Disentangle-then-Align: Non-Iterative Hybrid Multimodal Image Registration via Cross-Scale Feature Disentanglement

Supplementary Material

1. ILDA Boundary Case Handling

For clarity, Eq. (3) in the main paper defines the ILDA attention using three neighboring scales $(i-1, i, i+1)$, which applies to all intermediate scales. To handle boundary scales (i.e., the coarsest and finest levels) in a principled way, we adopt a replicate-padding strategy instead of zero-padding or dropping one neighbor.

Concretely, let the scales be indexed by $i \in \{0, 1, \ldots, L\}$ from fine to coarse. For intermediate scales $0 < i < L$, the query-key-value triplets are constructed as:

$$
\begin{aligned}
Q_i^s &= W_q^i\, G(F_i^s), \\
K_i^s &= \big[\, W_k^{i-1} G(F_{i-1}^s),\ W_k^i G(F_i^s),\ W_k^{i+1} G(F_{i+1}^s) \,\big], \\
V_i^s &= \big[\, W_v^{i-1} G(F_{i-1}^s),\ W_v^i G(F_i^s),\ W_v^{i+1} G(F_{i+1}^s) \,\big],
\end{aligned}
$$

where $G(\cdot)$ denotes global average pooling and W_q^i, W_k^i, W_v^i are learnable projections at scale i.

For the boundary scales, we fill the missing neighbor by replicating the boundary feature itself:

$$
F_{-1}^s = F_0^s, \quad F_{L+1}^s = F_L^s,
$$

and construct K_i^s and V_i^s using the same three-scale pattern as above. In other words, at the finest scale $i = 0$, ILDA attends to $\{F_0^s, F_0^s, F_1^s\}$, and at the coarsest scale $i = L$, it attends to $\{F_{L-1}^s, F_L^s, F_L^s\}$. All subsequent attention, gating, and projection operations remain unchanged.

This replicate-padding scheme keeps the attention formulation consistent across all scales, avoids introducing zero-valued features that could distort the attention distribution, and empirically leads to more stable cross-scale gating at the boundaries.

2. Detailed Structure of HPPM

The Hybrid Parameter Prediction Module (HPPM) performs hybrid registration over five scales in a coarse-to-fine manner. At each scale $i \in \{4, 3, 2, 1, 0\}$ (from coarse to fine), HPPM predicts an incremental transformation parameter ϕ_i', accumulates it with the upsampled transformation from the previous scale, and uses the resulting transformation ϕ_i to drive both feature warping and deformation refinement. The detailed structure of HPPM is shown in Fig. 1

Per-scale pipeline. At each scale i, a Hybrid Registration Block (HRB) takes as input the shared features of the fixed and moving images at that scale, together with the fused representation from the coarser scale (except for the coarsest scale, where only the current features are used). Concretely:

- **Input composition.** For the coarsest scale $i = 4$, we concatenate the shared features $\hat{M}_4^s$ and $\hat{F}_4^s$ as the input to the HRB. For finer scales $i < 4$, we first warp the moving feature $\hat{M}_i^s$ using the accumulated transformation from the previous scale (see the ST step below), obtaining the aligned feature $\hat{M}_i^{s'}$. We then concatenate $\hat{M}_i^{s'}$, $\hat{F}_i^s$, and the fused feature f_{i+1} from the coarser scale, and feed the result into the HRB at scale i.
- **Initial convolution.** A 3×3 convolution is applied to the concatenated input to mix channel information and provide a refined representation for subsequent processing.
- **Residual State Space Blocks (RSSB).** Each HRB contains two Residual State Space Blocks (RSSB) [5], which model long-range dependencies within each scale with a state-space formulation. The output of the second RSSB is regarded as the fused representation f_i at scale i.
- **Registration heads.** On top of f_i, we employ specialized heads for rigid and non-rigid registration. At the coarsest scale $i = 4$, we only use a rigid head to estimate a low-dimensional rigid parameter vector H via global average pooling followed by two fully connected layers, and then encode H into a coarse rigid flow ϕ_4. This provides a global initialization for the subsequent refinement. At finer scales $i < 4$, we no longer estimate rigid parameters and instead focus on non-rigid refinement: a convolutional head predicts an incremental non-rigid flow ϕ_i' from the fused feature f_i.
- **Hybrid accumulation.** The hybrid deformation at scale i is obtained by accumulating the upsampled deformation from the previous (coarser) scale and the current incremental flow:

$$
\phi_i = \mathrm{upsample}(\phi_{i+1}) + \phi_i'. \tag{1}
$$

  At the coarsest scale $i = 4$, we simply set ϕ_4 to the rigid flow encoded from H. In this way, HPPM follows a $1:4$ schedule with one coarse rigid step at $i = 4$ and four non-rigid refinement steps at $i = 3, 2, 1, 0$.
  For clarity, we further summarize the inputs, outputs, and roles of all five scales in HPPM in Table 1.
- **Spatial Transformer (ST) warping.** The accumulated transformation ϕ_i is then fed into a Spatial Transformer (ST) module, which warps the moving feature at the next finer scale. Specifically, ϕ_i is upsampled to the resolution of scale $i-1$ and used to warp $\hat{M}_{i-1}^s$, yielding the aligned

Table 1. Per-scale inputs and outputs of HPPM. $\hat{M}_i^s$ / $\hat{F}_i^s$ denote shared moving / fixed features at scale i, $\hat{M}_i^{s'}$ is the warped moving feature using the accumulated flow from the previous scale.

Scale i	Role in HPPM	HRB input	Outputs at scale i
4	Global rigid initialization	$[\hat{M}_4^s,\ \hat{F}_4^s]$	f_4; rigid flow ϕ_4 encoded from H
3	Non-rigid refinement	$[\hat{M}_3^{s'},\ \hat{F}_3^s,\ f_4]$	f_3; ϕ_3'; $\phi_3 = \text{upsample}(\phi_4) + \phi_3'$
2	Non-rigid refinement	$[\hat{M}_2^{s'},\ \hat{F}_2^s,\ f_3]$	f_2; ϕ_2'; $\phi_2 = \text{upsample}(\phi_3) + \phi_2'$
1	Non-rigid refinement	$[\hat{M}_1^{s'},\ \hat{F}_1^s,\ f_2]$	f_1; ϕ_1'; $\phi_1 = \text{upsample}(\phi_2) + \phi_1'$
0	Non-rigid refinement (finest)	$[\hat{M}_0^{s'},\ \hat{F}_0^s,\ f_1]$	f_0; ϕ_0'; $\phi = \text{upsample}(\phi_1) + \phi_0'$

Table 2. Ablation of the rigid/non-rigid step ratio in HPPM on RGB-TIR and RGB-SAR datasets.

Steps in HPPM		RGB-TIR		RGB-SAR	
N_{rigid}	$N_{nonrigid}$	RE↓	NCC↑	RE↓	NCC↑
1	3	0.929	0.7898	4.242	0.8205
1	4	0.744	0.7960	3.161	0.8664
1	5	1.024	0.7859	5.077	0.8046

feature $\hat{M}_{i-1}^{s'}$ that serves as part of the input to the HRB at scale $i-1$. In this way, each scale operates on moving features that have already been globally aligned and progressively refined by all previous scales.

3. Dataset Overview and Examples

We evaluate HRNet on four multimodal datasets covering both natural and remote-sensing scenarios, with aligned image pairs serving as the reference for registration.

RGB-NIR. For the RGB-NIR task, we use the RGB-NIR Scene dataset [1], which contains 442 high-resolution RGB-near-infrared image pairs captured in outdoor scenes such as buildings, vegetation, and urban environments. These pairs are provided in an approximately aligned form. Following the setting in the main paper, we crop the original images into 256×256 patches and construct 3,000 pairs for training and 300 pairs for testing.

RGB-TIR. For the RGB-TIR task, we utilize the TBBR dataset [10], which contains aligned RGB and thermal infrared image pairs captured from low-altitude UAV flights over building roofs. We randomly sample 3,000 pairs for training and 300 pairs for testing.

RGB-IR and RGB-SAR. For remote-sensing experiments, we use the MRSR dataset [8], which contains co-registered RGB-infrared (IR) and RGB-SAR image pairs. The RGB-IR subset consists of 4,000 pairs, and the RGB-SAR subset consists of 3,850 pairs, covering various land-cover types such as urban areas, farmland, water bodies, and forests. For both modality pairs, we use 3,500 pairs for training and the remaining pairs for testing.

Fig. 2 shows representative image pairs from these four datasets.

Table 3. Complexity comparison of rigid registration methods. Para. denotes the number of parameters.

Method	IHN	InMIRNet	RHWF	SCPNet	MCNet	MMRNet	Ours
Para. (M)	1.7	135.1	1.29	1.6	0.85	32.6	128.1
FLOPs (G)	51.8	35.5	137.6	100.4	36.4	122	186.6
Time (ms)	37.6	24.3	133.2	20.6	28.5	9.8	25.2

Table 4. Complexity comparison of non-rigid registration methods. Para. denotes the number of parameters.

Method	SuperFusion	InMIRNet	NBRNet	ADRNet	MMRNet	Ours
Para. (M)	1.9	0.14	13.6	75.6	28.8	128.1
FLOPs (G)	45.3	4.6	146.2	53.8	202	186.6
Time (ms)	7.5	6.7	68.1	6.4	7.8	25.2

4. More Experimental Results

4.1. Structure Ablation

As shown in Table 2, the 1:4 configuration used in the main paper achieves the best performance on both datasets. When we increase the number of non-rigid steps to 1:5, the coarsest scale becomes too small to provide a reliable global view, making the rigid alignment less effective and degrading the overall registration quality. In contrast, the 1:3 configuration leaves fewer scales for non-rigid refinement, which leads to insufficient local correction and higher residual errors. These results indicate that allocating one coarse rigid step followed by four non-rigid refinement scales strikes a good balance between global alignment and fine-grained deformation modeling.

4.2. Complexity Comparison

To provide a comprehensive view of the computational cost of existing registration models, we report the model size, FLOPs, and inference time for representative methods designed for two different task types: rigid registration and non-rigid registration. These comparisons are independent of any specific dataset and instead reflect the inherent complexity of each model family.

Table 3 summarizes the complexity of rigid registration methods, including IHN [2], InMIRNet [4], RHWF [3], SCPNet [11], MCNet [14], and MMRNet [8]. Table 4 further reports the complexity of non-rigid registration methods, such as SuperFusion [6], NBRNet [9], ADRNet [7], and MMRNet. Our HRNet, although equipped with a unified hybrid registration design, is included in both tables for reference. Overall, these results show that HRNet maintains a competitive computational profile in both categories. While HRNet has a larger parameter size than lightweight rigid-registration networks, it remains within a reasonable FLOPs range and achieves practical inference speed. Compared with high-capacity non-rigid registration

	RGB–NIR		PET–MRI	
	RE ↓	NCC ↑	RE ↓	NCC ↑
PGMR	$4.491_{0.069}$	$0.631_{0.004}$	$8.236_{0.356}$	$0.317_{0.004}$
HRNet	$1.633_{0.044}$	$0.755_{0.002}$	$3.329_{0.001}$	$0.606_{0.004}$

Table 5. Quantitative comparison between HRNet and PGMR on RGB–NIR and PET–MRI. Results are reported as $mean_{std}$.

	PET–MRI	
	SSIM ↑	VIF ↑
PGMR	1.12	0.41
HRNet	1.32	0.79

Table 6. Fusion performance on PET–MRI using the aligned image pairs produced by different registration methods.

models, HRNet achieves comparable FLOPs with a more balanced hybrid architecture, supporting both rigid and non-rigid transformations within a single pipeline.

4.3. Additional Qualitative Comparisons

To provide a more comprehensive evaluation of registration performance, we present additional qualitative comparisons on all four multimodal datasets (RGB-NIR, RGB-TIR, RGB-IR, and RGB-SAR). As shown in Figs. 3–10, for each dataset, we show the results of rigid-only registration and full non-rigid registration separately, and include side-by-side comparisons with representative baseline methods. In total, eight sets of qualitative results are provided, covering both rigid and non-rigid transformation settings across all datasets. Each visualization contains the fixed image, the moving image before registration, and the registered results produced by HRNet as well as competing approaches. These comparisons allow us to examine the strengths and limitations of each method from a visual perspective. Overall, rigid-only methods reduce large global offsets but often leave noticeable residual local distortions. HRNet produces cleaner global alignment than baseline rigid methods, and, under the full non-rigid setting, further corrects local deformations, yielding sharper boundaries, more consistent structures, and better semantic correspondence across modalities. Across all datasets, HRNet consistently provides more accurate and visually coherent alignment than competing methods under both transformation types.

4.4. Additional Evaluation

To further verify the generalization ability of the our method, we provide additional evaluation on a medical multimodal registration and its downstream fusion task. Specifically, we include the PET–MRI (Harvard) dataset[1] and compare our method with PGMR [13]. For completeness, we also report the comparison on RGB–NIR under the same evaluation protocol. As shown in Table 5, HRNet consistently outperforms PGMR on both datasets across all metrics. In particular, on RGB–NIR, HRNet reduces RE from 4.491 to 1.633, improves NCC from 0.631 to 0.755. On the PET–MRI dataset, HRNet also achieves clear gains, decreasing RE from 8.236 to 3.329, improving NCC from 0.317 to 0.606. To further verify the practical utility of improved registration quality, we conduct an additional downstream evaluation on the PET–MRI fusion task. After registration, the aligned image pairs are fed into the same fusion network [12] to assess whether better alignment leads to better fusion quality. The results are summarized in Table 6. Compared with PGMR, the aligned pairs produced by HRNet lead to superior fusion performance, improving SSIM from 1.12 to 1.32 and VIF from 0.41 to 0.79. This observation indicates that the gains of HRNet are not limited to registration metrics alone, but can also translate into more favorable downstream multimodal fusion results.

s

[1] https://www.med.harvard.edu/AANLIB/home.html

References

[1] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 177–184. IEEE, 2011. 2

[2] Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative deep homography estimation. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1879–1888, 2022. 2

[3] Si-Yuan Cao, Runmin Zhang, Lun Luo, Beinan Yu, Zehua Sheng, Junwei Li, and Hui-Liang Shen. Recurrent homography estimation using homography-guided image warping and focus transformer. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 9833–9842, 2023. 2

[4] Xin Deng, Enpeng Liu, Shengxi Li, Yiping Duan, and Mai Xu. Interpretable multi-modal image registration network based on disentangled convolutional sparse coding. *IEEE Trans. Image Process. (TIP)*, 32:1078–1091, 2023. 2

[5] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *Eur. Conf. Comput. Vis. (ECCV)*, 2024. 1

[6] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE-CAA J AUTOMATIC (JAS)*, 9(12):2121–2137, 2022. 2

[7] Yun Xiao, Chunlei Zhang, Yuan Chen, Bo Jiang, and Jin Tang. Adrnet: Affine and deformable registration networks for multimodal remote sensing images. *IEEE Trans. Geosci. Remote Sens. (TGRS)*, 62:1–13, 2024. 2

[8] Yun Xiao, Chunlei Zhang, Bo Jiang, Yuan Chen, and Jin Tang. Multi-modal remote sensing image registration via modality perception and self-supervised position estimation. *IEEE Trans. Geosci. Remote Sens. (TGRS)*, 2025. 2

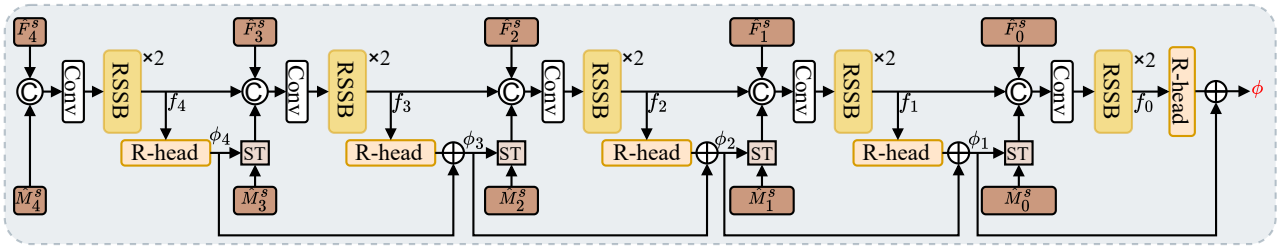

Figure 1. Detailed architecture of the Hybrid Parameter Prediction Module (HPPM). Conv: convolution; RSSB: Residual State Space Block; ST: Spatial Transformer; R-head: rigid registration head; M_i^s and F_i^s denote shared moving and fixed features at scale i.

[9] Yingxiao Xu, Jun Li, Chun Du, and Hao Chen. Nbr-net: A nonrigid bidirectional registration network for multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens. (TGRS)*, 60:1–15, 2022. 2

[10] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: post-capture control of resolution, dynamic range, and spectrum. *IEEE Trans. Image Process.(TIP)*, 19(9):2241–2253, 2010. 2

[11] Runmin Zhang, Jun Ma, Si-Yuan Cao, Lun Luo, Beinan Yu, Shu-Jie Chen, Junwei Li, and Hui-Liang Shen. Scpnet: Unsupervised cross-modal homography estimation via intra-modal self-supervised learning. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 460–477, 2024. 2

[12] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5906–5916, 2023. 3

[13] Tianheng Zheng, Guanglu Dong, Pingping Zhang, Xiaohai He, and Chao Ren. Plug-and-play general image registration for misaligned multi-modal image fusion. *IEEE Trans. Circuits Syst. Video Technol.*, 2025. 3

[14] Haokai Zhu, Si-Yuan Cao, Jianxin Hu, Sitong Zuo, Beinan Yu, Jiacheng Ying, Junwei Li, and Hui-Liang Shen. Mcnet: Rethinking the core ingredients for accurate and efficient homography estimation. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 25932–25941, 2024. 2

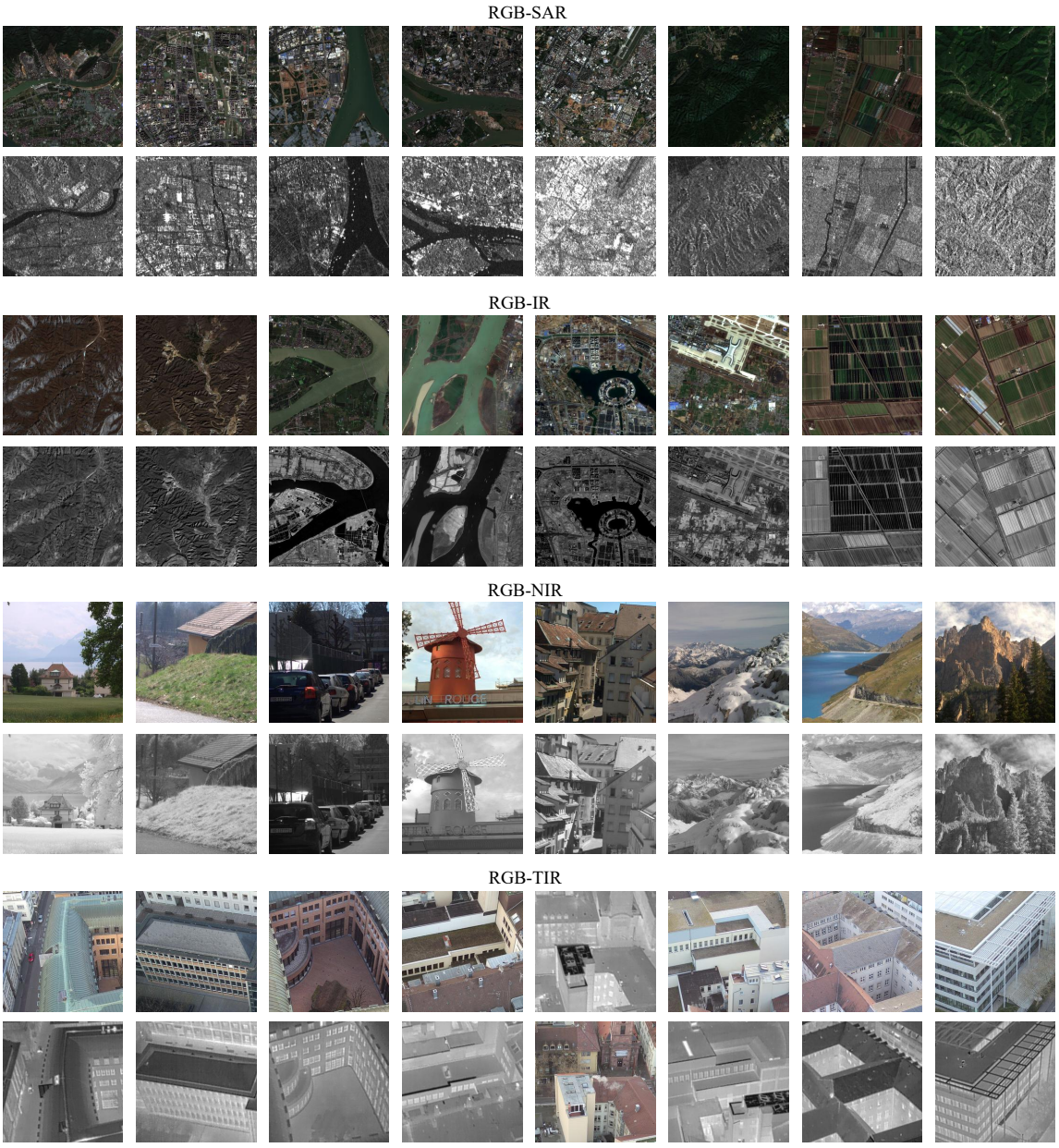

Figure 2. Representative multimodal image pairs from the four datasets used in our experiments.

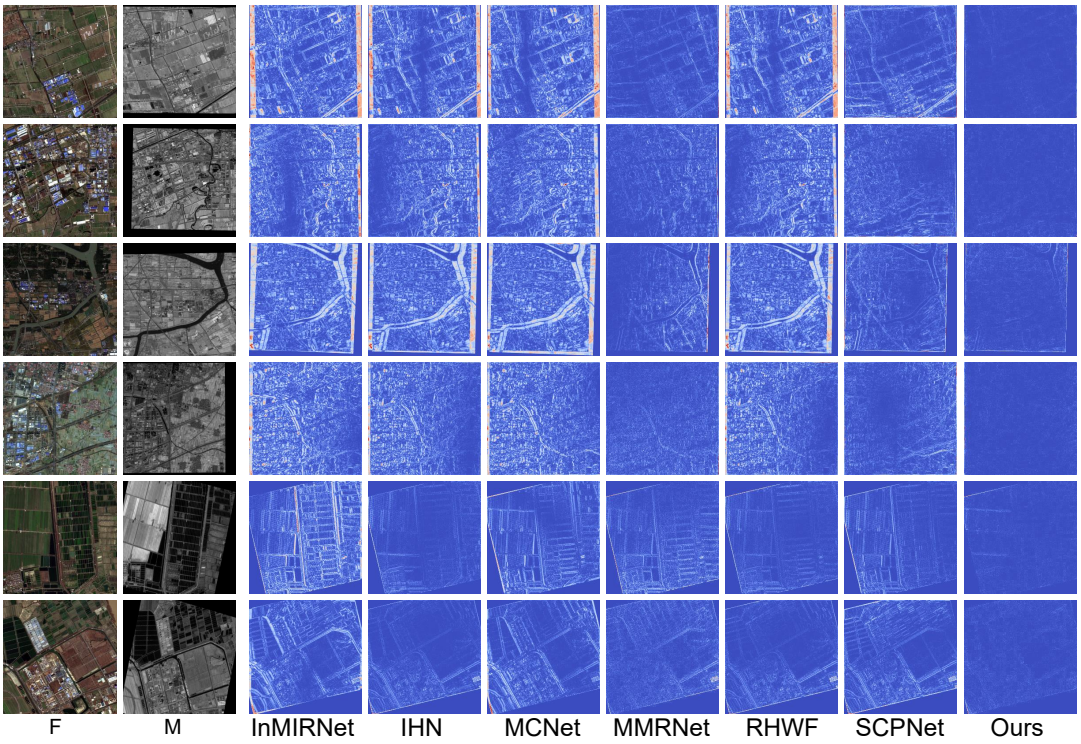

Figure 3. Qualitative comparison of rigid registration on the RGB-IR dataset. F: fixed image; M: moving image; the other columns show the results of different methods.

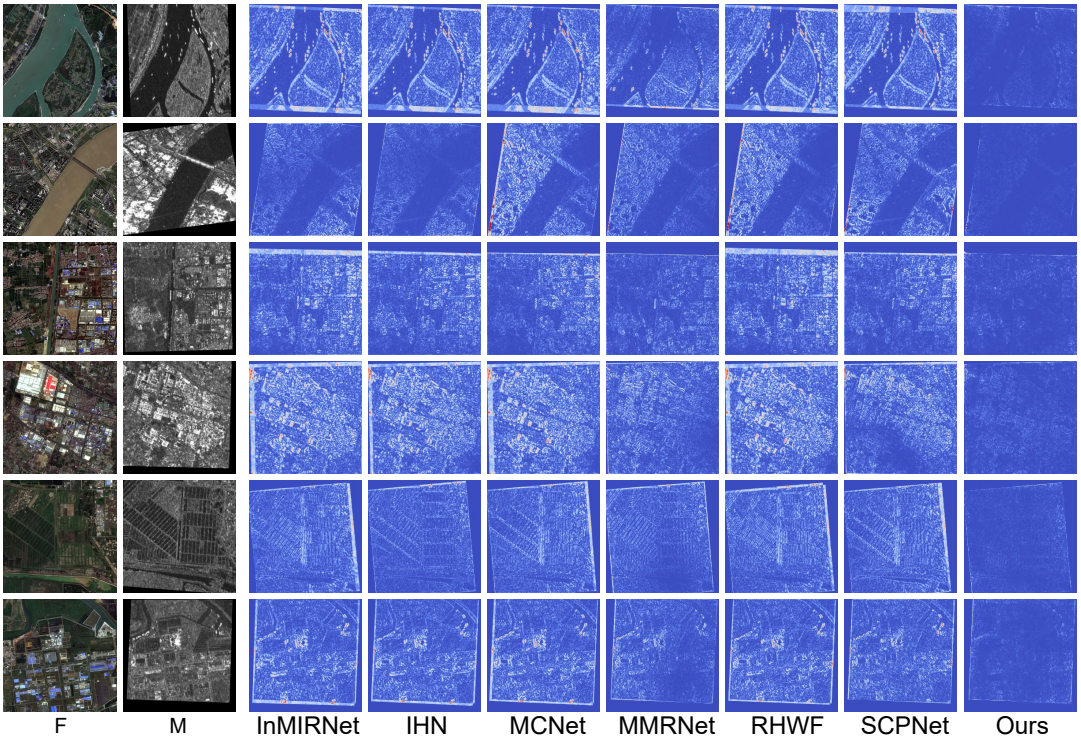

Figure 4. Qualitative comparison of rigid registration on the RGB-SAR dataset. F: fixed image; M: moving image; the other columns show the results of different methods.

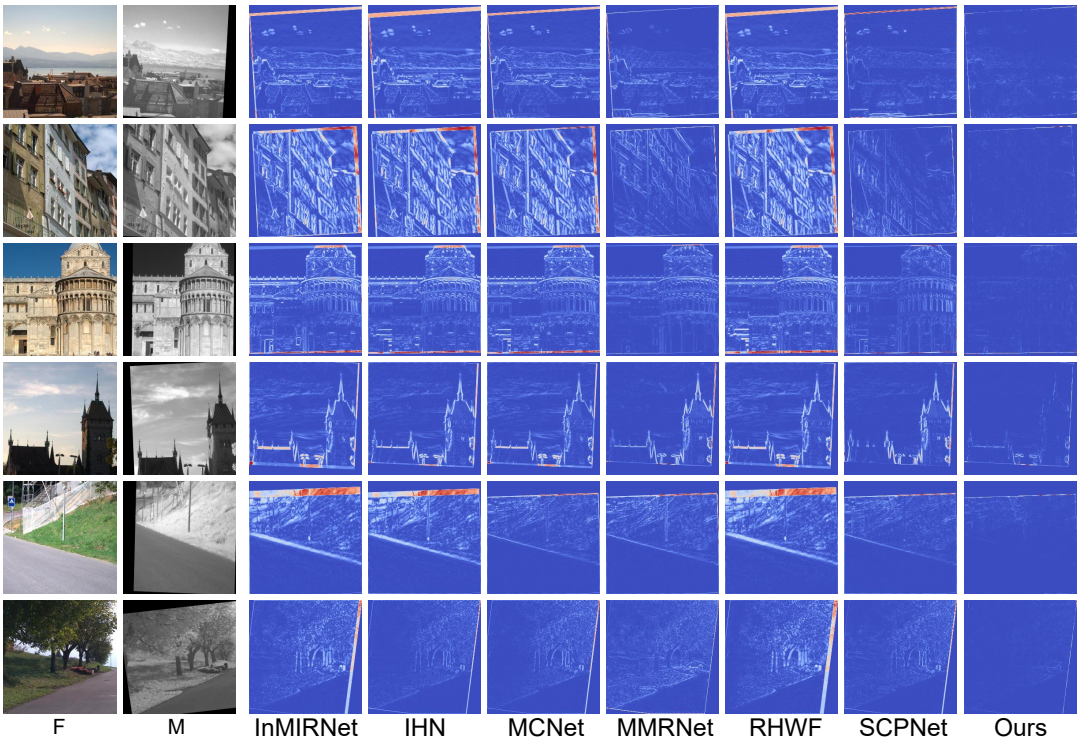

Figure 5. Qualitative comparison of rigid registration on the RGB-NIR dataset. F: fixed image; M: moving image; the other columns show the results of different methods.

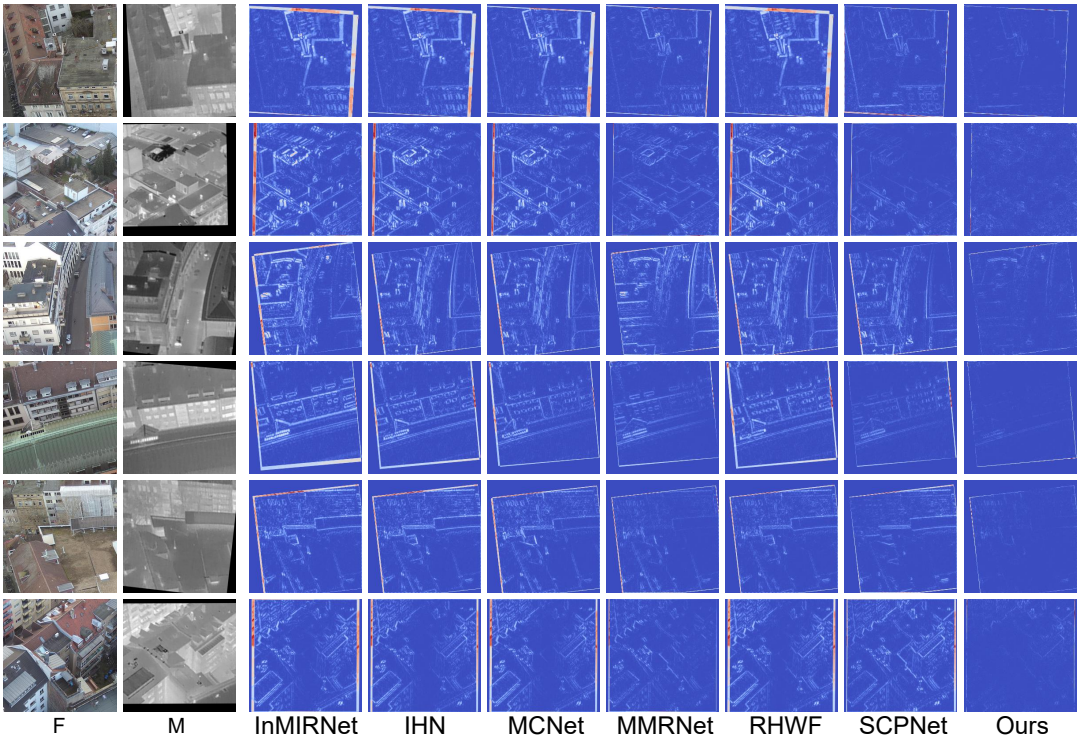

Figure 6. Qualitative comparison of rigid registration on the RGB-TIR dataset. F: fixed image; M: moving image; the other columns show the results of different methods.

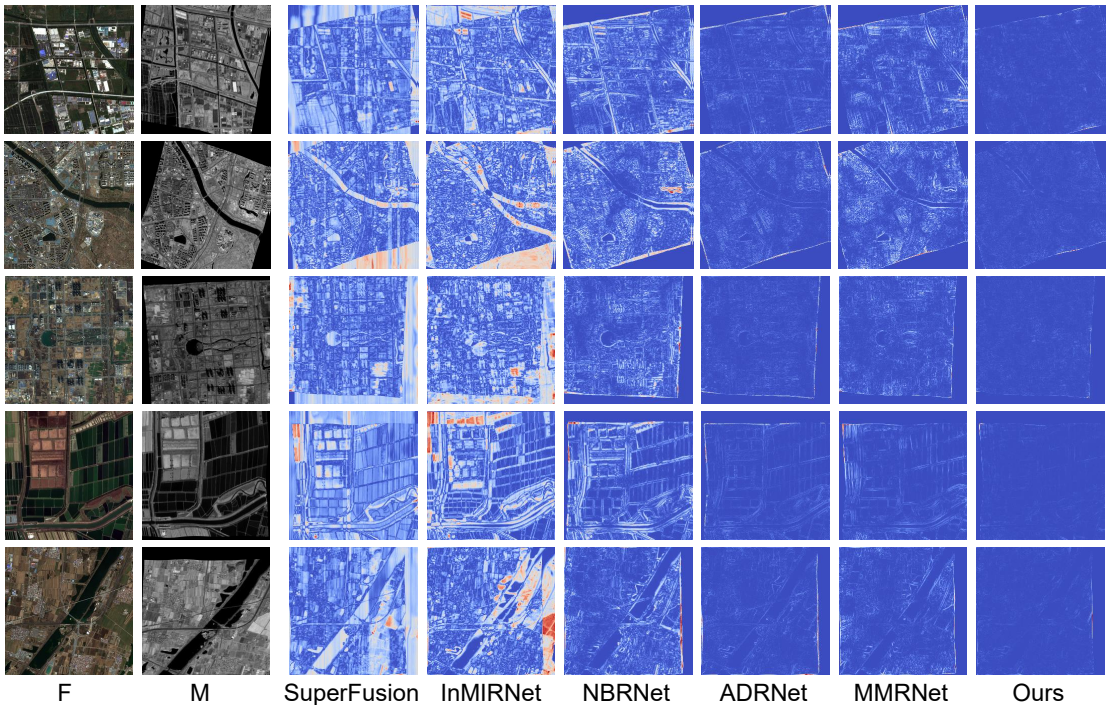

Figure 7. Qualitative comparison of non-rigid registration on the RGB-IR dataset. F: fixed image; M: moving image; the other columns show the results of different methods.

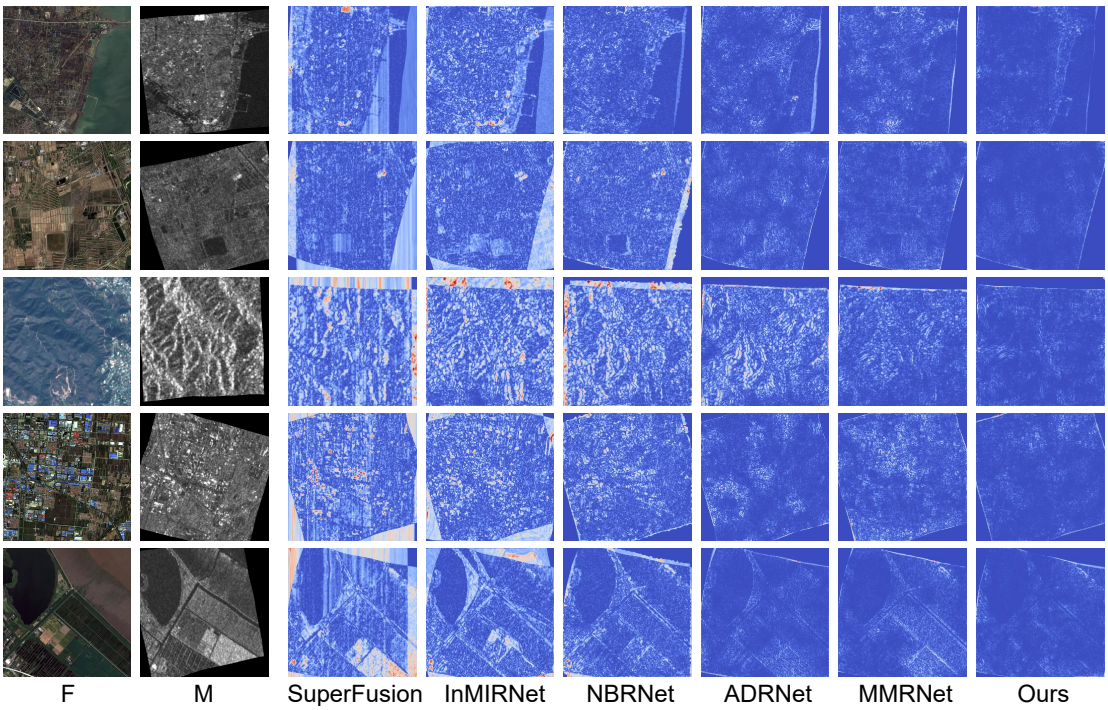

Figure 8. Qualitative comparison of non-rigid registration on the RGB-SAR dataset. F: fixed image; M: moving image; the other columns show the results of different methods.

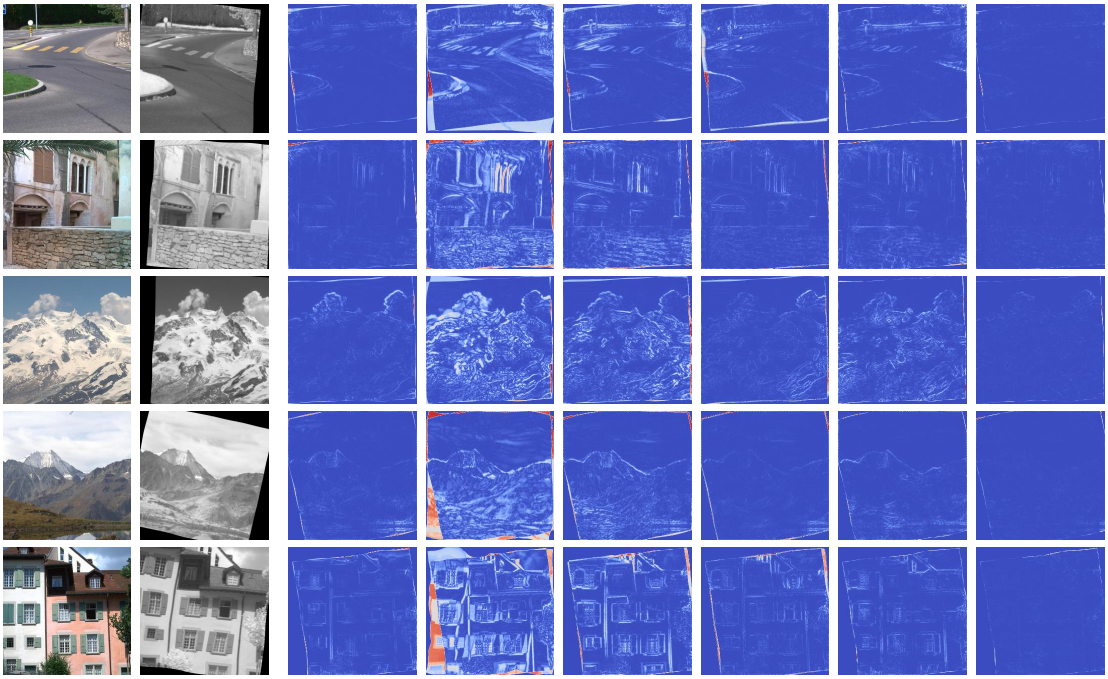

Figure 9. Qualitative comparison of non-rigid registration on the RGB-NIR dataset. F: fixed image; M: moving image; the other columns show the results of different methods.

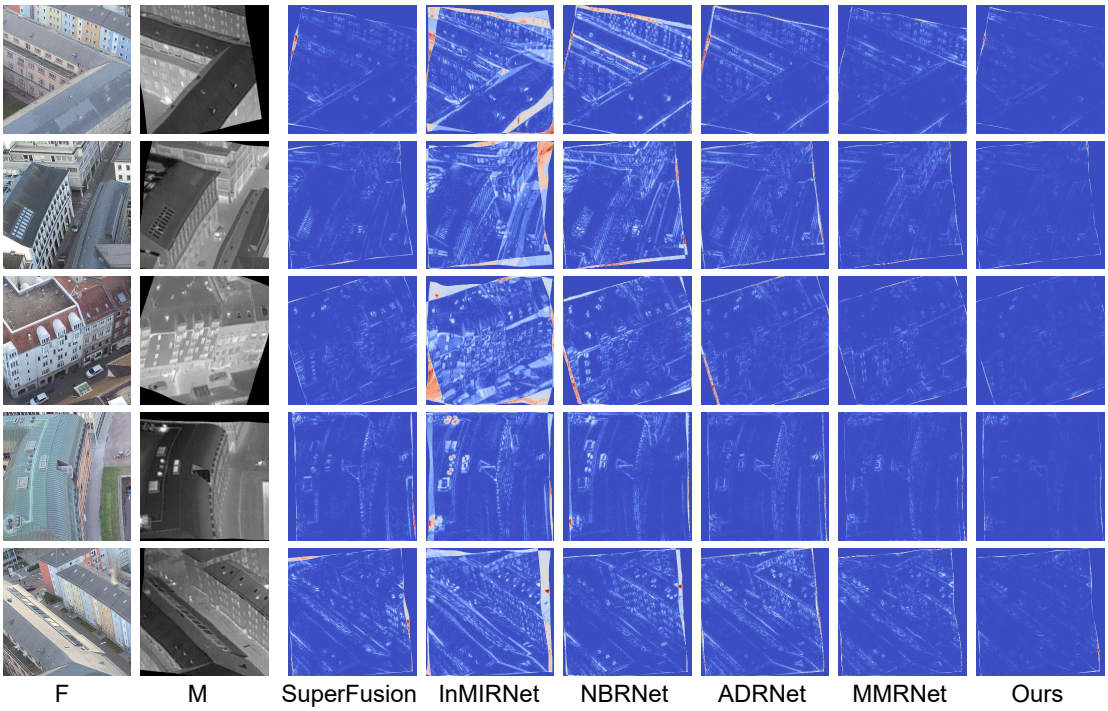

Figure 10. Qualitative comparison of non-rigid registration on the RGB-TIR dataset. F: fixed image; M: moving image; the other columns show the results of different methods.