CRFT: Consistent–Recurrent Feature Flow Transformer for Cross-Modal Image Registration

Supplementary Material

A. Visualization of Registration Results

To provide deeper insight into the behavior of CRFT, we present additional qualitative visualizations on both OSdataset and RoadScene. These examples highlight the model's robustness across heterogeneous modalities and challenging geometric conditions.

Flow prediction visualization. Figs. 7 and 8 show dense flow predictions on OSdataset and RoadScene. CRFT produces smooth and structurally coherent flow fields that closely follow the scene geometry. The difference maps contain only small residuals, indicating accurate displacement estimation despite strong modality-induced appearance variations. These results confirm the stability of CRFT's iterative discrepancy-guided flow refinement mechanism across different sensing modalities.

Registration comparison across modalities. Fig. 9 presents large-scale qualitative comparisons on both OSdataset (top) and RoadScene (bottom). Classical methods such as HOWP and MSG, as well as deep learning-based approaches including GMFlow and XoFTR, exhibit reduced robustness under large rotations, scale variations, and substantial optical-SAR/infrared appearance gaps, often resulting in angular or translational misalignments. In contrast, CRFT produces registration results that align closely with the ground-truth quadrilaterals, yielding lower end-point errors and more stable geometric fidelity.

Checkerboard fusion and fine-scale alignment. To further evaluate pixel-level correspondence, Fig. 10 visualizes checkerboard fusion results on OSdataset (left) and RoadScene (right). CRFT achieves coherent grid alignment across both datasets without tearing or ghosting artifacts. The zoomed-in patches demonstrate precise matching of fine structural elements such as field boundaries, edges, and textures. These observations validate CRFT's ability to recover fine-scale geometric correspondence in complex cross-modal registration scenarios.

B. Limitations and Failure Cases.

While our method demonstrates strong robustness under standard augmentations, we acknowledge a performance drop under extreme geometric distortions. when subjected to extreme 90° rotations coupled with random scaling in the range of $[0.5, 1.5]$, the AEPE on the OSdataset rises significantly to 9.75. This reveals a vulnerability to extreme affine variations, highlighting the need for scale- and rotation-invariant representations in future research.

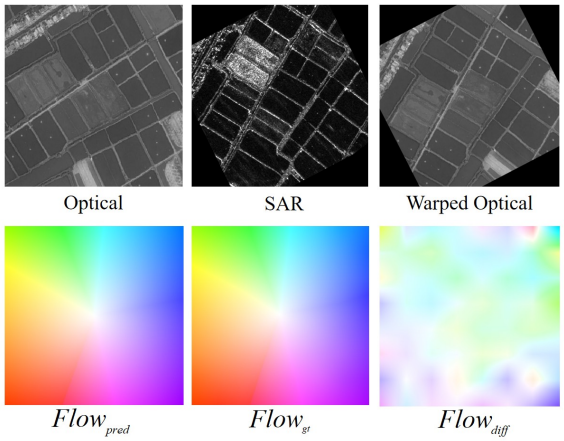

Figure 7. **Flow prediction visualization on OSdataset.** For each pair, we visualize the predicted dense flow, the ground-truth flow, and the corresponding difference map. CRFT produces smooth and geometrically consistent flow fields across optical-SAR modality gap, indicating accurate geometric correspondence and robustness to strong appearance variations

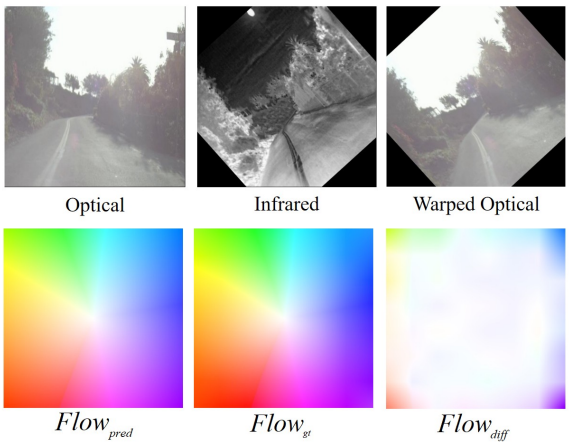

Figure 8. **Flow prediction visualization on RoadScene.** For each pair, we visualize the predicted dense flow, the ground-truth flow, and the corresponding difference map. CRFT maintains stable and consistent flow behavior across heterogeneous modalities, demonstrating robustness to illumination changes, noise patterns, and texture inconsistencies in complex driving environments.

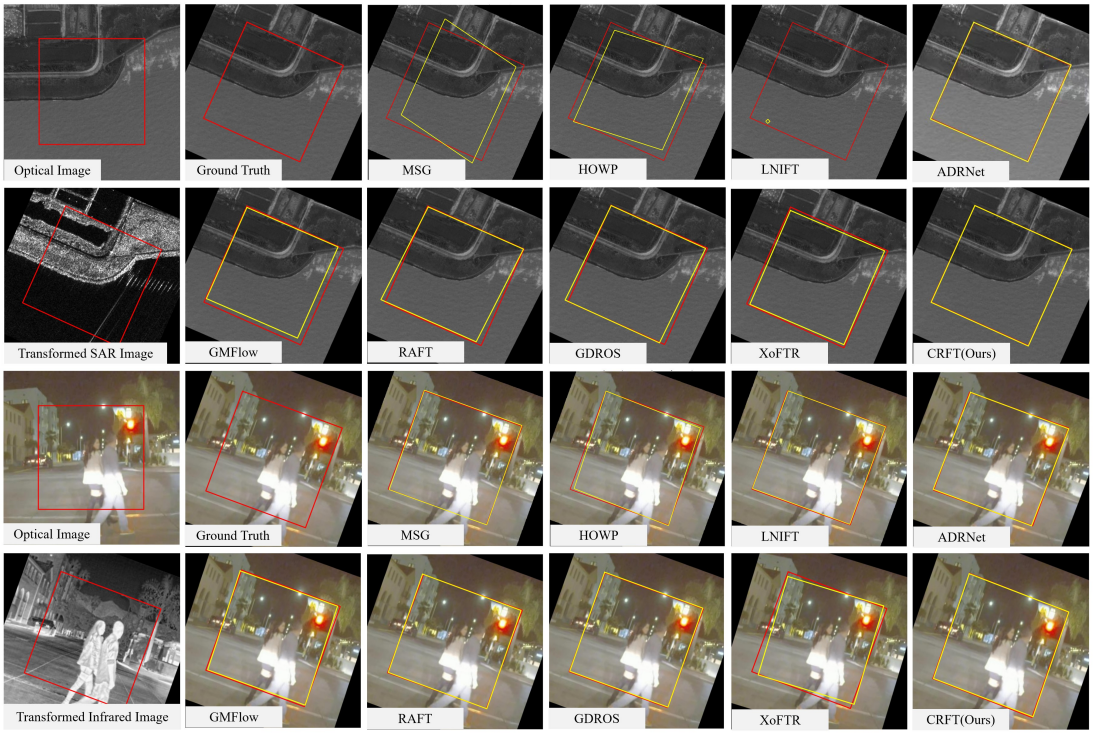

Figure 9. **Qualitative comparison on the OSdataset (top) and RoadScene (Bottom).** The red quadrilateral denotes the ground-truth alignment, while the yellow quadrilateral shows the predicted registration result. CRFT achieves more accurate and geometrically consistent registration under large geometric deformation and modality gaps compared with existing optical-SAR registration baselines.

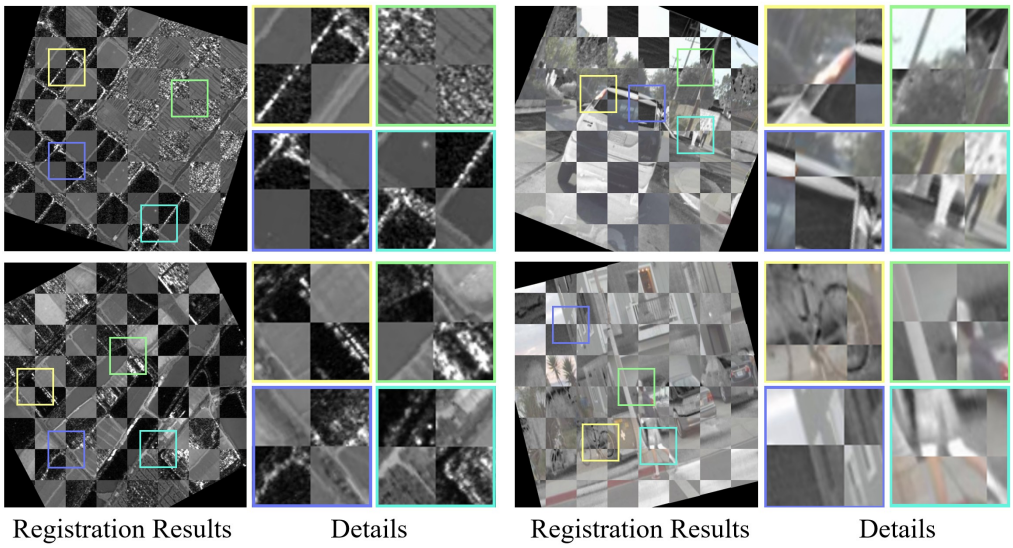

Figure 10. **Checkerboard registration on OSdataset (left) and RoadScene (right).** CRFT yields geometrically coherent checkerboard fusion across both optical-SAR and optical-infrared modalities. Zoomed-in regions demonstrate precise alignment of boundaries and textures, confirming the model's ability to recover fine-scale geometric correspondence under large cross-modal appearance differences.