Supplementary Materials

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

Md Tanvir Islam[1[0009−0007−9405−5684]], Inzamamul Alam[1[0009−0004−6564−4867]], Simon S. Woo[1,*[0000−0002−8983−1542]], Saeed Anwar[2[0000−0002−0692−8411]], IK Hyun Lee[3[0000−0002−0605−7572]], and Khan Muhammad[3,*[0000−0002−5302−1150]]

[1] Department of Software, Sungkyunkwan University, South Korea
[2] The Australian National University, Acton, Canberra, Australia
[3] Department of Mechatronics Engineering, Tech University of Korea, South Korea
[4] Department of Human-AI Interaction, Sungkyunkwan University, South Korea
*Corresponding authors:{swoo, 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 $I_{\text{high0}}, I_{\text{high1}}$, 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:

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

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

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.

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

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

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

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 $\tilde{x}_t$:

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

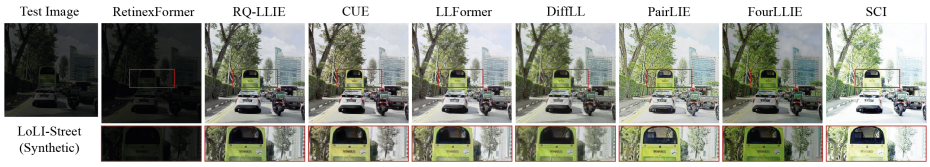
end for
Obtain the final denoised image $\tilde{x}_0$
Apply the edge sharpening module (ESM) to $\tilde{x}_0$ to enhance edges:

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

End Procedure

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

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



(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.

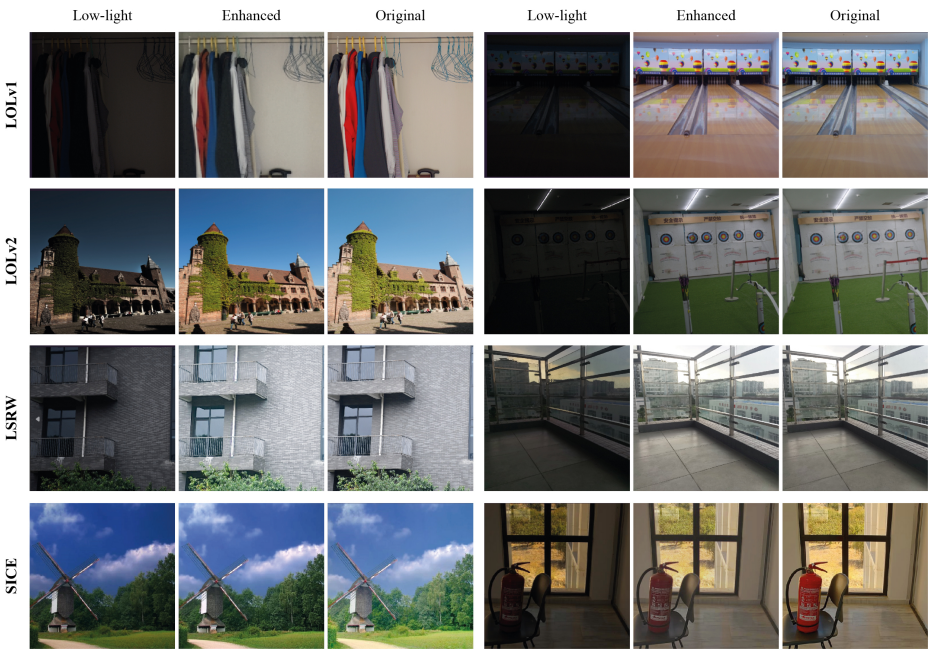


Fig. 2: Inference on some randomly picked images from different mainstream datasets (LOLv1 [6], LOLv2 [7], LSRW [3], and SICE [1]) using our proposed TriFuse model.

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.

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

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

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. 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 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.

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)
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
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. 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)
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
6. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (2018)
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)