Supplementary Materials

A. Overview

- In Sec. B we provide training details for training densification and fusion modules.
- In Sec. C we provide more analysis, including effects of thresholds, choices for hyper-parameters, comparison with 3D re-projection based baseline with estimated depth and pose, and effects of 3D reconstruction from large models such as VGGT and MapAnything.
- In Sec. D we provide more evaluations following the main experiments.
- In Sec. E we further experiment on RGB-Normal map as a style to transfer between views.

B. Training Details

Details of D and F training. To train D, we use RGB image and downsample its paired X-image by random sparsity as the input that uniformly samples a density within [0.05, 0.85]. The low density simulates situation such as larger homogeneous areas, and the high density simulates situations with more matched keypoints. We combine this uniform sampling with ORB and SIFT keypoints extraction on the X-images. The original X-images is the groundtruth. We use a batch size of 4, an image size of 512x512, a learning rate of 0.001, and an epoch number of 24 on a NVIDIA RTX-4090 GPU. We adopt losses that are combined with commonly used L1, Laplacian [13], and gradient matching [5, 8], which are widely used in depth estimation and completion tasks. Then, we freeze D during the test time.

F is pretrained on image enhancement with lightweight architecture [12]. During the test time, we perform test-time training (TTT) to train F to fuse multi-level X-images. Self-supervised losses with RGB images are computed, including the cosine similarity loss and self-matching loss. For each modality, F is trained for 100 steps for the fusion step initialization, with data randomly drawn from the pool. Then a 2-step update is applied to each inference sample. A learning rate of 0.00006 is used for this high-dimensional task.

We keep the larger size D as a frozen foundation, which is unaffected by specific test scenes and still maintains test-time efficiency. Meanwhile, we update lightweight F to efficiently adopt test-time information to better synthesize X-views built upon D's performance.

RGBX-3DGS. For training RGBX-3DGS, we follow the official and most general 3DGS setting and add X-channels in the CUDA kernels in addition to RGB channels. We keep one set of all the other GS parameters, such as scale, rotation, or opacity. This is different from [4], where they use disentanglement to have one set of scale, rotation, and opacity for RGB and the other set for another modality (See paper Section 3.3)

The learning rate for the additional X-channel is set to 0.0025, which is the same as the RGB channel, and the rendering follows the alpha composition from RGB-channels and X-channels separately to render RGB-images and X-images. The hyper-parameters such as iterations, density control, and learning rates follow the official settings. Note that there are 3DGS methods tailored for individual modalities, such as RGB-Thermal 3DGS [7], Thermal-only 3DGS [2], or RGB-NIR [3].

In contrast, we retain the original formulation of 3DGS, as our goal is to address general RGB-X modalities rather than design yet another modality-specific variant. Our novelty lies in developing a scalable framework, built upon the proposed match–densify–consolidate paradigm, to enable cross-sensor view synthesis. We are the first scalable work for this practical but widely-overlooked task of aligning sensor pairs without requiring 3D priors of X sensors. Our purpose is not to build tailored and new 3DGS formulations for each individual modality.

C. More Analysis

Analysis on Impact from Thresholds. As mentioned in the paper Section 3.2, a higher threshold yields sparser initial points and needs to rely on RGB cues for densification, whereas a lower one yields more points as initialization but they are noisier. We show examples in Fig. S2 to describe the observations. Using the fusion module and trained with the given self-supervised losses with respect to RGB images, our proposed F successfully preserves the advantages of both, showing details, smoother regions, and less noise.

Study on the Hyper-Parameters. We experiment with different hyper-parameter choices and show results in Fig. S1 using RGB-NIR-Stereo 09-28-16-48-17/1, including the number of levels K, the number of frames N, and the density in area sampling. For K, $K = 3$ attains similar scores with more levels used and becomes the choice. A lower K value is not efficient for utilizing the advantages of multi-level confidence as mentioned earlier. For N, $N = 7$ attains the best score. Specifically, using more frames risks matching across low-overlap views, leading to more noisy matches, and thus $K = 9$ drops the score.

For density in area sampling, we also find 5% is a good spot for peak performance, and a higher sampling rate performs slightly worse due to errors from homography warping. Further, based on main paper Tab. 5's last entry, we add an experiment to control sampling density of pixels and report PSNR. 0%: 16.45, 1%:18.70, 2%: 18.91, 5%: 19.22, 10%: 18.88, where a 5% rate seeds enough points.

Compare to 3D Reprojection by Estimated Pose and Depth. Here we provide more details about experiments on 3D reprojection-based method. Given unpaired RGB and thermal images, we first use DepthAnythingV2 to estimate depth, MINIMA's cross-modal matcher to match keypoints

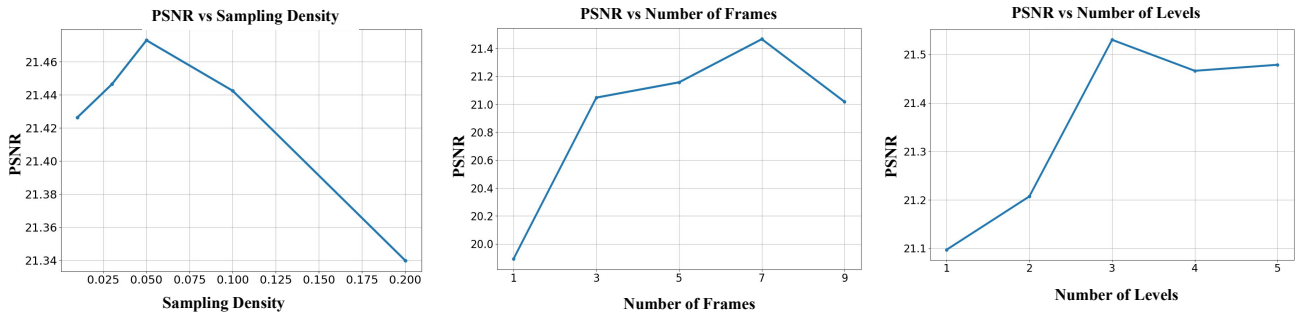

Figure S1. **PSNR Changes with Hyper-Parameters.** We study the hyper-parameter choices and PSNR changes, while controlling all the other factors.

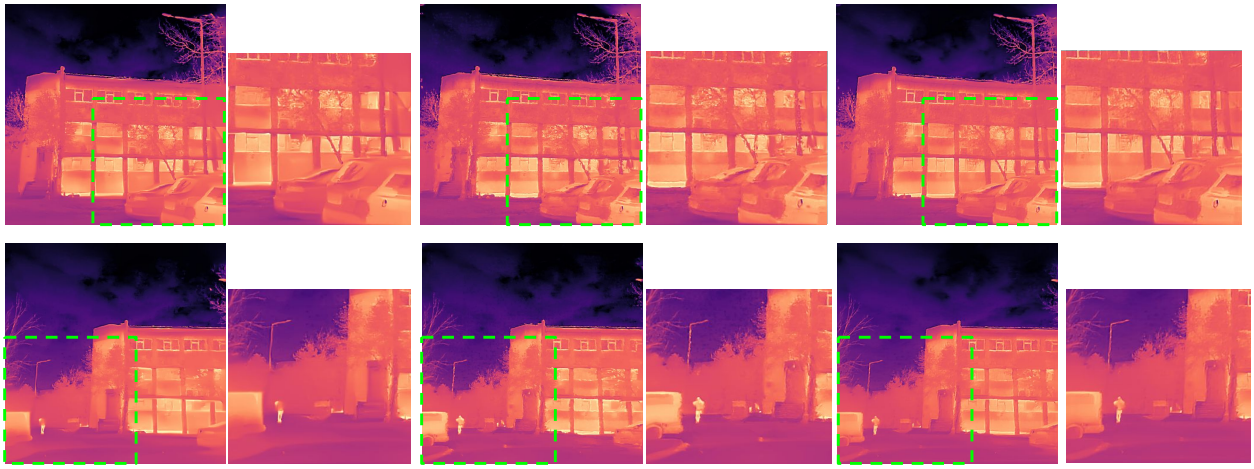

Figure S2. **Effects of Different Thresholds.** Higher thresholds yield less noise but also fewer details; lower thresholds yield more details but also a higher noise level. Our fusion strategy combines the advantages across thresholds and shows better quality.

and compute relative pose between RGB and thermal images, and intrinsics from the dataset's groundtruth.

We experiment with two settings to cover depth estimation from thermal images directly or from RGB images. The first (Setting A) is to estimate depth from RGB's view and reproject RGB's view onto thermal camera's view space by the following relation.

$$\begin{bmatrix} u_{\mathrm{T}} \\ v_{\mathrm{T}} \\ 1 \end{bmatrix} = K_{\mathrm{T}} \left(R_{\mathrm{RGB}\to\mathrm{T}} \cdot K_{\mathrm{RGB}}^{-1} \cdot D_{\mathrm{RGB}} \begin{bmatrix} u_{\mathrm{RGB}} \\ v_{\mathrm{RGB}} \\ 1 \end{bmatrix} + t_{\mathrm{RGB}\to\mathrm{T}} \right),$$

where $[u_{\mathrm{RGB}}, v_{\mathrm{RGB}}, 1]$ and $[u_{\mathrm{T}}, v_{\mathrm{T}}, 1]$ mean homogeneous coordinates for RGB and thermal images, $K_{\mathrm{RGB}}, K_{\mathrm{T}}$ mean intrinsics for RGB and thermal cameras from the dataset, which also consider lens distortion, R and t represent estimated relative rotation and translation between two views, and D_{RGB} is the estimated depth map. We infill the hole with the nearest neighbors after re-projection.

We show visualization and error analysis in Fig. S3. Different errors in depth estimation are marked, including wrong depth by comparing the neighboring regions (Box A and B), faint and irregular structures (Box C), and also irregular sky areas. For the pose error, the translation RMSE is about 0.99m and the rotation error is about 0.8°. However, this slight pose error still results in a visible global misalignment in the reprojected RGB, which does not align well with the thermal view.

Then we use the reprojected depth map as depth for thermal images and reproject thermal images back to RGB's view by

$$\begin{bmatrix} u_{\mathrm{RGB}} \\ v_{\mathrm{RGB}} \\ 1 \end{bmatrix} = K_{\mathrm{RGB}} \left(R_{\mathrm{T}\to\mathrm{RGB}} \cdot K_{\mathrm{T}}^{-1} \cdot D_{\mathrm{T}} \begin{bmatrix} u_{\mathrm{T}} \\ v_{\mathrm{T}} \\ 1 \end{bmatrix} + t_{\mathrm{T}\to\mathrm{RGB}} \right).$$

Then we train with 3DGS for consolidation.

The second (Setting B) is to estimate depth from thermal images directly, and then we re-project thermal images with estimated depth back to the RGB views. We show a result in Fig. S4, where depth from thermal views is not reliable, as DepthAnything primarily operates in the RGB domain,

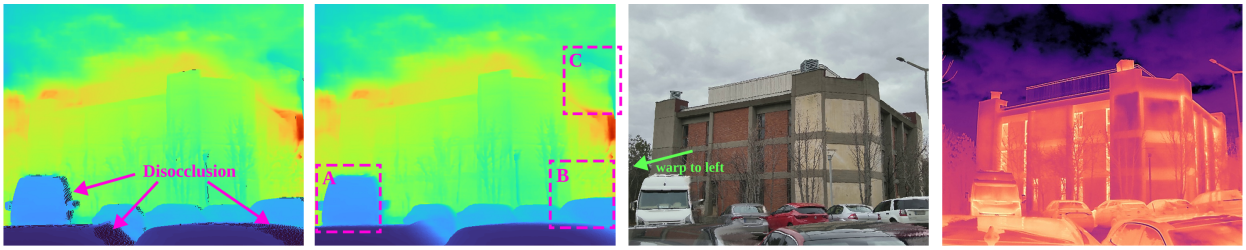

Figure S3. **3D Re-projection Artifacts: From RGB to Thermal**. Pseudo-color for the depth map captures 3-75m. We infill the disocclusion areas with the nearest values. Visible Depth Errors: In Box A, the van is roughly at the same depth as the neighboring car, but the colormap shows a visibly closer depth since the size is larger. (The metric depth for the car inside Box A is 16.1 m, and the depth for the car on its right-hand side is 20.4m. The difference is 4.3m, much larger than it should be from the parallel parking situation.) In Box B, the car should be slightly farther than the neighboring car, but it shows a closer depth (3.2m closer than its neighboring car on the left). In Box C, the pole structure cannot be reflected. Further, the sky areas are associated with wrong and irregular depth values. Visible Pose Error: The car is reprojected to the left-facing, but the groundtruth from the target thermal view is right-facing, and the projected view is misaligned globally with the target view, showing pose estimation error.

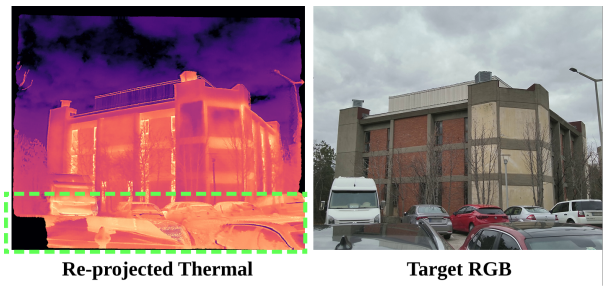

Figure S4. **3D Re-projection Artifacts: From Thermal to RGB**. Visible depth errors are shown in the car roof areas and the pole with skewness or larger shifts.

leading to larger errors such as a skewed pole and shifted car roofs. We also train 3DGS for consolidation in this setting.

Video or Multiview Depth Estimators. We show numerical and visual comparison on METU-VisTIR-Cloudy sc-1 in Tab. S1 and Fig. S5. Our performance shows much better results from both numerical and visual aspects. Low numerical values and poor rendering quality are affected by imprecise depth and pose.

To further strengthen the depth reprojection-based baseline, we replace the monocular DepthAnything with dense multiview stereo, MapAnything (MA), or video depth, VideoDepthAnything (VDA) methods. We use the networks to predict metric depth for each frame from multiviews, which potentially improves regions where monocular one struggles. Following the Setting A, we use the predicted RGB's depth and reproject the scene depth to the X-view and infill holes. Then, we back-project the X-view to RGB using the scene depth. We use METU-VIsTIR-Cloudy 6 scenes for evaluation and show the results in Table S2, where we also include results using groundtruth relative pose for reprojection. From the table, our results still outperform the

Table S1. Comparison with 3D reprojection-based settings on METU-VisTIR-Cloudy sc-1.

Method	Icos ↑	p30↑	p50↑	p70↑	p90↑	ITM↑	ITcos↑
Setting A	0.658	30.94	34.23	36.24	38.24	0.860	0.414
Setting B	0.651	30.22	33.39	35.35	37.71	0.648	0.393
Ours	**0.702**	**35.67**	**39.56**	**41.14**	**42.64**	**0.988**	**0.460**

Table S2. Comparison with 3D reprojection-based methods with depth from multi-view estimators.

Method	Icos ↑	p30↑	p50↑	p70↑	p90↑	ITM↑	ITcos↑
VDA	0.66	29.71	32.56	34.56	36.68	0.88	0.44
VDA (+ gt pose)	0.68	30.40	33.18	35.12	37.20	0.90	0.44
MA	0.66	29.42	32.06	34.20	36.38	0.87	0.42
MA (+gt pose)	0.67	30.09	32.89	34.91	36.72	0.88	0.44
Ours	**0.69**	**31.18**	**34.39**	**36.43**	**38.72**	**0.92**	**0.45**

those baselines with a large margin, even with groundtruth poses. The challenges of inexact depth for thin or cropped objects persists that cause object deformation, indicating getting precise metric depth from images is still challenging.

Foundation Model-based Reconstruction on RGB-X. We experiment with two foundation models, including VGGT and MapAnything. We feed in RGB and thermal images together, alternatively sampled with three frames as a step (e.g., frame-1 RGB, frame4-thermal, frame7-RGB, frame10-thermal, and so on.), and show the VGGT's 3D reconstruction in Fig. S6 and MapAnything's in Fig. S7. Both reconstructions have cleared out the sky areas. VGGT reconstructs each modality separately without aligning their features. MapAnything though can recognize cross-modal features and reconstruct a single building; it still suffers from large errors, such as unaligned and multiple layers for each object from each modality (e.g., two separated car layers from RGB and thermal views in Fig. S7).

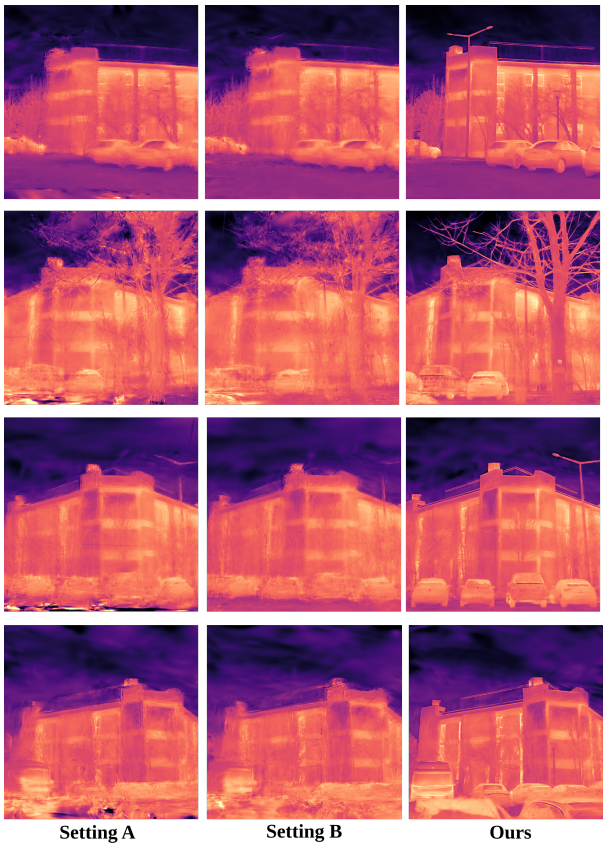

Figure S5. **Visual Comparison to 3D Re-projection Methods**. Our results show much better structures.

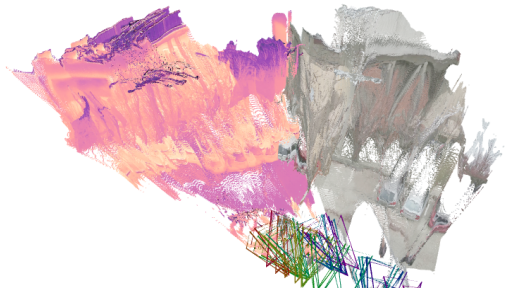

Figure S6. **3D Reconstruction by VGGT.** The reconstruction are separated without alignment.

D. More Results

RGBT-Scenes. We show the visual results for RGBT-Scenes in Fig. S8, which includes both train view and novel view results. Numerical results are in Table 3 of the paper. Our synthesis is closer to the groundtruth for both train and novel views.

Evaluation on Image Quality. Fig. 6 of the paper shows improved rendering quality with better thermal images. Here Table S3 also shows the numerical results on image quality metrics. Our RGB rendering also slightly outperforms others due to better thermal images involved in 3DGS training, as our side benefit.

More RGB-SAR Visualization. Following Fig. 8 in the paper, we show more RGB-SAR results in Fig. S10, where our SAR maps are closer to the groundtruth compared with other methods.

Source and Target View Visualization. We show the source current frame and the synthesized target frame in Fig. S9, to help understand how information propagate.

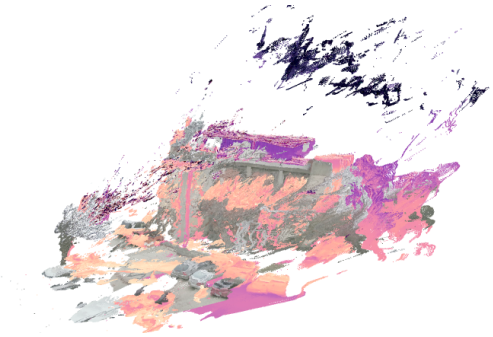

Figure S7. **3D Reconstruction by MapAnything.** The method suffers from large 3D reconstruction prediction errors, which can be seen in two separated car rows from RGB and thermal on the bottom.

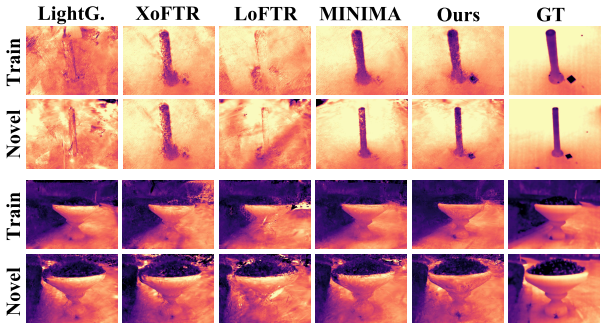

Figure S8. **Visual Comparison on RGBT-Scenes.** Ours shows closer structures to the groundtruth in the thermal domain.

Table S3. **RGB Image Quality on METU-VisTIR-Cloudy**.

Method	PSNR↑	SSIM↑	LPIPS↓
XoFTR [11]	26.838	0.833	0.170
LightGlue [6]	26.843	0.828	0.172
LoFTR [10]	26.834	0.830	0.171
MINIMA [9]	26.841	0.832	0.169
Ours	**26.990**	**0.842**	**0.145**

E. More RGB-X Modality

. Although our work is presented in the context of sensor-based captures, the method is not restricted to sensor modality only and can be applied to general RGB-X settings. In particular, it naturally supports scenarios involving two unpaired **Camera A and B**, where modality X is captured or derived from **Camera B**, and the goal is to fuse X into the

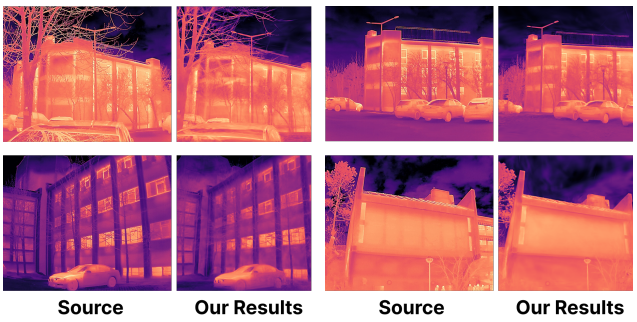

Figure S9. **Visualization of Source and Targeted Views.**

3D fields of **Camera A**. This also covers practical cases where X originates from external sources—for example, a third-party map or reconstruction provided without the original viewpoints—yet one still wishes to integrate it into a self-captured scene.

We further conduct an experiment on **RGB–Normal** by treating the normal map as an additional style modality. We adopt the METU-VisTIR-Cloudy dataset and use its original RGB images as **Camera A**. Then, using our trained RGBX-3DGS model, we render novel RGB views to serve as **Camera B**. To introduce larger viewpoint displacement, the pose of each novel view is set to the mean of the current frame's pose and that of the frame 10 steps ahead. Next, we apply DSINE [1] to estimate normal maps from Camera B's RGB images. Finally, we use our method to fuse Camera B's estimated normals with Camera A's RGB observations into a unified 3DGS field.

We show the results in Fig. S11 and compare with MINIMA's warping + 3DGS. MINIMA is trained with higher robustness across different modalities. Other image matchers like LoFTR, XoFTR, or LightGLUE fail to match across RGB and normal maps to produce reasonable results to consolidate in 3DGS. Our method produces much better, more structured normal maps. A numerical comparison is shown in Tab. S4, where we compare using the ITM and ITcos.

Table S4. **Comparison on RGB-Normal using METU-VisTIR-Cloudy**.

Method	**ITM↑**	**ITcos↑**
MINIMA [9]	0.412	0.333
Ours	**0.906**	**0.422**

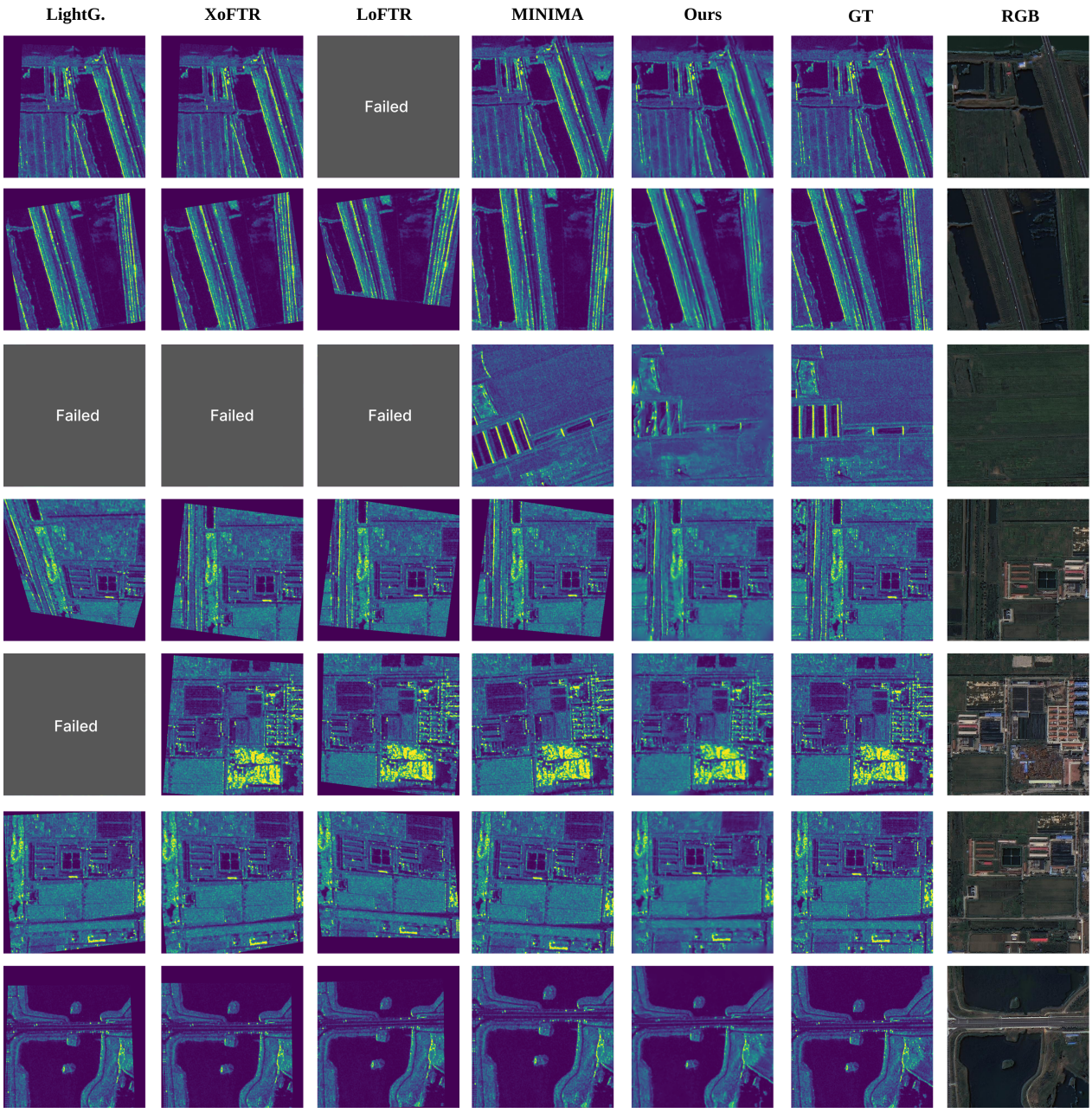

Figure S10. **More RGB-SAR Results and Comparison.**

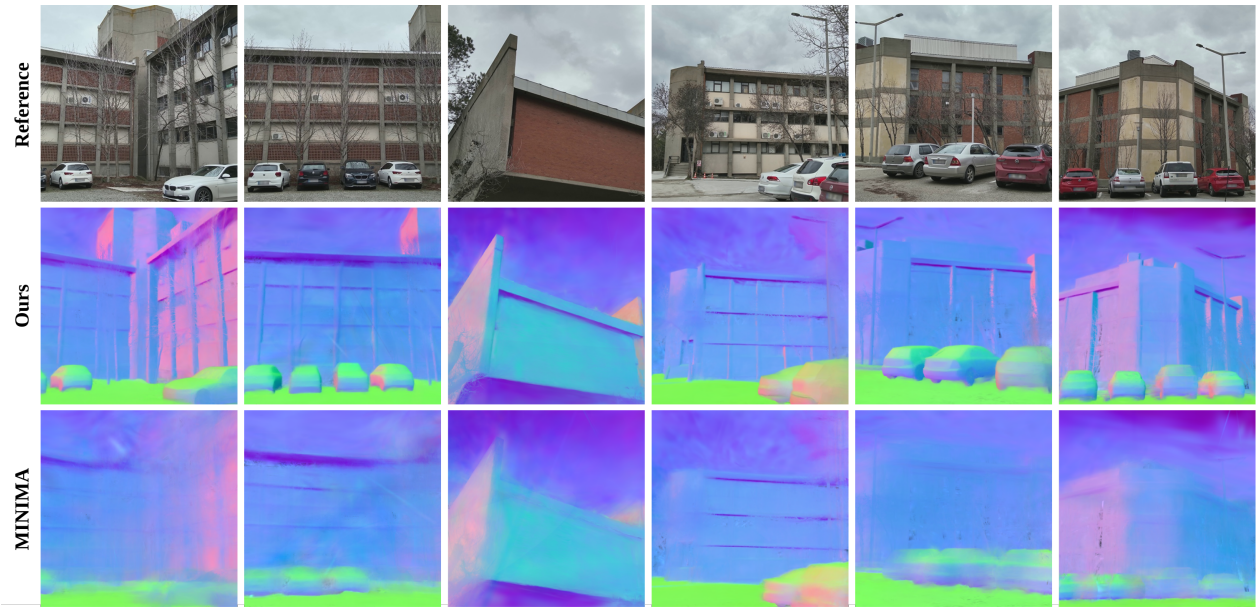

Figure S11. **Visual Comparison on RGBT-Normal.** Our method produces much better and structured normal maps.

References

[1] Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9535–9545, 2024. 5

[2] Qian Chen, Shihao Shu, and Xiangzhi Bai. Thermal3d-gs: Physics-induced 3d gaussians for thermal infrared novel-view synthesis. In *European Conference on Computer Vision*, pages 253–269. Springer, 2024. 1

[3] Josef Grün, Lukas Meyer, Maximilian Weiherer, Bernhard Egger, Marc Stamminger, and Linus Franke. Towards integrating multi-spectral imaging with gaussian splatting. *arXiv preprint arXiv:2509.00989*, 2025. 1

[4] Saimouli Katragadda, Cho-Ying Wu, Yuliang Guo, Xinyu Huang, Guoquan Huang, and Liu Ren. Online language splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1

[5] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1

[6] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17627–17638, 2023. 4

[7] Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Chenggang Yan, et al. Thermalgaussian: Thermal 3d gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[8] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1

[9] Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, and Xiang Bai. Minima: Modality invariant image matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23059–23068, 2025. 4, 5

[10] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 4

[11] Önder Tuzcuoğlu, Aybora Köksal, Buğra Sofu, Sinan Kalkan, and A Aydin Alatan. Xoftr: Cross-modal feature matching transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition Workshops*, pages 4275–4286, 2024. 4

[12] Yan Wang. Edge-enhanced feature distillation network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 777–785, 2022. 1

[13] Yiming Zuo, Willow Yang, Zeyu Ma, and Jia Deng. Omni-dc: Highly robust depth completion with multiresolution depth integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9297, 2025. 1