Sparse-to-dense Multimodal Image Registration via Multi-Task Learning

Kaining Zhang [1] **Jiayi Ma** [1]

Abstract

Aligning image pairs captured by different sensors or those undergoing significant appearance changes is crucial for various computer vision and robotics applications. Existing approaches cope with this problem via either Sparse feature Matching (SM) or Dense direct Alignment (DA) paradigms. Sparse methods are efficient but lack accuracy in textureless scenes, while dense ones are more accurate in all scenes but demand for good initialization. In this paper, we propose SDME, a Sparse-to-Dense Multimodal feature Extractor based on a novel multi-task network that simultaneously predicts SM and DA features for robust multimodal image registration. We propose the sparse-to-dense registration paradigm: we first perform initial registration via SM and then refine the result via DA. By using the well-designed SDME, the sparse-to-dense approach combines the merits from both SM and DA. Extensive experiments on MSCOCO, GoogleEarth, VIS-NIR and VIS-IR-drone datasets demonstrate that our method achieves remarkable performance on multimodal cases. Furthermore, our approach exhibits robust generalization capabilities, enabling the fine-tuning of models initially trained on single-modal datasets for use with smaller multimodal datasets. Our code is available at `https://github.com/KN-Zhang/SDME`.

1. Introduction

Pixel-wise alignment of multimodal images is of essential importance in computer vision (Ma et al., 2021; Lu et al., 2023), robotics (Cadena et al., 2016; Nguyen et al., 2020), remote sensing (Audebert et al., 2018; Gómez-Chova et al., 2015), medical imaging (Collignon et al., 1995), and many others. For example, a SLAM system is likely to be equipped with a RGB and an infrared camera to work at night; a drone needs to navigate based on the satellite maps and instant photos. Aligning these images suffers from multimodal color representations along with appearance changes. So the challenge of multimodal image registration not only lies in the large motion variations between image pairs, but also how to deal with the modality difference.

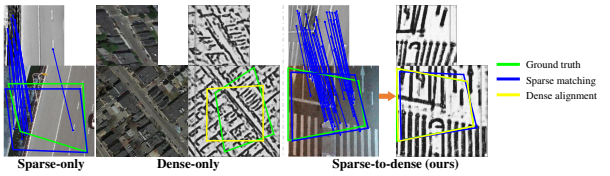

Figure 1. Visualization of homography estimation on three multimodal image pairs when sparse matching (**sparse-only**), dense alignment (**dense-only**), and our **sparse-to-dense** strategies are used. All operations are performed with our learned features. Each image pair is composed of an input image I_I (the larger one) and a template image I_T (the smaller one). Green, blue and yellow polygons denote the positions of I_T on I_I with the homography given by ground truth, sparse matching and dense alignment.

In this paper, we aim to estimate homography to align image pairs under the flat world assumptions (Goforth & Lucey, 2019; Zhao et al., 2021). Most approaches in this field can be categorized into two classes: Sparse feature Matching (**SM**) and Direct Alignment (**DA**). SM consists of three separate phases: keypoint detection, description and matching. Then homography is recovered from the sparse matches. This pipeline is very efficient and can deal with large motion variations since it only considers sparse keypoints with viewpoint-invariant descriptors (Revaud et al., 2019; Xue et al., 2023). However, as the sparse-only case shown in Fig. 1, such sparsity introduces robust issue under a textureless scene. DA regards homography as an optimization variable and finds the optimal solution by minimizing geometric (Schonberger & Frahm, 2016), photometric (Engel et al., 2014) or feature-metric (Tang & Tan, 2018; Zhao et al., 2021; Sarlin et al., 2021; Zhang et al., 2023) errors over all pixels. This minimization consists of aggregating the matching costs over all image pixels iteratively, so it is more accurate yet computationally more demanding than the SM counterparts. Moreover, it is easily trapped in local minima with an inappropriate initialization, as depicted in the dense-only case in Fig. 1.

[1]Electronic Information School, Wuhan University, Wuhan, China. Correspondence to: Jiayi Ma <jyma2010@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

We propose to combine SM and DA paradigms to perform multimodal image registration in a Sparse-to-Dense (**S2D**) manner, *i.e.*, we first conduct SM, and the estimation results from this step serve as the initialization for DA. In order to do this, two key issues should be carefully investigated: *i) difference between different modalities*, and *ii) difference between SM and DA registration.* We use advanced contrastive learning that has shown great promise in multimodal data (Xu et al., 2023; Xing et al., 2023) for the first issue. Turning to the latter, a natural question emerges: can we use the same features for both SM and DA successively? Unfortunately, it is not feasible since features for SM are required to be locally discriminative for reliable matches, while those for DA are designed to be locally smooth with regard to image pixels to widen the convergence basin.

Given that two kinds of features need to be extracted in the S2D pipeline, we introduce multi-task learning (MTL) to save memory and inference speed. In other words, we treat learning features for SM and DA as two tasks and design a unified network with sharing and task-specific parameters to predict them. Prior works have shown that MTL allows better performance than learning each task independently since it favours the exchange of complementary signals across tasks (Guo et al., 2020; Bansal et al., 2023). But it struggles with learning all tasks equally (Standley et al., 2020; Wang et al., 2023). Therefore, to enjoy the merit of MTL and prevent it from being dominated by one of the tasks, especially two competing ones, we employ an off-the-shelf multiple gradient descent algorithm (Sener & Koltun, 2018) for training. Besides, we propose mutual guidance of different tasks to boost task interaction, leading to more robust features. As Fig. 1 shows, features for DA highlight invariant structures between different modalities, which can help keypoint learning in SM. Meanwhile, keypoints in SM reveal which locations in the image are more reliable, helping weight the feature-metric loss in DA to steer the optimization direction.

In summary, we present the following contributions:

- We propose an S2D strategy for multimodal image registration. It combines the advantages of SM and DA paradigms, at the same time allowing high efficiency and high accuracy even under large motion variations.
- We propose an MTL network termed as Sparse-to-Dense Multimodal feature Extractor (SDME) to predicts SM and DA features. During training, we introduce a variant of the multiple gradient descent method to balance the conflict objectives of the two tasks, and design mutual guidance to enhance task interaction.
- Extensive experiments show that our method outperforms the state-of-the-art competitors on multimodal datasets and exhibits robust generalization abilities.

2. Related Work

Either homography or other transformation such as affine and 6-DoF camera pose can be parametrized in some forms, followed by being estimated in the feature-based or direct paradigm. Thus we mainly make a brief literature review of these two paradigms, not limited to homography estimation.

Feature-based. Methods belonging to this pipeline proceed by i) detecting and describing keypoints, ii) matching based on the similarity between descriptors, and iii) estimating transformation via RANSAC (Fischler & Bolles, 1981) or its variants (Chum et al., 2003; Barath et al., 2020). The used feature extractor affects the final results a lot, which has transformed from the handcrafted SIFT (Lowe, 2004), ORB (Rublee et al., 2011) to the deep learning-based D2 (Dusmanu et al., 2019), R2D2 (Revaud et al., 2019), ASLFeat (Luo et al., 2020), and SFD2 (Xue et al., 2023). Although these learning-based ones have shown superiority over the handcrafted counterparts when dealing with extreme viewpoint and illumination changes, they are trained on correspondences with single modality. This is not the case for multimodal images since correspondences between different modalities may have a completely different look. To address this issue, some feature extractors (Xiang et al., 2018; Li et al., 2018; Deng & Ma, 2023) are later designed specifically for multimodal data. In this work, we learn invariant structures between different modalities and use them to guide how to learn a robust feature extractor for SM.

Direct. This paradigm aims at minimizing some costs over all image pixels via the Gauss-Newton or Levenberg-Marquardt algorithm (Nocedal & Wright, 1999). Specifically, the geometric cost that minimizes re-projection errors is the golden standard for structure-from-motion in the last two decades, but its performance is limited by its using single image information (Schonberger & Frahm, 2016) (*i.e.*, image corners, blobs). The photometric cost tries to minimize pixel intensity difference of aligned pixels, which is easily affected by illumination changes and has a narrow convergence due to its high non-convexity (Engel et al., 2014). Recently, training models that can minimize feature-metric errors across different views have shown great promise (Chang et al., 2017; Tang & Tan, 2018; Sarlin et al., 2021). These approaches have wide convergence and are robust against appearance changes. However, they are not efficient because the used features are usually of high dimension which leads to a large computational burden during optimization. DeepLK (Zhao et al., 2021) addresses this by designing a single-channel feature map. During DA, it first constructs a feature pyramid with three unshared networks and then optimizes the homography in a coarse-to-fine manner. Although this single-channel feature map improves optimization efficiency, its convergence is limited and highly dependent on initialization conditions. In this

work, we learn features for DA based on DeepLK, and solve the initialization problem via SM.

Deep Homography Estimation. As this work focuses on homography estimation, some deep approaches that do not belong to any of the two paradigms need to be mentioned. They use a single or cascade VGG-style network to directly or iteratively output homography parametrized by eight independent parameters or the locations of four image corners (DeTone et al., 2016; Le et al., 2020; Cao et al., 2022).

Multi-task Learning. MTL aims to improve the average performance of multiple target tasks from training together. Hard parameter sharing is the most common setting in MTL. It means a subset of the parameters is shared between tasks while other parameters are task-specific. However, among tasks with incompatible objectives, improper design of networks or training losses easily leads to imbalance task-wise performance (Sener & Koltun, 2018; Standley et al., 2020; Wang et al., 2023). This is because task gradients may interfere and multiple summed losses may make the optimization landscape more difficult. This issue also exists in our work. We use (Sener & Koltun, 2018) to dynamically modify the gradient direction for mitigating conflict.

3. Method

Given an input image I_I and a template image I_T, we aim to achieve pixel-wise alignment by estimating the underlying 8-DoF homography between them. It can be denoted as $\hat{\mathbf{x}}_I = \mathbf{H}\hat{\mathbf{x}}_T$, where $\hat{\mathbf{x}}_I$ and $\hat{\mathbf{x}}_T$ are homogeneous coordinates of the pixels in I_I and I_T, $\mathbf{H}$ is a non-singular 3×3 homography matrix to be estimated.

3.1. Sparse-to-dense Image Registration Pipeline

To combine the advantages of SM and DA, we propose a sparse-to-dense (S2D) strategy. It first estimates $\mathbf{H_0}$ via SM, which is then regarded as initialization and refined by DA.

SM. In this step, a set of keypoints together with their corresponding descriptors are first extracted from I_I and I_T. Then a putative set is built based on the descriptor similarity with a mutually nearest neighbor standard. Afterwards, $\mathbf{H_0}$ is estimated by a robust estimator such as RANSAC.

DA. We vectorize $\mathbf{H}$ and normalize it such that the last element is equal to one. Then homography is represented by the first eight elements, denoted as $\mathbf{p} \in \mathbb{R}^8$. Let $\mathbf{X}_T, \mathbf{X}_I \in \mathbb{R}^{H\times W\times C}$ denote the features of I_T and I_I, respectively. The goal of DA is to find a more accurate homography by iteratively minimizing a feature-metric projection error:

$$E(\mathbf{p}) = \textstyle\sum_i \|\mathbf{r}_i\|_2^2, \text{ where } \mathbf{r}_i = \mathbf{X}_T[i] - \mathbf{X}_I[W(i;\mathbf{p})], \quad (1)$$

where i is a pixel in I_T, $W(\cdot;\mathbf{p}) : \mathbf{R}^2 \to \mathbf{R}^2$ is the warp function parametrized by $\mathbf{p}$ and $[\cdot]$ is a lookup with sub-pixel interpolation. Following (Chang et al., 2017; Zhao et al., 2021), we use Inverse Compositional Lucas-Kanade (IC-LK) to improve the efficiency of optimization. Specifically, we optimize Eq. (1) from $\mathbf{p}_0$, which is given by $\mathbf{H}_0$. Then in the k-th iteration, $\mathbf{p}_k$ is updated to $\mathbf{p}_k + (\Delta\mathbf{p})^{-1}$ with

$$\mathbf{J}_i = \frac{\partial \mathbf{r}_i}{\partial \mathbf{p}} = \frac{\partial \mathbf{X}_T[i]}{\partial i}\frac{\partial i}{\partial \mathbf{p}} \in \mathbb{R}^{C\times 8}, \mathbf{G}_i = \mathbf{J}_i^T\mathbf{J}_i \in \mathbb{R}^{8\times 8},$$
$$\Delta\mathbf{p} = \left(\textstyle\sum_i \mathbf{G}_i\right)^{-1}\left(\textstyle\sum_i \mathbf{J}_i^T\mathbf{r}_i\right). \quad (2)$$

More details of Eqs. (1)-(2) are provided in *Appendix* A.

What We Want to Do. We aim to learn Sparse-to-Dense Multimodal feature Extractor (SDME) for our S2D strategy. Specifically, the extracted keypoints and descriptors should be robust to modality differences to enable SM to provide good initialization. Meanwhile, the feature $\mathbf{X}$ used in DA should help the objective function in Eq. (1) converge better, regardless its highly non-linear nature. We design SDME in an MTL manner, with the goal of reducing parameters, enhancing efficiency, and simultaneously improving the robustness of features through task interaction.

Network Design. For SM, R2D2 (Revaud et al., 2019) proposes a 9-layer lightweight network which is dominated by the dilation convolution for cheap runtime and momery cost. We follow them and split the network into two branches starting from the 5-th layer to build our multi-task network, which is shown in Fig. 2.

3.2. Learning Features for Sparse Matching (Task 1)

Features for SM are output from the sparse branch and we denote them as SDME-S. In this branch, two tensors are predicted for an image of size $H \times W$. The first one is a 3D tensor $\mathbf{D} \in \mathbb{R}^{H\times W\times 128}$ that corresponds to per-pixel local descriptor. The second one is a heatmap $\mathbf{S} \in [0,1]^{H\times W}$ that indicates sparse yet repeatable keypoint locations. SDME-S is composed of a keypoint and its corresponding descriptor.

Modality-invariant Transformer Block (MITB). Inspired by (Weinzaepfel et al., 2022), we apply the attention mechanism between descriptors and a set of **learnable** modality-invariant elements to enlarge the receptive field of descriptors in $\mathbf{D}$ in an efficient way. Specifically, let $\hat{\mathbf{D}} \in \mathbb{R}^{128\times HW}$ denote the flattened $\mathbf{D}$ and $\mathbf{A} \in \mathbb{R}^{128\times M}$ denote M learnable elements, then the queries, keys and values in the attention operation can be calculated by:

$$\mathbf{Q} = \mathbf{W}^{\mathcal{Q}}\mathbf{A}, \mathbf{K} = \mathbf{W}^{\mathcal{K}}\hat{\mathbf{D}}, \mathbf{V} = \mathbf{W}^{\mathcal{V}}\hat{\mathbf{D}}, \quad (3)$$

where $\mathbf{W}^{\mathcal{Q}} \in \mathbb{R}^{d_q\times 128}, \mathbf{W}^{\mathcal{K}} \in \mathbb{R}^{d_k\times 128}, \mathbf{W}^{\mathcal{V}} \in \mathbb{R}^{d_v\times 128}$ indicate learned linear transformations without bias, and we set $d_q = d_k = d_v = 128$. In this way, $\mathbf{A}$ can be updated by

$$\mathbf{A} = \mathbf{V} \cdot Softmax\left(\mathbf{K}^T\mathbf{Q}\right). \quad (4)$$

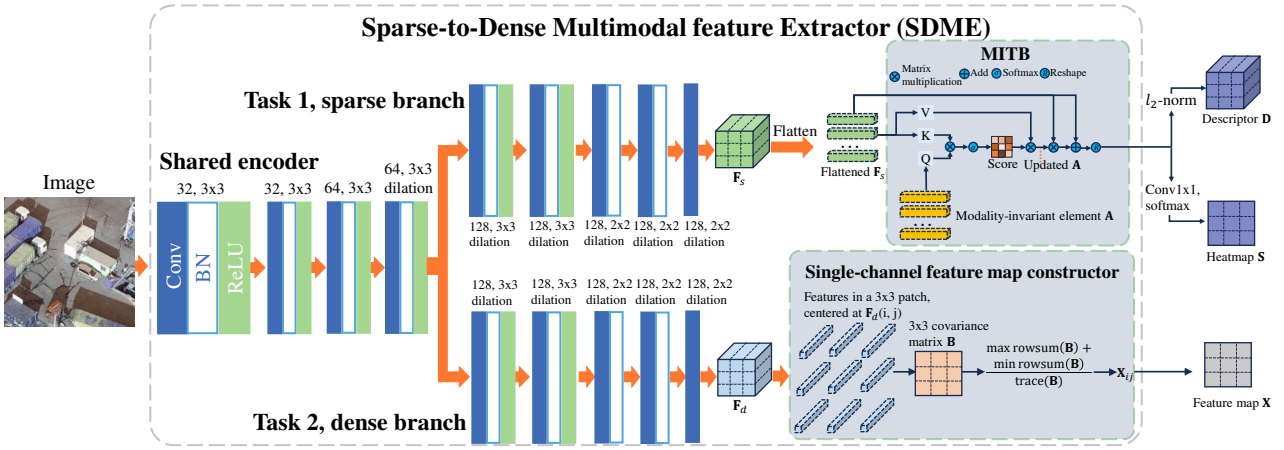

Figure 2. The architecture of the proposed sparse-to-dense multimodal feature extractor. We provide the number and size of convolution kernels. MITB refers to the modality-invariant transformer block. The single-channel feature map constructor transforms features of dimensions $H \times W \times 128$ into a final size of $H \times W$.

The updated $\mathbf{A}$ has interacted with all descriptors in $\mathbf{D}$, thus based on it the contextual descriptors can be acquired by

$$\hat{\mathbf{D}} = \hat{\mathbf{D}} + \mathbf{AP}, \text{ where } \mathbf{P} = \mathbf{A}^T\hat{\mathbf{D}}. \tag{5}$$

Each element in $\mathbf{A}$ is a function of all local features in $\mathbf{D}$, thus it further enlarges the receptive field of $\mathbf{D}$. To make M elements in $\mathbf{A}$ attend different areas in the images, the following loss is adopted:

$$\mathcal{L}_{attn} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \frac{\langle \mathbf{A}_i, \mathbf{A}_j \rangle}{\|\mathbf{A}_i\|_2 \|\mathbf{A}_j\|_2}, \tag{6}$$

where $\langle \cdot, \cdot \rangle$ is an operation for the inner product and $\| \cdot \|_2$ refers to L2-normalization.

Discriminative Descriptors. Descriptors in $\mathbf{D}$ are expected to be as discriminative as possible to make feature matching reliable. To this end, we maximize the Average-Precision (AP) metric for all local descriptors in $\mathbf{D}$. We start by randomly sampling N points in the first image and identifying their corresponding descriptors in $\mathbf{D}$ as queries, denoted as $\{\mathbf{d}_i\}_{i=1}^{N}$. Then each point is mapped to the second image according to $\mathbf{H}_{gt}$, centered on which some points are sampled within different radius neighborhoods and their corresponding descriptors are identified as positive and negative samples, *i.e.*, $\{\mathbf{d}_{ij}^{+}\}_{j=1}^{N_p}$ and $\{\mathbf{d}_{ij}^{-}\}_{j=1}^{N_n}$. Besides, we randomly sample some points in the second image and regard their descriptors $\{\mathbf{d}_{ij}^{*}\}_{j=1}^{N_n}$ as distractors to further strengthen the robustness of local descriptors. We calculate the similarity scores among $\mathbf{d}_i$ and all other samples and obtain a differentiable approximation of AP ranking loss, denoted as $\widetilde{AP}$ (He et al., 2018). Then we minimize

$$\mathcal{L}_{AP} = \textstyle\sum_i 1 - \widetilde{AP}(\mathbf{d}_i) \tag{7}$$

to ensure that local descriptors are accurate enough for SM.

Repeatable Keypoints. Good keypoints should be with high repeatability, which means their positions should be invariant to natural image transformations such as viewpoint or illumination changes. We follow $\mathcal{L}_{cosim}$ and $\mathcal{L}_{peaky}$ proposed in R2D2 to achieve this goal. Concretely, $\mathcal{L}_{cosim}$ aims to enforce heatmaps of two images to have high similarity in corresponding local patches, while $\mathcal{L}_{peaky}$ tries to maximize the local peakness of the heatmap within each patch. More details can be found in *Appendix* B.

3.3. Learning Features for Dense Alignment (Task 2)

For an image of size $H \times W$, the dense branch will first output $\mathbf{F}_d$ of size $H \times W \times 128$. Directly using $\mathbf{F}_d$ for Eq. (2) results in $O(128^3)$ time complexity for the calculation of $\mathbf{G}_i^{-1}$. To reduce the complexity, we follow (Zhao et al., 2021) and regard $\mathbf{F}_d$ as $W \cdot H$ vectors of 128-dimension, and cast each vector as the center and calculate the covariance matrix in its 3×3 patch. Then for each covariance matrix, a value is computed by taking the ratio of the sum between the maximum and minimum row sums to the trace of the matrix. This process is depicted in Fig. 2 and it builds a single-channel feature map $\mathbf{X} \in \mathbb{R}^{H \times W}$ for each image by traversing all vectors in $\mathbf{F}_d$. We denote $\mathbf{X}$ as SDME-D and use it for the following DA. In this case, the time complexity decreases from $O(128^3)$ to $O(1)$. To guarantee Eq. (1) converge better during DA, we follow $\mathcal{L}_{conv1}$ and $\mathcal{L}_{conv2}$ proposed in (Zhao et al., 2021) to learn $\mathbf{X}$. More details can be referred to *Appendix* C.

Besides, a modality consistency loss $\mathcal{L}_{mc}$ is introduced to build connection between different modalities:

$$\mathcal{L}_{mc} = \textstyle\sum_i \|\mathbf{X}_T[i] - \mathbf{X}_I[W(i; \overline{\mathbf{p}})]\|_2^2, \tag{8}$$

where $\overline{\mathbf{p}}$ is the ground truth.

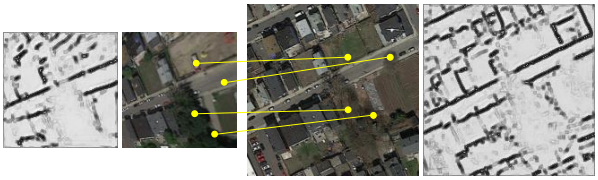

Figure 3. Four correspondences & modality-invariant structures of a multimodal pair with different plant conditions.

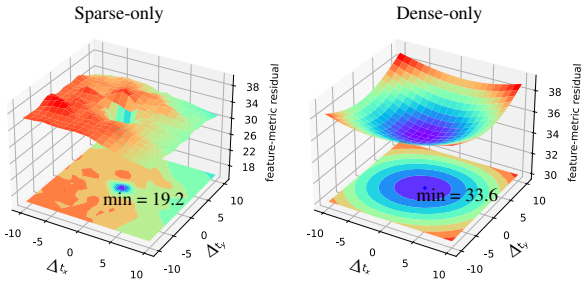

Figure 4. Left: only the sparse branch is trained. Right: only the dense branch is trained. Keeping the other six parameters in the homography matrix fixed, we perturb the translation vector with $(\Delta t_x, \Delta t_y)$ in two directions around the ground truth and calculate the feature-metric residual by replacing $\mathbf{X}$ in Eq. (1) with $\mathbf{D}$ and $\mathbf{F}_d$ shown in Fig. 2.

3.4. Training Strategy: Mutual Guidance and Multi-task Learning

We have introduced how to learn SDME-S and SDME-D separately. In the following, we will explore the interaction of the two tasks for better performance.

Task 2 Guides Task 1. Given that the same object can be displayed in vastly different forms in multimodal scenarios (Fig. 3), demanding a high likelihood of detecting keypoints in such areas can lead to challenges in descriptor learning and consequently hinder the performance of SM. According to (Zhao et al., 2021), features learned by $\mathcal{L}_{conv1}$ and $\mathcal{L}_{conv2}$ can spontaneously recognize invariant structures between different modalities. So we use SDME-D (denoted as $\mathbf{X}$) to guide the learning of the heatmap $\mathbf{S}$ in Task 1. We design the following loss to achieve this goal:

$$\mathcal{L}_{guide} = \left\| \mathbf{S} - Softmax\big(Relu(\lambda)(1-\tilde{\mathbf{X}})\big) \right\|_2, \quad (9)$$

where $\tilde{\mathbf{X}} \in [0,1]$ is the normalized $\mathbf{X}$, λ is a learnable parameter that eliminates scale difference between $\mathbf{S}$ and $\mathbf{X}$. In this way, modality-invariant structures, such as corners and edges, will exhibit higher detection scores in $\mathbf{S}$ compared to other ambiguous areas.

Task 1 Guides Task 2. The heatmap $\mathbf{S}$ learned by our sparse branch reveals reliable and prominent structures in the image, which can help steer the optimization in DA towards the correct result. To achieve this goal, we weight the objective function in Eq. (1) with $\mathbf{S}$, such that:

$$E(\mathbf{p}) = \textstyle\sum_i w_i \|\mathbf{r}_i\|_2^2, \quad (10)$$

where $w_i = \mathbf{S}_T[i] \cdot \mathbf{S}_I[W(i;\mathbf{p})] \in [0,1]$. Notably, w_i tends to 1 if the pixel is of high repeatability in both I_T and I_I, otherwise 0 if it locates at ambiguous areas which will impair the optimization. Then Eq. (2) turns to:

$$\mathbf{J}_i = \frac{\partial \mathbf{r}_i}{\partial \mathbf{p}} = \frac{\partial \mathbf{X}_T[i]}{\partial i}\frac{\partial i}{\partial \mathbf{p}} \in \mathbb{R}^{C\times 8}, \mathbf{G}_i = w_i \mathbf{J}_i^T \mathbf{J}_i \in \mathbb{R}^{8\times 8},$$
$$\Delta\mathbf{p} = \big(\textstyle\sum_i \mathbf{G}_i\big)^{-1}\big(\sum_i w_i \mathbf{J}_i^T \mathbf{r}_i\big). \quad (11)$$

Multi-task Learning. We denote parameters in the shared encoder as θ^{sh}, while parameters in task-specific decoders as θ^s (Task 1) and θ^d (Task 2), respectively. Then the loss in the sparse branch can be denoted as $\mathcal{L}^s(\theta^{sh},\theta^s) = w_{s1}\mathcal{L}_{attn} + w_{s2}\mathcal{L}_{ap} + w_{s3}\mathcal{L}_{cosim} + w_{s4}\mathcal{L}_{peaky} + w_{s5}\mathcal{L}_{guide}$, while that in the dense branch is $\mathcal{L}^d(\theta^{sh},\theta^d) = w_{d1}\mathcal{L}_{conv1} + w_{d2}\mathcal{L}_{conv2} + w_{d3}\mathcal{L}_{mc}$. These losses are summarized in Appendixes B and C.

Generally, the network can be learned by optimizing a weighted combination, *i.e.*, $\alpha\mathcal{L}^s + \beta\mathcal{L}^d$, where α and β are weights to balance the emphasis between the two tasks. However, it is difficult to determine the optimal weights manually as the following interferences exist in the training of SM and DA features.

First, SM features need larger receptive fields than DA features. Because in SM, each pixel in one image needs to find a match among all pixels in the second one. Instead, in DA, each pixel in one image only needs to find a match within a local window centered at the initial matched point identified by SM in the other image.

Second, feature-metric residuals in SM and DA are different. As shown in Fig. 4 , two images are totally aligned when $\Delta t_x = \Delta t_x = 0$. In this case, the sparse branch exhibits smaller feature residuals than the dense branch. When $(\Delta t_x, \Delta t_y)$ deviates from the origin, the feature residuals change rapidly for the sparse-only case, while the dense-only case changes smoothly. This can be attributed to the fact that the sparse branch is trained with contrastive learning ($\mathcal{L}_{AP}$ in Eq. (7)), where negative samples are found within a small radius from the positive samples. This ensures the accuracy of SM. In contrast, features emanating from the dense branch are subject to broader constraints due to the wide convergence required by $\mathcal{L}_{conv1}$ and $\mathcal{L}_{conv2}$.

To avoid negative transfer (Standley et al., 2020) caused by the competing training objectives analyzed above, we use Multiple Gradient Descent Algorithm - Upper Bound (MGDA-UB) (Sener & Koltun, 2018) to find a descent direction that improves both tasks, resulting in a solution θ that can achieve a trade-off between the two feature spaces. Concretely, the task-specific parameters θ^s and θ^d are up-

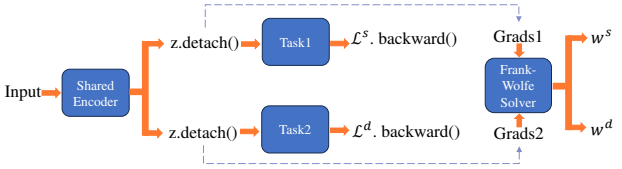

Figure 5. The process of how weights of different tasks are dynamically modulated to update the shared parameters θ^{sh}.

dated by $\theta^s - \eta\nabla_{\theta^s}\mathcal{L}^s(\theta^{sh}, \theta^s)$ and $\theta^d - \eta\nabla_{\theta^d}\mathcal{L}^d(\theta^{sh}, \theta^d)$, respectively, with η being the learning rate. Regarding θ^{sh}, its procedure of updating is shown in Fig. 5. Given an input, we pass it through the shared encoder and regard the resulting output (with gradients detached) as the representation z. Then z is fed into task-specific branches, giving rise to $\mathcal{L}^s$ and $\mathcal{L}^d$, respectively, which are then back-propagated to compute $\nabla_z\mathcal{L}^s(\theta^{sh}, \theta^s)$ and $\nabla_z\mathcal{L}^d(\theta^{sh}, \theta^d)$. These two gradients are subsequently employed in the Frank-Wolfe solver (Jaggi, 2013) to determine w^s and w^d. Then θ^{sh} is updated by $\theta^{sh} - \eta w^s\nabla_{\theta^{sh}}\mathcal{L}^s(\theta^{sh}, \theta^s) - \eta w^d\nabla_{\theta^{sh}}\mathcal{L}^d(\theta^{sh}, \theta^d)$.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct experiments on four datasets, including MSCOCO (Lin et al., 2014), GoogleEarth (Zhao et al., 2021), VIS-NIR (Brown & Süsstrunk, 2011) and VIS-IR-drone (Sun et al., 2022). Specifically, GoogleEarth displays different traffic and plant conditions due to the time of the satellite image capture. VIS-NIR involves RGB and near-infrared images while VIS-IR-drone involves RGB and infrared images, the modality differences of which mainly occur in color representations. All these three datasets provide aligned multimodal image pairs. Regarding MSCOCO, it is a widely used unimodal dataset and we generate aligned image pairs by naive replication.

We refer to (Chang et al., 2017; Zhao et al., 2021) to generate training and test data by randomly sampling a homography between an aligned image pair. Concretely, we first resize an aligned image pair to 192×192 and regard one of it as the input image. Then on the other image, we randomly choose four points in four 64×64 boxes at the corner and warp the chosen area to 128×128 as the template image. We show some examples in Fig. 6 and the green polygons denote where the four corners of the template should be on the input. Overall, we split the training and test sets of MSCOCO and GoogleEarth according to (Zhao et al., 2021), which results in around 30000/6000 and 8000/1000 training/test samples. Regarding VIS-NIR and VIR-IR-drone, we first partition the aligned image pairs in a 7:3 ratio for training and test data generation. Then random sampling is performed 5 times, which results in around 7600/1300 and

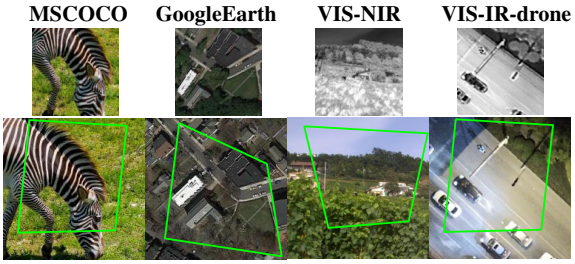

Figure 6. Top: template image. Bottom: input image.

8000/1600 training/test samples, respectively. Besides, to test the ability to cope with large deformation, we generate two subsets consisting of approximately 600 images from MSCOCO and VIS-NIR and denote them as MSCOCO* and VIS-NIR*, respectively.

Training. All the experiments are conducted on a single TITAN RTX. We train with AdamW with an initial learning rate of $1e-4$, a weight decay of $5e-4$, and a batch size of 8 image pairs. We train 100 epochs and use CosineAnnealing to schedule the learning rate. In the sparse branch, we calculate $\mathcal{L}_{AP}$ by sampling positive samples within a radius of 3 pixels and negative ones between a radius of 5 and 7 pixels. The patch size is set to 16 to calculate L_{cosim} and L_{peaky}. The total loss in this branch is set to $\mathcal{L}_{AP} + 5\mathcal{L}_{cosim} + \mathcal{L}_{peaky} + 0.008\mathcal{L}_{guide}$, while that in the dense branch is $\mathcal{L}_{conv1} + \mathcal{L}_{conv2} + \mathcal{L}_{mc}$. During training, the multi-objective optimization algorithm MGDA-UB is used to update the parameters of the network.

Inference. Given a pair of unaligned images I_T, I_I, we use our feature extractor to extract their heatmaps $\mathbf{S}_T, \mathbf{S}_I$ and descriptors $\mathbf{D}_T, \mathbf{D}_I$ in the sparse branch, along with the feature maps $\mathbf{X}_T, \mathbf{X}_I$ in the dense branch, all in one forward propagation. Then for each image, we find local maxima in $\mathbf{S}$ as keypoints and gather the corresponding descriptors from $\mathbf{D}$ and regard a keypoint along with a descriptor constitute a feature. A shortlist of the best 1000 features are kept in terms of the score ranking in $\mathbf{S}$. Our SM starts by building a putative set based on the descriptor similarity with a mutually nearest neighbor (MNN) standard, followed by MAGSAC++ (Barath et al., 2020) with the reprojection threshold set to 1 pixel and the maximum iteration number to 10K. Then the output of SM is regarded as the input of DA, where the maximum iteration number is set to 15.

Evaluation Metric. We use the same evaluation metrics as in recent works. Pixel Error (PE) is the average L_2 distance between the 4 ground-truth perturbation points and the 4 output point location predictions from an algorithm.

4.2. State-of-the-art Comparison

We classify the state-of-the-art methods into **deep-based** (DHN (DeTone et al., 2016), MHN (Le et al., 2020),

Table 1. Comparative results on MSCOCO, MSCOCO*, VIS-NIR, VIS-NIR*, GoogleEarth and VIS-IR-drone. PE<1 is the percentage of testing image pairs that satisfy PE<1, and so on. APE means the Average PE of all testing image pairs. Success rate (SR) is the percentage of the testing image pairs whose predicted PEs are smaller than the initial ones. PE< 0.5, ..., 20 and APE are only calculated on the success cases. **Bold** means first and underline means second.

Dataset	Method		PE<0.5 ↑	PE<1 ↑	PE<3 ↑	PE<5 ↑	PE<10 ↑	PE<20 ↑	APE ↓	SR ↑
MSCOCO&MSCOCO*	Sparse	SIFT	62.79 / 0.22	88.39 / 3.79	98.26 / 40.53	99.29 / 63.03	99.65 / 82.63	99.95 / 93.54	0.63 / 6.36	99.69 / 74.96
		R2D2	8.96 / 0.00	45.36 / 6.93	92.72 / 81.25	97.98 / 93.07	99.67 / 96.79	99.91 / 97.80	1.42 / 2.96	99.75 / 98.83
		SFD2	0.13 / 0.00	4.95 / 0.00	57.74 / 13.32	82.77 / 47.39	96.69 / 86.85	99.67 / 97.30	3.42 / 6.57	99.59 / 99.00
		ReDFeat	82.82 / 57.19	97.13 / 92.14	**99.71 / 98.32**	**99.81 / 98.99**	**99.95 / 99.85**	**100.00 / 100.00**	0.38 / 0.63	99.65 / 99.58
	Dense	DeepLK	82.59 / 79.76	89.92 / 90.31	93.28 / 95.07	94.45 / 95.92	96.55 / 97.11	99.09 / 98.81	1.02 / 1.02	96.67 / 98.16
		PRISE	**95.50 / 92.39**	**97.71 / 96.11**	98.72 / 97.12	98.97 / 97.29	99.41 / 97.63	99.79 / 99.49	**0.28 / 0.56**	99.18 / 98.83
	Deep	DHN	0.00 / 0.00	0.00 / 0.00	1.25 / 0.00	12.34 / 0.00	66.21 / 6.01	98.62 / 87.98	8.99 / 15.50	99.90 / 99.83
		MHN	32.10 / 0.50	70.14 / 5.84	92.79 / 36.39	96.06 / 64.61	98.44 / 90.32	99.73 / 98.33	1.27 / 5.01	**99.97 / 100.00**
		IHN	93.31 / 89.64	97.14 / 95.65	98.26 / 97.49	98.76 / 97.66	99.45 / 98.49	99.84 / 99.16	0.37 / 0.65	99.93 / **100.00**
	Ours		70.85 / 63.42	91.44 / 90.60	98.67 / 98.15	99.38 / 98.82	99.78 / 99.32	99.93 / 99.32	0.53 / 0.80	99.71 / 99.49
VIS-NIR&VIS-NIR*	Sparse	SIFT	**23.31** / 0.00	54.38 / 5.39	86.86 / 31.95	93.14 / 45.23	97.19 / 66.39	99.58 / 82.99	1.77 / 10.39	93.22 / 38.56
		R2D2	3.27 / 0.00	25.84 / 1.51	81.57 / 55.56	94.14 / 79.62	98.04 / 93.26	99.76 / 96.63	2.17 / 4.49	98.69 / 95.04
		SFD2	0.00 / 0.00	1.74 / 0.00	37.09 / 7.11	63.22 24.40	91.81 / 68.98	98.96 / 89.93	4.91 / 9.97	96.99 / 94.40
		ReDFeat	9.04 / 6.57	48.65 / 44.71	88.34 / 85.51	95.27 / 93.39	97.79 / 95.35	99.06 / 97.65	1.78 / 2.33	99.84 / 94.53
	Dense	DeepLK	3.17 / 0.28	16.53 / 14.05	56.73 / 64.19	66.81 / 79.34	75.19 / 88.71	91.05 / 93.66	6.39 / 4.97	68.02 / 58.08
		PRISE	0.92 / 0.17	10.22 / 5.34	54.93 / 44.38	68.75 / 58.46	77.46 / 65.77	91.17 / 77.71	6.25 / 10.0	66.28 / 89.76
	Deep	DHN	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.53 / 0.00	13.41 / 1.60	64.79 / 52.48	17.83 / 20.41	80.82 / **100.00**
		MHN	0.00 / 0.00	0.07 / 0.16	3.85 / 2.40	8.63 / 6.88	25.12 / 24.96	79.98 / 66.72	13.33 / 16.65	87.33 / 99.52
		IHN	12.09 / 7.68	51.15 / 46.24	91.91 / 91.68	97.38 / 95.72	99.53 / 97.68	99.92 / **99.84**	1.39 / 2.19	99.84 / 99.26
	Ours		21.49 / **12.42**	**60.78 / 51.45**	**93.14 / 93.87**	**97.99 / 95.96**	**99.92 / 98.54**	**100.00** / 99.19	**1.21 / 1.58**	**100.00** / 99.20
GoogleEarth	Sparse	SIFT	1.73	17.40	72.38	86.45	95.48	99.73	2.84	88.59
		R2D2	0.12	10.68	75.17	92.51	98.81	99.76	2.50	99.05
		SFD2	0.12	0.48	31.07	65.41	91.77	99.03	4.93	97.29
		ReDFeat	7.61	44.86	92.78	96.69	99.05	99.51	1.82	99.52
	Dense	DeepLK	0.37	12.55	73.43	88.31	94.59	99.14	3.04	95.65
		PRISE	8.5	45.55	85.50	90.13	93.42	98.41	2.49	96.58
	Deep	DHN	0.00	0.00	0.00	0.61	21.05	92.41	13.45	96.12
		MHN	0.00	0.00	0.60	8.29	57.57	96.75	10.05	97.88
		IHN	**10.23**	**55.05**	92.58	97.00	99.05	99.73	1.82	**100.00**
	Ours		8.00	49.52	**93.05**	**97.64**	**99.53**	**99.88**	**1.38**	**100.00**
VIS-IR-drone	Sparse	SIFT	0.00	5.01	47.19	66.67	86.62	97.37	5.14	49.97
		R2D2	0.00	1.63	44.31	69.08	92.54	98.95	4.54	91.34
		SFD2	0.00	0.06	13.32	43.55	84.69	98.32	6.57	96.35
		ReDFeat	0.00	9.19	66.68	83.40	97.75	99.75	3.02	98.20
	Dense	DeepLK	0.00	3.19	54.06	77.88	91.44	97.75	4.31	95.52
		PRISE	0.06	4.27	56.21	78.01	91.08	98.24	4.23	95.04
	Deep	DHN	0.00	0.00	0.00	0.09	7.57	75.44	16.63	67.82
		MHN	0.00	0.00	0.00	1.48	31.03	92.91	12.60	96.78
		IHN	0.12	5.97	66.12	84.66	98.28	99.88	3.00	99.64
	Ours		**0.17**	**10.34**	**68.49**	**87.87**	**99.58**	**100.00**	**2.75**	**99.88**

Table 2. Runtime and model size analysis of each method. All methods are tested on VIS-IR-drone. The runtime of sparse methods and ours (SM) calculates the total runtime of feature extraction + MNN + MAGSAC++.

	Sparse				Dense		Deep			Ours		
	SIFT	R2D2	SFD2	ReDFeat	DeepLK	PRISE	DHN	MHN	IHN	SM	DA	Total
Parameters (M)	-	0.49	4.04	1.13	33.94	33.94	10.99	32.97	1.71	-	-	1.07
Runtime (ms)	20.19	50.56	72.86	69.26	191.58	204.67	11.09	35.91	44.94	29.72	34.64	66.37

IHN (Cao et al., 2022)), **sparse-based** (SIFT (Lowe, 2004), R2D2 (Revaud et al., 2019), SFD2 (Xue et al., 2023), ReDFeat (Deng & Ma, 2023)) and **dense-based** (DeepLK (Zhao et al., 2021), PRISE (Zhang et al., 2023)) ones for comparison. Deep-based ones are those predicting homography matrices or four corner points via networks directly. Sparse-based approaches follow the pipeline of feature extraction + MNN + MAGSAC++, with the keypoint number and MAGSAC++ details consistent with ours. Dense-based methods optimize homography matrices via DA. We use the pretrained models of R2D2[1] and SFD2[2] and fine-tune them on our datasets for evaluation. Other methods are trained from scratch for evaluation. Following the original papers, DeepLK and PRISE are initialized by MHN.

Comparative results are reported in Table 1 and Table 2

[1] https://github.com/naver/r2d2/blob/master/models/r2d2_WAF_N16.pt

[2] https://github.com/feixue94/sfd2

Table 3. Comparison with state-of-the-art matchers. We report results of APE↓ (SR%↑), best in bold. * in the first column means we use the corresponding dataset to fine-tune the pre-trained model.

	MSCOCO	MSCOCO*	VIS-NIR	VIS-NIR*	GoogleEarth	VIS-IR-drone
SuperPoint+SuperGlue	1.31 (**99.87%**)	2.06 (99.83%)	1.26 (99.84%)	2.71 (**99.68%**)	1.77 (99.52%)	2.84 (99.40%)
RedFeat+SuperGlue*	0.68 (99.68%)	6.75 (**100%**)	1.69 (99.84%)	2.05 (99.33%)	2.04 (**100%**)	3.63 (**100%**)
LoFTR*	1.16 (97.83%)	4.38 (96.82%)	3.36 (94.60%)	7.15 (94.08%)	3.68 (97.52)	11.39 (53.55%)
Ours	**0.53** (99.71%)	**0.80** (99.49%)	**1.21** (**100%**)	**1.58** (99.20%)	**1.38** (**100%**)	**2.75** (99.88%)

Table 4. Ablation study. We use $\alpha\mathcal{L}^s+\beta\mathcal{L}^d$ with $\alpha=\beta=0.5$ to optimize network when training jointly w/o MGDA-UB.

Model	Description	APE ↓	
		GoogleEarth	VIS-IR-drone
A	Baseline	2.56	3.45
B	+ MITB	2.45	3.37
C	+ MGDA-UB	1.78	3.03
D	+ Mutual guidance	**1.38**	**2.75**

Table 5. APEs *w.r.t.* different weight combinations of the sparse and dense branches when MITB, MGDA-UB and mutual guidance are not involved. DA results of the same dataset are initialized by the same SM results.

Weight setting		APE ↓			
		VIS-IR-drone		GoogleEarth	
α (sparse)	β (dense)	SM	DA	SM	DA
0.1	0.9	5.74	**2.76**	4.26	**1.92**
0.5	0.5	4.28	2.87	3.92	2.01
0.9	0.1	**3.50**	2.92	**3.22**	2.14

and visualization results are provided in *Appendix* D. Overall, our method achieves comparable performance to the state-of-the-art on the MSCOCO dataset, while consistently outperforming them across the remaining three multimodal datasets, and showcasing robustness against substantial deformations. More specifically, we observe that: i) Deep-based methods including DHN and MHN lack accuracy due to the neglect of the geometric constraint. Although IHN performs comparable with our method in most cases, its performance drops greatly when the deformation becomes large while ours changes little. ii) Our novel S2D pipeline yields lower APEs, *i.e.*, higher accuracy, than the sparse-only and dense-only methods in most cases. This demonstrates the superiority of our S2D pipeline, *i.e.*, SM serves as a robust initialization for DA, with DA subsequently refining the outcomes from SM. iii) Our model comprises 1.07M parameters and the S2D pipeline operates within 100 ms, which strikes a balance between accuracy and efficiency.

We also compare with state-of-the-art matchers including SuperPoint (DeTone et al., 2018)+SuperGlue (Sarlin et al., 2020), RedFeat+SuperGlue and LoFTR (Sun et al., 2021), and results are shown in Table 3. We use the official outdoor model for SuperPoint+SuperGlue[3], and fine-tune SuperGlue with our trained RedFeat for RedFeat+SuperGlue*. As the official outdoor model of LoFTR[4] nearly fails on our datasets, we only report the results of the fine-tuned model. It can be seen that although our method only involves the simplest matcher (*i.e.*, MNN), it still outperforms the sophisticated GNN- or Transformer-based matchers. This occurs as matchers merely offer an initial setup with limited accuracy, and the substantial enhancement in accuracy is derived from DA in our case.

[3] https://github.com/magicleap/SuperGluePretrainedNetwork

[4] https://github.com/zju3dv/LoFTR

4.3. Ablation Study on GoogleEarth and VIS-IR-drone

We ablate the proposed components (*i.e.*, MITB, MGDA-UB and mutual guidance) in this work to show their effectiveness. Results are reported in Table 4.

MITB. We observed that MITB can bring a slight performance improvement since it enhances descriptors with a broader global receptive field.

MGDA-UB. Compared model B with model C in Table 4, we find that MGDA-UB can improve the performance by 27% and 10% on GoogleEarth and VIS-IR-drone, respectively. To further demonstrate the effectiveness of MGDA-UB, we additionally train three models without MITB, MGDA-UB and Mutual guidance, and set α and β to $(0.1, 0.9)$, $(0.5, 0.5)$ and $(0.9, 0.1)$, respectively, to test different weight combinations. Results can be found in Table 5. Note that for DA, we use the same initialization for the three settings in order to only test the performance of the dense branch. It can be seen that larger weight of one branch will impair the performance of the other branch, and it is an expensive operation to search the optimal combination. MGDA-UB addresses this issue by dynamically determining the weights according to the gradient information and learning a trade-off between the two conflict tasks.

Mutual Guidance. Table 4 shows that mutual guidance can also contribute to the performance and we intend to explore how SM and DA are exactly guided. We employ models C and D, as outlined in Table 4, for evaluation. We measure the features predicting by our sparse branch in terms of Mean Matching Accuracy (MMA), *i.e.*, the ratio of matches with a reprojection error below a threshold, from 1 to 10 pixels, and averaged across all image pairs. Results are reported

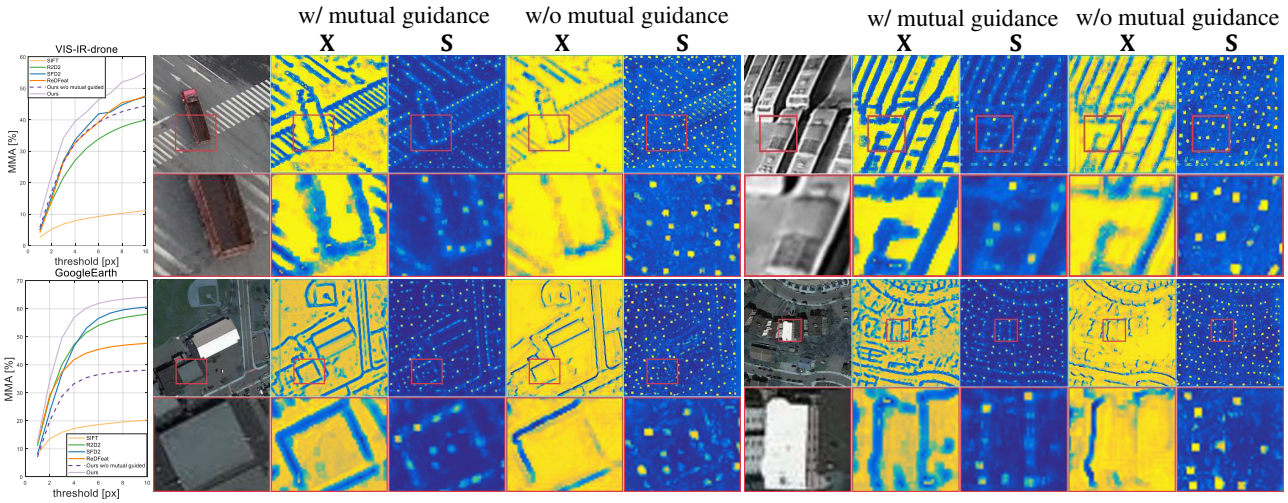

Figure 7. Left: MMA *w.r.t.* 10 pixel thresholds. Right: Visualization of single-channel feature maps in dense branch (**X**) and heatmaps in sparse branch (**S**) of two images from VIS-IR-drone (top two) and GoogleEarth (bottom two), when mutual guidance is or is not adopted.

Table 6. Performance of DA w/ and w/o SM guidance.

Method	DA APE ↓	
	VIS-IR-drone	GoogleEarth
w/ mutual guidance	**2.75**	**1.38**
w/o mutual guidance	2.96	1.52

Table 7. Cross-dataset results of APE↓ (SR%↑). Best in bold.

Train \ Test	MSCOCO	VIS-NIR	GoogleEarth	VIS-IR-drone
MSCOCO	**0.53 (99.71%)**	1.86 (94.60%)	1.87 (98.94%)	4.06 (88%)
VIS-NIR	0.87 (99.51%)	**1.21 (100%)**	2.04 (98.11%)	3.28 (98.62%)
GoogleEarth	1.23 (99.15%)	2.67 (92.21%)	**1.38 (100%)**	4.34 (89.61%)
VIS-IR-drone	1.40 (99.08%)	2.23 (97.14%)	2.46 (97.88%)	**2.75 (99.88%)**

in the left column in Fig. 7. The introduction of the mutual guidance strategy leads to a significant performance boost, with our features achieving the highest level of performance. As shown in the red boxes in Fig. 7, this phenomenon can be attributed to the modality-invariant structures highlighted by the dense branch, which effectively guide the detection of keypoints across these structures. As a result, the overall robustness of the features is greatly enhanced. Regarding DA, we use the same initialization and the results when the mutual guidance is adopted or not are reported in Table 6. It is evident that incorporating information from the sparse branch into the DA objective function steers the optimization towards more accurate outcomes.

5. Generalization Analysis

We first conduct cross-dataset evaluation of our method, with results listed in Table 7. It is shown that our method generalizes well to unseen datasets. Then we perform fine-tuning with small datasets and results are shown in Table 8. Specifically, we start from the pre-trained model of MSCOCO and fine-tune it using only 10% of the original training data from other datasets. It shows performance is improved after fine-tuning on small datasets. This further shows the practicability of our method, since in practice single-modal datasets are easier to acquire than the multi-modal ones. So in practice, we can first pre-train our model on a large-scale single-modal dataset and then fine-tune it on the small multi-modal dataset.

Table 8. Fine-tuning results of APE↓ (SR%↑). Pre-trained refers to the model trained on MSCOCO.

	VIS-NIR	GoogleEarth	VIS-IR-drone
Pre-trained	1.86 (94.60%)	1.87 (98.94%)	4.06 (88%)
Fine-tune	1.35 (99.61%)	1.49 (100%)	3.18 (99.64%)
Ours	1.21 (100%)	1.38 (100%)	2.75 (99.88%)

6. Conclusion

In this work, we propose to combine the advantages of sparse matching (SM) and direct alignment (DA) to perform multimodal image registration in a sparse-to-dense (S2D) manner. For this end, we design a multi-task network termed as sparse-to-dense multimodal feature extractor (SDME) to predict features for SM and DA. During training, we employ the Multiple Gradient Descent Algorithm (MGDA) to strike a balance between tasks, while introducing mutual guidance to enhance interaction between them. Experiments validate the effectiveness of the mutual guidance and MGDA strategies. Our S2D strategy based on SDME outperforms state-of-the-art methods, demonstrating superior efficiency and maintaining high performance even in scenes with extremely large deformation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant no. 62276192.

Impact Statement

This work reveals the relationship between existing multimodal image registration paradigms, meanwhile proposing a new paradigm for this field. This may give inspiration to the following researchers.

References

Audebert, N., Le Saux, B., and Lefèvre, S. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140:20–32, 2018.

Baker, S. and Matthews, I. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56:221–255, 2004.

Bansal, N., Ji, P., Yuan, J., and Xu, Y. Semantics-depth-symbiosis: Deeply coupled semi-supervised learning of semantics and depth. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5828–5839, 2023.

Barath, D., Noskova, J., Ivashechkin, M., and Matas, J. Magsac++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1304–1312, 2020.

Brown, M. and Süsstrunk, S. Multi-spectral sift for scene category recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 177–184, 2011.

Cadena, C., Dick, A. R., and Reid, I. D. Multi-modal auto-encoders as joint estimators for robotics scene understanding. In *Proceedings of the Robotics: Science and Systems*, pp. 1–9, 2016.

Cao, S.-Y., Hu, J., Sheng, Z., and Shen, H.-L. Iterative deep homography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1879–1888, 2022.

Chang, C.-H., Chou, C.-N., and Chang, E. Y. Clkn: Cascaded lucas-kanade networks for image alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2213–2221, 2017.

Chum, O., Matas, J., and Kittler, J. Locally optimized ransac. In *Proceedings of the Joint Pattern Recognition Symposium*, pp. 236–243, 2003.

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., and Marchal, G. Automated multi-modality image registration based on information theory. In *Proceedings of the Information Processing in Medical Imaging*, pp. 263–274, 1995.

Deng, Y. and Ma, J. Redfeat: Recoupling detection and description for multimodal feature learning. *IEEE Transactions on Image Processing*, 32:591–602, 2023.

DeTone, D., Malisiewicz, T., and Rabinovich, A. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

DeTone, D., Malisiewicz, T., and Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., and Sattler, T. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8092–8101, 2019.

Engel, J., Schöps, T., and Cremers, D. Lsd-slam: Large-scale direct monocular slam. In *Proceedings of the European Conference on Computer Vision*, pp. 834–849, 2014.

Fischler, M. A. and 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.

Goforth, H. and Lucey, S. Gps-denied uav localization using pre-existing satellite imagery. In *Proceedings of the International Conference on Robotics and Automation*, pp. 2974–2980, 2019.

Gómez-Chova, L., Tuia, D., Moser, G., and Camps-Valls, G. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, 2015.

Guo, P., Lee, C.-Y., and Ulbricht, D. Learning to branch for multi-task learning. In *Proceedings of the International Conference on Machine Learning*, pp. 3854–3863, 2020.

He, K., Lu, Y., and Sclaroff, S. Local descriptors optimized for average precision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 596–605, 2018.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, pp. 427–435, 2013.

Le, H., Liu, F., Zhang, S., and Agarwala, A. Deep homography estimation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7652–7661, 2020.

Li, J., Hu, Q., and Ai, M. Rift: Multi-modal image matching based on radiation-invariant feature transform. *arXiv preprint arXiv:1804.09493*, 2018.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60: 91–110, 2004.

Lu, X., Wang, X., and Fan, J. E. Learning dense correspondences between photos and sketches. In *Proceedings of the International Conference on Machine Learning*, pp. 22899–22916, 2023.

Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T., and Quan, L. Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6589–6598, 2020.

Ma, J., Jiang, X., Fan, A., Jiang, J., and Yan, J. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021.

Nguyen, A., Nguyen, N., Tran, K., Tjiputra, E., and Tran, Q. D. Autonomous navigation in complex environments with deep multimodal fusion network. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5824–5830, 2020.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, 1999.

Revaud, J., De Souza, C., Humenberger, M., and Weinzaepfel, P. R2d2: Reliable and repeatable detector and descriptor. *Advances in Neural Information Processing Systems*, 32, 2019.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. Orb: An efficient alternative to sift or surf. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2564–2571, 2011.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., and Rabinovich, A. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020.

Sarlin, P.-E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Hammarstrand, L., Kahl, F., et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3247–3257, 2021.

Schonberger, J. L. and Frahm, J.-M. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

Sener, O. and Koltun, V. Multi-task learning as multi-objective optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., and Savarese, S. Which tasks should be learned together in multi-task learning? In *Proceedings of the International Conference on Machine Learning*, pp. 9120–9132, 2020.

Sun, J., Shen, Z., Wang, Y., Bao, H., and Zhou, X. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021.

Sun, Y., Cao, B., Zhu, P., and Hu, Q. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022.

Tang, C. and Tan, P. Ba-net: Dense bundle adjustment networks. In *Proceedings of the International Conference on Learning Representations*, pp. 1–11, 2018.

Wang, Y., Chen, Y., Jamieson, K., and Du, S. S. Improved active multi-task representation learning via lasso. In *Proceedings of the International Conference on Machine Learning*, pp. 35548–35578, 2023.

Weinzaepfel, P., Lucas, T., Larlus, D., and Kalantidis, Y. Learning super-features for image retrieval. In *Proceedings of the International Conference on Learning Representations*, pp. 1–12, 2022.

Xiang, Y., Wang, F., and You, H. Os-sift: A robust sift-like algorithm for high-resolution optical-to-sar image registration in suburban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6):3078–3090, 2018.

Xing, B., Ying, X., Wang, R., Yang, J., and Chen, T. Cross-modal contrastive learning for domain adaptation in 3d semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2974–2982, 2023.

Xu, H., Yuan, J., and Ma, J. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (10):12148–12166, 2023.

Xue, F., Budvytis, I., and Cipolla, R. Sfd2: Semantic-guided feature detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5206–5216, 2023.

Zhang, Y., Huang, X., and Zhang, Z. Prise: Demystifying deep lucas-kanade with strongly star-convex constraints for multimodel image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13187–13197, 2023.

Zhao, Y., Huang, X., and Zhang, Z. Deep lucas-kanade homography for multimodal image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15950–15959, 2021.

A. Inverse Compositional Lucas-Kanade (IC-LK)

During direct alignment (DA), we aim to minimize the feature-metric residual between two unaligned images to refine the homography between them. As illustrated in the original paper, let $\mathbf{p} \in \mathbb{R}^8$ denote the homography of 8-DoF, and $\mathbf{X}_T, \mathbf{X}_I \in \mathbb{R}^{H\times W\times C}$ denote the features of $I_T, I_I \in \mathbb{R}^{H\times W}$, then the highly non-linear objective that we aim to optimize during DA is given by:

$$\min_{\mathbf{p}} E(\mathbf{p}) = \sum_i \left\|\mathbf{X}_T[i] - \mathbf{X}_I[W(i;\mathbf{p})]\right\|_2^2, \tag{A1}$$

where i is a pixel in I_T, $W(\cdot;\mathbf{p}) : \mathbf{R}^2 \to \mathbf{R}^2$ is the warp function parametrized by $\mathbf{p}$, and $[\cdot]$ is a lookup with sub-pixel interpolation. The Lucas-Kanade algorithm iteratively solves for the warp parameters $\mathbf{p}_{k+1} = \mathbf{p}_k + \Delta\mathbf{p}$. At every iteration k, the warp increment $\Delta\mathbf{p}$ is obtained by linearizing:

$$\min_{\Delta\mathbf{p}} \sum_i \left\|\mathbf{X}_T[i] - \mathbf{X}_I[W(i;\mathbf{p}_k + \Delta\mathbf{p})]\right\|_2^2, \tag{A2}$$

with the first order Taylor expansion being:

$$\min_{\Delta\mathbf{p}} \sum_i \left\|\mathbf{X}_T[i] - \mathbf{X}_I[W(i;\mathbf{p}_k)] - \frac{\partial \mathbf{X}_I[W(i;\mathbf{p}_k)]}{\partial \mathbf{p}}\Delta\mathbf{p}\right\|_2^2. \tag{A3}$$

Since $\mathbf{X}_I[W(i;\mathbf{p}_k)]$ changes with $\mathbf{p}_k$, $\frac{\partial \mathbf{X}_I[W(i;\mathbf{p}_k)]}{\partial \mathbf{p}}$ needs to be recomputed in each iteration. IC-LK (Baker & Matthews, 2004) addresses this issue by solving for the warp parameters $\Delta\mathbf{p}$ to make $\mathbf{p}_{k+1} = \mathbf{p}_k + (\Delta\mathbf{p})^{-1}$, *i.e.*, it applys $\Delta\mathbf{p}$ on I_T instead of I_I. Thus Eq. (A2) turns into:

$$\min_{\Delta\mathbf{p}} \sum_i \left\|\mathbf{X}_T[W(i;\Delta\mathbf{p})] - \mathbf{X}_I[W(i;\mathbf{p}_k)]\right\|_2^2. \tag{A4}$$

Correspondingly, the Taylor expansion in Eq. (A3) turns into:

$$\min_{\Delta\mathbf{p}} \sum_i \left\|\mathbf{X}_T[i] - \frac{\partial \mathbf{X}_T[W(i;\mathbf{0})]}{\partial \mathbf{p}}\Delta\mathbf{p} - \mathbf{X}_I[W(i;\mathbf{p}_k)]\right\|_2^2. \tag{A5}$$

In this case, $\frac{\partial \mathbf{X}_T[W(i;\mathbf{0})]}{\partial \mathbf{p}}$ is independent on $\mathbf{p}_k$ and can be pre-computed, resulting in a more efficient algorithm.

Afterwards, we expand the contents within $\|\cdot\|_2^2$ in Eq. (A5) and calculate the derivative of the unfolded equation *w.r.t.* $\Delta\mathbf{p}$. Making the derivative equal to 0, we acquire

$$\begin{aligned}
\Delta\mathbf{p} &= \Big(\sum_i \mathbf{J}_i^T\mathbf{J}_i\Big)^{-1}\sum_i \mathbf{J}_i^T\mathbf{r}_i, \\
\mathbf{J}_i &= \frac{\partial \mathbf{X}_T[W(i;\mathbf{0})]}{\partial \mathbf{p}} = \frac{\partial \mathbf{X}_T[W(i;\mathbf{0})]}{\partial i}\frac{\partial i}{\partial \mathbf{p}} \in \mathbb{R}^{C\times 8}, \\
\mathbf{r}_i &= \mathbf{X}_T[i] - \mathbf{X}_I[W(i;\mathbf{p}_k)] \in \mathbb{R}^C.
\end{aligned} \tag{A6}$$

B. Loss in The Sparse Branch

The total loss in this branch is set to $\mathcal{L}_{AP} + 5\mathcal{L}_{cosim} + \mathcal{L}_{peaky} + 0.008\mathcal{L}_{guide}$, where $\mathcal{L}_{AP}$ in Eq. (7) is a metric loss for descriptor learning, $\mathcal{L}_{guide}$ in Eq. (9) aims to guide keypoints towards distribution over modality-invariant structures. $\mathcal{L}_{cosim}$ and $\mathcal{L}_{peaky}$ are referred to R2D2 (Revaud et al., 2019) and we introduce them in the following.

$\mathcal{L}_{cosim}$ facilitates the detection of highly repeatable keypoints, thereby ensuring their positions remain invariant to common natural image transformations, including changes in viewpoint and illumination. We denote the heatmaps of I_I and I_T as $\mathbf{S}_I$ and $\mathbf{S}_T$, respectively. Then $\mathbf{S}_T$ is transformed to $\mathbf{S}_T^{'}$ based on the known $\mathbf{H}_{gt}$. A set of $N \times N$ overlapping patches $\mathcal{P} = \{p\}$ can be obtained by traversing all pixel locations of $\mathbf{S}_T^{'}$. Since $\mathbf{S}_T^{'}$ aligns with $\mathbf{S}_I$, keypoints with high repeatability can be learned by making all local maxima in $\mathbf{S}_T^{'}$ correspond to the ones in $\mathbf{S}_I$, *i.e.*, minimizing L_{cosim}:

$$L_{cosim} = 1 - \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} cosim\big(\mathbf{S}_T^{'}[p], \mathbf{S}_I[p]\big). \tag{A7}$$

Here $cosim(\cdot,\cdot)$ refers to the cosine similarity, $\mathbf{S}_T^{'}[p] \in \mathcal{R}^{N^2}$ denotes the flattened $N \times N$ patch p extracted from $\mathbf{S}_T^{'}$, and likewise for $\mathbf{S}_I[p]$.

$\mathcal{L}_{peaky}$ is introduced on both $\mathbf{S}_T^{'}$ and $\mathbf{S}_I$ to maximize the local peakiness of the heatmap:

$$L_{peaky} = 1 - \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\left(\max_{(i,j)\in p}\mathbf{S}_{ij} - \underset{(i,j)\in p}{mean}\,\mathbf{S}_{ij}\right). \tag{A8}$$

Implementation. we set the patch size to 16 when calculating L_{cosim} and L_{peak}.

C. Loss in The Dense Branch

The total loss in this branch is set to $\mathcal{L}_{conv1} + \mathcal{L}_{conv2} + \mathcal{L}_{mc}$, where $\mathcal{L}_{mc}$ is introduced in Eq. (8) to bridge the gap between different modalities. The terms $\mathcal{L}_{conv1}$ and $\mathcal{L}_{conv2}$ aim to shape the optimization landscape around the ground truth of Eq. (A1) for better convergence (Zhao et al., 2021).

Firstly, Eq. (A1) is expected to achieve the local minimum at ground truth $\overline{\mathbf{p}}$. For this, *Condition 1* is proposed:

$$\forall(\overline{\mathbf{p}} + \Delta\mathbf{p}) \in \Theta,\ E(\overline{\mathbf{p}} + \Delta\mathbf{p}) - E(\overline{\mathbf{p}}) \geq g(\overline{\mathbf{p}} + \Delta\mathbf{p}) - g(\overline{\mathbf{p}}), \tag{A9}$$

where Θ is a small region around $\overline{\mathbf{p}}$, the supportive convex function $g(\mathbf{p}) = \|\mathbf{p} - \overline{\mathbf{p}}\|_2^2 = \sum_{i=1}^{8}(\mathbf{p}_i - \overline{\mathbf{p}}_i)^2$ can achieve minimum at $\overline{\mathbf{p}}$. This condition guarantees that $E(\overline{\mathbf{p}})$ is the local minimum and the gradient of $E(\mathbf{p})$ is larger than that of $g(\mathbf{p})$ at $\overline{\mathbf{p}}$. Then, Eq. (A9) can be transformed to:

$$\mathcal{L}_{conv1} = -\frac{1}{M}\sum_{m=1}^{M} minimum\left(0, E(\overline{\mathbf{p}} + \Delta\mathbf{p}^m) - E(\overline{\mathbf{p}}) - \sum_{i=1}^{8}(\Delta\mathbf{p}_i^m)^2\right), \tag{A10}$$

where $\Delta\mathbf{p}^m$ is the m-th sampled random noise, $\Delta\mathbf{p}_i^m$ is the i-th variable of $\Delta\mathbf{p}^m$, and M times are sampled for each training batch.

Secondly, we hope that $E(\mathbf{p})$ in Eq. (A1) has a steep directional derivative (∇_v) in Θ, which introduces *Condition 2*:

$$\forall\mathbf{p} \in \Theta, \nabla_v E(\mathbf{p}) \geq \nabla_v g(\mathbf{p}), \tag{A11}$$

where $\nabla_v E(\mathbf{p})$ can be calculated by $E(\mathbf{p}+\Delta\mathbf{p})$ and $E(\mathbf{p}+\lambda\Delta\mathbf{p})$ with $\lambda \in (0,1)$, and likewise for $\nabla_v g(\mathbf{p})$. This condition guarantees that $E(\mathbf{p})$ has a smooth surface around $\overline{\mathbf{p}}$ without other minima. Then, Eq. (A11) can be transformed to:

$$\mathcal{L}_{conv2} = -\frac{1}{M}\sum_{m=1}^{M} minimum\left(0, E(\overline{\mathbf{p}} + \Delta\mathbf{p}^m) - E(\overline{\mathbf{p}} + \lambda\Delta\mathbf{p}^m) - (1-\lambda^2)\sum_{i=1}^{8}(\Delta\mathbf{p}_i^m)^2\right). \tag{A12}$$

Implementation. We sample 4 small perturbations, *i.e.*, $M = 4$, around $\overline{\mathbf{p}}$ to simulate the small region Θ, and λ is set by multiplying $\overline{\mathbf{p}}$ with a factor randomly sampled within $[-0.083, 0.083]$.

D. Visualization

We show some visualization results of image registration in Fig. A1 and Fig. A2. It can be observed that our method yields satisfying results in single-modal, multimodal and large deformation cases.

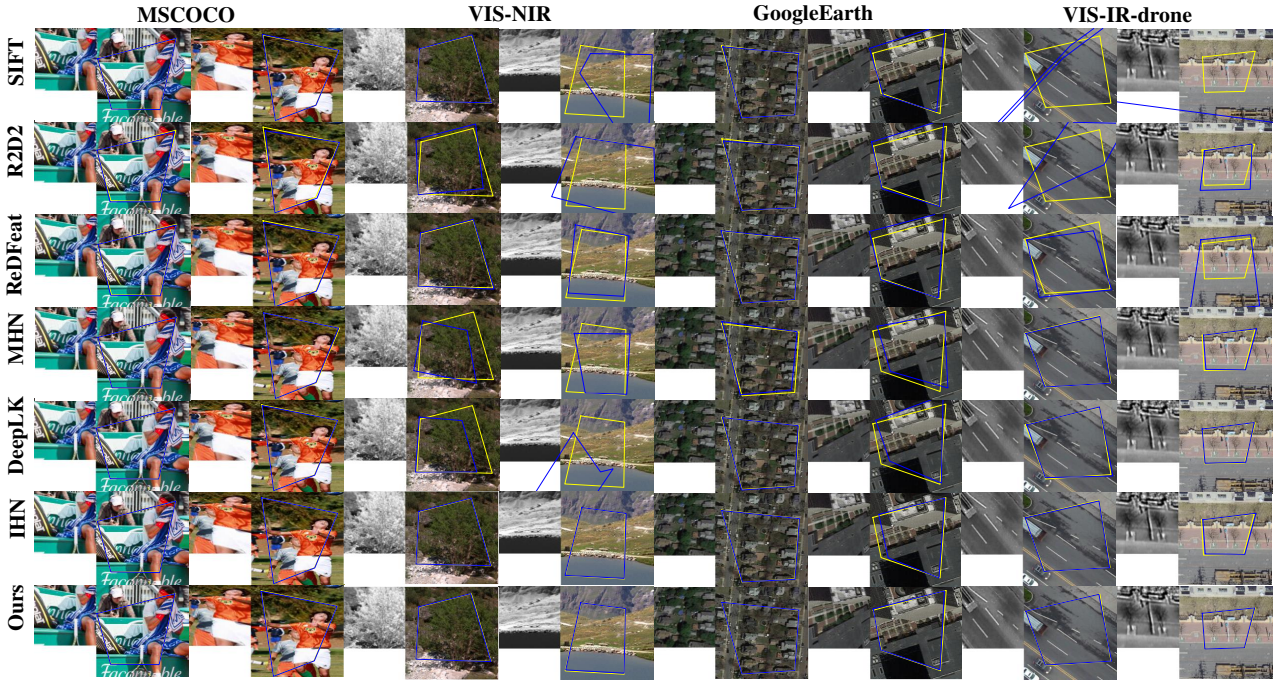

Figure A1. Visualization results of different methods on two image pairs from MSCOCO, VIS-NIR, GoogleEarth and VIS-IR-drone, respectively. Yellow: ground truth positions of the four corners of the template image (smaller one) on the input image (larger one). Blue: the predicted results.

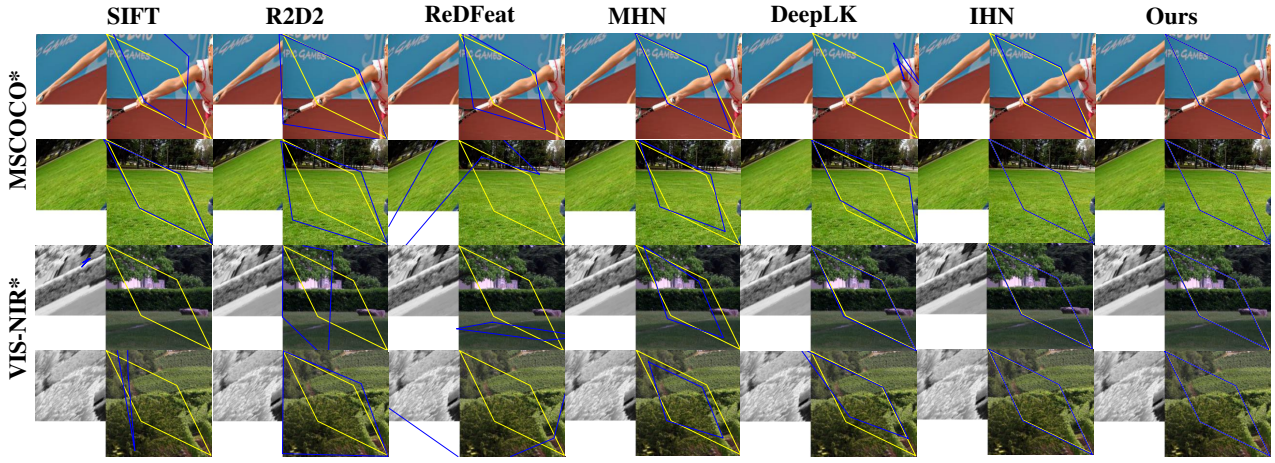

Figure A2. Visualization results of different methods on two datasets with large deformation, *i.e.*, MSCOCO* and VIS-NIR*. Yellow: ground truth positions of the four corners of the template image (smaller one) on the input image (larger one). Blue: the predicted results.