Supplementary of "MINIMA: Modality Invariant Image Matching"

Jiangwei Ren[1], Xingyu Jiang[1†], Zizhuo Li[2], Dingkang Liang[1], Xin Zhou[1], Xiang Bai[1]
[1] Huazhong University of Science and Technology, [2] Wuhan University
{jwren, jiangxy998, dkliang, xbai}@hust.edu.cn

We first provide more details of our data engine and the proposed MINIMA. Then we conduct additional experiments, including more ablation studies, more quantitative and qualitative matching results, and applying our MINIMA to the *Visual Localization*.

A. Details of Our Data Generation

A.1. Quality Verification of Modality Generation

The generation models for several modalities, excluding infrared (IR), have achieved significant success in their respective domains. Therefore, we additionally evaluate the quality of our infrared generation model. We use a diffusion-based method [7] for fine-tuning due to its significant performance in style transfer. The fine-tuning process utilizes 80% real RGB-IR pairs from LLVIP [8] and M3FD [14], while the rest 20% is used for the test. These two datasets provide over 10,000 real RGB-IR image pairs, which are fully aligned by manual labeling.

Evaluation Protocol. In addition to LLVIP and M3FD, we additionally evaluate the generation performance on the MSRS dataset [23] by randomly selecting 120 RGB-IR pairs. Specifically, we generate the pseudo-IR image for one RGB and then compare it with the corresponding real IR image. For comparison, we adopt XoFTR (CVPR 24) [24] and CPSTN (IJCAI 22) [26] as baseline methods. XoFTR used a handcrafted method to transfer RGB to IR, while CPSTN is a cycle-consistent perceptual network. We employ quantitative metrics including PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure) [29], LPIPS (Learned Perceptual Image Patch Similarity) [30] with AlexNet [11], and PyTorch FID (Frechet Inception Distance) [20]. As for FID, we compute the dimensionality of features with sizes 2048 to serve as an evaluation indicator. The results are presented in Tab. A1 with visualizations provided in Fig. A1 and Fig. A2

Results Analysis. From both qualitative and quantitative results, we find our infrared generation achieves huge improvements. Specifically, our generated infrared images are closer to the real sensors, and the contents are clear and even better than ground truths. In addition, almost all the metrics demonstrate superiority to others by large margins. The promising performance helps a lot for our data engine to generate high-quality cross-modal image pairs. It is also critical in training a unified matching model, making our MINIMA obtain high generalization ability.

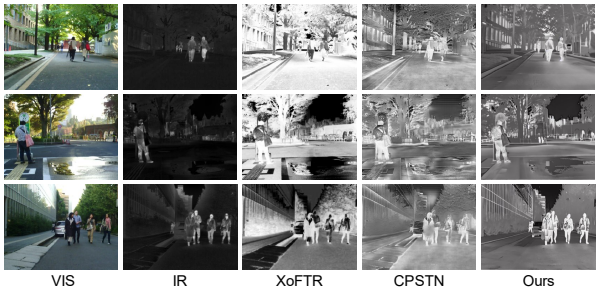

Figure A1. **Visualization Results of Infrared generation on MSRS**. The first two columns are real RGB and Infrared images.

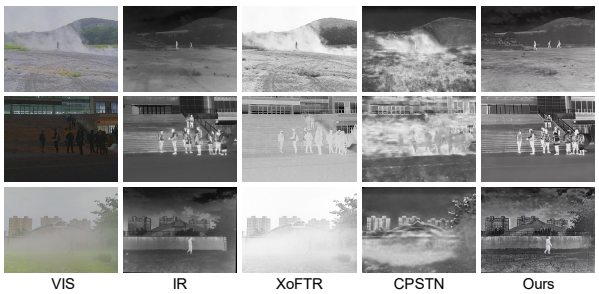

Figure A2. **Visualization Results of Infrared Generation on M3FD**. The first two columns are real RGB and Infrared images.

A.2. Data Cleaning

It is necessary to clean up the synthetic data to reduce the impacts of abnormal ones since we can not ensure the quality of the generated images. To this end, and for each RGB image and its corresponding pseudo modalities, we use our matching model (fine-tuned on the target modality) to recover the homographies (the ground truths are the identity matrix) for them. Any image pair with the mean projection error of corner points larger than 10 pixels is regarded as dirty data and dropped. Finally, 0.91% of the matching

† Corresponding author.

Table A1. **Quantitative Evaluation of Infrared Generation with Different Metrics**. The test datasets are LLVIP [8], M3FD [14] and MSRS [23]. CPSTN (IJCAI 22) [26] and XoFTR (CVPR 24) [24] are used for comparison. **Bold** indicates the best.

Data	Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID-2048 ↓
LLVIP	CPSTN	27.91	0.32	0.66	303.55
	XoFTR	27.90	0.29	0.71	204.44
	Ours	**28.28**	**0.55**	**0.42**	**145.93**
M3FD	CPSTN	27.82	0.37	0.56	161.71
	XoFTR	27.86	0.33	0.59	125.07
	Ours	**28.14**	**0.53**	**0.46**	**119.96**
MSRS	CPSTN	**27.95**	0.15	0.74	204.37
	XoFTR	27.84	0.16	0.77	167.39
	Ours	27.87	**0.19**	**0.73**	**161.37**

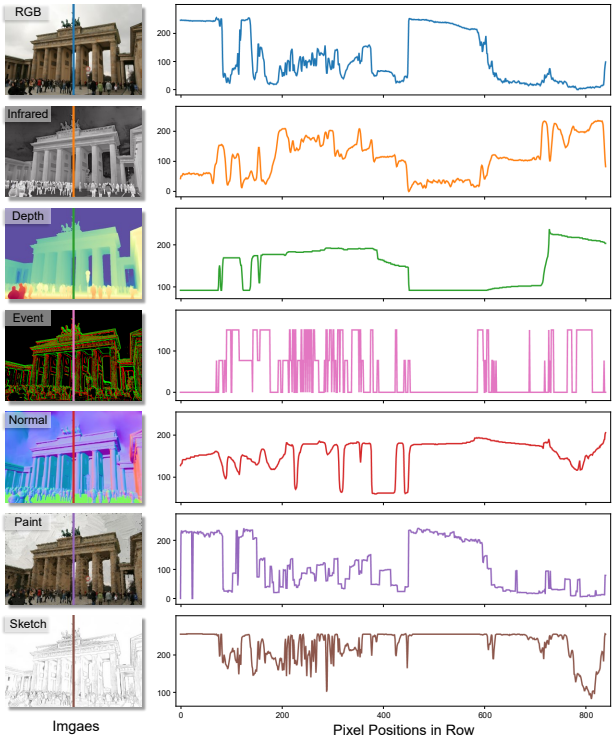

Figure A3. **Pixel Intensity Statistic for Generated Modalities**. The statistical differences reveal the excellent ability of our data engine to generate modality gaps.

pairs are dropped in the training set.

A.3. Analysis of the Generated Modalities

We show a group of generated modalities of our data engine in A3, which show high-quality visible results. We also provide their pixel intensity statistics, which reveal the modality gap among them. These high-quality images together with their modality gaps enable existing matching models to easily obtain cross-modal ability. Not only image matching, our MINIMA can also deliver significant reference and insight for other multimodal perception tasks.

B. Details of MINIMA

The details of our fine-tuning stage are as follows. i) **LightGlue** (LG) [13]. We use SuperPoint [4] to extract 2048 keypoints and freeze its parameters, then only fine-tune LightGlue. Because SuperPoint is verified to extract matchable features for cross-modal images [10]. Just fine-tuning LightGlue can achieve promising performance, as demonstrated by our MINIMA$_{\text{LG}}$. Here we directly adopted the initial learning rate, *i.e.,* 1×10^{-4}, in the fine-tuning stage. In practice, we fine-tune the LG model for 50 epochs as the authors suggested, which also shows converged performance in our study. ii) **LoFTR** [22] and **RoMa** [6]. We lower the learning rate to the $1/10$ of the original, with the linear scaling rule to account for batch size differences. Specifically, the initial learning rate is set as 1×10^{-4} for LoFTR. And we set it as 1.5×10^{-5} for the RoMa decoder, and 7.5×10^{-7} for the RoMa encoder. Note that we maintain the default learning rate decay strategies for all methods during the fine-tuning. For LoFTR, we fine-tune for 30 epochs. In contrast, we fine-tune RoMa for only 4 epochs due to its inherent capabilities, which have already achieved amazing gains. For better understanding, we also fine-tune ELoFTR [28] and XoFTR [24], denoted as MINIMA$_{\text{ELoFTR}}$ and MINIMA$_{\text{XoFTR}}$. And they are fine-tuned for 20 epochs and 5 epochs, respectively. Their learning rates are similar to our MINIMA$_{\text{LoFTR}}$. The corresponding results are in Tab. A3, Tab. A4 and Tab. A5.

C. Additional Experimental Results

C.1. More Studies on Different Training Data

To better understand our MINIMA, we fine-tune the basic models on different combinations of our generated cross-modal data. The obtained models are evaluated on different real scenes, and the results are reported in Tab. A2. For each baseline, we first report the AUCs of the official models (without any fine-tuning) in the first row. Then we fine-tune each model on a single type of modality pair (RGB+X), which shows large enhancements compared with the basic models. Finally, we fine-tune the models on two or more modality types. The results demonstrate that different modalities can cooperate to train a better model. Using RGB-IR/Depth/Normal can achieve the best overall performance; hence, we use them as our final models. Additionally, using artistic data (Paint and Sketch) can not further enhance the performance because the artistic type has no physical property and is different from other modality types.

C.2. More Results of Semi-dense Matching

For a better understanding of our MINIMA, we further fine-tune ELoFTR [28] and XoFTR [24] on the generated data, obtaining MINIMA$_{\text{ELoFTR}}$ and MINIMA$_{\text{XoFTR}}$. The corresponding results of semi-dense matching are reported

Table A2. **Ablation Studies with Different Training Data.** The basic models are LG, LoFTR, and RoMa. The training sets are different combinations of our generated cross-modal data. We evaluated the fine-tuned models on real cross-modal cases. For each baseline, the model trained on the original MegaDepth is reported in the first row. The average performance is shown in the last column.

Models	Generated Modalities						Rel IR	Rel Depth	Rel Event	RS	Medical	Average
	Infrared	Depth	Normal	Event	Paint	Sketch	AUC@10°	AUC@10px	AUC@10px	AUC@10px	AUC@10px	
$\text{MINIMA}_{\text{LG}}$							5.37	11.26	7.02	44.62	49.48	23.55
	✓						35.55	47.27	**13.39**	55.12	**52.73**	40.81
		✓					30.54	51.78	12.08	57.73	52.17	40.86
			✓				32.66	48.66	10.44	58.15	52.43	40.47
				✓			23.33	38.37	10.01	55.72	51.32	35.75
	✓	✓	✓				**37.17**	**55.97**	12.82	**58.74**	52.50	**43.44**
				✓	✓	✓	13.05	31.09	10.89	51.52	49.67	31.24
	✓	✓	✓	✓	✓	✓	36.34	55.93	12.74	58.41	52.45	43.17
$\text{MINIMA}_{\text{LoFTR}}$							6.94	15.16	5.91	50.79	50.13	25.79
	✓						29.55	39.23	11.12	48.79	51.71	36.08
		✓					15.12	36.06	5.32	53.64	52.40	32.51
			✓				23.14	39.79	10.28	54.73	52.53	36.09
				✓			14.96	32.97	12.19	45.77	50.28	31.24
	✓	✓	✓				**30.84**	44.85	11.38	**56.81**	**52.77**	39.33
				✓	✓	✓	10.10	21.96	11.33	48.59	49.88	35.47
	✓	✓	✓	✓			30.80	**48.55**	**12.44**	56.04	51.82	**39.93**
	✓	✓	✓	✓	✓	✓	30.61	45.10	11.83	55.33	52.19	39.01
$\text{MINIMA}_{\text{RoMa}}$							48.12	49.31	10.71	57.84	53.75	43.95
	✓						57.28	57.49	10.49	60.37	57.08	48.54
		✓					60.42	72.63	11.00	62.95	56.72	52.74
			✓				60.36	72.51	10.89	63.23	55.26	52.45
				✓			59.11	69.11	11.71	64.30	57.75	52.40
		✓	✓				58.89	72.88	12.36	63.91	55.50	52.71
	✓	✓	✓				60.70	72.54	**17.07**	64.38	55.09	**53.96**
				✓	✓	✓	58.14	65.91	8.79	59.66	57.73	50.05
	✓	✓	✓	✓			**61.27**	**73.80**	11.02	**65.01**	55.04	53.23
	✓	✓	✓	✓	✓	✓	60.43	72.83	12.98	64.80	**57.92**	53.79

Table A3. **Semi-dense Matching Results on Our Synthetic Dataset.** The AUC of the pose error in percentage is reported. The best and second are masked as **Bold** and Underline, respectively.

Category	Method	RGB-IR			RGB-Depth			RGB-Normal			RGB-Event			RGB-Sketch			RGB-Paint		
		@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°
Semi-Dense	LoFTR [22]	5.44	12.58	24.28	0.13	0.44	1.88	5.72	12.07	23.14	4.90	12.43	26.45	37.81	54.82	69.52	5.93	12.22	22.19
	XoFTR [24]	17.85	32.21	49.53	12.82	23.10	36.02	22.74	38.35	54.71	**33.33**	**51.61**	**67.49**	**44.18**	**61.39**	**75.07**	3.73	7.54	14.48
	ELoFTR [28]	6.73	14.59	27.36	0.25	0.79	3.32	11.20	21.67	36.86	9.25	20.39	37.56	43.86	61.09	74.84	14.09	25.11	39.44
	$\text{GIM}_{\text{LoFTR}}$ [21]	2.60	6.79	15.50	0.00	0.04	0.27	0.35	1.06	4.01	0.44	1.43	5.28	17.30	31.82	48.79	4.84	10.64	21.82
	$\text{MINIMA}_{\text{LoFTR}}$	18.07	32.36	48.42	14.70	28.81	46.23	27.65	44.26	59.88	18.14	32.74	49.11	36.07	53.54	68.47	7.79	15.45	27.39
	$\text{MINIMA}_{\text{XoFTR}}$	**18.97**	**34.36**	**51.72**	**24.47**	**40.90**	**58.36**	**30.47**	**47.90**	**64.64**	31.14	49.39	65.71	42.91	60.77	75.00	5.61	11.56	20.95
	$\text{MINIMA}_{\text{ELoFTR}}$	13.14	26.36	43.63	16.59	32.26	50.37	29.72	47.47	63.72	15.66	30.72	48.73	41.64	59.63	73.73	**15.02**	**27.02**	**41.62**

in Tab. A3 and Tab. A4. The results reveal that with better pipelines, our MINIMA can achieve further enhancements of overall performance.

C.3. Results on Original MegaDepth Dataset

In this part, we will evaluate the performance degradation on RGB-only matching tasks for those cross-modal matching methods. To this end, we test these methods back to the original MegaDepth-1500 [12]. We use the same settings as described in [13, 22]. Following previous testing, the RANSAC threshold is still set to 0.5. For semi-dense and dense methods, the longest edge of the input images is resized to 1200 pixels, while for sparse methods, it is resized to 1600 pixels. The results are summarized in Tab. A5, revealing that our MINIMA can well maintain the ability of RGB-only matching, except for LoFTR.

C.4. Scratch Training v.s. Fine-tuning

We report the loss values and AUC@5° performance with respect to epochs, by using scratch training and fine-tuning. The test set is our synthetic RGB-IR data generated from MegaDepth-1500 [12]. We use LoFTR as the basic model, and the training set is our synthetic RGB-IR/Depth/Normal. Statistic results are shown in Fig. A4, which reveal that the fine-tuning strategy can converge more rapidly since the pre-trained model can provide good matching priors for

Table A4. **Semi-dense Matching Results on Real Dataset.** The AUC of the pose error in percentage is reported. The best and second are masked as **Bold** and Underline, respectively.

Category	Method	Real RGB-IR			Real RGB-Depth			Medical			Remote Sensing			Real RGB-Event		
		@5°	@10°	@20°	@3px	@5px	@10px	@3px	@5px	@10px	@3px	@5px	@10px	@3px	@5px	@10px
Semi-Dense	LoFTR [22]	2.88	6.94	14.95	0.97	4.20	15.16	38.42	43.89	50.13	24.13	33.80	50.79	0.00	0.00	3.59
	XoFTR [24]	18.47	34.64	51.5	11.03	27.24	51.60	**39.67**	**45.60**	52.32	**27.35**	**39.58**	56.63	0.00	1.37	12.64
	ELoFTR [28]	2.88	7.88	17.72	0.82	4.09	16.69	34.57	41.66	49.08	16.45	29.65	46.74	0.64	1.34	7.78
	$\text{GIM}_{\text{LoFTR}}$ [21]	0.43	1.06	2.99	0.00	0.25	1.15	39.51	44.40	48.94	17.96	27.41	37.29	0.00	0.55	1.19
	$\text{MINIMA}_{\text{LoFTR}}$	15.61	30.84	47.87	5.35	18.65	44.85	**39.67**	45.33	**52.77**	23.32	35.18	**56.81**	**0.81**	**2.49**	11.75
	$\text{MINIMA}_{\text{XoFTR}}$	**19.38**	**35.82**	**52.94**	**11.76**	**29.48**	**55.05**	39.33	44.92	52.09	25.19	37.86	54.36	0.00	1.92	**15.23**
	$\text{MINIMA}_{\text{ELoFTR}}$	12.11	28.07	47.25	3.96	16.42	44.03	39.12	44.61	52.12	19.70	33.78	53.83	0.37	1.04	9.66

Table A5. **Evaluation on Original Megadepth-1500 for Pose Estimation.** The AUC of the pose error in percentage is reported. This mainly demonstrates that our MINIMA can well preserve the RGB-only matching performance except when using LoFTR.

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Sparse	SuperGlue [16](CVPR 20)	49.7	67.1	80.6
	LightGlue (LG) [13](ICCV 23)	49.9	67.0	80.1
	GIM_{LG} [21](ICLR 24)	41.3	60.7	75.9
	$\text{MINIMA}_{\text{LG}}$	47.3	65.0	78.6
Semi-Dense	LoFTR [22](CVPR 21)	53.6	69.9	82.0
	$\text{GIM}_{\text{LoFTR}}$ [21](ICLR 24)	51.3	68.5	81.1
	ELoFTR [28](CVPR 24)	56.4	72.2	83.5
	XoFTR [24](CVPR 24)	45.8	61.7	74.0
	$\text{MINIMA}_{\text{LoFTR}}$	29.9	45.3	59.5
	$\text{MINIMA}_{\text{ELoFTR}}$	51.0	68.1	80.3
	$\text{MINIMA}_{\text{XoFTR}}$	44.5	60.0	72.3
Dense	DKM [5](CVPR 23)	60.4	74.9	85.1
	GIM_{DKM} [21](ICLR 24)	60.7	75.5	85.9
	RoMa [6](CVPR 24)	62.6	76.7	86.3
	$\text{MINIMA}_{\text{RoMa}}$	61.7	76.5	86.4

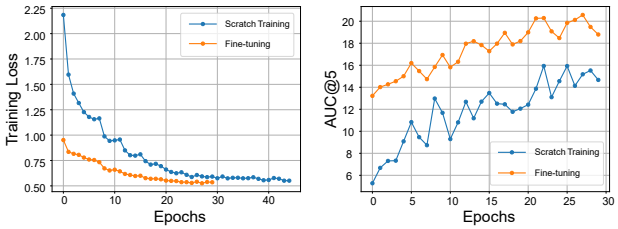

Figure A4. **Training Loss and AUC@5° *w.r.t.* Epochs, using Scratch Training and Fine-tuning**. The basic model is LoFTR. The test set is our synthetic RGB-IR of MD-syn.

challenging cross-modal tasks.

C.5. Apply to Visual Localization

Vision localization (VL) is a critical downstream task of image matching. The target is to recover the 6−degree-of-freedom (6−DOF) camera pose from a query image related to a known 3D scene model. We perform it on the Aachen v1.0 dataset, which is a challenging large-scale outdoor dataset for localization with large-viewpoint and day-night illumination changes, making the localization largely rely on the robustness of matching methods. We adopt its full localization track for benchmarking.

Table A6. **Visual Localization on Aachen Day-Night V1.0** [17]

Method	Day	Night
	(0.25m,2°) / (0.5m,5°) / (5m,10°)	
MNN	86.9 / 92.0 / 95.5	73.5 / 79.6 / 88.8
SuperGlue [16](CVPR 20)	87.9 / **95.0** / 97.9	84.7 / **92.9** / 99.0
SGMNet [2](ICCV 21)	86.5 / 93.7 / 97.2	82.7 / 91.8 / 99.0
LightGlue (LG) [13](ICCV 23)	88.0 / 93.8 / 97.5	84.7 / 91.8 / 99.0
ConvMatch [31](TPAMI 23)	88.1 / 94.4 / 97.3	79.6 / 88.8 / 96.9
$\text{MINIMA}_{\text{LG}}$	**88.3** / 94.7 / **98.3**	**85.7** / **92.9** / **100.0**

Following [13, 15], we integrate different matching methods into the official HLoc pipeline [15] to achieve localization. Specifically, with COLMAP [18, 19] toolbox, we first triangulate a 3D point cloud for all reference images with known poses and calibration, then retrieve 20 reference images for each query image with NetVLAD [1] on Aachen Day-Night v1.0. Then, we match the query image and the retrieved images with image matching methods, where the feature points are extracted up to 4096 by SuperPoint [4]. Finally, the camera poses are estimated by RANSAC and a Perspective-n-Point solver. We report the pose recall at different scales of distance and angular thresholds, *i.e.,* (0.25m,2°) / (0.5m,5°) / (5m,10°). The sparse matchers, including SuperGlue [16], SGMNet [2], LightGlue (LG) [13], ConvMatch [31] and our MINIMA fine-tuned with LightGlue, are used for comparison. We also report the raw results of SuperPoint directly with Mutual Nearest Neighbor (MNN) matching.

The localization results are summarized in Tab. A6, which demonstrate the good ability of our MINIMA for downstream applications. Since our MINIMA is additionally trained on high-quality multimodal image pairs, it can be more robust in complex scenarios.

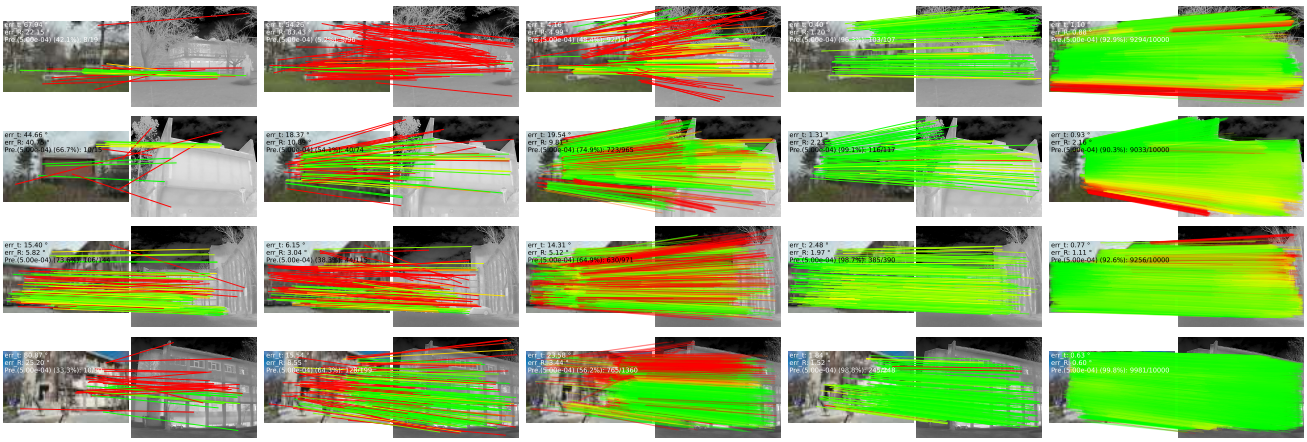

Figure A5. **Qualitative Results on Real RGB-IR Image Pairs of METU-VisTIR** [24]. The red lines indicate false matches.

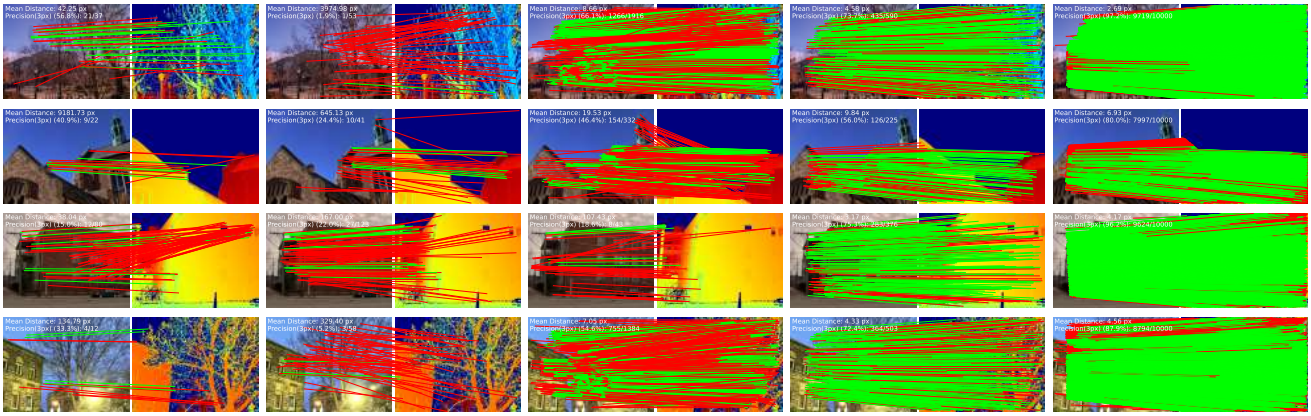

Figure A6. **Qualitative Results on Real RGB-Depth Image Pairs of DIODE Dataset** [25]. The red lines indicate false matches.

C.6. More Visible Results on Real Datasets

We also show more qualitative results, which are selected from real RGB-IR [24], RGB-Depth [25], RGB-Event [27] and Remote Sensing [10] (including Optical-SAR, optical-Map, and Day-Night) datasets. For each pair, we show the raw matching results before RANSAC. The red lines indicate false matches whose epipolar error (pose) or projection error (homography) is beyond 5×10^{-4} and 3 pixels, respectively. Visible results are shown in Fig. A5, Fig. A6, Fig. A7 and Fig. A8. Our methods MINIMA$_{\text{LG}}$ (sparse) and MINIMA$_{\text{RoMa}}$ (dense) are compared with the sparse pipeline ReDFeat [3] and OmniGlue [9], and semi-dense matcher XoFTR [24]. ReDFeat and XoFTR are cross-modal methods, and OmniGlue is known for its generalization. The results reveal that our MINIMA can produce a high number and ratio of correct matches (green lines).

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, pages 5297–5307, 2016. 4

[2] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *ICCV*, pages 6301–6310, 2021. 4

[3] Yuxin Deng and Jiayi Ma. Redfeat: Recoupling detection and description for multimodal feature learning. *IEEE Trans. Image Process.*, 32:591–602, 2022. 5

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, pages 224–236, 2018. 2, 4

[5] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *CVPR*, pages 17765–17775, 2023. 4

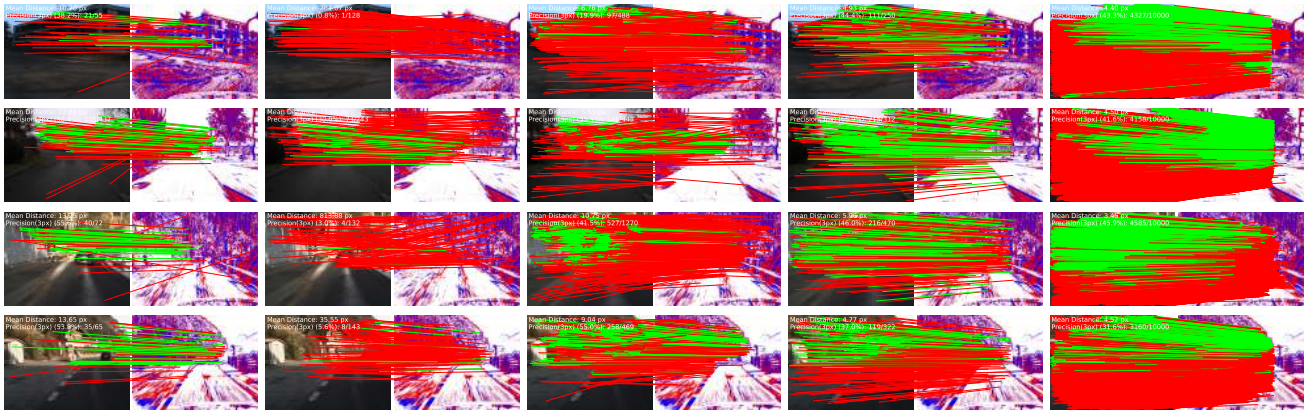

Figure A7. **Qualitative Results on Real RGB-Event Image Pairs of DSEC Dataset** [27]. The red lines indicate false matches.

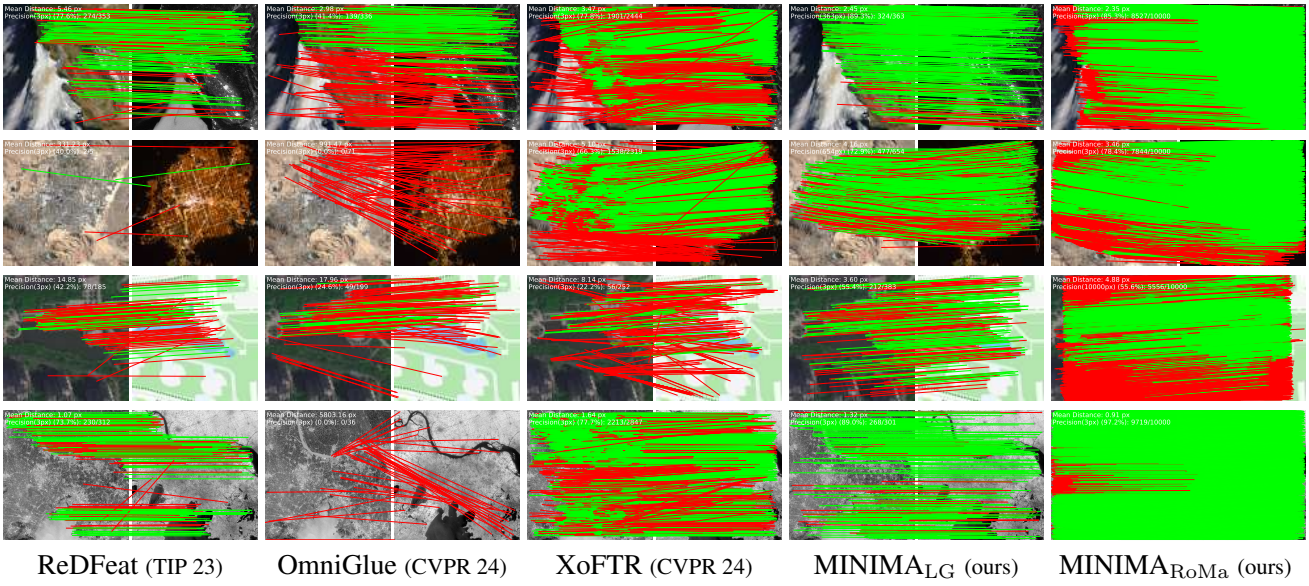

Figure A8. **Qualitative Results on Real Image Pairs of Cross-modal Remote Sensing Dataset** [10]. The red lines indicate false matches.

[6] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, pages 19790–19800, 2024. 2, 4

[7] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. *arXiv preprint arXiv:2404.12154*, 2024. 1

[8] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *ICCV*, pages 3496–3504, 2021. 1, 2

[9] Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. Omniglue: Generalizable feature matching with foundation model guidance. In *CVPR*, pages 19865–19875, 2024. 5

[10] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021. 2, 5, 6

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. 2012. 1

[12] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 3

[13] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, pages 17627–17638, 2023. 2, 3, 4

[14] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, pages 5802–5811, 2022. 1, 2

[15] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, pages 12716–12725, 2019. 4

[16] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 4

[17] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, pages 8601–8610, 2018. 4

[18] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 4

[19] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 4

[20] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2020. Version 0.3.0. 1

[21] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. In *ICLR*, 2024. 3, 4

[22] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, pages 8922–8931, 2021. 2, 3, 4

[23] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 2022. 1, 2

[24] Önder Tuzcuoğlu, Aybora Köksal, Buğra Sofu, Sinan Kalkan, and A Aydin Alatan. Xoftr: Cross-modal feature matching transformer. In *CVPR*, pages 4275–4286, 2024. 1, 2, 3, 4, 5

[25] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 5

[26] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pages 3508–3515, 2022. 1, 2

[27] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Trans. Cybern.*, 2023. 5, 6

[28] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *CVPR*, pages 21666–21675, 2024. 2, 3, 4

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 1

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 1

[31] Shihua Zhang and Jiayi Ma. Convmatch: Rethinking network design for two-view correspondence learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 4