CVPR
#16226

3M-TI: High-Quality Mobile Thermal Imaging via Calibration-free Multi-Camera Cross-Modal Diffusion

Anonymous CVPR submission

Paper ID 16226

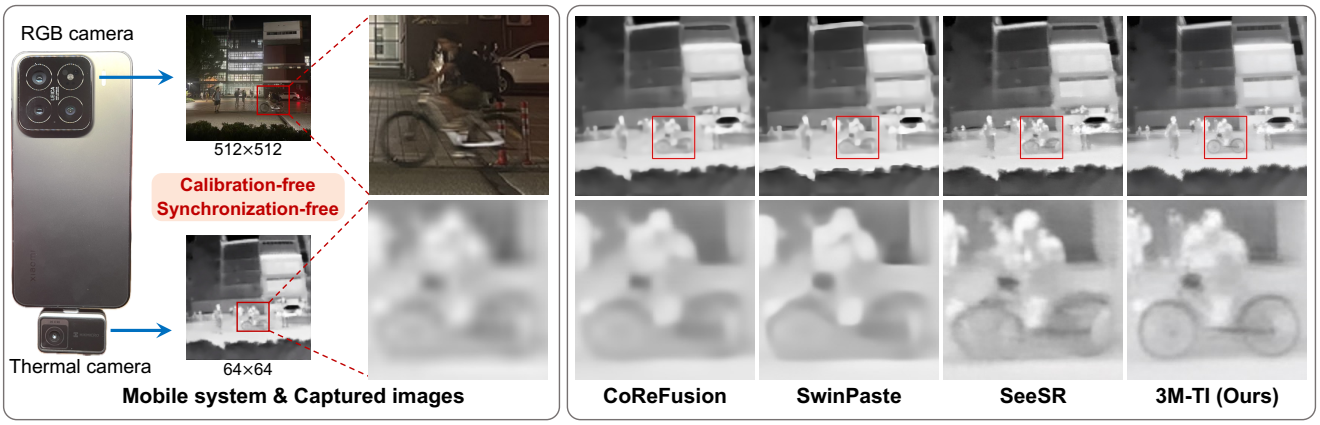

Figure 1. A smartphone-based mobile imaging system integrating calibration-free and synchronization-free RGB and thermal cameras. The proposed 3M-TI method delivers superior thermal image quality compared with state-of-the-art restoration approaches.

Abstract

The miniaturization of thermal sensors for mobile platforms inherently limits their spatial resolution and textural fidelity, leading to blurry and less informative images. Existing thermal super-resolution (SR) methods can be grouped into single-image and RGB-guided approaches: the former struggles to recover fine structures from limited information, while the latter relies on accurate and laborious cross-camera calibration, which hinders practical deployment and robustness. Here, we propose 3M-TI, a calibration-free Multi-camera cross-Modality diffusion framework for Mobile Thermal Imaging. At its core, 3M-TI integrates a cross-modal self-attention module (CSM) into the diffusion UNet, replacing the original self-attention layers to adaptively align thermal and RGB features throughout the denoising process, without requiring explicit camera calibration. This design enables the diffusion network to leverage its generative prior to enhance spatial resolution, structural fidelity, and texture detail in the super-resolved thermal images. Extensive evaluations on real-world mobile thermal cameras and public benchmarks validate our superior performance, achieving state-of-the-art results in both visual quality and quantitative metrics. More importantly, the thermal images enhanced by 3M-TI lead to substantial gains in critical downstream tasks like object detection and segmentation, underscoring its practical value for robust mobile thermal perception systems.

1. Introduction

Thermal imaging detects infrared radiation emitted by objects, extending perception beyond the visible spectrum for enhanced scene understanding. It is particularly effective under challenging conditions such as darkness, fog, or smoke, providing reliable scene information that complements RGB imaging for multimodal perception and fusion [30, 31, 43]. Owing to its robustness, thermal imaging has become increasingly valuable in safety-critical applications, such as autonomous driving, robotic navigation, and situational awareness [40, 41, 44].

Despite rapid advances in infrared sensing, mobile thermal imaging remains fundamentally constrained by hardware. Miniaturization for compact devices reduces aperture size, lowering spatial resolution and textural detail. In addition, the long wavelength of thermal radiation limits the minimum pixel size, while the high cost of infrared sensors further restricts pixel counts compared with RGB cameras.

These factors jointly result in blurry and less informative thermal images on mobile platforms, motivating extensive efforts to enhance their resolution and perceptual quality through computational reconstruction.

A representative example of such computational reconstruction is thermal image super-resolution (SR), which aims to reconstruct high-resolution spatial details from low-resolution thermal observations [49]. However, the inherently limited information in low-resolution observations, aggravated by the lack of large-scale thermal datasets, makes thermal super-resolution highly underconstrained and hinders network training and generalization compared with RGB-based tasks. To compensate for this deficiency, RGB-guided SR methods exploit cross-modal cues to enhance reconstruction [1, 12, 17]. Nevertheless, due to the fundamental differences in imaging principles, thermal and RGB images often exhibit substantial discrepancies, and directly merging their features may introduce unrealistic details [29, 59, 62]. Moreover, most existing methods rely on pixel-level aligned RGB-thermal pairs, requiring laborious calibration, which severely limits robustness and scalability in practical deployment.

To address these challenges, we propose 3M-TI, a calibration-free **M**ulti-camera cross-**M**odal diffusion framework for **M**obile **T**hermal **I**maging. Firstly, rather than relying on pixel-level calibration between RGB and thermal cameras, 3M-TI performs alignment in the latent space of a variational autoencoder (VAE) [19], where continuous and disentangled representations facilitate robust cross-modal correspondence. An attention-based fusion module further integrates RGB and thermal representations in this latent domain, achieving reliable alignment and fusion without explicit registration. Secondly, because existing RGB–thermal datasets are typically small, homogeneous, and strictly aligned [16, 28], we design a data augmentation strategy that intentionally offsets RGB–thermal pairs to simulate real-world conditions. This encourages the attention module to learn robust cross-modal correspondence and fusion without relying on explicit pixel-level alignment. Furthermore, we incorporate a one-step latent-space diffusion [39] to suppress fusion artifacts, enhance visual fidelity, and improve robustness under limited-data conditions.

Finally, we validate the proposed 3M-TI framework on a real mobile thermal imaging system. Experiments show that it significantly outperforms existing state-of-the-art methods, highlighting both its effectiveness and practical potential. The main contributions of 3M-TI can be summarized as follows:

- **Calibration-free Fusion**: Fully automatic alignment and fusion of uncalibrated RGB–thermal pairs via a cross-modal self-attention module, without requiring pixel-level image registration.
- **Misalignment Augmentation**: Camera-pose augmentation that generates intentionally misaligned RGB counterparts, enabling the cross-modal self-attention module to handle multi-camera parallax and unsynchronized captures.
- **Cross-modal Diffusion**: Integration of latent-space diffusion to suppress fusion artifacts, enhance detail realism, and leverage pretrained priors to compensate for limited thermal data.
- **Practical Validation**: Evaluation on both public datasets and a real mobile thermal imaging system, demonstrating effectiveness and practical potential.

2. Related Work

Computational thermal imaging intersects several areas of computer vision and image processing. In this section, we review the most relevant work in three directions: classical and learning-based image restoration, reference-guided image super-resolution, and recent diffusion-based approaches.

2.1. Image Restoration

Image restoration seeks to recover a high-quality image from a degraded observation by inverting an imaging model. Classical image restoration methods are typically model-based, formulated as inverse problems with explicit priors. Variational methods regularize the solution with smoothness or edge-preserving constraints [38]. Sparse representation approaches exploit dictionary learning and sparsity priors to recover fine details [10, 32]. Non-local self-similarity models, such as non-local means [3] and BM3D [8], leverage the redundancy of similar patches across the image for collaborative filtering. Deep learning has fundamentally reshaped image restoration by learning powerful data-driven priors directly from large-scale datasets. CNN-based models (e.g., SRCNN[9], EDSR[26]) pioneered end-to-end restoration pipelines, while recent Transformer architectures such as SwinIR[25] further improved long-range dependency modeling. These advances have led to state-of-the-art performance in denoising [24, 25, 45], deblurring [20, 27], and super-resolution [5, 6, 9, 21, 25]. Nevertheless, single-image restoration remains challenging under severe degradation or large upscaling factors, due to the limited high-frequency information in a single observation, motivating approaches that leverage additional cues.

2.2. Reference-Guided Image Restoration

Reference-guided image restoration leverages an auxiliary high-resolution modality (typically RGB) to recover fine details lost in degraded observations. This paradigm has been explored across multiple tasks, including RGB image super-resolution [11, 42, 58, 61], hyperspectral image super-resolution [12, 57], and thermal image enhance-

ment [1, 63]. Early approaches, such as CoReFusion [17], employ a dual-encoder UNet to jointly extract and fuse RGB–thermal features. A contrastive loss is further introduced to improve cross-modal consistency and overall restoration quality. With the advent of Transformers [46], attention-based fusion has become the dominant paradigm. MGNet [60] mines multi-level cues (appearance, edge, and semantics) from RGB guidance to enhance UAV thermal super-resolution. SwinFuSR [1] introduces a lightweight Swin-Transformer backbone with robust training under missing guidance, and SwinPaste [63] further improves it via data mixing and multi-scale supervision. MSFFCT [35] combines channel-based Transformers with multi-scale feature fusion to capture long-range dependencies and rich thermal-RGB correlations. However, the large domain shift between RGB and thermal images poses a significant challenge to establishing reliable cross-modal alignment, especially under uncalibrated conditions.

2.3. Diffusion-Powered Image Restoration

Diffusion models have recently demonstrated remarkable capabilities in image restoration and super-resolution by learning powerful generative priors that capture rich textures and structural details [13]. This generative nature allows diffusion models to recover fine structures and high-frequency details that are challenging for CNN- and Transformer-based approaches. However, the original diffusion model is computationally expensive due to its long iterative denoising process, limiting practical deployment. To address this, several efficient variants have been proposed: latent-space diffusion methods [47, 51, 52, 54] perform denoising within the compact latent representation of a VAE, reducing computation overhead while preserving high-fidelity reconstructions; distillation-based approaches [15, 33, 39] transfer the generative ability of a large teacher model to a lightweight student, enabling faster inference with minimal quality loss.

Diffusion models have also been applied to thermal image super-resolution [7, 23]. DifIISR [23] incorporates gradient-based priors and a thermal spectrum distribution regulation into the diffusion process, guiding the model to capture the unique frequency characteristics of infrared images and achieving strong visual quality and downstream task performance. [7] adapts the ResShift diffusion model[53] to thermal imaging with an uncertainty-aware approach. This method produces a confidence map that evaluates the reliability of each reconstructed region, thereby enhancing pixel-level fidelity and facilitating texture reconstruction.

However, existing methods still rely on strictly pixel-aligned datasets and explicit calibration between cameras, making them sensitive to spatial and temporal misalignment and difficult to generalize to real-world mobile scenarios. These limitations motivate our design of 3M-TI, a calibration-free latent-space diffusion framework that leverages multi-modal information for robust mobile thermal imaging.

3. Method

We propose 3M-TI, a cross-modal diffusion frame, for high-quality thermal image reconstruction from uncalibrated RGB-thermal image pairs (Fig. 2). 3M-TI specifically addresses three key challenges: (1) uncalibrated and unsynchronized RGB-thermal captures, (2) fusion of heterogeneous RGB-thermal representations, and (3) limited data scale and diversity of RGB-thermal datasets.

Section 3.1 provides an overview of the 3M-TI framework. Section 3.2 introduces the cross-modal self-attention (CSA) module, which performs automatic alignment and fusion within the VAE latent space. Section 3.3 presents a misalignment augmentation strategy that simulates realistic inter-camera parallax and temporal offset to improve the robustness of CSA. Finally, Section 3.4 details the practical implementation and validation of 3M-TI on a smartphone-based multi-camera thermal imaging system.

3.1. 3M-TI Framework Overview

The 3M-TI framework reconstructs a high-resolution thermal image from a low-resolution thermal input and an uncalibrated high-resolution RGB reference, as illustrated in Fig. 2(a). Our approach is built upon the one-step diffusion model, SD-Turbo [39], to ensure efficient inference. The process begins by encoding the thermal and RGB images into latent representations using the frozen VAE encoder. To establish cross-modal correspondence and fusion in the latent space, we introduce the cross-modal self-attention module (CSM). This module replaces the original self-attention layers in the diffusion UNet with our cross-modal self-attention layers, which are designed to learn multiscale correspondences between the RGB and thermal latents (Fig. 2(b)). During training, we apply misalignment augmentation to the RGB images to enhance robustness against uncalibration and temporal unsynchronization (Fig. 2(c)). Furthermore, a skip connection [34] is incorporated to enhance structural consistency and mitigate geometric distortions. Specifically, the feature maps from 4 encoder downsampling blocks are passed through a 1×1 zero-initialized convolutional layer before being added to the corresponding decoder upsampling blocks. Since extracting reliable semantics directly from low-resolution thermal images is suboptimal, we generate text prompts from the corresponding RGB images using the Recognize Anything Model (RAM) [56]. Finally, we employ low-rank adaptation (LoRA) [14] to fine-tune both the UNet and the VAE decoder.

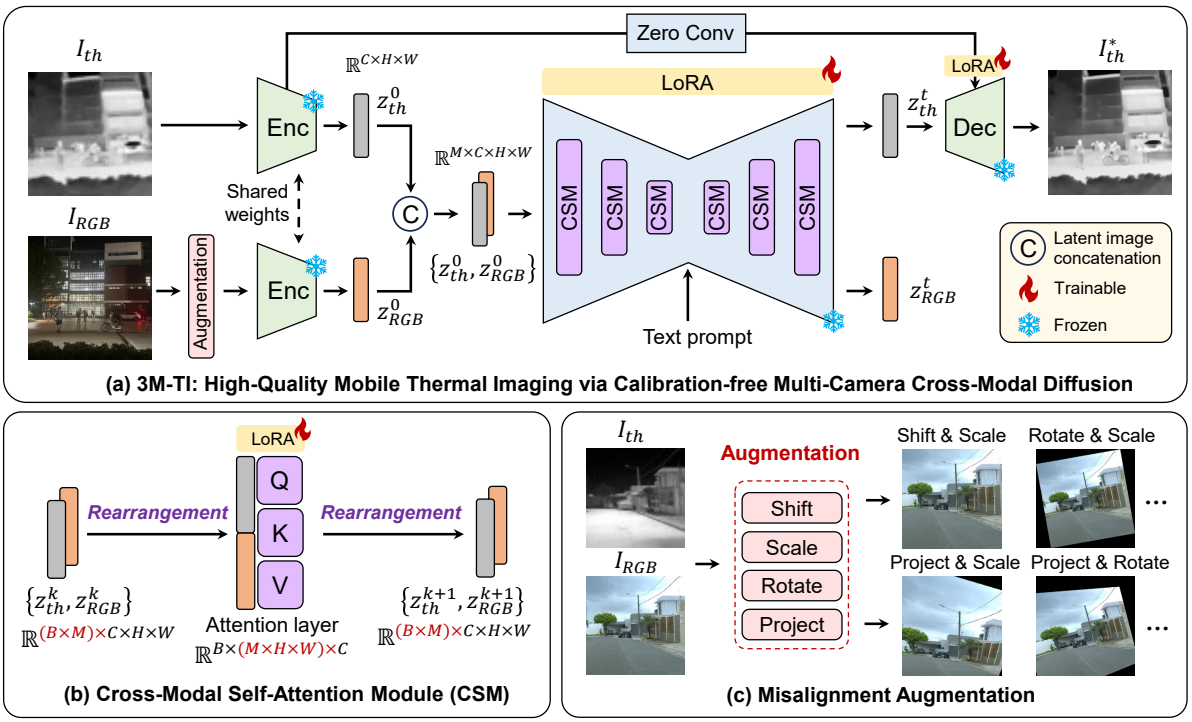

Figure 2. Overview of the 3M-TI architecture. (a) 3M-TI framework. The core of 3M-TI is a one-step diffusion-based model equipped with a cross-modal self-attention module (CSM) and a misalignment augmentation strategy. LoRA fine-tuning is applied to both the UNet and the VAE decoder. (b) Cross-modal self-attention module (CSM). Two rearrangement layers are inserted before and after the original self-attention layers to capture cross-modal correspondences. (c) Misalignment augmentation. A data augmentation strategy designed to enhance model robustness against camera parallax and temporal misalignment between RGB and thermal inputs.

3.2. Cross-Modal Self-Attention Module

To establish cross-modal correspondence between uncalibrated RGB and thermal images, we introduce the cross-modal self-attention module (CSM). Inspired by video- and multi-view diffusion models [2, 50], our key idea is to repurpose the transformer blocks within the pretrained diffusion UNet for cross-modal RGB-thermal correspondence and fusion. The RGB and thermal inputs are first encoded into a concatenated latent tensor $\{z^0_{RGB}, z^0_{th}\} \in \mathbb{R}^{B\times M\times C\times H\times W}$, where B is the batch size, $M = 2$ denotes the number of images (RGB and thermal, can be extended to multiple RGB images), and C, H, W are the channel and spatial dimensions.

For non-transformer blocks in the diffusion UNet, the image-number dimension M is folded into the batch dimension so that RGB and thermal images are processed independently. Within the transformer blocks (Fig. 2b), two tensor rearrangements are applied. Before entering a block, the latent tensor is reshaped to $\mathbb{R}^{B\times(M\times H\times W)\times C}$, treating each pixel as a token and merging all pixels from both modalities into a single sequence with a length of $M \times H \times W$. Self-attention then computes pairwise dependencies among all RGB and thermal pixels, naturally enabling the transformer to learn cross-modal correspondences and integrate complementary cues. After the block, the latent tensor is reshaped back to $\mathbb{R}^{(B\times M)\times C\times H\times W}$ for subsequent layers.

At the output stage, the refined thermal latent is merged with the initial latent in the VAE encoder via a zero-initialized skip connection, producing the final high-fidelity thermal image. The cross-modal self-attention module introduces no additional parameters, allowing the UNet and VAE decoder to be efficiently fine-tuned using LoRA [14].

3.3. Misalignment Augmentation

Existing RGB-thermal datasets are typically small and strictly pixel-aligned, limiting generalization to practical multi-camera configurations. In practice, spatial misalignment arise from camera parallax and temporal unsynchronization: parallax varies with scene depth and baseline, while temporal offsets are determined by object motion and capture delay.

Since our goal is to learn a model that is robust to these geometric variations, rather than overfitting to a specific calibrated setup, we propose a misalignment augmentation strategy without physical simulation. 3M-TI applies a set of controlled spatial transformations to the RGB images, including translation, scaling, rotation, and perspec-

tive warping. The transformation parameters are chosen to reflect typical deviations encountered in handheld and multi-camera scenarios (see Figure 2(c)).

This misalignment augmentation encourages the CSM to learn robust cross-modal correspondences under uncalibrated and unsynchronized conditions, effectively bridging the gap between constrained training data and diverse deployment environments.

3.4. Practical Multi-Camera System

To validate our method under realistic conditions, we constructed a multi-camera system comprising a HIKVISION P09 thermal camera module (costing less than 100 US Dollars) and a Xiaomi 15 smartphone, connected via a Type-C interface (Fig. 1). The thermal module provides a native resolution of 96×96, which is resized to 64×64 in this work, with a $50° \times 50°$ field of view (FOV) and a pixel pitch of 12 µm. For the RGB reference, we utilize the primary camera of the Xiaomi 15, which is equipped with an OV50H sensor and captures images at a resolution of 4096×3072. This RGB camera has an equivalent focal length of 23 mm, yielding a FOV of $74° \times 59°$, slightly wider than that of the thermal module. This system offers a realistic platform to evaluate the performance of 3M-TI under practical, uncalibrated mobile imaging conditions.

4. Experiments

4.1. Experimental Setup

Datasets Preparation. Our training and test sets are compiled from four public RGB-thermal datasets: IRVI [22], LLVIP [16], M^{3}FD [28], and the PBVS 2025 TISR Challenge Track 2 [37]. Among them, IRVI, LLVIP, and PBVS 2025 provide official training-validation/test splits, while M^{3}FD does not. To ensure a balanced and diverse dataset, we sample images from each source as follows: 3,200 pairs from IRVI, 3,200 from LLVIP, 3,822 from M^{3}FD, and 700 from the PBVS 2025 training split, forming a combined training set of 10,922 image pairs. For evaluation, we construct a test set of 1,176 pairs by sampling 300 pairs from the IRVI test set, 300 from LLVIP, 376 from M^{3}FD, and 200 from the PBVS 2025 validation set. Since the original datasets contain strictly aligned image pairs, we further augment the test set by applying controlled spatial transformations to simulate camera parallax and temporal offset. All thermal images are center-cropped to a square aspect ratio and resized to 64×64 pixels. To better mimic real sensor noise, we add Gaussian noise. Corresponding RGB images are cropped identically and resized to 512×512 pixels to serve as high-resolution references.

Smartphone Dataset. For real-world evaluation, we collected 40 image pairs across 30 diverse scenes using our smartphone-based multi-camera system, including 37 nighttime and the remaining daytime captures. Similarly, all thermal images are resized to 64×64. The corresponding RGB images are center-cropped to a square aspect ratio and resized to 512×512 as high-resolution references.

Implementation Details. The model is trained with the loss in Eq. (1), combining L2 and learned perceptual image patch similarity (LPIPS) [55] terms with $\lambda = 1$:

$$\mathcal{L} = \mathcal{L}_2 + \lambda \cdot \mathcal{L}_{\text{LPIPS}}. \tag{1}$$

Training is performed using the Adam optimizer [18] with a learning rate of 2×10^{-5} and a batch size of 4 on a single NVIDIA A800 (80 GB) GPU for about 4 hours (8000 iterations). LoRA is applied with ranks of 16 for the UNet and 4 for the VAE decoder.

Comparative Methods. We compare our model against 5 representative baselines: CoReFusion [17], SwinFuSR [1], SwinPaste [63], SeeSR [52], and OSEDiff [51]. CoReFusion is an RGB-guided super-resolution model built upon the UNet architecture. SwinFuSR employs a Swin Transformer backbone with RGB guidance, while SwinPaste is an enhanced variant. SeeSR and OSEDiff are diffusion-based image super-resolution methods without reference guidance. For fair comparison, all baseline models are retrained on our dataset using their publicly released codes.

Evaluation Metrics. We evaluate model performance using reference metrics, including PSNR and SSIM [48] for reconstruction fidelity and LPIPS [55] for perceptual quality. We also report no-reference metrics, MUSIQ [4] and MANIQA [4], to quantify overall image quality.

4.2. Image Quality Assesment

Table 1 summarizes the quantitative results. Our 3M-TI achieves the best performance on perceptual metrics, including LPIPS, MANIQA, and MUSIQ. For the reference-based LPIPS metric, 3M-TI, OSEDiff, and SeeSR outperform non-diffusion methods (CoReFusion, SwinFuSR, and SwinPaste), highlighting the power of diffusion priors for perceptual enhancement. In terms of fidelity metrics (PSNR, SSIM), 3M-TI ranks a close second, demonstrating better structural preservation than OSEDiff, SeeSR, SwinFuSR, and SwinPaste. The proposed misalignment augmentation significantly improves 3M-TI, particularly on perceptual metrics, while other methods see only marginal gains. This indicates that 3M-TI effectively learns cross-modal correspondences and generalizes well to diverse uncalibrated conditions.

As shown in Fig. 3, qualitative results reveal a key trend: non-diffusion methods tend to achieve high PSNR/SSIM yet produce overly smooth reconstructions that miss fine high-frequency structures. OSEDiff and SeeSR can synthesize high-frequency content but often introduces details that are inconsistent with the ground truth. In contrast, 3M-TI reconstructs fine details that are both visually plausible and

CVPR
#16226

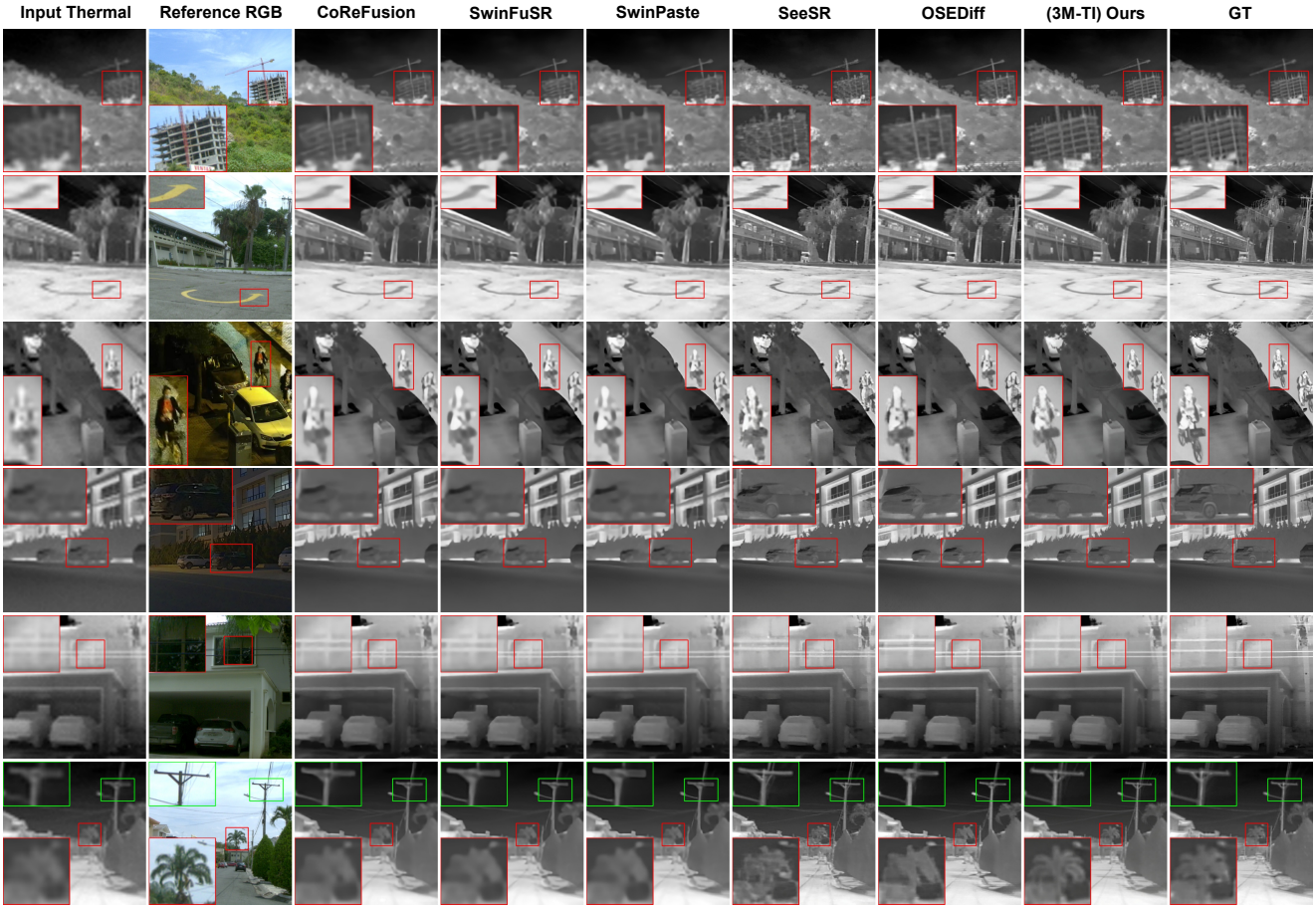

Figure 3. Qualitative comparison on our test set (zoom in for details). 3M-TI achieves the most faithful and visually consistent results, exhibiting sharp structures and accurate thermal patterns that best align with the GT.

closely aligned with the ground truth. For example, 3M-TI faithfully transfers geometric details such as partially constructed buildings (Row 1), road markings (Row 2), windows and wires (Row 5), and tree branches and utility poles (Row 6) from the RGB references to the thermal outputs, whereas other methods either miss these structures or reproduce them inaccurately. 3M-TI also yields more natural object contours (e.g., cyclists and cars in the Row 3 and Row 4), resulting in visually realistic thermal images.

Figure 4 presents qualitative results on our smartphone dataset, confirming 3M-TI's robustness to real-world, uncalibrated pairs. While CoReFusion, SwinFuSR, and SwinPaste yield blurry outputs; OSEDiff and SeeSR introduce unrealistic artifacts, our method produces geometrically consistent and visually plausible details. A key example is the bicycle wheels (Row 3), where 3M-TI accurately reconstructs their circular shape despite significant positional misalignment between the RGB and thermal captures due to unsynchronization. Moreover, 3M-TI can leverage weak cues from degraded references, as shown by its ability to recover the railing structure (Row 5) from an RGB image heavily corrupted by light flare. Table 2 reports no-reference evaluation on the captured smartphone dataset, where 3M-TI achieves the highest scores. More results and technical details are given in the supplementary material.

4.3. Downstream Applications

The high-fidelity outputs of 3M-TI provide structurally and semantically rich representations that directly benefit downstream vision tasks. To validate this claim, we evaluate two representative tasks: open-vocabulary object detection and semantic segmentation. In these experiments, we apply the pretrained Grounded-SAM model [36] in a zero-shot manner (no fine-tuning) and use identical text prompts for all methods to ensure fair comparisons.

Object detection. For the examples in Fig. 5, the text prompts are "automobile" and "person". 3M-TI demonstrates superior detection robustness and object discovery capability compared to SwinPaste and SeeSR. Table 3 summarizes detection performance (Precision, Recall, F1-score, and IoU) on the test set (LLVIP and M^3FD with GT an-

CVPR
#16226

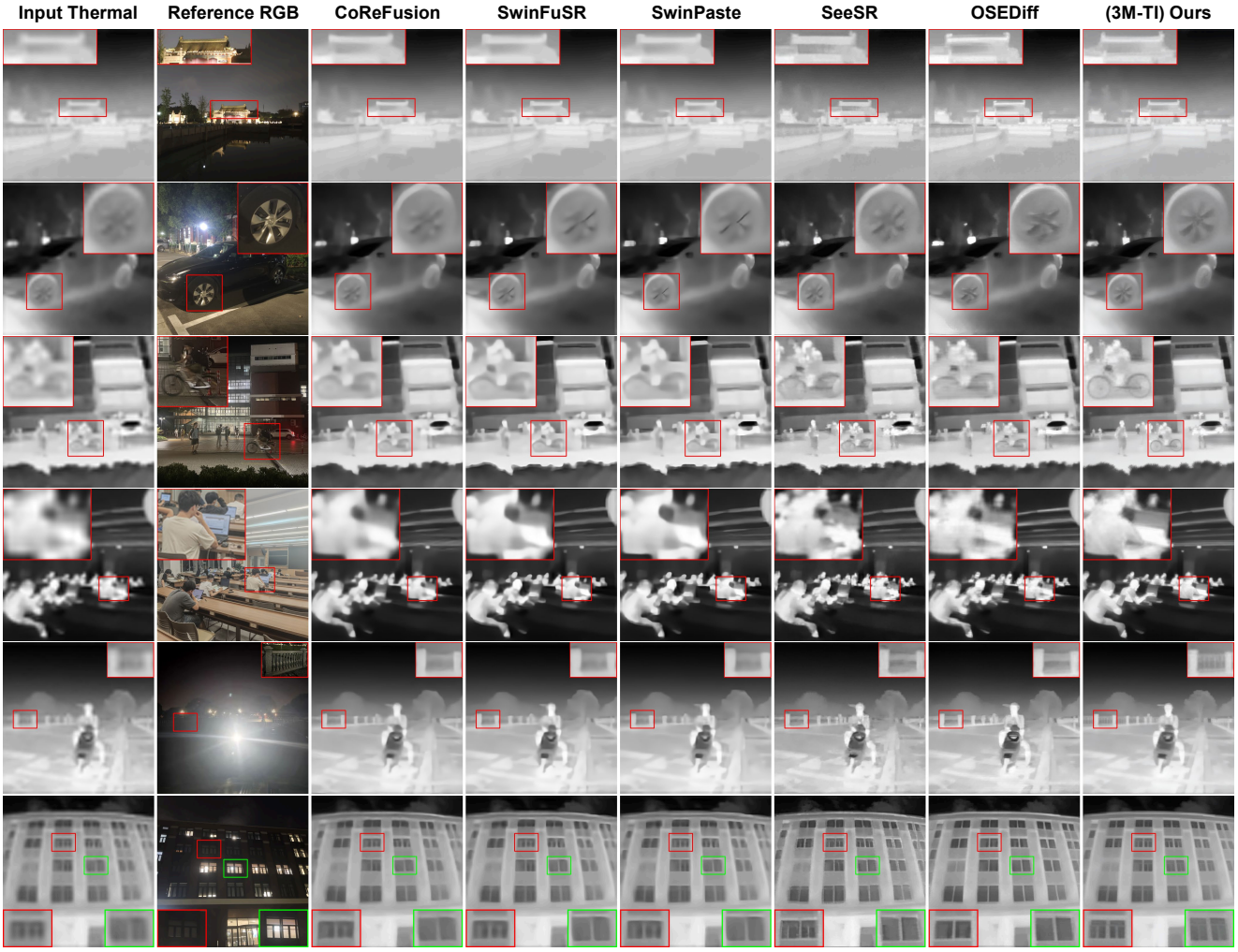

Figure 4. Qualitative comparison on our real-world smartphone dataset (zoom in for details). 3M-TI exhibits remarkable generalization capability, producing sharp and faithful thermal details that closely align with RGB images.

Methods \ Metrics	PSNR↑	SSIM↑	LPIPS↓	MANIQA↑	MUSIQ↑
CoReFusion [17]	30.11	0.8588	0.3214	0.2771	28.35
CoReFusion w/ Augment	**30.30**	**0.8634**	0.3174	0.2748	28.95
SwinFuSR [1]	29.85	0.8549	0.3085	0.2740	29.86
SwinFuSR w/ Augment	29.98	0.8581	0.3134	0.2753	30.33
SwinPaste [63]	29.83	0.8545	0.3075	0.2719	29.63
SwinPaste w/ Augment	29.91	0.8575	0.3084	0.2745	30.66
SeeSR [52]	29.41	0.8495	0.1828	0.4278	35.22
OSEDiff [51]	28.05	0.8422	0.2113	0.4014	36.30
Ours w/o Augment	30.04	0.8597	0.1917	0.4265	34.94
Ours w/ Augment	30.09	0.8610	**0.1787**	**0.4443**	**36.66**

Table 1. Performance comparison of different methods on public datasets. Gray cells indicate the best result, and light gray cells indicate the second-best result for each metric.

notations). 3M-TI achieves the best overall detection performance, marginally surpassing the reference RGB results and closely approaching the scores obtained on GT thermal images.

Metrics \ Methods	CoReFusion	SwinFuSR	SwinPaste	SeeSR	OSEDiff	3M-TI
MUSIQ↑	25.74	25.99	26.17	30.15	29.85	**30.62**
MANIQA↑	0.2701	0.2754	0.2749	0.3454	0.3285	**0.3589**

Table 2. Performance comparison on real-world smartphone dataset. Gray cells indicate the best.

Semantic segmentation. For the examples in Fig. 6, the prompts are "sky, tree, automobile, window" and "wheel, automobile, rider". 3M-TI produces more accurate and coherent segmentation maps than other methods, while complementing the RGB reference. In the first example, compared with the RGB results, 3M-TI segments trees more precisely though it misses a small automobile in the bottom-left. In the second low-light example, 3M-TI even outperforms the RGB reference (e.g., wheels and the upper automobile).

Overall, these results confirm that 3M-TI generates se-

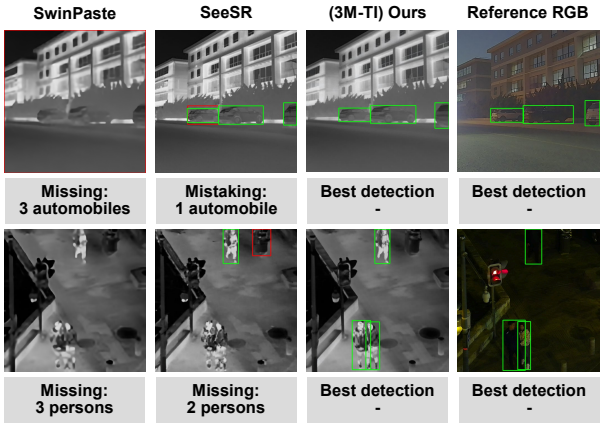

Figure 5. Visualization of detection results, where green bounding boxes indicate the correct detection, red bounding boxes indicate the wrong detection.

mantically meaningful details that enhance performance in critical vision tasks, demonstrating its practical utility beyond perceptual quality.

Methods \ Metrics	Precision↑	Recall↑	F1-score↑	IoU↑
Thermal (SwinPaste)	0.1800	0.2109	0.1765	0.1941
Thermal (SeeSR)	0.3832	0.4637	0.3849	0.3022
Thermal (3M-TI)	**0.4565**	0.5455	**0.4724**	**0.3427**
Reference RGB	0.4322	**0.5708**	0.4643	0.3359
Thermal (GT)	0.4582	0.5793	0.4887	0.3494

Table 3. Detection performance comparison across different methods, reference RGB, and GT, evaluated by Precision, Recall, F1-score, and IoU.

Methods \ Metrics	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑
w/o Reference	29.97	0.8592	0.2106	32.85
w/o Augmentation	30.04	0.8597	0.1917	34.94
w/o Skip	29.86	0.8572	0.1795	36.58
Ours w/ All	**30.09**	**0.8610**	**0.1787**	**36.66**

Table 4. Ablation study of 3M-TI components. Gray cells indicate the best result for each metric.

4.4. Ablation Study

We conduct an ablation study to investigate the contributions of each component. Table 4 and Fig. 7 report quantitative drops and representative visual examples when individual modules are removed. Removing the RGB reference produces markedly blurrier reconstructions (e.g., bicycle spokes and shrub textures), indicating the importance of cross-modal cues. Removing the misalignment augmentation reduces robustness to geometric and temporal offsets, leading to visibly degraded high-frequency details. Removing the skip connection degrades structural fidelity: circular wheels become distorted and geometric consistency is lost. With all components enabled, 3M-TI attains the best balance of fidelity and perceptual quality, producing sharp, semantically consistent thermal reconstructions.

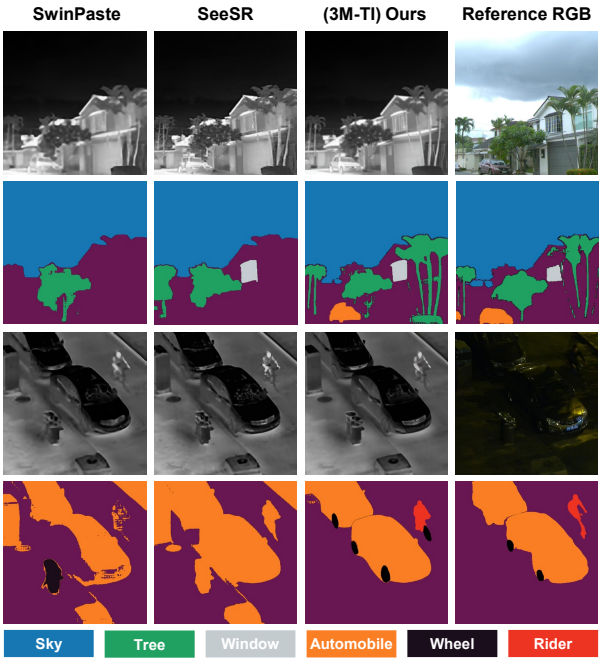

Figure 6. Visualization of segmentation results, where different colors represent different object categories.

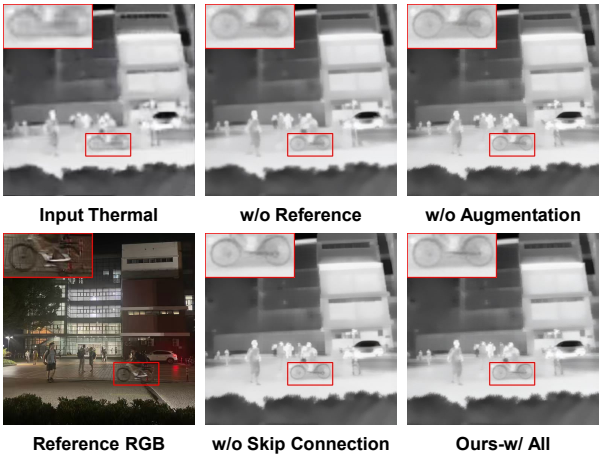

Figure 7. Qualitative results of the ablation study.

5. Conclusion

We proposed 3M-TI, a cross-modal diffusion framework that integrates calibration-free RGB guidance to enhance thermal image reconstruction. With the cross-modal self-attention module and misalignment augmentation, 3M-TI effectively aligns and fuses information across modalities, producing thermal images with sharper, more realistic, and semantically consistent details. Comprehensive experiments and real-system evaluations demonstrate that 3M-TI not only achieves superior perceptual quality but also significantly benefits downstream machine vision tasks.

References

[1] Cyprien Arnold, Philippe Jouvet, and Lama Seoud. Swinfusr: an image fusion-inspired model for rgb-guided thermal image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3027–3036, 2024. 2, 3, 5, 7

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 4

[3] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local means denoising. *Image processing on line*, 1: 208–212, 2011. 2

[4] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. 5

[5] 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. 2

[6] 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. 2

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3123–3130, 2024. 3

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 2

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2

[10] Weisheng Dong, Lei Zhang, Guangming Shi, and Xin Li. Nonlocally centralized sparse representation for image restoration. *IEEE transactions on Image Processing*, 22(4): 1620–1630, 2012. 2

[11] Lu Fang, Mengqi Ji, Xiaoyun Yuan, Jing He, Jianing Zhang, Yinheng Zhu, Tian Zheng, Leyao Liu, Bin Wang, and Qionghai Dai. Engram-driven videography. *Engineering*, 25:101–109, 2023. 2

[12] Ying Fu, Tao Zhang, Yinqiang Zheng, Debing Zhang, and Hua Huang. Hyperspectral image super-resolution with optimized rgb guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11661–11670, 2019. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3, 4

[15] Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation. *Advances in Neural Information Processing Systems*, 36:65299–65316, 2023. 3

[16] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. 2, 5

[17] Aditya Kasliwal, Pratinav Seth, Sriya Rallabandi, and Sanchit Singhal. Corefusion: Contrastive regularized fusion for guided thermal super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 507–514, 2023. 2, 3, 5, 7

[18] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[20] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 2

[21] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2

[22] Shuang Li, Bingfeng Han, Zhenjie Yu, Chi Harold Liu, Kai Chen, and Shuigen Wang. I2v-gan: Unpaired infrared-to-visible video translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3061–3069, 2021. 5

[23] Xingyuan Li, Zirui Wang, Yang Zou, Zhixin Chen, Jun Ma, Zhiying Jiang, Long Ma, and Jinyuan Liu. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7534–7544, 2025. 3

[24] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18278–18289, 2023. 2

[25] 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. 2

[26] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE confer-*

ence on computer vision and pattern recognition workshops, pages 136–144, 2017. 2

[27] Hanzhou Liu, Binghan Li, Chengkai Liu, and Mi Lu. Deblurdinat: A lightweight and effective transformer for image deblurring. *CoRR*, 2024. 2

[28] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5811, 2022. 2, 5

[29] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2226–2235, 2025. 2

[30] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2226–2235, 2025. 1

[31] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019. 1

[32] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *IEEE Transactions on image processing*, 17(1):53–69, 2007. 2

[33] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14297–14306, 2023. 3

[34] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 3

[35] Raghunath Sai Puttagunta, Birendra Kathariya, Zhu Li, and George York. Multi-scale feature fusion using channel transformers for guided thermal image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3086–3095, 2024. 3

[36] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 6

[37] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4630–4639, 2025. 5

[38] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 2

[39] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 2, 3

[40] Mark Sheinin, Aswin C Sankaranarayanan, and Srinivasa G Narasimhan. Projecting trackable thermal patterns for dynamic computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25223–25232, 2024. 1

[41] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep depth estimation from thermal image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1043–1053, 2023. 1

[42] Yang Tan, Haitian Zheng, Yinheng Zhu, Xiaoyun Yuan, Xing Lin, David Brady, and Lu Fang. Crossnet++: Cross-scale large-parallax warping for reference-based super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4291–4305, 2020. 2

[43] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. Divfusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023. 1

[44] Zitian Tang, Wenjie Ye, Wei-Chiu Ma, and Hang Zhao. What happened 3 seconds ago? inferring the past with thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17111–17120, 2023. 1

[45] Zhijun Tu, Kunpeng Du, Hanting Chen, Hailing Wang, Wei Li, Jie Hu, and Yunhe Wang. Ipt-v2: Efficient image processing transformer using hierarchical attentions. *arXiv preprint arXiv:2404.00633*, 2024. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[47] 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

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[49] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020. 2

[50] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26024–26035, 2025. 4

[51] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37:92529–92553, 2024. 3, 5, 7

[52] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024. 3, 5, 7

[53] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36:13294–13307, 2023. 3

[54] Jianing Zhang, Jiayi Zhu, Feiyu Ji, Xiaokang Yang, and Xiaoyun Yuan. Degradation-modeled multipath diffusion for tunable metalens photography. *arXiv preprint arXiv:2506.22753*, 2025. 3

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[56] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 3

[57] Yingkai Zhang, Zeqiang Lai, Tao Zhang, Ying Fu, and Chenghu Zhou. Unaligned rgb guided hyperspectral image super-resolution with spatial-spectral concordance: Y. zhang et al. *International Journal of Computer Vision*, pages 1–21, 2025. 2

[58] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7982–7991, 2019. 2

[59] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13955–13965, 2023. 2

[60] Zhicheng Zhao, Yong Zhang, Chenglong Li, Yun Xiao, and Jin Tang. Thermal uav image super-resolution guided by multiple visible cues. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. 3

[61] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 88–104, 2018. 2

[62] Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing synergistic high-order interaction in infrared and visible image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26384–26395, 2024. 2

[63] Hang Zhong, Yu Wang, and Shengjie Zhao. Swinpaste: A swin transformer-based framework for rgb-guided thermal image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4589–4594, 2025. 3, 5, 7