Thermal Diffusion Matters: Infrared Spatial-Temporal Video Super-Resolution through Heat Conduction Priors

Supplementary Material

A. Overview

This supplementary material is organized as follows. § B offers additional details about the *IRVAL* dataset. § C outlines the explicit implementation details of the experiments. § D provides additional ablation studies. Further information on infrared object detection experiments and qualitative comparisons is presented in Section E. § F summarizes the computational efficiency of THERIS with detailed FLOPs, parameter counts, and runtime comparisons. Additional visualization results of super-resolution are presented in § G. Finally, we discuss the limitations of our approach in § H.

B. IRVAL Dataset

We introduce IRVAL, a large-scale, high-resolution infrared video dataset specifically designed to benchmark and advance spatiotemporal super-resolution methods in the infrared domain. IRVAL comprises 108,512 high-quality frames captured at 512 × 512 resolution using vanadium oxide (VOx) uncooled focal plane array detectors operating in the long-wave infrared (LWIR) band (8–14 μm). These detectors provide rich thermal contrast information while maintaining fine spatial detail, making IRVAL highly representative of real-world infrared imaging conditions.

To ensure broad scene diversity and realistic motion dynamics, our dataset includes video sequences captured from both vehicle-mounted and fixed surveillance platforms. The vehicle-mounted cameras traverse urban thoroughfares, residential areas, and commercial districts at varying speeds, capturing dynamic interactions among vehicles, cyclists, and pedestrians under varying traffic conditions. Fixed cameras are strategically positioned to monitor busy intersections, building entrances, and roadside infrastructure.

All frames in our dataset are provided in lossless PNG format, with example video frames showcased in Figure 2. We hope that **IRVAL** will serve as a valuable resource for advancing development in infrared video processing.

C. More Implementation Details

We perform $\times 4$ spatial and $\times 2$ temporal SR tasks. Specifically, the HR frames are downsampled by a factor of 4 using bicubic interpolation. The resulting frames, specifically the odd-numbered frames (e.g., $1^{st}, 3^{rd}, \cdots$), are used as inputs to train the model, and the corresponding HR frames $(1^{st}, 2^{nd}, 3^{rd}, \cdots)$ are used as supervision.

In TDIM, we stack three temporal layers, with the first two layers preserving the time axis. This can be efficiently implemented by modifying the formulation of $\mathrm{CT}_{\mathrm{out}}[n, t]$. The final temporal operator expands the temporal dimension, enabling physically grounded frame interpolation directly within the learned feature space. In TSSM, we adopt the bidirectional scan strategy [14] to accommodate the characteristics of video data. Different scanning axis orders induce varying neighborhood conditions, resulting in different levels of feature connectivity. So we employ three distinct scan orders within each TSSM: *space-first scan*, *time-first scan*, and *Hilbert scan*, as illustrated in Figure 1. The space-first scan traverses the data in a raster order along the *width-height-time* axes, while the time-first scan proceeds along the *time-width-height* order. The Hilbert scan leverages a space-filling curve to better preserve spatial locality when flattening high-dimensional data for infrared data.

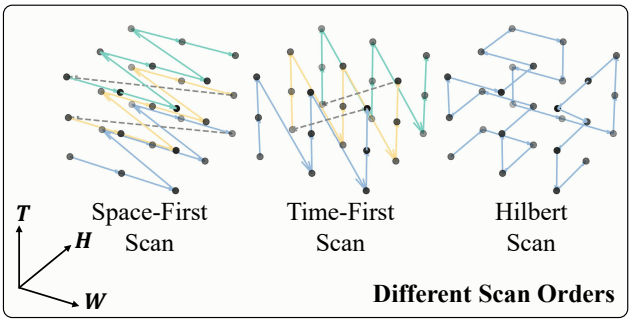

Figure 1. Different scan orders within each TSSM: *space-first scan*, *time-first scan*, and *Hilbert scan*.

The proposed THERIS model is trained using Adam optimizer and decoupled weight decay, setting the hyperparameters to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 2×10^{-4} and is gradually reduced to 10^{-7} following the scheme of cosine annealing. The entire training is conducted on an NVIDIA RTX 4090 GPU.

D. More Ablation Studies

Rationale for Frequency Masks in TSSM. As illustrated in Figure 3, applying a high-pass filter to both the original HR infrared image and a simply upsampled LR counterpart reveals a pronounced degradation of high-frequency information in the latter. To mitigate this issue, each TSSM module incorporates a learnable spectral mask. Figure 4 shows the log-amplitude difference of feature maps with and without the proposed spectral mask. Specifically, the $\triangle$ log amplitude of signals is computed as the difference between the log amplitude at normalized frequency 0.0π (center)

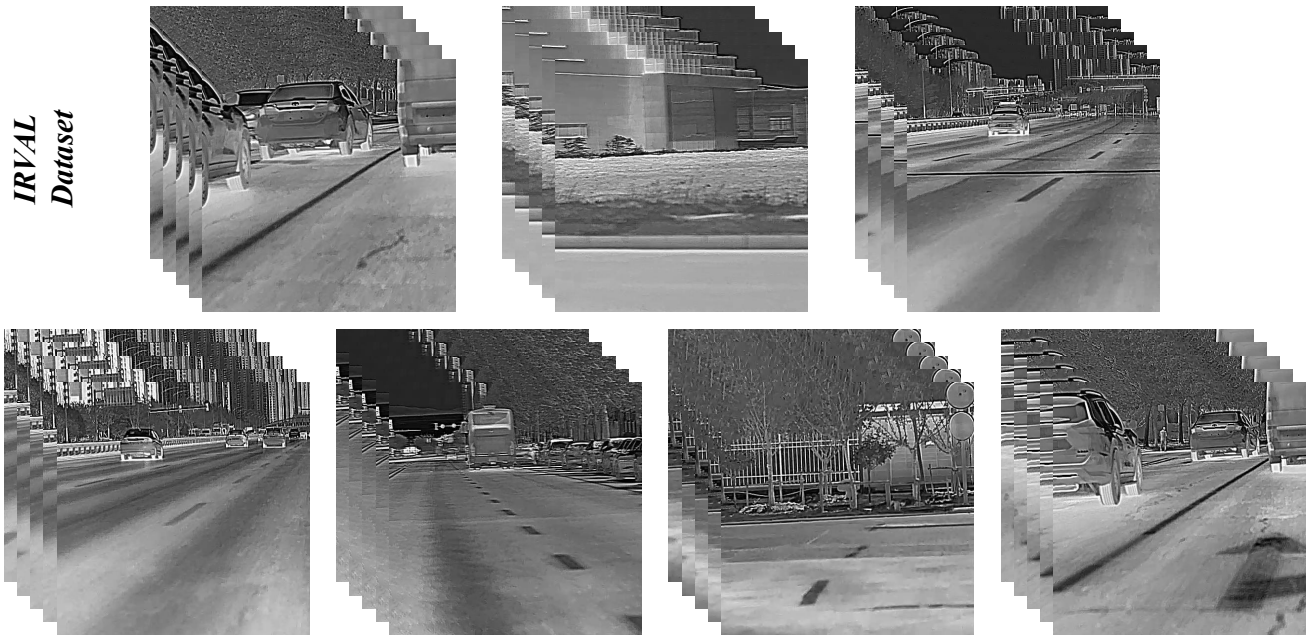

Figure 2. An overview of our proposed IRVAL dataset.

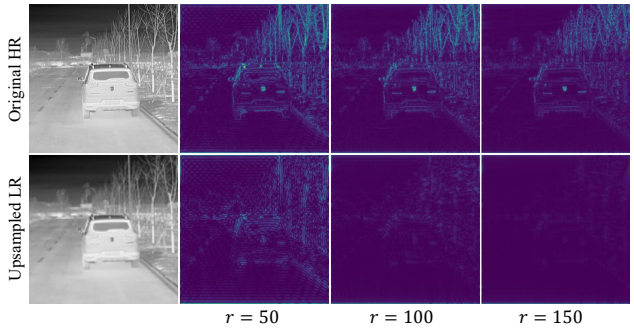

Figure 3. We conduct high-pass operations to infrared images in the Fourier Spectrum domain and then reverse the FFT process. r stands for the radius of the filter mask.

and 1.0π (boundary). The visualization in Figure 4 clearly demonstrates that the learnable frequency filter in TSSM effectively enhances high-frequency components, thereby recovering fine spatial details in infrared video frames.

The Scan Mechanism. To capture diverse spatial and temporal dependencies, TSSM integrates multiple Selective State Space Blocks following the learnable spectral filter, each employing distinct scan orders: space-first, time-first, and local Hilbert scan [10], corresponding to the Space Mamba Block (SMB), Time Mamba Block (TMB), and Hilbert Mamba Block (HMB), respectively. We further evaluate the impact of these scan strategies by conducting an ablation study in which one scan order is removed at a time, while maintaining the total number of state space transformations by proportionally increasing the others. As shown in Table 1, removing any individual scan type leads

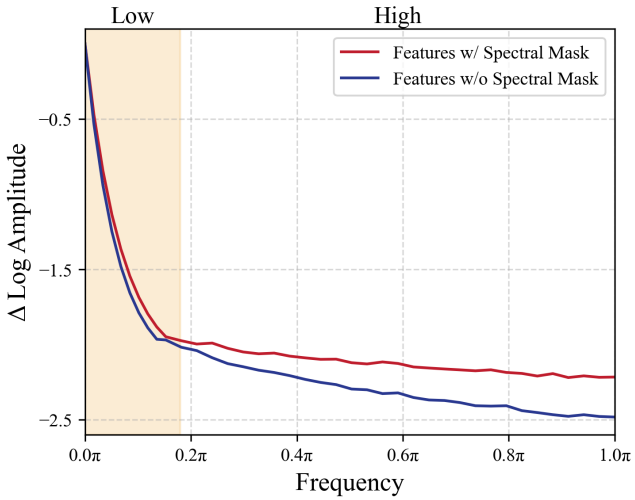

Figure 4. Relative log amplitudes of Fourier transformed feature maps w/ and w/o our proposed spectral mask.

to a performance drop. This confirms that hybridizing different scan orders enhances model performance by offering more complementary spatiotemporal connectivity.

E. Experiments on Infrared Object Detection

Experiments Setup. To assess the perceptual and structural fidelity of super-resolved frames in a task-oriented manner, we evaluate object detection performance as a downstream application. Specifically, we employ the YOLOv8 detector [6], pre-trained on the COCO dataset [8] and fine-tuned on the LLVIP dataset, which provides high-quality

Table 1. Ablation studies on different scan orders. BD denotes bidirectional scan, SF, TF, and LH denote space-first, time-first, and local Hilbert scan, respectively.

BD	SF	TF	LH	PSNR↑	SSIM↑	MUSIQ↑
✓	✗	✓	✓	20.68	0.7794	53.81
✓	✓	✗	✓	20.94	0.7758	52.26
✓	✓	✓	✗	20.87	0.7770	52.13
✗	✓	✓	✓	20.57	0.7728	51.91
✓	✓	✓	✓	21.37	0.7872	55.59

bounding-box annotations for pedestrian instances across diverse and challenging street-scene environments.

Super-resolved video frames generated by different methods are input into the YOLO detector under identical inference settings. The evaluated baselines include a combination of VFI models—SuperSloMo [5], DAIN [1], and VFIMamba [13]—as well as VSR methods such as Bicubic Interpolation, EDVR [9], and RealBasicVSR [2]. Furthermore, we incorporate several STVSR approaches, including ZoomingSlowMo [11], TMNet [12], VideoINR [4], MoTIF [3], and BF-STVSR [7]. By keeping the detection model fixed across all evaluations, we ensure that any variations in detection performance can be attributed solely to differences in the visual quality of the super-resolved frames. The detection performance is assessed using the mean Average Precision (mAP) metric across a range of Intersection-over-Union (IoU) thresholds (mAP@0.5:0.95).

By evaluating object detection performance on super-resolved outputs, we demonstrate that our thermodynamics-guided approach improves not only conventional pixel-level metrics (e.g., PSNR, SSIM) but also enhances performance in downstream vision tasks—producing frames that are both perceptually sharper and semantically more informative.

Quantitative Comparisons. Table 2 presents a quantitative comparison of object detection performance across various methods. Since higher-quality, detail-preserving super-resolved videos contribute to improved detection accuracy, this evaluation serves as an indirect yet effective measure of visual enhancement. THERIS achieves the highest detection score of 50.7, significantly outperforming the upsampled LR frames, which only achieve a score of 43.8. This improvement further underscores the effectiveness of THERIS in enhancing the visual quality of infrared videos.

Qualitative Comparisons. We further provide a qualitative comparison of object detection performance on the IRVAL dataset in Figure 5. Competing methods frequently miss detections or produce erroneous results. For instance, in the first row, most methods fail to detect the traffic lights located in the upper-right corner of the frame. In the second row, some methods overlook pedestrians near the track in the center, while all fail to identify the distant truck. In contrast, our proposed method consistently delivers the most accurate and complete detection results.

Table 2. Comparison of detection results on the LLVIP dataset.

VFI Method	VSR Method	mAP↑
SuperSloMo [5]	Bicubic	45.9
SuperSloMo [5]	EDVR [9]	44.9
SuperSloMo [5]	RealBasicVSR [2]	45.2
DAIN [1]	Bicubic	45.3
DAIN [1]	EDVR [9]	45.0
DAIN [1]	RealBasicVSR [2]	45.7
VFIMamba [13]	Bicubic	45.9
VFIMamba [13]	EDVR [9]	45.3
VFIMamba [13]	RealBasicVSR [2]	46.9
ZoomingSlowMo [11]		47.1
TMNet [12]		46.5
VideoINR [4]		45.8
MoTIF [3]		45.5
BF-STVSR [7]		47.4
Upsampled LR		43.8
THERIS		50.7

Table 3. Comparison of efficiency and performance across different VFI/VSR combinations and STVSR methods.

VFI Method	VSR Method	Params (M)	FLOPs (G)	Time (ms)	PSNR↑
SuperSloMo [5]	RealBasicVSR [2]	56.81	212.57	149.90	19.67
DAIN [1]	EDVR [9]	44.57	364.91	209.97	20.15
VFIMamba [13]	RealBasicVSR [2]	103.06	270.92	190.02	20.50
ZoomingSlowMo [11]		11.10	302.33	182.51	19.90
TMNet [12]		12.26	376.51	208.04	19.78
VideoINR [4]		11.31	417.68	232.68	19.94
MoTIF [3]		12.55	449.31	251.31	19.58
BF-STVSR [7]		13.47	419.11	192.11	20.18
THERIS		36.16	225.10	164.88	21.37

F. Computational Costs

Table 3 compares the computational efficiency and reconstruction performance of recent VFI/VSR and STVSR pipelines. Our proposed THERIS achieves the best overall performance, obtaining 21.37 dB PSNR, which surpasses all compared methods by a significant margin, while maintaining relatively low complexity (36.16M parameters and 225.10G FLOPs). Its inference speed (164.88 ms) is also faster than most two-stage frameworks and competitive with existing STVSR methods, demonstrating a superior balance between performance and efficiency.

G. More Visualization Results

Additional qualitative comparisons on the IRVAL, LLVIP, and SGMP test sets are presented in Figure 6 and Figure 7. Across all datasets, our proposed THERIS method consistently delivers superior visual quality, exhibiting enhanced fine-grained details and improved structural clarity. We rec-

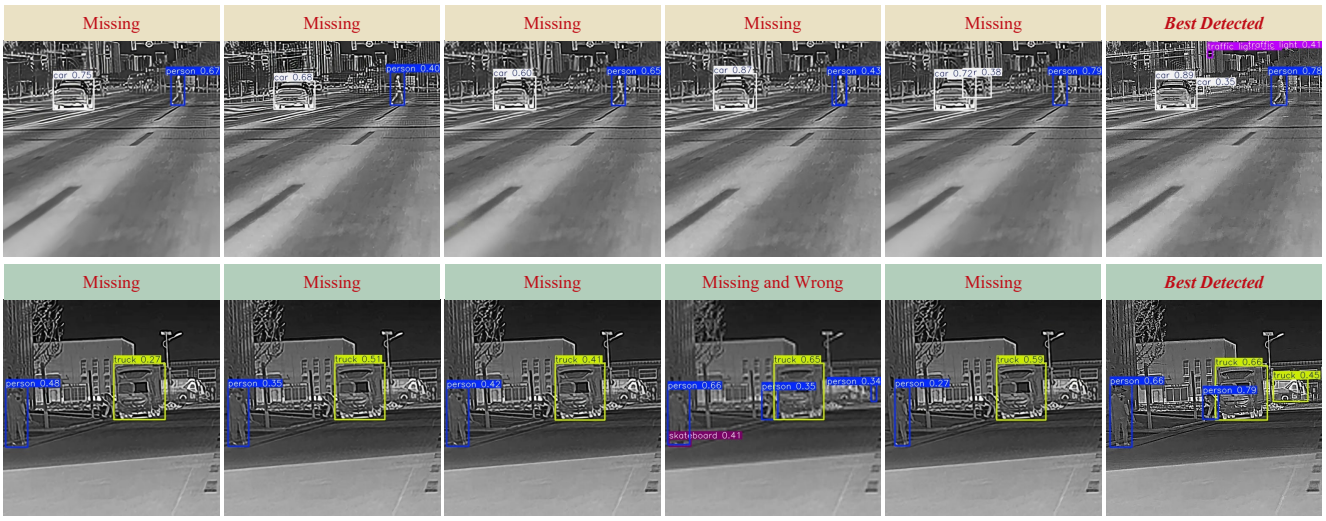

Figure 5. Visual comparison of object detection performance on infrared super-resolved frames.

ommend zooming in for the optimal visualization.

H. Discussion and Limitation

In this paper, we propose THERIS, a thermodynamics-inspired framework tailored for infrared spatiotemporal video super-resolution. Specifically, the *Thermal Diffusion Interpolation Module* models feature sequences as one-dimensional heat fields, enabling temporally coherent frame alignment and synthesis. On top of this, the hierarchical *Thermo-Aware State Space Module* and a dedicated *Temperature Field Modeling Loss* introduce physics-guided constraints that enhance spatial sharpness and temporal consistency. Experimental results demonstrate that our method achieves SOTA performance across multiple benchmark datasets, validating its utility and effectiveness.

Despite the growing interest in video SR, prior research has largely overlooked the spatiotemporal enhancement of infrared data. In the visible domain, several continuous STVSR methods [3, 4, 7] have been proposed, enabling interpolation at arbitrary spatial and temporal resolutions. In contrast, our proposed THERIS framework represents the preliminary dedicated attempt at addressing infrared STVSR. The current design is constrained to fixed scaling factors defined during training, limiting its flexibility in real-world scenarios. Extending our THERIS framework to support continuous spatiotemporal SR for infrared video represents a promising direction for future research.

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *CVPR*, pages 3703–3712, 2019. 3

[2] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, pages 5962–5971, 2022. 3

[3] Yi-Hsin Chen, Si-Cun Chen, Yen-Yu Lin, and Wen-Hsiao Peng. Motif: Learning motion trajectories with local implicit neural functions for continuous space-time video super-resolution. In *CVPR*, pages 23131–23141, 2023. 3, 4

[4] Zeyuan Chen, Yinbo Chen, Jingwen Liu, Xingqian Xu, Vidit Goel, Zhangyang Wang, Humphrey Shi, and Xiaolong Wang. Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In *CVPR*, pages 2047–2057, 2022. 3, 4

[5] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, pages 9000–9008, 2018. 3

[6] Glenn Jocher et al. Ultralytics yolov8, 2023. 2

[7] Eunjin Kim, Hyeonjin Kim, Kyong Hwan Jin, and Jaejun Yoo. Bf-stvsr: B-splines and fourier—best friends for high fidelity spatial-temporal video super-resolution. In *CVPR*, pages 28009–28018, 2025. 3, 4

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2

[9] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, pages 0–0, 2019. 3

[10] Hongtao Wu, Yijun Yang, Huihui Xu, Weiming Wang, Jinni Zhou, and Lei Zhu. Rainmamba: Enhanced locality learning with state space models for video deraining. In *ACM MM*, pages 7881–7890, 2024. 2

[11] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and

accurate one-stage space-time video super-resolution. In *CVPR*, pages 3370–3379, 2020. 3

[12] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In *CVPR*, pages 6388–6397, 2021. 3

[13] Guozhen Zhang, Chuxnu Liu, Yutao Cui, Xiaotong Zhao, Kai Ma, and Limin Wang. Vfimamba: Video frame interpolation with state space models. In *NeurIPS*, pages 107225–107248, 2024. 3

[14] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: efficient visual representation learning with bidirectional state space model. In *ICML*, 2024. 1

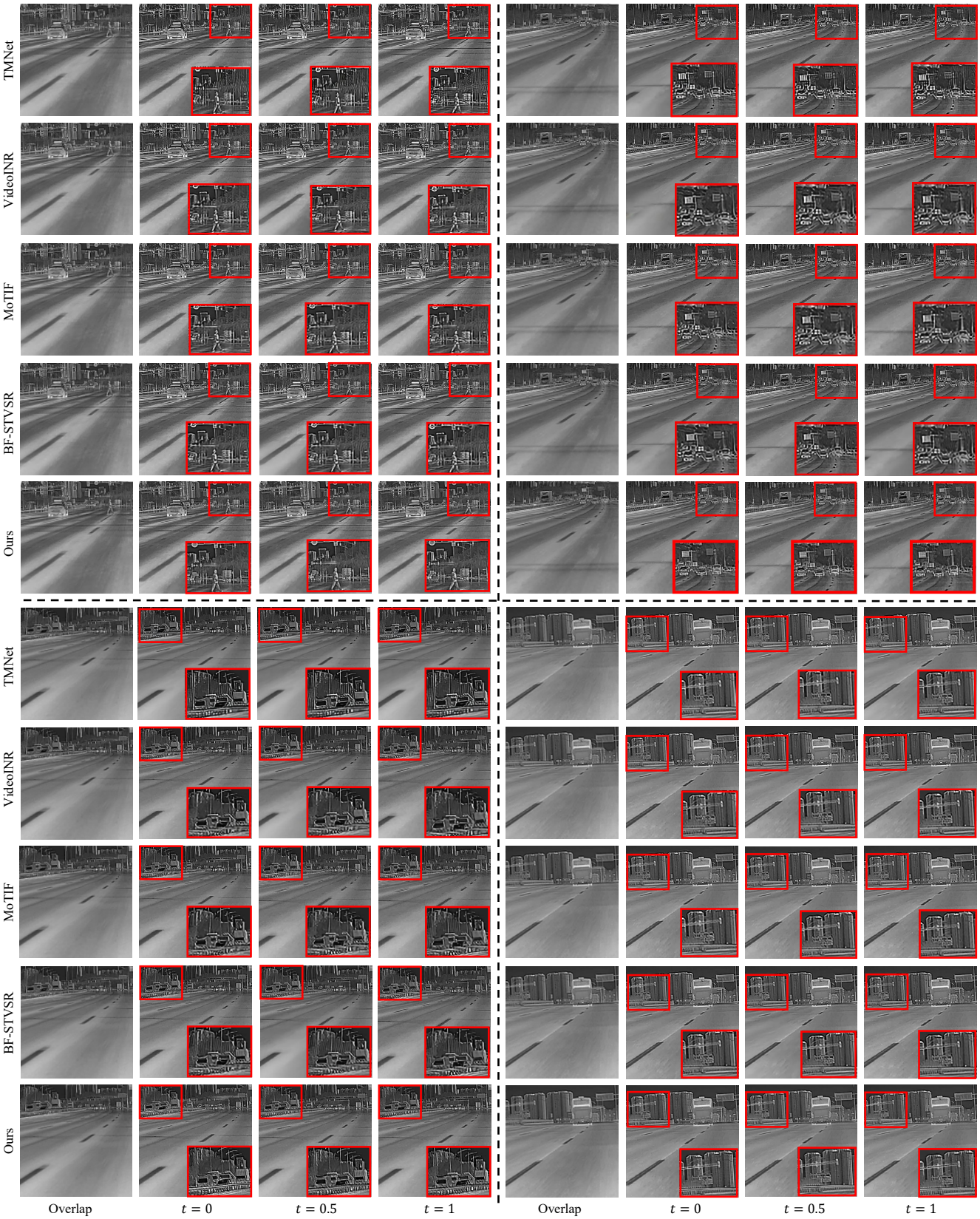

Figure 6. More qualitative comparisons with different methods on the IRVAL test set. Please zoom in for details.

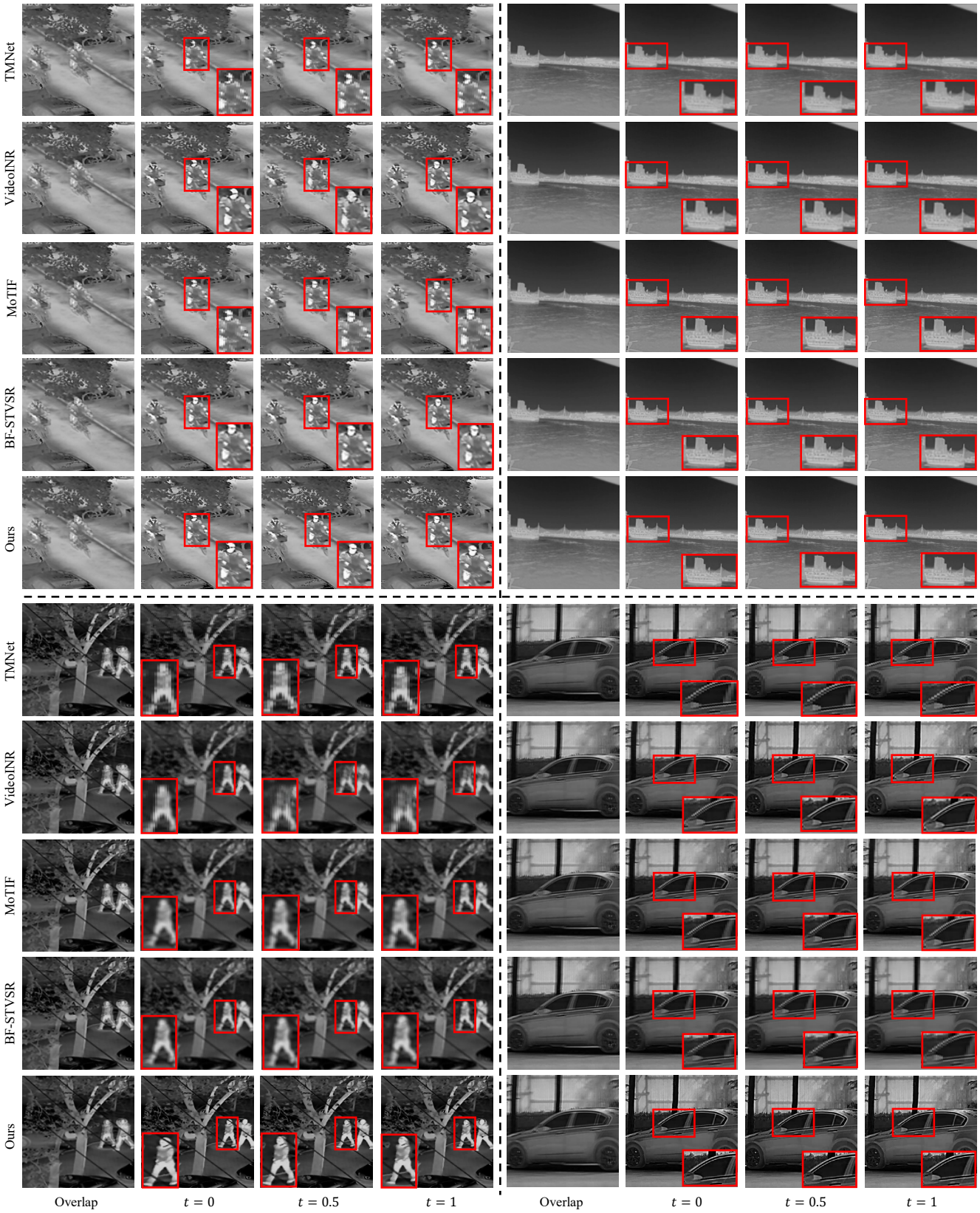

Figure 7. More qualitative comparisons with different methods on LLVIP and SGMP test sets. Please zoom in for details.