

MINIMA: Modality Invariant Image Matching

Xingyu Jiang¹, Jiangwei Ren¹, Zizhuo Li², Xin Zhou¹, Dingkang Liang¹, Xiang Bai^{1†}

¹Huazhong University of Science and Technology, ² Wuhan University
{jiangxy998, jwren, dkliang, xbai}@hust.edu.cn

Abstract

Image matching for both cross-view and cross-modality plays a critical role in multimodal perception. In practice, the modality gap caused by different imaging systems/styles poses great challenges to the matching task. Existing works try to extract invariant features for specific modalities and train on limited datasets, showing poor generalization. In this paper, we present MINIMA, a unified image matching framework for multiple cross-modal cases. Without pursuing fancy modules, our MINIMA aims to enhance universal performance from the perspective of data scaling up. For such purpose, we propose a simple yet effective data engine that can freely produce a large dataset containing multiple modalities, rich scenarios, and accurate matching labels. Specifically, we scale up the modalities from cheap but rich RGB-only matching data, by means of generative models. Under this setting, the matching labels and rich diversity of the RGB dataset are well inherited by the generated multimodal data. Benefiting from this, we construct MD-syn, a new comprehensive dataset that fills the data gap for general multimodal image matching. With MD-syn, we can directly train any advanced matching pipeline on randomly selected modality pairs to obtain cross-modal ability. Extensive experiments on in-domain and zero-shot matching tasks, including 19 cross-modal cases, demonstrate that our MINIMA can significantly outperform the baselines and even surpass modality-specific methods. The dataset and code are available at <https://github.com/LSXI7/MINIMA>.

1. Introduction

Image matching refers to establishing pixel-wise correspondences from two-view images, which serves as a prerequisite for a wide range of visual applications [32]. Recently, matching two images of different imaging systems/styles plays a vital role in multimodal perceptions [21], including image fusion and enhancement [51,

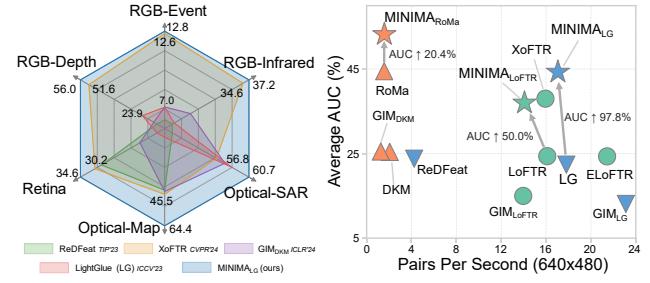


Figure 1. **Overall Image Matching Accuracy and Efficiency on Six Datasets of Real Cross-modal Image Pairs.** AUC of the pose error (@10°) or reprojection error (@10px) is used for accuracy evaluation, while Pairs Per Second is used for efficiency test. **Left:** AUCs on each dataset of representative methods are reported. **Right:** average performance is summarized, wherein different colors indicate matching pipelines of sparse, semi-dense and dense matching, while our MINIMA is marked as ★. Only with synthetic multimodal data created by our data engine, MINIMA can generalize to real cross-modal scenes with large improvements.

[57], visual localization/navigation [1, 63], target detection/recognition/tracking [13, 49, 52, 61, 62, 64], etc. They benefit from gathering the advantages of different modalities by aligning them, thereby yielding more comprehensive representations for better accuracy and robustness. However, the cross-view and cross-modality nature makes the matching task much more challenging, particularly using a single model for different modalities such as *RGB-Infrared (IR)*, *RGB-Depth*, and *RGB-Event*.

Existing studies focus more on RGB-only image matching due to the accessible training sets [5, 26], which have given birth to many advanced matching architectures [9, 10, 35, 41, 47]. By contrast, cross-modal matching dataset is weak in scale and scene coverage, as concluded in Tab. 1. The main reasons are as follows: i) It is laborious to capture a large number of multimodal images of the same target/scene, and also hard to ensure rich scene coverage. Therefore, existing datasets are often captured from driving or fixed camera views [19, 43], and the number of modality types is merely two or three for each dataset. ii) Creating precise dense labels is also expensive. Researchers of

† Corresponding author.

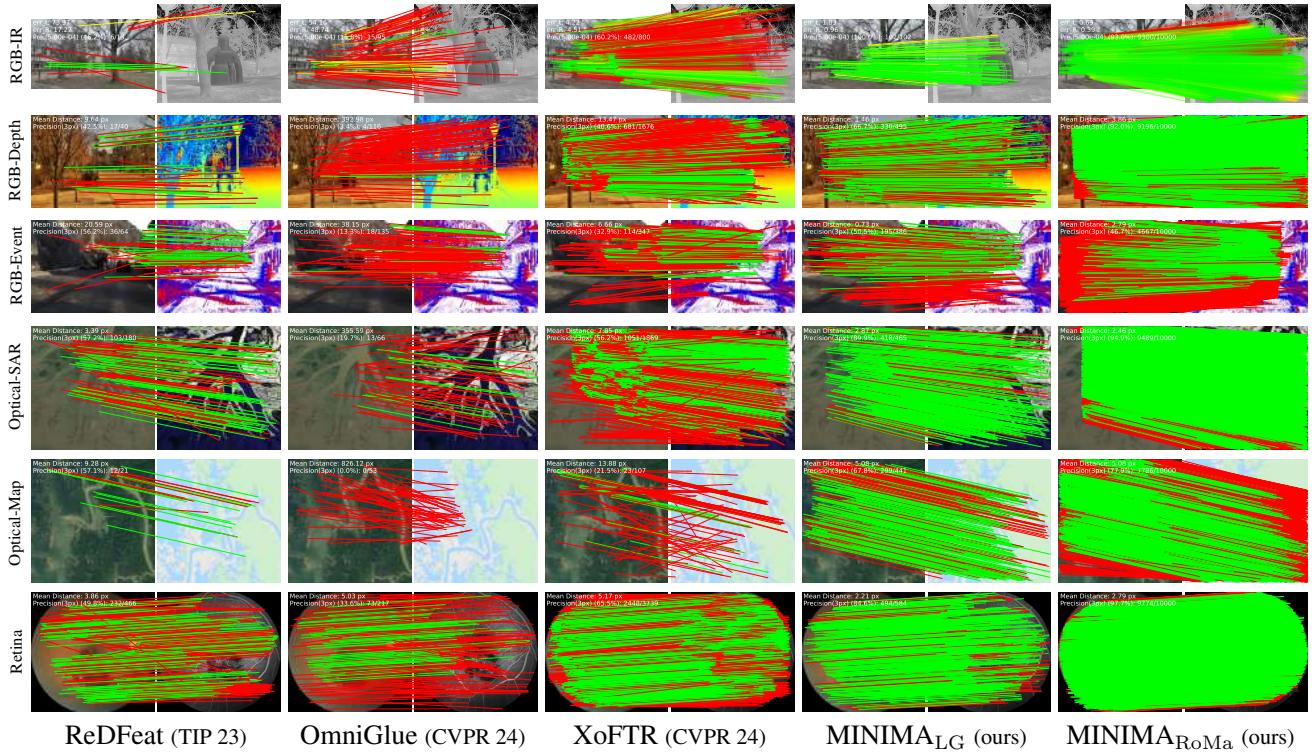


Figure 2. Qualitative Results on Real Cross-modal Image Pairs. Our methods $\text{MINIMA}_{\text{LG}}$ (sparse) and $\text{MINIMA}_{\text{RoMa}}$ (dense) are compared with the sparse matching pipeline ReDFeat [7] and OmniGlue [20], and semi-dense matcher XoFTR [43]. ReDFeat and XoFTR are cross-modal methods, and OmniGlue is known for its generalization ability. Matches generated by each method are drawn, where the red lines indicate epipolar error (pose) or projection error (homography) beyond 5×10^{-4} or 3 pixels. Details are recorded in the top-left of each image pair, including the geometric errors created by default RANSAC estimation and the (# correct match / # match).

ten manually label matched landmarks [19, 21, 29], or use camera calibration¹ to produce *approximate* poses [43, 53]. These limited datasets can not support the training of a general matching method well due to the imbalances among them, which make the models easily dominated by simple datasets. In addition, to enlarge the data scale, researchers often generate pseudo transformations from aligned image pairs [6, 58]. However, this approach is still limited by the original data, where the stimulated deformations are not consistent with practical viewpoint changes. That is why existing works can only extract matchable features for specific modalities and show poor generalization.

In this paper, we try to develop a unified matching framework for multiple cross-modal cases, by filling the data gap with an effective data engine. This engine helps us to freely scale up cheap RGB images to a large multimodal dataset with rich scenarios and accurate labels. The introduced dataset can well support the training of any matching pipelines, and largely enhance the cross-modal performance and zero-shot ability. *Our contributions are as follows:*

- We are the first to develop a unified matching framework

¹Making two cameras as close as possible, thus considering they share a common camera pose.

MINIMA for any cross-view and cross-modality image pairs, and achieve amazing performance enhancement.

- We introduce a simple yet effective data engine to freely build a high-quality multimodal dataset for image matching. Based on this, we construct a comprehensive dataset MD-syn with large scene coverages and precise labeling, which fills the data gap for the matching community.
- We conduct extensive experiments on in-domain and zero-shot matching tasks including 19 cross-modal cases, which demonstrate the high quality of our MD-syn and the promising generalization of our MINIMA .

2. Related Work

2.1. RGB-Only Image Matching

Image matching is a fundamental problem in computer vision, which has aroused numerous matching methods over the past decades. Conventional pipelines start with hand-crafted designs of keypoints detection, description then matching, which are recently updated with deep learning [8, 28, 32, 35]. These detector-based methods can establish keypoint matches with high efficiency but often struggle in textureless regions. Recently, detector-free meth-

Table 1. Overview of Representative Datasets. It contains RGB-only and multimodal matching datasets, and our proposed MD-syn. The number (#) of Pairs, Scene (Type: Indoor or Outdoor), Modality type, and the forms of Match Label are summarized.

| Dataset | #Pairs | # Scene (Type) | #Modality | Match Label |
|----------------------------|--------|----------------|-----------|-------------|
| <i>RGB Matching</i> | | | | |
| MegaDepth [26] | 40M | 196 (Out.) | 1 | Depth, Pose |
| ScanNet [5] | 230M | 1513 (In.) | 1 | Depth, Pose |
| <i>Multimodal Matching</i> | | | | |
| METU-VisTIR [43] | 2.5K | 6 (Out.) | 2 | Pose |
| M3FD [29] | 4.2K | 15 (Out.) | 2 | Pre-aligned |
| LLVIP [19] | 15K | 26 (Out.) | 2 | Pre-aligned |
| DIODE [53] | 25K | 20 (In.&Out.) | 3 | Pre-aligned |
| MD-syn (ours) | 480M | 196 (Out.) | 7 | Depth, Pose |

ods [41, 47] have been introduced to produce semi-dense or dense [9, 10] pixel matches, and achieve dominant performance on RGB image matching in terms of match number and downstream applications. Since these methods regard each pixel as matchable points within the coarse and fine matching stage, they commonly produce a huge computational burden. Driven by sufficient datasets, those deep methods enjoy great success in building more accurate point matches. Supported by our data engine, those advanced matching pipelines can be easily fine-tuned to multimodal cases with large enhancements.

2.2. Multimodal Image Matching

Image matching for multi-modalities is more challenging, due to the domain gap between two images. It often shows variations in pixel intensity distributions (as shown in Fig. 4), making it difficult to search matchable cues. Existing studies still rely on handcrafted designs [17, 23, 55, 56], focusing on extracting matchable information such as shape, gradient, or phase. However, these low-level features are not consistently effective and time-consuming to extract. Data-driven methods exhibit powerful abilities to extract matchable features for multimodal images. They commonly utilized off-the-shelf matching pipelines [7, 31, 43] as the backbone, then adapted them to the target modalities with specific designs. For example, ReDFeat [7] re-coupled independent constraints of detection and description of multimodal feature learning with a mutual weighting strategy. It performs for three cross-modality cases, but is merely trained and tested on each dataset separately. Recently, XoFTR [43] utilizes a two-stage training approach for RGB-IR image matching. It is first pre-trained on a collected real dataset containing 95K aligned multi-spectral images, then fine-tuned on pseudo infrared images. XoFTR achieves large enhancement on RGB-IR image matching, by using abundant training data and a tailored matching rule. In this paper, we contribute to filling the data gap of the general image matching task. We demonstrate that

our MINIMA can outperform modality-specific approaches and shows superior performance in zero-shot tasks, solely relying on high-quality synthetic training data.

2.3. Existing Datasets

It is necessary to analyze the data gap between RGB-only image matching and the cross-modal cases. Specifically, multi-viewed RGB images of the same target/scene are extremely cheap and easy to collect, such as directly collecting from the internet [26] or capturing video frames [40]. Open-source tools like COLMAP [37, 38] are widely used to generate precise matching labels, such as depths and camera poses. These good datasets give birth to advanced models for RGB image matching [28, 32, 35]. However, capturing a large number of multimodal images of the same scene is laborious, since some imaging devices should be gathered for shooting together. This limits the scale of available image pairs. Moreover, the matching labels cannot be directly obtained by tools, which are often labeled manually [19, 21, 29], or approximated from camera calibrations [43, 53]. We conclude representative public datasets in Tab. 1. It shows that these multimodal datasets exhibit significant variability from each other, and are all limited by the scale and scene coverage. This impedes us from training a unified matching model for multiple cross-modal cases.

Recently, data scale-up has shown great success in general vision tasks [53, 54]. They typically enlarge the training set by generating pseudo labels from huge wild RGB images. In contrast, our challenges lie in getting numerous paired images of different modalities and rich scenarios. For such purposes, we propose a data engine to generate multiple pseudo modalities from cheap RGB image pairs. On this basis, we can generate a high-quality dataset for cross-view and cross-modality image matching, that may fill the data gap for the universal matching community and will encourage more excellent matching techniques.

3. Cross-Modal Generation with Data Engine

In this paper, we contribute to exploring a unified image matching framework for all possible image modalities by generating a large multimodal matching dataset. To achieve this, several key challenges we will face:

- How to obtain a large scale of image pairs with viewpoint and modality changes, and ensure the rich diversity.
- How to freely generate dense labels of matching for those image pairs, such as depths and camera poses.
- How to ensure the balance of different modalities in terms of scale and scene coverage.

Next, we will introduce a data engine to alleviate these concerns. It mainly benefits from the powerful ability of recent generative methods [15, 16]. The proposed engine can freely produce various pseudo modalities from real RGB

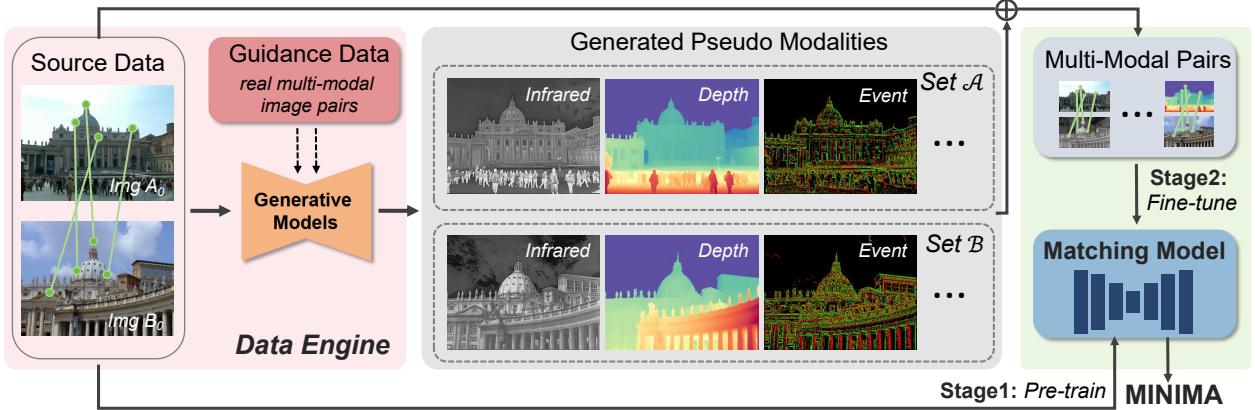


Figure 3. **Overview of the proposed MINIMA pipeline: a single model for any cross-modal matching tasks.** Wherein the *Data Engine* is to generate a large multimodal image matching dataset, which supports the training of matching models to obtain cross-modal ability.

image pairs, whose matching labels and scene diversity would be well inherited by the generated data.

3.1. Advantages of Cross-Modal Generation

The ideal strategy is to capture real images of multiple modalities in the wild. But obviously, it is impractical to arrange multiple imaging systems together. Additionally, it is more troublesome to obtain dense labels for raw image pairs, such as depth and pose information. Another common strategy [6, 7, 58] involves augmenting existing aligned image pairs by randomly generating homography matrices to simulate geometric distortions. However, this is still limited by the small scale of the used dataset in diversity. The synthetic deformations are not consistent with real viewpoint changes, resulting in weak generalization of the trained model, *i.e.*, it can only work for the test set separated from the same dataset as the training set [7, 58].

To this end, we try to generate pseudo modalities to obtain a large-scale multimodal dataset, which may help to train a unified matching model for multiple cross-modal cases. Cross-modal generation from multi-viewed RGB images has distinct advantages. **a) Cheap:** Those RGB images are easy to collect, such as capturing from the internet [26] or video frames [40]. This allows us to avoid capturing raw multimodal images in the wild. **b) Flexible:** We can obtain any pseudo modality we want by only giving some real image pairs as guidance. With cheap RGB images, we can freely define the scale and scene of the target modality to generate. This helps us to obtain sufficient multimodal image pairs and ensure the balance of scale and scene diversity among different modalities, preventing model bias toward specific modalities. **c) High-quality:** First, the generated images have the same resolution as RGB, breaking the limits of real sensors such as infrared or depth. Second, the matching labels of RGB images can be easily obtained by open-source tools [37, 38]. Those accurate and dense labels

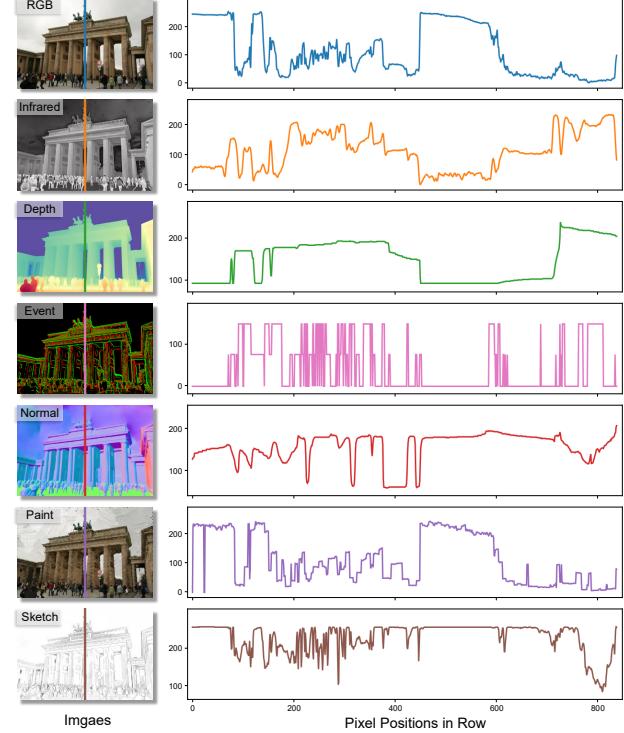


Figure 4. **Pixel Intensity Statistic for Generated Modalities.** The statistic differences reveal the excellent ability of our data engine to generate modality gaps.

can be directly inherited by the generated data.

3.2. Scaling Up from MegaDepth

There are many RGB image matching benchmarks, represented by MegaDepth (outdoor) [26] and ScanNet (indoor) [5]. These datasets contain millions of image pairs with depth and pose information, which are widely used and have facilitated the development of advanced matching pipelines [9, 10, 28, 47]. Here we choose MegaDepth [26]

as the basic dataset for the following reasons. i) Multimodal perception tasks are typically performed outdoors, and also, the corresponding datasets [21] are from outdoor scenes. ii) MegaDepth demonstrates strong generalization capabilities due to its rich scene coverage and accurate labeling, making it a popular choice for training the models of existing methods [9, 10, 47] to test their generalization. iii) The synthetic MegaDepth makes it convenient to fine-tune those advanced matching methods for multimodal cases. Obviously, we can also generate the data from videos as GIM [40]. However, GIM uses several times the scale of images but merely achieves slight gains in outdoor performance. Considering the high computational cost of generative models, using long videos is not economical.

3.3. Details of Our Data Engine

We subsequently introduce how to use our data engine to generate different modalities from the public MegaDepth dataset. Here we consider the target modalities as commonly used *Infrared*, *Depth*, *Event*, *Normal*, and two *Artistic Styles*. Each modality is combined with RGB to construct a cross-modal pair. Actually, we can combine any two of these modalities to form a matching pair if needed, and any other new modalities we want can also be added.

As depicted in Fig. 3, our data engine consists of three parts: *Source Data*, *Guidance Data*, and *Generative Models*. The source data is multi-viewed RGB images that we want to scale up, *i.e.*, MegaDepth. The guidance data is real image pairs of our target cross-modality, mainly for fine-tuning the generative models. Here, we use publicly aligned data introduced in Tab. 1. As for generative models, we first leverage existing models to directly obtain corresponding modalities for convenience, since recent generative methods have achieved great success in image style transfer [15, 50] and depth or normal generation [3, 54]. As for other modalities, such as infrared, we use the guidance data to fine-tune advanced generative models [15]. Fig. 3 gives the overall process, and the details are as follows:

Infrared: Transferring RGB to infrared is challenging due to the significant variations in their imaging systems, making existing works hard to produce satisfying results [19]. To this end, we turn to a diffusion-based model for help. We use StyleBooth [15] as the basis due to its remarkable performance in style transfer. StyleBooth was originally used to generate artistic styles controlled by an image or text description. In our study, we fine-tuned it using aligned RGB-IR image pairs from the LLVIP [19] and M3FD [29] datasets. We then implement the style tuner with LoRA [18] of rank 256 and standardize the resolution² as 1024×1024 for both input and output. We fine-tune it on 1 NVIDIA 3090 GPU for $210k$ steps with a fixed learning rate 1×10^{-4}

²To meet the resolution, we upscale the longer side of each image to 1024 pixels, then the short side is padded with zero.

and batch size 2. (*More Details Are in Sec. A.1*).

Depth: We directly use DepthAnything V2 [54] of the official model (the large one) to generate high-quality depth images, due to the outstanding performance of monocular depth estimation and the zero-shot ability.

Event: The imaging principle of an event camera is simple, which has independent pixels that respond to brightness changes in their log photocurrent $L \doteq \log(I)$. Specifically, an event $e_k \doteq (\mathbf{x}_k, t_k, p_k)$ is triggered at pixel $\mathbf{x}_k \doteq (x_k, y_k)^\top$ and at time t_k as soon as the brightness increment reaches a temporal contrast threshold $\pm C$, *i.e.*,

$$\Delta L(\mathbf{x}_k, t_k) \doteq L(\mathbf{x}_k, t_k) - L(\mathbf{x}_k, t_k - \Delta t_k), \quad (1)$$

with $\Delta L(\mathbf{x}_k, t_k) = p_k C$, where $C > 0$, Δt_k is the time elapsed since the last event at the same pixel, and the polarity p_k is the sign of the brightness change [12, 27]. In our study, we randomly set $C \in [0.05, 0.5]$, $p_k = \pm 1$ as suggested in [14] to simulate varied sensors and give a random slight motion to compute the event responses.

Normal: The surface normal images are directly generated with DSINE [3], an advanced approach that utilizes the per-pixel ray direction and recasts surface normal estimation as relative rotation estimation between pixels.

Artistic: Our artistic styles include oil paint and sketch, which are implemented with open-source models, *i.e.*, Paint Transformer [30] and Anime2Sketch [50], respectively. Each of them is selected for the stylistic specialization.

With the above settings, we can obtain our data engine $\{\mathcal{F}_{\theta_i}\}_{i=1}^K$ corresponding to above $K = 6$ models. On this basis, and for a pair of RGB images $\{A_0, B_0\}$, we will create two image sets $\mathcal{A} = \{A_i\}_{i=1}^K$, $\mathcal{B} = \{B_i\}_{i=1}^K$ of K modality types. Since our source data MegaDepth [26] contains $40M$ image pairs for image matching, we will create over $480M$ cross-modal image pairs in total, with $\{A_0, B_i\}_{i=1}^K$ or $\{A_i, B_0\}_{i=1}^K$. We term the new dataset as MD-syn. Notably, we can also create any modality pair, such as *Infrared-Event*, if needed. The training and testing sets are split similarly to the original MegaDepth.

4. Modality Invariant Image Matching Model

After constructing MD-syn, the training of our *Modality Invariant Image Matching* (MINIMA) is easy and clear. As shown in Fig. 3, it consists of the following two stages:

- **Stage 1:** Pre-train advanced matching models on multi-view RGB data until they are converged.
- **Stage 2:** Fine-tune on randomly selected cross-modal image pairs with a small learning rate.

We adopt a pre-training and then fine-tuning strategy for the following reasons. First, training from scratch on MD-syn is challenging due to the high variance among different modalities. This requires extensive iterations for convergence. By contrast, training on the RGB dataset is easy.

The pre-trained models can provide good matching priors for hard tasks like multimodal matching, making it converge rapidly, as verified in Fig. 5. In addition, the training on the RGB dataset is well studied [10, 28, 47], whose officially trained models can directly support our fine-tuning.

Since MegaDepth has given birth to numerous matching methods with the taxonomy of sparse, semi-dense, and dense matching, we use three representative models from them as our basic models, *i.e.*, LightGlue (LG) [28], LoFTR [41], and RoMa [10]. We will release our three models for sparse, semi-dense, and dense matching, termed as $\text{MINIMA}_{\text{LG}}$, $\text{MINIMA}_{\text{LoFTR}}$, and $\text{MINIMA}_{\text{RoMa}}$. Those models will be evaluated with in-domain and zero-shot matching on synthetic and real cross-modal datasets.

5. Experiments

5.1. Implementation Details

We directly use the official models of LightGlue [28], LoFTR [41] and RoMa [10] as the pre-trained models, and then fine-tune them with our MD-syn. All the models are trained on 4 RTX 3090 GPUs, with batch sizes being 32, 8, and 12, respectively. Based on the pre-trained models, we first individually fine-tune for each modality. The unified model is fine-tuned on randomly selected modality pairs in each iteration, which is used for the generalization ability test. Notably, we only use *RGB-IR*, *RGB-Depth*, *RGB-Normal* modality pairs for training, which is sufficient to achieve satisfying performance. (*Refer to Appendices for More Details*).

Datasets. The used datasets include our synthetic MD-syn and 5 multimodal datasets of real images, which contain 19 cross-modal cases in total. In particular, 1) *MD-syn* splits two scenes for test, which consist of 1500 image pairs for each cross-modal case following the setting of the original Megadepth. It contains 6 cross-modal cases: 3 of them (RGB-IR/-Depth/-Normal) are used for the in-domain test, while the rest are for zero-shot evaluation. 2) *METU-VisTIR* [43] is a real RGB-IR dataset containing 2590 real image pairs with camera poses attached. 3) *DIODE* [44] is a real RGB-Depth/Noraml dataset, containing 27858 fully aligned image pairs. 4) *DSEC* [46] provides 60 sequences of RGB-Event videos. Three sequences are selected as our test set, which generates 100 RGB-Event pairs for testing. We rectify the frames following the instructions of the authors to obtain aligned image pairs. To test the generalization in 5) *Remote Sensing* and 6) *Medical* domains, we use MMIM datasets [21] for evaluation, where the ground truths are manually labeled matches for homography estimation. The *Remote Sensing* domain consists of 7 cross-modal cases such as Optical-SAR, Optical-Map, Optical-Depth, *etc.* The *Medical* domain consists of 6 cross-modal

cases such as Retina, MRI-PET, CT-SPECT, *etc.*

Evaluation Protocols. The used datasets exhibit different labels for matching, such as camera poses and 2-D homography matrices. For two-viewed datasets, such as our MD-syn and METU-VisTIR, the recovered poses by matches are evaluated to measure the matching accuracy. We report the area under the curve (AUC) of the pose error at thresholds $\{5^\circ, 10^\circ, 20^\circ\}$. As for homography, similar to [41], we collect the mean projection error of four corner points and report the AUC under thresholds $\{3\text{px}, 5\text{px}, 10\text{px}\}$ for evaluation. Notably, for those aligned image pairs, we impose synthetic homography matrices on one image to imitate deformations, which are finished before evaluation to maintain fairness. Then, we try to recover the homography matrix. And we uniformly resize all images with their long dimension equal to 640. All the experiments are performed on a single RTX 3090 GPU for accuracy and runtime tests. For all baselines, we employ the same RANSAC [11] settings as a robust homography or pose estimator for fair comparison.

Baselines. Following [47], we select representative methods from the matching pipelines of sparse, semi-dense, and dense matching. 1) For spare keypoint detection and matching methods, we choose SuperGlue [35], LightGlue (LG) [28], OmniGlue [20], and LG-based GIM [40] for comparison. All of them (including our $\text{MINIMA}_{\text{LG}}$) use SuperPoint as the keypoint detector (the maximum number of extracted keypoints is set as 2048). We also take ReDFeat [7] into account as it is a deep method designed for multimodal image matching. In addition, three handcrafted multimodal matching methods, including RIFT [23], SRIT [25], LNIFT [24], are also used. However, we only test them (including OmniGlue) on the real RGB-IR dataset due to their poor accuracy and huge time costs. 2) Semi-dense matching methods includes LoFTR [41], ELoFTR [47], XoFTR [43], and GIM_{LoFTR}, where XoFTR is tailored for RGB-IR image matching. 3) As for dense matching, DKM [9], GIM_{DKM} [40] and recent SOTA matcher RoMa [10] are used for comparison.

5.2. Evaluate on Our MD-syn

We first test the matching methods on MD-syn, a multimodal image matching dataset synthesized by our data engine. Tab. 2 reports the qualitative results. It shows that our MINIMA can largely enhance the cross-modal ability of the baselines. However, we achieve weak advantages for RGB-Sketch and RGB-Paint since these two artistic modalities are more similar to RGB. As the table revealed, GIM shows poor generalization for multimodal cases, since it is overfitted on RGB videos. ReDFeat performs not well on new scenes and even fails in the event case. As for the LoFTR series, the original LoFTR and ELoFTR are worse than SuperGlue and LG. Because Edge or shape information is more crucial for multimodal image matching, it is difficult for

Table 2. **Full Results on Our Synthetic Dataset.** The AUC of the pose error in percentage is reported. The best and second of each category 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° |
| Sparse | SuperGlue [35] | 7.49 | 17.51 | <u>33.54</u> | <u>3.06</u> | <u>6.94</u> | <u>13.70</u> | 11.53 | 24.42 | 41.85 | 10.38 | <u>23.48</u> | <u>41.63</u> | 21.52 | 37.99 | 56.17 | 11.35 | 24.15 | 42.51 |
| | LightGlue (LG) [28] | 7.64 | 17.73 | 32.86 | 1.19 | 2.87 | 6.42 | 12.32 | 24.93 | 41.86 | 10.11 | 22.40 | 39.33 | 26.77 | 44.47 | 62.00 | <u>13.93</u> | 27.99 | 46.16 |
| | ReDFeat [7] | 2.75 | 8.56 | 20.90 | 2.20 | 6.36 | 15.25 | 2.56 | 7.25 | 17.79 | 0.00 | 0.00 | 0.00 | 5.26 | 13.91 | 29.01 | 2.73 | 7.32 | 17.83 |
| | GIM _{LG} [40] | 8.40 | <u>18.88</u> | 33.20 | 0.00 | 0.00 | 0.12 | 12.03 | 23.93 | 38.53 | 6.75 | 14.19 | 23.81 | 28.80 | 46.82 | 63.94 | 13.18 | 26.84 | 43.45 |
| Semi-Dense | MINIMA _{LG} | 14.74 | 30.24 | 49.22 | 16.19 | 32.53 | 51.76 | 20.47 | 37.33 | 56.17 | 19.00 | 36.27 | 54.97 | <u>27.51</u> | <u>45.71</u> | <u>63.77</u> | 16.39 | 32.85 | 51.65 |
| | LoFTR [41] | 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 [43] | <u>17.85</u> | <u>32.21</u> | 49.53 | <u>12.82</u> | <u>23.10</u> | <u>36.02</u> | <u>22.74</u> | <u>38.35</u> | <u>54.71</u> | 33.33 | 51.61 | 67.49 | 44.18 | 61.39 | 75.07 | 3.73 | 7.54 | 14.48 |
| | ELOFTR [47] | 6.73 | 14.59 | 27.36 | 0.25 | 0.79 | 3.32 | 11.20 | 21.67 | 36.86 | 9.25 | 20.39 | 37.56 | <u>43.86</u> | <u>61.09</u> | <u>74.84</u> | 14.09 | 25.11 | 39.44 |
| | GIM _{LoFTR} [40] | 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 |
| Dense | MINIMA _{LoFTR} | 18.07 | 32.36 | <u>48.42</u> | 14.70 | 28.81 | 46.23 | 27.65 | 44.26 | 59.88 | <u>18.14</u> | <u>32.74</u> | <u>49.11</u> | <u>36.07</u> | <u>53.54</u> | <u>68.47</u> | <u>7.79</u> | <u>15.45</u> | <u>27.39</u> |
| | DKM [9] | 15.68 | 29.46 | 46.11 | 0.10 | 0.38 | 1.92 | 23.23 | 39.28 | 55.22 | 10.18 | 18.14 | 27.78 | 56.91 | 72.25 | 83.31 | 29.64 | 44.73 | 58.57 |
| | GIM _{DKM} [40] | 11.23 | 22.72 | 37.93 | 1.42 | 4.07 | 10.86 | 14.09 | 25.81 | 40.55 | 22.86 | 38.30 | 53.58 | 50.89 | 67.12 | 79.02 | 28.22 | 43.49 | 58.06 |
| | RoMa [10] | <u>20.27</u> | <u>35.99</u> | <u>54.02</u> | <u>10.21</u> | <u>22.75</u> | <u>39.43</u> | <u>40.99</u> | <u>59.48</u> | <u>74.19</u> | <u>40.86</u> | <u>58.87</u> | <u>73.35</u> | <u>58.49</u> | <u>73.90</u> | <u>84.80</u> | 41.30 | 58.36 | 72.70 |
| Dense | MINIMA _{RoMa} | 24.33 | 40.94 | 58.33 | 29.56 | 48.58 | 65.87 | 47.10 | 64.48 | 77.90 | 43.83 | 61.48 | 75.21 | 59.17 | 74.30 | 84.86 | <u>40.09</u> | <u>57.21</u> | <u>71.96</u> |

Table 3. **Evaluation on Real RGB-IR Dataset (METU-VisTIR) [43] with Pose Estimation.** The AUC of the pose error in percentage is reported. The average runtime is listed in the last column.

| Category | Method | Pose estimation AUC | | | Time (ms) |
|------------|------------------------------------|---------------------|--------------|--------------|--------------|
| | | @5° | @10° | @20° | |
| Sparse | RIFT [23](TIP 19) | 0.05 | 0.27 | 0.90 | 13k |
| | SRIT [25](ISPRS 23) | 0.00 | 0.08 | 0.37 | 1.9k |
| | LNIFT [24](TGRS 22) | 0.02 | 0.09 | 0.43 | 1.2k |
| | SuperGlue [35](CVPR 20) | 4.30 | 9.26 | <u>17.21</u> | 86.1 |
| | ReDFeat [7](TIP 23) | 1.71 | 4.57 | 10.85 | 235.8 |
| | LightGlue (LG) [28](ICCV 23) | 2.17 | 5.37 | 11.21 | 57.7 |
| | GIM _{LG} [40](ICLR 24) | 2.43 | 5.85 | 10.58 | 42.9 |
| | OmniGlue [20](CVPR 24) | 1.48 | 4.13 | 10.11 | 3k |
| Semi-Dense | MINIMA _{LG} | 19.14 | 37.17 | 55.51 | 58.6 |
| | LoFTR [41](CVPR 21) | 2.88 | 6.94 | 14.95 | 61.6 |
| | GIM _{LoFTR} [40](ICLR 24) | 0.43 | 1.06 | 2.99 | 69.5 |
| | ELOFTR [47](CVPR 24) | 2.88 | 7.88 | 17.72 | 46.6 |
| | XoFTR [43](CVPR 24) | 18.47 | 34.64 | 51.50 | 62.7 |
| Dense | MINIMA _{LoFTR} | <u>15.61</u> | <u>30.84</u> | <u>47.87</u> | 71.6 |
| | DKM [9](CVPR 23) | 6.76 | 13.69 | 22.53 | 485.3 |
| | GIM _{DKM} [40](ICLR 24) | 5.08 | 12.30 | 23.69 | 792.2 |
| | RoMa [10](CVPR 24) | <u>25.61</u> | <u>48.12</u> | <u>68.37</u> | 639.1 |
| Dense | MINIMA _{RoMa} | 37.45 | 60.70 | 78.00 | 633.3 |

semi-dense methods to build matches among textureless areas. XoFTR achieves competitive results, as it is pre-trained on sufficient multi-spectral image pairs and equipped with many advanced designs. As for dense matching, DKM and GIM_{DKM} perform poorly on four cross-modal cases due to the huge modality gaps among them. While original RoMa exhibits good generalization, mainly because of the use of DINOv2 [33] that has seen numerous types of images during pre-training. Our MINIMA still obtains significant enhancement over RoMa. Moreover, we also evaluate those methods back to the original Megadepth-1500

Table 4. **Evaluation on Real RGB-Depth Dataset (DIODE) [44] with Homography Estimation.** The AUC of the projective error in percentage is reported.

| Category | Method | Homo. estimation AUC | | |
|------------|------------------------------------|----------------------|--------------|--------------|
| | | @3px | @5px | @10px |
| Sparse | SuperGlue [35](CVPR 20) | <u>1.77</u> | 6.83 | <u>21.15</u> |
| | ReDFeat [7](TIP 23) | 1.01 | 4.58 | 16.30 |
| | LightGlue (LG) [28](ICCV 23) | 0.79 | 3.30 | 11.26 |
| | GIM _{LG} [40](ICLR 24) | 0.30 | 1.14 | 3.65 |
| | MINIMA _{LG} | 8.71 | 26.80 | 55.97 |
| Semi-Dense | LoFTR [41](CVPR 21) | 0.97 | 4.20 | 15.16 |
| | GIM _{LoFTR} [40](ICLR 24) | 0.00 | 0.25 | 1.15 |
| | ELOFTR [47](CVPR 24) | 0.82 | 4.09 | 16.69 |
| | XoFTR [43](CVPR 24) | 11.03 | 27.24 | 51.60 |
| | MINIMA _{LoFTR} | <u>5.35</u> | <u>18.65</u> | <u>44.85</u> |
| Dense | DKM [9](CVPR 23) | 1.29 | 4.23 | 11.78 |
| | GIM _{DKM} [40](ICLR 24) | 1.90 | 6.34 | 17.96 |
| | RoMa [10](CVPR 24) | <u>9.21</u> | <u>24.64</u> | <u>49.31</u> |
| | MINIMA _{RoMa} | 28.98 | 50.88 | 72.54 |

(Refer to Sec. C.3 Tab. A5 for Details).

5.3. In Domain Image Matching

We next conduct in-domain tests, *i.e.*, training on synthetic data but testing on real data of the same modality. Two real datasets are used, including RGB-IR (METU-VisTIR [43]) for pose estimation and RGB-Depth (DIODE [44]) for homography estimation. The results are in Tab. 3 and Tab. 4.

For RGB-IR test, we add three handcrafted methods, *i.e.*, RIFT [23], SRIT [25] and LNIFT [24], and a unified sparse matching model OmniGlue [20]. We find they achieve extremely low accuracy and require dozens of time, thus we drop them on other experiments. For sparse methods, our MINIMA_{LG} achieves amazing enhancement, with

Table 5. **Zero-shot Matching on Real Dataset with Homography Estimation.** The AUC of the corner error in percentage is reported. The best and second of each category are masked as **Bold** and Underline, respectively.

| Category | Method | Medical | | | Remote Sensing | | | RGB-Event | | |
|------------|---------------------------|--------------|--------------|--------------|----------------|--------------|--------------|-------------|-------------|--------------|
| | | @3px | @5px | @10px | @3px | @5px | @10px | @3px | @5px | @10px |
| Sparse | SuperGlue [35] | 30.72 | 36.18 | 44.66 | <u>18.34</u> | 27.47 | <u>45.59</u> | 0.00 | 0.67 | <u>8.00</u> |
| | LightGlue (LG) [28] | 35.47 | 42.37 | 49.48 | <u>16.22</u> | <u>27.51</u> | <u>44.62</u> | 0.00 | 0.67 | 7.02 |
| | ReDFeat [7] | 38.55 | 44.26 | <u>50.93</u> | 15.99 | 23.95 | 43.72 | <u>0.55</u> | 0.97 | 6.07 |
| | GIM _{LG} [40] | 24.32 | 27.88 | 33.84 | 11.09 | 17.44 | 27.18 | 0.57 | <u>1.08</u> | 5.54 |
| | MINIMA _{LG} | <u>37.95</u> | <u>44.08</u> | 52.50 | <u>23.53</u> | 38.40 | 58.74 | 0.52 | 2.27 | 12.82 |
| Semi-Dense | LoFTR [41] | 38.42 | 43.89 | 50.13 | <u>24.13</u> | 33.80 | 50.79 | 0.00 | 0.00 | 3.59 |
| | XoFTR [43] | 39.67 | 45.60 | <u>52.32</u> | 27.35 | 39.58 | <u>56.63</u> | 0.00 | <u>1.37</u> | 12.64 |
| | ELoFTR [47] | 34.57 | 41.66 | 49.08 | 16.45 | 29.65 | 46.74 | <u>0.64</u> | 1.34 | 7.78 |
| | GIM _{LoFTR} [40] | <u>39.51</u> | 44.40 | 48.94 | 17.96 | 27.41 | 37.29 | 0.00 | 0.55 | 1.19 |
| | MINIMA _{LoFTR} | 39.67 | <u>45.33</u> | 52.77 | 23.32 | <u>35.18</u> | 56.81 | 0.81 | 2.49 | <u>11.75</u> |
| Dense | DKM [9] | 39.43 | 45.00 | 51.78 | 26.44 | 35.82 | 51.20 | 0.00 | 0.00 | 0.00 |
| | GIM _{DKM} [40] | 37.78 | 43.46 | 48.87 | 21.19 | 30.28 | 47.68 | 0.00 | 0.66 | 7.04 |
| | RoMa [10] | 39.62 | <u>45.13</u> | <u>53.75</u> | <u>29.24</u> | <u>40.50</u> | <u>57.84</u> | 0.85 | <u>1.69</u> | <u>10.71</u> |
| | MINIMA _{RoMa} | 39.17 | 45.92 | 57.55 | <u>32.55</u> | 44.68 | 64.38 | 0.54 | <u>3.51</u> | 17.07 |

AUC increased over 400%. Mostly, it even beats the SOTA semi-dense method XoFTR. As for semi-dense matching, XoFTR achieves the best performance. This is attributed to its pre-training on sufficient multi-spectral image pairs, the use of an effective data augmentation strategy, and specific designs incorporated in both the training and matching stages. In dense matching, our MINIMA combined with RoMa outperforms all other pipelines consistently with large margins. The runtime of each method is also listed. The results show that sparse and semi-dense methods (except for handcrafted methods, ReDFeat, and OmniGlue) are often more efficient due to their fewer points to match. ELoFTR is faster than the sparse methods due to its efficient designs. This trend is consistent with existing works [10, 47].

The same trends are obtained in the RGB-Depth matching as in Tab. 4. To be specific, our semi-dense method is worse than XoFTR. That is because depth data is more challenging, and our MINIMA is based on LoFTR, a representative but old model without any fancy designs. But we largely enhance the original LoFTR from 4.2 to 18.65 @5 px. The overall performance on all real cross-modal data is concluded in Fig. 1, which reveals the promising generalization of our MINIMA.

5.4. Zero-shot Matching for Unseen Modality

We next extend to zero-shot matching. 1) *Medical tasks* consist of 6 modality pairs such as Retina, CT-SPECT, etc. 2) *Remote Sensing tasks* consist of 7 cases such as Optical-SAR, Optical-Map, etc. 3) *RGB-Event* case is from DSEC [46]. 1) and 2) are both from MMIM [21] datasets.

The quantitative results are outlined in Tab. 5. For medical scenes, almost all the methods have close accuracy since the datasets are either too easy or too difficult. But our MINIMA_{LG} still exhibits a few advantages. As for re-

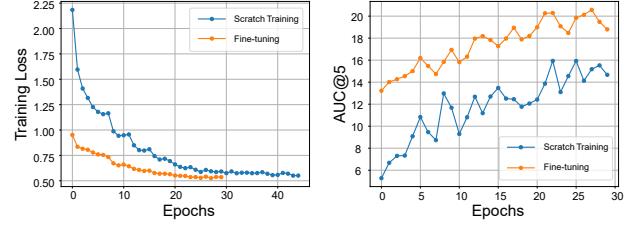


Figure 5. **Training Loss and AUC@5° w.r.t. Epochs, using Scratch Training and Fine-tuning.** The basic model is LoFTR. The test set is our synthetic RGB-IR of MD-syn.

mote sensing cases, our MINIMA achieves large gains for sparse and dense matching. While the semi-dense matcher MINIMA_{LoFTR} falls behind XoFTR for the same reason. As for the RGB-Event matching, the task is extremely challenging due to the large modality gap. Despite this, our proposed MINIMA performs good capacity of matching them. Some qualitative results are shown in Fig. 2, and more results are in Sec. C.5, which demonstrate that our MINIMA can establish a high number and ratio of correct matches for real cross-modal image pairs.

5.5. 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 [26]. We use LoFTR as the basic model, and the training set is our synthetic RGB-IR/Depth/Nominal. Statistic results are shown in Fig. 5, which reveal that the fine-tuning strategy can converge more rapidly since the pre-trained model can provide good matching priors for challenging cross-modal tasks.

Table 6. **Ablation Studies.** Test on Synthetic RGB-IR, Real RGB-IR, and Real RGB-Depth data, with different training settings.

| Training Strategy | Syn RGB-IR | Rel RGB-IR | Rel RGB-D |
|----------------------------------|--------------|--------------|--------------|
| | AUC@10° | AUC@10° | AUC@10px |
| Basic Model: LoFTR (LT) [41] | 12.58 | 6.94 | 15.16 |
| (1) Train from scratch on syn IR | 23.63 | 21.41 | 30.04 |
| (2) LT + real IR | 6.28 | 9.78 | 32.93 |
| (3) LT + syn IR | 29.43 | 29.55 | 39.23 |
| (4) LT + syn Depth | 17.30 | 15.12 | 36.06 |
| (5) LT + syn IR/Depth/Normal | 32.36 | 30.84 | 44.85 |

5.6. Ablation Studies

In this part, we conduct ablation studies to analyze the superiority of our MINIMA. The results on synthetic RGB-IR, real RGB-IR, and real RGB-Depth data are reported in Tab. 6. We use LoFTR (LT) as the basic model, which serves as the pre-trained model for (2)-(5). (1) Directly train LT on synthetic RGB-IR from scratch. (2) Fine-tune LT on real RGB-IR datasets (LLVIP and M3FD). (3) Fine-tune LT only with our synthetic RGB-IR. (4) Fine-tune LT only with our synthetic RGB-Depth. (5) Fine-tune LT on mixed data of our synthetic RGB-IR, RGB-Depth, and RGB-Normal. The results of (1) and (3) reveal that training from scratch is worse than fine-tuning. (2) and (3) demonstrate the advantages of our synthetic data against real datasets. (4) reveals that only fine-tuning on synthetic RGB-Depth can well generalize to other cross-modal cases, even better than test (2) on real RGB-IR data. (5) and (3) reveal that different synthetic data can cooperate for better performance. Our full model can largely enhance the generalization ability. More combinations of training datasets are evaluated in Tab. A2.

5.7. Discussion on Possible Limitations

Our objective is to generate pseudo modalities to form a large multimodal dataset, which would produce two possible limitations: *i) The gap between real and pseudo modality.* *ii) The fake information during generation.* Fortunately, these two possible limitations have little impact on our task. First, multimodal images intrinsically vary in pixel intensity distributions [51]. This property is well exhibited in our generated modalities (see Fig. 4), which plays an important role in training a general matching model. Existing diffusion-based methods [15, 16] can generate high-quality images of the target modality, making the pseudo modality much closer to the real. Extensive experiments will verify the high quality of our generated data. As for the generated fake information, it can well imitate the multimodal cases, e.g. the target is visible in infrared but not in RGB, which may help to enhance the robustness of the trained model.

6. Conclusion

This paper presents a unified matching framework, named MINIMA, for any cross-modal cases. It is achieved by fill-

ing the data gap using an effective data engine that freely scales up cheap RGB data into a large multimodal one. The constructed MD-syn dataset contains rich scenarios and precise match labels, and supports the training of any advanced matching models, significantly improving cross-modal performance and zero-shot ability in unseen cross-modal cases.

References

- [1] NG Aditya, PB Dhrupal, Jehan Shalabi, Shubhankar Jape, Xueji Wang, and Zubin Jacob. Thermal voyager: A comparative study of rgb and thermal cameras for night-time autonomous navigation. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, pages 14116–14122. IEEE, 2024.
- [2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016.
- [3] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2024.
- [4] 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 *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 6301–6310, 2021.
- [5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nie10ner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.
- [6] Xin Deng, Enpeng Liu, Chao Gao, Shengxi Li, Shuhang Gu, and Mai Xu. Crossromo: Cross-modality and cross-resolution homography estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [7] Yuxin Deng and Jiayi Ma. Redfeat: Recoupling detection and description for multimodal feature learning. *IEEE Transactions on Image Processing*, 32:591–602, 2022.
- [8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabenovich. Superpoint: Self-supervised interest point detection and description. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition Workshop*, pages 224–236, 2018.
- [9] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023.
- [10] Johan Edstedt, Qiyu Sun, Georg Bökmä, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024.
- [11] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to

- image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 6
- [12] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2020. 5
- [13] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024. 1
- [14] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 5
- [15] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. *arXiv preprint arXiv:2404.12154*, 2024. 3, 5, 9, 12
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 3, 9
- [17] Zhuolu Hou, Yuxuan Liu, and Li Zhang. Pos-gift: A geometric and intensity-invariant feature transformation for multimodal images. *Information Fusion*, 102:102027, 2024. 3
- [18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proc. of Intl. Conf. on Learning Representations*, 2022. 5
- [19] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 3496–3504, 2021. 1, 2, 3, 5, 12
- [20] Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. Omnidglue: Generalizable feature matching with foundation model guidance. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 19865–19875, 2024. 2, 6, 7, 15
- [21] 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. 1, 2, 3, 5, 6, 8, 12, 15, 16
- [22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of Advances in Neural Information Processing Systems*, 2012. 12
- [23] Jiayuan Li, Qingwu Hu, and Mingyao Ai. Rift: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29: 3296–3310, 2019. 3, 6, 7
- [24] Jiayuan Li, Wangyi Xu, Pengcheng Shi, Yongjun Zhang, and Qingwu Hu. Lnift: Locally normalized image for rotation invariant multimodal feature matching. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 6, 7
- [25] Jiayuan Li, Qingwu Hu, and Yongjun Zhang. Multimodal image matching: A scale-invariant algorithm and an open dataset. *ISPRS J Photogramm*, 204:77–88, 2023. 6, 7
- [26] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 1, 3, 4, 5, 8, 14
- [27] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbrück. A 128 × 128 120db 15μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 5
- [28] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 17627–17638, 2023. 2, 3, 4, 6, 7, 8, 12, 14, 15
- [29] 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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022. 2, 3, 5, 12
- [30] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Ruifeng Deng, Xin Li, Errui Ding, and Hao Wang. Paint transformer: Feed forward neural painting with stroke prediction. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 6598–6607, 2021. 5
- [31] Yuyan Liu, Wei He, and Hongyan Zhang. Grid: Guided refinement for detector-free multimodal image matching. *IEEE Transactions on Image Processing*, 2024. 3
- [32] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021. 1, 2, 3
- [33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7
- [34] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 14
- [35] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 1, 2, 3, 6, 7, 8, 14, 15
- [36] 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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018. 14
- [37] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2016. 3, 4, 14
- [38] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proc. of European Conference on Computer Vision*, 2016. 3, 4, 14

- [39] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, 2020. Version 0.3.0. 12
- [40] 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 *Proc. of Intl. Conf. on Learning Representations*, 2024. 3, 4, 5, 6, 7, 8, 14
- [41] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 1, 3, 6, 7, 8, 9, 13, 14
- [42] 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. 12
- [43] Önder Tuzcuoğlu, Aybora Köksal, Buğra Sofu, Sinan Kalkan, and A Aydin Alatan. Xoftr: Cross-modal feature matching transformer. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 4275–4286, 2024. 1, 2, 3, 6, 7, 8, 12, 13, 14, 15
- [44] 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. 6, 7, 15
- [45] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proc. of Intl. Joint Conf. on Artificial Intelligence*, pages 3508–3515, 2022. 12
- [46] 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 Transactions on Cybernetics*, 2023. 6, 8, 15, 16
- [47] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 21666–21675, 2024. 1, 3, 4, 5, 6, 7, 8, 13, 14
- [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. 12
- [49] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluiu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 19156–19166, 2024. 1
- [50] Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P Allebach. Adversarial open domain adaptation for sketch-to-photo synthesis. In *Proc. of IEEE Winter Conf. on Applications of Computer Vision*, pages 1434–1444, 2022. 5
- [51] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023. 1, 9
- [52] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 11069–11079, 2023. 1
- [53] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2, 3
- [54] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Proc. of Advances in Neural Information Processing Systems*, 2024. 3, 5
- [55] Yongxiang Yao, Yongjun Zhang, Yi Wan, Xinyi Liu, Xiaohu Yan, and Jiayuan Li. Multi-modal remote sensing image matching considering co-occurrence filter. *IEEE Transactions on Image Processing*, 31:2584–2597, 2022. 3
- [56] Yuanxin Ye, Lorenzo Bruzzone, Jie Shan, Francesca Bovolo, and Qing Zhu. Fast and robust matching for multimodal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9059–9070, 2019. 3
- [57] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1
- [58] Kaining Zhang and Jiayi Ma. Sparse-to-dense multimodal image registration via multi-task learning. In *Proc. of Intl. Conf. on Machine Learning*, 2024. 2, 4
- [59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 12
- [60] Shihua Zhang and Jiayi Ma. Convmatch: Rethinking network design for two-view correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 14, 15
- [61] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023. 1
- [62] Yuxiang Zhang, Yang Zhao, Yanni Dong, and Bo Du. Self-supervised pretraining via multimodality images with transformer for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023. 1
- [63] Kaichen Zhou, Changhao Chen, Bing Wang, Muhamad Risqi U Saputra, Niki Trigoni, and Andrew Markham. Vm-loc: Variational fusion for learning-based multimodal camera localization. In *Proc. of the AAAI Conf. on Artificial Intelligence*, pages 6165–6173, 2021. 1
- [64] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023. 1

Appendices

We first provide more details of our data engine and proposed MINIMA. We also provide additional experimental results, including applying our MINIMA to the *Visual Localization* and more visible matching results.

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 [15] for fine-tuning due to its significant performance in style transfer. The fine-tuning process utilizes 80% real RGB-IR pairs from LLVIP [19] and M3FD [29], 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 [42] 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) [43] and CPSTN (IJCAI 22) [45] 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) [48], LPIPS (Learned Perceptual Image Patch Similarity) [59] with AlexNet [22], and PyTorch FID (Frechet Inception Distance) [39]. 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.

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



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

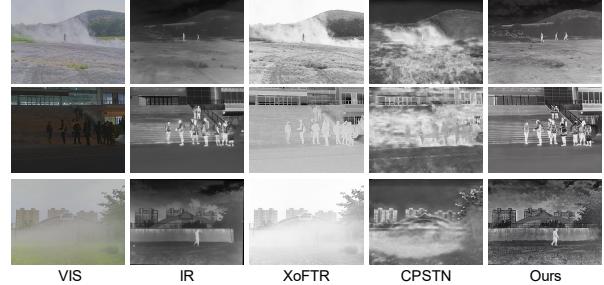


Figure A2. **Visualization Results of Infrared Generation on M3FD.** The first two columns are real RGB and Infrared images.

Table A1. **Quantitative Evaluation of Infrared Generation with Different Metrics.** The test datasets are LLVIP [19], M3FD [29] and MSRS [42]. CPSTN (IJCAI 22) [45] and XoFTR (CVPR 24) [43] are used for comparison. **Bold** indicates the best.

| Data | Method | PSNR \uparrow | SSIM \uparrow | LPIPS \downarrow | FID-2048 \downarrow |
|-------|-------------|-----------------|-----------------|--------------------|-----------------------|
| 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 |

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 pairs are dropped in the training set.

B. Details of MINIMA

The details of our fine-tuning stage are as follows. i) **Light-Glue** (LG) [28]. We use SuperPoint [8] 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 [21]. Just fine-tuning LightGlue can achieve promising performance, as demon-

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 | | |
| MINIMA _{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 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 36.34 | 55.93 | 12.74 | 58.41 | 52.45 | 43.17 |
| MINIMA _{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 |
| | | | | ✓ | ✓ | ✓ | 30.80 | 48.55 | 12.44 | 56.04 | 51.82 | 39.93 |
| | | | | ✓ | ✓ | ✓ | 30.61 | 45.10 | 11.83 | 55.33 | 52.19 | 39.01 |
| MINIMA _{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 |
| | | | | ✓ | ✓ | ✓ | 61.27 | 73.80 | 11.02 | 65.01 | 55.04 | 53.23 |
| | | | | ✓ | ✓ | ✓ | 60.43 | 72.83 | 12.98 | 64.80 | 57.92 | 53.79 |

strated by our MINIMA_{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** [41] and **RoMa** [10]. 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 [47] and XoFTR [43], denoted as MINIMA_{ELoFTR} and MINIMA_{XoFTR}. And they are fine-tuned for 20 epochs and 5 epochs, respectively. Their learning rates are similar to our MINIMA_{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 [47] and XoFTR [43] on the generated data, obtaining MINIMA_{ELoFTR} and MINIMA_{XoFTR}. The corresponding results of semi-dense matching are reported 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

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 [41] | 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 [43] | 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 [47] | 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 | <u>14.09</u> | <u>25.11</u> | <u>39.44</u> |
| | GIM _{LoFTR} [40] | 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 |
| | MINIMA _{LoFTR} | <u>18.07</u> | <u>32.36</u> | 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 |
| | MINIMA _{XoFTR} | 18.97 | 34.36 | 51.72 | 24.47 | 40.90 | 58.36 | 30.47 | 47.90 | 64.64 | <u>31.14</u> | <u>49.39</u> | <u>65.71</u> | <u>42.91</u> | <u>60.77</u> | <u>75.00</u> | 5.61 | 11.56 | 20.95 |
| | MINIMA _{ELoFTR} | 13.14 | 26.36 | 43.63 | <u>16.59</u> | <u>32.26</u> | <u>50.37</u> | <u>29.72</u> | <u>47.47</u> | <u>63.72</u> | 15.66 | 30.72 | 48.73 | 41.64 | 59.63 | 73.73 | 15.02 | 27.02 | 41.62 |

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 [41] | 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 [43] | <u>18.47</u> | <u>34.64</u> | <u>51.5</u> | <u>11.03</u> | <u>27.24</u> | <u>51.60</u> | 39.67 | 45.60 | <u>52.32</u> | 27.35 | 39.58 | <u>56.63</u> | 0.00 | 1.37 | <u>12.64</u> |
| | ELoFTR [47] | 2.88 | 7.88 | 17.72 | 0.82 | 4.09 | 16.69 | 34.57 | 41.66 | 49.08 | 16.45 | 29.65 | 46.74 | <u>0.64</u> | 1.34 | 7.78 |
| | GIM _{LoFTR} [40] | 0.43 | 1.06 | 2.99 | 0.00 | 0.25 | 1.15 | <u>39.51</u> | 44.40 | 48.94 | 17.96 | 27.41 | 37.29 | 0.00 | 0.55 | 1.19 |
| | MINIMA _{LoFTR} | 15.61 | 30.84 | 47.87 | 5.35 | 18.65 | 44.85 | 39.67 | <u>45.33</u> | 52.77 | 23.32 | 35.18 | 56.81 | 0.81 | 2.49 | 11.75 |
| | MINIMA _{XoFTR} | 19.38 | 35.82 | 52.94 | 11.76 | 29.48 | 55.05 | 39.33 | 44.92 | 52.09 | <u>25.19</u> | <u>37.86</u> | 54.36 | 0.00 | <u>1.92</u> | 15.23 |
| | MINIMA _{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 [35] (CVPR 20) | 49.7 | 67.1 | 80.6 |
| | LightGlue (LG) [28] (ICCV 23) | 49.9 | 67.0 | 80.1 |
| | GIM _{LG} [40] (ICLR 24) | 41.3 | 60.7 | 75.9 |
| | MINIMA _{LG} | 47.3 | 65.0 | 78.6 |
| Semi-Dense | LoFTR [41] (CVPR 21) | 53.6 | 69.9 | 82.0 |
| | GIM _{LoFTR} [40] (ICLR 24) | 51.3 | 68.5 | 81.1 |
| | ELoFTR [47] (CVPR 24) | 56.4 | 72.2 | 83.5 |
| | XoFTR [43] (CVPR 24) | 45.8 | 61.7 | 74.0 |
| Dense | MINIMA _{LoFTR} | 29.9 | 45.3 | 59.5 |
| | MINIMA _{ELoFTR} | 51.0 | 68.1 | 80.3 |
| | MINIMA _{XoFTR} | 44.5 | 60.0 | 72.3 |
| | DKM [9] (CVPR 23) | 60.4 | 74.9 | 85.1 |
| | GIM _{DKM} [40] (ICLR 24) | 60.7 | 75.5 | 85.9 |
| | RoMa [10] (CVPR 24) | 62.6 | 76.7 | 86.3 |
| | MINIMA _{RoMa} | 61.7 | 76.5 | 86.4 |

original MegaDepth-1500 [26]. We use the same settings as described in [28, 41]. 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, re-

Table A6. **Visual Localization on Aachen Day-Night V1.0** [36]

| Method | Day | | Night | | |
|-------------------------------|-------------|-------------|-------------|-------------|-------------|
| | (0.25m, 2°) | (0.5m, 5°) | (5m, 10°) | (0.25m, 2°) | (0.5m, 5°) |
| MNN | 86.9 | 92.0 | 95.5 | 73.5 | 79.6 |
| SuperGlue [35] (CVPR 20) | 87.9 | 95.0 | 97.9 | 84.7 | 92.9 |
| SGMNet [4] (ICCV 21) | 86.5 | 93.7 | 97.2 | 82.7 | 91.8 |
| LightGlue (LG) [28] (ICCV 23) | 88.0 | 93.8 | 97.5 | 84.7 | 91.8 |
| ConvMatch [60] (TPAMI 23) | 88.1 | 94.4 | 97.3 | 79.6 | 88.8 |
| MINIMA _{LG} | 88.3 | 94.7 | 98.3 | 85.7 | 92.9 |
| | | | | | 100.0 |

vealing that our MINIMA can well maintain the ability of RGB-only matching, except for LoFTR.

C.4. 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.

Following [28, 34], we integrate different matching methods into the official HLoc pipeline [34] to achieve localization. Specifically, with COLMAP [37, 38] toolbox, we first triangulate a 3D point cloud for all reference images with known poses and calibration, then retrieve 20 ref-

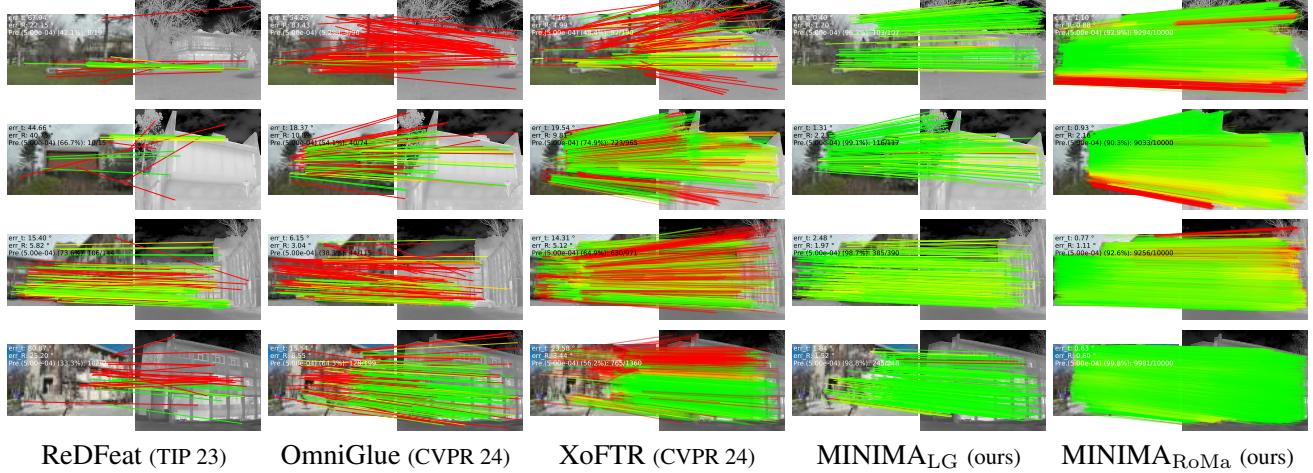


Figure A3. Qualitative Results on Real RGB-IR Image Pairs of METU-VisTIR [43]. The red lines indicate false matches.

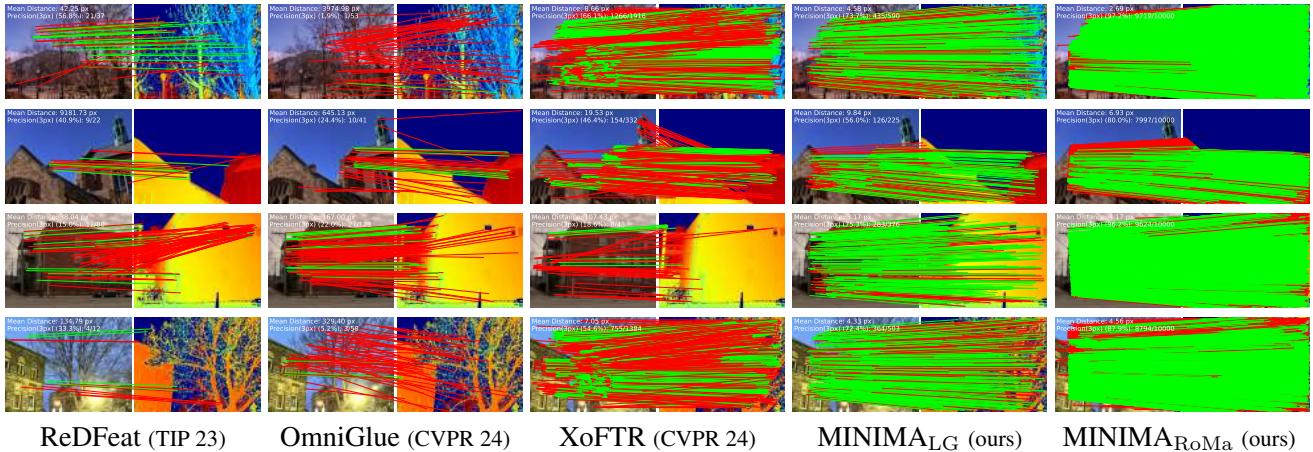


Figure A4. Qualitative results on real RGB-Depth image pairs of DIODE dataset [44]. The red lines indicate false matches.

erence images for each query image with NetVLAD [2] 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 [8]. 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.*, $(0.25m, 2^\circ) / (0.5m, 5^\circ) / (5m, 10^\circ)$. The sparse matchers, including SuperGlue [35], SGMNet [4], LightGlue (LG) [28], ConvMatch [60] 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 demonstrates 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 to complex scenarios.

C.5. More Visible Results on Real Datasets

We also show more qualitative results, which are selected from real RGB-IR [43], RGB-Depth [44], RGB-Event [46] and Remote Sensing [21] (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. A3, Fig. A4, Fig. A5 and Fig. A6. Our methods MINIMA_{LG} (sparse) and MINIMA_{RoMa} (dense) are compared with the sparse pipeline ReDFeat [7] and OmniGlue [20], and semi-dense matcher XoFTR [43]. 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).

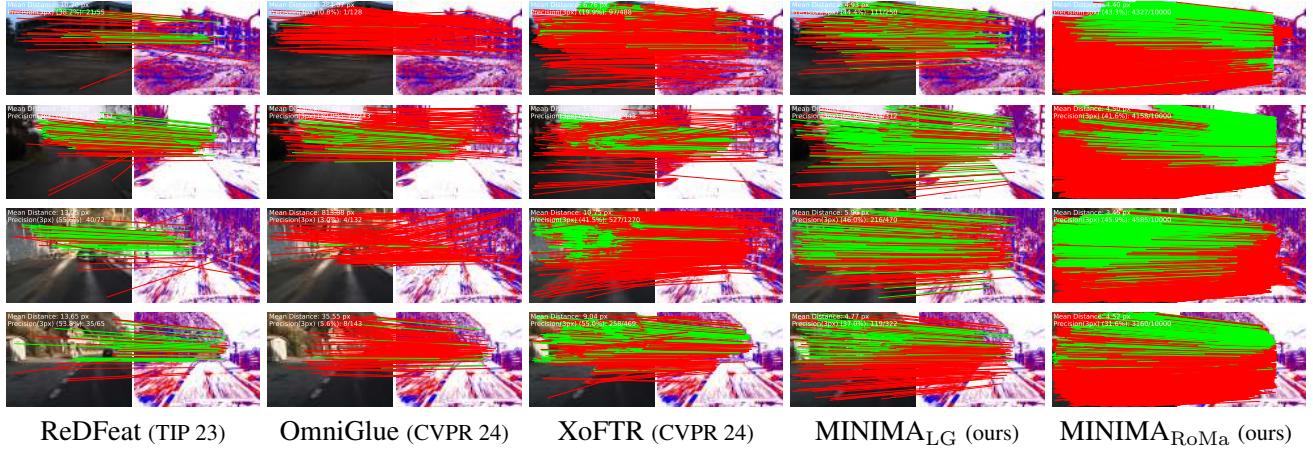


Figure A5. Qualitative results on real RGB-Event image pairs of DSEC dataset [46]. The red lines indicate false matches.

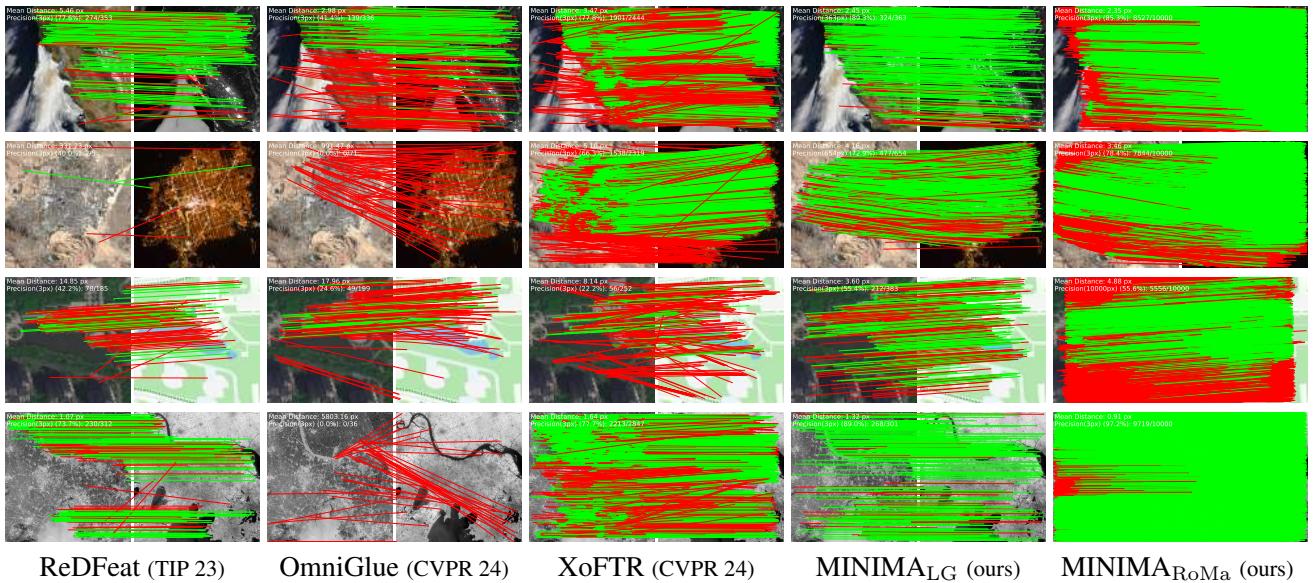


Figure A6. Qualitative results on real image pairs of cross-modal remote sensing dataset [21]. The red lines indicate false matches.