenarios such as nighttime driving or tunnels, images captured by cameras often suffer from insufficient brightness, severe noise contamination, and significant detail loss, leading to a sharp decline in the performance of downstream tasks such as object detection and semantic segmentation [1]–[3]. Therefore, developing efficient and high-quality low-light image enhancement algorithms has become a critical requirement for autonomous driving vision systems.

In recent years, deep learning–based methods for low-light image enhancement have achieved remarkable progress [4]. In particular, diffusion models and flow models have demonstrated breakthroughs in image quality, driven by their powerful generative capabilities [5]–[11]. However, these generative approaches suffer from a fundamental design limitation: they inherit the paradigm of general image generation, starting from pure Gaussian noise and gradually approaching the target distribution through multi-step iterations. This paradigm introduces two critical issues in low-light enhancement. First, the generation trajectory from noise to image is inherently nonlinear, where small drifts at early steps are progressively amplified, eventually leading to color distortions. Second, the multi-step sampling process not only incurs significant computational overhead but also disrupts the inherently direct mapping between low-light and normal-light images. Although existing works attempt to alleviate these issues through techniques such as conditional guidance [12], [13] and customized noise scheduling strategies [14], [15], they fail to fundamentally resolve the mismatch between the generative paradigm and the nature of the enhancement task.

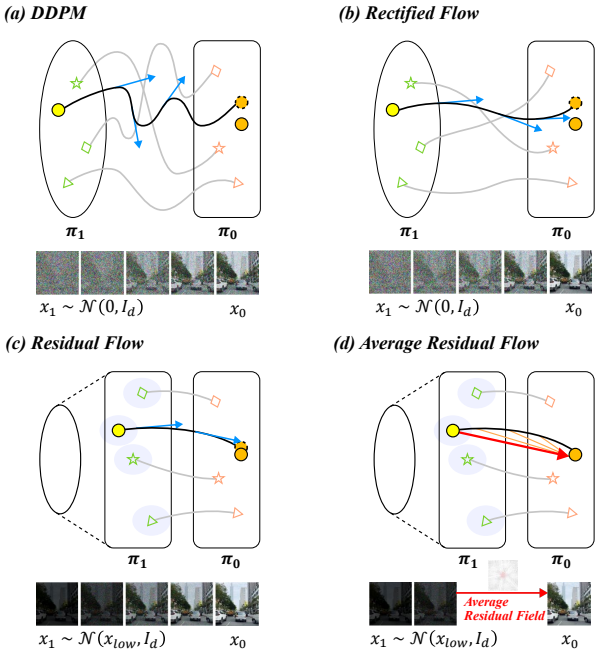

Fig. 1. (a) illustrates that DDPM is affected by stochastic noise during sampling, while (b) performs ODE-based flow matching for smoother trajectories. (c) replaces the Gaussian prior with a sample-centered distribution, shifting flow matching to the residual space for more stable sampling. (d) shows the proposed ARF, which averages the residual flow to estimate the average residual field and achieve one-step enhancement.

*Corresponding author: Yiyang Li (2252078@tongji.edu.cn)

Through systematic experiments, we identify a phenomenon of "trajectory drift accumulation" in diffusion and flow models for low-light image enhancement. Specifically, we designed the following comparative studies: first, visualizing the generation trajectories across different time steps shows a clear difference between paths starting from pure noise and those initialized from low-light images—the former tends to be less smooth and more prone to deviate from the target; second, examining the color distributions of intermediate states indicates that color shifts appear early in the generation process and continue to accumulate over subsequent steps (see Fig. 2). These observations indicate that the conventional noise-to-image generation paradigm is not well-suited for low-light enhancement, which requires a direct mapping between input and output. Crucially, the multi-step iterative process not only fails to improve generation quality but also amplifies distortions due to accumulated errors, while significantly reducing inference efficiency.

To address the core limitation of these issues, we propose the **Average Residual Flow (ARF)** and the **Conditioned Illumination-Temporal Enhancement Network (CITE-Net)**, enabling high-fidelity, single-step low-light enhancement through a redefined generation process. The core innovation of ARF lies in modeling the enhancement task as a direct mapping in the residual space, rather than following the conventional noise-to-image generation paradigm. ARF theoretically redefines the generation process by setting the mean of the initial distribution to the input low-light image, effectively transforming the problem into learning the image residual. To better describe the fitting process of residual flow, a *residual field* is defined, representing both the rate and direction of residual changes. By averaging residual fields across multiple time steps, ARF eliminates temporal dependencies between states, compressing a previously multi-step iterative process into a single-step transformation. This design not only significantly accelerates inference but also prevents the accumulation of trajectory drifts. CITE-Net is designed to support the prediction of the *average residual field*, incorporating simple yet effective conditional injection and a specialized attention mechanism to capture joint features of temporal evolution and illumination changes. By explicitly modeling dependencies across different time steps and illumination levels, CITE-Net provides rich semantic guidance for accurate residual field prediction.

Extensive experiments on six benchmarks covering both traditional LLIE and autonomous driving scenarios demonstrate that ARF and CITE-Net achieve an average **5.54% improvement in image quality metrics** on the LoLI-Street valid set compared to the best existing methods, while delivering **2–5× faster inference** than current diffusion and flow models, offering an efficient and reliable solution for nighttime autonomous driving perception.

Our main contributions are as follows:

- We identify and systematically validate the phenomenon of "trajectory drift accumulation" in diffusion and flow models for low-light enhancement, revealing a key mismatch between noise-based generation paradigms and the enhancement task, and providing new insights into the limitations of existing methods.
- We propose **Average Residual Flow (ARF)** and **CITE-Net**. ARF compresses multi-step iterations into a single-step residual mapping, while CITE-Net enables temporal-illumination joint modeling of the residual field, fundamentally addressing color distortion and low efficiency.
- Experiments on six benchmarks demonstrate that ARF and CITE-Net achieve simultaneous improvements in both quality and speed, offering a 2–5× acceleration over comparable methods, and an average performance gain of 5.54% on the LoLI-Street valid set, fully validating the effectiveness and practical value of our approach.

II. Related Work

Low-light Image Enhancement (LLIE). Deep learning has been widely applied to low-light image enhancement. Transformer-based modular networks, such as LLFormer [16], Restormer [17], and Retinex-inspired models like URetinex-Net [18] excel in capturing global illumination consistency for accurate color recovery but often incur halo artifacts. To achieve better generalization and enhance the model's ability to capture noise characteristics, methods such as DiffLL [9], PyDiff [10], AGLLDiff [19], and LLFlow [11] leverage the generative capability of diffusion and flow models, however, their perceptual quality remains limited due to drifts in the denoising trajectory.

Diffusion Models & Flow Models for Image Restoration. The introduction of denoising diffusion models such as DDPM [20] and DDIM [21] has brought significant breakthroughs to image generation. Equally significant are flow models, such as Flow Matching [22] and Rectified Flow [23]. In recent years, they have shown remarkable success in low-level vision applications. For instance, DiffIR [24] accelerates the restoration process via a conditional denoising network. The Reversing Flow framework [25] inverts the image degradation process into a flow-based generative pathway, and RDDM [26] proposes a dual-diffusion framework that decouples residuals and noise to achieve directional diffusion from the target image to the conditional input. Nevertheless, both diffusion and flow models suffer from slow inference due to multiple reverse sampling steps, and their unstable trajectories can further degrade image quality.

III. Methodology

In this section, we provide a detailed overview of the principles and design of ARF and CITE-Net. Section III-A provides the theoretical background necessary for this work. Section III-B presents the core concepts and theoretical derivation of ARF, which constitutes the primary contribution of this paper. Section III-C introduces the design rationale of CITE-Net based on ARF theory. The overall workflow is illustrated in Fig. III-B1.

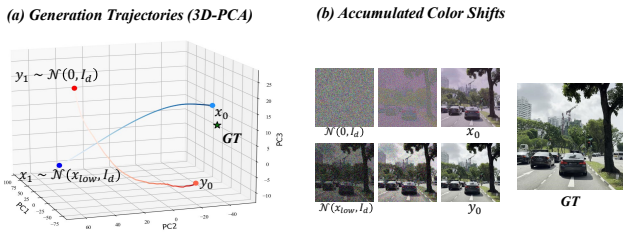

Fig. 2. (a) **Trajectory Visualization in PCA Space.** We flatten sampling sequences into vectors and project the concatenated paths of Ground Truth (GT), Gaussian-based, and our residual-centered trajectories onto the top-3 principal components. The trajectory starting from the low-light image (blue) is smoother and converges closer to GT than the Gaussian-initiated path (red), which exhibits significant non-linear drift. (b) **Color Distortion Analysis.** Inverting from Gaussian noise accumulates mid-stage color shifts (e.g., pinkish artifacts), whereas our approach, by centering the prior on the low-light input, constrains the sampling path and ensures global color consistency.

A. Preliminaries

The essence of diffusion and flow models is learning a mapping from a source to a target distribution. In conventional diffusion models, the local distribution is typically a standard Gaussian, $\mathcal{N}(0, I_d)$. The generative process can be defined as an ODE or an SDE:

$$dX_t = \mu(X_t, t), dt \tag{1}$$

$$dX_t = \mu(X_t, t), dt + \sigma(X_t, t), dW_t \tag{2}$$

where X_t is the trajectory, μ the drift term, and σ the diffusion term. Flow models are deterministic, so ODE-based flows produce straighter trajectories, making low-light enhancement faster and more accurate, and providing a suitable basis for incorporating the average residual flow. The trajectory of the generative process via ODE is also referred to as the flow.

B. Average Residual Flow

1) Residual Field and Residual Flow: Conventional flow models typically initialize from Gaussian noise, which is suboptimal for image enhancement tasks where a strong structural prior (the low-light image) is available. Unlike iterative denoising, we formulate the enhancement as a conditional probability path p_t that bridges the low-light distribution p_1 and the target manifold p_0. As shown in Fig. 2, trajectories initiated from x_{low} exhibit significantly lower curvature and better convergence to the ground truth compared to noise-based baselines, effectively mitigating the color drift accumulation common in generative paradigms.

Hence, the initial distribution of the model is defined as a normal distribution with its mean centered at x_{low}, formulated as:

$$x_1 = x_{low} + \epsilon \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, I_d)$, I_d denotes the identity matrix, and $x_1 \sim \mathcal{N}(x_{low}, I_d)$. In this work, the low-light image is denoted as x_{low}, and the corresponding ground truth image is denoted as x_0. At time step t ($t \in [0, 1]$), the image generated by the modified flow model can be expressed as:

$$x_t = \alpha_t x_0 + \beta_t x_1 \tag{4}$$

where α_t and β_t are functions of t and satisfy $\alpha_t + \beta_t = 1$, commonly referred to as schedulers. Next, we define the residual $\delta = x_0 - x_{low}$, i.e., the difference between the ground-truth image and the low-light image. From the above expression, we have:

$$x_t = (\alpha_t + \beta_t)x_0 - \beta_t \delta + \beta_t \epsilon \tag{5}$$

This yields a sampling process that is inherently constrained by the residual geometry.

We now define the *target residual field* as:

$$F_{res}^{tgt}(x_t, t) = \xi(t)(\epsilon - \delta) \tag{6}$$

where ξ characterizes the instantaneous velocity of the residual flow along the probability path. Unlike noise-only flows, this formulation forces the velocity field to align with the global enhancement vector δ, suppressing the chaotic drift.

By solving the target residual field, both the residual direction and the rate of change between any location along the residual flow toward the target distribution can be determined at any time step. Flow Matching adopts a multi-step sampling strategy (Euler method), which integrates the residual field over small time intervals to progressively approximate the ground-truth image and thus achieve low-light enhancement.

However, such a multi-step sampling strategy is undoubtedly time-consuming and may introduce approximation bias. From a field theory perspective, these recursive iterations are redundant for restoration tasks if the global trajectory can be analytically pre-computed.

Inspired by the theoretical framework of Mean Flows [27]—an advanced paradigm enabling one-step inference for flow models—we reformulate the enhancement task by analytically averaging the residual field over the entire probability path. This transformation allows us to collapse the temporal integration into a single-pass, closed-form mapping, enabling the model to recover the target residual directly without incremental updates. By bypassing the iterative sampling process, ARF achieves a superior balance between generative quality and inference efficiency.

2) ARF Evolution Equation: (6) defines the target residual field F_{res}^{tgt}, representing the ideal velocity at each time step. To enable single-step inference, we define the *average residual field* $\overline{F}_{res}$ as the temporal average of F_{res}^{tgt} over the window $[t-h, t]$:

$$h\overline{F}_{res}(x_t, t, h) = \int_{t-h}^{t} F_{res}^{tgt}(x_t, \tau) d\tau \tag{7}$$

By applying the Leibniz integral rule to the definition in (7), we derive the temporal dynamics of the average field:

$$\frac{d}{dt}[h\overline{F}_{\text{res}}] = \frac{d}{dt}\int_{t-h}^{t} F_{\text{res}}^{\text{tgt}}(z_\tau, \tau)\, d\tau \tag{8}$$

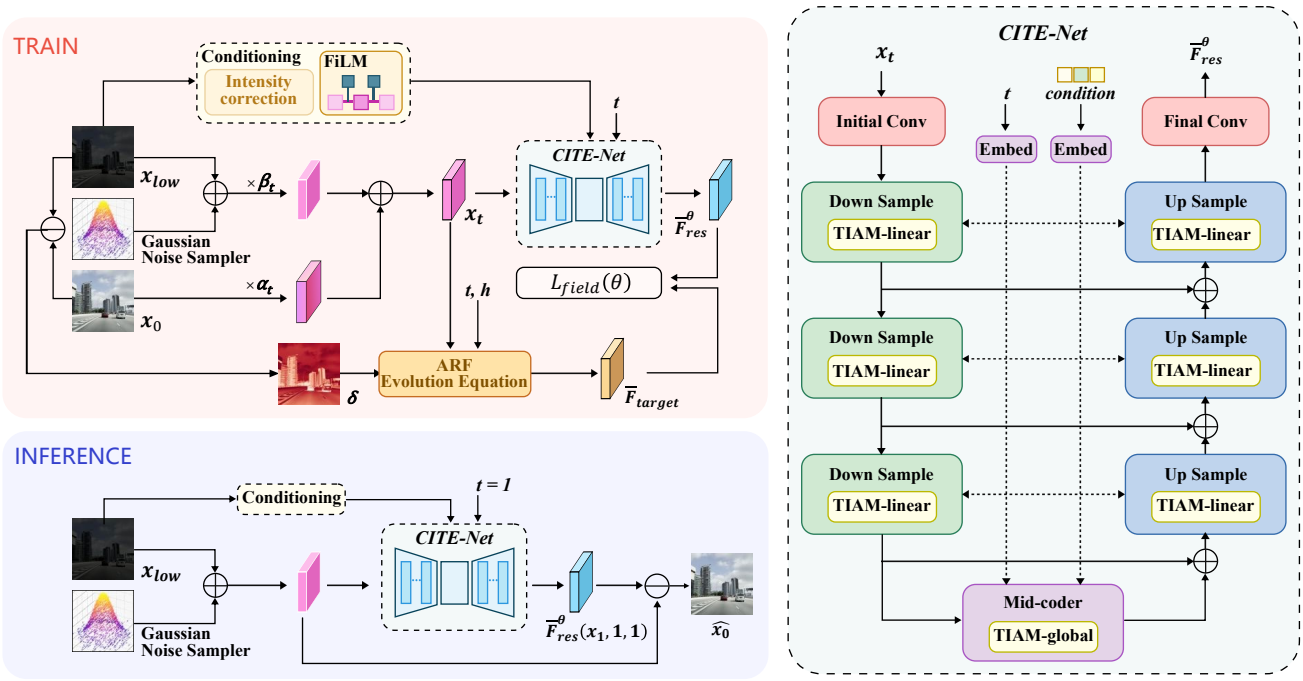

Fig. 3. Overall framework. During training, CITE-Net learns to fit the average residual field at different time steps, correcting the residual flow. At inference, it outputs the residual field at $t = 1$, recovering the true image residual and completing the enhancement. The architecture of CITE-Net is shown on the right; TIAMs denote Temporal–Illumination Attention Modules.

The relationship between the average and target residual field is thus obtained:

$$\overline{F}_{\text{res}} + h\frac{d\overline{F}_{\text{res}}}{dt} = F^{\text{tgt}}_{\text{res}}(x_t, t) \tag{9}$$

We now expand the total derivative term:

$$\frac{d\overline{F}_{\text{res}}}{dt} = \nabla_x \overline{F}_{\text{res}} \cdot v(x_t, t) + \frac{\partial \overline{F}_{\text{res}}}{\partial t} \tag{10}$$

Here, $v(x_t, t)$ denotes the velocity field, which can be interpreted as $\frac{dx_t}{dt}$ or $\frac{d\psi(x_t)}{dt}$. Substituting into the previous relation yields the final field evolution equation, also termed the Average Residual Flow Evolution Equation (**ARF Evolution Equation**):

$$\overline{F}_{\text{res}} = F^{\text{tgt}}_{\text{res}}(x_t, t) - h\left[\nabla_x \overline{F}_{\text{res}} \cdot v(x_t, t) + \frac{\partial \overline{F}_{\text{res}}}{\partial t}\right] \tag{11}$$

This equation consists of the target residual field minus a spatiotemporal correction term, while the factor h reflects the influence of the temporal scale on long-term equilibrium.

In solving the ARF Evolution Equation, the target residual field is predefined (see (6)), whereas the spatiotemporal correction term requires computing the total derivative of the average residual field. This derivative is obtained efficiently using the Jacobian–Vector Product (JVP).

The solution of the equation provides the average residual field, which measures the deviation from the normally illuminated image at a given time step, and thus computes the corresponding residual. For one-step generation ($t = 1, h = 1$), the cumulative residual is directly given by

$$\overline{\delta} = ((t-h)-t) \cdot \overline{F}^{\theta}_{\text{res}}(x_1, 1, 1) = -\overline{F}^{\theta}_{res}(x_1, 1, 1) \tag{12}$$

Following our residual centering, the final enhanced image is reconstructed as:

$$\hat{x}_0 = x_1 + \overline{\delta} \tag{13}$$

The above constitutes the fundamental principles of the Average Residual Flow theory. The following section builds upon these principles to describe the training and inference procedures.

C. CITE-Net

1) Framework of CITE-Net: Based on the above ARF Evolution Equation (11), the training objective is reformulated. A neural network is used to predict the average residual field, parameterized by θ, whose output is denoted as $\overline{F}^{\theta}_{res}$ and referred to as the Conditioned Illumination-Temporal Enhancement Network (CITE-Net). The field evolution equation can then be rewritten as follows:

$$\overline{F}_{target} = F^{tgt}_{res}(x_t, t) - h\left[\nabla_x \overline{F}^{\theta}_{\text{res}} \cdot v(x_t, t) + \frac{\partial \overline{F}^{\theta}_{\text{res}}}{\partial t}\right] \tag{14}$$

where F^{tgt}_{res} is defined the same way as in 6.

From (14), it can be seen that the objective of the neural network is to learn and predict the average residual field during

the enhancement process. By sampling different time steps with varying values of t and h, the network can accurately approximate the average residual field that characterizes the transformation between the two data distributions.

We design the following loss function for the average residual field network:

$$L_{\text{field}}(\theta) = \mathbb{E}_{t,h,x_t} \left\| \overline{F}^{\theta}_{res}(x_t, t, h) - \text{sg}(\overline{F}_{\text{target}}) \right\|_2^2 \quad (15)$$

where $sg(\cdot)$ denotes the stop-gradient operation.

To efficiently model the average residual field, we adopt the neural network architecture illustrated in Fig. III-B1. The time step t and step size h are transformed into temporal embedding vectors via sinusoidal positional encoding, which are used for feature modulation.

The low-light image, refined through intensity correction (e.g., gamma enhancement and histogram equalization) and further modulated by FiLM [28], serves as the conditional input to the U-shaped network.

The encoder consists of three downsampling stages, each comprising two residual blocks and a linear Temporal–Illumination Attention Module (TIAM). In the midcoder stage, global TIAM and lightweight residual blocks are iteratively applied to capture long-range dependencies and refine features. The decoder restores spatial resolution through three upsampling stages, where each stage concatenates the corresponding skip connection and processes it with residual blocks and TIAMs.

To maintain consistency with the conventional Flow Matching algorithm, our training and sampling procedures adopt the specific case of $\alpha_t = 1 - t$ and $\beta_t = t$.

2) Design of TIAM: To enhance the CITE-Net's ability to capture both temporal evolution and spatial illumination, we propose the Temporal–Illumination Attention Module (TIAM), consisting of two variants: Linear Attention for efficient feature modulation and Global Attention for precise contextual aggregation.

Each TIAM uses a time-step embedding to adapt feature aggregation across noise levels and a lightweight illumination estimator to guide attention toward informative regions. Gating parameters γ_I and γ_T balance the contributions of illumination and temporal modulation, initialized to zero so the module starts as standard attention and gradually learns optimal weights.

Linear and global attention adopt multiplicative and additive gating mechanisms, respectively, to integrate temporal and illumination information.

IV. Experiments

A. Datasets and Settings

We conducted extensive comparative experiments across six benchmark settings on the LOLv1 [36], LOLv2 [37], and LoLI-Street [38] datasets. The following section provides an overview of the datasets and the fundamental experimental configurations.

Datasets. LOLv1 and LOLv2 (real and synthetic) provide paired low-/normal-light images for evaluating low-light enhancement methods. LoLI-Street contains 30K training and 3K validation pairs of street scenes with light, moderate, and dense darkness, enabling models to handle diverse real-world low-light scenarios.

Metrics. To evaluate the enhancement performance of ARF and CITE-Net, we adopt Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [44], and Learned Perceptual Image Patch Similarity (LPIPS) [45] as perceptual quality metrics. PSNR measures pixel-level distortions, while SSIM evaluates structural and perceptual similarity in terms of luminance and contrast. LPIPS, computed with AlexNet [46] features, assesses perceptual similarity in deep feature space, where higher PSNR/SSIM and lower LPIPS indicate better visual quality.

Implementation Details. All experiments were conducted on a single NVIDIA RTX 4090 GPU (VRAM 48 GB) using the PyTorch framework. The model was trained with the Adam [47] optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$), starting from a learning rate of 1×10^{-4}, which was gradually decayed to 0 via cosine annealing. Unless otherwise stated, we used a batch size of 4 with a gradient accumulation step of 2 (effective batch size = 8), and trained for 160 K iterations. Both t and $t - h$ follow a uniform distribution over $0, 1$, ensuring that the residual field effectively captures dynamic averages along the probabilistic trajectory and refines intermediate states.

B. Results Comparison

Results on LOL Datasets. Our method achieves consistently strong results across the LOLv1, LOLv2-real, and LOLv2-synthetic datasets. On LOLv1, ARF and CITE-Net surpass the second-best approach by 1.361 dB in PSNR. As summarized in Table I, we attain state-of-the-art (SOTA) results on all three benchmarks; crucially, all results are reported without the use of gt-mean. In Table III, we further compare model parameters and FLOPs with other representative methods. Notably, ARF and CITE-Net reduce the parameter count by nearly 50% compared to the leading flow-based model LLFlow, and by about 10% compared to the best diffusion-based model PyDiff, while still achieving superior SSIM performance. The FLOPs and average inference time (AIT) are evaluated at a resolution of 256 × 256, where AIT is computed over 20 random patches from both LOLv1 and LoLI-Street. Compared to other models, our method attains a better balance between computational complexity and visual quality, maintaining high performance with improved efficiency. Visual comparisons are presented in Fig. 4.

Results on LoLI-Street Dataset. To evaluate our model in real-world driving scenarios, we conducted experiments on three low-light benchmarks (light, moderate, dense) from the LoLI-Street dataset. As shown in Table II, Some models show notable SSIM fluctuations, likely because they can only recover low-light images under specific illumination conditions. In contrast, ARF with CITE-Net achieves the highest SSIM

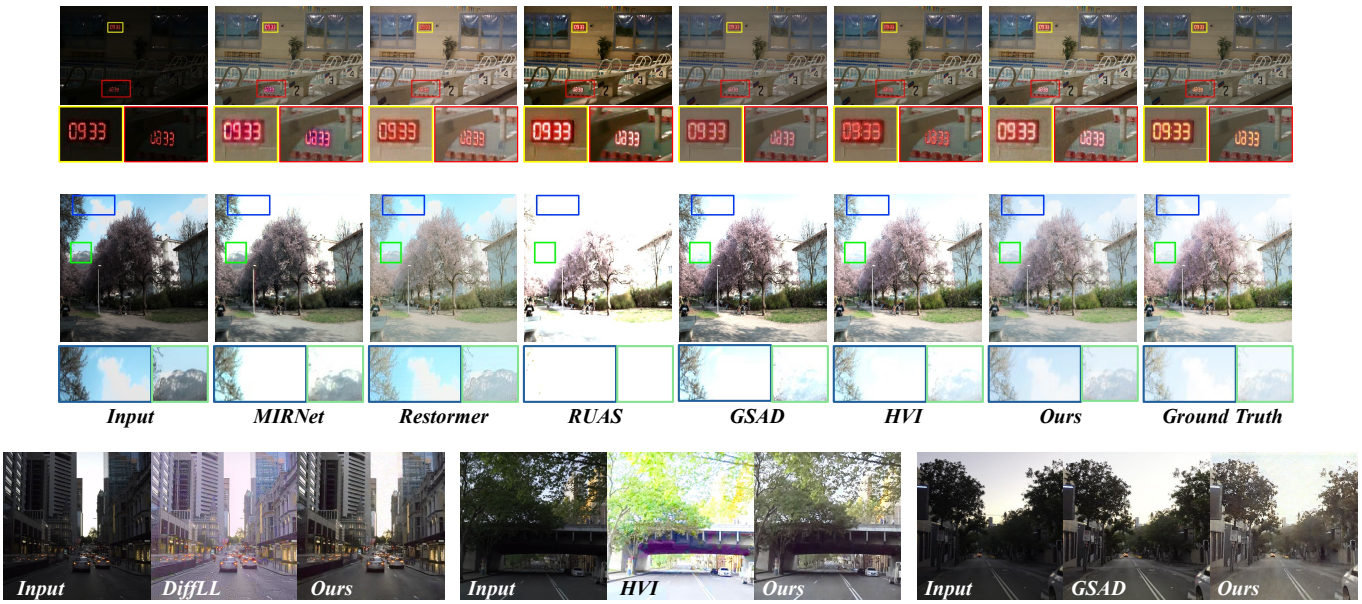

Fig. 4. Qualitative comparison of ARF+CITE-Net with other methods. The three rows, from top to bottom, are sampled from the LOLv2-real, LOLv2-synthetic, and LoLI-Street (unpaired test sets). Our method better preserves fine details and restores realistic illumination.

TABLE I
COMPARISON OF DIFFERENT METHODS ON LOLV1, LOLV2-REAL, AND LOLV2-SYNTHETIC DATASETS. METRICS: PSNR (HIGHER IS BETTER), SSIM (HIGHER IS BETTER), LPIPS (LOWER IS BETTER). IN EACH COLUMN, THE BEST RESULTS ARE **BOLDED** AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Methods	Source	LOLv1			LOLv2-Real			LOLv2-Synthetic		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MIRNet [29]	ECCV'20	22.122	0.830	0.250	20.021	0.820	0.233	22.520	0.899	0.110
RUAS [30]	CVPR'21	18.654	0.518	0.270	15.326	0.488	0.310	13.765	0.638	0.305
LLFlow [11]	AAAI'22	24.064	0.860	**0.117**	17.433	0.831	0.176	24.807	0.919	0.067
CLIP-LIT [31]	ICCV'23	12.394	0.493	0.397	15.262	0.601	0.398	16.190	0.772	0.200
Diff-Retinex [32]	ICCV'23	22.690	0.853	0.191	21.190	0.833	0.271	24.336	0.921	0.061
GSAD [7]	NeurIPS'23	23.224	0.855	0.193	20.177	0.847	0.205	24.098	0.827	0.073
PyDiff [10]	IJCAI'23	23.298	0.858	0.215	23.364	0.834	0.208	25.126	0.917	0.098
FourierDiff [5]	CVPR'24	17.560	0.607	0.359	18.673	0.602	0.362	13.699	0.631	0.252
LightenDiffusion [6]	ECCV'24	20.188	0.814	0.316	21.097	0.847	0.305	21.542	0.866	0.188
AnlightenDiff [33]	TIP'24	21.762	0.811	0.252	20.423	0.825	0.267	20.581	0.874	0.146
URetinexNet++ [34]	TPAMI'25	23.826	0.839	0.231	21.967	0.836	0.203	24.603	0.927	0.102
HVI [35]	CVPR'25	23.809	0.857	0.188	23.427	0.862	0.169	25.129	0.939	0.070
AGLLDiff [19]	AAAI'25	21.813	0.841	0.153	22.122	0.837	0.206	21.113	0.868	0.130
Ours	—	**25.425**	**0.923**	0.121	**23.572**	**0.870**	**0.161**	**25.155**	**0.945**	**0.060**

and PSNR across all scenarios, enabling stable performance with SSIM differences below 0.01 across scenarios.

C. Ablation Study

Parameter ξ and Scheduler . We adopt $\xi = 1$ with a linear schedule $\alpha_t = 1-t$ and $\beta_t = t$ for most experiments, owing to its simplicity and smooth trajectories. To validate this choice, we compare it with a cosine scheduler $\xi = \frac{\pi}{2}\sin(\pi t)$, which increases the residual field weights initially and then decreases them, and a quadratic scheduler $\xi = 2t$, which continuously increases the weights over time. Results on LOLv1 confirm that the linear schedule remains superior, enabling CITE-Net to fit the average residual field more effectively (Table IV).

CITE-Net. To verify the effectiveness of each module in CITE-Net, we conducted ablation studies under the ARF framework. Specifically, we replaced CITE-Net with SongUNet [48], removed Condition extraction in favor of Condition concatenation, and disabled the TIAM by keeping the parameters of its two gating units at 0. Table V presents the detailed results of our module ablations. It can be observed that CITE-Net achieves a PSNR improvement of 3.918dB over the standard UNet on the LOLv1 datasett, demonstrating the effectiveness of the proposed architectural design..

TABLE II
QUANTITATIVE COMPARISON ON THREE LOW-LIGHT BENCHMARKS FROM THE LOLI-STREET DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Dataset	Metric	RetinexF. [39]	RQ-LLIE [40]	PairLIE [41]	FourLLIE [42]	SCI [43]	LLFlow [11]	AGLLDiff [19]	PyDiff [10]	Ours
LoLI-Street-light	PSNR↑	27.92	27.66	28.78	28.06	27.84	31.04	27.56	30.78	**32.69**
	SSIM↑	0.8767	0.8811	0.9169	0.8363	0.8759	0.9327	0.9053	0.9254	**0.9469**
LoLI-Street-moderate	PSNR↑	28.44	27.59	28.42	27.96	28.38	30.23	27.11	30.01	**32.16**
	SSIM↑	0.6197	0.8829	0.9242	0.8693	0.6192	0.9104	0.8521	0.8856	**0.9406**
LoLI-Street-dense	PSNR↑	27.87	29.03	27.66	28.07	27.79	29.55	27.32	29.62	**31.08**
	SSIM↑	0.3396	0.9167	0.8702	0.8828	0.3394	0.8956	0.8207	0.8402	**0.9305**

TABLE III
COMPUTATIONAL COMPLEXITY COMPARISON BETWEEN OUR METHOD AND PREVIOUS FLOW AND DIFFUSION METHODS. SSIM IS EVALUATED ON THE LOLV1 DATASET. AIT DENOTES THE AVERAGE INFERENCE TIME(S).

Method	Parameter (M)	FLOPs (G)	SSIM↑	AIT(s)
LLFlow [11]	17.42	358.40	0.860	0.2134
GSAD [7]	17.36	442.02	0.855	0.3066
PyDiff [10]	97.19	708.68	0.858	0.2541
DiffLL [9]	20.09	89.60	0.845	0.1453
ReDDiT [8]	17.43	67.02	0.872	0.2064
Ours	9.67	37.60	**0.923**	0.0591

TABLE IV
ABLATION STUDY ON ξ AND SCHEDULERS.

ξ	Schedulers	SSIM↑	LPIPS↓
$2t$	$\alpha_t = 1 - t^2,\ \beta_t = t^2$	0.896	0.141
$\frac{\pi}{2}sin(\pi t)$	$\alpha_t = cos^2(\frac{\pi}{2}t),\ \beta_t = sin^2(\frac{\pi}{2}t)$	0.914	0.128
1	$\alpha_t = 1 - t,\ \beta_t = t$	0.923	0.121

Revised DDIM and Rectified Flow. We compare the proposed ARF method with representative diffusion and flow approaches, including DDIM and Rectified Flow, as shown in Table VI. In all comparisons, we adopt CITE-Net as the backbone network, where CITE-Net is used to predict the noise in DDIM and the velocity field in flow matching on LOLv1. The "Revised" variants utilize the ARF-style initial distribution—a Gaussian distribution centered at the low-light image—while the standard versions use a typical zero-centered normal distribution. Experimental results show that ARF achieves the best performance with the fewest time steps.

V. CONCLUSION

In this work, we propose ARF, a novel flow matching method for low-light enhancement, and CITE-Net, a model that accurately predicts the average residual field. By averaging residual fields in the residual space and leveraging CITE-Net to predict the averaged residual through temporal and illumination cues, our approach compresses the original multi-step iteration into a single forward pass. This significantly improves inference speed, enabling ARF and CITE-Net to meet the real-time requirements of nighttime autonomous driving. Experiments on six benchmarks show our method achieves state-of-the-art performance, offering new insights and theoretical support for low-light enhancement in nighttime autonomous driving.

TABLE V
ABLATION STUDY ON CITE-NET MODULES UNDER THE ARF FRAMEWORK.

Backbone	Condition	Attention	PSNR↑	SSIM↑	LPIPS↓
✗	✗	✗	22.020	0.791	0.241
✓	✗	✗	24.337	0.873	0.153
✓	✓	✗	24.879	0.901	0.136
✓	✓	✓	25.425	0.923	0.121

TABLE VI
COMPARISON OF REVISED DDIM, REVISED RECTIFIED FLOW, AND ARF USING CITE-NET ON LOLV1.

Method	Num_steps	PSNR↑	SSIM↑	LPIPS↓
DDIM	10	22.472	0.855	0.201
Revised DDIM	10	23.523	0.869	0.187
Rectivied Flow	5	24.803	0.885	0.154
Revised Rectified Flow	5	25.126	0.909	0.129
ARF	1	25.425	0.923	0.121

REFERENCES

[1] Z. Du, M. Shi, and J. Deng, "Boosting object detection with zero-shot day-night domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12666–12676, 2024.

[2] H. Gong, Z. Li, C. Lu, G. Du, and J. Gong, "Leveraging multi-stream information fusion for trajectory prediction in low-illumination scenarios: A multi-channel graph convolutional approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 5, pp. 3854–3869, 2023.

[3] J. Li, B. Li, Z. Tu, X. Liu, Q. Guo, F. Juefei-Xu, R. Xu, and H. Yu, "Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15205–15215, 2024.

[4] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9396–9416, 2021.

[5] X. Lv, S. Zhang, C. Wang, Y. Zheng, B. Zhong, C. Li, and L. Nie, "Fourier priors-guided diffusion for zero-shot joint low-light enhancement and deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25378–25388, 2024.

[6] H. Jiang, A. Luo, X. Liu, S. Han, and S. Liu, "Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models," in *European Conference on Computer Vision*, pp. 161–179, Springer, 2024.

[7] J. Hou, Z. Zhu, J. Hou, H. Liu, H. Zeng, and H. Yuan, "Global structure-aware diffusion process for low-light image enhancement," *Advances in Neural Information Processing Systems*, vol. 36, pp. 79734–79747, 2023.

[8] G. Lan, Q. Ma, Y. Yang, Z. Wang, D. Wang, X. Li, and B. Zhao, "Efficient diffusion as low light enhancer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21277–21286, 2025.

[9] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–14, 2023.

[10] D. Zhou, Z. Yang, and Y. Yang, "Pyramid diffusion models for low-light image enhancement," *arXiv preprint arXiv:2305.10028*, 2023.

[11] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. Kot, "Low-light image enhancement with normalizing flow," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 2604–2612, 2022.

[12] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[13] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[14] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*, pp. 8162–8171, PMLR, 2021.

[15] S. Lin, B. Liu, J. Li, and X. Yang, "Common diffusion noise schedules and sample steps are flawed," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5404–5411, 2024.

[16] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu, "Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 2654–2662, 2023.

[17] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

[18] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "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, 2022.

[19] Y. Lin, T. Ye, S. Chen, Z. Fu, Y. Wang, W. Chai, Z. Xing, W. Li, L. Zhu, and X. Ding, "Aglldiff: Guiding diffusion models towards unsupervised training-free real-world low-light image enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 5307–5315, 2025.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[21] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[22] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[23] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," *arXiv preprint arXiv:2209.03003*, 2022.

[24] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13095–13105, 2023.

[25] H. Qin, W. Luo, L. Wang, D. Zheng, J. Chen, M. Yang, B. Li, and W. Hu, "Reversing flow for image restoration," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7545–7558, 2025.

[26] J. Liu, Q. Wang, H. Fan, Y. Wang, Y. Tang, and L. Qu, "Residual denoising diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2773–2783, 2024.

[27] Z. Geng, M. Deng, X. Bai, J. Z. Kolter, and K. He, "Mean flows for one-step generative modeling," *arXiv preprint arXiv:2505.13447*, 2025.

[28] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[29] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *European conference on computer vision*, pp. 492–511, Springer, 2020.

[30] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10561–10570, 2021.

[31] Z. Liang, C. Li, S. Zhou, R. Feng, and C. C. Loy, "Iterative prompt learning for unsupervised backlit image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8094–8103, 2023.

[32] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "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, 2023.

[33] C.-Y. Chan, W.-C. Siu, Y.-H. Chan, and H. A. Chan, "Anlightendiff: Anchoring diffusion probabilistic model on low light image enhancement," *IEEE Transactions on Image Processing*, 2024.

[34] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "Interpretable optimization-inspired unfolding network for low-light image enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[35] Q. Yan, Y. Feng, C. Zhang, G. Pang, K. Shi, P. Wu, W. Dong, J. Sun, and Y. Zhang, "Hvi: A new color space for low-light image enhancement," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5678–5687, 2025.

[36] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[37] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE Transactions on Image Processing*, vol. 30, pp. 2072–2086, 2021.

[38] M. T. Islam, I. Alam, S. S. Woo, S. Anwar, I. Lee, and K. Muhammad, "Loli-street: Benchmarking low-light image enhancement and beyond," in *Proceedings of the Asian Conference on Computer Vision*, pp. 1250–1267, 2024.

[39] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12504–12513, 2023.

[40] Y. Liu, T. Huang, W. Dong, F. Wu, X. Li, and G. Shi, "Low-light image enhancement with multi-stage residue quantization and brightness-aware attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12140–12149, 2023.

[41] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding, and K.-K. Ma, "Learning a simple low-light image enhancer from paired low-light instances," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22252–22261, 2023.

[42] C. Wang, H. Wu, and Z. Jin, "Fourllie: Boosting low-light image enhancement by fourier frequency information," in *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7459–7469, 2023.

[43] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "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, 2022.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[47] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[48] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.