Self-supervised Diffusion-guided Hallucination-free Thermal Infrared Image Denoising *[Supplementary Material]*

Félix Hazebrouck[1,2], Alexander Schock-Schmidtke[1,3], Norbert Stuhrmann[2], Johannes Fottner[1], Michael Teutsch[2]

[1] Technical University of Munich (TUM), Germany
[2] HENSOLDT, Germany [3] digital workbench, Germany
michael.teutsch@hensoldt.net alexander.schock@digital-workbench.de

Abstract

This supplementary material provides additional details and results supporting the main paper, ***Self-supervised Diffusion-guided Hallucination-free Thermal Infrared Image Denoising****. We include details for enabling appropriate use of the results of this work, implementation specifics, extended experimental evaluations, and further qualitative comparisons to complement the findings presented in the main manuscript. These materials aim to improve transparency and reproducibility, and to provide deeper insights into the proposed diffusion-guided self-supervised denoising approach.*

1. Inference Parameters for Diffusion Models

As described in Secion 3 of the main paper, diffusion models from the Visual-optical (VIS) are applied to real thermal infrared (TIR) images to produce synthetic TIR images of higher perceptual quality than any publically available TIR dataset. Several diffusion models from VIS are tested on the HDRT-TIR dataset introduced by [8], and for each we selected the inference parameter assortment producing the best perceptual outputs. These assortments are reported in Table 1.

The main drawback of diffusion models are the inherent hallucinations in the enhanced output images with respect to the input, which makes them unsuitable for direct use in content-critical image enhancement applications, as we seek for in this work. Attempts to minimize the hallucinations by varying the inference parameters, for example strengthening control or minimizing the number of diffusion step, resulted in the parameter assortment reported in Table 2. These inference parameters qualitatively reduce the hallucinations in the output at maximum, while still preserving satisfactory perceptual quality. Some qualitative samples with these parameter arrangements are displayed in Figure 1 and still exhibit major hallucinations.

Model	Architecture	Parameters
DiffBIR	v2	`strength = 1, cfg_scale = 6.0, steps = 50, pos_prompt = '', neg_prompt = 'low quality, blurry, low-resolution, noisy, unsharp, weird textures'`
PASD	pasd (base)	`guidance_scale = 9.0, conditioning_scale = 1.0, num_inference_steps = 20, pos_prompt = ''`
StableSR	"Turbo" ; 512 generative backbone	`dec_w = 0.5, ddpm_steps = 4`

Table 1. Architecture-parameter assortment producing the qualitatively best results on each dataset.

The results of this comparative study on inference parameters of different diffusion-based image-enhancement methods show the impossibility to achieve fidelity with diffusion based methods, therefore justifying the choice of using the output as clean target images, unrelated to the original input due to the hallucinations.

2. Generalizability Study for Diffusion-Enhancing TIR Datasets

The approach described in Section 3 of the main paper for the HDRT dataset would be very powerful if generalizing to other datasets. If any real TIR dataset could be 1)en-

Model	Architecture	Parameters
DiffBIR	v2	`strength = 2, cfg_scale = 8.0, steps = 50, pos_prompt = '', neg_prompt = 'low quality, blurry, low-resolution, noisy, unsharp, weird textures'`
PASD	pasd (base)	`guidance_scale = 7.0, conditioning_scale = 1.2, num_inference_steps = 20, pos_prompt = ''`
StableSR	"beyond 512" ; 512 generative backbone	`dec_w = 0.75, ddpm_steps = 200`

Table 2. Architecture-parameter assortment producing the most faithful, yet visually satisfactory results on the HDRT-TIR dataset

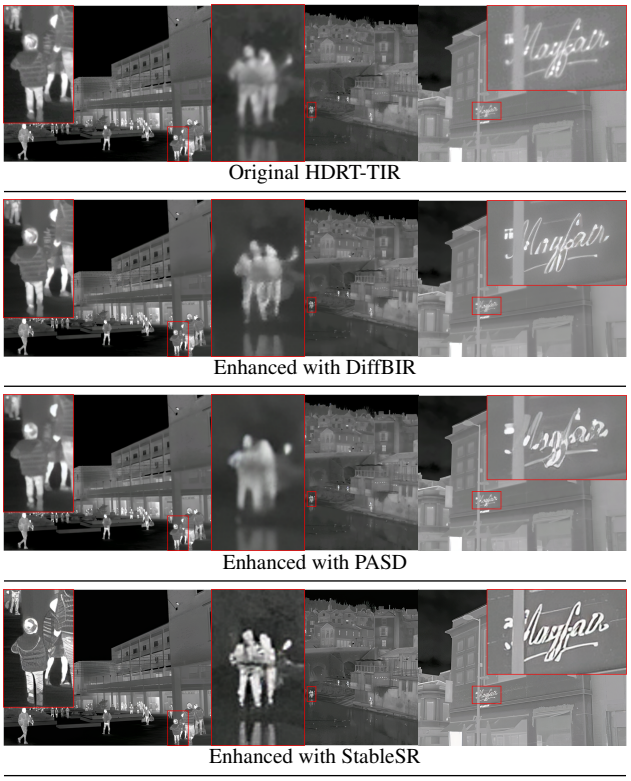

Figure 1. Images from original HDRT-TIR dataset and from its enhanced versions obtained with DiffBIR, PASD, and StableSR, with inference parameters minimizing hallucinations with respect to original image (Table 2). **(Zoom in for details)**

hanced with a fixed procedure to create clean reference targets and 2)these clean references degraded using a task-specific degradation model, then the described approach would be a dataset-agnostic self-supervised training framework like Noise2Noise (N2N) or Noisy-As-Clean (NAC).

In order to evaluate if the whole approach can be transferred to other datasets, we study the effects of the diffusion based image enhancement method *StableSR* [9] on the real TIR datasets FLIR [3], M3FD [6] and MassMIND [7]. We choose StableSR as the resulting increase in perceptual quality is most visible and because as a relatively basic diffusion model, the results should be representative for other diffusion models too. Additionally, we try different preprocessing before inputting the images into StableSR to better match the training resolution of StableSR. This is achieved by doubling the image resolution with respectively bilinear interpolation and the EDSR [5] super-resolution method used by [8] for producing parts of the HDRT dataset.

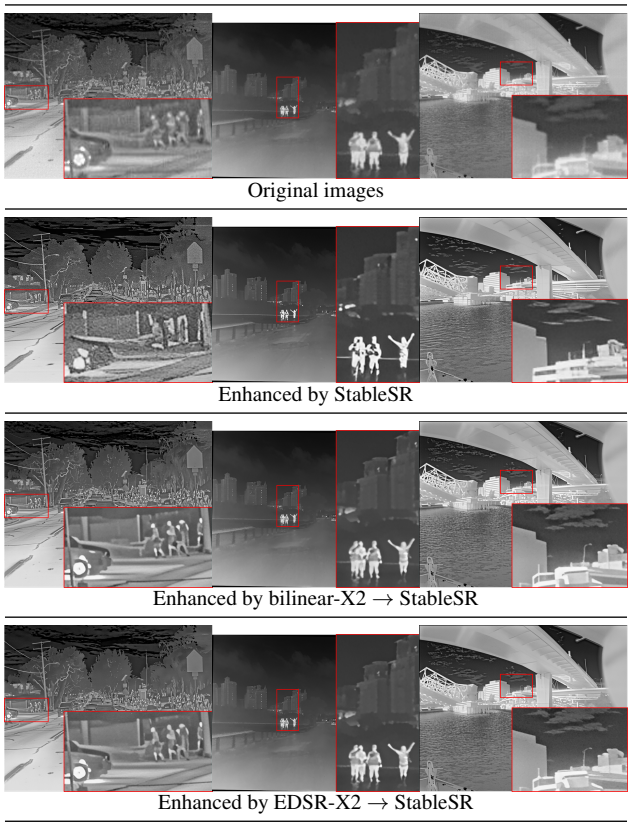

Figure 2. Attempts to apply StableSR (in maximizing visual quality setup) on other real TIR datasets, with and without prior upsampling show diverse output quality. Images are taken from the FLIR (left), M3FD (center) and MassMIND (right) dataset.

The samples displayed in Figure 2 show that StableSR produces strong artifacts on the FLIR dataset [3] while generating visually appealing results on the MassMIND dataset [7] and satisfactory results on M3FD [6]. Upscaling the images prior to processing with StableSR partially mitigates artifacts on the FLIR images, particularly when upsampled with the EDSR [5] method[1] instead of a naïve

[1] According to [8], the EDSR super-resolution method was used to create the high-resolution TIR images of the HDRT dataset.

bilinear upsampling. However, this preprocessing reduces contrast, blurs edges, and limits the visibility of fine details in the MassMIND and M3FD images.

Consequently, the dataset enhancement method presented here cannot be generalized from the HDRT-TIR to other real TIR datasets at this time and, hence, cannot be used as a general self-supervised training framework such as NAC [10] or N2N [4]. We have to content ourselves with one clean dataset, the *HDRT-TIR-DE*, which is still a strong contribution to the TIR image restoration field.

3. Statistics of the HDRT-TIR-DE and other TIR Datasets

This section provides supplementary details on the proposed *HDRT-TIR-DE* dataset. Most dataset characteristics are not altered by the enhancement with StableSR [9] and are directly taken from the description of the original dataset, the TIR part of the HDRT [8] dataset. The HDRT-TIR-DE proposed in this work is publically available and can be accessed through the project repository.

The images from the HDRT-TIR dataset from [8] were captured with an *Optris PI 640i* camera with 640 $\times$ 480 pixel resolution and working wavelengths from 8 to 14 µm in the Long-Wave Infrared (LWIR) spectrum. This corresponds to a captured temperature range from -20°C to 100°C. [8] internally employed the EDSR super-resolution method introduced by [5] to upscale the resolution to 1280 $\times$ 960. The released dataset from [8] comprises both resolutions. The images are available with a depth of 8 bit. Unfortunately, [8] does not explicitly tell how the tone-mapping from the raw 14-bit image returned by the sensor is achieved. The paper uses the Durand tone mapping operator [2], but as it is designed for the VIS spectrum and as the paper tasks on both VIS and TIR images, we cannot tell which tone-mapping operator (TMO) is used. Furthermore, the Optris software allows different settings for tone-mapping, and the raw data of the HDRT-TIR images is not published, making a reverse-engineering approach for determining the TMO impossible. Anyway, the TMO used internally by [8] is irrelevant for our use case, as the use of the images as-is produces the desired High Quality (HQ) TIR targets. Therefore, the matter was not further investigated in this work.

Regarding the image content, Table 3 shows a more detailed overview on the content of the HDRT dataset, which transfers directly to the HDRT-TIR-DE dataset. As pointed out in Section 4 of the main paper, the HDRT dataset was captured across three distinct seasons over a period of six months and in eight different cities located at various latitudes, ensuring a wide range of environmental conditions and thermal contexts.

Table 4 shows statistics on the TIR datasets used in this work, including the HDRT-TIR and TIR datasets used in Section 5 of the main paper for the generalization study of the proposed training pipeline basing on the HDRT-TIR-DE on other datasets against dataset-specific training frameworks.

4. TIR specific sensor noise model

The TIR sensor-noise model developed in this work relies on the well-described and well-documented model from [1]. The code for their original noise model can be found in their GitHub-Project: `https://github.com/cailijing/MDIVDnet`. The computationally optimized version, adapted for single image sensor-noise degradation can be found on our project repository, given in the abstract. A validation of the noise model with respect to real-world TIR noise con be found in [1].

According to [1], the sensor noise in TIR imaging is mainly composed of the following components:

- *Gaussian noise*: caused by *photon noise* (randomness of photons arriving at the detector), thermal movement of electrons in the electrical components of the sensor, *dark current noise* (random variations in the detector-output-current) and *readout noise* (random variations of sensor-voltage before and after analog amplification), the last two being due to the "immaturity of manufacturing process" of the infrared focal plane array (IRFPA) [1]
- *Spacial strip noise*: caused by inconsistency in the amplification rate of line and column amplifiers and by thermal interference in the analog-to-digital converter (ADC) ([1])
- *Hill noise*: caused by a temperature gradient in the sensor, makes the hotter parts of the IRFPA produce brighter pixel-values

The physical origin of these noises is not the focus of this work and a comprehensive description with related literature can be found in [1].

The sensor-noise model used in this work is the following:

$$\begin{aligned}
&\text{Gaussian noise:} && N_g \sim \mathcal{N}(0, \sigma_g) \\
&\text{Spatial stripe noise row:} && N_r \sim \mathcal{N}(0, \sigma_r) \\
&\text{Spatial stripe noise column:} && N_c \sim \mathcal{N}(0, \sigma_c) \\
&\text{Additional row or column strip noise:} && N_{l,t} \sim \mathcal{N}(0, \sigma_{l,t}) \\
&\text{Hill noise:} && H(x, y, \eta_h) = \eta_h \ln\left(\frac{r_2(x,y)}{r_1(x,y)}\right) - \frac{\eta_h}{2}
\end{aligned} \tag{1}$$

with

[2] According to [8], the HDRT dataset was initially recorded with a resolution of 640 $\times$ 480 pixels and upscaled with the Enhanced Deep residual networks image Super-Resolution (EDSR) method [5]. For this work, however, we consider the HDRT-TIR dataset as-is.

Features	Daytime	Nighttime	Buildings	Churches	Skyscrapers	Castles	Bridges	Pedestrians	Boats
Number	7571	2429	8660	583	1289	222	452	641	218
Features	Cars	Water Bodies	Urban	Rural	Natural Scenery	Trees	Parks	Sky	Clouds
Number	689	2008	4711	2191	3098	4871	937	1646	4669

Table 3. HDRT dataset statistics (Numbers and caption from [8])

Dataset name	No. of Images Resolution	Image Content
HDRT-TIR [8]	10,000 1280×960^2	Automotive scenes, urban area, countryside, coastal area
FLIR V2.0 [3]	15,635 640×512	Automotive scenes, urban environment
M3FD [6]	4200 1024×768	Automotive scenes, urban area, countryside
MassMIND [7]	2916 640×512	Coastal area

Table 4. Main characteristics of the used TIR image datasets.

- N_g the random variable describing the noise to add independently to each pixel value, with standard deviation σ_g
- N_r and N_c the random variables describing the noise to add independently to each row and column respectively, with standard deviation σ_r and σ_c respectively
- $N_{l,t}$ the random variable describing the noise to add independently to, either each row, or each column with a probability 0.5 for each image (both exist, the direction depends on the sensor circuit design), with standard deviation $\sigma_{l,t}$
- $H(x, y, \eta_h)$ the deterministic noise to add to the pixel in position (x, y) in the image, with scale parameter η_h representing "the intensity of the simulation noise"

The parameter value-ranges from [1] for variable image degradation during training are:

$$\sigma_g \in [5, 35] \quad ; \quad (\sigma_r, \sigma_c) \in [0, 15]^2 \quad ; \quad \sigma_{l,t} \in [5, 35] \quad \text{and} \quad \eta_h \in [0, 35] \quad (2)$$

The synthetically degraded images can be "generated by uniformly sampling the parameters within their predefined ranges" [1]. By setting the uniform-sampling range for the parameters large enough, the noise model can account for the wide range of different sensor-noise strengths and composition-weightings represented in existing TIR cameras.

References

[1] Lijing Cai, Xiangyu Dong, Kailai Zhou, and Xun Cao. Exploring video denoising in thermal infrared imaging: Physics-inspired noise generator, dataset, and model. *IEEE Transactions on Image Processing*, 33:3839–3854, 2024. 3, 4

[2] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *ACM SIGGRAPH*, 2002. 3

[3] Teledyne FLIR. FLIR ADAS Thermal Dataset V2.0.0, 2022. 2, 4

[4] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning Image Restoration without Clean Data. In *ICML*, 2018. 3

[5] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE CVPR Workshops*, 2017. 2, 3

[6] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware Dual Adversarial Learning and a Multi-scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection. In *IEEE CVPR*, 2022. 2, 4

[7] Shailesh Nirgudkar, Michael DeFilippo, Michael Sacarny, Michael Benjamin, and Paul Robinette. MassMIND: Massachusetts Maritime INfrared Dataset. *The International Journal of Robotics Research*, 42(1-2):21–32, 2023. 2, 4

[8] Jingchao Peng, Thomas Bashford-Rogers, Francesco Banterle, Haitao Zhao, and Kurt Debattista. HDRT: A large-scale dataset for infrared-guided HDR imaging. *Elsevier Information Fusion*, 120, 2025. 1, 2, 3, 4

[9] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *IJCV*, 132(12):5929–5949, 2024. 2, 3

[10] Jun Xu, Yuan Huang, Ming-Ming Cheng, Li Liu, Fan Zhu, Zhou Xu, and Ling Shao. Noisy-as-clean: Learning self-supervised denoising from corrupted image. *IEEE Transactions on Image Processing*, 29:9316–9329, 2020. 3