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

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 [30]. Recently, matching two images of different imaging systems/styles plays a vital role in multimodal perceptions [20], including image fusion and enhancement [44, 50], visual localization/navigation [1, 54], target detection/recognition/tracking [12, 42, 45, 52, 53, 55], *etc*. They benefit from gathering the advantages of different modalities by aligning them, thereby yielding more comprehensive representations. However, the cross-view and cross-modality nature makes the matching task more challenging, particularly using a single model for different modalities such as *RGB-Infrared (IR), RGB-Depth, and RGB-Event*.

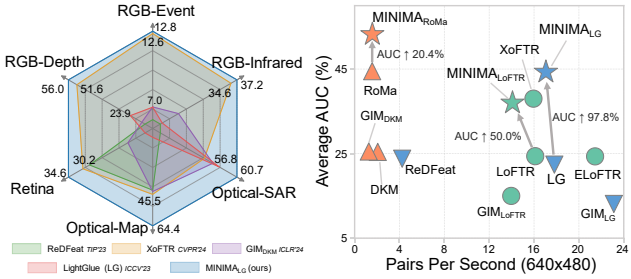

Figure 1. **Overall Image Matching Accuracy and Efficiency on Six Datasets of Real Cross-modal Image Pairs.** AUC of the pose error (@10°) or reprojection error (@10px) is used for accuracy evaluation, while Pairs Per Second is used for efficiency test. *Left*: AUCs on each dataset of representative 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 ★. Using only synthetic multimodal data created by our data engine, MINIMA can generalize to real cross-modal scenes with large improvements.

Existing studies focus more on RGB-only image matching due to the accessible training sets [4, 24], which have given birth to many advanced matching architectures [8, 9, 32, 37, 41]. By contrast, cross-modal matching datasets are 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 [18, 38], and the number of modality types is merely two or three for each dataset. ii) Creating precise dense labels is also expensive. Researchers often *manually* label matched landmarks [18, 20, 27], or use

† Corresponding author.

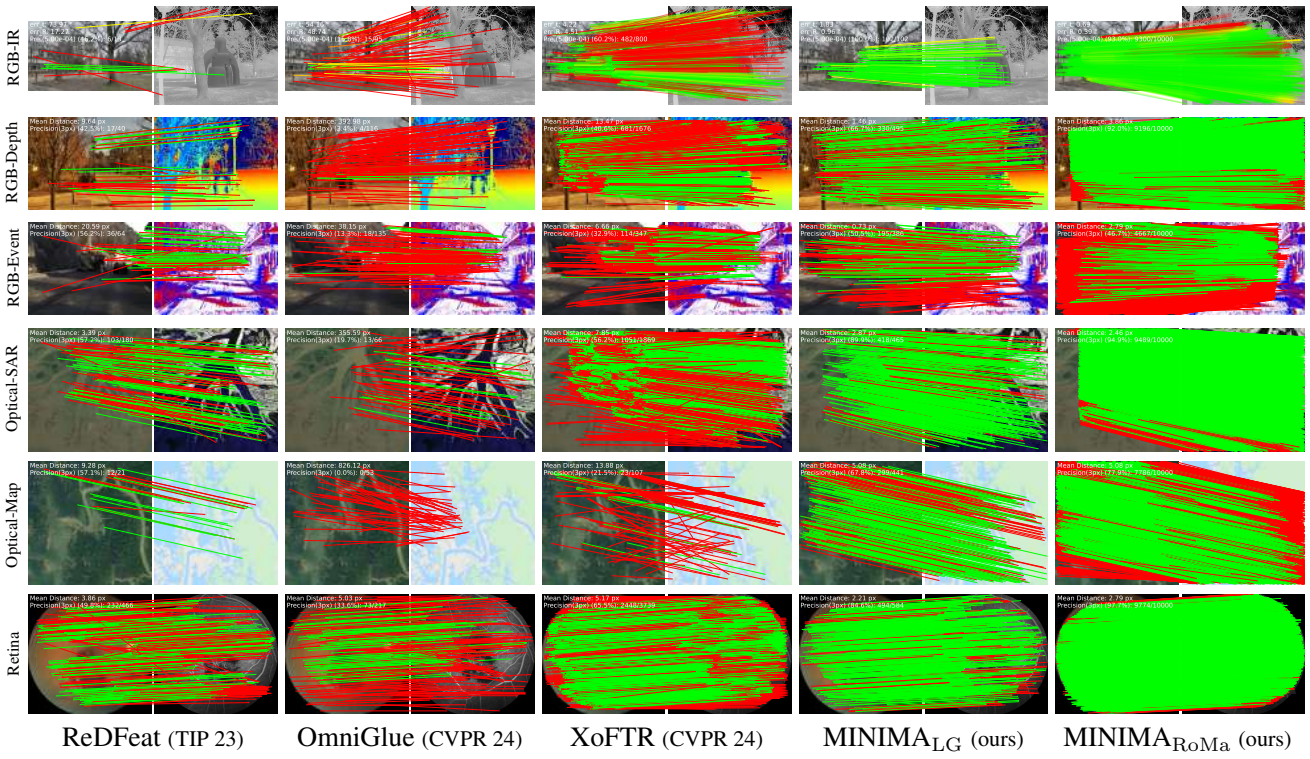

Figure 2. **Qualitative Results on Real Cross-modal Image Pairs.** Our methods MINIMA$_{\text{LG}}$ (sparse) and MINIMA$_{\text{RoMa}}$ (dense) are compared with the sparse matching pipeline ReDFeat [6] and OmniGlue [19], and semi-dense matcher XoFTR [38]. ReDFeat and XoFTR are cross-modal methods, and 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).

camera calibration to produce *approximate* poses [38, 46]. 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 [5, 51]. 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 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 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 MD-syn, a comprehensive dataset 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 handcrafted designs of keypoints detection, description then matching, which are recently updated with deep learning [7, 26, 30, 32]. These detector-based methods can establish keypoint matches with high efficiency but often struggle in textureless regions. Recently, detector-free methods [37, 41] have been introduced to produce semi-dense or dense [8, 9] pixel matches, and achieve dominant perfor-

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

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
RGB Matching				
MegaDepth [24]	40M	196 (Out.)	1	Depth, Pose
ScanNet [4]	230M	1513 (In.)	1	Depth, Pose
Multimodal Matching				
METU-VisTIR [38]	2.5K	6 (Out.)	2	Pose
M3FD [27]	4.2K	15 (Out.)	2	Pre-aligned
LLVIP [18]	15K	26 (Out.)	2	Pre-aligned
DIODE [46]	25K	20 (In. & Out.)	3	Pre-aligned
NYU-Depth V2[36]	408k	464 (In.)	2	Pre-aligned
DVS128[2]	1.3k	122 (In.)	2	Pre-aligned
MD-syn (ours)	480M	196 (Out.)	7	Depth, Pose

mance 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 stages, 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, making it difficult to search for matchable cues. Existing studies still rely on handcrafted designs [16, 21, 48, 49], focusing on extracting matchable information such as shape, gradient, or phase. However, these low-level features are not consistently effective and are time-consuming to extract. Data-driven methods exhibit powerful abilities to extract matchable features for multimodal images. They commonly utilize off-the-shelf matching pipelines [6, 29, 38] as the backbone, then adapt them to the target modalities with specific designs. For example, ReDFeat [6] recoupled 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 [38] utilizes a two-stage training approach for RGB-IR image matching. It 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. We demonstrate that our MINIMA can outperform modality-specific approaches and show 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 [24] or capturing video frames [35]. Open-source tools like *COLMAP* [33, 34] 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 [26, 30, 32]. However, capturing a large number of multimodal images of the same scene is laborsome, 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 [18, 20, 27], or approximated from camera calibrations [38, 46]. 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 [46, 47]. 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 [14, 15]. The proposed engine can freely produce various pseudo modalities from real RGB image pairs, whose matching labels and scene diversity would be well inherited by the generated data.

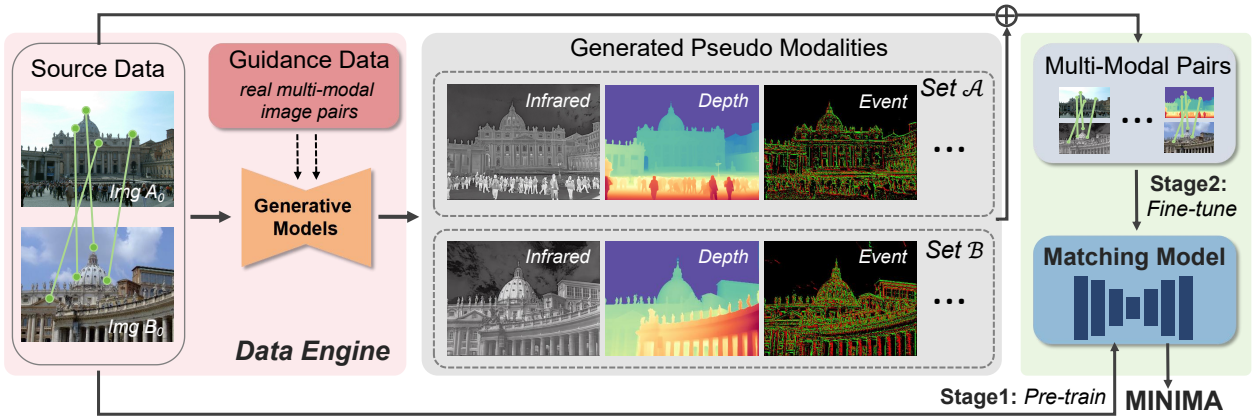

Figure 3. **Overview of the Proposed MINIMA Pipeline: Trained Once to Achieve Any Cross-modal Matching Tasks**. Wherein the *Data Engine* is to generate a large multimodal matching dataset, supporting the training of matching models to obtain cross-modal ability.

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 [5, 6, 51] 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 [6, 51].

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 [24] or video frames [35]. 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 [33, 34]. Those accurate and dense labels 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) [24] and ScanNet (indoor) [4]. 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 [8, 9, 26, 41]. Here we choose MegaDepth [24] as the basic dataset because: i) Multimodal perception tasks are typically performed outdoors, and also, the corresponding datasets [20] 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 [8, 9, 41] to test their generalization. iii) The synthetic MegaDepth makes it convenient to fine-tune those advanced matching methods. Obviously, we can also generate from videos as GIM [35]. 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 [14, 43] and depth or normal generation [3, 47]. As for other modalities, such as infrared, we use the guidance data to fine-tune advanced generative models [14]. 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 [18]. To this end, we turn to a diffusion-based model for help. We use StyleBooth [14] 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 [18] and M3FD [27] datasets. We then implement the style tuner with LoRA [17] of rank 256 and standardize the resolution as 1024×1024 for both input and output. We fine-tune it on a single GPU for $210k$ steps with a fixed $lr = 1 \times 10^{-4}$ and batch size 2.

Depth: We directly use DepthAnything V2 [47] 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.,*

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

with $\triangle L(\mathbf{x}_k, t_k) = p_k C$, where $C > 0$, $\triangle t_k$ is the time elapsed since the last event at the same pixel, and the polarity p_k is the sign of the brightness change [11, 25]. In our study, we randomly set $C \in [0.05, 0.5]$, $p_k = \pm 1$ as suggested in [13] 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 [28] and Anime2Sketch [43], 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 [24] 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.

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

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 the multimodal matching task, making it converge rapidly (verified in our supplementary). In addition, the training on the RGB dataset is well studied [9, 26, 41], 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) [26], LoFTR [37], and RoMa [9]. We will fine-tune them and release our three models, 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 [26], LoFTR [37] and RoMa [9] 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.

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-*

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 [32]	7.49	17.51	33.54	3.06	6.94	13.70	11.53	24.42	41.85	10.38	23.48	41.63	21.52	37.99	56.17	11.35	24.15	42.51
	LightGlue (LG) [26]	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	13.93	27.99	46.16
	ReDFeat [6]	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} [35]	8.40	18.88	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
	$MINIMA_{LG}$	**14.74**	**30.24**	**49.22**	**16.19**	**32.53**	**51.76**	**20.47**	**37.33**	**56.17**	**19.00**	**36.27**	**54.97**	27.51	45.71	63.77	**16.39**	**32.85**	**51.65**
Semi-Dense	LoFTR [37]	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 [38]	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 [41]	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**
	GIM_{LoFTR} [35]	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
	$MINIMA_{LoFTR}$	**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
Dense	DKM [8]	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} [35]	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 [9]	20.27	35.99	54.02	10.21	22.75	39.43	40.99	59.48	74.19	40.86	58.87	73.35	58.49	73.90	84.80	**41.30**	**58.36**	**72.70**
	$MINIMA_{RoMa}$	**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**	40.09	57.21	71.96

Table 3. **Evaluation on Real RGB-IR Dataset (METU-VisTIR) [38] 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 [21]$_{(TIP\ 19)}$	0.05	0.27	0.90	13k
	SRIT [23]$_{(ISPRS\ 23)}$	0.00	0.08	0.37	1.9k
	LNIFT [22]$_{(TGRS\ 22)}$	0.02	0.09	0.43	1.2k
	SuperGlue [32]$_{(CVPR\ 20)}$	4.30	9.26	17.21	86.1
	ReDFeat [6]$_{(TIP\ 23)}$	1.71	4.57	10.85	235.8
	LightGlue (LG) [26]$_{(ICCV\ 23)}$	2.17	5.37	11.21	57.7
	GIM_{LG} [35]$_{(ICLR\ 24)}$	2.43	5.85	10.58	42.9
	OmniGlue [19]$_{(CVPR\ 24)}$	1.48	4.13	10.11	3k
	$MINIMA_{LG}$	**19.14**	**37.17**	**55.51**	58.6
Semi-Dense	LoFTR [37]$_{(CVPR\ 21)}$	2.88	6.94	14.95	61.6
	GIM_{LoFTR} [35]$_{(ICLR\ 24)}$	0.43	1.06	2.99	69.5
	ELoFTR [41]$_{(CVPR\ 24)}$	2.88	7.88	17.72	46.6
	XoFTR [38]$_{(CVPR\ 24)}$	**18.47**	**34.64**	**51.50**	62.7
	$MINIMA_{LoFTR}$	15.61	30.84	47.87	71.6
Dense	DKM [8]$_{(CVPR\ 23)}$	6.76	13.69	22.53	485.3
	GIM_{DKM} [35]$_{(ICLR\ 24)}$	5.08	12.30	23.69	792.2
	RoMa [9]$_{(CVPR\ 24)}$	25.61	48.12	68.37	639.1
	$MINIMA_{RoMa}$	**37.45**	**60.70**	**78.00**	633.3

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* [38] is a real RGB-IR dataset containing 2590 real image pairs with camera poses attached. 3) *DIODE* [39] is a real RGB-Depth/Normal dataset, containing 27858 fully aligned image pairs. 4) *DSEC* [40] 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 [20] 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-view 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 $\{5°, 10°, 20°\}$. As for homography, similar to [37], we collect the mean projection error of four corner points and report the AUC under thresholds $\{3px,5px,10px\}$ 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 [10] settings as a robust homography or pose estimator for fair comparison.

Baselines. Following [41], we select representative methods from the matching pipelines of sparse, semi-dense, and dense matching. 1) For sparse keypoint detection and matching methods, we choose SuperGlue [32], LightGlue (LG) [26], OmniGlue [19], and LG-based GIM [35] for comparison. All of them (including our $MINIMA_{LG}$) use SuperPoint as the keypoint detector (the maximum

Table 4. **Evaluation on Real RGB-Depth Dataset (DIODE) [39] with Homography Estimation.** The AUC of the projective error in percentage is reported.

Category	Method	Homo. estimation AUC		
		@3px	@5px	@10px
Sparse	SuperGlue [32]$_{(CVPR\ 20)}$	1.77	6.83	21.15
	ReDFeat [6]$_{(TIP\ 23)}$	1.01	4.58	16.30
	LightGlue (LG) [26]$_{(ICCV\ 23)}$	0.79	3.30	11.26
	GIM$_{\mathrm{LG}}$ [35]$_{(ICLR\ 24)}$	0.30	1.14	3.65
	MINIMA$_{\mathrm{LG}}$	**8.71**	**26.80**	**55.97**
Semi-Dense	LoFTR [37]$_{(CVPR\ 21)}$	0.97	4.20	15.16
	GIM$_{\mathrm{LoFTR}}$ [35]$_{(ICLR\ 24)}$	0.00	0.25	1.15
	ELoFTR [41]$_{(CVPR\ 24)}$	0.82	4.09	16.69
	XoFTR [38]$_{(CVPR\ 24)}$	**11.03**	**27.24**	**51.60**
	MINIMA$_{\mathrm{LoFTR}}$	5.35	18.65	44.85
Dense	DKM [8]$_{(CVPR\ 23)}$	1.29	4.23	11.78
	GIM$_{\mathrm{DKM}}$ [35]$_{(ICLR\ 24)}$	1.90	6.34	17.96
	RoMa [9]$_{(CVPR\ 24)}$	9.21	24.64	49.31
	MINIMA$_{\mathrm{RoMa}}$	**28.98**	**50.88**	**72.54**

number of extracted keypoints is set as 2048). We also take ReDFeat [6] into account as it is a deep method designed for multimodal image matching. In addition, three handcrafted multimodal matching methods, including RIFT [21], SRIT [23], LNIFT [22], 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 include LoFTR [37], ELoFTR [41], XoFTR [38], and GIM$_{\mathrm{LoFTR}}$, where XoFTR is tailored for RGB-IR image matching. 3) As for dense matching, DKM [8], GIM$_{\mathrm{DKM}}$ [35] and recent SOTA matcher RoMa [9] 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 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$_{\mathrm{DKM}}$ perform poorly on four cross-modal cases due to the huge modality gaps among them. The original RoMa exhibits good generalization, mainly because of the use of DINOv2 [31] that has seen numerous types of images during pre-training. Our MINIMA still obtains significant enhancement over RoMa.

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 [38]) for pose estimation and RGB-Depth (DIODE [39]) for homography estimation. The results are in Tab. 3 and Tab. 4.

For the RGB-IR test, our MINIMA$_{\mathrm{LG}}$ enhances sparse matching, with AUC increasing 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 [9, 41].

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 @5px. 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 [40]. 1) and 2) are both from MMIM [20] 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$_{\mathrm{LG}}$ still exhibits a few advantages. As for remote sensing cases, our MINIMA achieves large gains for sparse and dense matching. While the semi-dense matcher MINIMA$_{\mathrm{LoFTR}}$ 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 for matching them.

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 [32]	30.72	36.18	44.66	18.34	27.47	45.59	0.00	0.67	8.00
	LightGlue (LG) [26]	35.47	42.37	49.48	16.22	27.51	44.62	0.00	0.67	7.02
	ReDFeat [6]	**38.55**	**44.26**	50.93	15.99	23.95	43.72	0.55	0.97	6.07
	GIM_{LG} [35]	24.32	27.88	33.84	11.09	17.44	27.18	**0.57**	1.08	5.54
	$\text{MINIMA}_{\text{LG}}$	37.95	44.08	**52.50**	**23.53**	**38.40**	**58.74**	0.52	**2.27**	**12.82**
Semi-Dense	LoFTR [37]	38.42	43.89	50.13	24.13	33.80	50.79	0.00	0.00	3.59
	XoFTR [38]	**39.67**	**45.60**	52.32	**27.35**	**39.58**	56.63	0.00	1.37	**12.64**
	ELoFTR [41]	34.57	41.66	49.08	16.45	29.65	46.74	0.64	1.34	7.78
	$\text{GIM}_{\text{LoFTR}}$ [35]	39.51	44.40	48.94	17.96	27.41	37.29	0.00	0.55	1.19
	$\text{MINIMA}_{\text{LoFTR}}$	**39.67**	45.33	**52.77**	23.32	35.18	**56.81**	**0.81**	**2.49**	11.75
Dense	DKM [8]	39.43	45.00	51.78	26.44	35.82	51.20	0.00	0.00	0.00
	GIM_{DKM} [35]	37.78	43.46	48.87	21.19	30.28	47.68	0.00	0.66	7.04
	RoMa [9]	**39.62**	45.13	53.75	29.24	40.50	57.84	**0.85**	1.69	10.71
	$\text{MINIMA}_{\text{RoMa}}$	39.17	**45.92**	**57.55**	**32.55**	**44.68**	**64.38**	0.54	**3.51**	**17.07**

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** AUC@10°	**Rel RGB-IR** AUC@10°	**Rel RGB-D** AUC@10px
Basic Model: LoFTR (LT) [37]	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**

Some qualitative results are shown in Fig. 2, which demonstrate that our MINIMA can establish a high number and ratio of correct matches for real cross-modal image pairs.

5.5. 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.

5.6. 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 [44]. This property is well exhibited in our generated modalities (see the suppl.), which plays an important role in training a general matching model. Existing diffusion-based methods [14, 15] can generate high-quality images of the target modality, making the pseudo modality much closer to the real. Extensive experiments 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 filling 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.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grant 62406117, U234120202, and 62225603), the China Postdoctoral Science Foundation (Grant 2023M741263 and GZC20230895), and the Postdoctor Project of Hubei Province (Grant 2024HBBHCXA014).

References

[1] NG Aditya, PB Dhruval, 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 *ICRA*, pages 14116–14122, 2024. 1

[2] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *CVPR*, pages 7243–7252, 2017. 3

[3] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *CVPR*, 2024. 5

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nie10ner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 1, 3, 4

[5] Xin Deng, Enpeng Liu, Chao Gao, Shengxi Li, Shuhang Gu, and Mai Xu. Crosshomo: Cross-modality and cross-resolution homography estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 2, 4

[6] Yuxin Deng and Jiayi Ma. Redfeat: Recoupling detection and description for multimodal feature learning. *IEEE Trans. Image Process.*, 32:591–602, 2022. 2, 3, 4, 6, 7, 8

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

[8] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *CVPR*, pages 17765–17775, 2023. 1, 2, 4, 6, 7, 8

[9] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, pages 19790–19800, 2024. 1, 2, 4, 5, 6, 7, 8

[10] 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

[11] 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 Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2020. 5

[12] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024. 1

[13] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *CVPR*, pages 3586–3595, 2020. 5

[14] 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, 8

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 3, 8

[16] 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

[17] 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 *ICLR*, 2022. 5

[18] 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, 3, 5

[19] 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. 2, 6

[20] 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, 3, 4, 6, 7

[21] Jiayuan Li, Qingwu Hu, and Mingyao Ai. Rift: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Trans. Image Process.*, 29:3296–3310, 2019. 3, 6, 7

[22] Jiayuan Li, Wangyi Xu, Pengcheng Shi, Yongjun Zhang, and Qingwu Hu. Lnift: Locally normalized image for rotation invariant multimodal feature matching. *IEEE Trans. Geosci. Remote Sens.*, 60:1–14, 2022. 6, 7

[23] 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

[24] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 1, 3, 4, 5

[25] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 $120db$ $15\mu s$ latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2):566–576, 2008. 5

[26] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, pages 17627–17638, 2023. 2, 3, 4, 5, 6, 7, 8

[27] 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, 3, 5

[28] 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 *ICCV*, pages 6598–6607, 2021. 5

[29] Yuyan Liu, Wei He, and Hongyan Zhang. Grid: Guided refinement for detector-free multimodal image matching. *IEEE Trans. Image Process.*, 2024. 3

[30] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *Int. J. Comput. Vis.*, 129(1):23–79, 2021. 1, 2, 3

[31] 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. *Transac. Machine Learning Research*, 2023. 7

[32] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 1, 2, 3, 6, 7, 8

[33] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3, 4

[34] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 3, 4

[35] 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, 6, 7, 8

[36] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012. 3

[37] 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. 1, 2, 5, 6, 7, 8

[38] 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, 6, 7, 8

[39] 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

[40] 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. 6, 7

[41] 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. 1, 2, 4, 5, 6, 7, 8

[42] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *CVPR*, pages 19156–19166, 2024. 1

[43] Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P Allebach. Adversarial open domain adaptation for sketch-to-photo synthesis. In *WACV*, pages 1434–1444, 2022. 5

[44] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):12148–12166, 2023. 1, 8

[45] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *ICCV*, pages 11069–11079, 2023. 1

[46] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024. 2, 3

[47] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024. 3, 5

[48] Yongxiang Yao, Yongjun Zhang, Yi Wan, Xinyi Liu, Xiaohu Yan, and Jiayuan Li. Multi-modal remote sensing image matching considering co-occurrence filter. *IEEE Trans. Image Process.*, 31:2584–2597, 2022. 3

[49] Yuanxin Ye, Lorenzo Bruzzone, Jie Shan, Francesca Bovolo, and Qing Zhu. Fast and robust matching for multimodal remote sensing image registration. *IEEE Trans. Geosci. Remote Sens.*, 57(11):9059–9070, 2019. 3

[50] 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

[51] Kaining Zhang and Jiayi Ma. Sparse-to-dense multimodal image registration via multi-task learning. In *ICML*, 2024. 2, 4

[52] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *CVPR*, pages 2153–2162, 2023. 1

[53] Yuxiang Zhang, Yang Zhao, Yanni Dong, and Bo Du. Self-supervised pretraining via multimodality images with transformer for change detection. *IEEE Trans. Geosci. Remote Sens.*, 61:1–11, 2023. 1

[54] Kaichen Zhou, Changhao Chen, Bing Wang, Muhamad Risqi U Saputra, Niki Trigoni, and Andrew Markham. Vmloc: Variational fusion for learning-based multimodal camera localization. In *AAAI*, pages 6165–6173, 2021. 1

[55] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *CVPR*, pages 9516–9526, 2023. 1