Supplementary Material

7. Ablation studies for ISPT

In this section, we conduct ablation experiments to assess the individual contributions of components in ISPT. Fig. 8 illustrates the data augmentation process of ISPT, where each component is systematically ablated to verify its effectiveness.

Specifically, we first reduce the influence of the gamma transform and contrast adjustment operations separately. For the gamma transform, the parameter γ is constrained within the range [1, 5], and the contrast adjustment factor α is limited to [0.55, 1]. In addition, we respectively conduct experiments by completely removing the gamma transform and contrast adjustment to further evaluate their roles. Finally, the noise introduction process is respectively replaced with impulse noise with the noise density being in the range of [0, 0.2] and directly removed to evaluate its contribution.

As shown in Tab. 5, our model outperforms alternative configurations of ISPT, demonstrating the effectiveness of its individual components. Additionally, Fig. 9 highlights key observations: the absence of the gamma transform significantly impairs the model's ability to enhance image quality. When contrast adjustment is omitted, the generated images exhibit brightness and color distortions. Finally, the exclusion of noise introduction of ISPT reduces the model's ability to effectively address the inherent noise in the original images.

8. User studies

We conduct user studies to evaluate the performance of our method. We select 20 images from the testing set of MSRS [6], KAIST-MS [2]. For each image, it is enhanced by seven methods (EnlightenGAN [3], PyDiff [9], EMMA [8], EMD [3–5], PDMD [4, 5, 9], and ours). Subsequently, 15 participants are invited to choose the best method regarding contrast and luminance, artifacts (*e.g.* noise, blurring, color deviation, and texture distortion), object information (*e.g.* pedestrians, vehicles, and road markings), and overall perceptual quality. All of the participants have good English proficiency and most of them have basic knowledge of computer vision. An example of the questionary interface for different models in different dimensions can be seen in Fig. 10.

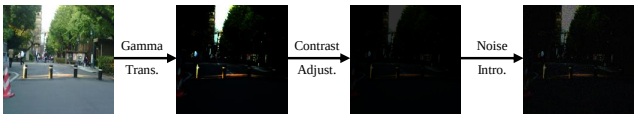

Figure 8. The data augmentation process of ISPT.

Figure 9. Qualitative results on discussing components of ISPT.

As shown in Tab. 6, except for the salience of object information, our model achieves much better results than the other six methods. Although VIF methods can generate images with thermal radiation foreground objects contrasting with the background, more participants think our model achieves better overall object salience.

9. Analysis for information fidelity

To evaluate the information fidelity between the model outputs and authentic information, in this part, we conduct experiments to test the consistency between output images and ground truth. Since the absence of ground truth in the existing aligned visible and infrared image datasets, we

Configurations	SD↑	EN↑	NIQE↓	BRISQUE↓
Reduced γ	45.56	7.04	4.02	20.72
Reduced α	44.60	7.13	3.68	18.74
w/o Gamma Transform	12.71	4.47	9.01	37.91
w/o Contrast Adjustment	42.49	7.08	3.91	22.01
Impulse Noise	46.02	6.94	3.67	17.51
w/o Noise	46.92	**7.23**	4.00	20.96
Original	**48.81**	7.15	**3.57**	**14.44**

Table 5. Ablation experiment results on MSRS dataset. The best results are marked with **bold**.

Dimension	EnGAN [3]	PyDiff [9]	MUGAN [5]	EMMA [8]	EMD	PDMD	Ours
Contrast and Luminance	5.33%	7.67%	11%	0.33%	7%	10.33%	**58.33%**
Artifacts	4%	8.67%	3%	6%	5%	6.33%	**67%**
Object Information	2.33%	3%	6.67%	31%	6%	7.67%	**43.33%**
Overall Quality	7%	8.67%	5.67%	5.67%	8%	9.67%	**55.33%**

Table 6. The result for user study, where the values represent the percentage of votes obtained by each method in a certain dimension for all images and participants. The best result is marked with **bold**.

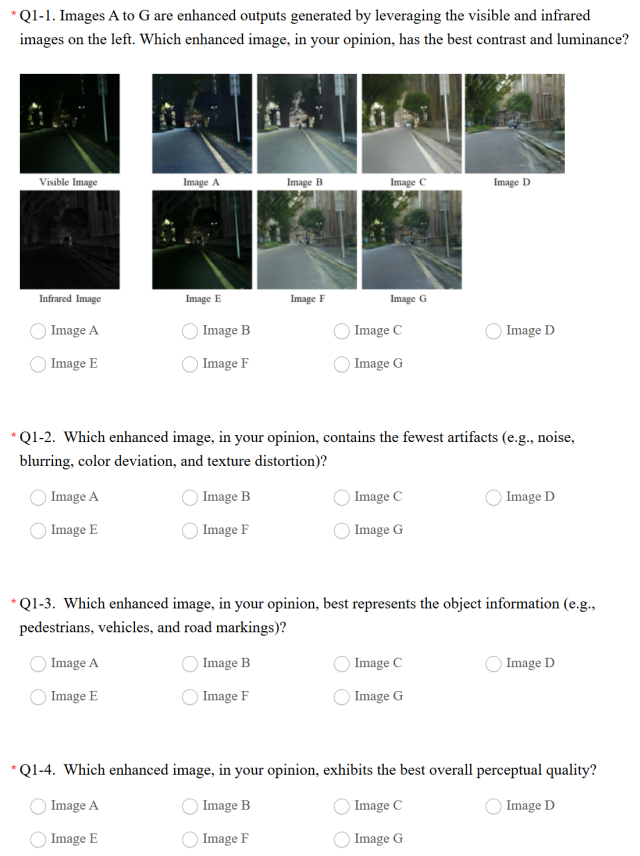

Figure 10. Screenshot of the questionary interface from user study for different models in different dimensions.

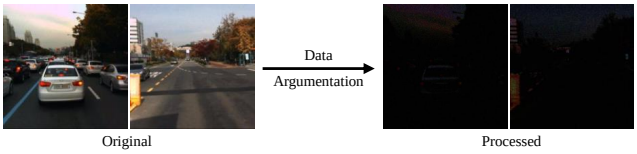

Figure 11. Some instances in the testing dataset where the high-quality original images undergo data augmentation to get the pseudo-low-light images.

construct a synthetic dataset. Specifically, we conduct the fidelity experiment on KAIST-MS [2] dataset and employ the image augmentation method of the information synthesis pretext task (ISPT) to degrade clear high-quality images in the daytime testing set. Using the degraded pseudo-low-light visible images and the corresponding infrared images, our model and baselines are employed to regenerate the original images. Subsequently, we calculate the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) metrics between the generated images and the original images.

Baseline	Enhancement	Colorization	Fusion
EMD	EnGAN [3]	MUGAN [5]	Defusion [4]
UMD	Uformer [7]		
RMD	Reformer [1]		
PDMD	PyDiff [9]		

Table 7. The newly designed baselines for information fidelity experiment.

9.1. Experimental settings

Baselines. We compare our model against four low-light image enhancement methods EnlightenGAN [3], Uformer [7], Retinexformer [1] and PyDiff [9] and one infrared image colorization method MUGAN [5]. In addition, we also utilize the versatile self-supervise-based image fusion model, Defusion [4] to combine the low-light image enhancement and infrared image colorization methods to build up new baselines which are shown in Tab. 7.

Implementation details. The training dataset and parameter configuration of the data argumentation of our model and the four low-light image enhancement methods are the same as those in Sec. 4.1 of the main paper. For the testing dataset, as the KAIST-MS dataset is derived from videos and displays frame similarities, we select one image pair from every 400 frames to construct the testing dataset. Additionally, the parameters of the image augmentation for the testing dataset are set as follows: the gamma parameter γ is configured at 6, the contrast factor α is set to 0.5, and the noise parameters are set as $\lambda = 10, \sigma = 5$. Some instances in the testing dataset are shown in Fig. 11.

9.2. Model comparisons.

As shown in Fig. 12, the low-light image enhancement methods fail to restore these severely corrupted image regions which lack essential information and introduce artifacts and blurring. The infrared image colorization method

Metric	EnGAN [3]	Uformer [7]	Reformer [1]	PyDiff [9]	MUGAN [5]	EMD	UMD	RMD	PDMD	Ours
SSIM	0.52	0.59	0.59	0.62	0.54	0.59	0.61	0.62	0.62	**0.69**
PSNR	14.35	15.71	18.87	17.44	16.29	16.56	17.53	18.74	18.44	**21.88**

Table 8. Quantitative results for fidelity experiment. The best result is marked with **bold**.

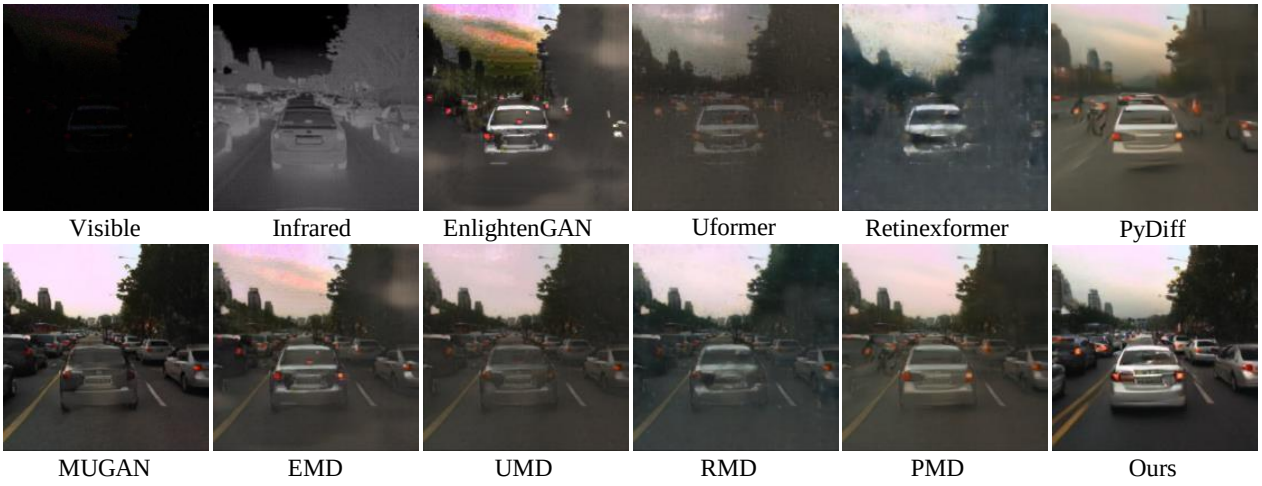

Figure 12. Qualitative comparisons for fidelity experiment. The ground truth is shown on the left of Fig. 11.

MUGAN [5] is unable to generate authentic color and texture information. Although the newly designed baselines realize information complementation between the above two sorts of methods, they can't avoid the intrinsic problems in them (*e.g.*, unauthentic color and artifacts). In contrast, our method restores the processed image to the original with high fidelity. Additionally, according to Tab. 8, our model outperforms all of the aforementioned methods with significant margins (+3 and +0.07 in PSNR and SSIM metrics, respectively).

10. Diagram of sparse cross-attention module

In this section, as shown in Fig. 13, we illustrate the sparse cross-attention module with an intermediate feature F_e^l of the encoder of the Unet-based denoising network.

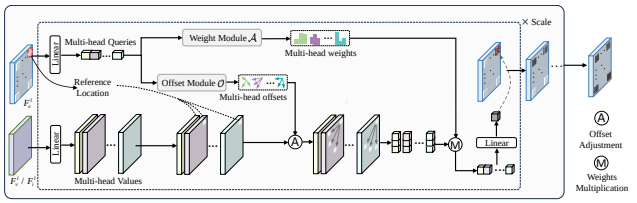

Figure 13. In SCAM, the intermediate feature F_e^l from the Unet-based denoising network and visible/infrared features F_v^l, F_i^l are initially embedded to get muti-head queries and values. Subsequently, each query element attends to elements surrounding its reference location in values, utilizing offset module $\mathcal{O}$ and weight module $\mathcal{A}$, both of which are linear blocks. Finally, the muti-head outputs undergo a linear projection operation to update the element of F_e^l.

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 International Conference on Computer Vision*, pages 12504–12513, 2023. 2, 3

[2] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1037–1045, 2015. 1, 2

[3] 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. 1, 2, 3

[4] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision*, pages 719–735, 2022. 1, 2

[5] Hangying Liao, Qian Jiang, Xin Jin, Ling Liu, Lin Liu, Shin-Jye Lee, and Wei Zhou. Mugan: thermal infrared image colorization using mixed-skipping unet and generative adversarial network. *IEEE Transactions on Intelligent Vehicles*, 8(4):2954–2969, 2023. 1, 2, 3

[6] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022. 1

[7] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17662–17672, 2022. 2, 3

[8] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25912–25921, 2024. 1, 2

[9] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. In *IJCAI*, pages 1795–1803, 2023. 1, 2, 3