Improving Visual and Downstream Performance of Low-Light Enhancer with Vision Foundation Models Collaboration

Supplementary Material

In the supplementary material, we provide a more comprehensive analysis through detailed experiments, systematically divided into three sections. Section A outlines the experimental setup in detail, including the composition of the loss function, the selection of pre-trained models, and the hyperparameter settings. Section B integrats DINOv2 into the proposed FoFA. Section C presents an in-depth evaluation of GaTA, along with visual comparisons across various tasks. Section D offers an extensive collection of visual comparisons, highlighting the practical advantages of the proposed method.

A. Detailed Experiment Settings

As described in Section 3, the loss function used in our experiments is formulated as follows:

$$\mathcal{L} = \lambda_C * \mathcal{L}_C + \lambda_{FoFA} * \mathcal{L}_{FoFA} + \lambda_{GaTA} * \mathcal{L}_{GaTA} \quad (14)$$

Here, the first term represents the illumination consistency loss, while the second and third terms correspond to the FoFA and GaTA loss components, respectively. The parameters λ are weighting coefficients. During the experiments, λ_C, λ_{FoFA} and λ_{GaTA} are set to 1, 0.5, 0.1 for application, 1, 0.3, and 0.1 for Nikon and Huawei and 2, 0.3, and 0.1 for FiveK respectively, with the illumination consistency loss defined as:

$$\mathcal{L}_s = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}(i)} w_{i,j} \left| G(x)_i - G(x)_j \right|, s.t. x \sim p_{low}(x), \quad (15)$$

$$\mathcal{L}_c = \left\| E - G(x) \right\|, s.t. x \sim p_{low}(x), \quad (16)$$

$$\mathcal{L}_C = \lambda_c * \mathcal{L}_c + \mathcal{L}_s, \quad (17)$$

where $\mathcal{N}$ is the total number of pixels. i is the i-th pixel. $\mathcal{N}(i)$ denotes the adjacent pixels of i in its 5×5 window. α was set to 7. Additionally, the FoFA loss is formulated as:

$$\mathcal{L}_C = \lambda_G * \mathcal{L}_G + \lambda_D * \mathcal{L}_D, \quad (18)$$

where λ_G is set to 1 and λ_D is set to 0.1. The Adam optimizer is employed for model optimization. The generator used a learning rate of 1×10^{-4}, while the discriminator's learning rate is set to 1×10^{-4}. The discriminator is optimized every 10 times the generator is optimized.

For the pre-trained backbone models used during training, we adopt imagenet-pretrained ResNet18 for ResNet, the official CLIP ViT-B/32 for CLIP, and SAM2.1/hiera-tiny from the official implementation for SAM.

B. FoFA with Other Model

In this section, we conduct experiments to investigate the compatibility of FoFA with DINOv2, a foundational model renowned for its rich feature knowledge and robust extraction capabilities. As illustrated in Table 5 (where the experimental settings are consistent with those in Table 3), DINOv2 exhibits superior performance and further enhances results when integrated with other models. This observation is consistent with our analysis in Section 4.5, where we noted that the diversity of features across different models contributes to enhanced performance. Although DINOv2 provides rich representations, other foundational models with distinct feature characteristics can still complement its capabilities.

Table 5. Results with DINOv2.

	PSNR	SSIM	LPIPS	AP	mAP∼Segm	Top-1 Acc
Ours	**19.58**	**0.83**	**0.13**	**0.68**	0.28	**0.59**
DINOv2	18.49	0.81	0.24	0.67	0.29	0.58
Ours+DINOv2	18.92	0.82	0.24	0.68	**0.29**	0.58

C. GaTA for Various Tasks

In Section 3.3, we employ segmentation and detection tasks to implement GaTA. This section provides a detailed explanation of the process. PSPNet and YOLOX are used as detectors for GaTA. Through data filtering and augmentation techniques, we train both their F^l and F^s variants, with the detailed experimental settings summarized in Table 6. The HQ ratio indicates the proportion of the highest-quality images selected from the training set; for instance, YOLOX's F^s variant selects the top 30% of high-quality images for training. The Gamma range specifies the range of γ values used during data augmentation.

Table 6. Comparison of PSPNet and YOLOX configurations.

Parameter	**PSPNet**	**YOLOX**
Backbone	ResNet50	CSPDarknet
Optimizer	SGD	SGD
LR max	0.01	0.01
Steps	8×10^4	4×10^6
LR Scheduler	PolyLR	Cosine
Training Set	ADE20K	COCO2017
Image Resolution	512×512	640×640
Gamma Range	3–0.1	3–0.1
HQ Ratio	50%	30%

D. More visual Comparisons

In this section, we provide a detailed visual comparison of every benchmark, with additional visual results of the proposed method on each dataset. All methods listed in Table 1 and 2, including Retinexformer, LLFlow, SKF, KinD++, SGZ, Zero-DCE, SCL-LLE, RUAS, EnlightGAN, CoLIE and SCI are included in the comparison, further demonstrating the superiority of the proposed method across multiple tasks. While alternative methods may struggle with incomplete degradation removal or loss of semantic information, our method consistently demonstrates its ability to effectively recover image details. This further substantiates the advanced capabilities of the proposed framework.

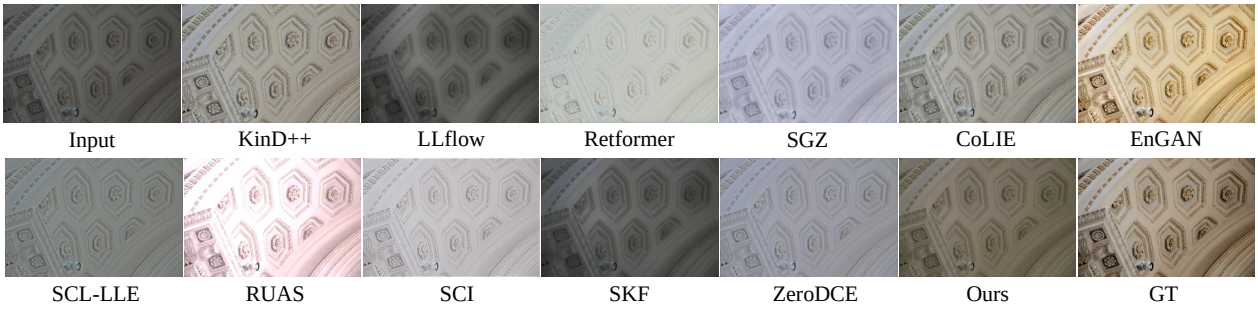

Figure 8. Visual comparison on FiveK Dataset.

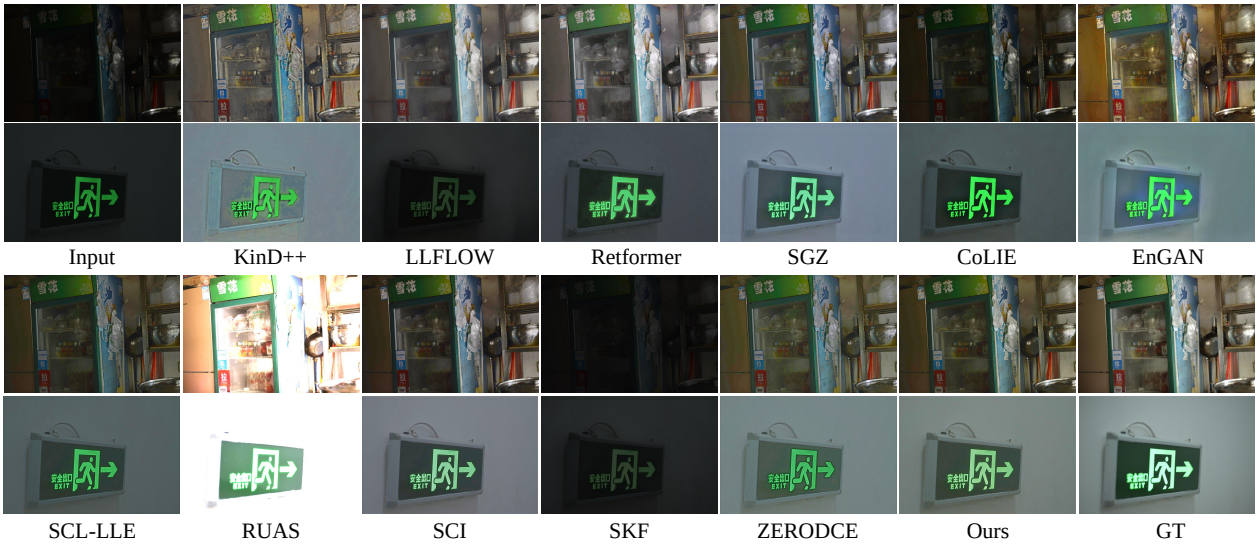

Figure 9. Visual comparison on Nikon Dataset.

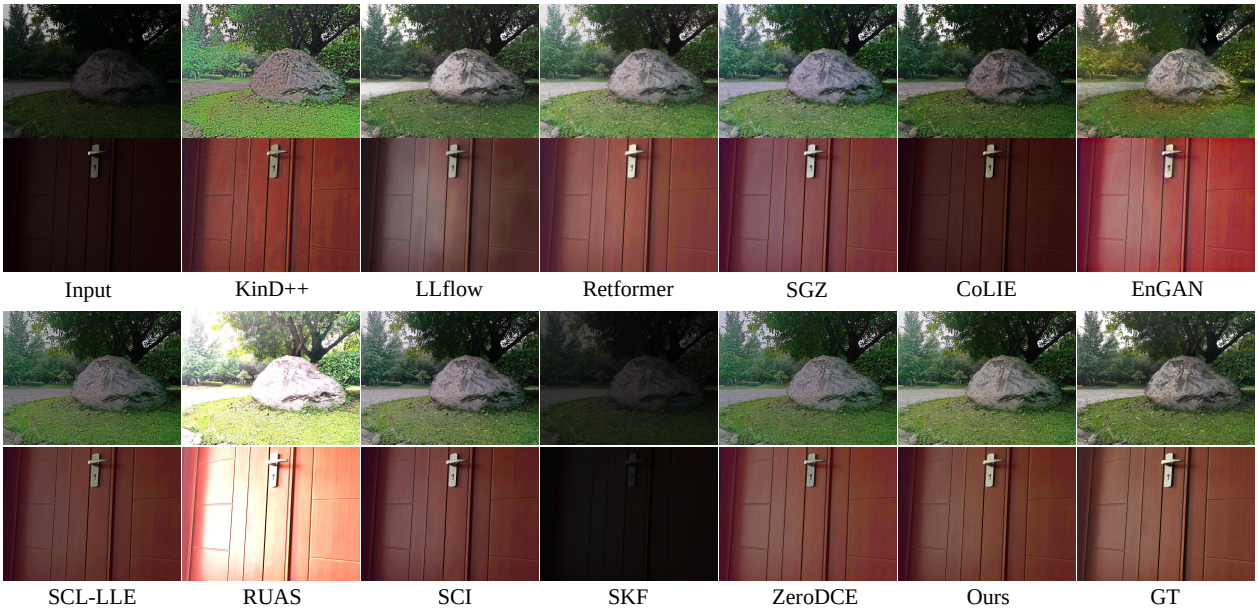

Figure 10. Visual comparison on Huawei Dataset.

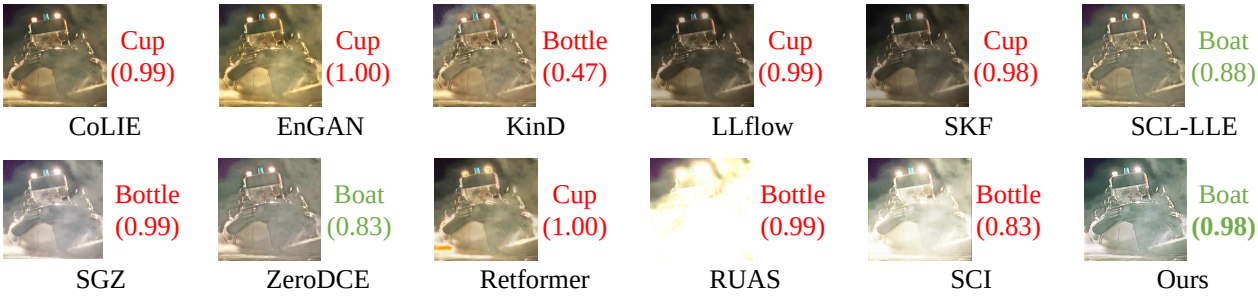

Figure 11. Visual comparison on CODaN Dataset.

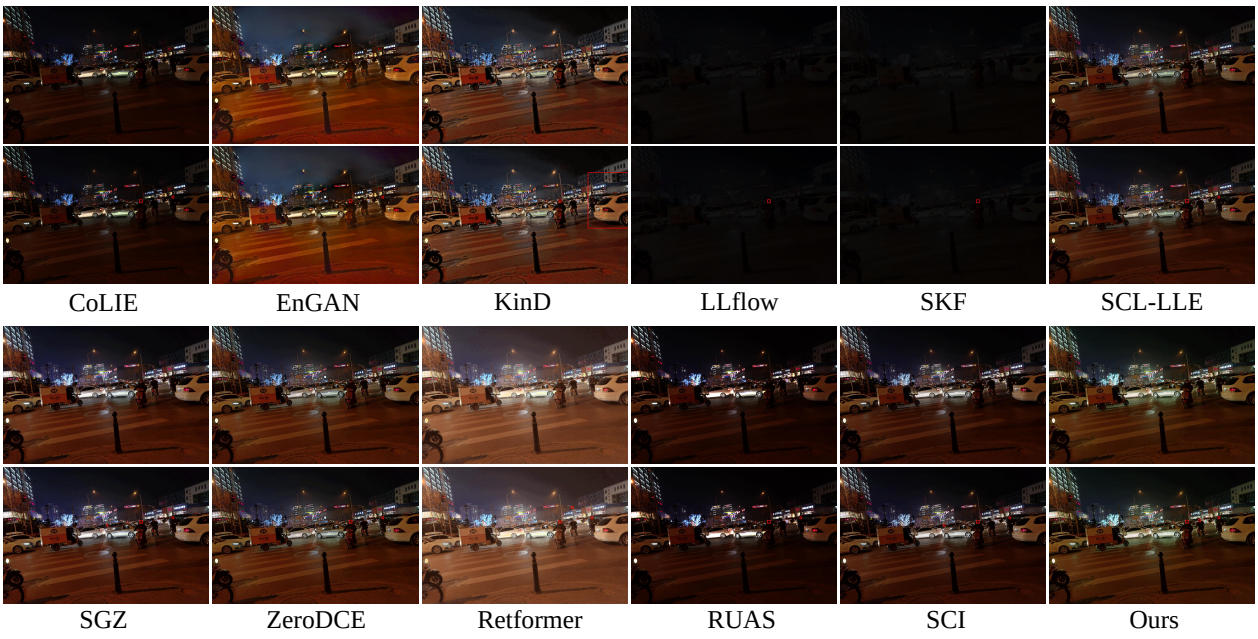

Figure 12. Visual comparison on Darkface Dataset.

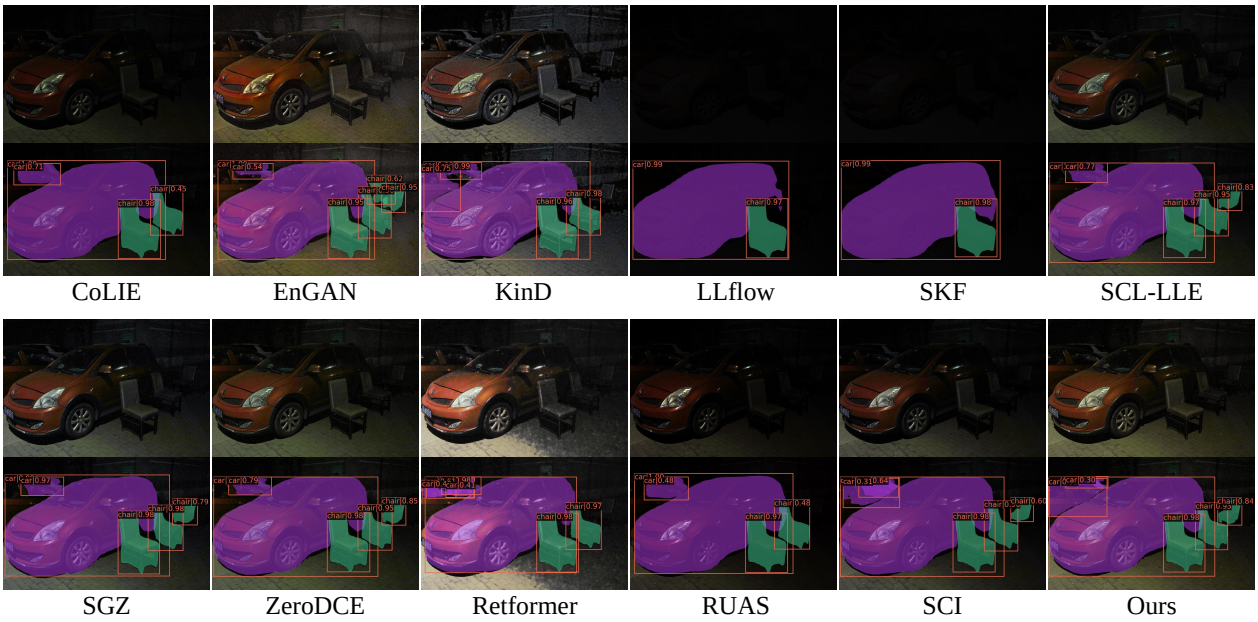

Figure 13. Visual comparison on LIS Dataset.