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

Minchong Chen[1,2,†] Xiaoyun Yuan[1,2,†,*] Junzhe Wan[1] Jianing Zhang[3,4] Jun Zhang[4]
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]School of Computer Science, Shanghai Jiao Tong University [3]Fudan University [4]Tsinghua University
yuanxiaoyun@sjtu.edu.cn

1. Dataset Collection

Datasets Preparation Details. For our training set, we sample 3,200 thermal-RGB pairs out of 22,079 pairs in the training set of IRVI [4]; 3,200 out of 12,025 pairs in the training set of LLVIP [2]; 3,822 pairs split from 4,198 pairs of M^{3}FD [5]; all 700 pairs in the training set of PBVS 2025 TISR Challenge Track 2 [6], forming a combined training set of 10,922 thermal-RGB pairs. For our test set, we sample 300 thermal-RGB pairs out of 2,272 pairs in the test set of IRVI; 300 out of 3,463 pairs in the test set of LLVIP; the remaining 376 pairs of M^{3}FD; all 200 pairs in the validation set of PBVS 2025 TISR Challenge Track 2, forming a combined test set of 1,176 thermal-RGB pairs.

We design dataset-specific sampling strategies to ensure sufficient scene diversity and avoid overlapping scenarios across splits:

- **IRVI and LLVIP datasets**: Since the images are extracted from continuous video frames, random sampling may result in densely clustered samples from the same scene. To mitigate this, we **uniformly** sample frames from the training/test subsets of both datasets.
- **M^{3}FD dataset**: To avoid scenario overlap between training and testing, we adopt a scene-level partition. Based on the dataset's ordering, the first 3,822 images cover complete scenarios and are assigned to the training set, while the remaining images constitute the test set.
- **PBVS 2025 TISR Challenge Track 2 dataset**: Given its small scale and the existence of an official split, we directly use all images provided in the official set without additional sampling.

Datasets Preprocessing. The original resolution of IRVI images is 256×256, and we directly upsample them to 512×512. For LLVIP (1280×1024), M^{3}FD (1024×768), and PBVS 2025 TISR Challenge Track 2 (640×448), we first apply a center crop to preserve the primary scene content, followed by resizing to 512×512. We apply these preprocessing steps to both the RGB and thermal images. Thermal GT images are further downsampled to 64×64 and corrupted with Gaussian noise to serve as input. RGB images are processed with our misalignment augmentation while retaining a resolution of 512×512 as reference images.

2. Metrics Selection

We evaluate model performance using both reference and no-reference image quality metrics. For reference evaluation, we adopt PSNR and SSIM [7] to measure reconstruction fidelity, and LPIPS [10] for perceptual similarity. For no-reference evaluation, we report MUSIQ and MANIQA to assess overall visual quality. All metrics are computed using the IQA-PyTorch (pyiqa) package. Specifically, we use the `musiq-spaq` and `maniqa-pipal` models as implemented in IQA-PyTorch.

3. More Qualitative Results

We provide additional qualitative results on public datasets in Fig. S1, as well as on our smartphone dataset in Fig. S2. Our 3M-TI model produces thermal images with sharper, more realistic, and visually faithful details.

For example, 3M-TI successfully reconstructs the traffic cone (Row 3 in Fig. S1) and preserves the circular shape of the sign, which is severely degraded in the input thermal image (Row 3 in Fig. S2). By comparison, OSEDiff [8] and SeeSR [9] frequently introduce unrealistic artifacts, while CoReFusion [3], SwinFuSR [1], and SwinPaste [11] tend to yield blurrier results.

4. More Results on Object Detection

We present additional thermal object detection results in Fig. S3. In Row 1, 3M-TI delivers the most accurate de-

† These authors have contributed equally to this work.
* Corresponding author. Mail: yuanxiaoyun@sjtu.edu.cn.

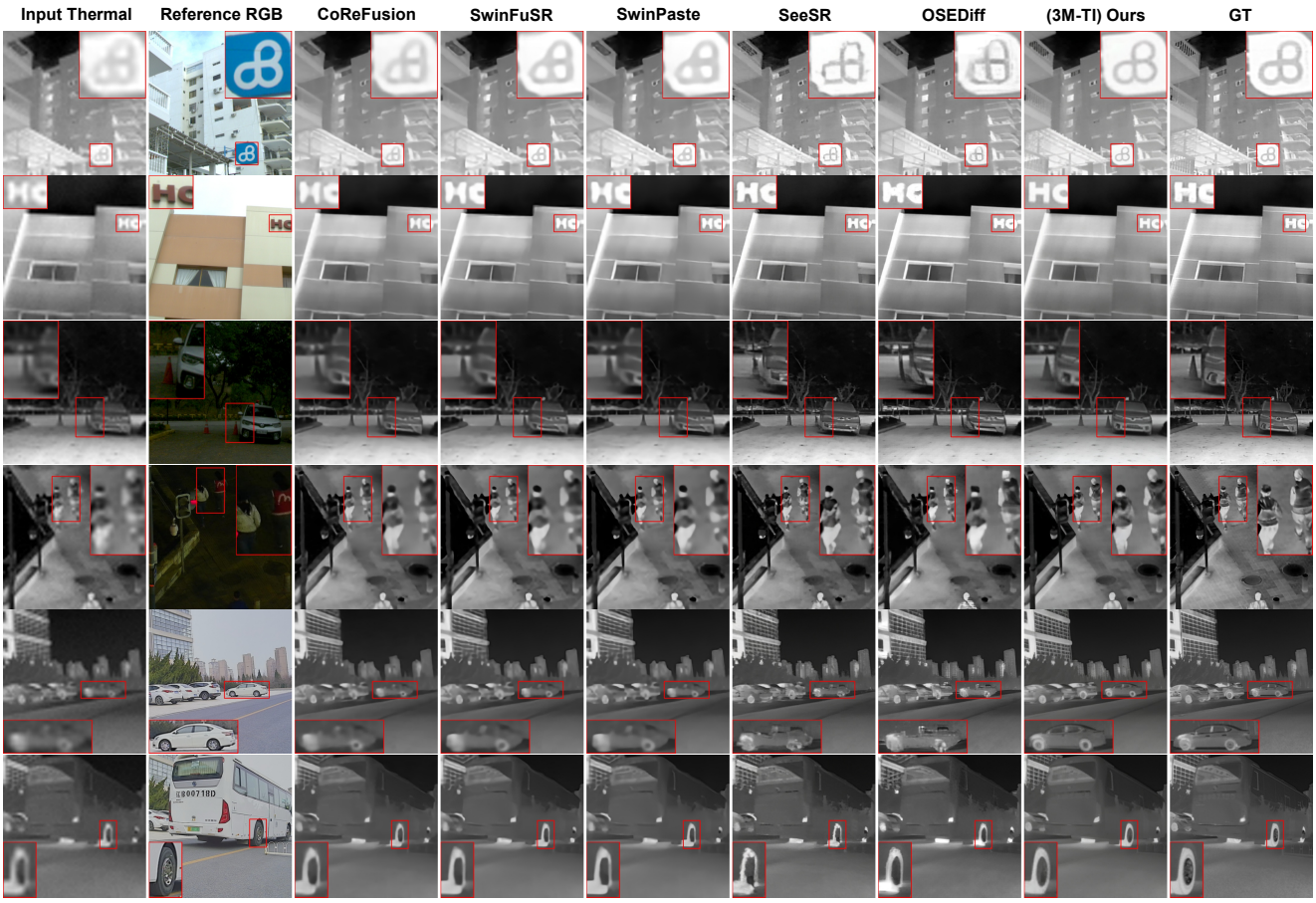

Figure S1. Qualitative comparison on our test set (zoom in for details). 3M-TI produces the most faithful and visually consistent results, delivering sharp structures and accurate thermal patterns that align closely with the GT. Representative examples include the sign (Row 1), English characters (Row 2), the traffic cone and car (Rows 3 and 5), and the pedestrians (Row 4).

tection performance, whereas the RGB image and SeeSR introduce 1 and 3 false positives, respectively. SwinPaste fails to detect any individual. Moreover, the reconstructed thermal and RGB detections exhibit complementary characteristics. In Row 2, the RGB image successfully separates two heavily overlapping individuals but yields 1 false positive, while the reconstructed thermal image of 3M-TI tends to merge them into a single entity but produces no false alarms (the positions of the two overlapping individuals are indicated in the figure).

5. More Results on Semantic Segmentation

We present additional thermal semantic segmentation result in Fig. S4. The prompts are "automobile, road". 3M-TI produces the most accurate and comprehensive segmentation map, even outperforming the map produced by the RGB reference. While RGB misses the middle automobile, SeeSR and SwinPaste fail to segment the two automobiles occluded by the front automobile.

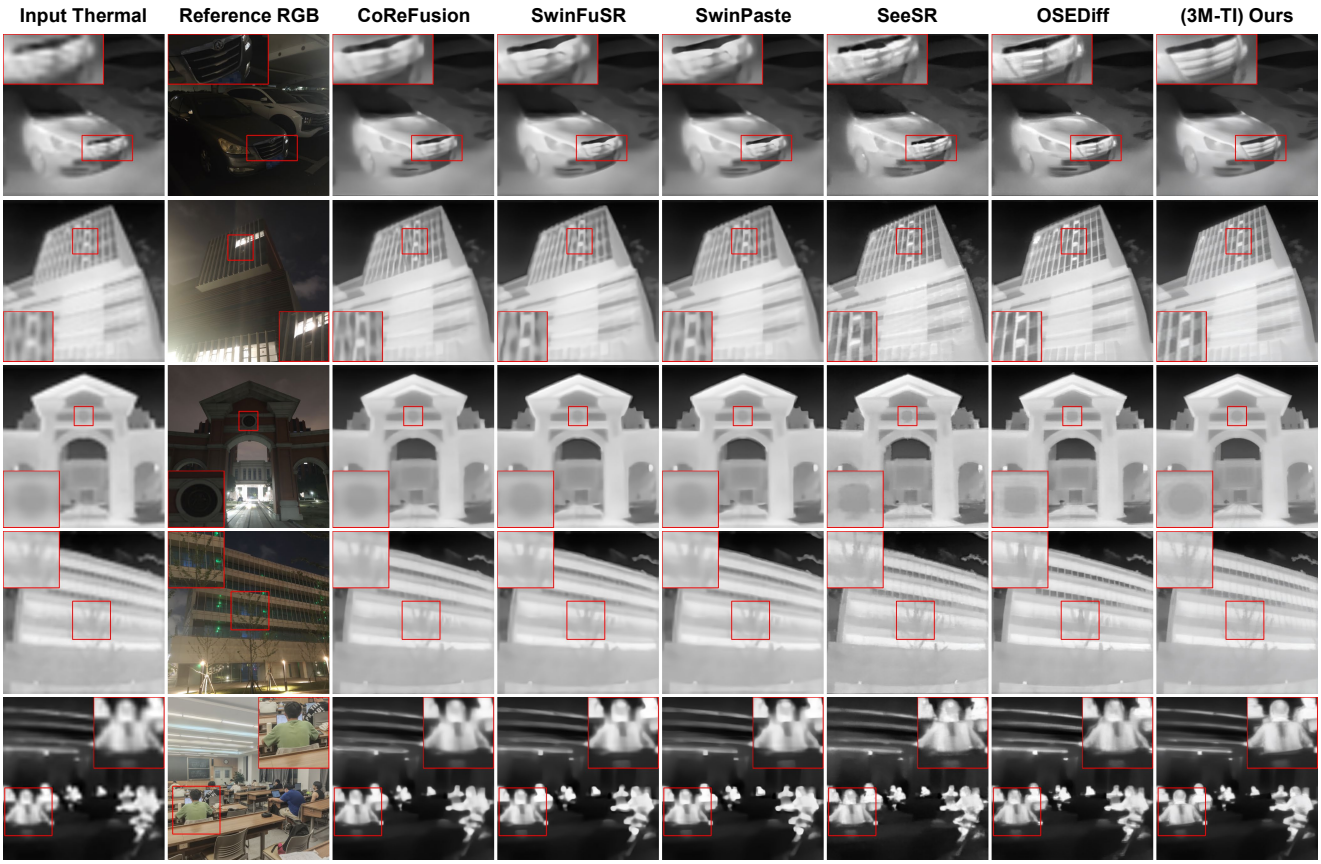

Figure S2. Qualitative comparison on our smartphone dataset (zoom in for details). 3M-TI demonstrates strong generalization ability, producing sharp and faithful thermal details that remain highly consistent with the corresponding RGB images.

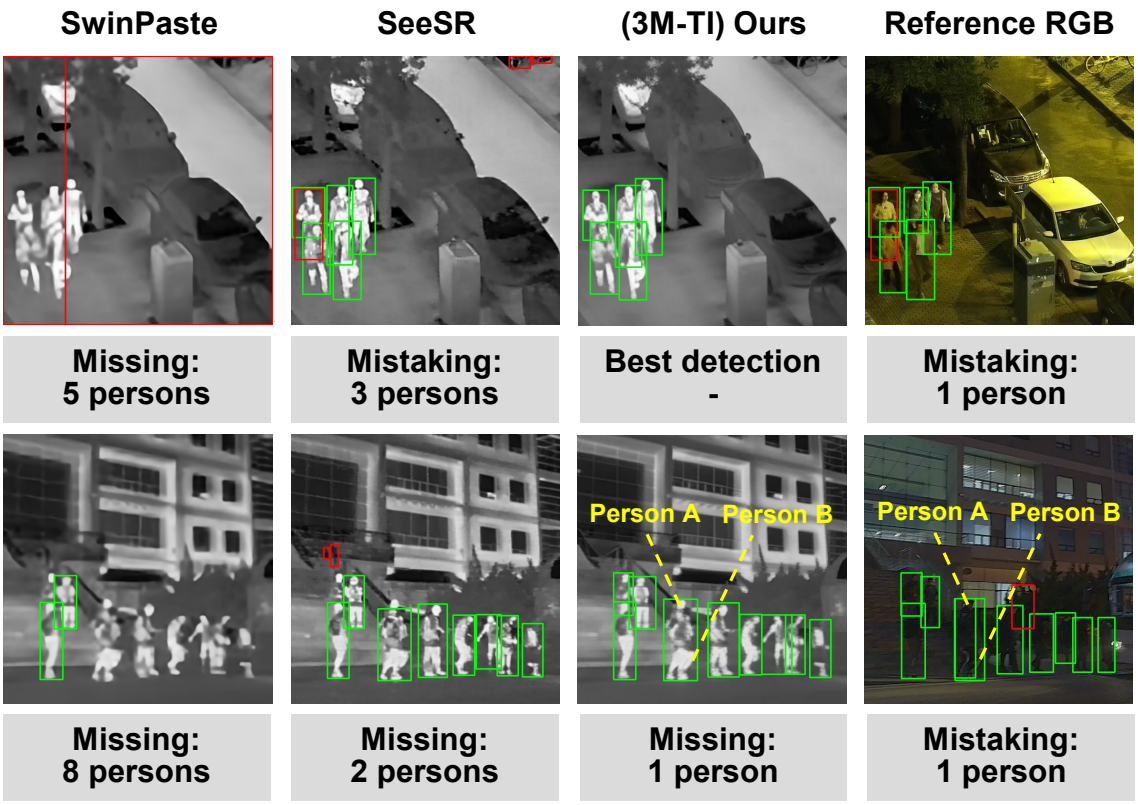

Figure S3. Visualization of detection results, where green bounding boxes denote correct detections and red bounding boxes denote incorrect ones. In Row 1, 3M-TI yields the most accurate detection results, whereas the RGB image and SeeSR introduce 1 and 3 false positives, respectively. SwinPaste fails to detect any individual. In Row 2, the RGB image successfully separates heavily overlapping individuals but yields 1 false positive, while the reconstructed thermal image of 3M-TI tends to merge them into a single entity but produces no false alarms.

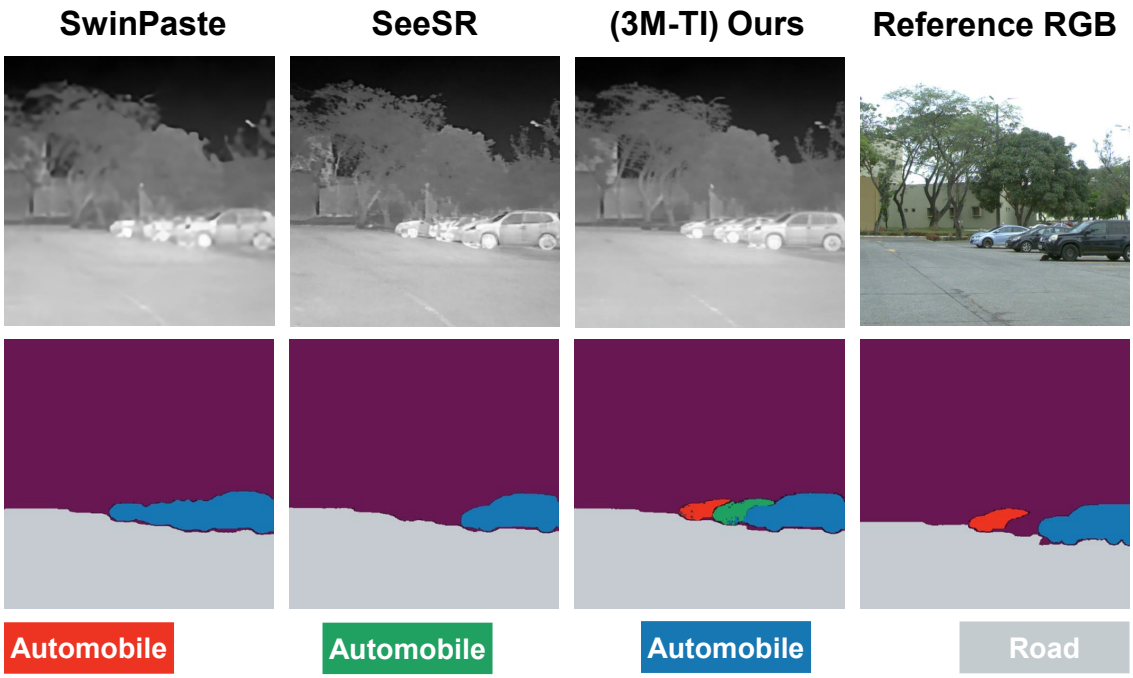

Figure S4. Visualization of segmentation results. 3M-TI produces the most accurate and comprehensive segmentation map, even outperforming the map produced by the RGB reference. While RGB misses the middle automobile, SeeSR and SwinPaste fail to segment the two automobiles occluded by the front automobile.

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. 1

[2] 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. 1

[3] 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. 1

[4] 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. 1

[5] 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. 1

[6] 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. 1

[7] 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. 1

[8] 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. 1

[9] 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. 1

[10] 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. 1

[11] 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. 1