From Enhancement to Understanding: Build a Generalized Bridge for Low-light Vision via Semantically Consistent Unsupervised Fine-tuning (Supplementary Material)

Sen Wang[1*] Shao Zeng[2*] Tianjun Gu[1] Zhizhong Zhang[1] Ruixin Zhang[2†] Shouhong Ding[2]
Jingyun Zhang[3] Jun Wang[3] Xin Tan[1†] Yuan Xie[1] Lizhuang Ma[1]

[1]East China Normal University [2]Tencent Youtu Lab [3]Tencent WeChat Pay Lab 33

Contents

*Equal contribution. This work was done by Sen Wang during an internship at Tencent Youtu Lab.
†Corresponding author.

A. Implementation details

A.1. Hyper Parameters

We use the AdamW optimizer to train our model, the weight decay and epsilon are set to 1e-2 and 1e-8, respectively. During training, the weights of λ_{idt} and λ_{GAN} are 0.5 and 1, respectively, we use gradient clipping with a max norm of 10, and low light and normal light images are randomly paired to ensure the generalization of the model.

A.2. Algorithm Flow

We use pseudo code as shown in Algorithm 1 to illustrate the process of our method more fully, and Algorithm 2 describes in more detail the specific process of the proposed caption consistency and reflectance consistency.

Algorithm 1 Pipeline of the Proposed Method SCUF

Input:
(1) the low-light image I_l and normal-light image I_n
(2) the V channel image $I_{l,v}$, $I_{n,v}$ and reverse image $I^r_{l,v}$, $I^r_{n,v}$ from I_l and I_n in HSV color space, respectively.
(3) the text prompt T_l and T_d for lightening and darkening, respectively.
Networks: The lighten encoder E_l and decoder D_l, the draken encoder E_d and decoder D_d, and initial fixed Unet U, lightening and darkening discriminators Dis_l and Dis_n
for i in $1 : iterations$ **do**
  for the cycle generation do
    Input I_l and obtain the generated normal-light image $\hat{I}_n$ and low-light image $I_l^{'}$ by:
    $\hat{I}_n = D_l(U(E_l(I_l), (T_l, I^r_{l,v})))$
    $I_l^{'} = D_d(U(E_d(\hat{I}_n), (T_d, I_{l,v})))$
    Do caption and reflectance consistency
    Input I_n and obtain the generated low-light image $\hat{I}_l$ and normal-light image $I_n^{'}$ by:
    $\hat{I}_l = D_d(U(E_d(I_n), (T_d, I^r_{n,v})))$
    $I_n^{'} = D_l(U(E_l(\hat{I}_l), (T_l, I_{n,v})))$
    Do caption and reflectance consistency
    Compute the L1 loss $\mathcal{L}_{l1}$ for $\mathcal{L}_{l1}(I_l, I_l^{'})$ and $\mathcal{L}_{l1}(I_n, I_n^{'})$
  end for
  for the identity regularization do
    Input I_l and obtain the generated low-light image $\hat{I}_l$ by:
    $\hat{I}_l = D_d(U(E_d(I_l), (T_d, I_{l,v})))$
    Do caption and reflectance consistency
    Input I_n and obtain the generated normal-light image $\hat{I}_n$ by:
    $\hat{I}_n = D_l(U(E_l(I_l), (T_l, I_{n,v})))$
    Do caption and reflectance consistency
    Compute the L1 loss $\mathcal{L}_{l1}$ for $\mathcal{L}_{l1}(I_l, \hat{I}_l)$ and $\mathcal{L}_{l1}(I_n, \hat{I}_n)$
  end for
  for the discriminator learning do
    learn from the fake predictions $Dis_l(\hat{I}_n)$ and $Dis_d(\hat{I}_l)$
    learn from the real inputs $Dis_d(I_l)$ and $Dis_l(I_n)$
  end for
end for

Algorithm 2 Caption and Reflectance Consistency

Input:
(1) the caption prompt Cap_l and Cap_n from I_l and I_n, respectively.
(2) the reflectance map $I_{ref,l}$ and $I_{ref,n}$ from I_l and I_n, respectively.
Networks: The reflectance decoder D_r.
Loss: The cosine similarity loss $\mathcal{COS}$, L1 loss $\mathcal{L}_{l1}$, and MSE loss $\mathcal{L}_{mse}$.
for i in 1 : $iterations$ **do**
  for the cycle generation do
    Input I_l and compute the caption consistency loss $\mathcal{L}_{cap,I_l}$ and reflectance consistency loss $\mathcal{L}_{ref,I_l}$ by:
    $Z_l = U(E_l(I_l), (T_l, I^r_{l,v}))$
    $Z_d = U(E_d(\hat{I}_n), (T_d, I_{l,v}))$
    $\mathcal{L}_{cap,I_l} = \mathcal{COS}(U(E_l(I_l), Cap_l), Z_d)$
    $\mathcal{L}_{ref,I_l} = \mathcal{L}_{mse}(D_r(Z_l), D_r(Z_d)) + \mathcal{L}_{l1}(D_r(Z_d), I_{ref,l})$
    Input I_n and compute the caption consistency loss $\mathcal{L}_{cap,I_n}$ and reflectance consistency loss $\mathcal{L}_{ref,I_n}$ by:
    $Z_d = U(E_d(I_n), (T_d, I^r_{n,v}))$
    $Z_l = U(E_l(\hat{I}_l), (T_l, I_{n,v}))$
    $\mathcal{L}_{cap,I_n} = \mathcal{COS}(U(E_d(I_n), Cap_n), Z_l)$
    $\mathcal{L}_{ref,I_n} = \mathcal{L}_{mse}(D_r(Z_d), D_r(Z_l)) + \mathcal{L}_{l1}(D_r(Z_l), I_{ref,n})$
  end for
  for the identity regularization do
    Input I_l and compute the caption consistency loss $\mathcal{L}_{cap,I_l}$ and reflectance consistency loss $\mathcal{L}_{ref,I_l}$ by:
    $Z_d = U(E_d(I_l), (T_d, I_{l,v}))$
    $\mathcal{L}_{cap,I_l} = \mathcal{COS}(U(E_l(I_l), Cap_l), Z_d)$
    $\mathcal{L}_{ref,I_l} = \mathcal{L}_{l1}(D_r(Z_d), I_{ref,l})$
    Input I_n and compute the caption consistency loss $\mathcal{L}_{cap,I_n}$ and reflectance consistency loss $\mathcal{L}_{ref,I_n}$ by:
    $Z_l = U(E_l(I_n), (T_l, I_{n,v}))$
    $\mathcal{L}_{cap,I_n} = \mathcal{COS}(U(E_d(I_n), Cap_n), Z_l)$
    $\mathcal{L}_{ref,I_n} = \mathcal{L}_{l1}(D_r(Z_l), I_{ref,n})$
  end for
end for

B. Experimental Comparisons

B.1. Detailed Quantitative Analysis Results.

Since training datasets used by unsupervised low-light enhancement methods are different, we follow [14] to explain training sets of all methods. In the paper, we only show results of RUAS[11] and SCI[12] trained on LOL[19]. We can see from Tab. 1 that our method performs best on high-level vision tasks and shows the best generalization. We also show the result of the model trained on the LSRW[5] dataset, where we can see that its performance on high-level vision tasks is not as good as that trained on EnlightGan[7], but still outperforms most existing low-light enhancement methods.

Table 1. Compare with existing low-light enhancement methods. 'T', 'S', and 'U' indicate traditional, supervised, and unsupervised methods, respectively. $*$ denotes our re-implementation with the same training data we use. The best results are highlighted in **bold**.

Type	Method	Venue & Years	Train Set	LSRW[5]			LOL[19]			CODaN[8]	DARK FACE[18]	BDD100K-night[21]
				PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Top-1(%)	mAP(%)	mIoU(%)
T	LIME [4]	TIP'16	N/A	14.88	0.3487	0.4030	16.90	0.4917	0.4022	14.09	11.0	14.2
	DUAL [22]	CGF'19	N/A	13.76	0.3532	0.4150	16.76	0.4911	0.4060	14.67	11.0	14.1
S	RetinexNet [15]	BMCV'18	LOL	15.59	0.4176	0.3998	17.68	0.6477	0.4433	47.48	13.2	13.2
	Retinexformer [1]	ICCV'23	LOL	17.19	0.5093	0.3314	22.79	0.8397	0.1707	52.81	16.4	15.9
	CIDNet [16]	CVPR'25	LOL	18.00	0.5198	0.2962	20.68	0.8411	0.1156	58.32	14.5	17.4
U	EnlightenGan [7]	TIP'21	own data	17.59	0.4867	0.3117	18.68	0.6728	0.3013	56.42	14.2	16.6
	Zero-DCE [3]	CVPR'20	own data	15.86	0.4536	0.3176	18.06	0.5736	0.3125	57.76	15.9	16.6
	Zero-DCE++ [9]	TPAMI'21	own data	16.21	0.4571	0.3266	17.37	0.4373	0.3118	59.88	15.2	17.7
	RUAS_upe [11]	CVPR'21	MIT	13.00	0.3442	0.3989	13.97	0.4656	0.3401	57.26	12.8	18.6
	RUAS_lol [11]	CVPR'21	LOL	14.33	0.4841	0.4800	15.33	0.4876	0.3097	51.60	14.0	15.2
	RUAS_dark [11]	CVPR'21	DARK FACE	14.11	0.4183	0.3811	14.89	0.4553	0.3722	55.42	12.0	16.5
	SCI_easy [12]	CVPR'22	MIT	11.79	0.3174	0.4004	11.98	0.3986	0.3543	59.76	14.0	17.5
	SCI_medium [12]	CVPR'22	LOL	15.24	0.4240	0.3218	17.30	0.5335	0.3079	58.84	14.7	18.0
	SCI_difficult [12]	CVPR'22	DARK FACE	15.16	0.4080	0.3259	17.25	0.5462	0.3171	59.56	14.8	17.4
	PairLIE [2]	CVPR'23	own data	17.60	0.5118	0.3290	19.88	0.7777	0.2341	52.29	16.0	16.4
	SADG [23]	AAAI'23	own data	16.32	0.4564	0.3471	16.93	0.5372	0.3513	56.80	14.9	14.8
	CLIP-LIT [10]	ICCV'23	own data	13.47	0.4089	0.3572	15.18	0.5290	0.3689	54.64	14.1	17.3
	NeRCo [17]	ICCV'23	LSRW	**19.46**	0.5506	0.3052	19.66	0.7172	0.2705	54.15	12.4	18.1
	QuadPrior [14]	CVPR'24	COCO	16.90	0.5429	0.3459	20.30	0.7909	**0.1858**	59.48	15.7	14.9
	ZERO-IG_LSRW [13]	CVPR'24	LSRW	18.21	**0.5665**	0.4946	18.65	0.4819	0.3819	47.60	15.6	14.9
	ZERO-IG_LOL [13]	CVPR'24	LOL	16.44	0.5033	0.3744	18.13	0.7455	0.2478	53.48	15.2	14.7
	LightenDiffusion [6]	ECCV'24	own data	18.42	0.5334	0.3209	22.79	0.8540	0.1666	57.40	16.3	16.0
	LightenDiffusion*	ECCV'24	EnlightenGan data	16.92	0.5250	0.3824	18.27	0.7944	0.2457	57.32	16.4	16.8
	Ours		EnlightenGan data	18.41	0.5341	0.2974	**21.32**	**0.8073**	0.1928	**60.92**	**16.9**	**20.1**
	Ours-LSRW		LSRW	18.96	0.5438	**0.2673**	20.22	0.7649	0.2157	60.56	16.3	18.0

B.2. Visual Quality Comparison.

We also show enhancement results of different low-light enhancement methods, as shown in Fig. 1, Fig. 2 and Fig. 3. Our model achieves relatively high-fidelity results.

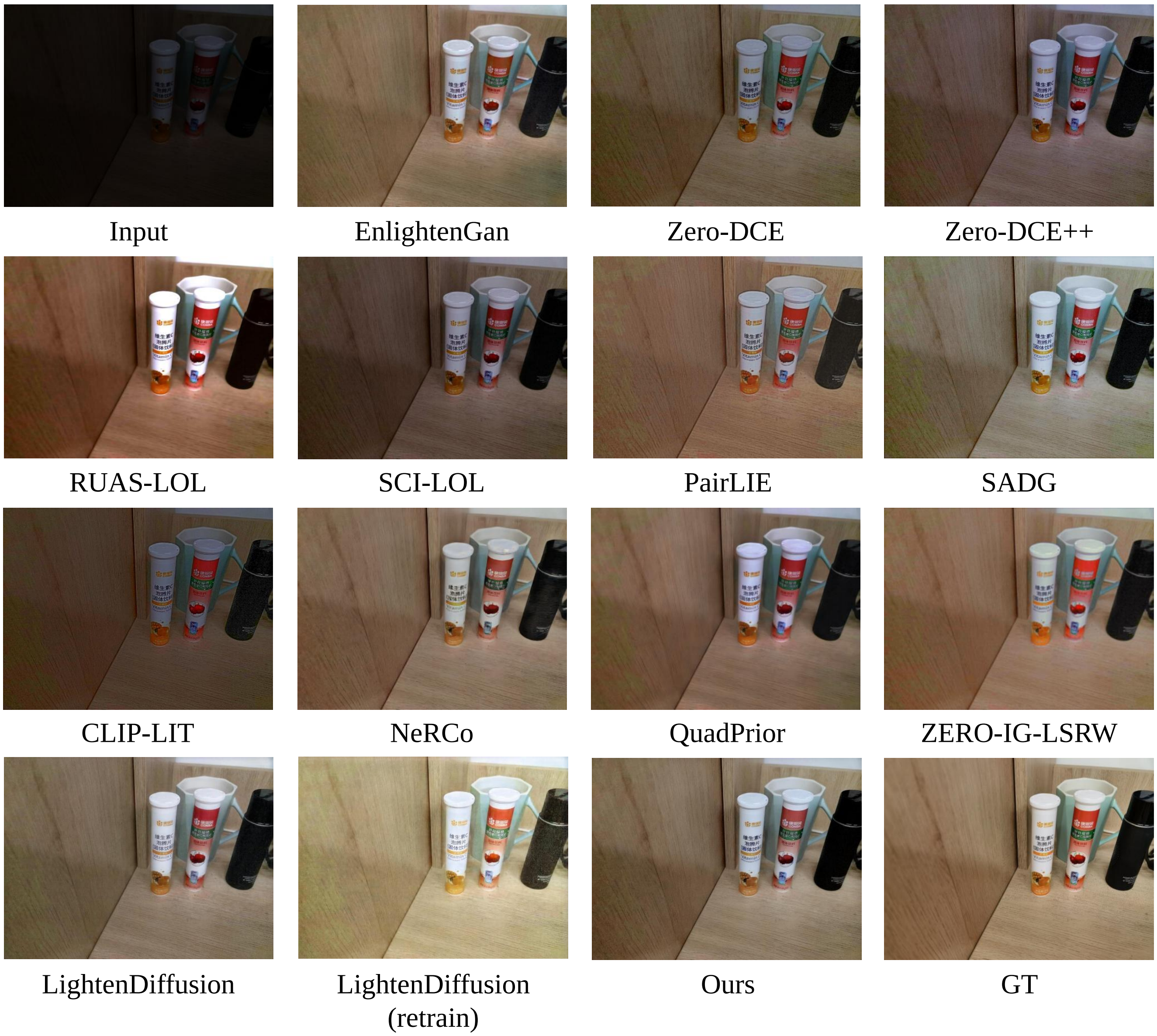

Figure 1. Visual quality comparison between the proposed method and other state-of-the-art methods on the LSRW[5].

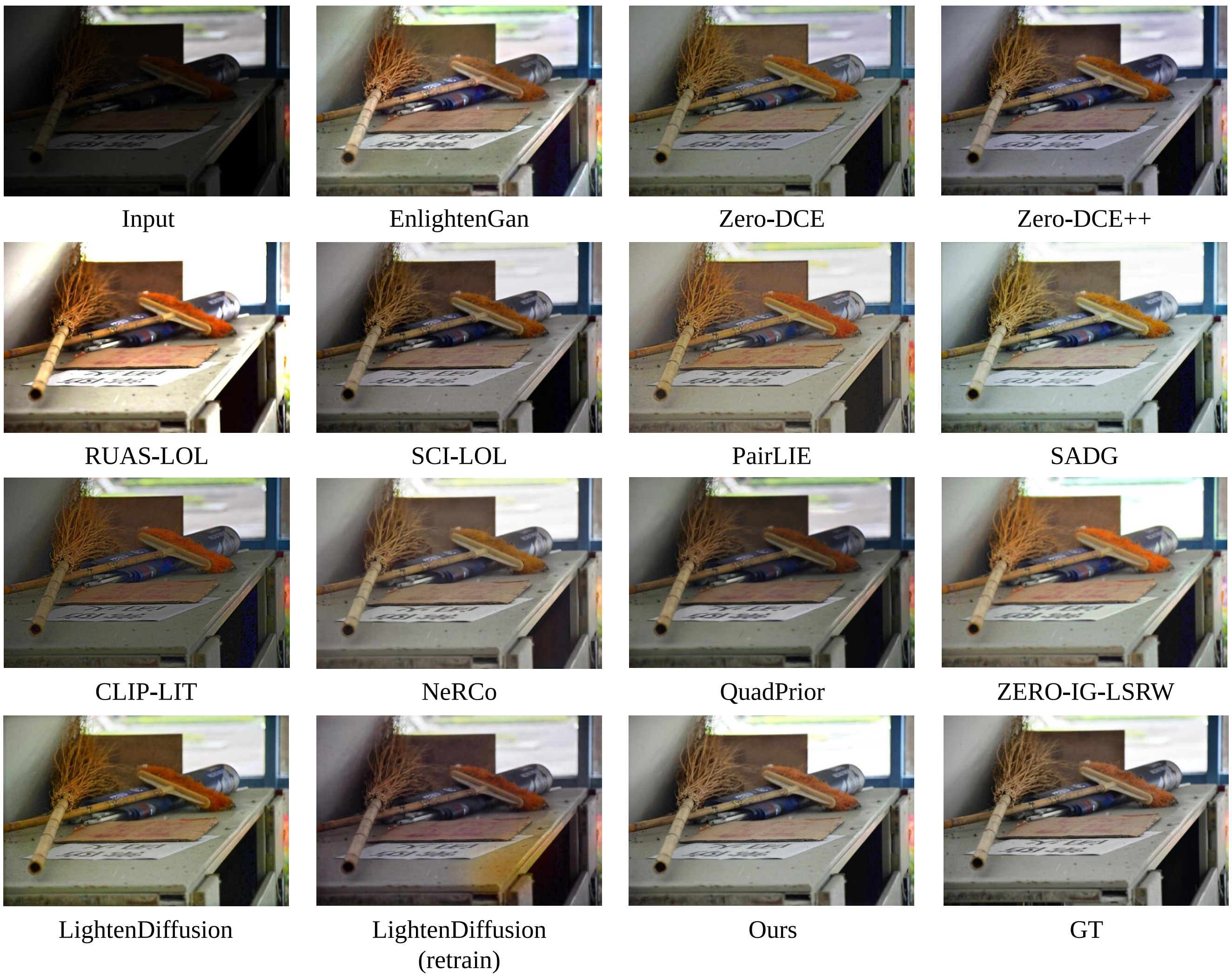

Figure 2. Visual quality comparison between the proposed method and other state-of-the-art methods on the LSRW[5].

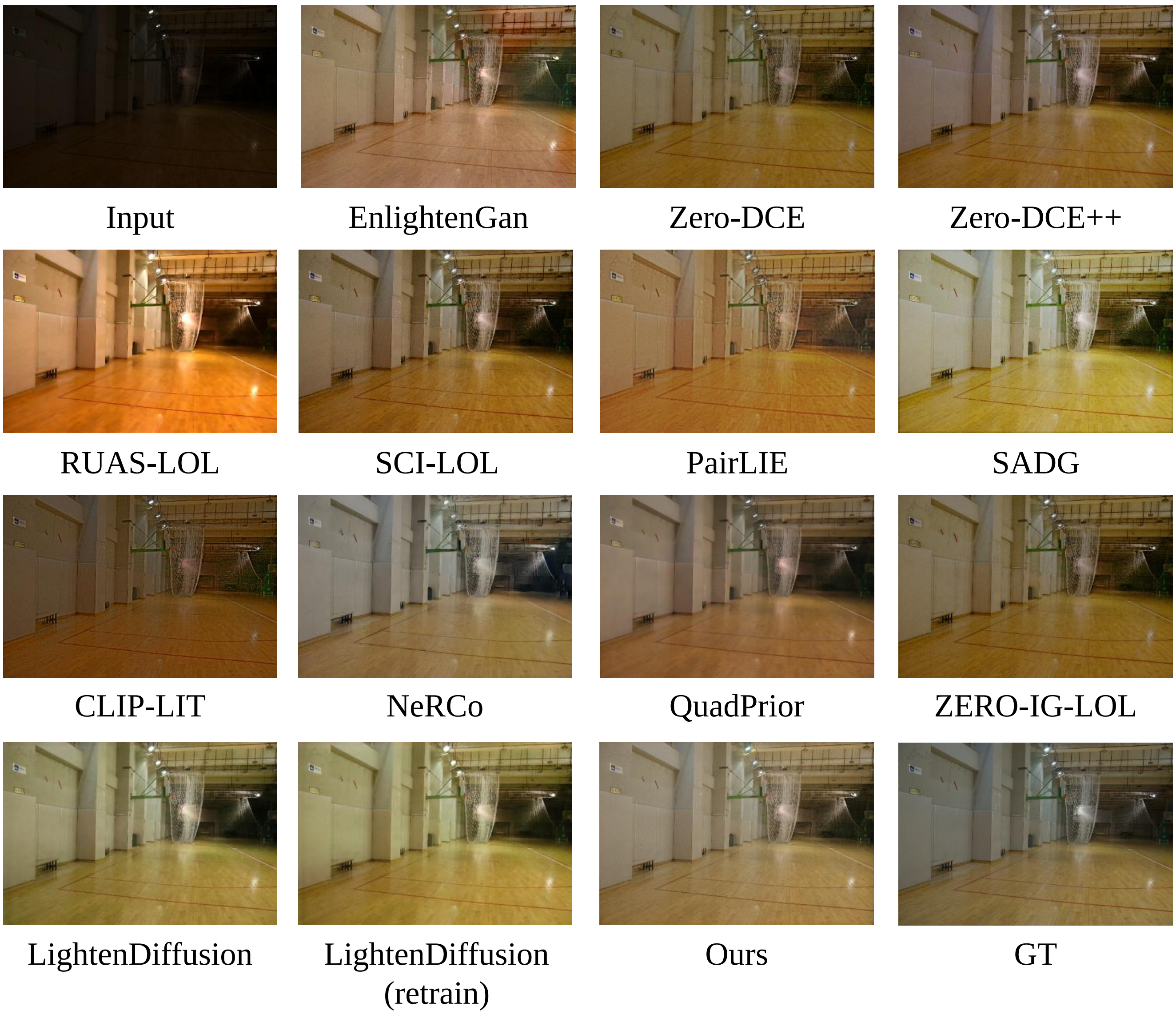

Figure 3. Visual quality comparison between the proposed method and other state-of-the-art methods on the LOL[19].

B.3. High-level Vision Comparison.

We show results of existing low-light enhancement methods on night image classification in Fig. 4, low-light face detection in Fig. 5 and Fig. 6, and nighttime semantic segmentation in Fig. 7. We can see that our method achieves the best performance.

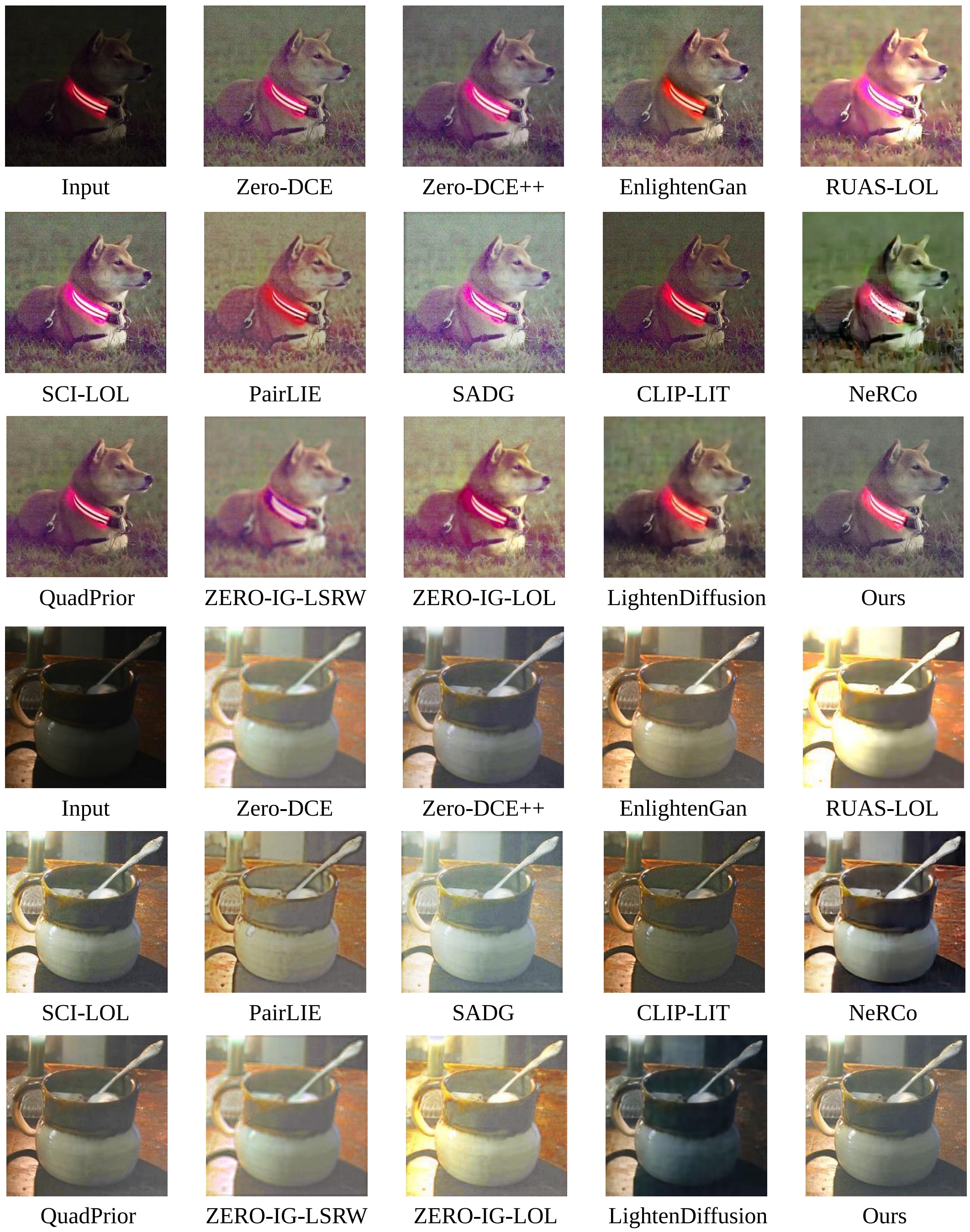

Figure 4. Qualitative comparison of the proposed method with other state-of-the-art low-light enhancement methods on night image classification on CODaN[8].

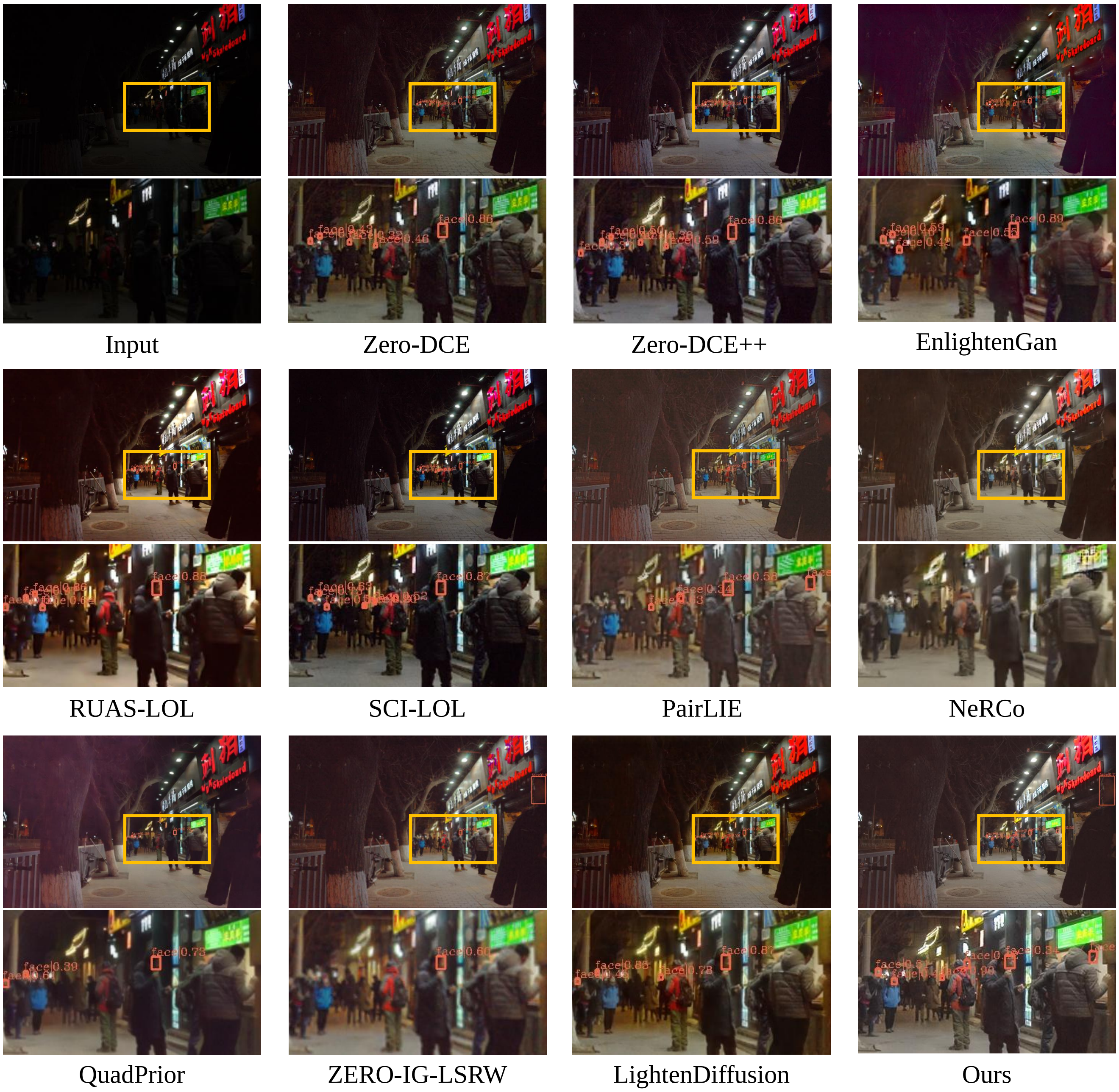

Figure 5. Qualitative comparison of the proposed method with other state-of-the-art low-light enhancement methods on dark face detection on DARK FACE[18].

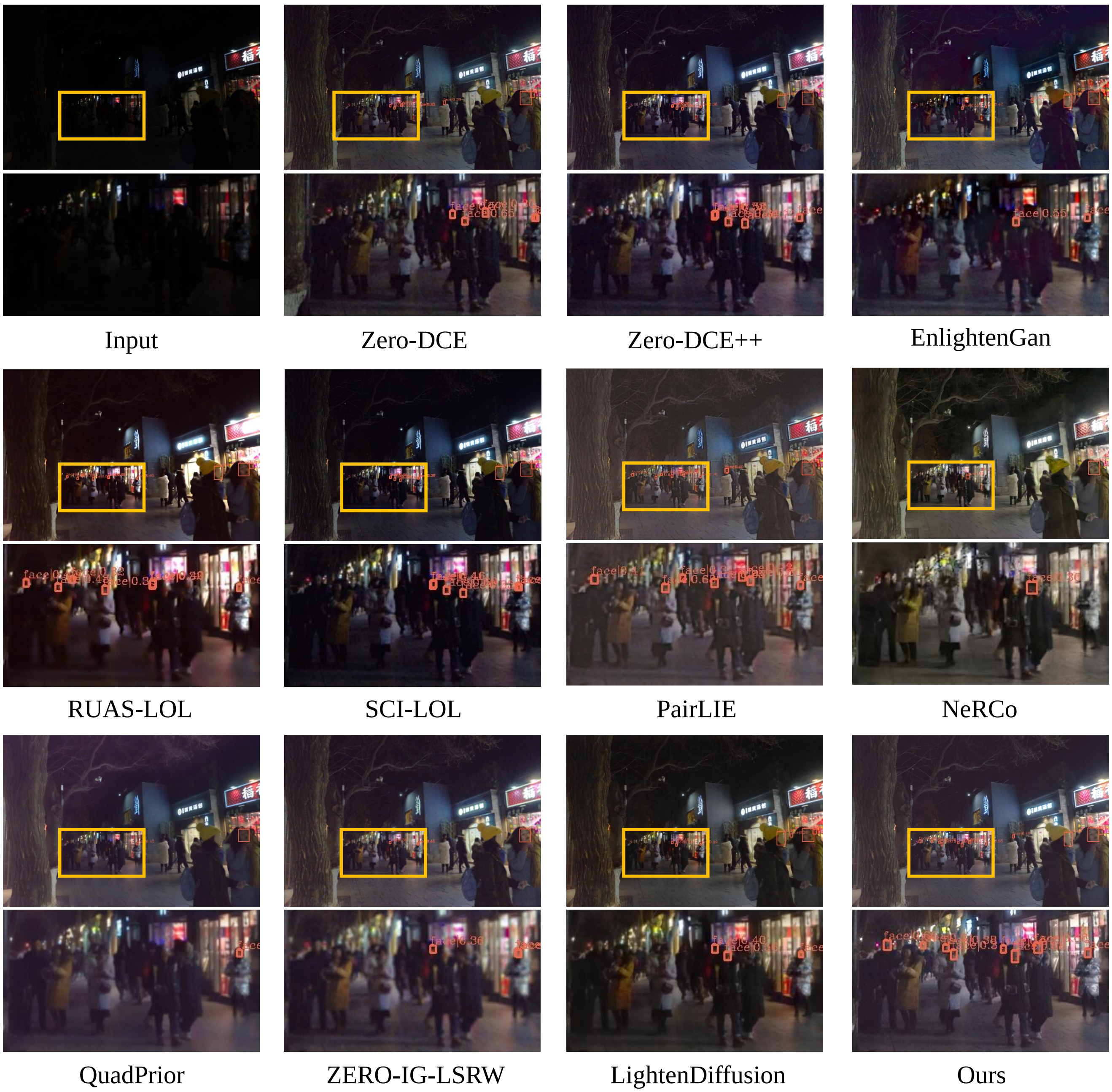

Figure 6. Qualitative comparison of the proposed method with other state-of-the-art low-light enhancement methods on dark face detection on DARK FACE[18].

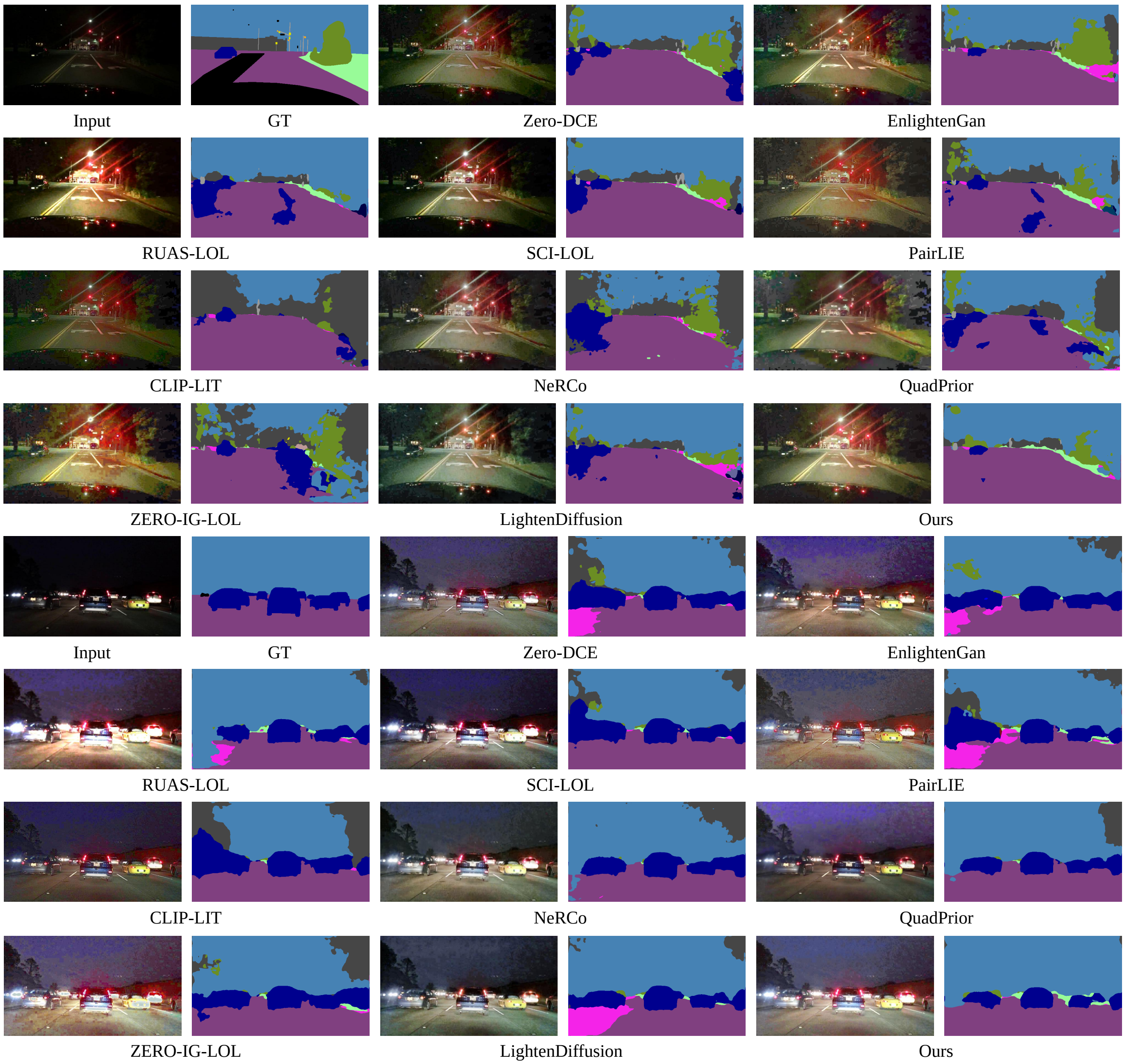

Figure 7. Qualitative comparison of the proposed method with other state-of-the-art low-light enhancement methods on nighttime image semantic segmentation on BDD100k-night[21].

B.4. Visual Analysis of Different Adapters.

In the paper, we compared the visual analysis of image feature extraction by different components. We supplement the visual analysis experiments using different adapters. As shown in Fig. 8, results using the IP-Adapter[20] are even worse than the original adapter that directly connects text and image features. This fully demonstrates that the IP-Adapter is not suitable for the illumination-aware image prompt. The cycle-attention adapter we proposed fully exploits the semantic features of the illumination-aware image prompt and achieves the best result.

Figure 8. Visual analysis of the different adapters.

B.5. Another Baseline on Normal Light Images.

The pre-trained downstream models tend to overfit on training data, such as classification and detection results as shown in Sec. B.5, while the normal light segmentation results are close to our enhanced low light images because they are tested on BDD100k, which the model has not seen.

Setting	Cls Top-1(%)	Det mAP(%)	Seg mIoU(%)
Pretrained Data	CODaN-day	WIDER FACE	Cityscapes
Normal light Data	CODaN-day	WIDER FACE	BDD100k-day
Baseline (Normal)	82.52	55.9	23.1
Low light Data	CODaN-dark	DARK FACE	BDD100k-night
Baseline (Low)	53.24	10.8	11.4
Ours	60.92	16.9	20.1

C. Failure cases.

While our method generalizes better than existing approaches, two challenges remain as shown in Fig. 9. First, due to detail loss in low-light images, small object detection remains difficult, which is a limitation across most methods. Second, under extreme low-light degradation, enhanced outputs may still contain noise and artifacts, hindering complete restoration (e.g., the tree in segmentation). Transferring semantic knowledge from normal-light conditions to guide restoration remains a challenging problem. Future work needs to explore more diverse low-light scenarios and the performance upper bound.

Figure 9. Fail cases.

References

[1] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12504–12513, 2023. 4

[2] Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22252–22261, 2023. 4

[3] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020. 4

[4] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016. 4

[5] Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation*, 90:103712, 2023. 4, 5, 6

[6] Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. *arXiv preprint arXiv:2407.08939*, 2024. 4

[7] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021. 4

[8] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021. 4, 8

[9] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4225–4238, 2021. 4

[10] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8094–8103, 2023. 4

[11] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10561–10570, 2021. 4

[12] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5637–5646, 2022. 4

[13] Yiqi Shi, Duo Liu, Liguo Zhang, Ye Tian, Xuezhi Xia, and Xiaojing Fu. Zero-ig: Zero-shot illumination-guided joint denoising and adaptive enhancement for low-light images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3015–3024, 2024. 4

[14] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26057–26066, 2024. 4

[15] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement, 2018. 4

[16] Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. *arXiv preprint arXiv:2502.20272*, 2025. 4

[17] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12918–12927, 2023. 4

[18] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, , and et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020. 4, 9, 10

[19] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. 4, 7

[20] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 12

[21] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 4, 11

[22] Qing Zhang, Yongwei Nie, and Wei-Shi Zheng. Dual illumination estimation for robust exposure correction. In *Computer graphics forum*, pages 243–252. Wiley Online Library, 2019. 4

[23] Naishan Zheng, Jie Huang, Man Zhou, Zizheng Yang, Qi Zhu, and Feng Zhao. Learning semantic degradation-aware guidance for recognition-driven unsupervised low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3678–3686, 2023. 4