AnyMatch: Supercharging Universal Multi-Modal Image Matching with Large-Scale Single-View Images

Meng Yang[1*], Zizhuo Li[1*†], Linfeng Tang[2], Fan Fan[2] ✉, and Jiayi Ma[2]

Electronic Information School[1], Wuhan University, Wuhan 430072, China
School of Robotics, Wuhan University[2], Wuhan 430072, China
{2024102120059, zizhuo_li, fanfan}@whu.edu.cn, {linfeng0419, jyma2010}@gmail.com

Abstract. Multi-modal image matching is essential for visual localization and multi-sensor fusion, but it is hindered by the scarcity of large-scale training data with precise geometric annotations. Existing real-world datasets suffer from prohibitive costs, limited scene diversity, and errors in SfM-MVS pipelines, while synthetic methods struggle to maintain 3D geometric consistency or achieve photorealistic appearance. To address this, we propose AnyMatch, a novel framework that leverages abundant, easily accessible single-view images at minimal cost to generate rich multi-modal training data. AnyMatch integrates monocular depth estimation, 3D reprojection, diffusion-based inpainting, and cross-modal image translation to synthesize multi-view, multi-modal image pairs with 3D geometric fidelity. Crucially, our method provides annotations that strictly adhere to 3D geometric consistency through explicit 3D reprojection, avoiding SfM-MVS error accumulation. Furthermore, AnyMatch offers strong scalability, enabling controllable scene diversity and annotation difficulty via adjustable input and camera parameters. We construct Any-syn, a large-scale synthetic multi-modal dataset using AnyMatch. Experimental results show that matching networks (*e.g.*, LoFTR, EDM, RoMa) fine-tuned on Any-syn achieve substantial performance gains on multi-modal benchmarks, exhibiting superior generalization and robustness compared to models trained on existing data.

Keywords: Multi-Modal Image Matching · Synthetic Data Generation · Single-View Image

1 Introduction

Feature matching across different imaging modalities is a cornerstone of many computer vision systems, enabling critical capabilities such as visual localization

* Equal contribution. † Project leader. ✉ Corresponding author.

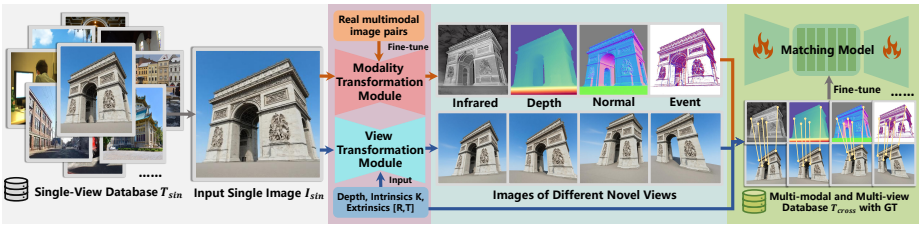

Fig. 1: Overview of AnyMatch pipeline. AnyMatch aims to generate a large-scale multi-modal and multi-view image matching dataset for training matching models, thereby enabling robust cross-modal and cross-view matching capabilities.

under varying conditions [45], cross-spectral object detection [24], and multi-sensor fusion for autonomous systems [1]. Traditional feature matching methods, from handcrafted descriptors like SIFT [25] and SURF [2] to modern deep learning approaches like LoFTR [33], EDM [20] and RoMa [7], have primarily focused on single-modality (RGB-RGB) scenarios. However, real-world applications often require matching between images captured by different sensors (*e.g.*, RGB-thermal, RGB-depth, or RGB-event cameras), where modal differences dramatically challenge conventional single-modal matching methods.

The fundamental bottleneck hindering the development of robust multi-modal matching algorithms lies in the scarcity of large-scale training data with precise geometric annotations. However, both existing real-world and synthetic multi-modal datasets suffer from significant limitations.

Existing multi-modal image datasets face three critical challenges: prohibitive acquisition costs, limited scene diversity, and inaccurate geometric annotations. First, capturing genuine multi-modal image pairs requires specialized, synchronized sensor suites (*e.g.*, RGB and thermal cameras) with precise cross-modal calibration, resulting in substantial hardware and labor expenditures. Second, geometric ground truth (GT) for existing single-modal datasets is typically derived via Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipelines from multi-view imagery [4, 21]. However, such reconstruction approaches inherently favor static scenes with high appearance consistency, minimal occlusion, and good illumination conditions, thereby failing to reflect the complexity of real-world environments and limiting coverage of diverse scene content. These constraints also preclude direct application of SfM-MVS pipelines to GT generation of multi-modal images. Third, SfM-MVS-derived datasets suffer from error accumulation—including pose estimation inaccuracies, reconstruction failures in textureless regions, and violations of photometric consistency assumptions—yielding sparse and unreliable geometric constraints and depth estimation. Consequently, models trained on such data, which exhibit severely limited diversity and accuracy, often undergo catastrophic performance degradation when deployed in unseen scenarios or under challenging conditions.

To mitigate the scarcity of real-world data, existing synthetic data generation methods offer a viable alternative. First, image translation–based synthesis

methods such as MINIMA [29] and MatchAnything [13] synthesize multi-modal image pairs by applying cross-modal translation to pre-existing single-modal matching datasets (*e.g.*, MegaDepth [21]), followed by fine-tuning to improve matching performance on the generated data. However, it relies on SfM for dataset construction, with three limitations: static-region reconstruction only (scene biases), tourist attractions-centric scenes (limited diversity), and time-consuming multi-view image acquisition. Second, 2D homography–based synthesis methods operate solely within the image plane via homography transformations [5,44], thereby discarding cross-view 3D geometric consistency (such as the physical relationship between parallax and depth) and failing to produce occlusion patterns that adhere to real-world physics. Third, game engine–driven synthesis methods [11,26] enable explicit control over scene parameters but struggle to achieve photorealistic appearance and accurate physical effects, particularly in material reflectance properties and global illumination. Although synthetic data generation circumvents the high costs of real data acquisition and manual annotation, thereby significantly reducing dataset construction expenses, it introduces non-negligible limitations. These constraints restrict the scale, geometric accuracy, and physical authenticity of current synthetic multi-modal datasets.

In contrast to these limitations, vast repositories of single-view images are readily available on the internet and across diverse research domains. Datasets such as GLDv2 [38], SA-1B [17], and COCO [22] collectively comprise millions of images spanning heterogeneous environments—including indoor/outdoor settings and urban/natural scenes—under varying conditions and domains. These resources inherently embody the complexity and long-tail distribution characteristics of the real world, theoretically offering feature matching models unprecedented scene coverage and generalization potential. Nevertheless, the absence of an effective mechanism to synthesize multi-view, multi-modal image pairs with 3D spatially consistent geometry and pixel-level precise supervision has left these abundant resources in a state of untapped potential: despite their richness, they remain largely inaccessible for advancing visual matching tasks.

Inspired by recent advances in monocular depth estimation, image generation, and cross-modal image translation, we propose AnyMatch, a novel framework that transforms abundant, easily accessible single-view images at minimal cost into a rich source of multi-modal training data with high 3D geometric fidelity. AnyMatch comprises two core components: view transformation and modality transformation, as shown in Fig. 1. The view transformation module lifts a single 2D image into 3D space by estimating depth and camera intrinsics, followed by novel view synthesis via 3D reprojection based on the camera extrinsics, as shown in Fig. 2. Using diffusion-based inpainting, we complete occluded regions to produce visually coherent novel views. The modality transformation module generates different modal images (infrared, depth, normal and event modalities) through cross-modal image translation techniques. By jointly performing view transformation and modality transformation, we synthesize multi-modal image pairs and annotations that strictly adhere to 3D geometric consis-

tency, avoiding SfM-MVS error accumulation. This pipeline can be applied to any existing image dataset, enabling the generation of arbitrarily large, diverse multi-modal training sets with pixel-perfect geometric supervision.

AnyMatch offers three distinct advantages over existing methods. First, it decouples data generation from expensive multi-sensor hardware by leveraging the abundance of readily accessible single-view images at minimal cost. Second, it provides stereo-geometrically precise GT through explicit 3D reprojection, ensuring annotations that strictly adhere to 3D geometric consistency and avoiding error accumulation. Third, it can control scene diversity and annotation difficulty levels via adjustable input and camera parameters, thereby supporting diverse application requirements and facilitating curriculum learning strategies.

To validate our method, we fine-tune existing image matching networks (LoFTR, EDM, and RoMa) on datasets synthesized by AnyMatch. Experimental results demonstrate that the fine-tuned models achieve substantial performance gains on cross-modal benchmark evaluations.

In summary, our contributions are threefold:

- We propose AnyMatch, the first framework capable of synthesizing large-scale multi-modal image pairs with high 3D geometric fidelity from abundant, easily accessible single-view images at minimal cost.
- We present a novel pipeline that integrates monocular depth estimation, 3D reprojection, diffusion-based inpainting, and cross-modal image translation to synthesize diverse multi-modal datasets with annotations that strictly adhere to 3D geometric consistency.
- We construct Any-syn, a synthetic multi-modal dataset generated via AnyMatch. Matching networks fine-tuned on Any-syn achieve superior performance on multi-modal matching benchmarks, with significant improvements in generalization and robustness. The supplementary material presents some data synthesis results.

2 Related Work

2.1 Multi-modal Image Matching

Cross-modal image matching aims to establish pixel-wise correspondences between images captured under different sensing modalities, serving as a fundamental building block for visual localization [1, 45] and multi-sensor fusion tasks [40,43]. Traditional methods primarily rely on handcrafted feature descriptors that extract shape-, gradient-, or phase-based features [10,14,18,19], yet exhibit limited robustness in textureless regions or under drastic illumination variations. Recently, data-driven deep learning methods have demonstrated superior feature representation capabilities. Existing works typically adopt off-the-shelf matching architectures (*e.g.*, EDM [20], LoFTR [33], and RoMa [7]) as backbone networks, with task-specific adaptations for target modalities. For instance, ReDFeat [6] decouples detection and description constraints in multi-modal feature learning via a mutual weighting strategy. XoFTR [34] and XCP-Net [42] employ

a two-stage training paradigm that first pre-trains on real multi-modal data and subsequently fine-tunes on pseudo-infrared images, significantly boosting RGB-infrared matching performance. SCFeat [3] and SAMFeat [39] leverage semantic priors extracted from vision foundation models (*e.g.*, DINOv2 [27] and SAM [17]) to refine feature representation and enhance matching accuracy. PYM translates cross-modal images into the matcher's native domain via correspondence-aware adversarial learning, enabling competitive cross-modal matching without modifying the pre-trained matcher [9]. Nevertheless, these methods are predominantly tailored to specific modality pairs and remain severely constrained by the scale and diversity of available training data, hindering their generalization capability across unseen modalities and real-world scenarios.

2.2 Synthetic Data Generation for Matching

Current large-scale RGB image matching datasets (*e.g.*, MegaDepth [21], ScanNet [4]) primarily rely on SfM-MVS pipelines to generate geometric GT. MegaDepth reconstructs scene depth and camera poses from internet-sourced multi-view photographs using open-source tools such as COLMAP [31, 32]. ScanNet similarly generates annotations from indoor RGB-D video sequences through analogous reconstruction workflows. The core advantage of these methods lies in their ability to leverage real-world imagery and mature geometric reconstruction tools to derive correspondence supervision. However, the SfM-MVS pipeline suffers from several inherent limitations: (1) it intrinsically favors static scenes with high photometric consistency, minimal occlusions, and favorable lighting conditions, struggling with dynamic objects, drastic illumination changes, or textureless regions; (2) the multi-stage reconstruction process accumulates errors across stages; and (3) it critically depends on multi-view image inputs. Owing to significant discrepancies in pixel intensity distributions and fundamental gaps in imaging mechanisms across modalities, conventional SfM-MVS pipelines fail to operate robustly on multi-modal image pairs. Consequently, acquiring accurate geometric annotations for real-world multi-modal data remains highly challenging. Coupled with the high cost of acquiring real multi-modal data, this has led to a severe scarcity of annotated multi-modal image matching datasets. To mitigate this data scarcity, researchers have proposed diverse synthetic data generation strategies, which can be broadly categorized into three classes:

Image translation-based synthesis methods. These methods generate multi-modal image pairs by applying cross-modal translation to existing single-modality matching datasets. As a representative work, MINIMA [29] proposes a data synthesis engine that leverages existing multi-view RGB datasets and employs generative models (StyleBooth [12] and DepthAnything v2 [41]) to synthesize pseudo-modality images including infrared (IR), depth, and event modalities, *etc.* The geometric GT (depth maps and camera poses) from the source dataset is directly inherited to serve as correspondence supervision for the synthesized multi-modal pairs. This method demonstrates that leveraging large-scale synthetic multi-modal data can effectively enhance the cross-modal generalization

capability of matching models. Nevertheless, MINIMA exhibits inherent limitations: (1) its source data is constrained by the scene coverage and scale of existing multi-view datasets; (2) its geometric ground truth is directly inherited from SfM-MVS reconstruction results, inevitably subject to pose estimation errors and reconstruction failures. These constraints propagate into the synthesized multi-modal pairs, compromising the geometric accuracy of the training supervision.

Homography-based synthesis methods. These methods simulate geometric deformations by applying homography transformations within the image plane [5, 44]. Although computationally efficient and straightforward to implement, they discard cross-view 3D geometric consistency and fail to generate occlusion patterns that conform to real-world physics. As a result, models trained on such synthetic data exhibit weak generalization capability under view variations encountered in real-world scenarios.

Game engine-based synthesis methods. Several works leverage game engines to generate synthetic data [11, 26]. While these methods enable explicit control over scene parameters and sensor models, they struggle to achieve photorealistic appearance fidelity—particularly in modeling material reflectance properties, global illumination effects, and sensor noise characteristics. This results in a significant domain gap between synthesized and real-world images, thereby limiting the transferability of models trained on such data to practical applications.

To address the aforementioned limitations, we propose AnyMatch, a framework that synthesizes multi-view, multi-modal image pairs with high 3D geometric fidelity from large-scale real-world single-view images. Compared with the existing data synthesis methods, AnyMatch utilizes real scene images, does not rely on expensive acquisition hardware, and provides GT that strictly adheres to 3D geometric consistency, without the error accumulation of SfM-MVS.

3 Method

Existing image matching datasets predominantly rely on real-world multi-view captures; however, they are inherently constrained by limited data scale, insufficient scene diversity, narrow modality coverage, and sub-optimal annotation quality. These limitations hinder the development and training of robust, general-purpose multi-modal matching models. To address this challenge, we propose AnyMatch, a geometry-supervised generative framework for multi-modal image synthesis. The framework explicitly integrates two core components: view transformation and modality transformation.

3.1 View Transformation

We construct a large-scale, diverse, and high-quality single-view image dataset from publicly available internet sources across multiple research domains (*e.g.*,

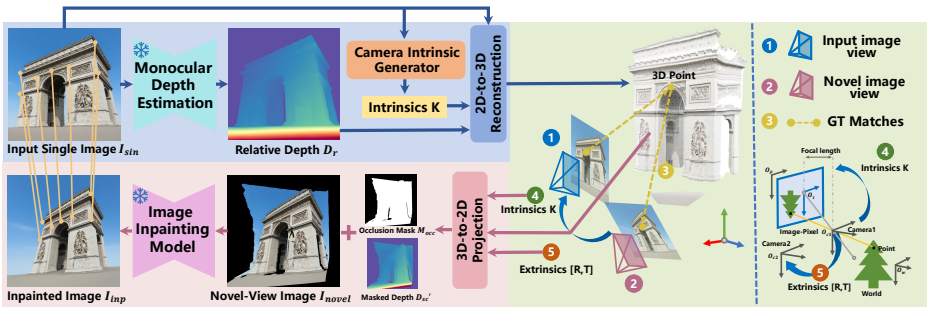

Fig. 2: Overview of view transformation module. The view transformation module is designed to generate an arbitrary number of multi-view image pairs that strictly adhere to 3D geometric consistency.

GLDv2 and SA-1B), predominantly in the visible modality, to establish the single-view database T_{sin}. All images are uniformly resized to 512×512 via isotropic scaling and center cropping. As shown in Fig. 2, AnyMatch takes a single-view image $I_{\text{sin}} \in T_{\text{sin}}$ as input. First, we apply a pre-trained monocular relative depth estimation model M_{rd} to predict the dense relative depth map $D_r = M_{\text{rd}}(I_{\text{sin}})$, which captures the depth of fine-grained geometric structures of the image and preserves rich local details. Due to the inherent scale ambiguity in monocular depth estimation, we apply random affine perturbations to the depth to enhance the algorithm's robustness against depth scale uncertainty. Specifically, the normalized depth map $D_r \in [0, 1]$ is transformed via:

$$D = 1/(\alpha \cdot (1/D_r) + \beta + \epsilon), \tag{1}$$

where the scaling factor $\alpha \in (0.5, 2.0)$ simulates scale uncertainty arising from variations in camera position and focal length. The bias $\beta \in (0, 0.3)$ models sensor systematic errors and the non-zero mean noise inherent in the depth estimation network. ϵ is introduced for numerical stability. Subsequently, to lift the 2D image into 3D space, we design a random camera intrinsic generator. Specifically, we initialize a intrinsic $K = \begin{bmatrix} f & 0 & 0.5 \\ 0 & f & 0.5 \\ 0 & 0 & 1 \end{bmatrix}$, where the principal point is fixed at the normalized image center $(0.5, 0.5)$, and the focal length f is randomly sampled from $(0.58, 0.88)$. For each pixel (u_i, v_i) in I_{sin}, the corresponding 3D coordinates in the camera coordinate system are computed using K and the depth value $D(u_i, v_i)$, thereby constructing the 3D point cloud P:

$$P = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = D(u_i, v_i) \cdot K^{-1} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}. \tag{2}$$

Subsequently, we design a random camera extrinsic generator to enable switching to a novel view by sampling camera extrinsic parameters $[R \mid T]$. For the rotation component of extrinsic, the rotation angles are within $[-7.5°, +7.5°]$.

For the translation component $t = [t_x, t_y, t_z]^\top$, the perturbation ranges are set to $[-0.3, +0.3]$ for t_x and t_y, and $[-0.5, +0.5]$ for t_z. The data difficulty level can be modulated by adjusting the variation range of extrinsics. The point cloud P is transformed into the novel camera coordinate system via $P' = R \cdot P + T$. Using differentiable rendering, P' is reprojected into the novel view to generate a preliminary rendered novel-view image I_{novel} and an occlusion mask M_{occ}, while the Z-coordinates of P' directly constitute the corresponding depth map D' for I_{novel}. To inpaint occluded regions in I_{novel} while preserving semantic continuity, we apply an image inpainting model M_{inpaint} for context-aware completion:

$$I_{\text{inp}} = M_{\text{inpaint}}(I_{\text{novel}}, M_{\text{occ}}). \tag{3}$$

The view transformation module outputs a novel-view image I_{inp} with high 3D geometric fidelity, camera parameters $\{K, [R \mid T]\}$ and depth maps $\{D, D'\}$, which strictly follow 3D geometric consistency and circumvent the limitations of SfM. Using these outputs, we can establish the pixel-level GT correspondence between the synthetic image pair $\{I_{\text{sin}}, I_{\text{inp}}\}$. In our current implementation, to maintain dense supervision, we uniformly treat all pixels (including inpainted ones) in the loss formulation.

3.2 Modality Transformation

Another critical component in constructing a multi-modal image matching database lies in leveraging a modality translation mechanism to transform the input single-view image into target modalities (*e.g.*, infrared, depth, normal, and event), thereby simulating the modality differences inherent in real-world multi-modal image pairs, as shown in Fig. 1. Specifically, for the original image I_{sin}, multiple modality translation models are deployed in parallel to perform modality transformation, yielding a diverse set of pseudo-multi-modal images. In AnyMatch, we implement four representative target modalities: infrared, depth, normal, and event modalities.

For the **infrared** modality, we directly employ a pre-trained RGB-to-IR diffusion model. In the first stage, it collects a large-scale infrared image dataset for training. The text prompt for the diffusion model is uniformly set to "an infrared image", and the Latent Diffusion Model (LDM) [30] is fine-tuned using Low-Rank Adaptation (LoRA) [15] to generate infrared images under text-conditioned guidance, thereby establishing a semantic association between the concept of "infrared" and its characteristic visual patterns. In the second stage, it collects a large-scale dataset of RGB-IR image pairs; during training, the RGB image serves as the conditional input to synthesize the corresponding infrared image, enabling the model to produce outputs that faithfully reflect the physical and perceptual characteristics of infrared imaging based on visible inputs.

For the **depth and normal** modalities, we directly employ pre-trained monocular relative depth estimation and normal estimation models, respectively.

For the **event** modality, we adopt a motion simulation method to synthesize event streams from static visible image pairs. Each pixel of the event camera inde-

pendently monitors the logarithmic change in brightness. The logarithmic intensity is defined as $L = \log(I_{\text{sin}})$. When the change in brightness exceeds a temporal contrast threshold C, pixel $\mathbf{x}_k = (x_k, y_k)^\top$ triggers an event $e_k = (\mathbf{x}_k, t_k, p_k)$ at timestamp t_k. Therefore, the logarithmic change in brightness for each pixel is defined as $\Delta L(\mathbf{x}_k, t_k) = L(\mathbf{x}_k, t_k) - L(\mathbf{x}_k, t_k - \Delta t_k)$, where $\Delta L(\mathbf{x}_k, t_k) = p_k C$, with $C > 0$. Δt_k is the time elapsed since the last triggered event for the same pixel, and $p_k = \pm 1$ represents the polarity of the brightness change (+1 for brightening, -1 for darkening). The threshold parameter $C \in [0.05, 0.5]$ and p_k are set randomly to simulate different sensor characteristics. Furthermore, small random motions are applied to the input image frame I_{sin} to simulate camera motion in real-world scenarios for calculating the event response.

The process of modality transformation can be simplified as:

$$I_{\text{modal}} = F(I_{\text{sin}}), \tag{4}$$

where F denotes the modality transformation module, and I_{modal} represents the synthesized pseudo-target-modality image.

3.3 Integration of View and Modality Transformation

Since modality transformation preserves geometric structure without inducing distortions, we construct multi-modal and multi-view image pairs by combining the modality-transformed image I_{modal} with the view-transformed image I_{inp}, while directly inheriting the GT supervision information generated during view transformation. Through the synergistic coupling of view and modality transformation, we ultimately synthesize the multi-modal and multi-view image matching dataset T_{cross}.

3.4 Sample-level Geometric Consistency Verification

During occlusion inpainting, image restoration may generate spurious content that violates the geometric and semantic consistency of the original scene—a phenomenon commonly termed "hallucination" [23] in generative vision literature. Although such content may appear locally coherent, it fundamentally deviates from the true scene structure and can adversely affect the learning of discriminative feature representations. To mitigate this issue, we introduce a geometric consistency verification mechanism for sample-level quality screening. As shown in Fig. 3, for the single-modal image pair composed of the original image I_{sin} and the inpainted novel-view image I_{inp}, we first employ a high-performance dense matching model to extract a dense correspondence set:

$$\mathcal{M}_{\text{init}} = \{(x^i_{\text{sin}}, x^i_{\text{inp}})\}_{i=1}^{N}, \tag{5}$$

where N denotes the total number of extracted correspondences. Subsequently, leveraging the known camera parameters and depth maps, the endpoint error (EPE) for each correspondence is computed as:

$$EPE(x^i_{\text{sin}}) = \left\| x^i_{\text{inp}} - \hat{x}^i_{\text{inp}} \right\|^2, \tag{6}$$

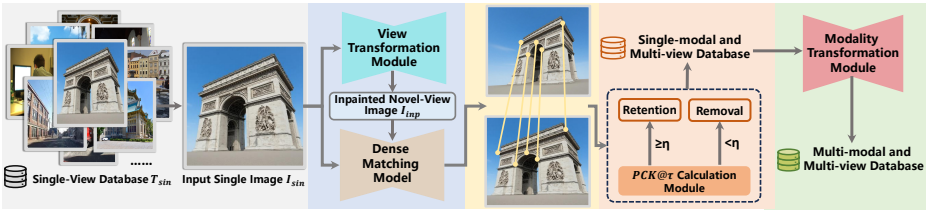

Fig. 3: Overview of sample-level geometric consistency verification module. The image pair is retained if PCK@$\tau \geq \eta$ and removed if PCK@$\tau < \eta$.

where x^i_{inp} is the predicted correspondence in I_{inp} (from the matching model), and $\hat{x}^i_{inp}$ is the geometrically projected GT correspondence. The proportion PCK@τ of correct correspondences at threshold τ is defined as:

$$\text{PCK@}\tau = \frac{\left|\{(x^i_{\text{sin}}, x^i_{\text{inp}})\}_{i=1}^{N} | EPE(x^i_{\text{sin}}) < \tau\right|}{N} \times 100\%. \tag{7}$$

If PCK@τ of the synthesized image pair satisfies PCK@$\tau \geq \eta$, the sample is retained as it demonstrates satisfactory geometric consistency and semantic coherence; otherwise, it is filtered out as an unreliable sample.

Through sample-level geometric consistency verification (SGCV), the quality of single-modal image pairs can be effectively optimized prior to modality transformation, thereby establishing a reliable foundation for generating trustworthy multi-modal data and ultimately enhancing the overall usability of the synthesized dataset. Based on the aforementioned pipeline, we construct a synthetic dataset named Any-syn, which comprises both training and testing subsets.

4 Experiments

4.1 Implementation Details

Training Strategy: We directly adopt the official pre-trained models of EDM, LoFTR, and RoMa as initialization weights and perform fine-tuning on the Any-syn dataset. All models are trained on four NVIDIA RTX 4090 GPUs, with batch sizes of 4, 4, and 2 assigned to EDM, LoFTR, and RoMa, respectively. Tailored to the distinct characteristics of each baseline model, we employ different learning rates: for EDM, the original learning rate of 1×10^{-4} is maintained throughout 10 epochs of fine-tuning; for LoFTR, a learning rate of 8×10^{-4} is applied over 10 fine-tuning epochs; and for RoMa, the encoder and decoder learning rates are set to 6.25×10^{-8} and 1.25×10^{-6}, respectively, achieving convergence within only 4 fine-tuning epochs.

Model Details: For AnyMatch, we employ Moge V1 [35] as the monocular relative depth estimation model, Moge V2 [36] for surface normal map generation, and Stable Diffusion V2 [30] for image inpainting. The visible-to-infrared diffusion model is DiffV2IR [28]. The high-performance dense matching model

Table 1: Quantitative results (AUC of pose error) on the Any-syn dataset. The top-2 results of each category are highlighted as first and second.

Method	RGB-IR			RGB-Depth			RGB-Normal			RGB-Event		
	@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°
ReDFeat	3.80	10.64	23.19	0.20	0.81	2.93	0.90	3.15	8.94	0.19	0.84	3.23
XoFTR	27.03	47.37	66.11	1.55	4.05	9.29	10.33	21.62	35.49	0.89	2.34	5.51
ELoFTR$_{\text{PYM}}$	8.20	18.51	32.36	0.34	1.38	4.23	8.43	18.74	32.99	1.15	3.28	7.94
EDM	9.20	19.63	32.78	0.11	0.41	1.64	6.25	14.81	27.35	0.21	0.83	2.76
EDM$_{\text{MINIMA}}$	14.70	31.56	51.20	4.01	11.25	23.98	11.60	25.99	44.21	1.18	3.30	8.04
EDM$_{\text{AnyMatch}}$	28.80	50.00	68.53	8.32	20.36	38.09	18.23	36.65	56.60	3.31	9.32	19.86
LoFTR	8.35	17.92	30.13	0.06	0.17	0.73	3.20	8.70	17.48	0.32	1.08	2.71
LoFTR$_{\text{MINIMA}}$	13.09	28.75	48.20	2.29	9.27	19.98	9.89	22.59	39.96	1.06	3.43	8.67
LoFTR$_{\text{AnyMatch}}$	22.30	41.45	60.91	3.75	9.78	20.89	12.17	26.61	45.08	1.39	3.94	9.37
RoMa	30.08	48.65	65.94	2.44	7.00	16.03	19.63	36.24	52.50	1.09	3.10	6.45
RoMa$_{\text{MA}}$	34.45	55.68	72.82	10.22	22.76	39.21	26.35	45.79	63.59	3.42	9.16	18.31
RoMa$_{\text{MINIMA}}$	40.00	60.62	76.27	9.76	22.16	38.59	29.97	49.33	65.81	2.52	7.29	14.84
RoMa$_{\text{AnyMatch}}$	50.91	69.88	82.75	14.05	29.39	47.39	36.28	56.50	72.23	4.81	11.87	22.13

used for geometric verification is RoMa [7]. For sample-level geometric consistency verification, the pixel threshold τ is set to $\tau = 5$ and the minimum correct match ratio η is set to $\eta = 0.6$.

Datasets: We evaluate on five datasets, covering 17 cross-modal combinations. In the synthetic dataset Any-syn, the resolution of the images is 512×512 pixels, covering five modal combinations of RGB-RGB/IR/Depth/Normal/Event. The synthetic training subset is constructed from the GLDv2 dataset [38], comprising 500000 image pairs, while the synthetic test subset is derived from the SA-1B dataset [17], containing 10000 image pairs. Following MINIMA, we train exclusively on three modality pairs: RGB-IR, RGB-Depth, and RGB-Normal. For real-world evaluation, METU-VisTIR [34] provides 2590 RGB-IR image pairs with pose annotations for pose estimation benchmarks. DIODE [41] contains 27858 pre-aligned RGB-Depth/Normal image pairs. DSEC [37] supplies 100 rectified RGB-Event image pairs. To simulate geometric deformations while preserving evaluation fairness, synthetic homography transformations are applied to one image in each aligned pair prior to testing. Additionally, the MMIM dataset [16] encompasses 13 cross-modal combinations from remote sensing and medical imaging domains, with ground-truth correspondences manually annotated for homography transformation estimation tasks.

Evaluation Protocol: For pose estimation tasks, we report the Area Under the Curve (AUC) of the pose error at rotation error thresholds of 5°, 10°, 20°. For homography estimation tasks, we report the AUC of the maximum corner reprojection error across the four image corners at pixel thresholds of 3px, 5px, 10px. To ensure fair comparison across methods, all robust geometric estimation procedures employ identical RANSAC [8] hyperparameters.

4.2 Evaluation on Any-syn

To validate the quality and learnability of Any-syn and the effectiveness of AnyMatch on synthetic data, we first conduct systematic pose estimation evaluations on the Any-syn test subset, comparing three model variants: the original

Table 2: Quantitative results (AUC of pose error and reprojection error) in real RGB-IR/Depth/Normal/Event datasets. The top-2 results of each category are highlighted as first and second.

Method	RGB-IR			RGB-Depth			RGB-Normal			RGB-Event		
	@5°	@10°	@20°	@3px	@5px	@10px	@3px	@5px	@10px	@3px	@5px	@10px
ReDFeat	1.71	4.57	10.85	1.01	4.58	16.30	0.19	1.59	8.71	0.55	0.97	6.07
XoFTR	18.47	34.64	51.50	11.03	27.24	51.60	19.79	40.52	64.47	0.00	1.37	12.64
ELoFTR$_{\text{PYM}}$	14.45	29.73	46.55	25.70	42.90	62.45	1.75	7.02	21.37	0.31	0.87	10.64
EDM	5.20	15.82	31.71	1.21	5.75	19.44	6.08	18.25	41.82	0.60	1.60	8.89
EDM$_{\text{MINIMA}}$	15.32	32.80	53.47	8.44	24.89	52.09	13.69	33.72	59.62	0.27	1.41	**13.63**
EDM$_{\text{AnyMatch}}$	17.13	35.80	55.24	10.71	28.75	55.64	18.42	40.04	65.22	0.50	2.12	13.57
LoFTR	4.48	8.56	15.35	0.79	4.37	15.87	0.06	0.19	0.50	0.30	1.09	5.53
LoFTR$_{\text{MINIMA}}$	15.61	30.84	47.87	2.14	8.85	26.22	14.19	33.37	58.72	0.35	1.21	8.56
LoFTR$_{\text{AnyMatch}}$	14.48	25.83	40.39	4.08	13.33	34.63	13.83	34.03	60.69	0.34	1.83	12.78
RoMa	25.61	48.12	68.37	12.57	29.47	54.47	24.28	46.01	69.09	0.25	0.91	8.72
RoMa$_{\text{MA}}$	30.42	52.77	71.60	26.33	48.80	71.07	35.71	58.35	78.37	0.26	1.51	7.66
RoMa$_{\text{MINIMA}}$	37.45	60.70	78.00	29.77	51.09	72.64	37.91	59.24	77.88	0.25	1.37	13.16
RoMa$_{\text{AnyMatch}}$	40.49	62.61	77.76	29.12	49.83	71.50	39.26	60.02	78.35	0.32	1.64	14.25

pre-trained models, models fine-tuned by MINIMA, and models fine-tuned by AnyMatch. Furthermore, we also compare with ReDFeat, XoFTR, ELoFTR improved by PYM, and RoMa fine-tuned with MatchAnything (RoMa$_{\text{MA}}$). As shown in Tab. 1, models fine-tuned with AnyMatch consistently achieve superior performance across all multi-modal image pairs compared to baseline methods, MINIMA-fine-tuned counterparts, and other state-of-the-art multi-modal matchers, demonstrating that our framework effectively enhances cross-modal matching capabilities. Taking RoMa as an example, RoMa$_{\text{AnyMatch}}$ attains an AUC@10° of 69.88% on the RGB-IR task, representing an absolute improvement of 9.26% over RoMa$_{\text{MINIMA}}$ (60.62%) and a substantial gain of 21.23% (43.6% relative improvement) over the original RoMa baseline (48.65%). Owing to the fundamentally different sensing principle of event cameras—which record asynchronous per-pixel brightness changes as temporal event streams rather than full-frame intensities—the modality gap in RGB-Event pairs is significantly larger than in other modalities, resulting in generally lower performance across all methods. Nevertheless, even under this highly challenging setting, EDM$_{\text{AnyMatch}}$ achieves an AUC@10° of 9.32%, substantially outperforming EDM$_{\text{MINIMA}}$ (3.30%) by 6.02%. These results collectively validate the efficacy of AnyMatch in generating geometrically consistent and semantically plausible synthetic data for robust cross-modal matching.

4.3 Evaluation on Real Dataset

To validate the generalization capability of models trained by AnyMatch in real-world scenarios and provide more compelling evidence, we further conduct both in-domain and zero-shot evaluations across multiple real-world cross-modal datasets. Specifically, the RGB-IR dataset (METU-VisTIR) is employed for pose estimation benchmarks, while the remaining datasets (DIODE, DSEC, and MMIM) are utilized for homography estimation tasks.

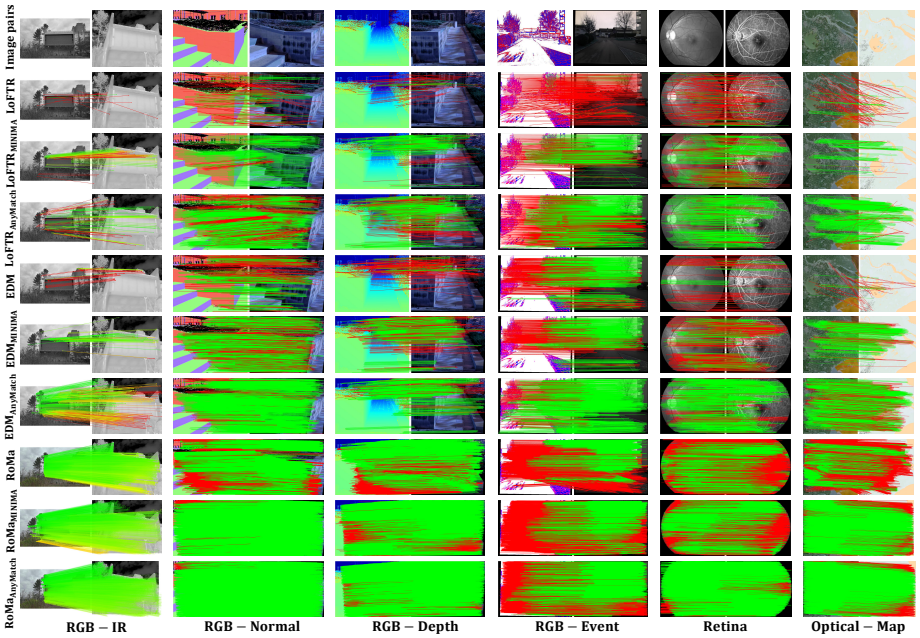

Fig. 4: Qualitative results on real multi-modal images. Comparison of LoFTR, LoFTR$_{\text{MINIMA}}$, LoFTR$_{\text{AnyMatch}}$, EDM, EDM$_{\text{MINIMA}}$, EDM$_{\text{AnyMatch}}$, RoMa, RoMa$_{\text{MINIMA}}$, and RoMa$_{\text{AnyMatch}}$ in multiple image pairs, including RGB-IR, RGB-Normal, RGB-Depth, RGB-Event, Retina, and Optical-Map (from left to right). 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.

In-domain evaluation: The results on real-world RGB-IR/RGB-Depth/ RGB-Normal datasets are presented in Tab. 2 and Fig. 4. For EDM, EDM$_{\text{AnyMatch}}$ consistently outperforms EDM$_{\text{MINIMA}}$ on nearly all modalities. Regarding RoMa and LoFTR, owing to the influence of inpainting hallucinations and systematic errors in depth estimation within the AnyMatch pipeline, MINIMA-fine-tuned models achieve marginally higher or comparable results on certain specific tasks (*e.g.*, LoFTR on RGB-IR and RoMa on RGB-Depth). Nevertheless, AnyMatch maintains strong competitiveness against state-of-the-art multi-modal matchers.

Zero-shot evaluation: Zero-shot results on the unseen medical and remote sensing image datasets (MMIM) are reported in Tab. 3 and Fig. 4. In the results, AnyMatch demonstrates consistently superior generalization capability compared to MINIMA and other state-of-the-art matchers.

AnyMatch not only delivers superior performance on known modalities but also establishes state-of-the-art results in zero-shot tasks involving unseen modalities. This demonstrates that models fine-tuned on AnyMatch learn modality-invariant feature representations and enhance 3D geometric awareness, achieving stronger generalization capability across diverse cross-modal matching scenarios.

Table 3: Quantitative results (AUC of reprojection error) in real medical and remote sensing dataset. The top-2 results of each category are highlighted as first and second.

Method	Medical			Remote Sensing		
	@3px	@5px	@10px	@3px	@5px	@10px
ReDFeat	31.44	39.16	45.62	1.92	5.91	19.68
XoFTR	38.97	45.23	51.89	23.31	34.44	53.56
ELoFTR$_{\text{PYM}}$	14.59	21.90	32.15	0	0	10.12
EDM	38.79	44.10	50.61	16.95	26.50	44.12
EDM$_{\text{MINIMA}}$	39.57	45.42	52.09	18.29	30.52	52.81
EDM$_{\text{AnyMatch}}$	39.76	45.79	55.18	25.17	38.69	57.90
LoFTR	37.89	42.73	47.36	5.94	9.34	12.57
LoFTR$_{\text{MINIMA}}$	38.86	44.79	52.35	20.93	33.96	54.30
LoFTR$_{\text{AnyMatch}}$	38.97	45.17	52.93	24.36	39.76	58.77
RoMa	39.16	44.79	53.89	26.29	37.55	55.45
RoMa$_{\text{MA}}$	37.84	43.92	55.42	27.60	42.83	60.28
RoMa$_{\text{MINIMA}}$	37.92	44.67	57.07	29.27	44.19	63.08
RoMa$_{\text{AnyMatch}}$	38.10	46.20	59.08	30.25	44.06	63.48

4.4 Ablation Studies

To analyze the impact of various factors in AnyMatch on model performance, we adopt EDM as the baseline model and perform a series of ablation studies on the METU-VisTIR dataset.

(1) **Number of modalities.** As shown in Tab. 4, we fine-tune EDM using datasets containing either a single modality pair (RGB-IR) or three modality pairs (RGB-IR, RGB-Depth, RGB-Normal) to validate the impact of modality diversity on matching performance. The results demonstrate that joint training across multiple modalities consistently yields significant performance improvements compared to single-modality training, highlighting the benefit of multi-modal collaborative learning in enhancing cross-modal generalization.

Table 4: Ablation Studies. Test on real METU-VisTIR dataset using AUC of pose error.

Training Strategy	@5°	@10°	@20°
1 modality w/o SGCV	13.55	27.71	44.50
3 modalities w/o SGCV	15.63	31.24	48.29
3 modalities w/ SGCV	17.13	35.80	55.24
3 modalities w/ SGCV + from scratch	5.69	14.16	27.45
3 modalities w/ SGCV + homo	15.45	30.78	46.16

(2) **Sample-level geometric consistency verification.** We apply the verification threshold η with values of 0.7, 0.6, 0.5, 0.4, 0.3, and 0 (no filtering) to progressively filter the original synthetic dataset, and subsequently fine-tune EDM on each filtered subset. As reported in Tab. 5, the optimal performance is achieved when $\eta = 0.6$. This result demonstrates that selectively retaining samples with high geometric consistency effectively eliminates "hallucination" artifacts introduced during the generative process, thereby substantially enhancing model robustness and matching accuracy.

(3) **Fine-tuning *vs.* From Scratch.** Comparing the strategies of training from scratch and fine-tuning, results in Tab. 4 reveal that models trained from scratch achieve substantially inferior performance compared to fine-tuned counterparts, with the AUC@20° metric reaching only approximately 50% of the fine-tuned model's score. This finding validates that leveraging the matching

Table 5: Results of sample-level geometric consistency verification using varying thresholds. Models fine-tuned on Any-syn subsets filtered by different verification thresholds ($\eta \in \{0.7, 0.6, 0.5, 0.4, 0.3, 0\}$) are evaluated on the METU-VisTIR dataset.

Threshold	0.7	0.6	0.5	0.4	0.3	0
AUC@5°	15.50	17.13	15.39	15.36	15.73	15.63
AUC@10°	32.52	35.80	31.47	32.04	31.32	31.24
AUC@20°	51.67	55.24	48.97	50.08	48.55	48.29

prior knowledge embedded in pre-trained models is crucial for effective convergence and robust adaptation in cross-modal matching tasks.

(4) **2D Homography-based Synthesis *vs.* 3D View Transformation-based Synthesis.** We construct a homography-synthetic dataset (homo-syn) by applying synthetic homography transformations to original RGB images followed by the same modality transformation pipeline in AnyMatch. In Tab. 4, models fine-tuned on homo-syn exhibit substantially degraded performance compared to those fine-tuned on Any-syn. This result validates that disparity and occlusion relationships generated via 3D reprojection better conform to real-world physical constraints and geometric priors, thereby yielding more realistic and geometrically consistent training data than simplistic planar transformations.

5 Conclusion

In conclusion, to address the scarcity of large-scale training data with precise geometric annotations for multi-modal image matching, we propose AnyMatch—a framework that transforms abundant, easily accessible single-view images at minimal cost into multi-modal training data with high 3D geometric fidelity. By decoupling view and modality transformation, AnyMatch integrates monocular depth estimation, 3D reprojection, diffusion-based inpainting, and cross-modal translation to synthesize multi-view, multi-modal pairs and annotations that adhere to 3D geometric consistency, while employing Sample-level Geometric Consistency Verification (SGCV) to filter generative artifacts. Moreover, AnyMatch offers strong scalability, enabling controllable scene diversity and annotation difficulty levels via adjustable input and camera parameters. Experiments on the Any-syn dataset demonstrate that matching networks (LoFTR, EDM, RoMa) fine-tuned on Any-syn achieve substantial gains on cross-modal benchmarks, exhibiting superior generalization and robustness over baselines and MINIMA-fine-tuned models, including in zero-shot evaluations on unseen modalities.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62276192 and 62475199) and National Key R&D Program of China (Grant No. 2024YFE0202500).

References

1. Aditya, N., Dhruval, P., et al.: Thermal voyager: A comparative study of rgb and thermal cameras for night-time autonomous navigation. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 14116–14122 (2024)
2. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Proceedings of the European Conference on Computer Vision. pp. 404–417 (2006)
3. Cadar, F., Potje, G., Martins, R., Demonceaux, C., Nascimento, E.R.: Leveraging semantic cues from foundation vision models for enhanced local feature correspondence. In: Proceedings of the Asian Conference on Computer Vision. pp. 1268–1283 (2024)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5828–5839 (2017)
5. Deng, X., Liu, E., Gao, C., Li, S., Gu, S., Xu, M.: CrossHomo: Cross-modality and cross-resolution homography estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5725–5742 (2024)
6. Deng, Y., Ma, J.: Redfeat: Recoupling detection and description for multimodal feature learning. IEEE Transactions on Image Processing **32**, 591–602 (2022)
7. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust dense feature matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 19790–19800 (2024)
8. Fischler, M.A., Bolles, R.C.: 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)
9. Frolov, A., Rodehorst, V.: Patch your matcher: Correspondence-aware image-to-image translation unlocks cross-modal matching via single-modality priors. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 7913–7924 (2026)
10. Gao, C., Li, W., Weng, D.: HOMO-feature: Cross-arbitrary-modal image matching with homomorphism of organized major orientation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10538–10548
11. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3749–3761 (2022)
12. Han, Z., Mao, C., Jiang, Z., Pan, Y., Zhang, J.: Stylebooth: Image style editing with multimodal instruction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1947–1957 (2025)
13. He, X., Yu, H., Peng, S., Tan, D., Shen, Z., Bao, H., Zhou, X.: MatchAnything: Universal cross-modality image matching with large-scale pre-training. arXiv preprint arXiv:2501.07556 (2025)
14. Hou, Z., Liu, Y., Zhang, L.: POS-GIFT: A geometric and intensity-invariant feature transformation for multimodal images. Information Fusion **102**, 102027 (2024)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-Rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (2022)

16. Jiang, X., Ma, J., Xiao, G., Shao, Z., Guo, X.: A review of multimodal image matching: Methods and applications. Information Fusion **73**, 22–71 (2021)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4015–4026 (2023)
18. Li, J., Hu, Q., Ai, M.: RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. IEEE Transactions on Image Processing **29**, 3296–3310 (2019)
19. Li, J., Hu, Q., Zhang, Y.: Multimodal image matching: A scale-invariant algorithm and an open dataset. ISPRS Journal of Photogrammetry and Remote Sensing **204**, 77–88 (2023)
20. Li, X., Rao, T., Pan, C.: EDM: Efficient deep feature matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 26198–26208 (2025)
21. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2041–2050 (2018)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
23. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
24. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5802–5811 (2022)
25. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
26. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4040–4048 (2016)
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
28. Ran, L., Wang, L., Wang, G., Wang, P., Zhang, Y.: DiffV2IR: visible-to-infrared diffusion model via vision-language understanding. arXiv preprint arXiv:2503.19012 (2025)
29. Ren, J., Jiang, X., Li, Z., Liang, D., Zhou, X., Bai, X.: MINIMA: Modality invariant image matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23059–23068 (2025)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
31. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4104–4113 (2016)
32. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision. pp. 501–518 (2016)

33. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8922–8931 (2021)
34. Tuzcuoğlu, Ö., Köksal, A., Sofu, B., Kalkan, S., Alatan, A.A.: Xoftr: Cross-modal feature matching transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4275–4286 (2024)
35. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5261–5271 (2025)
36. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: MoGe-2: Accurate monocular geometry with metric scale and sharp details. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 35928–35959 (2026)
37. Wang, X., Li, J., Zhu, L., Zhang, Z., Chen, Z., Li, X., Wang, Y., Tian, Y., Wu, F.: VisEvent: Reliable object tracking via collaboration of frame and event flows. IEEE Transactions on Cybernetics **54**(3), 1997–2010 (2023)
38. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2575–2584 (2020)
39. Wu, J., Xu, R., Wood-Doughty, Z., Wang, C., Xu, S., Lam, E.Y.: Segment anything model is a good teacher for local feature learning. IEEE Transactions on Image Processing **34**, 2097–2111 (2025)
40. Xu, H., Yuan, J., Ma, J.: Murf: Mutually reinforcing multi-modal image registration and fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12148–12166 (2023)
41. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 21875–21911 (2024)
42. Yang, M., Fan, F., Huang, J., Ma, Y., Mei, X., Cai, Z., Ma, J.: Multimodal image matching based on cross-modality completion pre-training. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 2206–2214 (2025)
43. Zhang, H., Xu, H., Tian, X., Jiang, J., Ma, J.: Image fusion meets deep learning: A survey and perspective. Information Fusion **76**, 323–336 (2021)
44. Zhang, K., Ma, J.: Sparse-to-dense multimodal image registration via multi-task learning. In: Proceedings of the International Conference on Machine Learning (2024)
45. Zhou, K., Chen, C., Wang, B., Saputra, M.R.U., Trigoni, N., Markham, A.: VM-Loc: Variational fusion for learning-based multimodal camera localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 6165–6173 (2021)