Supplementary Material for Paper: LightenDiffusion: Unsupervised Low-Light Image Enhancement with Latent-Retinex Diffusion Models

Hai Jiang[1,5], Ao Luo[2,5], Xiaohong Liu[4],
Songchen Han[1], and Shuaicheng Liu[3,5,†]

[1] Sichuan University, [2] Southwest Jiaotong University,
[3] University of Electronic Science and Technology of China,
[4] Shanghai Jiao Tong University, [5] Megvii Technology
{jianghai@stu.,hansongchen@}scu.edu.cn, aoluo_uestc@gmail.com,
xiaohongliu@sjtu.edu.cn, liushuaicheng@uestc.edu.cn
[†] Corresponding Author

This supplementary material is organized as follows:

- Sec. 1 provides a more comprehensive analysis of our method and the recently published diffusion-based low-light image enhancement method.
- Sec. 2 conducts more experiments including detailed ablation studies of our overall framework, quantitative comparison on the VE-LOL dataset [15], computational efficiency analysis, and the application for low-light instance segmentation on the LIS dataset [2].
- Sec. 3 provides more qualitative results on paired and unpaired benchmarks, instance segmentation results, and face detection results.
- Sec. 4 presents the limitation of our method.

1 More Detailed Analysis

In this section, we provide a more comprehensive analysis of our method and the recently published diffusion-based LLIE method including PyDiff [30], DiffLL [11], GSAD [10], and Diff-Retinex [28].

- PyDiff [30] adopts a pyramid resolution strategy that progressively increases the image resolution in one reverse process to achieve efficient sampling and uses a global corrector to alleviate the global degradation that may occur in the reverse denoising process for performance improvement and computational consumption reduction.
- DiffLL [11] leverages 2D discrete wavelet transformation to convert images into the wavelet domain to accelerate inference and reduce computational resource usage without sacrificing information, followed by performing diffusion operations on the average coefficient for robust and efficient low-light image enhancement.

Table 1: Quantitative comparisons of our method and other supervised diffusion-based LLIE methods on four unpaired datasets including DICM [13], VV [21], LoLi-Huawei [14] and LoLi-iPhone [14]. Two non-reference perceptual metrics NIQE [20] and PI [1] are adopted to measure the visual quality. The best results are highlighted in **bold**. Note that Diff-Retinex [28] is not open-source, so we cannot reproduce its performance on unseen real-world datasets.

Method	DICM [13]		VV [21]		LoLi-Huawei [14]		LoLi-iPhone [14]	
	NIQE #	PI #	NIQE #	PI #	NIQE #	PI #	NIQE #	PI #
PyDiff [30]	4.499	3.792	4.360	3.678	6.217	4.900	5.524	4.810
DiffLL [11]	3.806	**3.011**	3.507	2.562	5.071	4.259	4.797	3.709
GSAD [10]	4.496	3.593	5.252	3.657	6.378	5.248	5.597	4.894
Ours	**3.724**	3.144	**2.941**	**2.558**	**4.755**	**4.019**	**4.119**	**3.702**

- GSAD [10] proposes a global structure-aware regularization scheme for low-light image enhancement, which preserves the intrinsic structural integrity in the reconstructed samples by performing rank-informed regularization and uncertainty-guided noise learning on learnable closed-form samples in terms of minimizing the curvature of the ODE trajectory.
- Diff-Retinex [28] performs decomposition in image space following the typical Retinex-based supervised learning pipeline [7, 23] and employs two separate diffusion models to reconstruct the decomposed components for contrast enhancement and noise removal, respectively.

However, unlike previous Retinex-based methods, we convert images into latent space and employ the proposed content-transfer decomposition network to obtain content-rich reflectance maps and content-free illumination maps specified for unsupervised learning, without any efficiency linkage, while using the reflectance map of the low-light image and the illumination map of the unpaired normal-light image as inputs to the unsupervised setting of the specialized diffusion model for restoration.

Moreover, all the above methods are designed for supervised LLIE, which requires large-scale paired data and evidently cannot be trained using unpaired images. Although such methods enable obtaining excellent distortion metrics on paired datasets, their generalization to unseen real-world scenes is limited. In contrast, all our contributions aim to realize the development of a new diffusion-based unsupervised LLIE approach that learns degraded representations of various scenes by training on extensive unpaired real-world data, ultimately achieving visually favorable LLIE and improving the generalization of diffusion models. As shown in Table 1, our method achieves better perceptual quality on four real-world unpaired datasets including DICM [13], VV [21], LoLi-Huawei [14] and LoLi-iPhone [14], proving that our method can generalize better to unknown scenes compared to supervised diffusion-based methods.

Table 2: Quantitative results of ablation studies, please refer to the text for more details. The results using default settings are underlined.

Method	LOL [23]			DICM [13]	
	PSNR "	SSIM "	LPIPS #	NIQE #	PI #
Baseline #1	18.342	0.783	0.207	4.137	3.459
Baseline #2	19.340	0.797	0.198	3.934	3.379
Baseline #3	19.096	0.780	0.212	4.102	3.421
Default	20.453	0.803	0.192	3.724	3.144

2 More Experiments

2.1 More Detailed Results of the Ablation Study

In this section, we conduct more detailed experiments to evaluate the impact of different component choices of our method. We use the implementation details described in Sec. 4.1 of the main paper for training, and quantitative results on paired dataset LOL [23] and unpaired dataset DICM [13] are illustrated in Table 2. Detailed experiment settings are discussed below.

Baseline #1. Some Retinex-based methods [5,7] directly employ the Gamma correction (GC) to improve the contrast of the estimated illumination map, which is subsequently fused with the reflectance map to obtain the restored result. Following the above approaches, we form Baseline #1 by feeding our constructed pseudo label $F_{low} = R_{low} \quad L_{low}^{\gamma}$ into the decoder D to produce the final result $\hat{I}_{low}$, which can be expressed as:

$$\hat{I}_{low} = D(R_{low} \quad L_{low}^{\gamma}); \tag{1}$$

where we set the illumination correction factor $= 0.2$ following [5].

Baseline #2. We take the pseudo labels as the input of the diffusion model for training with the guidance of the low-light feature to form Baseline #2. Accordingly, we do not need the illumination map of the normal-light image during the training process, which could avoid the interference of unpaired content information on the final result.

Baseline #3. As mentioned in the main paper, one straightforward way to obtain the enhanced feature in the ideal case is to multiply the reflectance

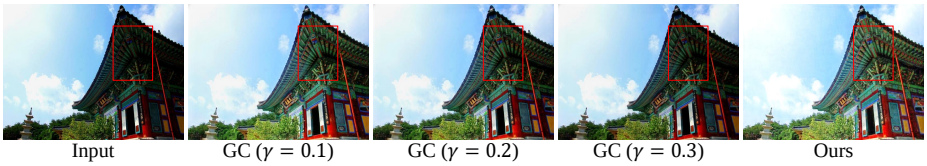

Fig. 1: illustration of the results obtained by Gamma correction (GC) with different illumination correction factor γ and our method.

Fig. 2: Illustration of the restored result obtained through Baseline #3, where the illumination map L_{high} of the normal-light image contains certain content information that leads to the restored result being disrupted.

map of the low-light feature with the illumination map of the normal-light one as $\hat{F}_{low} = R_{low} \quad L_{high}$. Therefore, we form Baseline #3 that feeds the fused result $\hat{F}_{low}$ into the decoder $D(\ )$ to produce the nal result $\hat{I}_{low}$, which can be expressed as:

$$\hat{I}_{low} = D(R_{low} \quad L_{high}); \tag{2}$$

where R_{low} and L_{high} are the decomposition results of unpaired low-light and normal-light images, respectively.

As reported in Table 2, when comparing Baseline #1 and Baseline #3, Gamma correction achieves better SSIM and LPIPS on the LOL dataset while showing inferior on the DICM dataset, due to the fact that the results obtained by gamma correction are usually unrealistic and an appropriate correction factor is di cult to choose for various illumination scenes, as shown in Fig. 1. Therefore, in Baseline #2, using the distorted results obtained by Gamma correction as the input to the di usion model would cause the learned distribution to deviate from the real-world distribution, leading to poor generalization ability.

In contrast, the illumination maps of normal-light images show more natural lighting conditions and therefore have advantages in generalizability. However, the restored results would be disturbed once the illumination map still presents the content information of the normal-light image, resulting in unsatisfactory visual quality, as shown in Fig. 2. Nevertheless, we take advantage of Gamma correction, which does not introduce extra content information, to design the self-constrained consistency loss to eliminate potential artifacts and ensure the restored result has the same inherent content information as the input low-light image. As shown in Table 2, our method presents overall performance improvement by incorporating the strengths of Baseline #1 and Baseline #3.

2.2 More Quantitative Results

To further validate the generalization ability of our method, we compare our method with competitive learning-based methods on the VE-LOL [15] test set and two real-world unpaired datasets LoLi-Huawei [14] and LoLi-iPhone [14]. The two unpaired datasets contain low-light images captured under challenging scenes with signi cant environment noise and the VE-LOL [15] test set contains

Table 3: Quantitative comparisons on the paired dataset VE-LOL [15] as well as two unpaired datasets including LoLi-Huawei [14] and LoLi-iPhone [14]. The best results are highlighted in bold .

Methods	VE-LOL [15]			LoLi-Huawei [14]		LoLi-iPhone [14]	
	PSNR "	SSIM "	LPIPS #	NIQE #	PI #	NIQE #	PI #
RetinexNet [23]	12.353	0.121	1.123	6.297	5.180	6.051	5.208
KinD++ [29]	15.307	0.370	0.692	5.527	5.290	5.390	5.617
URetinexNet [24]	15.493	0.339	0.851	6.245	5.742	5.815	5.648
PyDi [30]	18.999	0.805	0.251	6.217	4.900	5.524	4.810
GSAD [10]	17.423	0.407	0.716	6.378	5.248	5.597	4.894
DRBN [26]	16.907	0.673	0.399	6.384	6.331	5.164	5.334
BL [18]	10.785	0.103	1.022	7.834	6.120	6.314	5.358
Zero-DCE [6]	12.028	0.142	1.035	4.981	4.447	4.495	4.082
EnlightenGAN [12]	13.875	0.232	0.956	5.327	4.861	4.532	4.381
RUAS [16]	9.557	0.128	0.984	6.549	5.688	6.344	5.809
SCI [19]	10.905	0.116	1.069	7.509	5.743	6.228	5.318
GDP [4]	12.437	0.182	1.021	9.159	7.574	7.581	6.401
PairLIE [5]	14.120	0.342	0.906	5.082	4.596	4.631	4.294
NeRCo [25]	16.422	0.405	0.727	4.964	4.332	4.683	4.327
Ours	20.228	0.782	0.374	4.755	4.019	4.119	3.702

100 pairs of real-captured images and 100 pairs of synthetic images. However, the real part of the VE-LOL dataset overlaps with the training set of the LOL [23] dataset and supervised methods typically trained on the LOL dataset. For fair comparisons, We follow [30] that combines the training and test sets of the real part and lters out the images that overlap with the LOL training set based on the IDs. For the ltered images, we sort them by ID and select 100 images with the smallest ID as the test set to make the image count equal to the original one.

As reported in Table 3, supervised methods show poor performance since the VE-LOL dataset has noise and illumination distributions not present in the LOL dataset. In contrast, the semi-supervised method and unsupervised methods present better generalization ability than supervised ones, where our method achieves the best PSNR compared to all methods, proving the superior generalization ability of our method to unseen scenes. Moreover, unsupervised methods also show better performance on the two unpaired benchmarks than supervised ones, in which our method achieves the best perceptual scores, proving that our method has better generalization capability to unseen scenes.

2.3 Computational E ciency

In this section, we report the model parameters, inference time, and memory consumption on RTX 2080Ti when performing inference on the LOL [23] test set, where the image size is 600400 3. As illustrated in Table 4, the model parameters of our method is about 27.837MB. The computational e ciency and memory consumption are 0.314s and 1.772GB, respectively.

Table 4: Parameters (Para.), inference time (Time), and memory consumption (Mem.) on RTX 2080Ti when performing inference on an image with 400 600 3 resolution.

Resolution	Para. (MB)	Time (s)	Mem. (GB)
400 600 3	27.837	0.314	1.772

2.4 Low-Light Instance Segmentation

In this section, we conduct experiments on the LIS dataset [2], in which 2,230 images are randomly split into a train set of 1,561 images and a test set of 669 images following [2], to investigate the impact of LLIE methods as a pre-processing step in improving the low-light instance segmentation task. We employ our method and competitive learning-based LLIE methods to restore images of the training set to retrain the Mask R-CNN [9] as a baseline and subsequently evaluate it on the restored images of the test set. The performance is measured by using COCO-style AP^{seg} (averaged over thresholds from 0.5 to 0.95 with an interval of 0.05), AP^{seg}_{50} and AP^{seg}_{75} (i.e., AP at an IoU threshold of 0.5) [9]. We also provide the detection results, which are represented as AP^{box}, AP^{box}_{50}, and AP^{box}_{75}.

As reported in Table 5, without any pre-processing steps, the baseline normal instance segmentation model only has 34.2% AP^{seg} and 41.3% AP^{box}, which shows limited accuracy in low-light conditions. After improving the visual quality of low-light images using the LLIE methods, we intuitively expected an improvement in performance, where our method achieved the best performance on all metrics, with improvements of 5.4%(=39.6-34.2) and 5.6%(=46.9-41.3) for AP^{seg} and AP^{box}, respectively. In particular, our method achieves 40.8% on AP^{seg}_{75} which outperforms the second-best EnlightenGAN by 1.6 points, as well as

Table 5: Quantitative comparisons of instance segmentation results before and after enhanced by di erent methods on the LIS dataset [2]. The best results are highlighted in bold .

Method	AP^{seg} "	AP^{seg}_{50} "	AP^{seg}_{75} "	AP^{box} "	AP^{box}_{50} "	AP^{box}_{75} "
Mask R-CNN [9]	34.2	55.6	34.7	41.3	63.9	44.6
RetinexNet [23]	37.6	60.3	38.6	44.6	68.3	48.6
URetinexNet [24]	37.6	60.1	38.3	45.3	67.9	50.1
DRBN [26]	37.5	60.2	38.4	45.0	68.6	49.2
Zero-DCE [6]	38.7	62.0	39.0	46.4	70.0	50.9
EnlightenGAN [12]	38.4	61.5	39.2	45.8	69.5	49.7
RUAS [16]	36.1	58.6	36.4	43.8	66.7	48.0
SCI [19]	36.5	59.5	37.0	44.3	67.3	48.4
PairLIE [5]	38.4	61.7	39.0	46.0	69.1	50.8
NeRCo [25]	36.7	59.0	37.2	44.2	66.7	48.1
Ours	39.6	62.9	40.8	46.9	70.3	52.0

52.0% on AP_{75}^{box} which outperforms the second-best Zero-DCE by 1.1 points. The above ndings demonstrate the e ectiveness of our method as a pre-processing step for high-level vision tasks in weakly illuminated environments.

3 More Qualitative Results

In this section, we provide more qualitative results of our method and other competitive methods on the paired datasets including LOL [23], VE-LOL [15], and LSRW [8], as shown in Fig. 3 and Fig. 4, as well as unpaired benchmarks including DICM [13], NPE [22], VV [21], LoLi-Huawei [14], and LoLi-iPhone [14], as shown in Fig. 5-9. we can see that previous methods appear to have incorrect exposure, color distortion, noise ampli cation, or artifacts, thereby undermining the overall visual quality. In contrast, our method e ectively improves the contrast and presents vivid color without introducing artifacts. Moreover, we also provide the instance segmentation and face detection results of di erent methods on the LIS dataset [2] and DARK FACE dataset [27], as shown in Fig. 10 and Fig. 11, where our method can detect more targets, but other methods cannot.

4 Limitations

Although our method is capable enough for restoring images captured in weakly illuminated conditions, while unable to handle extremely low-light environments since images captured in such conditions su er from greater information loss and environment noise, being more challenging to restore. Moreover, since di usion-based methods iteratively denoise the randomly sampled Gaussian noise to obtain restored results in the inference phase, leading to the inference speed is a ected by the sampling step. Therefore, our method shows inferior e ciency compared to some lightweight methods. In the future, we will explore more e ective sampling strategies, such as DPM-Solver [17], to improve inference e ciency and investigate the e ectiveness of our method for other image restoration tasks.

References

1. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: The 2018 pirm challenge on perceptual image super-resolution. In: ECCV (2018)
2. Chen, L., Fu, Y., Wei, K., Zheng, D., Heide, F.: Instance segmentation in the dark. IJCV pp. 1 21 (2023)
3. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: CVPR. pp. 5203 5212 (2020)
4. Fei, B., Lyu, Z., Pan, L., Zhang, J., Yang, W., Luo, T., Zhang, B., Dai, B.: Generative di usion prior for uni ed image restoration and enhancement. In: CVPR. pp. 9935 9946 (2023)
5. 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: CVPR. pp. 22252 22261 (2023)

Fig. 3: Qualitative comparison of our method and competitive methods on the LOL [23] (row 1) and VE-LOL [15] (rows 2) test sets. Best viewed by zooming in.

Fig. 4: Qualitative comparison of our method and competitive methods on the LSRW [8] test set. Best viewed by zooming in.

6. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: CVPR. pp. 1780 1789 (2020)
7. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. IEEE TIP 26(2), 982 993 (2016)
8. 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)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: CVPR. pp. 2961 2969 (2017)

Fig. 5: Qualitative comparison of our method and competitive methods on the DICM [13] dataset. Best viewed by zooming in.

Fig. 6: Qualitative comparison of our method and competitive methods on the NPE [22] dataset. Best viewed by zooming in.

10. Hou, J., Zhu, Z., Hou, J., Liu, H., Zeng, H., Yuan, H.: Global structure-aware di usion process for low-light image enhancement. NeurIPS 36 (2023)

Fig. 7: Qualitative comparison of our method and competitive methods on the VV [21] dataset. Best viewed by zooming in.

Fig. 8: Qualitative comparison of our method and competitive methods on the LoLi-Huawei [14] dataset. Best viewed by zooming in.

11. Jiang, H., Luo, A., Han, S., Fan, H., Liu, S.: Low-light image enhancement with wavelet-based di usion models. ACM TOG 42(6), 1 14 (2023)
12. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE TIP 30, 2340 2349 (2021)
13. Lee, C., Lee, C., Kim, C.S.: Contrast enhancement based on layered di erence representation of 2d histograms. IEEE TIP 22(12), 5372 5384 (2013)
14. 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 TPAMI 44(12), 9396

Fig. 9: Qualitative comparison of our method and competitive methods on the LoLi-iPhone [14] dataset. Best viewed by zooming in.

Fig. 10: Visual Comparison of instance segmentation results using Mask R-CNN [9] on the LIS dataset [2]. Best viewed by zooming in.

9416 (2021)
15. Liu, J., Xu, D., Yang, W., Fan, M., Huang, H.: Benchmarking low-light image enhancement and beyond. IJCV 129, 1153 1184 (2021)
16. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: CVPR. pp. 10561 10570 (2021)
17. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for di usion probabilistic model sampling in around 10 steps. NeurIPS 35, 5775 5787 (2022)