LoLI-Street: Benchmarking Low-Light Image Enhancement and Beyond

Md Tanvir Islam[1], Inzamamul Alam[1], Simon S. Woo[1], Saeed Anwar[2], IK Hyun Lee[3], and Khan Muhammad[3,*]

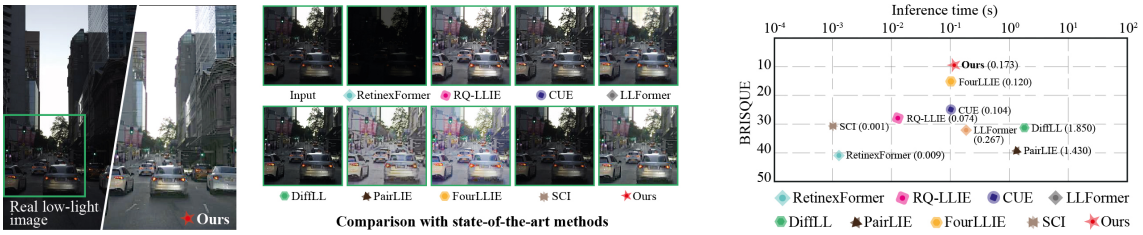
[1] Department of Software, Sungkyunkwan University, Suwon 16419, South Korea
[2] The Australian National University, Canberra 0200, Australia
[3] Department of Mechatronics Engineering, Tech University of Korea, Siheung-Si, 15073, South Korea
[4] Department of Human-AI Interaction, Sungkyunkwan University, Seoul 03063, South Korea
*Corresponding author: khanmuhammad@g.skku.edu
Code and dataset: https://github.com/tanvirnwu/TriFuse

Abstract. Low-light image enhancement (LLIE) is essential for numerous computer vision tasks, including object detection, tracking, segmentation, and scene understanding. Despite substantial research on improving low-quality images captured in underexposed conditions, clear vision remains critical for autonomous vehicles, which often struggle with low-light scenarios, signifying the need for continuous research. However, LLIE models and LLIE-paired datasets are scarce, particularly for street scenes, limiting the development of robust LLIE methods. Despite using advanced transformers and/or diffusion-based models, current LLIE methods struggle in real-world low-light conditions and lack training on street-scene datasets, limiting their effectiveness for autonomous vehicles. To bridge these gaps, we introduce a new large-scale dataset "LoLI-Street" (Low-Light Images of Streets) with $33k$ paired low-light and well-exposed images from street scenes in developed cities, covering $19k$ object classes for object detection, including Person, Bicycle, Car, Bus, Motorcycle, and Traffic Light, etc. LoLI-Street dataset also features 1,000 real low-light test images, providing a benchmark for evaluating models under real-world conditions. Furthermore, we propose a transformer and diffusion-based LLIE model named "TriFuse". Leveraging the LoLI-Street dataset, we train and evaluate our TriFuse and other SOTA models to benchmark our dataset. Comparing various models, the feasibility of our dataset for generalization is evident in testing across different mainstream datasets by significantly enhancing low-quality images and object detection for practical applications in autonomous driving and surveillance systems.

Keywords: Low-light image enhancement · LoLI-Street dataset· Conditional noise diffusion · Diffusion denoising · Transformers

1 Introduction

Low-light environments can pose significant challenges for a wide range of computer vision tasks in our daily lives. For most computer vision tasks, models are typically trained on datasets collected during the day with sufficient lighting, making them less effective in dark or low-light environments. This limitation poses a significant challenge as the underlying datasets do not account for the variations and complexities found in real-world low-light conditions.



(a) (b)

Fig. 1: Comparison between our TriFuse method and SOTA models using a sample real low-light test image from our LoLI-Street dataset. (a) Qualitative comparison: Visually, PairLIE and RQ-LLIE produce brighter outputs but lack realism. In contrast, TriFuse ensures high visual quality with realistic enhancements. (b) Quantitative comparison based on the no-reference metric BRISQUE (↓) and inference time (↓).

Thus, as daylight fades into night, the reduced visibility can hinder the ability to perform even the most basic tasks for computer vision systems. This issue is a matter of convenience, safety, and efficiency. To address these practical challenges, advancements in computer vision technology are crucial. Such systems can significantly assist in low-light conditions, enhance the vision capabilities of autonomous vehicles [35,29], and improve safety and security measures [29]. The importance of computer vision in mitigating the effects of low-light conditions underscores its potential impact on a wide range of applications [42]. For instance, recent advancements in image processing and machine learning have led to sophisticated algorithms that enhance image clarity [27], detecting [36] and recognizing [45] objects in near-darkness, significantly advancing computer vision. Additionally, with the rise of deep learning [40,51,37], transformers [46,32,6,46], and diffusion methods [55,25,54,47], its powerful feature representation capabilities led to the rapid adoption of LLIE. Moreover, in addition to traditional methods, researchers are now exploring the latest transformer and diffusion-based methods for LLIE by utilizing synthetic datasets and reporting significant improvements in LLIE.

Moreover, the models struggle to perform in real-world low-light conditions, which is a huge gap that leaves the scope to develop robust methods for effective LLIE in real-world scenarios. Thus, the full potential of these methods for LLIE has not yet been fully explored and requires further research. Furthermore, these learning-based methods heavily rely on high-quality labeled data for training to perform accurately in real-world scenarios. In the current literature, different datasets are available with different scene types of images under various low-light conditions [48,31,10,11,28,14]. Despite having several LLIE datasets, there is still a lack of datasets, especially for urban street image scene types, which can be used to train the LLIE models for autonomous vehicles to use navigation and surveillance cameras in urban street scenarios where accurate object detection, recognition, and navigation are crucial for safety.

Recognizing the aforementioned gaps, we introduce a unique street-scene LLIE dataset and propose a novel method that outperforms mainstream SOTA LLIE methods. Our contributions are summurized as follows:

– We introduce a unique and challenging dataset named LoLI-Street consisting of 30,000 train, 3,000 validation, and 1,000 real low-light test (RLLT) images for training and evaluating LLIE models. LoLI-Street dataset has street scene types, which are rare among the existing datasets and feature three intense levels (high, moderate, and light) of low-light effect.

- We propose "TriFuse", a transformer and diffusion-based low-light enhancement model that integrates a vision transformer, wavelets, and an edge sharpening module. TriFuse reduces the number of sampling steps in the diffusion process by using the transformer as an accurate noise predictor.
- Benchmarking our proposed TriFuse method against SOTA LLIE models on LoLI-Street real low-light testset and mainstream datasets, we found it excels both quantitatively and qualitatively in LLIE and object detection, as shown in Fig. 1 and detailed in Section 5.
- Leveraging the LoLI-Street dataset, we benchmark the SOTA LLIE methods, which will work as a foundation for future LLIE research, especially for street-type scenes where object detection tasks are vital for various autonomous vehicles and surveillance cameras under low-light conditions.

2 Related Works

2.1 LLIE Datasets

Several video and image-based LLIE datasets are available in the literature. This section covers mainstream image-based LLIE datasets related to LoLI-Street. For example, the ExDARK dataset [31] includes 7,363 annotated images across 12 classes, crucial for low-light object detection. The LLVIP dataset [24] provides 15,488 pairs of visible and infrared images, essential for image fusion and pedestrian detection. The MIT-Adobe FiveK dataset [3] offers 5,000 indoor and outdoor images for various enhancement tasks. The SICE [4] dataset synthesizes 589 images across varied illumination conditions, while the SID [10] dataset pairs 5,094 short-exposure images with long-exposure references. Additionally, the LIME [17] dataset features 10,000 images for LLIE in low-light conditions, and the DPED [23] dataset enhances mobile photo quality. The LOLv1 [49] and LOLv2 [52] datasets contain paired high and low-light images, and the LSRW dataset [19] includes paired low-light images. These datasets have indoor and outdoor scenes, as presented in Table 1a. To the best of our knowledge, there is no dataset that presents the street scene types, unlike our proposed LoLI-Street dataset, which is crucial for autonomous vehicles under real-world low-light street scenarios. Moreover, our LoLI-Street provides a test set of 1000 street scene-type images under real-life low-light conditions unparalleled in literature, which can be used to benchmark the LLIE models.

2.2 LLIE Methods

Transformer-based LLIE. Initially proposed for natural language processing [43], transformers have recently shown remarkable performance in computer vision tasks, such as image classification [1,2,13], semantic segmentation [7,50,59], and object detection [8,12]. They have also proven effective in low-level vision tasks like image restoration [6,46] and image synthesis [22,26,57]. Recent studies highlight transformers' effectiveness in image enhancement by utilizing illumination-guided multi-head self-attention mechanisms to improve interactions between regions of different exposure levels [6].

Diffusion-based LLIE. Diffusion-based models have shown significant potential in LLIE by leveraging their generative capabilities to handle various degradations, including noise [56], low contrast [33], color correction [9], and medical-image denoising [16,39]. Recent advancements include

(a) (b)

Fig. 2: Sample images and distribution of our dataset. (a) Sample images of LoLI-Street. **Green:** train and validation sets, **Red:** real low-light test set. (b) Distribution of the low-light images across various subsets of the LoLI-Street dataset.

Table 1: Quantatitive comparison of mainstream datasets and our LoLI-Street dataset.

(a) Datasets comparison.

Dataset Name	Venue	Image Quantity
LOLv1 [49]	BMVC'18	1,500
SICE [5]	IEEE TIP'18	5,389
ExDARK [31]	CVIU'19	7,363
LOLv2 [53]	CVPR'20	3,576
LLVIP [24]	ICCV'21	15,488
LSRW [19]	JVCIR'23	11,300
LoLI-Street (Ours)	**ACCV'24**	**34,000**

(b) Quantitative analysis of our LoLI-Street dataset.

Metrics	Train			Validation			Test
	Light	Moderate	Dense	Light	Moderate	Dense	
PSNR↑	28.35	27.88	27.89	28.44	27.91	27.87	-
SSIM↑	0.8564	0.6045	0.3528	0.8767	0.6196	0.3398	-
MS-SSIM↑	0.9531	0.7943	0.5854	0.9422	0.7818	0.5621	-
LPIPS↓	0.0410	0.1547	0.2988	0.04199	0.1563	0.2490	-
MSE↓	106.14	95.71	105.99	106.79	94.77	108.02	-
MAE↓	204.51	166.09	137.21	206.09	169.68	142.15	-
BRISQUE↓	21.99	24.82	33.46	15.80	18.00	26.34	30.99
NIQE↓	11.045	12.119	13.352	10.49	10.49	10.49	12.334

Diff-Retinex [54], which combines Retinex with generative diffusion networks for enhanced detail and noise reduction. Integrating generative networks with physical models has led to effective restoration of scene structures [21]. Other recent research has introduced a conditional diffusion model incorporating multi-scale patch-based training [38] and a wavelet-based conditional diffusion model [25], improving visual quality. Also, CLE Diffusion [55] offers controllable light enhancement using classifier-free guidance, while ExposureDiffusion [47] combines a diffusion model with a physics-based exposure model for enhanced performance and reduced inference time. And, LDM integrates a denoising diffusion probabilistic model with a light enhancement network, achieving state-of-the-art performance for LLIE [33].

Despite recent advancements, existing LLIE methods struggle with real-world low-light images, especially on our challenging LoLI-Street dataset, making them unsuitable for autonomous vehicles and surveillance cameras. Our proposed model, TriFuse, leverages this street-scene dataset to achieve significant improvements over SOTA models on the RLLT.

3 Methodology

3.1 Our Dataset: LoLI-Street

We introduce the benchmark dataset "Low-light Images of Streets (LoLI-Street)", containing three subsets: train, validation, and test. The train and validation sets consist of $30k$ and $3k$ paired low and high-light images and the real low-light testset (RLLT) contains $1k$ images under real-world low-light conditions, totaling $33k$ images. We collected high-resolution videos (4K/8K at 60fps) from various cities under low-light conditions, extracting and manually reviewing frames to create the Real Low-light Testset (RLLT) of our LoLI-Street dataset, ensuring high quality and excluding any with motion blur. As shown in Table 1b, LoLI-Street encompasses three levels of low-light intensity, resulting in different quantitative metrics. Sample images are presented in Fig. 2a, and Fig. 2b shows the average pixel distribution across subsets. Inspired by [41,**?**], we used Photoshop v25.0 to generate the synthetic images of our dataset and examined the distribution of the images. As evident from Fig. 2b, the distribution of our dataset varied across the subsets, which is crucial for generalizing LLIE models.

3.2 Our Proposed Method

Our proposed method, TriFuse, integrates a custom vision transformer, wavelet-based conditional diffusion denoising, and an edge-sharpening module for effective LLIE. Each module of TriFuse is detailed below:

Discrete Wavelet Transformation (DWT). We use DWT to decompose a given low-light image $I_{\text{low}} \in \mathbb{R}^{H\times W\times C}$ in various low and high-frequency components. The 2D-DWT with Haar wavelets [18] decomposes the image into four sub-bands: A_1^{low}, V_1^{low}, H_1^{low}, and D_1^{low}, as illustrated in Fig. 3. The mathematical formulation for the 2D-DWT is provided in Eq. (1):

$$\{A_1^{\text{low}}, V_1^{\text{low}}, H_1^{\text{low}}, D_1^{\text{low}}\} = \text{2D-DWT}(I_{\text{low}}), \tag{1}$$

where A_1^{low} is the approximation coefficient representing the low-frequency information, and V_1^{low}, H_1^{low}, and D_1^{low} are the coefficients representing the vertical, horizontal, and diagonal high-frequency information, respectively. Focusing the diffusion process on these components, especially the average coefficients, our method enhances the model's ability to handle global image structures effectively.

TriFuse. TriFuse integrates a custom transformer, CNN, Encoder, and Decoder block, incorporating the diffusion process for predicting noise at each timestamp, forming the cornerstone of our conditional noise generation for the diffusion denoising mechanism. This approach leverages the power of transformers to accurately predict and adjust noise at each diffusion timestep of denoising diffusion probabilistic models (DDPM) [20], enhancing the denoising process and ultimately improving LLIE.

In the forward diffusion process in Eq. (2), the input image x_0 is progressively corrupted into a noisy version x_T over T steps, governed by a variance schedule $\{\beta_1, \beta_2, \ldots, \beta_T\}$ in the following way:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{2}$$

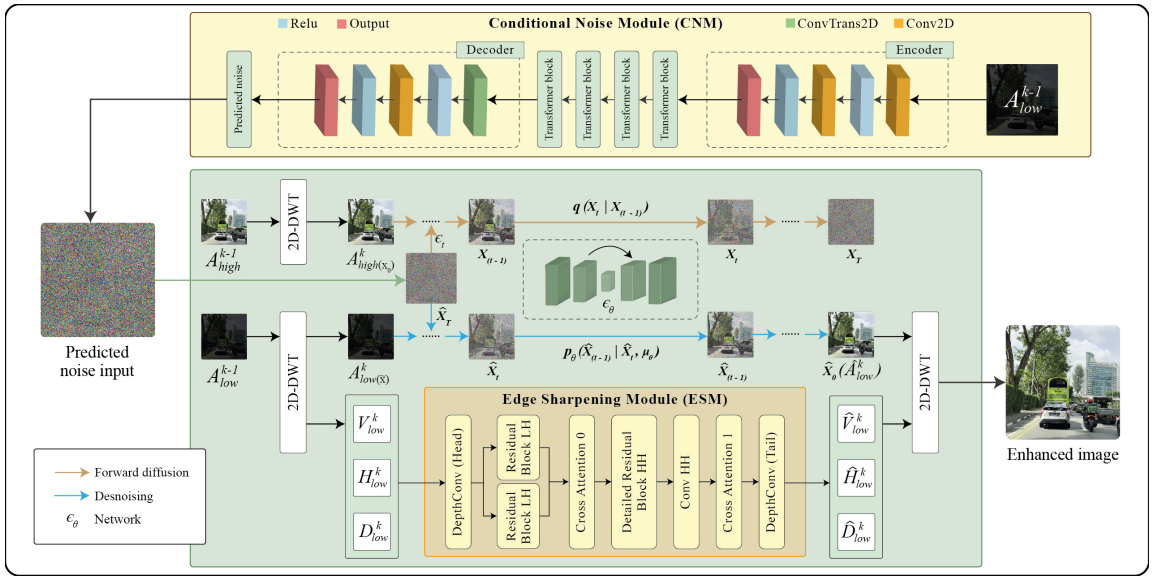


Fig. 3: Overview of our TriFuse model, featuring the Conditional Noise Module (CNM) and Edge Sharpening Module (ESM) for effective LLIE. The CNM generates noise, refined through forward and backward passes by the diffusion process. The ESM sharpens the output image's edges. The process starts with predicted noise, undergoes wavelet transformation and denoising within TriFuse, and results in a visually enhanced image.

where X_t is noisy data at timestep t, and β_t is the variance schedule parameters.

The reverse diffusion process in Eq. (3) involves learning to denoise the noisy image x_T back to a clean image x_0 through a series of Gaussian denoising transitions as follows:

$$p_\theta(\hat{x}_{0:T}) = p(\hat{x}_T) \prod_{t=1}^{T} p_\theta(\hat{x}_{t-1} \mid \hat{x}_t), \quad p_\theta(\hat{x}_{t-1} \mid \hat{x}_t) = \mathcal{N}(\hat{x}_{t-1}; \mu_\theta(\hat{x}_t, t), \sigma_t^2 \mathbf{I}). \tag{3}$$

Here, μ_θ is the predicted mean, and σ_t is the variance, both of which are learned parameters.

Conditional Noise Module (CNM) for Diffusion Denoising. The CNM is designed to predict the noise ϵ_t at each timestep t, utilizing a transformer-based architecture to grasp the intricate patterns in noise and image details. Our model utilizes self-attention mechanisms to capture long-range dependencies and contextual information, unlike traditional diffusion models that rely on random Gaussian noise at each timestep. By conditioning the noise on the input image and the timestep, our CNM significantly enhances the denoising process.

The CNM architecture begins by encoding the input image into a higher-dimensional space using convolutional layers, which extract detailed feature representations. These encoded features are then flattened and processed through a series of transformer blocks. Within these blocks, the self-attention mechanism enables the model to assess the importance of different image parts, effectively predicting the noise to be added or removed. After transforming the features through self-attention and feed-forward layers, the output is reshaped back to the original feature map dimensions and

passed through a decoder. This decoder reconstructs the predicted noise map, guiding the diffusion process.

The CNM's ability to model complex dependencies and incorporate contextual information results in superior image restoration, particularly in challenging low-light conditions. By accurately predicting and controlling the noise at each diffusion step, the CNM ensures an effective and precise denoising process, preserving fine details and maintaining contextual awareness throughout.

This integration enhances image quality by preserving fine details, maintaining contextual awareness, and providing adaptive denoising. Mathematically, the noise prediction is expressed as $\epsilon_\theta(\hat{x}_t, t) = \text{CNM}(\hat{X}_T)$. After integrating our custom CNM with the diffusion denoising process shown in Eq. (3), it can be expressed as Eq. (4) as follows:

$$\hat{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \text{CNM}(\hat{X}_T) \right) + \sigma_t \eta \tag{4}$$

where α_t and $\bar{\alpha}_t$ are predefined noise schedules, and η represents Gaussian noise.
Overall, this novel approach ensures that the denoising process is both effective and precise by accurately predicting and controlling the noise at each diffusion step. The integration of the CNM enhances image quality by ensuring that the noise prediction is conditional on both the image content and the timestep, leading to superior restoration of image details in low-light conditions.

Edge Sharpening Module (ESM). ESM plays a critical role in enhancing the sharpness and clarity of edges in the restored images. It focuses on the high-frequency components obtained from the DWT, ensuring that fine details and textures are well preserved during the restoration process.

The ESM comprises several sophisticated components designed to handle high-frequency information efficiently. Depthwise convolutions capture channel-wise spatial information effectively, ensuring that the model can focus on intricate details without increasing computational complexity. Dilated Residual Blocks (ϕ) preserve the input's spatial resolution while capturing multi-scale features as provided in Eq. (5). Using dilated convolutions allows the network to have a larger receptive field, which is essential for capturing contextual information at multiple scales without losing fine details.

$$\mathbf{Y} = \mathbf{X} + \text{Conv}(\text{ReLU}(\text{BN}(\text{Conv}(\text{ReLU}(\text{BN}(\mathbf{X})))))), \tag{5}$$

where $\mathbf{X}$ denotes the input feature map that enters the Dilated Residual Block, and $\mathbf{Y}$ is the output feature map after processing through the block. Conv, ReLU, and BN denote convolution, Rectified Linear Unit, and Batch Normalization, respectively. Cross-attention mechanisms are used to align and integrate contextual information across different directions (vertical, horizontal, and diagonal). The cross-attention mechanism is defined in Eq. (6) as follows:

$$\mathbf{A}_{\text{attn}} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \right) \mathbf{V}, \tag{6}$$

where $\mathbf{Q} = \text{Conv}(\mathbf{X})$, $\mathbf{K} = \text{Conv}(\mathbf{X})$, $\mathbf{V} = \text{Conv}(\mathbf{X})$ are the query, key, and value matrices, and d_k is the dimensionality of the key vectors. The ESM processes the high-frequency components as given in Eq. (7):

$$\text{ESM}(x) = x + \text{Conv}(\text{Concat}(\phi_{\text{HL}}(x_{\text{HL}}), \phi_{\text{LH}}(x_{\text{LH}}), \phi_{\text{HH}}(x_{\text{HH}}))), \tag{7}$$

where $x_{\text{HL}}, x_{\text{LH}}, x_{\text{HH}}$ are the high-frequency components and ϕ_{HL}, ϕ_{LH}, ϕ_{HH} are the corresponding dilated residual blocks, respectively. By integrating these components, the ESM effectively enhances the sharpness of edges and preserves the fine details in the restored images, addressing one of the critical challenges in LLIE.

Table 2: Quantitative evaluation of the SOTA models on the validation set of LoLI-Street dataset using the pre-trained weights of each model.

Models	RetinexF. [6]	RQ-LLIE [30]	CUE [58]	LLFormer [46]	DiffLL [25]	PairLIE [15]	FourLLIE [44]	SCI [34]
Venue	ICCV'23	ICCV'23	ICCV'23	AAAI'23	TOG'23	CVPR'23	ACM MM'23	CVPR'22
Metrics				Light				
PSNR↑	27.89	28.11	27.66	**28.42**	28.16	<u>28.17</u>	28.06	27.62
SSIM↑	0.1900	0.7382	<u>0.9000</u>	**0.9274**	0.8932	0.8374	0.8363	0.7877
MS-SSIM↑	0.4200	0.8506	<u>0.9331</u>	**0.9497**	0.9148	0.8562	0.8196	0.7890
LPIPS↓	0.4300	0.1637	<u>0.0603</u>	**0.0486**	0.1069	0.1574	0.1996	0.2415
MSE↓	110.33	101.30	111.58	**94.49**	99.48	<u>99.36</u>	101.99	112.51
MAE↓	93.01	93.01	**45.36**	86.41	114.12	70.74	87.12	<u>69.35</u>
Metrics				Moderate				
PSNR↑	27.73	27.88	<u>28.02</u>	27.99	**28.16**	28.00	27.96	27.81
SSIM↑	0.0900	0.5145	0.6389	**0.9018**	0.8709	0.8651	0.5397	<u>0.8737</u>
MS-SSIM↑	0.3000	0.7552	0.7964	**0.9468**	0.8914	0.8764	0.8335	<u>0.9083</u>
LPIPS↓	0.6000	0.2321	0.1606	**0.0530**	0.1365	0.1294	0.1831	<u>0.1021</u>
MSE↓	106.87	106.20	103.98	104.02	**99.41**	<u>103.09</u>	104.14	107.61
MAE↓	124.52	124.51	85.66	**44.76**	119.65	101.05	113.81	<u>48.02</u>
Metrics				Dense				
PSNR↑	27.75	27.87	27.94	**28.67**	<u>28.56</u>	28.06	28.07	28.16
SSIM↑	0.0300	0.3498	0.3651	**0.9056**	0.8744	<u>0.8915</u>	0.8828	0.8758
MS-SSIM↑	0.2100	0.5987	0.5916	<u>0.9488</u>	0.9322	0.8963	0.8780	**0.9633**
LPIPS↓	0.7900	0.2889	0.2971	<u>0.0441</u>	0.0795	0.1016	0.1187	**0.0274**
MSE↓	110.89	106.29	104.52	**89.01**	<u>91.05</u>	101.65	101.52	99.36
MAE↓	116.85	116.85	<u>111.97</u>	**80.34**	160.96	144.18	186.56	212.06

Overall, our proposed TriFuse model produces high-quality, sharp images by combining the ESM and CNM modules in the diffusion denoising process, making it an efficient solution for LLIE and suitable for various real-world applications.

4 Experimental Setup

In this section, we comprehensively investigate and validate the performance of our proposed TriFuse model across various benchmark datasets.

Datasets. We use the train set of $30k$ paired images from our LoLI-Street dataset to train the models and validate the models' performance on the synthetic validation set of $3k$ paired images. Furthermore, we evaluate the models on the real test set of the LoLI-Street dataset, including $1k$ unpaired images. We used the well-known LOLv1 and LOLv2 datasets to evaluate the pre-trained and trained weights of each model and compare the existing models' performance with our TriFuse model. From LOLv1, we use the validation set, which features 15 paired real low-light images to evaluate the models. Similarly, from LOLv2, we take the synthetic and real validation subsets of 100 paired images from each to assess the models. The LSRW dataset features two paired subsets of Huawei and Nikon camera-captured images, which we combine to get 50 paired images to evaluate the models' performance. ExDark and LLVIP unpaired datasets check the models' effectiveness under highly dark conditions. For all the paired datasets, we used the full-reference metrics to evaluate the performance of the models, while for unpaired datasets, we used the no-reference metrics.

Table 3: Performance comparison of mainstream SOTA models and our proposed TriFuse on the LoLI-Street validation set using LoLI-Street-trained weights.

Methods	RetinexF. [6]	RQ-LLIE [30]	CUE [58]	LLFormer [46]	DiffLL [25]	PairLIE [15]	FourLLIE [44]	SCI [34]	TriFuse
Venue	ICCV'23	ICCV'23	ICCV'23	AAAI'23	TOG'23	CVPR'23	ACM MM'23	CVPR'22	(Ours)
Metrics				Light					
PSNR↑	27.92	27.66	28.77	**33.40**	32.59	28.78	28.06	27.84	32.89
SSIM↑	0.8767	0.8811	0.6493	**0.9648**	0.9560	0.9169	0.8363	0.8759	0.9585
MS-SSIM↑	0.9422	0.9035	0.3177	0.9876	0.9889	0.9372	0.8196	0.9413	**0.9899**
LPIPS↓	0.0419	0.1038	0.4194	**0.0039**	0.0139	0.0625	0.1996	0.0429	0.0107
MSE↓	106.79	112.01	89.34	**30.29**	38.55	86.62	101.99	106.87	34.06
MAE↓	206.09	**55.35**	125.41	93.59	115.68	84.86	87.12	206.12	107.63
				Moderate					
PSNR↑	28.44	27.59	30.58	**32.15**	31.87	28.42	27.96	28.38	**32.15**
SSIM↑	0.6197	0.8829	0.9100	0.9386	0.9352	0.9242	0.8693	0.6192	**0.9462**
MS-SSIM↑	0.7819	0.9372	0.9668	**0.9837**	0.9789	0.9374	0.8336	0.7809	0.9819
LPIPS↓	0.1563	0.0668	0.0201	**0.0061**	0.0142	0.0447	0.1831	0.1572	0.0142
MSE↓	94.78	113.49	57.41	**40.14**	40.38	93.75	104.14	94.88	43.16
MAE↓	169.68	**44.33**	148.72	80.25	138.32	162.81	113.81	169.69	107.29
				Dense					
PSNR↑	27.87	29.03	30.58	31.62	31.04	27.66	28.07	27.79	**31.67**
SSIM↑	0.3398	0.9167	0.9100	**0.9274**	0.9165	0.8702	0.8828	0.3394	0.9214
MS-SSIM↑	0.5621	0.9616	0.9668	**0.9738**	0.9734	0.9413	0.8780	0.5614	0.9734
LPIPS↓	0.3037	0.0326	0.0201	**0.0131**	0.0273	0.0357	0.1188	0.3048	0.0201
MSE↓	108.02	82.51	57.41	45.39	51.66	111.57	101.52	108.11	**45.01**
MAE↓	142.15	107.57	148.72	108.04	123.79	214.17	186.56	142.17	**78.88**

Implementation. We implemented the models using PyTorch on a server with four NVIDIA RTX 2080 GPUs (24GB each). All SOTA models were trained with default settings for fair comparison. Our TriFuse model was trained with a batch size of 12 and a patch size of 256×256. The initial learning rate of 1×10^{-4} decayed by 0.8 every 5×10^3 iterations. For efficient restoration, the time step T was set to 200, and the implicit sampling step S was set to 5 for both the training and inference phases.

Evaluation Strategy. We calculated full-reference metrics (PSNR, SSIM, MS-SSIM, MSE, and MAE) and no-reference metrics (BRISQUE, and NIQE) to evaluate existing models and our TriFuse model. Due to the lack of clean ground truth images in real-world testing, no-reference metrics were also employed. We assessed SOTA LLIE models with pre-trained weights on our LoLI-Street dataset to evaluate their quality. Additionally, we trained these models on our dataset and tested them to determine its suitability for generalization. Finally, we quantitatively and qualitatively compared our proposed model with recent SOTA models across various benchmark datasets.

Table 4: Quantitative comparison of mainstream SOTA models and our TriFuse on the LoLI-Street real low-light testset using LoLI-Street-trained weights.

Models	RetinexF. [6]	RQ-LLIE [30]	CUE [58]	LLFormer [46]	DiffLL [25]	PairLIE [15]	FourLLIE [44]	SCI [34]	TriFuse
Venue	ICCV'23	ICCV'23	ICCV'23	AAAI'23	TOG'23	CVPR'23	ACM MM'23	CVPR'22	(Ours)
Metrics	Pre-trained weights								
BRISQUE↓	54.12	29.76	<u>13.65</u>	**12.69**	18.54	39.65	15.05	38.05	-
NIQE↓	**11.79**	12.57	14.36	16.40	<u>12.16</u>	12.28	12.57	<u>12.16</u>	
	Trained weights								
BRISQUE↓	41.69	29.76	25.97	30.44	30.11	35.25	<u>14.50</u>	30.96	**10.32**
NIQE↓	11.83	12.57	12.44	11.84	12.30	<u>11.78</u>	11.89	12.32	**10.61**

5 Comparative Analysis

We compare our TriFuse with multiple SOTA LLIE methods, including RetinexFormer [6], RQ-LLIE [30], CUE [58], LLFormer [46], DiffLL [25], PairLIE [15], FourLLIE [44], and SCI [34], covering transformer and diffusion-based models.

Quantitative Analysis. We present a quantitative analysis of SOTA models on the LoLI-Street and existing datasets. Table 2 shows the performance of these models against the validation set using pre-trained weights with full-reference metrics under various lighting conditions. LLFormer performs robustly across all subsets, achieving the highest PSNR of 28.67 for the dense variety of our validation set. Table 3 evaluates SOTA models on the LoLI-Street validation set using LoLI-Street-trained weights, showing significant performance improvements and model generalization. Our proposed TriFuse achieves the highest scores in various metrics, demonstrating its robustness and effectiveness in LLIE tasks.

The performance of the SOTA models is presented in Table 4 on the real low-light test set of LoLI-Street, using both pre-trained and trained weights for each model. The evaluation metrics include BRISQUE and NIQE. Our proposed model, TriFuse, stands out with the lowest BRISQUE and NIQE scores, indicating superior visual quality and naturalness of the enhanced images compared to the existing models. Table 5 provides a performance comparison of the SOTA models and our proposed TriFuse on existing datasets (LOLv1, LOLv2 (real), LOLv2 (synthetic), LSRW, SICE, ExDark, and LLVIP). As shown, we observed that our model consistently achieved either the best or second-best performance across multiple datasets, as indicated by both full-reference and no-reference metrics. This further validates the effectiveness of our model and emphasizes its ability to generalize well from the training dataset. Table 6 summarizes the computational complexity, demonstrating our model's balance between efficiency and performance with competitive FLOPS and inference time metrics. Overall, the quantitative analysis establishes that our proposed TriFuse model consistently outperforms existing SOTA models across various metrics and datasets, proving its effectiveness and robustness for LLIE tasks. Moreover, Table 7 presents object detection results on the synthetic low-light validation dataset. Our model achieves the highest mAP(0.5) and mAP(0.5-0.9) values, where 0.5 indicates the Intersection over Union (IoU) threshold and 0.5-0.9 represents the average mAP over multiple IoU thresholds, significantly enhancing object detection performance.

Table 5: Performance comparison between SOTA models and our proposed TriFuse based on the mainstream LLIE datasets.

Dataset	Models	RetinexF. [6]	RQ-LLIE [30]	CUE [58]	LLFormer [46]	DiffLL [25]	PairLIE [15]	FourLLIE [44]	SCI [34]	TriFuse
	Venue	ICCV'23	ICCV'23	ICCV'23	AAAI'23	TOG'23	CVPR'23	ACM MM'23	CVPR'22	(Ours)
	Metrics	Full-reference metrics								
LOLv1	PSNR↑	27.89	28.00	27.97	27.77	27.88	27.82	27.93	27.95	**28.01**
	SSIM↑	0.6299	0.8181	0.8724	0.7778	0.8207	0.7111	0.7074	0.2333	**0.8756**
	MS-SSIM↑	0.7542	**0.8642**	0.8562	0.8389	0.8555	0.8313	0.7932	0.4956	0.8578
	LPIPS↓	0.2072	**0.1157**	0.1418	0.1502	0.1473	0.1448	0.1595	0.4177	0.1410
	MSE↓	106.44	105.14	104.59	109.08	103.02	108.18	105.25	104.42	**102.83**
	MAE↓	174.08	172.88	176.54	179.50	169.74	179.70	182.58	150.58	**145.05**
LOLv2 (R)	PSNR↑	27.82	27.82	**28.13**	27.72	27.88	27.77	27.62	27.73	27.88
	SSIM↑	0.5650	0.7799	0.8437	0.7670	0.7875	0.7246	0.7473	0.2543	**0.8966**
	MS-SSIM↑	0.3261	0.8753	**0.8951**	0.8651	0.8695	0.8612	0.8167	0.5542	0.8823
	LPIPS↓	0.5583	0.0988	0.1057	0.1163	0.1219	0.1021	0.1580	0.3518	**0.0888**
	MSE↓	107.39	109.26	**102.06**	110.77	108.32	109.78	112.71	109.94	107.32
	MAE↓	170.48	180.82	174.28	185.34	175.12	199.54	205.84	176.08	**170.28**
LOLv2 (S)	PSNR↑	28.02	28.07	28.01	28.03	28.03	28.14	**29.61**	28.03	28.70
	SSIM↑	0.6511	0.6392	0.6516	0.6759	0.5967	0.8223	**0.9623**	0.4857	0.8593
	MS-SSIM↑	0.7756	0.7741	0.7905	0.7799	0.7644	0.8623	**0.9656**	0.7059	0.8236
	LPIPS↓	0.2449	0.2470	0.2397	0.2374	0.2567	0.1697	**0.0403**	0.2969	0.0727
	MSE↓	103.06	102.03	103.26	103.01	102.85	101.06	**76.78**	103.17	98.48
	MAE↓	158.50	164.45	159.11	156.94	159.62	167.85	**121.24**	176.12	175.73
LSRW	PSNR↑	27.96	28.00	27.93	27.95	27.98	27.94	27.98	27.81	**28.03**
	SSIM↑	0.6160	0.6590	0.6751	0.6543	0.6099	0.6684	0.7298	0.2938	**0.7459**
	MS-SSIM↑	0.6832	0.6900	0.6959	0.6958	0.6846	**0.7045**	0.6970	0.5399	0.6988
	LPIPS↓	0.1873	0.1648	0.1592	0.1685	0.1805	0.1358	0.1748	0.3526	**0.1350**
	MSE↓	104.56	103.52	102.69	104.77	104.10	105.12	103.77	107.88	**100.14**
	MAE↓	182.63	176.41	177.46	182.88	172.70	185.33	174.83	172.28	**152.92**
SICE	PSNR↑	28.02	27.96	27.97	28.02	27.97	**28.34**	28.14	27.81	27.92
	SSIM↑	0.8078	0.7681	0.7605	0.8092	0.7530	**0.8843**	0.8796	0.6582	0.8001
	MS-SSIM↑	0.8396	0.8275	0.8279	0.8397	0.8333	0.8462	**0.8739**	0.8112	0.8240
	LPIPS↓	0.1537	0.1654	0.1679	0.1519	0.1606	0.1252	**0.1082**	0.1774	0.1587
	MSE↓	102.89	104.24	104.15	102.85	104.02	**96.10**	101.22	107.85	102.18
	MAE↓	157.14	163.18	155.44	156.13	161.02	**116.67**	122.22	183.48	118.72
		No-reference metrics								
ExDark	BRISQUE↓	22.90	33.56	17.81	**16.92**	22.19	32.29	18.08	38.95	17.29
	NIQE↓	13.93	15.83	**10.40**	16.250	13.56	14.93	13.844	14.392	13.47
LLVIP	BRISQUE↓	25.09	26.23	17.81	19.86	18.32	34.99	21.55	23.65	**10.32**
	NIQE↓	10.66	10.79	17.22	10.62	11.37	10.96	**10.64**	11.78	11.07

Table 6: Computational complexity of each model. **AIT:** Average inference time (↓).

Models	RetinexF. [6]	RQ-LLIE [30]	CUE [58]	LLF. [46]	DiffLL [25]	PairLIE [15]	FourLLIE [44]	SCI [34]	TriFuse
Venue	ICCV'23	ICCV'23	ICCV'23	AAAI'23	TOG'23	CVPR'23	ACM MM'23	CVPR'22	(Ours)
Params	1.61M	11.38M	0.25M	24.55M	20.09M	0.34M	0.12M	**384**	26.48M
FLOPS	15.57G	162.07G	157.32G	39.61G	89.60G	89.38G	1.95G	**0.14G**	147.03G
AIT	0.009s	0.074s	0.104s	0.267s	1.850s	1.4304s	0.120s	**0.001s**	0.173s

Qualitative Analysis. In addition to the quantitative analysis, we conducted a qualitative evaluation of the enhanced images produced by different models on various datasets. Figure 4 showcases

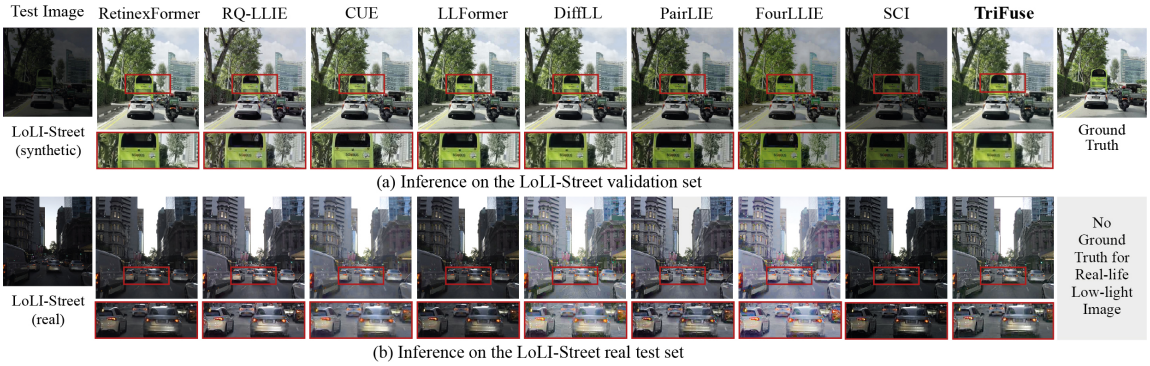


Fig. 4: Enhanced images by different models picking a random image from the (a) synthetic validation set and (b) real low-light test set of our LoLI-Street dataset.

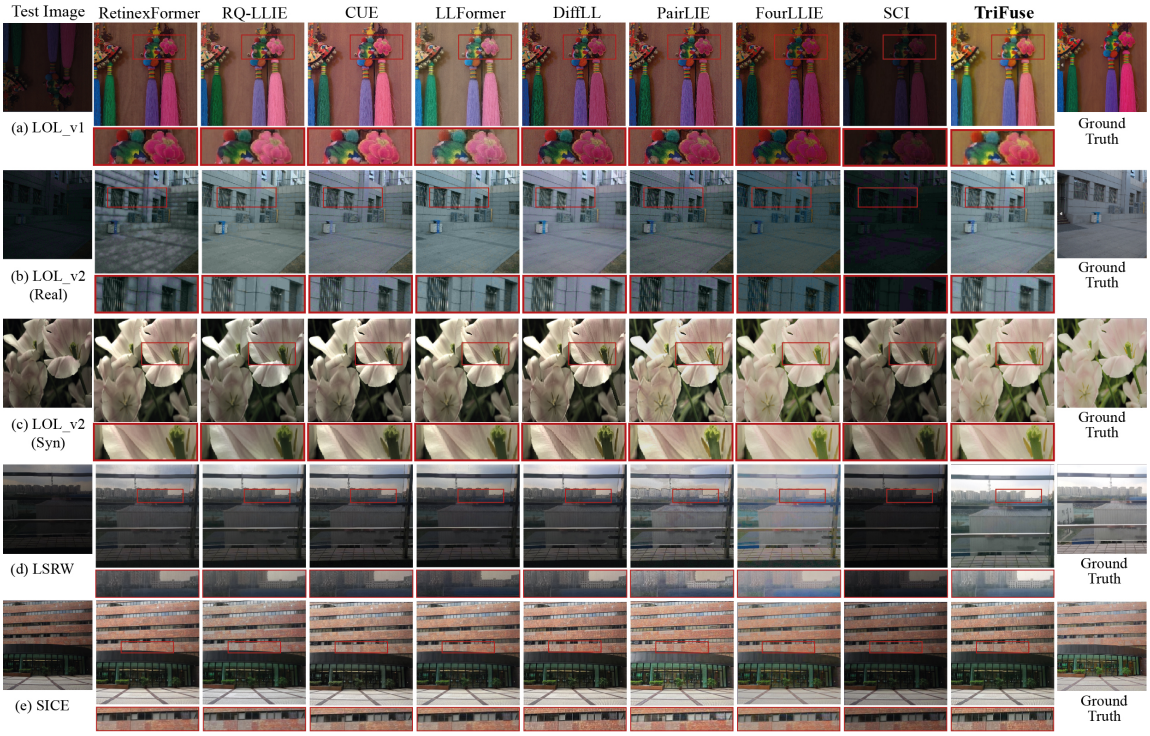


Fig. 5: Enhanced images by SOTA models and our proposed TriFuse picking a random image from the validation sets of mainstream LLIE datasets.

enhanced images from the LoLI-Street dataset's synthetic validation set and real low-light test set, demonstrating that our model consistently provides clearer and more detailed visual enhancements, especially in shadowed and low-light areas. Figure 5 presents enhanced images from the LOLv1 and LOLv2 (both real and synthetic), LSRW, and SICE validation sets, where our model excels in color fidelity and enhancing image details, as evident in the close-up views, revealing well-maintained

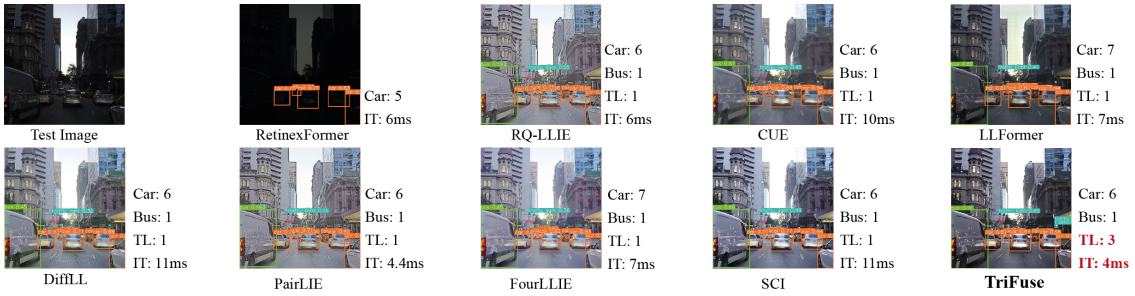


Fig. 6: The outcome of YOLOV10 inference on a sample RLLT image of LoLI-Street dataset after enhancement with each model. **TL:** Traffic light, **IT:** Inference time (↓).

texture details and reduced artifacts. Overall, the comparison highlights TriFuse's robustness and superior performance in enhancing low-light images across multiple datasets. Also, Fig. 6 illustrates the results of YOLOv10 inference on a randomly selected image from the LoLI-Street test set after enhancement by different models. Our model not only improves visual quality but also enhances object detection accuracy, detecting additional objects such as traffic lights and cars with faster inference times compared to other approaches. This qualitative analysis demonstrates our model's effectiveness in enhancing low-light images, significantly improving visual quality and object detection performance in real-world conditions.

Table 7: Performance of object detection using YOLOV10 on our labeled LoLI-Street after LLIE using different models. **MC:** Motorcycle, **TL:** Traffic light, **SS:** Stop sign.

Models	RetinexF. [6]	RQ-LLIE [30]	CUE [58]	LLF. [46]	DiffLL [25]	PairLIE [15]	FourLLIE [44]	SCI [34]	TriFuse
Venue	ICCV'23	ICCV'23	ICCV'23	AAAI'23	TOG'23	CVPR'23	ACM MM'23	CVPR'22	(Ours)
Metrics	Average of mAP for various IoU thresholds for object detection after enhancing images using each model								
mAP(0.5) ↑	0.548	0.602	0.586	0.623	<u>0.650</u>	0.548	0.483	<u>0.650</u>	**0.753**
mAP(0.5-0.9) ↑	0.476	0.521	0.510	0.562	0.568	0.476	0.388	<u>0.653</u>	**0.692**
mAP for varying IoU(0.5-0.9) values detecting some important classes after enhancing images using each model									
Person	0.762	0.723	0.736	<u>0.779</u>	0.749	0.737	0.586	0.760	**0.791**
Bicycle	0.600	0.508	0.584	0.601	0.570	0.541	0.405	**0.660**	<u>0.650</u>
Car	0.876	0.854	0.858	**0.906**	0.870	0.858	0.782	0.884	<u>0.891</u>
MC	0.641	0.618	0.646	0.648	0.649	0.642	0.540	<u>0.743</u>	**0.750**
Bus	0.793	0.722	0.716	0.797	<u>0.825</u>	0.714	0.616	0.820	**0.851**
TL	0.666	0.580	0.642	<u>0.695</u>	0.621	0.596	0.356	0.652	**0.852**
SS	0.447	0.343	0.458	0.518	0.635	0.496	0.206	**0.821**	<u>0.785</u>

Ablation Study. We perform a set of experiments as an ablation study with various combinations of components, such as ESM and CNM, wavelet scale parameters of DWT, and sampling steps of DDPM as presented in Table 8.

For the wavelet transformation scale, we compared the default setting $k(1)$ with $k(2)$ and $k(3)$. The results demonstrate that the ESM+ CNM+ k(1)+ S(5) configuration achieves supe-

Table 8: Ablation studies performed on TriFuse model with varying wavelet parameter, component, and diffusion sampling step. The underlined TriFuse represents the default setting (ESM+ CNM+ $k(1)$+ $S(5)$) of our model.

Dataset	Metrics	Wavlet Scale		Components			Sampling Steps	
		$k(2)$	$k(3)$	w/o-ESM	w/o-CNM	TriFuse	$S(10)$	$S(15)$
RLLT	BRISQUE↓	10.87	11.69	11.03	11.75	**10.32**	10.65	11.14
	NIQE↓	11.25	11.85	11.58	12.01	**10.61**	10.95	11.52
Synthetic Validation	PSNR↑	31.02	31.63	31.74	30.87	**32.24**	32.88	33.37
	SSIM↑	0.9369	0.9312	0.9364	0.9227	**0.9420**	0.9411	0.9470

rior BRISQUE and NIQE scores of 10.32 and 10.61, respectively, on the RLLT dataset, indicating enhanced visual quality compared to other settings. Evaluating the importance of ESM and CNM, comparisons with configurations excluding these components (w/o-ESM and w/o-CNM) highlight the superior performance of the default TriFuse setup. Analysis of different sampling steps (S(5), S(10), S(15)) reveals that increasing to S(15) enhances performance, achieving the highest PSNR of 33.37 and SSIM of 0.9470 on the validation set. Nevertheless, S(5) maintains competitive performance with superior computational efficiency and clarity in real low-light conditions, achieving the lowest BRISQUE (10.32) and NIQE (10.61) scores on the RLLT among all.

6 Conclusion

Identifying the growing need for LLIE solutions, we introduced LoLI-Street, a novel benchmark dataset featuring street scenes under diverse lighting conditions designed to enhance images and improve object detection in low-light environments, which is crucial for autonomous systems. Our proposed LLIE model, TriFuse, incorporates a unique approach utilizing wavelet-based CNM to generate accurate input noise in the diffusion denoising process. This results in effective denoising for real-world LLIE in lower diffusion sampling steps. Comprehensive evaluations demonstrate TriFuse's superiority over existing state-of-the-art models across multiple benchmarks, achieving top performance in visual quality and object detection under low-light conditions based on various metrics. This study emphasizes the crucial role of LLIE in applications such as autonomous vehicles, surveillance, and navigation systems. Future research directions include optimizing TriFuse for real-time applications and adapting it to diverse low-light scenarios, further advancing effective image enhancement solutions.

References

1. Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al.: Xcit: Cross-covariance image transformers. Advances in neural information processing systems **34**, 20014–20027 (2021) 3
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021) 3
3. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input / output image pairs. In: The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition (2011) 3

4. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing **27**(4), 2049–2062 (2018) 3
5. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing **27**(4), 2049–2062 (2018) 4
6. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12504–12513 (2023) 2, 3, 8, 9, 10, 11, 13
7. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022) 3
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 3
9. Cha, J., Haider, A., Yang, S., Jin, H., Yang, S., Uddin, A.S., Kim, J., Kim, S.Y., Bae, S.H.: Descanning: From scanned to the original images with a color correction diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 954–963 (2024) 3
10. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3291–3300 (2018) 2, 3
11. Chen, L., Dong, X., Xie, Y., Wang, S.: Waterpairs: a paired dataset for underwater image enhancement and underwater object detection. Intelligent Marine Technology and Systems **2**(1), 6 (2024) 2
12. Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L.: Dynamic detr: End-to-end object detection with dynamic attention. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2988–2997 (2021) 3
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 3
14. Duarte, A., Codevilla, F., Gaya, J.D.O., Botelho, S.S.: A dataset to evaluate underwater image restoration methods. In: OCEANS 2016-Shanghai. pp. 1–6. IEEE (2016) 2
15. Fu, Z., Yang, Y., Tu, X., Huang, Y., Ding, X., Ma, K.K.: 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) 8, 9, 10, 11, 13
16. Güngör, A., Dar, S.U., Öztürk, Ş., Korkmaz, Y., Bedel, H.A., Elmas, G., Ozbey, M., Çukur, T.: Adaptive diffusion priors for accelerated mri reconstruction. Medical Image Analysis **88**, 102872 (2023) 3
17. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. IEEE Transactions on image processing **26**(2), 982–993 (2016) 3
18. Haar, A.: Zur theorie der orthogonalen funktionensysteme. Mathematische Annalen **71**(1), 38–53 (1911) 5
19. Hai, J., Xuan, Z., Yang, R., Hao, Y., Zou, F., Lin, F., Han, S.: R2rnet: Low-light image enhancement via real-low to real-normal network. Journal of Visual Communication and Image Representation **90**, 103712 (2023) 3, 4
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 5
21. Huang, J., Liu, Y., Chen, S.: Bootstrap diffusion model curve estimation for high resolution low-light image enhancement. In: Pacific Rim International Conference on Artificial Intelligence. pp. 67–80. Springer (2023) 4
22. Hudson, D.A., Zitnick, L.: Generative adversarial transformers. In: International conference on machine learning. pp. 4487–4499. PMLR (2021) 3
23. Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., Van Gool, L.: Dslr-quality photos on mobile devices with deep convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 3277–3285 (2017) 3

24. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3496–3504 (2021) 3, 4
25. Jiang, H., Luo, A., Fan, H., Han, S., Liu, S.: Low-light image enhancement with wavelet-based diffusion models. ACM Transactions on Graphics (TOG) **42**(6), 1–14 (2023) 2, 4, 8, 9, 10, 11, 13
26. Jiang, Y., Chang, S., Wang, Z.: Transgan: Two pure transformers can make one strong gan, and that can scale up. Advances in Neural Information Processing Systems **34**, 14745–14758 (2021) 3
27. Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.M., Gu, J., Loy, C.C.: Low-light image and video enhancement using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(12), 9396–9416 (2021) 2
28. Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An underwater image enhancement benchmark dataset and beyond. IEEE transactions on image processing **29**, 4376–4389 (2019) 2
29. Li, G., Yang, Y., Qu, X., Cao, D., Li, K.: A deep learning based image enhancement approach for autonomous driving at night. Knowledge-Based Systems **213**, 106617 (2021) 2
30. Liu, Y., Huang, T., Dong, W., Wu, F., Li, X., Shi, G.: 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) 8, 9, 10, 11, 13
31. Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset. Computer Vision and Image Understanding **178**, 30–42 (2019) 2, 3, 4
32. Lv, X., Zhang, S., Wang, C., Zhang, W., Yao, H., Huang, Q.: Unsupervised low-light video enhancement with spatial-temporal co-attention transformer. IEEE Transactions on Image Processing (2023) 2
33. Lv, X., Dong, X., Jin, Z., Zhang, H., Song, S., Li, X.: L 2 dm: A diffusion model for low-light image enhancement. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 130–145. Springer (2023) 3, 4
34. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: 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) 8, 9, 10, 11, 13
35. Mandal, G., Bhattacharya, D., De, P.: Real-time fast low-light vision enhancement for driver during driving at night. Journal of Ambient Intelligence and Humanized Computing **13**(2), 789–798 (2022) 2
36. Mittal, P., Singh, R., Sharma, A.: Deep learning-based object detection in low-altitude uav datasets: A survey. Image and Vision computing **104**, 104046 (2020) 2
37. Moran, S., Marza, P., McDonagh, S., Parisot, S., Slabaugh, G.: Deeplpf: Deep local parametric filters for image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12826–12835 (2020) 2
38. Nguyen, C.M., Chan, E.R., Bergman, A.W., Wetzstein, G.: Diffusion in the dark: A diffusion model for low-light text recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4146–4157 (2024) 4
39. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging (2023) 3
40. Panetta, K., KM, S.K., Rao, S.P., Agaian, S.S.: Deep perceptual image enhancement network for exposure restoration. IEEE Transactions on Cybernetics (2022) 2
41. Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. IEEE Transactions on Image Processing **32**, 1927–1941 (2023) 5
42. Szeliski, R.: Computer vision: algorithms and applications. Springer Nature (2022) 2
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 3
44. Wang, C., Wu, H., Jin, Z.: Fourllie: Boosting low-light image enhancement by fourier frequency information. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7459–7469 (2023) 8, 9, 10, 11, 13

45. Wang, N., Wang, Y., Er, M.J.: Review on deep learning techniques for marine object recognition: Architectures and algorithms. Control Engineering Practice **118**, 104458 (2022) 2
46. Wang, T., Zhang, K., Shen, T., Luo, W., Stenger, B., Lu, T.: Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2654–2662 (2023) 2, 3, 8, 9, 10, 11, 13
47. Wang, Y., Yu, Y., Yang, W., Guo, L., Chau, L.P., Kot, A.C., Wen, B.: Exposurediffusion: Learning to expose for low-light image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12438–12448 (2023) 2, 4
48. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (2018) 2
49. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (2018) 3, 4
50. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J.E., Keutzer, K., Vajda, P.: Visual transformers: Where do transformers really belong in vision models? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 599–609 (2021) 3
51. Xia, Z., Gharbi, M., Perazzi, F., Sunkavalli, K., Chakrabarti, A.: Deep denoising of flash and no-flash pairs for photography in low-light environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2063–2072 (2021) 2
52. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3063–3072 (2020) 3
53. Yang, W., Wang, W., Huang, H., Wang, S., Liu, J.: Sparse gradient regularized deep retinex network for robust low-light image enhancement. In: IEEE Transactions on Image Processing. pp. 072–2086. IEEE (2021) 4
54. Yi, X., Xu, H., Zhang, H., Tang, L., Ma, J.: 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) 2, 4
55. Yin, Y., Xu, D., Tan, C., Liu, P., Zhao, Y., Wei, Y.: Cle diffusion: Controllable light enhancement diffusion model. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8145–8156 (2023) 2, 4
56. Zeng, H., Cao, J., Zhang, K., Chen, Y., Luong, H., Philips, W.: Unmixing diffusion for self-supervised hyperspectral image denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27820–27830 (2024) 3
57. Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., Guo, B.: Styleswin: Transformer-based gan for high-resolution image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11304–11314 (2022) 3
58. Zheng, N., Zhou, M., Dong, Y., Rui, X., Huang, J., Li, C., Zhao, F.: Empowering low-light image enhancer through customized learnable priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12559–12569 (2023) 8, 9, 10, 11, 13
59. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021) 3

LoLI-Street: Benchmarking Low-Light Image Enhancement and Beyond

Md Tanvir Islam[1], Inzamamul Alam[1], Simon S. Woo[1],
Saeed Anwar[2], IK Hyun Lee[3], and Khan Muhammad[3,*]

[1] Department of Software, Sungkyunkwan University, Suwon 16419, South Korea
[2] The Australian National University, Canberra 0200, Australia
[3] Department of Mechatronics Engineering, Tech University of Korea, Siheung-Si, 15073, South Korea
[4] Department of Human-AI Interaction, Sungkyunkwan University, Seoul 03063, South Korea
*Corresponding author: khanmuhammad@g.skku.edu
Code and dataset: https://github.com/tanvirnwu/TriFuse

Loss Calculation (Linked with TriFuse part of Section 3.2). Our primary objective function L_{diff} needs to be optimized by our TriFuse model, and the training process includes additional loss functions to enhance detail preservation and overall content accuracy of the restored images. These loss functions include a noise loss L_{noise}, a frequency loss $L_{\text{frequency}}$, and a photo loss L_{photo}.

The noise loss L_{noise} is formulated to minimize the difference between the predicted noise and the actual noise:

$$L_{\text{noise}} = \mathcal{L}_{\text{MSE}}(\epsilon_{\text{pred}}, \epsilon), \tag{1}$$

where ϵ_{pred} is the predicted noise and ϵ is the actual noise.

The frequency loss $L_{\text{frequency}}$ is designed to preserve high-frequency details and is a combination of MSE loss and Total Variation (TV) loss [2]:

$$\begin{aligned} L_{\text{frequency}} = 0.1\,(\mathcal{L}_{\text{MSE}}(I_{\text{high0}}, I_{\text{gt\_high0}}) + \mathcal{L}_{\text{MSE}}(I_{\text{high1}}, I_{\text{gt\_high1}}) \\ + \mathcal{L}_{\text{MSE}}(I_{\text{pred\_LL}}, I_{\text{gt\_LL}})) + 0.01\,(\text{TV}(I_{\text{high0}}) + \text{TV}(I_{\text{high1}}) + \text{TV}(I_{\text{pred\_LL}}))\,, \end{aligned} \tag{2}$$

where I_{high0}, I_{high1}, and $I_{\text{pred\_LL}}$ are the predicted high-frequency components and $I_{\text{gt\_high0}}$, $I_{\text{gt\_high1}}$, and $I_{\text{gt\_LL}}$ are the ground truth high-frequency components. The TV loss helps in reducing noise while preserving edges.

The photo loss L_{photo} combines L1 loss and SSIM loss [5] to maintain the content fidelity of the restored image as follows:

$$L_{\text{photo}} = |I_{\text{pred}} - I_{\text{gt}}|_1 + (1 - \text{SSIM}(I_{\text{pred}}, I_{\text{gt}})), \tag{3}$$

where I_{pred} is the predicted image and I_{gt} is the ground truth image. The L1 loss ensures pixel-wise accuracy, while the SSIM loss promotes structural similarity.

The total loss L_{diff} combines the diffusion objective function, the noise loss, the frequency loss, and the photo loss as follows:

$$L_{\text{diff}} = L_{\text{noise}} + L_{\text{frequency}} + L_{\text{photo}}. \tag{4}$$

This comprehensive loss function ensures that our TriFuse network not only focuses on the diffusion process but also effectively preserves fine details and maintains high content fidelity throughout the image enhancement process.

Algorithm 1 Training steps of our proposed TriFuse model.

Require: Average coefficients of low/normal-light image pairs $\tilde{A}^K_{low}$ and A^K_{high}, denoted as $\tilde{x}$ and x_0, respectively, the time step T, the number of implicit sampling steps S, and the model parameters θ.

Procedure: Train TriFuse
while Not converged **do**
  Forward diffusion process
  $t \sim \text{Uniform}\{1, \cdots, T\}$
  $\epsilon_t \sim \mathcal{N}(0, I)$
  Compute the noisy image:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t$$

  Perform a single gradient descent step to minimize the loss:

$$\mathcal{L}_{\text{diffusion}} = \|\epsilon_t - \epsilon_\theta(x_t, \tilde{x}, t)\|^2$$

  Here, ϵ_θ is the noise prediction model incorporating the conditional noise module (CNM).
  Denoising process
  $\tilde{x}_T \sim \mathcal{N}(0, I)$
  for $i = S : 1$ **do**
    $t = (i-1) \cdot \frac{T}{S} + 1$
    $t_{next} = (i-2) \cdot \frac{T}{S} + 1$ if $i > 1$, else 0
    $\text{CNM}(\tilde{x}_t) = \epsilon_\theta(\tilde{x}_{t+1}, \tilde{x}, t)$
    Update $\tilde{x}_t$:

$$\tilde{x}_t \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\tilde{x}_{t+1} - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\tilde{x}_{t+1}, \tilde{x}, t) \right) + \sigma_t \eta$$

  end for
  Obtain the final denoised image $\tilde{x}_0$
  Apply the edge sharpening module (ESM) to $\tilde{x}_0$ to enhance edges, producing $\tilde{x}_0^{\text{sharp}}$:

$$\tilde{x}_0^{\text{sharp}} = \text{ESM}(\tilde{x}_0)$$

  Perform a single gradient descent step to minimize the reconstruction loss:

$$\mathcal{L}_{\text{reconstruction}} = \|\tilde{x}_0^{\text{sharp}} - x_0\|^2$$

end while
End Procedure

Dataset (Linked with Fig.6 and Table 7 of Section 5). We also prepared our dataset for research related to object detection tasks under low-light conditions. We annotated the ground truth images of the synthetic validation set using YOLOv10 [4] and then tested the low-light images of the same subset using YOLOv10 [4]. The detected objects for high-light and low-light images are presented in Table 1. The results indicate that YOLOv10 [4] struggles to detect all objects accurately under low-light conditions, as evidenced by the significantly reduced number of detected objects compared to their corresponding high-light versions. This emphasizes the necessity

Algorithm 2 Inference steps of our proposed TriFuse model.

Require: Input image x_0, trained model parameters θ, time step T, and the number of implicit sampling steps S.

Procedure: Inference with TriFuse
Initialize $\tilde{x}_T \sim \mathcal{N}(0, I)$
for $i = S : 1$ **do**
  $t = (i-1) \cdot \frac{T}{S} + 1$
  $t_{next} = (i-2) \cdot \frac{T}{S} + 1$ if $i > 1$, else 0
  $\text{CNM}(\tilde{x}_t) = \epsilon_\theta(\tilde{x}_{t+1}, \tilde{x}, t)$
  Update $\tilde{x}_t$:

$$\tilde{x}_t \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\tilde{x}_{t+1} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\tilde{x}_{t+1}, \tilde{x}, t) \right) + \sigma_t \eta$$

end for
Obtain the final denoised image $\tilde{x}_0$
Apply the edge sharpening module (ESM) to $\tilde{x}_0$ to enhance edges:

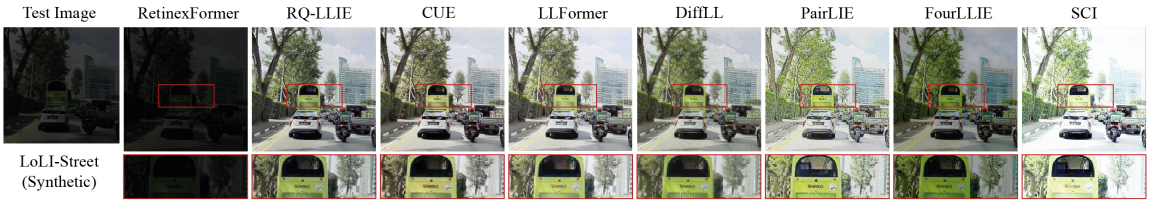
$$\tilde{x}_0^{\text{sharp}} = \text{ESM}(\tilde{x}_0)$$

End Procedure

ENSURE $\tilde{x}_0^{\text{sharp}}$

Table 1: Detected objects across different subsets of our LoLI-Street dataset using YOLOv10 [4], illustrating the challenges in accurately detecting objects under low-light conditions. The number of detected objects decreases in low-light subsets, whereas in high-light subsets, more objects are detected.

Class ID	Class Name	High		Low		Real Low-Light
		Train	Validation	Train	Validation	Testset
0	Person	53412	3957	44409	3082	1062
1	Bicycle	1671	108	1271	80	61
2	Car	171837	15744	133697	12067	4918
3	Motorcycle	2103	2400	1279	1578	45
4	Airplane	138	9	91	4	2
5	Bus	6255	852	4231	459	100
7	Truck	17976	2043	11930	1150	343
9	Traffic Light	41391	1176	32226	666	1890
10	Fire Hydrant	411	24	244	13	6
11	Stop Sign	549	66	394	24	3
12	Parking Meter	21	10	12	6	2
13	Bench	321	12	250	5	8
16	Dog	42	15	28	10	4
25	Umbrella	309	15	195	9	7
26	Handbag	399	30	286	9	14
33	Kite	156	12	103	4	2
36	Skateboard	135	5	95	2	2
58	Potted Plant	642	48	443	23	8
74	Clock	567	63	324	39	15



(a) Inference on the LoLI-Street validation set using the pre-trained weights of each SOTA model

(b) Inference on the LoLI-Street real low-light testset using the pre-trained weights of each SOTA model

Fig. 1: Enhanced images by using the pre-trained weights of SOTA low-light image enhancement models on a random image from the (a) synthetic validation set and (b) real low-light testset of our LoLI-Street dataset.

of low-light image enhancement, particularly for street-scene types where autonomous systems rely heavily on computer vision tasks such as object detection.

Visualizations (Linked with Qualitative Analysis of Section 5). The sample enhanced images from our LoLI-Street synthetic validation set and real low-light testset using pre-trained weights of various SOTA low-light image enhancement models are shown in fig:fig1. The figure clearly demonstrates that the SOTA low-light image enhancement models face difficulties in enhancing the images, particularly those from the real low-light testset, highlighting the challenges inherent in our LoLI-Street dataset. Therefore, it signifies the need to develop and train more robust models, particularly for street scene types. On the other hand, training the same models on our dataset improves the performance of the models, as presented in the main paper.

The inference results on some random images from different mainstream datasets (LOLv1 [6], LOLv2 [7], LSRW [3], and SICE [1]) using our proposed TriFuse model are presented in fig:fig2. The results illustrate the effectiveness of our proposed TriFuse model in enhancing visual quality across different types of low-light scenarios, demonstrating improved clarity and detail preservation.

To test the effectiveness of the proposed TriFuse model in enhancing low-light images from non-urban street scenes, we test a few randomly collected images as presented in Fig. 3, which demonstrates the model's capability to effectively enhance visibility and detail in various environments beyond urban streets.

Ablation (Linked with Ablation Study of Section 5). Table 2 evaluates the impact of different types of image degradations on our proposed TriFuse model using BRISQUE and NIQE metrics. The table categorizes performance under three degradation types: blur, noise, and JPEG compression. Each type is measured at varying levels, denoted by σ for blur, γ for noise, and η for JPEG compression. The results show that as the levels of these degradations increase, both BRISQUE and NIQE scores worsen, indicating a decline in the image quality enhanced by our TriFuse model.

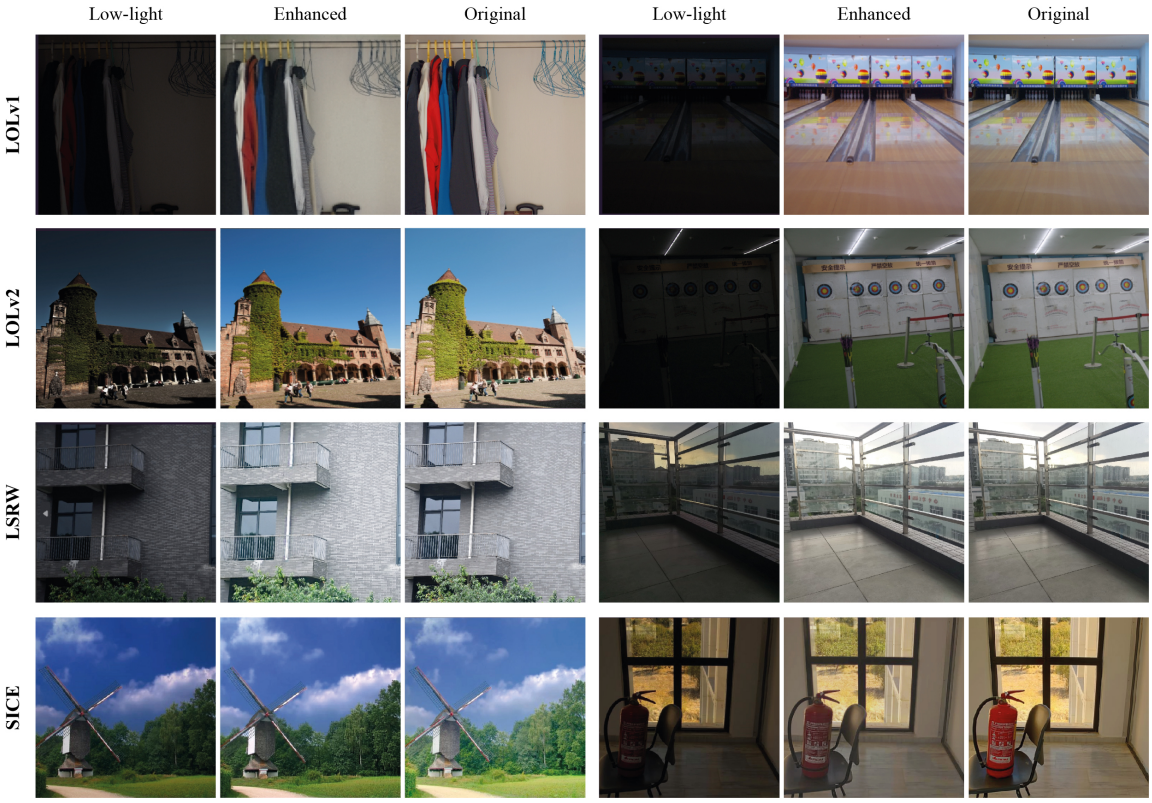


Fig. 2: Inference on some randomly picked images from different mainstream datasets (LOLv1 [6], LOLv2 [7], LSRW [3], and SICE [1]) using our proposed TriFuse model.

Table 2: Evaluating the impact of different types of image degradations on BRISQUE and NIQE metrics. The table presents performance under various levels of blur (σ), noise (γ), and JPEG compression (η).

Metrics	Blur			Noise			JPEG Compression		
	$\sigma(0.3)$	$\sigma(0.5)$	$\sigma(0.7)$	$\gamma(0.1)$	$\gamma(0.2)$	$\gamma(0.3)$	$\eta(20)$	$\eta(30)$	$\eta(50)$
BRISQUE↓	31.56	31.42	34.71	56.18	69.47	74.72	40.08	31.93	26.67
NIQE↓	12.16	11.79	12.11	18.78	30.12	35.44	14.26	14.09	18.78

For instance, BRISQUE values increase significantly from 31.56 to 74.72 as noise γ increases from 0.1 to 0.3, demonstrating the sensitivity of the model performance for adding Gaussian noise to the low-light images. Increasing the blurriness amount in the image from 0.3 to 0.7 resulted in only a slight increase in the BRISQUE value, from 31.56 to 34.71, and in the NIQE value, from 12.16 to 12.11. This contrasts with the significant changes observed when noise was added. Experiments with JPEG compression revealed that increasing the compression level η from 20 to 30 reduced the BRISQUE value from 40.08 to 26.67, while the NIQE value increased from 14.26 to 18.78. These findings

Fig. 3: Sample images from non-urban street scenes enhanced using the proposed TriFuse model. The top row shows the original low-light images, while the bottom row displays the enhanced versions, illustrating the model's effectiveness in diverse environments.

indicate that, beyond the effects of low light on various degradations, TriFuse encounters challenges because it was not primarily trained for these additional degradation types. This highlights the need for future research to develop image enhancement methods capable of handling a wider range of degradations. The analysis emphasizes the necessity for robust enhancement techniques to preserve image quality across different degradation scenarios, particularly under low-light conditions.

Impact of CNM:

Impact of ESM:

References

1. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing **27**(4), 2049–2062 (2018) 4, 5
2. Chan, S.H., Khoshabeh, R., Gibson, K.B., Gill, P.E., Nguyen, T.Q.: An augmented lagrangian method for total variation video restoration. IEEE Transactions on Image Processing **20**(11), 3097–3111 (2011). https://doi.org/10.1109/TIP.2011.2158229 1
3. Hai, J., Xuan, Z., Yang, R., Hao, Y., Zou, F., Lin, F., Han, S.: R2rnet: Low-light image enhancement via real-low to real-normal network. Journal of Visual Communication and Image Representation **90**, 103712 (2023) 4, 5
4. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458 (2024) 2, 3
5. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861 1
6. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (2018) 4, 5

7. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3063–3072 (2020) 4, 5