Exploring Diffusion-generated Guidance for Thermal Image Super-resolution

Leo Thomas Ramos[1,2] Angel D. Sappa[1,2,3]
[1]Computer Vision Center [2]Universitat Autònoma de Barcelona [3]ESPOL Polytechnic University
ltramos@cvc.uab.cat; sappa@ieee.org

Abstract

This work presents a strategy for guided thermal super-resolution that replaces the conventional RGB guidance with a pseudo-thermal image. The approach aims to enhance spectral alignment between the guidance and target images, improving the transfer of structural and thermal information during reconstruction. The pseudo-thermal guidance is generated using a diffusion model trained for thermal image synthesis and integrated into several state-of-the-art guided super-resolution methods. Experiments conducted on the M3FD and CIDIS datasets at $\times 8$ and $\times 16$ scaling factors demonstrate consistent performance gains over RGB-guided counterparts, including PSNR gains of up to +1.64 dB, SSIM improvements of +0.0290, and LPIPS reductions of 0.11 points. Visual results further confirm that the proposed strategy yields sharper details, cleaner contours, and perceptually superior reconstructions.

1. Introduction

Super-resolution (SR) is the process of transforming a low-resolution image into a high-resolution version [35], recovering fine details absent from the original input [1]. Current SR research is dominated by deep learning approaches [1], which achieve strong performance in reconstructing high-frequency information and producing visually coherent outputs [17]. These advances have broadened the range of applications, including remote sensing [12, 34], medical imaging [24], object detection [34], and surveillance [28, 35].

Guided super-resolution (GSR) is a variant of SR that incorporates an additional image to support the reconstruction process [31]. Rather than relying solely on a low-resolution input, the model uses information from a high-resolution reference image to enhance detail recovery [38]. This strategy is particularly effective when the guidance and target images share structural or semantic similarities [32, 38], enabling more accurate reconstruction of fine textures and edges. It also facilitates the handling of large scaling factors, such as $\times 8$, $\times 16$, or $\times 32$ [27], which are typically difficult for single-image methods.

GSR has been applied across different image modalities, including RGB images and depth maps [38]. However, GSR of thermal images has gained particular relevance in recent years [27]. This growing interest arises because thermal images are typically captured at low-resolution, unless high-end and costly sensors are used [27, 31]. This limitation affects their applicability in surveillance and related real-world settings [9], even though thermal imaging offers unique advantages by capturing the heat radiation emitted by objects [15, 25], and allows effective perception in darkness and other environments with poor visibility [15].

In thermal GSR, RGB imagery is commonly used as the high-resolution guidance [7, 23]. However, this strategy introduces challenges that can degrade reconstruction quality [31]. Most existing methods assume perfect pixel-level alignment between the low-resolution thermal input and the RGB guide [9], even though such alignment is difficult to achieve in practice [3, 9]. Thermal visible pairs often present substantial spectral and structural differences [9, 22], which can propagate errors during reconstruction and reduce the effectiveness of the guidance.

Based on the above, this work proposes a thermal GSR strategy in which the conventional RGB guidance is replaced by a synthetic thermal image. We argue that a pseudo-thermal guide provides stronger semantic correlation with the target thermal image and reduces domain mismatch compared to RGB, thereby increasing the relevance of the guidance features, particularly at high scaling factors. The synthetic guide is generated using a Diffusion Model (DM), given its demonstrated capacity for high fidelity image-to-image generation. Experiments on the M3FD and CIDIS datasets show that the proposed approach consistently outperforms standard RGB-based GSR. The contributions of this work are as follows:

- We introduce a thermal GSR strategy in which the conventional RGB guidance is replaced by a pseudo-thermal image generated using a diffusion model.
- We demonstrate that the proposed approach improves thermal GSR performance on the M3FD and CIDIS datasets at both $\times 8$ and $\times 16$ scaling factors compared to using an RGB guide.

- We show that generating the pseudo-thermal guidance with a diffusion model yields superior results compared to alternative approaches such as GANs and Pix2Pix.
- We demonstrate that the strategy is model-agnostic, successfully integrating with diverse SR architectures.

2. Related work

In thermal GSR, different approaches have been proposed. Early work such as [20] reframes the task as a pixel-to-pixel transformation from the RGB guide to the thermal domain, enforcing consistency with the low-resolution thermal input. The mapping is modelled with a multilayer perceptron that uses the guide's pixel values and spatial coordinates without spatial mixing to preserve sharpness. To address the misalignment between RGB and thermal images, [9] introduces a Siamese dense block network with spatial attention, offering two variants: one with a correlation based feature alignment loss, and another that estimates a misalignment map to warp the RGB guide before fusion.

Subsequent works adopt more sophisticated fusion architectures. CoReFusion [16] employs a dual-encoder U-Net with ResNet-34 backbones, merging RGB and thermal features via element-wise maximum operations and introducing contrastive regularization to enhance feature discrimination. MGNet [36] advances this idea by extracting complementary appearance, edge, and semantic cues from the RGB guide through dedicated networks, embedding them into the thermal representation via self- and cross-attention in a Multicue Guidance Module, and progressively fusing them to recover fine textures and structures.

Transformer-based designs have also been explored. SwinFuSR [2] adopts a dual-branch Swin Transformer model that merges RGB and thermal information through Attention-guided Cross-domain Fusion blocks, and improves robustness by randomly dropping RGB inputs during training. SwinPaste [37] extends this idea with a pre-training data mixing strategy for limited thermal data and multi-scale supervision to enhance detail recovery. Along similar lines, FW-SAT [13] combines global channel-spatial attention, localized windowed self-attention, and a regional aggregation module to capture multi-scale context.

Further advances include the MSFFCT model [23], which concatenates upsampled thermal and RGB images and uses multi-scale feature extraction with deformable convolutions to handle modality misalignment, fusing information through residual units with channel attention, transposed convolutions, and a channel-wise transformer. In the generative model category, DMs have been adapted to GSR. An example is in [5], where a ResShift-based formulation conditions the denoising process on RGB inputs, accelerates inference, and refines outputs with a lightweight U-Net.

Although the reviewed approaches differ in architectural design and fusion strategies, they largely rely on RGB imagery as the guiding modality, assuming it provides the most effective information. However, this assumption does not always hold. Methods such as [9] and [23] explicitly address misalignment and modality differences, acknowledging that RGB guidance can introduce artifacts or inconsistencies when geometric or photometric correspondence with the thermal domain is imperfect. Likewise, when the visible and thermal spectra capture fundamentally different scene characteristics, transferring high-frequency details from RGB to thermal may propagate irrelevant or misleading patterns.

Another observation is the growing interest in generative models within this domain, though their role remains mostly as components within SR pipelines rather than as stand-alone mechanisms for producing alternative guidance signals. Overall, these patterns indicate that progress in the field is shaped not only by architectural advances but also by a persistent dependence on RGB guidance, which may limit adaptability when RGB is suboptimal.

3. Methodology

3.1. Dataset description

M3FD The first dataset employed is the Multi-scenario Multi-modality Dataset (M3FD) [18], which provides pairs of spatially aligned visible and infrared images acquired with a binocular optical-infrared sensor. It includes diverse surveillance environments and objects, such as pedestrians, vehicles, buildings, buses, and roads. The original release contains 4,500 registered pairs with an average resolution of 1024×768 pixels. In this work, we adopt the cropped partitions from [29], where images are resized to 640×480 pixels. This split includes 826 pairs for training, 30 for validation, and 20 for testing.

CIDIS The second dataset used is the Cross-spectral Image Dataset for Image Super-resolution (CIDIS) [26]. It contains 1,000 pairs of visible-thermal images captured with a Balser camera and a TAU2 thermal sensor. Each pair has a resolution of 640×448 pixels, with 700 images for training, 200 for validation, and 100 for testing. The dataset covers a range of predominantly outdoor scenes, including vehicles, buildings, parks, pedestrians, and other elements typical of urban and semi-urban environments. CIDIS has also served as the benchmark for the Thermal Image Super-Resolution Challenge at CVPR 2024 and 2025.

3.2. Proposed strategy

3.2.1. Overview

The proposed strategy replaces the conventional RGB guidance used in thermal GSR with a synthetic thermal image to reduce semantic and spectral mismatch with the target. This enhances cross-modal correlation in feature space and helps

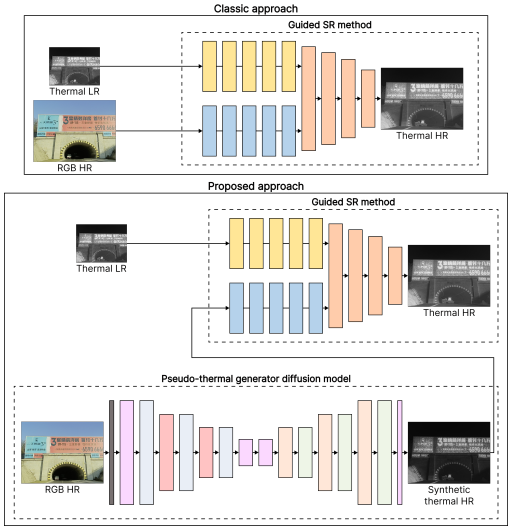


Figure 1. Overview of the thermal guided super-resolution strategy explored in this work.

the SR model recover finer details and sharper boundaries. As shown in Fig. 1, we extend the standard GSR framework by adding a pseudo thermal generation module that converts the available RGB image into a high-resolution synthetic thermal guide, which then replaces the RGB input in the SR pipeline. The design is model agnostic, allowing integration with multiple state-of-the-art (SOTA) GSR methods without modifying their original architectures.

3.2.2. Hypothesis and mathematical formulation

GSR aims to reconstruct a high-resolution target image I^T_{HR} from a low-resolution version I^T_{LR}, with the aid of a high-resolution guidance image I^G_{HR} from a different but related modality. This process can be expressed as in Eq. 1:

$$\hat{I}^T_{HR} = F_\theta(I^T_{LR}, I^G_{HR}), \tag{1}$$

where F_θ denotes a GSR model parameterized by θ, and $\hat{I}^T_{HR}$ is the reconstructed image.

A key factor in the effectiveness of GSR is the cross-modal correlation between I^G_{HR} and I^T_{HR}. In thermal GSR, I^G_{HR} usually corresponds to an RGB image, which often suffers from a domain gap with respect to the thermal modality. We define this gap as the distance between the representations of the two modalities in a shared feature space $\phi(\cdot)$, extracted by a suitable encoder as in Eq. 2:

$$\Delta_{G\to T} = \mathcal{D}(\phi(I^G_{HR}), \phi(I^T_{HR})), \tag{2}$$

where $\mathcal{D}(\cdot,\cdot)$ is a distance metric such as L_2 or cosine dissimilarity.

Our hypothesis is that replacing I^G_{HR} with a synthetic pseudo-thermal image I^{PT}_{HR}, generated to share the spectral characteristics of I^T_{HR}, reduces $\Delta_{G\to T}$, as in Eq. 3:

$$\Delta_{PT\to T} < \Delta_{RGB\to T}. \tag{3}$$

By narrowing this gap, the guidance features become more semantically aligned with the target, enabling F_θ to better transfer structural information and recover fine details, as in Eq. 4:

$$Q(F_\theta(I^T_{LR}, I^{PT}_{HR})) > Q(F_\theta(I^T_{LR}, I^{RGB}_{HR})), \tag{4}$$

where $Q(\cdot)$ denotes a quality metric such as PSNR or SSIM.

Our proposed strategy builds directly on this theoretical formulation, employing a DM to generate the pseudo-thermal images used as guidance. This approach is designed to enhance the relevance of the transferred features and unlock the potential for more accurate and perceptually faithful reconstructions.

3.2.3. Synthetic thermal generation

In this work, the generation of pseudo-thermal images is addressed using a diffusion-based model. DMs [10] have become the SOTA for image generation and image-to-image translation [19], offering strong semantic consistency and stability across diverse scenarios [33]. These advantages allow them to surpass Generative Adversarial Networks (GANs) [6, 8]. Integrating a DM into our strategy therefore enables the production of realistic and structurally coherent pseudo-thermals that provide effective guidance for subsequent SR.

Given that the goal of this study is to evaluate whether a pseudo-thermal image can serve as a more effective guide for GSR, and considering the wide availability of robust DMs trained on large-scale datasets, we adopt an existing model for generation. This choice isolates the variable of interest and avoids the bias that could arise from a newly designed, unvalidated generator. Using a well-established model provides a stable and reproducible component, keeping the analysis focused on the guidance itself. It also makes the experimental process more efficient, since training DM generators from scratch is computationally expensive [4], requires large amounts of data [19], and would narrow the scope of the study.

In this work, the Physics Informed Diffusion Model (PID) [21] is used for pseudo-thermal generation. PID is a conditional Latent Diffusion Model (LDM) that incorporates physics-based constraints into the denoising process to ensure that the generated images comply with the radiometric principles of infrared imaging. It follows the standard LDM formulation, where a clean image x_0 is encoded into a latent representation z_0, progressively corrupted with Gaussian noise over T steps, and then iteratively reconstructed from noise using a neural network parameterized by ϵ_θ. As shown in Fig. 2, PID incorporates three key components:

- **Latent diffusion backbone:** Based on U-Net, it takes as input the noisy latent vector z_t at timestep t, conditioned

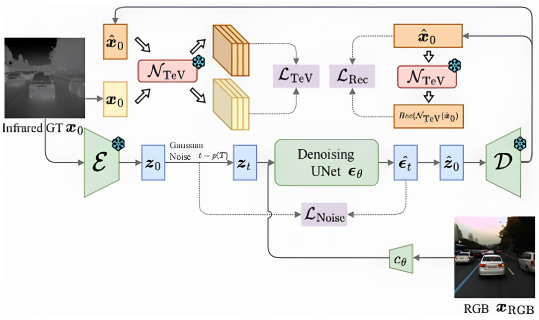


Figure 2. PID model used in this work for pseudo-thermal generation.

on a latent representation of the RGB guide. The network predicts the noise component ϵ_θ, which is used to approximate the previous denoising step z_{t-1}.

- **TeV decomposition network:** The TeV decomposition network (TeVNet) decomposes an image into temperature, emissivity, and thermal texture components. This is used to compute two physics-informed losses: the physical reconstruction loss L_{Rec}, which enforces pixel-space adherence to physical laws, and the TeV space loss L_{TeV}, which enforces consistency in the decomposed physical parameter space.
- **Multimodal conditioning:** The RGB input is encoded into a conditioning vector with the same spatial resolution as the latent infrared representation. This conditioning is concatenated with the noisy latent input, ensuring structural correspondence between the RGB and infrared domains and allowing the diffusion process to integrate visual patterns from the guidance modality while adhering to infrared physical characteristics.

PID training is organized into two sequential stages:

1. **TeVNet training**
   - TeVNet, is trained on infrared images to predict: temperature, emissivity, and thermal texture components.
   - The decomposition follows the TeV formulation derived from radiometric principles.
   - Once trained, TeVNet's parameters are frozen and reused in the following stage.
2. **LDM training**
   - The LDM is trained with RGB images as conditional input and infrared images, encoded by a VQGAN encoder, as the target output.
   - Gaussian noise is progressively added to the infrared latent vectors, and a U-Net predicts and removes this noise, conditioned on the RGB features.
   - Training is performed using the standard noise prediction loss, along with two physics-informed losses (L_{Rec} and L_{TeV}) computed with TeVNet.

During inference, the reverse diffusion process refines Gaussian noise in the latent space through the U-Net backbone conditioned on the encoded RGB image, relying solely on the knowledge learned during training to generate infrared outputs that remain physically consistent and structurally aligned with the guidance. In our setup, we use the TeVNet model pre-trained by its authors on the KAIST [11] dataset, which provides a strong prior from more than 95K paired samples, and train the LDM on M3FD to adapt it to the target domain. This approach yields a ready to use generator, enabling the production of pseudo-thermal images for each SR dataset, which are then used as guidance inputs for training the GSR methods.

3.3. Implementation details and testing

For the pseudo-thermal generation, we follow the training pipeline described in [30], which uses the M3FD dataset to train an image generator and then applies it for image synthesis on the same dataset and on others. Accordingly, we use the resulting generator to produce pseudo-thermal images for both M3FD and CIDIS. In addition to the selected DM, we also train Pix2Pix (P2P) and CycleGAN[1] to provide comparison baselines. All generators are trained with their default parameters, only setting the training epochs to 1000 in every case to ensure a fair comparison.

Then, we train SOTA GSR methods from scratch using the generated pseudo-thermals. Each model is trained independently on both datasets (M3FD and CIDIS) and at two scaling factors ($\times 8$ and $\times 16$), keeping all default parameters reported in the respective original works. For $\times 8$ we evaluate CoReFusion[16], SwinFuSR[2], and SwinPaste[37], and for $\times 16$ we employ FW-SAT[14], SwinFuSR, and SwinPaste.[2]

All experiments are conducted on an NVIDIA A100-SXM4 GPU with 40 GB of memory. The implementations are carried out in Python using the PyTorch framework, with each model sourced from its official repository. For evaluation, we employ a set of metrics that are widely used in image generation and enhancement tasks. For the evaluation of pseudo-thermal image generation, we use the Fréchet Inception Distance (FID) and the Structural Similarity Index Measure (SSIM). For the evaluation of guided GSR, we use the Peak Signal-to-Noise Ratio (PSNR), SSIM, and the Learned Perceptual Image Patch Similarity (LPIPS).

4. Experimental results and analysis

4.1. Pseudo-thermal generation

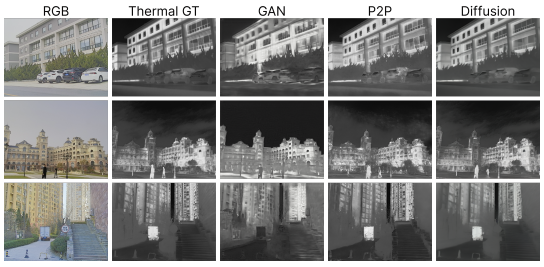
Table 1 summarizes the performance of the three generative models evaluated for pseudo-thermal synthesis on M3FD

[1]GAN and P2P implementations taken from: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.git

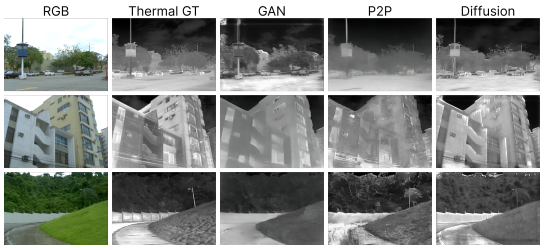
[2]CoReFusion provides only an $\times 8$ model, which is why FW-SAT is used for the $\times 16$ experiments.

Method	M3FD		CIDIS	
	FID ↓	SSIM ↑	FID ↓	SSIM ↑
GAN	57.60	0.5395	61.98	0.4428
P2P	42.36	0.6985	57.97	0.5849
Diffusion model	20.03	0.7427	27.86	0.6567

Table 1. Quantitative results of pseudo-thermal generation.



(a) Pseudo-thermal generation on the M3FD dataset.



(b) Pseudo-thermal generation on the CIDIS dataset.

Figure 3. Qualitative results of pseudo-thermal generation.

and CIDIS. The DM achieves the best results on both datasets, with substantially lower FID and higher SSIM than the GAN and P2P baselines. The gap is especially pronounced in FID, where PID remains below 30 points for both datasets, indicating pseudo-thermal images that are perceptually closer to real data and structurally more consistent. In contrast, the GAN model shows the weakest performance, with high FID and low SSIM, reflecting instability and limited ability to capture the physical and textural characteristics of thermal imagery. P2P achieves intermediate results, benefiting from its deterministic pixel-wise mapping but still lacking the semantic fidelity obtained through diffusion-based synthesis.

Fig. 3 presents qualitative examples of pseudo-thermal generation on M3FD and CIDIS. Consistent with the quantitative results, PID produces images visually closer to the ground-truth, preserving structural boundaries and relative intensity distributions. GAN-generated samples appear noisy and often fail to reproduce temperature gradients, while P2P outputs show overly sharp transitions and insufficient smoothness in homogeneous regions. Notably, P2P generalizes worse than GAN on CIDIS, opposite to what is observed on M3FD. Overall, these findings confirm that PID generates pseudo thermals that are both visually and statistically closer to real infrared data, providing a more reliable input for subsequent GSR.

4.2. Super-resolution

4.2.1. M3FD

Starting with M3FD at $\times 8$, replacing RGB with a diffusion-based pseudo-thermal guide yields consistent gains across all three SR methods, as shown in Table 2. PSNR increases by +0.19 for CoReFusion, +0.20 for SwinFuSR, and +0.38 for SwinPaste. SSIM also rises in every case, by +0.0057, +0.0056, and +0.0092 respectively. The perceptual metric shows the clearest effect, as LPIPS drops from 0.4398 to 0.3260 for CoReFusion (-0.1138), from 0.2443 to 0.2323 for SwinFuSR (-0.0120), and from 0.2443 to 0.2292 for SwinPaste (-0.0151). Since lower LPIPS indicates closer perceptual similarity, the diffusion guide improves both structure and perceived fidelity relative to RGB.

GAN guidance systematically underperforms the RGB baseline. For CoReFusion, SwinFuSR, and SwinPaste, PSNR drops by 0.33, 0.87, and 0.87, respectively, while SSIM decreases by 0.0092, 0.0224, and 0.0210; LPIPS also worsens in all cases. P2P-generated guidance is more stable, showing only minor deviations from RGB in PSNR and SSIM, yet it also fails to deliver consistent gains. These results indicate that not every synthetic modality is beneficial, since the guide must be both high-quality and thermally consistent. Diffusion satisfies both requirements, whereas GAN introduces artifacts and P2P lacks the thermal coherence needed to outperform RGB reliably.

Fig. 4 shows representative examples of GSR performance under different guidance modalities. Using the pseudo-thermal image generated by PID yields reconstructions that are visually sharper and more consistent with the thermal ground-truth. Fine structures such as building contours, window frames, and boundary transitions in shadowed areas are better preserved, with reduced texture degradation and over smoothing. In contrast, RGB and GAN guided outputs show stronger color texture inconsistencies and less defined geometric structures, particularly in high-frequency regions. P2P achieves intermediate quality, recovering some texture but lacking cross-scene stability.

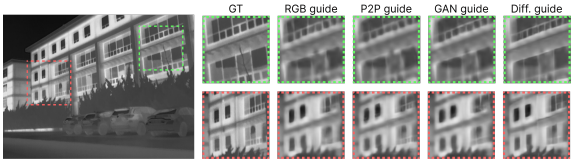
The quantitative comparison tables 4c and 4d in Fig. 4 reinforce these observations. For every SR model, the diffusion-guided results achieve the highest PSNR and SSIM, and the lowest LPIPS values, indicating better pixel fidelity, stronger structural similarity, and improved perceptual realism. These outcomes confirm the visual evidence that the diffusion-based pseudo-thermal guide contributes to clearer and more faithful thermal GSR.

At the higher $\times 16$ scaling factor, the trend observed at $\times 8$ remains consistent, reinforcing the robustness of diffusion guidance. Across all evaluated models, FW-SAT, SwinFuSR, and SwinPaste, the diffusion-guided versions

Method	RGB guide			Synt. GAN guide			Synt. P2P guide			Synt. Diffusion guide		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
CoReFusion[16]	26.56	0.7922	0.4398	26.23 (-0.33)	0.7830 (-0.0092)	0.4491 (+0.0093)	26.17 (-0.39)	0.7920 (-0.0002)	0.4233 (-0.0165)	26.75 (+0.19)	0.7979 (+0.0057)	0.3260 (-0.1138)
SwinFuSR[2]	27.36	0.8277	0.2443	26.49 (-0.87)	0.8053 (-0.0224)	0.2501 (+0.0058)	27.17 (-0.19)	0.8230 (-0.0047)	0.2495 (+0.0052)	27.56 (+0.20)	0.8333 (+0.0056)	0.2323 (-0.0120)
SwinPaste[37]	27.29	0.8258	0.2443	26.42 (-0.87)	0.8048 (-0.0210)	0.2563 (+0.0120)	27.24 (-0.05)	0.8255 (-0.0003)	0.2508 (+0.0065)	27.67 (+0.38)	0.8350 (+0.0092)	0.2292 (-0.0151)

Table 2. Super-resolution results with different guidance methods on the M3FD test set with a scale factor of $\times 8$. Values in parentheses represent the performance change relative to the RGB guide.

(a) Visual results using SwinFuSR model.



(b) Visual results using CoreFusion model.

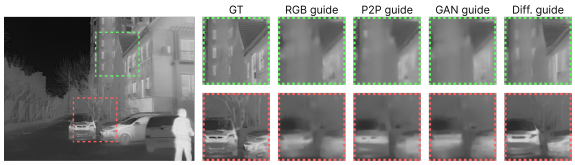
Guide	PSNR↑	SSIM↑	LPIPS↓
RGB	25.78	0.7819	0.3590
P2P	25.53	0.7721	0.3556
GAN	25.91	0.7839	0.3531
Diff.	27.41	0.8313	0.2270

(c) Quantitative results of Fig. 4a.

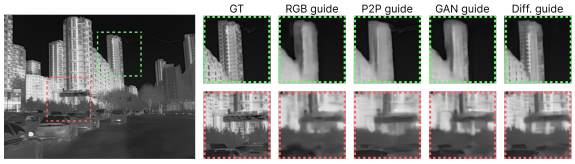
Guide	PSNR↑	SSIM↑	LPIPS↓
RGB	27.11	0.8562	0.3328
P2P	26.48	0.8410	0.3434
GAN	26.48	0.8539	0.3395
Diff.	26.79	0.8594	0.2079

(d) Quantitative results of Fig. 4b.

Figure 4. Guided super-resolution results at $\times 8$ scale on the M3FD dataset.



(a) Visual results using FW-SAT model.



(b) Visual results using SwinPaste model.

Guide	PSNR↑	SSIM↑	LPIPS↓
RGB	26.48	0.8041	0.3722
P2P	27.16	0.8249	0.3714
GAN	26.31	0.8020	0.4096
Diff.	27.66	0.8235	0.2477

(c) Quantitative results of Fig. 5a.

Guide	PSNR↑	SSIM↑	LPIPS↓
RGB	23.68	0.7218	0.4002
P2P	24.02	0.7188	0.3213
GAN	23.76	0.7259	0.3887
Diff.	24.96	0.7677	0.2972

(d) Quantitative results of Fig. 5b.

Figure 5. Guided super-resolution results at $\times 16$ scale on the M3FD dataset.

achieve the best results in every metric, as shown in Table 3. PSNR increases by +1.02, +0.67, and +0.51, while SSIM improves by +0.0290, +0.0223, and +0.0187. LPIPS shows the largest gains, with reductions of -0.1022, -0.0970, and -0.0870, indicating superior perceptual quality and sharper reconstruction. These consistent improvements demonstrate that diffusion-generated guidance remains effective even under more challenging upscaling conditions.

GAN-based guidance remains the weakest option, with the lowest PSNR and SSIM and the highest LPIPS, reflecting poor generalization at extreme upscaling. P2P offers moderate gains over RGB, mainly in SSIM, but still lags behind diffusion. These results show that, at higher scales, the guide's quality and thermal consistency become crucial. Diffusion-generated pseudo-thermals provide coherent structural cues that allow GSR models to recover finer textures and maintain stability even when the low-resolution input contains very limited information.

Fig. 5 shows representative visual results at $\times 16$, where the task becomes substantially harder due to the greater loss of spatial information. Even under these conditions, diffusion-guided outputs preserve sharper boundaries and finer textures in regions such as windows, building edges, and vehicles. By contrast, RGB, GAN, and P2P guidance produce noticeable blurring and structural deformation, especially in shadowed or high-frequency areas. Although overall metric values decrease at this scale, the diffusion guidance consistently provides the best reconstructions. More results in Supplementary Material.

4.2.2. CIDIS

Table 4 reports the $\times 8$ GSR results on CIDIS. Although this is a generalization setting in which the pseudo-thermals are generated by a model trained on M3FD rather than CIDIS, the diffusion-guided approach still outperforms the other modalities. Across CoReFusion, SwinFuSR, and SwinPaste, PSNR improves by +0.57, +1.64, and +1.02, with corresponding SSIM gains of +0.0037, +0.0064, and +0.0043. LPIPS also decreases notably for CoReFusion (-0.0804) and slightly for the other models, indicating that diffusion generated guidance maintains strong perceptual quality and structural coherence even under cross-dataset conditions.

Conversely, both GAN and P2P guidance suffer clear performance drops relative to RGB, with PSNR and SSIM decreasing across all architectures and LPIPS increasing substantially, particularly for P2P. This degradation reflects the limited generalization ability of these generative models in unseen domains. The diffusion pseudo-thermal guide, however, captures thermal-specific patterns more reliably,

Method	RGB guide			Synt. GAN guide			Synt. P2P guide			Synt. Diffusion guide		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
FW-SAT[14]	23.99	0.7538	0.3762	23.54 (-0.45)	0.7425 (-0.0113)	0.3977 (+0.0215)	24.18 (+0.19)	0.7648 (+0.0110)	0.3470 (-0.0292)	25.01 (+1.02)	0.7828 (+0.0290)	0.2740 (-0.1022)
SwinFuSR[2]	24.26	0.7578	0.3801	24.14 (-0.12)	0.7570 (-0.0008)	0.3947 (+0.0146)	24.60 (+0.34)	0.7697 (+0.0119)	0.3459 (-0.0342)	24.93 (+0.67)	0.7801 (+0.0223)	0.2831 (-0.0970)
SwinPaste[37]	24.34	0.7608	0.3878	24.24 (-0.10)	0.7605 (-0.0003)	0.3733 (-0.0145)	24.58 (+0.24)	0.7702 (+0.0094)	0.3443 (-0.0435)	24.85 (+0.51)	0.7795 (+0.0187)	0.3008 (-0.0870)

Table 3. Super-resolution results with different guidance methods on the M3FD test set with a scale factor of $\times 16$. Values in parentheses represent the performance change relative to the RGB guide.

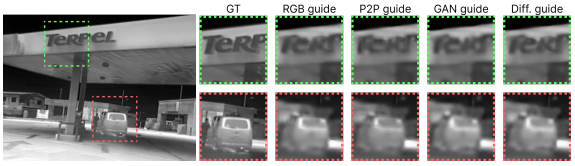
Method	RGB guide			Synt. GAN guide			Synt. P2P guide			Synt. Diffusion guide		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
CoreFusion[16]	26.55	0.7993	0.3907	26.39 (-0.16)	0.7954 (-0.0039)	0.4029 (+0.0122)	26.26 (-0.29)	0.7884 (-0.0109)	0.4278 (+0.0371)	27.12 (+0.57)	0.8030 (+0.0037)	0.3103 (-0.0804)
SwinFuSR[2]	28.06	0.8577	0.2098	27.83 (-0.23)	0.8518 (-0.0059)	0.2217 (+0.0119)	27.02 (-1.04)	0.8286 (-0.0291)	0.2649 (+0.0551)	29.70 (+1.64)	0.8641 (+0.0064)	0.2009 (-0.0089)
SwinPaste[37]	28.36	0.8579	0.2190	27.84 (-0.52)	0.8525 (-0.0054)	0.2243 (+0.0053)	27.06 (-1.30)	0.8303 (-0.0276)	0.2626 (+0.0436)	29.38 (+1.02)	0.8622 (+0.0043)	0.2067 (-0.0123)

Table 4. Super-resolution results with different guidance methods on the CIDIS test set with a scale factor of $\times 8$. Values in parentheses represent the performance change relative to the RGB guide.

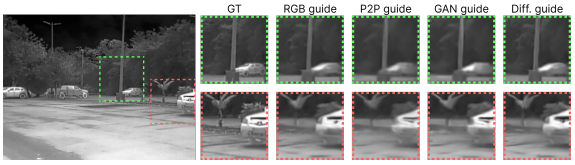
providing feature representations that transfer more effectively between datasets and enabling the GSR models to recover fine spatial details even without dataset specific tuning.

Fig. 6 shows qualitative results on CIDIS at $\times 8$, confirming the numerical trends. Diffusion guidance yields super-resolved thermal images with greater structural consistency and sharper detail, especially in edges, textures, and shadowed regions. Buildings, vehicles, and vegetation appear more clearly defined, and the reconstructed brightness distribution is closer to the ground-truth. In contrast, GAN and P2P guidance fail to generalize effectively, introducing artifacts and texture inconsistencies that blur fine structures or distort thermal contrast.

Although RGB guidance performs reasonably well, its limitations become more evident in areas where cross-spectral discrepancies are strong, leading to misalignment and inconsistent edge recovery. The diffusion-generated guide mitigates these issues by providing feature correlations more aligned with the target domain. This results in smoother temperature transitions and cleaner contours, reinforcing that the proposed synthetic thermal guidance enhances both perceptual quality and reconstruction accuracy, even when applied to unseen data during generator training.



(a) Visual results using CoreFusion model.



(b) Visual results using SwinPaste model.

Guide	PSNR↑	SSIM↑	LPIPS↓
RGB	26.44	0.8081	0.3508
P2P	26.06	0.7919	0.3515
GAN	26.11	0.7926	0.3521
Diff.	26.57	0.8094	0.3495

(c) Quantitative results of Fig. 6a.

Guide	PSNR↑	SSIM↑	LPIPS↓
RGB	26.84	0.8423	0.2176
P2P	26.19	0.8379	0.2262
GAN	26.37	0.8352	0.2216
Diff.	27.00	0.8461	0.2033

(d) Quantitative results of Fig. 6b.

Figure 6. Guided super-resolution results at $\times 8$ scale on the CIDIS dataset.

At the $\times 16$ scale, Table 5 shows the results. Differences among guidance types narrow at this magnification, which is expected given the increased reconstruction difficulty and the cross-dataset setting. The stronger scaling factor amplifies noise and information loss, making precise guidance more important. Even so, diffusion guidance retains a clear advantage, achieving PSNR gains of +0.27, +0.25, and +0.17 for FW-SAT, SwinFuSR, and SwinPaste, respectively. SSIM remains stable or slightly higher, and LPIPS decreases in all cases, confirming better perceptual fidelity.

The P2P and GAN guides, by contrast, show a more pronounced degradation as the scaling factor increases, with PSNR and SSIM dropping sharply. This suggests that their generated features are less transferable to unseen data and less informative for extreme upscaling. Even under these challenging conditions, the diffusion-generated pseudo-thermal guide provides a more coherent representation of thermal structure, which allows GSR models to retain detail and reduce perceptual distortion.

Next, Fig. 7 shows representative visual results for the $\times 16$ scale on the CIDIS dataset. As expected, none of the methods achieves outstanding reconstruction quality at this challenging magnification level, with all models showing some degree of blur and loss of fine texture compared to the $\times 8$ scale. This degradation is consistent with the increased complexity of the task, where the limited information available in the low-resolution thermal input constrains the models' ability to recover high-frequency detail.

Nevertheless, the diffusion-guided approach still provides visible improvements over the other guidance types.

Method	RGB guide			Synt. GAN guide			Synt. P2P guide			Synt. Diffusion guide		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
FW-SAT[14]	24.84	0.7918	0.2954	24.39 (-0.45)	0.7800 (-0.0118)	0.3254 (+0.0300)	22.96 (-1.88)	0.7464 (-0.0454)	0.3211 (+0.0257)	25.11 (+0.27)	0.7935 (+0.0017)	0.2873 (-0.0081)
SwinFuSR[2]	27.73	0.7850	0.3160	27.49 (-0.24)	0.7760 (-0.0090)	0.3142 (-0.0018)	27.03 (-0.70)	0.7646 (-0.0204)	0.3332 (+0.0172)	27.98 (+0.25)	0.7867 (+0.0017)	0.2920 (-0.0240)
SwinPaste[37]	27.74	0.7843	0.2995	27.52 (-0.22)	0.7774 (-0.0069)	0.3150 (+0.0155)	27.07 (-0.67)	0.7644 (-0.0199)	0.3135 (+0.0140)	27.91 (+0.17)	0.7854 (+0.0011)	0.2937 (-0.0058)

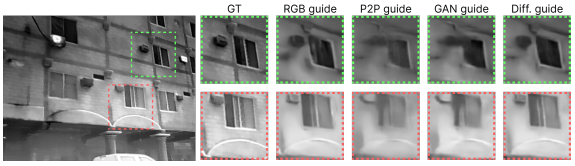
Table 5. Super-resolution results with different guidance methods on the CIDIS test set with a scale factor of $\times 16$. Values in parentheses represent the performance change relative to the RGB guide.

Its reconstructions show cleaner contours, and more consistent brightness in shadowed or reflective regions. By contrast, GAN and P2P guidance often fail to enhance the reconstruction and can distort structural elements, while the diffusion-guide preserves spatial coherence and produces perceptually more stable results. These observations match the quantitative trends, confirming that diffusion guidance remains the most effective choice even at extreme upscaling factors. More results in Supplementary Material.

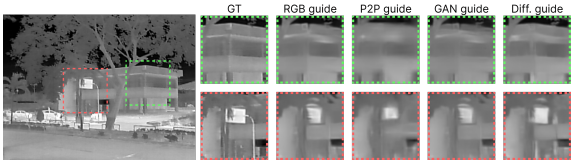
Overall, the experimental results across both datasets and scaling factors validate the hypothesis that replacing conventional RGB guidance with a synthetic pseudo-thermal image enhances the effectiveness of GSR. The consistent gains achieved by the diffusion-guided configuration, including cross dataset and large-scale scenarios, indicate that a guidance image with higher spectral and semantic affinity to the target domain enables more meaningful feature transfer during reconstruction.

5. Limitations and Future Work

Although the proposed strategy was evaluated on two well-established datasets and under different scaling factors, some limitations remain. M3FD and CIDIS primarily contain urban and semi-urban scenes, which restricts the assessment of generalization to broader surveillance contexts such as rural, or indoor environments, where texture complexity, thermal contrast, and noise characteristics may differ substantially. Future work should extend the evaluation to broader scene categories and additional datasets to fully assess cross-domain robustness. The range of scaling factors is also limited, since $\times 8$ and $\times 16$ offer a challenging yet realistic scope, exploring extreme magnifications, such as $\times 32$, could provide additional insights into the limits of pseudo-thermal guidance. Also, since the DM was trained only on M3FD, its generalization ability may vary when applied to thermal data with different spectral characteristics, potentially requiring dedicated training for other domains.



(a) Visual results using SwinFuSR model.



(b) Visual results using FW-SAT model.

Guide	PSNR↑	SSIM↑	LPIPS↓
RGB	23.01	0.7508	0.3701
P2P	22.81	0.7475	0.3831
GAN	22.98	0.7498	0.3781
Diff.	23.38	0.7607	0.3767

(c) Quantitative results of Fig. 7a.

Guide	PSNR↑	SSIM↑	LPIPS↓
RGB	23.15	0.7485	0.3193
P2P	22.90	0.7395	0.3271
GAN	22.99	0.7420	0.3101
Diff.	23.79	0.7514	0.3109

(d) Quantitative results of Fig. 7b.

Figure 7. Guided super-resolution results at $\times 16$ scale on the CIDIS dataset.

6. Conclusions

This work proposes a strategy for thermal GSR that replaces the conventional RGB guidance with a pseudo-thermal image, providing a synthetic modality that maintains stronger semantic and spectral consistency with the target. The pseudo-thermal guide is produced with a diffusion model trained for thermal synthesis and can be integrated into GSR architectures without modifying their design. The strategy was evaluated on the M3FD and CIDIS datasets at $\times 8$ and $\times 16$, consistently improving reconstruction quality across all tested models. The gains appeared in both pixel-based and perceptual metrics, with LPIPS reductions of up to 0.11 and notable increases in PSNR and SSIM. The approach remained effective even when the generator was trained on a different dataset, confirming its robustness and generalization. Visual results reinforced these findings, showing sharper edges, cleaner contours, and improved structural consistency in the diffusion-guided reconstructions.

Acknowledgement

This work was supported in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206, in part by Grant PID2024-162815NB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU, and Grant PID2021-128945NB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU. The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC's general activities, and the Departament de Recerca i Universitats from Generalitat de Catalunya to the SGR Research Group 2021 MACO (reference 2021 SGR 01499).

References

[1] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Comput. Surv.*, 53 (3):1–34, 2020. 1

[2] Cyprien Arnold, Philippe Jouvet, and Lama Seoud. Swinfusr: An image fusion-inspired model for rgb-guided thermal image super-resolution. In *CVPRW*, 2024. 2, 4, 6, 7, 8

[3] Junchi Bin, Heqing Zhang, Zhila Bahrami, Ran Zhang, Huan Liu, Erik Blasch, and Zheng Liu. The registration of visible and thermal images through multi-objective optimization. *Information Fusion*, 95:186–198, 2023. 1

[4] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. A survey on generative diffusion models. *IEEE TKDE*, 36(7): 2814–2830, 2024. 3

[5] Carlos Cortés-Mendez and Jean-Bernard Hayet. Exploring the usage of diffusion models for thermal image super-resolution: A generic uncertainty-aware approach for guided and non-guided schemes. In *CVPRW*, 2024. 2

[6] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE TPAMI*, 45(9):10850–10869, 2023. 3

[7] Yuan Fang, Lei Fan, and Yuanzhi Cai. Guided super-resolution for image fusion: A novel approach to enhancing crack segmentation in masonry structures. *IEEE Sensors Journal*, 25(7):11491–11507, 2025. 1

[8] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[9] Honey Gupta and Kaushik Mitra. Toward unaligned guided thermal super-resolution. *IEEE TIP*, 31:433–445, 2022. 1, 2

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[11] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *CVPR*, 2015. 4

[12] Mohamed Ramzy Ibrahim, Robert Benavente, Daniel Ponsa, and Felipe Lumbreras. Hyda-net: A hybrid dense attention network for remote sensing multi-image super-resolution. *IEEE JSTARS*, 18:7592–7614, 2025. 1

[13] Hongcheng Jiang and ZhiQiang Chen. Flexible window-based self-attention transformer in thermal image super-resolution. In *CVPRW*, 2024. 2

[14] Hongcheng Jiang and Zhiqiang Chen. Flexible window-based self-attention transformer in thermal image super-resolution. In *CVPRW*, 2024. 4, 7, 8

[15] Priya Kansal and Sabari Nathan. Dual-input frequency-aware network for high-quality thermal image super-resolution. In *CVPRW*, 2025. 1

[16] Aditya Kasliwal, Pratinav Seth, Sriya Rallabandi, and Sanchit Singhal. Corefusion: Contrastive regularized fusion for guided thermal super-resolution. In *CVPRW*, 2023. 2, 4, 6, 7

[17] Dawa Chyophel Lepcha, Bhawna Goyal, Ayush Dogra, and Vishal Goyal. Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 91:230–260, 2023. 1

[18] 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 *CVPR*, 2022. 2

[19] Xuhui Liu, Bohan Zeng, Sicheng Gao, Shanglin Li, Yutang Feng, Hong Li, Boyu Liu, Jianzhuang Liu, and Baochang Zhang. Ladiffgan: Training gans with diffusion supervision in latent spaces. In *CVPRW*, 2024. 3

[20] Riccardo De Lutio, Stefano D'aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as pixel-to-pixel transformation. In *ICCV*, 2019. 2

[21] Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. Pid: Physics-informed diffusion model for infrared image generation. *Pattern Recognition*, 169:111816, 2026. 3

[22] Mohamed Amine Marnissi, Hajer Fradi, Anis Sahbani, and Najoua Essoukri Ben Amara. Unsupervised thermal-to-visible domain adaptation method for pedestrian detection. *Pattern Recognition Letters*, 153:222–231, 2022. 1

[23] Raghunath Sai Puttagunta, Birendra Kathariya, Zhu Li, and George York. Multi-scale feature fusion using channel transformers for guided thermal image super resolution. In *CVPRW*, 2024. 1, 2

[24] Defu Qiu, Yuhu Cheng, and Xuesong Wang. Medical image super-resolution reconstruction algorithms based on deep learning: A survey. *Computer Methods and Programs in Biomedicine*, 238:107590, 2023. 1

[25] Leo Thomas Ramos and Angel D. Sappa. Multispectral semantic segmentation for land cover classification: An overview. *IEEE JSTARS*, 17:14295–14336, 2024. 1

[26] Rafael E Rivadeneira, Henry O Velesaca, and Angel Sappa. Cross-spectral image registration: a comparative study and a new benchmark dataset. In *ICICV*, 2024. 2

[27] Rafael E. Rivadeneira, Angel D. Sappa, Riad Hammoud, Jiyong Rao, Hang Zhong, Yu Wang, Shengjie Zhao, Zhiwei Zhong, Yung-Hui Li, Shiqi Wang, Qiangqiang Shen, Hanzhang Wang, and Xuanqi Zhang. Thermal image super-resolution challenge results - pbvs 2025. In *CVPRW*, 2025. 1

[28] Hu Su, Ying Li, Yifan Xu, Xiang Fu, and Song Liu. A review of deep-learning-based super-resolution: From methods to applications. *Pattern Recognition*, 157:110935, 2025. 1

[29] Patricia Suárez and Angel Sappa. Synthetic thermal image generation from multi-cue input data. In *VISIGRAPP*, 2025. 2

[30] Patricia L. Suárez and Angel Sappa. Depth-conditioned thermal-like image generation. In *ICPRS*, 2024. 4

[31] Patricia L. Suárez, Dario Carpio, and Angel D. Sappa. Enhancement of guided thermal image super-resolution approaches. *Neurocomputing*, 573:127197, 2024. 1

[32] Jiaxiang Tang, Xiaokang Chen, and Gang Zeng. Joint implicit image function for guided depth super-resolution. In *ACM MM*, 2021. 1

[33] Luan Thanh Trinh and Tomoki Hamagami. Latent denoising diffusion gan: Faster sampling, higher image quality. *IEEE Access*, 12:78161–78172, 2024. 3

[34] Yi Wang, Syed Muhammad Arsalan Bashir, Mahrukh Khan, Qudrat Ullah, Rui Wang, Yilin Song, Zhe Guo, and Yilong Niu. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Systems with Applications*, 197:116793, 2022. 1

[35] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. *IEEE TPAMI*, 43 (10):3365–3387, 2021. 1

[36] Zhicheng Zhao, Yong Zhang, Chenglong Li, Yun Xiao, and Jin Tang. Thermal uav image super-resolution guided by multiple visible cues. *IEEE TGRS*, 61:1–14, 2023. 2

[37] Hang Zhong, Yu Wang, and Shengjie Zhao. Swinpaste: A swin transformer-based framework for rgb-guided thermal image super-resolution. In *CVPRW*, 2025. 2, 4, 6, 7, 8

[38] Zhiwei Zhong, Xianming Liu, Junjun Jiang, Debin Zhao, and Xiangyang Ji. Guided depth map super-resolution: A survey. *ACM Comput. Surv.*, 55(14s):1–36, 2023. 1