Supplementary Material for DifIISR

Anonymous CVPR submission

Paper ID 5254

A. The Complete Derivation for Loss-gradient Guidance

As mentioned in Section 4.1, we provide a detailed derivation of our main method in the supplementary materials. We first provide an explanation for Equation (7) in our main text.

$$\nabla_{x_t} \log p(x_t \mid g) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(g \mid x_t). \tag{1}$$

According to Bayes' theorem, $p(x_t \mid g)$ can be expressed as:

$$p(x_t \mid g) = \frac{p(g \mid x_t) \times p(x_t)}{p(g)}. \tag{2}$$

Taking the logarithm on both sides of the equation, we obtain:

$$\log p(x_t \mid g) = \log p(g \mid x_t) + \log p(x_t) - \log p(g). \tag{3}$$

Differentiating both sides of the equation with respect to x_t, we can ignore $p(g)$ as it is independent of x_t. Thus, the equation can be expressed as:

$$\nabla_{x_t} \log p(x_t \mid g) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(g \mid x_t). \tag{4}$$

Thus, we have completed the derivation of Equation (7) in our main text. Next, we will continue with the detailed derivation of Equation (9) in Section 4.1, which can be represented as:

$$\begin{aligned} \epsilon'_\phi &= \epsilon_\phi(x_t, t) + \rho\sqrt{1-\alpha_t}\nabla_{x_t} \left\| g - \mathcal{M}(\hat{x}_0(x_t)) \right\|_2^2 \\ &= \epsilon_\phi(x_t, t) + \rho\sqrt{1-\alpha_t}\nabla\mathcal{L}_g, \end{aligned} \tag{5}$$

According to Equations (6) in our main text, we can deduce:

$$\begin{aligned} \epsilon_\phi(x_t, t) &= -\sqrt{1-\alpha_t}\nabla_{x_t} \log p(x_t), \\ \epsilon'_\phi &= -\sqrt{1-\alpha_t}\nabla_{x_t} \log p(x_t \mid g). \end{aligned} \tag{6}$$

Therefore, we can easily obtain:

$$\epsilon'_\phi = \epsilon_\phi(x_t, t) - \sqrt{1-\alpha_t}\nabla_{x_t} \log p(g \mid x_t). \tag{7}$$

It can be inferred from the main text that $\nabla_{x_t} \log p(g \mid x_t) \simeq -\rho\nabla_{x_t} \left\| g - \mathcal{M}(\hat{x}_0(x_t)) \right\|_2^2$. Thus, we can derive that:

$$\begin{aligned} \epsilon'_\phi &= \epsilon_\phi(x_t, t) + \rho\sqrt{1-\alpha_t}\nabla_{x_t} \left\| g - \mathcal{M}(\hat{x}_0(x_t)) \right\|_2^2 \\ &= \epsilon_\phi(x_t, t) + \rho\sqrt{1-\alpha_t}\nabla\mathcal{L}_g, \end{aligned} \tag{8}$$

We have also provided the pseudocode for our training phase, which can be found in Algorithm 1.

Algorithm 1 Pseudo Code for Our Guidance in the training stage

Require: The low-resolution image x_t, the high-resolution image x_0, the forward operator $\mathcal{M}$, no-guidance training iterations λ_t, total training iterations λ_T

if *training iterations* $< \lambda_t$ **then**
  Predict the noise $\epsilon_\phi(x_t, t)$ through the denoising model
  Take gradient descent step on:

$$\nabla_\phi \left\| \epsilon - \epsilon_\phi \left(x_t, t \right) \right\|^2 \tag{9}$$

  return
end if
if $\lambda_t \leq$ *training iterations* $< \lambda_T$ **then**
  Input the high-resolution image x_0 into the forward operator $\mathcal{M}$ to get guidance g

$$g = \mathcal{M}(x_0) \tag{10}$$

  Predict the noise $\epsilon_\phi(x_t, t)$ through the denoising model
  Optimize the predicted noise $\epsilon_\phi(x_t, t)$ using guidance g

$$\begin{aligned} \epsilon'_\phi &= \epsilon_\phi(x_t, t) + \rho\sqrt{1-\alpha_t}\nabla_{x_t} \left\| g - \mathcal{M}(\hat{x}_0(x_t)) \right\|_2^2 \\ &= \epsilon_\phi(x_t, t) + \rho\sqrt{1-\alpha_t}\nabla\mathcal{L}_g, \end{aligned} \tag{11}$$

  Take gradient descent step on:

$$\nabla_\phi \left\| \epsilon - \epsilon'_\phi \right\|^2 \tag{12}$$

  return
end if

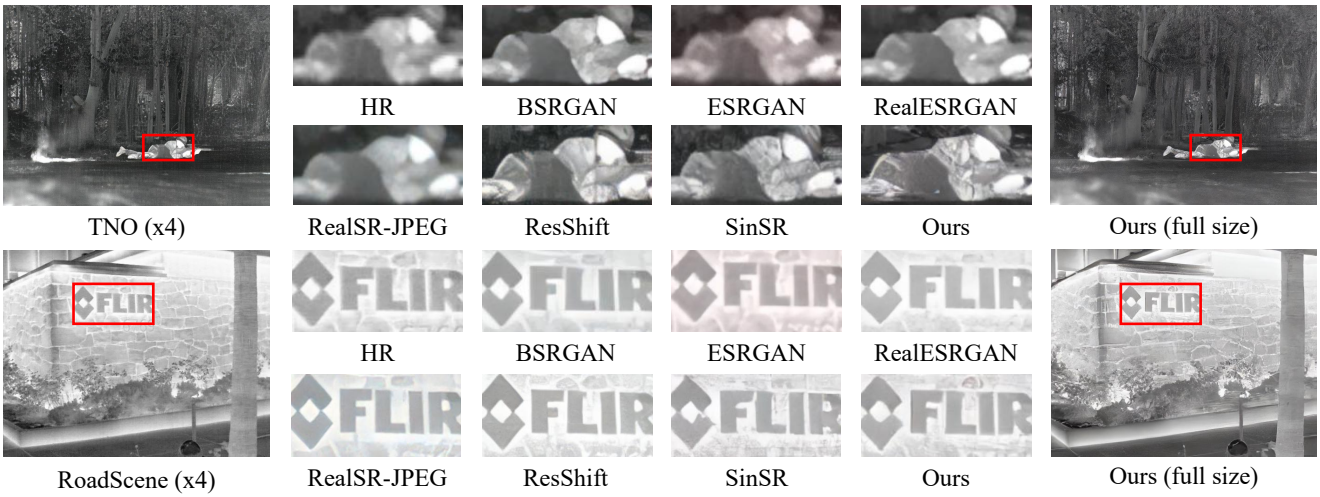

Figure 1. Visual comparison of infrared image super-resolution with SOTA methods on TNO and RoadScene datasets.

B. Additional Comparison for Our Proposed Approach

Qualitative Comparison. The qualitative results shown in Figure 1 highlight the superior visual performance of our method compared to other approaches. Additional examples are provided in the supplementary materials. For a comprehensive evaluation, we selected representative images from the TNO and RoadScene datasets for qualitative analysis. Our method preserves more accurate thermal details and clear object contours on the TNO dataset, avoiding artifacts and noise commonly observed in other methods. For the RoadScene dataset, our approach excels in reconstructing fine textures and maintaining structural consistency, particularly in complex traffic scenarios where other methods struggle with blurring and misalignment.

Datasets		TNO		RoadScene	
Methods		CLIP-IQA↑	MUSIQ↑	CLIP-IQA↑	MUSIQ↑
Low Resolution	-	0.2020	22.507	0.1487	22.349
ESRGAN [7]	ECCV'18	0.2625	33.033	0.2607	44.470
RealSR-JPEG [4]	CVPR'20	0.4271	46.562	0.4270	52.860
BSRGAN [11]	CVPR'21	0.4006	53.305	0.3139	53.491
SwinIR [6]	CVPR'21	0.2656	32.910	0.1890	34.677
RealESRGAN [8]	ICCV'21	0.3828	48.3307	0.2933	49.853
HAT [1]	CVPR'23	0.2791	33.961	0.2247	35.413
DAT [2]	ICCV'23	0.2854	34.130	0.2351	34.655
ResShift [10]	NeurIPS'23	0.4745	50.546	0.3607	51.528
CoRPLE [5]	ECCV'24	0.2601	30.516	0.2018	27.701
SinSR [9]	CVPR'24	<u>0.6159</u>	<u>54.027</u>	<u>0.5207</u>	<u>54.184</u>
Bi-DiffSR [3]	NeurIPS'24	0.3106	34.883	0.2519	38.543
DifIISR	Ours	**0.6218**	**54.601**	**0.5302**	**54.568**
High Resolution	-	0.2018	30.074	0.1604	40.864

Table 1. No-reference Metrics Comparison of infrared image super-resolution with SOTA methods on TNO and RoadScene datasets.

These results demonstrate the distinct advantages of our approach in producing high-quality visual outputs across diverse datasets.

Quantitative Comparison. Table 1 provides a quantitative analysis of our method on the TNO and RoadScene datasets compared to various approaches. Our method consistently achieves the best performance across all metrics on both datasets, demonstrating its effectiveness in evaluating and enhancing image quality. On the TNO dataset, our approach excels in capturing critical features under challenging scenarios, aligning well with human perceptual judgments. Similarly, on the RoadScene dataset, our method outperforms others across all evaluation metrics, showcasing its robustness and superior capability in handling diverse and complex scenes.

C. More Ablation for Our Dual Guidance

C.1. Ablation for different visual guidance

In the choices for visual guidance, we utilized MSE and SSIM as replacements. The following is the description of these loss functions. SSIM (Structural Similarity Index) loss is a visual loss function that measures the similarity between two images based on their contrast, and structure. The SSIM loss is defined as:

$$\mathcal{L}_{SSIM} = 1 - SSIM(x, y)$$

where $SSIM(x, y)$ measures the structural similarity between the images x and y. MSE (Mean Squared Error) loss is a common loss function that measures the average squared difference between predicted and target values. It is defined as:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

where x_i and y_i are the predicted and target values, respectively, and N is the total number of pixels. Table 2 presents our ablation results, demonstrating that our method outperforms other visual guidance approaches across all metrics.

C.2. Ablation for different perceptual guidance

LPIPS and Edge were adopted as substitutes for the replacement options of perceptual guidance. LPIPS (Learned Perceptual Image Patch Similarity) is a perceptual loss function that measures the perceptual similarity between two images using deep network features. It is defined as:

$$\mathcal{L}_{LPIPS} = \sum_{l} w_l \|\phi_l(x) - \phi_l(y)\|^2$$

Visual Guidance	PSNR	CLIP-IQA	mAP	mIoU
-	33.466	0.5102	31.2	40.9
SSIM Based	34.010	0.5189	31.8	41.5
MSE Based	34.244	0.5278	31.7	41.3
Ours	**34.575**	**0.5379**	**33.1**	**42.4**

Table 2. Ablation study for different visual guidance.

Visual Guidance	PSNR	CLIP-IQA	mAP	mIoU
-	33.466	0.5102	31.2	40.9
LPIPS Based	33.954	0.5189	32.7	41.6
EDGE Based	33.823	0.5217	32.5	42.0
Ours	**34.575**	**0.5379**	**33.1**	**42.4**

Table 3. Ablation study for different perceptual guidance.

where ϕ_l represents features from the l-th layer, and w_l are learned weights. Edge loss focuses on preserving edge details by comparing the edges of the generated image with the target image. Typically, edges are extracted using edge detection methods like Sobel or Canny filters, and the loss is defined as:

$$\mathcal{L}_{EDGE} = \|E(x) - E(y)\|^2$$

where $E(\cdot)$ represents the edge detection operator. The ablation results, shown in Table 3, indicate that our method performs better than other perceptual guidance strategies across all metrics.

References

[1] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 3

[2] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12312–12321, 2023. 3

[3] Zheng Chen, Haotong Qin, Yong Guo, Xiongfei Su, Xin Yuan, Linghe Kong, and Yulun Zhang. Binarized diffusion model for image super-resolution. *arXiv preprint arXiv:2406.05723*, 2024. 3

[4] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 466–467, 2020. 3

[5] Xingyuan Li, Jinyuan Liu, Zhixin Chen, Yang Zou, Long Ma, Xin Fan, and Risheng Liu. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 270–288, 2024. 3

[6] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 3

[7] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 0–0, 2018. 3

[8] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 3

[9] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024. 3

[10] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024. 3

[11] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 3