
SAGE: SOURCE-ANCHORED GUIDANCE VIA FREQUENCY EQUALIZATION FOR HIERARCHICAL RGB-T ALIGNMENT AND FUSION

Timing Li, Yiming Sun, Boan Tao, Xiyuan Gao, Haifang Cao, Pengfei Zhu

ABSTRACT

Spatial misregistration and cross-modal discrepancies often cause ghosting, structural blurring, and content imbalance in RGB-T fusion. Existing methods typically decouple appearance adaptation, geometric alignment, and information fusion, limiting dependency propagation across stages. We propose Source-Anchored Guidance via Frequency Equalization for Hierarchical RGB-T Alignment and Fusion (SAGE), a unified framework integrating frequency equalization, hierarchical alignment, and subband fusion. SAGE employs invertible joint encoding and source-specific low-frequency modulation to derive structural and gain guidance while preserving source information. Hierarchical frequency collaborative alignment estimates global affine geometry from low-frequency approximations and transfers geometric and contextual cues to high-frequency correlation reasoning for reliability-aware residual refinement. Guided subband fusion jointly aggregates the aligned frequency coefficients under propagated source and alignment guidance, coordinates complementary low- and high-frequency information, and reconstructs the fused image through the inverse wavelet transform. Extensive experiments on RGB-T datasets with real-world and synthetic misalignments demonstrate consistently competitive performance in alignment and fusion, validating the effectiveness of source-anchored guidance for weakly registered RGB-T images.

Keywords image registration, image alignment, image fusion, multi-modal learning

1 Introduction

RGB-T image fusion aims to integrate complementary information from visible and thermal sensors into a unified representation [1, 2]. Visible images provide rich textures and scene details, whereas thermal images emphasize radiation-sensitive targets and remain informative under adverse illumination. Their combination benefits vision tasks, including object detection, semantic segmentation, and scene understanding [3, 4, 5, 6]. In practice, however, differences in sensor placement, field of view, and sampling characteristics introduce residual geometric misregistration [7, 8]. Meanwhile, the distinct imaging mechanisms of visible and thermal sensors produce cross-modal appearance discrepancies. Directly combining heterogeneous observations can generate ghosting artifacts, structural blur, and biased information aggregation.

To address these challenges, methods reduce cross-modal discrepancy through a registration-friendly proxy or shared representation before estimating geometric correspondence [9, 10, 11, 12, 13, 14, 15]. Approaches couple registration and fusion via feature sharing, fusion feedback, or fusion-derived supervision [10, 11, 12, 14, 16, 17]. Although these developments improve correspondence learning and task coordination, intermediate representations facilitate registration, whereas fusion information serves as feedback or supervision. Cross-modal information established during discrepancy reduction is rarely retained and jointly exploited for alignment and fusion. Moreover, deformation is estimated from spatial images or mixed latent features through an undifferentiated process, although coarse scene layout and local structural details provide distinct geometric evidence. Consequently, cross-modal representation, alignment, and fusion remain insufficiently coordinated, leaving cross-stage and cross-frequency dependencies underexploited.

These limitations motivate the use of a representation that distinguishes global layout from local structure and supports information transfer across stages. Frequency decomposition is well suited here because it organizes image content

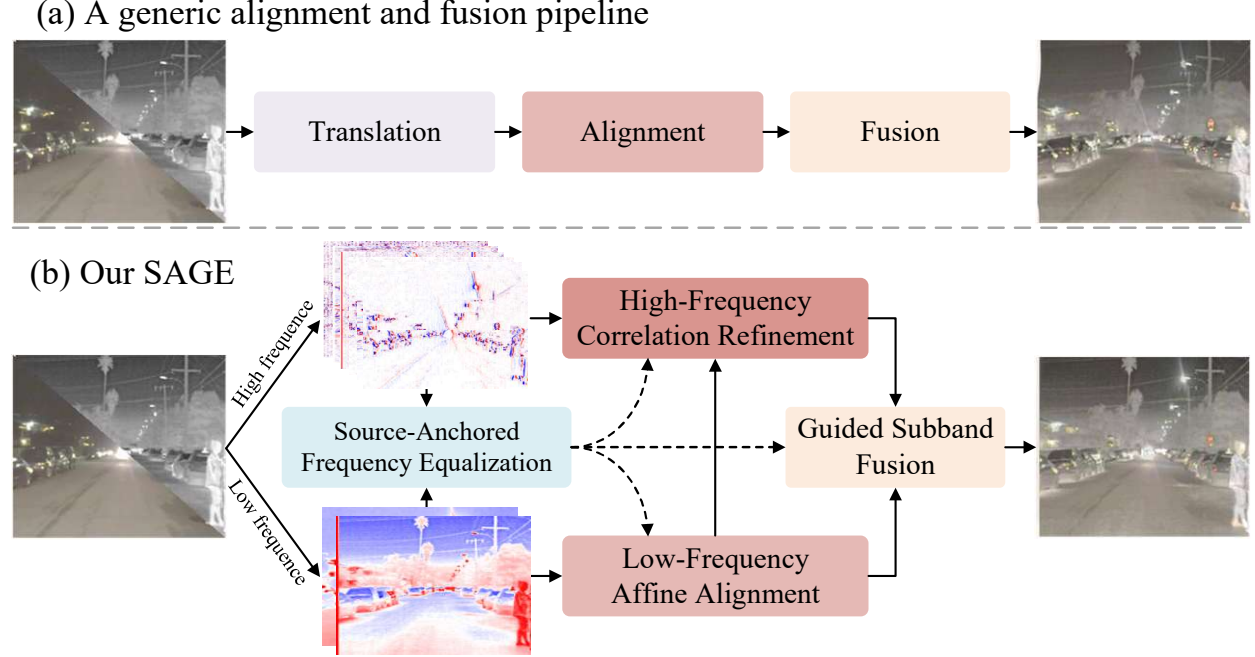


Figure 1: Comparison between a conventional pipeline based on image translation and SAGE. (a) Image translation, alignment, and fusion are performed sequentially. (b) SAGE uses source-anchored frequency equalization to guide hierarchical alignment from low to high frequencies and subband fusion. Solid arrows denote feature and dependency flows, while dashed arrows denote guidance propagation.

according to spatial frequency. Low frequency components retain coarse radiometric variation and global scene layout, making them suitable for global parametric estimation. High frequency components emphasize localized edges and textures, providing complementary evidence for refining local residual displacement after coarse warping. This distinction also applies to fusion because low frequency intensity organization and high frequency structural details require different aggregation behavior. Frequency representations can therefore provide a common basis for cross-modal discrepancy reduction, hierarchical alignment, and adaptive fusion. This shared basis allows information extracted during discrepancy reduction to guide subsequent alignment and fusion stages.

Based on this observation, we propose source-anchored guidance via frequency equalization for hierarchical RGB-T alignment and fusion (SAGE), as shown in Figure 1. Source-anchored frequency equalization (SAFE) jointly models paired wavelet coefficients while retaining source-indexed streams. It applies bounded modulation to low-frequency components, preserves high-frequency structures, and extracts structural and gain cues available to alignment and fusion. The resulting cues propagate information learned during discrepancy reduction to geometric alignment and subband fusion. Hierarchical frequency collaborative alignment (HFCA) assigns global geometry estimation to low-frequency affine alignment (LFA) and residual local correction to high-frequency correlation refinement (HCR), with LFA transferring affine geometry and low-frequency context to HCR. Guided subband fusion (GSF) then integrates aligned coefficients with propagated source cues and alignment reliability for frequency-specific aggregation and cross-frequency coordination before reconstruction. This organization links source preservation, low-to-high geometric reasoning, and reliability-aware fusion. We examine these design choices through component ablations, alternative frequency-branch assignments, and stage-wise guidance routing. The main contributions of this work are summarized as follows:

- We propose a unified source-anchored frequency guidance framework for RGB-T alignment and fusion. It combines invertible joint encoding with source-aware frequency modeling to preserve modality-specific structures and propagate structural and gain cues across stages.
- We develop a hierarchical frequency collaborative alignment architecture that models the global-to-local dependency between low- and high-frequency information. Low-frequency information establishes global geometric correspondence and provides contextual guidance for high-frequency local refinement.

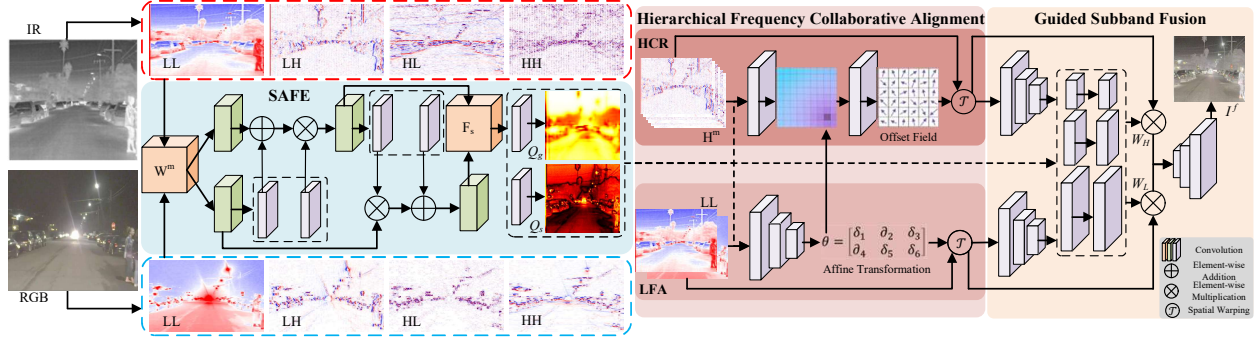


Figure 2: Overview of SAGE. The source anchored frequency guidance mechanism consistently propagates structural and gain cues across frequency equalization, hierarchical alignment, and subband fusion. SAFE reduces low frequency discrepancy while preserving source high frequency details. HFCA combines low frequency affine alignment with reliability guided high frequency refinement. GSF adaptively fuses the aligned subbands and reconstructs the final image through inverse Haar wavelet transform.

- We design a guided subband fusion that combines source-derived cues with correspondence and reliability information for frequency-specific aggregation and cross-frequency interaction. Extensive experiments demonstrate competitive overall performance across alignment and fusion metrics.

2 Related Work

Image Alignment. Geometric misalignment is a major obstacle to reliable infrared and visible image fusion, since spatial discrepancies can introduce ghosting and structural inconsistency into fused results. Existing methods generally alleviate cross modal appearance discrepancies before estimating geometric correspondence. UMF-CMGR [9] and IMF [10] progressively model cross modal deformation through transformation and refinement based registration pipelines, while HR4IR [13] constructs a harmonized representation domain and alternately searches for cross modal correspondences and geometric transformations. RFNet [11] and MURF [12] further formulate registration and fusion as mutually reinforcing tasks, allowing alignment and fusion to benefit from their reciprocal interaction. Recent studies increasingly couple registration with fusion to strengthen correspondence learning. AU-Net [14] jointly optimizes the two tasks at the feature level, C2RF [16] exploits commonality mining and fusion guided contrastive learning, and MulFS-CAP [17] introduces fusion supervision to facilitate alignment perception for unregistered inputs. Self supervised correspondence modeling has also received growing attention. B-SR [7] enforces deformation consistency through bidirectional proxy transformations, whereas Hy-CycleAlign [2] combines cyclic registration with hyperbolic correspondence modeling. Despite these advances, existing methods predominantly estimate deformation from spatial images or mixed latent representations, leaving heterogeneous frequency components largely subject to an undifferentiated alignment process. Our method instead performs hierarchical frequency aware alignment, where low frequency affine estimation provides a coarse geometric prior for subsequent high frequency correlation refinement.

Image Fusion. Infrared and visible image fusion aims to preserve thermal saliency while retaining visible structural details. Recent studies have increasingly explored adaptive interaction and information selection under diverse conditions [18, 19, 20]. DCEvo [21] improves multimodal integration through discriminative cross dimensional interaction, while EMMA [22] introduces equivariant learning for multimodal information integration. TG-ECNet [23] introduces task aware gating and multi expert collaboration for degraded multimodal fusion, whereas HCLFuse [24] explores generative fusion through information decomposition and physically guided generation. More recent methods further improve robustness and adaptivity through domain adaptation and region dependent fusion strategies, as demonstrated by DAFusion [25] and RegionFuse [26]. Despite these advances, fusion decisions are still mainly derived from current source or feature representations. Our method instead uses source anchored structural and gain cues generated before fusion to condition frequency dependent gating and cross frequency interaction, enabling guided information aggregation across different frequency components.

3 Method

SAGE is organized around a unified source-anchored frequency guidance mechanism that connects frequency equalization, hierarchical alignment, and subband fusion, as shown in Figure 2. This design preserves the distinct roles of low- and high-frequency information while consistently propagating source-derived cues across processing stages.

3.1 Overview

SAGE takes a weakly registered RGB-T pair (I^v, I^t) as input and produces a fused image I^f through source-anchored frequency equalization, hierarchical alignment, and guided subband fusion. We apply the Haar wavelet transform $\mathcal{D}$ to decompose each input into a low-low (LL) approximation subband and three directional detail subbands: low-high (LH), high-low (HL), and high-high (HH). For $m \in \{v, t\}$,

$$\begin{aligned}\mathcal{D}(I^m) &= (L_{LL}^m, \mathbf{H}^m), \\ \mathbf{H}^m &= [H_{LH}^m, H_{HL}^m, H_{HH}^m].\end{aligned}\quad (1)$$

We denote the complete wavelet representation as $\mathbf{W}^m = [L_{LL}^m, \mathbf{H}^m]$. An invertible joint encoder transforms $(\mathbf{W}^v, \mathbf{W}^t)$ into source-specific joint feature F_s . SAFE extracts structural and gain cues while preserving source-specific coefficients. HFCA integrates LFA to estimate affine geometry in the LL band and HCR to correct residual high-frequency misalignment. GSF then performs guidance-conditioned subband aggregation and cross-frequency coordination before inverse wavelet reconstruction.

SAFE performs pair conditioned equalization while preserving the coefficients associated with each source. For $m \in \{v, t\}$, the gain prediction head $\mathcal{P}_g^m$ processes $\mathbf{W}^v, \mathbf{W}^t$, their elementwise absolute difference, and the joint feature F_s . It predicts a bounded residual gain Γ^m to adaptively modulate the frequency representation while retaining the structural characteristics of the corresponding source.

$$\begin{aligned}\Gamma^m &= \gamma \tanh(\mathcal{P}_g^m(\mathbf{W}^v, \mathbf{W}^t, |\mathbf{W}^v - \mathbf{W}^t|, F_s)), \\ \hat{L}_{LL}^m &= L_{LL}^m \odot (1 + \Gamma^m), \\ \hat{\mathbf{H}}^m &= \mathbf{H}^m,\end{aligned}\quad (2)$$

$\odot$ denotes elementwise multiplication, γ limits modulation magnitude, and the hat indicates equalized coefficients. Restricting modulation to the LL subband reduces intensity discrepancies across modalities without altering source detail coefficients. SAFE derives a structural cue Q_s and a gain cue Q_g from F_s and propagates both to HFCA and GSF.

3.2 Hierarchical Frequency Collaborative Alignment (HFCA)

HFCA models RGB-T alignment through hierarchical frequency collaboration between LFA and HCR. LFA estimates global affine geometry from low-frequency cues, while HCR refines local residual displacement using high-frequency details after compensation. The transferred geometry and low-frequency context enable coordinated use of complementary frequency information across both components.

Low-Frequency Affine Alignment (LFA). Conditioned on the structural and gain cues Q_s and Q_g , LFA encodes the paired approximation subbands into a contextual representation F_L . Global aggregation of F_L predicts a 2×3 affine transformation θ , parameterized by six bounded residuals $\delta_1, \dots, \delta_6$ around the identity transformation. For notational clarity, we present the visible-to-thermal alignment below, taking the thermal modality as the reference. The estimated affine transformation maps the visible coefficients to the thermal coordinate system and provides the initial geometric compensation for the visible high-frequency coefficients. The global alignment is formulated as

$$\begin{aligned}L_a^v &= \mathcal{T}_\theta(\hat{L}_{LL}^v), \quad L_a^t = \hat{L}_{LL}^t, \\ \overline{\mathbf{H}}^v &= \mathcal{T}_\theta(\hat{\mathbf{H}}^v), \quad \overline{\mathbf{H}}^t = \hat{\mathbf{H}}^t,\end{aligned}\quad (3)$$

$\mathcal{T}_\theta$ denotes affine warping, and the overline indicates high-frequency coefficients after global compensation.

High-Frequency Correlation Refinement (HCR). HCR constructs a normalized local correlation volume from the globally compensated high-frequency features over a restricted search neighborhood. Combined with transferred low-frequency context and source-derived cues, the correlation evidence estimates a residual offset field D . HCR further derives a structural reliability map $R \in [0, 1]$ from coherent high-frequency responses and low-frequency gradients. The reliability-gated residual deformation and aligned high-frequency coefficients are computed as

$$\begin{aligned}\tilde{D} &= R \odot D, \\ \mathbf{H}_a^v &= \mathcal{T}_{\tilde{D}}(\overline{\mathbf{H}}^v), \quad \mathbf{H}_a^t = \overline{\mathbf{H}}^t.\end{aligned}\quad (4)$$

Through LFA and HCR, HFCA establishes a frequency-directed geometric dependency in which source-derived cues and structural reliability jointly govern a robust, adaptive residual displacement recovery process.

3.3 Guided Subband Fusion (GSF)

GSF incorporates guidance from SAFE and HFCA into frequency-specific fusion. The low-frequency gate is estimated from aligned approximation coefficients, Q_s , and Q_g , while the high-frequency gate also incorporates local correlation responses and the reliability map. Gates W_L and W_H are applied to aligned low- and high-frequency coefficients, respectively. The initial fused coefficients are computed as

$$\begin{aligned}\tilde{L}^f &= W_L \odot L_a^v + (1 - W_L) \odot L_a^t, \\ \tilde{\mathbf{H}}^f &= W_H \odot \mathbf{H}_a^v + (1 - W_H) \odot \mathbf{H}_a^t.\end{aligned}\tag{5}$$

The frequency-specific gates integrate global intensity organization and local structural evidence into the initial subband fusion. The fused coefficients are subsequently refined through cross-frequency interaction, with the reliability map regulating the high-frequency update according to local correspondence confidence. The final coefficients L^f and $\mathbf{H}^f$ are reconstructed as $I^f = \mathcal{D}^{-1}(L^f, \mathbf{H}^f)$.

3.4 Loss Function

SAGE is trained on weakly registered RGB-T image pairs. Its loss comprises equalization, alignment, and fusion terms. The equalization term $\mathcal{L}_{\text{eq}}$ applies Charbonnier loss [27] to invertible reconstruction and source-consistency errors, with predicted gain-field regularization. The alignment term $\mathcal{L}_{\text{align}}$ enforces correspondence between aligned low-frequency approximations and high-frequency energy representations via normalized cross-correlation [11], while regularizing affine transformation and residual offsets. The fusion term $\mathcal{L}_{\text{fus}}$ comprises maximum-intensity, maximum-gradient, and edge-preservation constraints derived from aligned source images [28]. The overall training loss is defined as

$$\mathcal{L} = \lambda_{\text{eq}} \mathcal{L}_{\text{eq}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{fus}} \mathcal{L}_{\text{fus}},\tag{6}$$

where λ_{eq} , λ_{align} , and λ_{fus} denote the weighting coefficients of the corresponding loss terms.

4 Experiments

4.1 Setup

Datasets. We evaluate SAGE on three public RGB-T datasets, including DroneVehicle [29], MFNet [30], and RoadScene [31]. DroneVehicle contains aerial RGB-T pairs with inherent cross-modal misalignment arising from platform motion, sensor parallax, and viewpoint variation. Its original image pairs are used without additional geometric perturbation. MFNet and RoadScene comprise ground-level traffic scenes captured under diverse illumination conditions.

Misalignment Settings. We evaluate DroneVehicle, MFNet, and RoadScene under complementary real-world, non-rigid, and rigid RGB-T misalignment scenarios. DroneVehicle retains its inherent aerial-view misalignment without additional perturbation. On MFNet, thermal images serve as references, while visible images undergo translations and elastic deformations using Gaussian kernels of sizes 85–101 with standard deviations of 24–32 pixels. On RoadScene, visible images serve as references, while thermal images are shifted leftward by 0.5% to 1.5% of image width. These settings cover both modality directions under real, synthetic non-rigid, and rigid conditions.

Metrics. We report five metrics to evaluate both fusion quality and cross-modal alignment accuracy. Average gradient (AG) and spatial frequency (SF) characterize the preservation of local details and spatial activity in the fused image. The sum of correlations of differences (SCD) assesses complementary information transfer and overall consistency with the source observations. To specifically evaluate alignment performance, we further report the 95th-percentile Hausdorff distance (HD95) and average symmetric surface distance (ASSD), which measure the geometric discrepancy between corresponding structural boundaries after alignment. Higher AG, SF, and SCD values indicate better fusion quality, while lower HD95 and ASSD values indicate more accurate structural alignment.

Implementation details. SAGE is implemented and trained on an NVIDIA GeForce RTX 3090. The network is optimized for 3000 epochs using AdamW with an initial learning rate of 5×10^{-5} . The frequency equalization parameter is set to $\gamma = 0.2$, while the loss weights are set to $\lambda_{\text{eq}} = 0.6$, $\lambda_{\text{align}} = 1$, and $\lambda_{\text{fus}} = 0.8$.

Quantitative evaluation. *Alignment quality.* As shown in Table 1, SAGE demonstrates competitive alignment performance across three datasets, which cover aerial observations and spatially varying and rigid deformations. Its

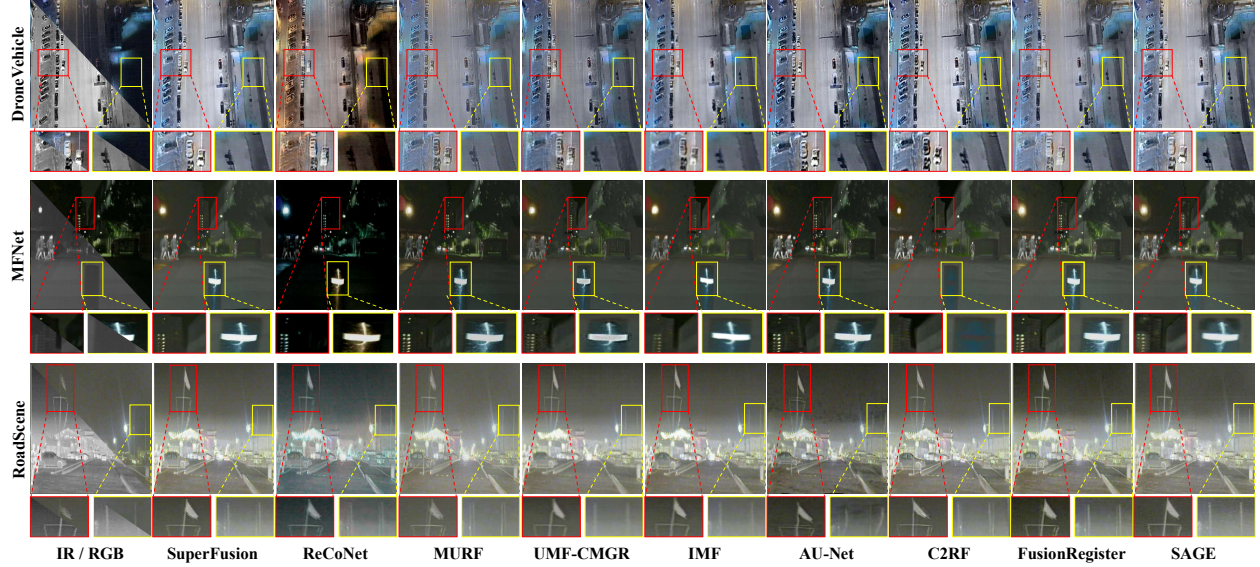


Figure 3: Qualitative comparisons on DroneVehicle, MFNet, and RoadScene.

advantage is most evident on MFNet, where local displacement varies across the image. The results indicate that SAGE accommodates deformation patterns while maintaining structural correspondence. This consistency is in line with the use of source-anchored frequency cues and frequency-aware global and local correspondence estimation.

Fusion quality. SAGE demonstrates competitive overall fusion performance across three benchmarks while preserving fine structures and complementary visible and thermal information. Its strong RoadScene results and generally stable performance elsewhere suggest a favorable balance among detail preservation structural fidelity and cross modal information retention under diverse imaging conditions. This balance arises from the complete SAGE pipeline where source frequency information is retained during equalization geometric correspondence is refined hierarchically and aligned content is aggregated through guided subband fusion.

4.2 Comparing with SOTA Methods

We compare SAGE against eight representative state-of-the-art methods developed for misaligned RGB-T image fusion. The selected approaches account for geometric inconsistency through correction, explicit registration, or joint registration and fusion. They include SuperFusion [32], ReCoNet [33], MURF [12], UMF-CMGR [9], IMF [10], AU-Net [14], C2RF [16], and FusionRegister [34].

Qualitative evaluation. *Alignment quality.* As shown in Figure 3, SAGE achieves sharper, spatially coherent alignment across the three datasets. On DroneVehicle, it preserves separable contours among densely parked vehicles and accurately localizes the isolated road target, while competing methods produce blurred boundaries or duplicated structures. On MFNet, SAGE maintains compact contours around the bright target and consistent facade patterns despite spatially varying displacement. On RoadScene, it reconstructs the flag and pole structures with clear single edges and minimal ghosting. These results indicate that SAGE handles both local and spatially varying misalignment.

Fusion quality. SAGE also provides balanced integration of infrared saliency and visible details. On DroneVehicle, it preserves vehicle structures, the isolated target, and road context without noticeable distortion. On MFNet, salient thermal targets are retained with tree and facade textures. On RoadScene, the flag boundary, roadside structures, and lane markings remain clear without excessive brightness spreading. Overall, SAGE produces fewer ghosting artifacts and less source bias by adaptively regulating frequency-specific aggregation according to alignment reliability.

4.3 Ablation Studies

We conduct ablation studies on MFNet to examine source-aware frequency equalization, hierarchical frequency collaborative alignment, and guided subband fusion, with Table 2 demonstrating the competitive overall performance of the complete SAGE model. The two alignment stages are removed separately to distinguish global affine correction from high-frequency residual refinement. Removing source-aware guidance weakens structural preservation and source

Method	DroneVehicle					MFNet					RoadScene				
	AG $\uparrow$	SCD $\uparrow$	SF $\uparrow$	HD95 $\downarrow$	ASSD $\downarrow$	AG $\uparrow$	SCD $\uparrow$	SF $\uparrow$	HD95 $\downarrow$	ASSD $\downarrow$	AG $\uparrow$	SCD $\uparrow$	SF $\uparrow$	HD95 $\downarrow$	ASSD $\downarrow$
SuperFusion	6.11	0.91	17.50	<u>30.12</u>	<u>7.55</u>	2.96	1.06	7.90	63.83	14.49	4.33	1.21	11.67	44.43	<u>10.38</u>
ReCoNet	5.20	1.48	13.60	30.53	7.71	<u>3.15</u>	1.53	9.51	69.72	17.71	3.65	<u>1.27</u>	9.02	45.72	11.57
MURF	2.34	0.89	5.72	44.95	12.06	3.02	<u>1.36</u>	<u>10.47</u>	<u>30.47</u>	<u>6.77</u>	3.49	<u>1.27</u>	9.79	<u>39.36</u>	8.43
UMF-CMGR	4.28	1.02	10.96	31.73	8.93	1.81	1.07	5.25	100.17	39.25	2.90	0.61	3.86	65.27	22.19
IMF	4.11	1.01	8.44	30.97	8.62	1.23	0.77	3.77	68.03	25.70	3.50	1.25	7.43	56.76	18.32
AU-Net	<u>6.15</u>	1.01	18.16	31.34	7.76	2.96	1.22	9.05	77.34	20.34	3.77	1.26	9.68	43.34	10.47
C2RF	5.66	0.82	<u>18.42</u>	35.24	9.63	1.63	0.41	4.48	92.36	24.41	<u>4.56</u>	1.19	<u>13.42</u>	44.06	11.10
FusionRegister	5.54	1.08	14.95	31.10	7.72	3.10	1.04	8.87	30.76	12.74	3.88	1.22	10.70	51.13	20.16
SAGE	6.18	<u>1.13</u>	18.73	29.19	7.42	3.17	1.28	11.49	21.65	4.24	5.18	1.28	15.05	38.52	14.28

Table 1: Quantitative comparison with SOTA methods. The **bold/underline** indicates the best and runner-up.

Figure 4: Qualitative ablation results on MFNet.

consistency. Disabling either alignment stage introduces greater misalignment, confirming the complementary roles of global geometry estimation and local residual correction. Removing guided subband fusion reduces detail preservation and disrupts the coordinated aggregation of aligned frequency information. The visual comparisons in Figure 4 support these findings. Variants without either alignment stage exhibit more visible ghosting and contour displacement around the vehicle. Removing source-aware guidance or guided fusion produces blurrier details and less balanced source integration. In contrast, the complete model preserves sharper boundaries, more coherent structures, and a better balance between thermal saliency and visible appearance. These results validate the coordinated contributions of frequency guidance, hierarchical alignment, and subband fusion.

4.4 Effect of Guidance Propagation

Figure 5 compares guidance delivered to low frequency alignment, high frequency alignment, fusion, and all stages with the no guidance reference. Using identical test pairs and misalignment conditions, we vary only the recipient stages. For each metric, the paired difference δ_i is oriented so that positive values indicate improvement and standardized by its cross-sample deviation as $d_z = \bar{\delta}/s_\delta$.

Alignment. Low frequency guidance produces a broad LL response but degrades HF alignment, while high frequency guidance improves the HF response and fusion guidance leaves both measures near the baseline. Guidance across all stages yields the strongest positive HF response while maintaining the LL response near the reference, supporting the coarse to fine dependency in HFCA.

Fusion. Fusion guidance improves source mutual information, edge overlap, and gradient fidelity without materially changing alignment, while high frequency guidance also benefits these measures and low frequency guidance degrades the edge and gradient responses. Guidance across all stages maintains positive mean effects for all three indicators together with the strongest HF response, supporting coordinated propagation throughout SAGE.

Methods	AG $\uparrow$	SCD $\uparrow$	SF $\uparrow$	HD95 $\downarrow$	ASSD $\downarrow$
w/o SAFE	2.72	1.19	9.86	29.46	5.82
w/o LFA	3.09	1.36	7.97	35.40	7.09
w/o HCR	2.68	1.14	8.58	36.20	7.36
w/o GSF	2.76	1.35	8.11	35.08	6.85
SAGE	3.17	1.28	11.49	21.65	4.24

Table 2: Ablation results on MFNet. LFA, HCR, and GSF denote low-frequency affine alignment, high-frequency correlation refinement, and guided subband fusion, respectively.

Methods	Recall $\uparrow$	Precision $\uparrow$	mAP@0.5 $\uparrow$
SuperFusion	23.40	14.11	11.43
ReCoNet	24.59	13.28	10.31
MURF	<u>27.51</u>	11.17	10.69
UMF-CMGR	26.94	11.96	10.80
IMF	21.85	11.67	8.83
AU-Net	24.87	12.54	10.56
C2RF	18.62	10.06	9.38
FusionRegister	23.51	<u>16.23</u>	<u>12.00</u>
SAGE	33.63	19.31	16.04

Table 3: Quantitative object detection results on DroneVehicle. The **bold/underline** indicates the best and runner-up.

4.5 Frequency Specialized Hierarchical Collaboration

Low frequency components mainly encode global scene structure, whereas high frequency components preserve edges and local details that are more sensitive to spatial misalignment. We therefore compare three alignment configurations under the same evaluation setting. LFA+LFA applies global affine alignment to both frequency branches, HCR+HCR applies local correlation refinement to both branches, and SAGE assigns LFA to the low frequency branch and HCR to the high frequency branch. As shown in Figure 6, SAGE achieves the strongest high frequency alignment improvement while preserving the highest low frequency correspondence. LFA+LFA provides limited improvement because a global transformation cannot adequately correct spatially varying high frequency displacement. HCR+HCR performs better but requires local refinement to compensate for both global and local misalignment. In contrast, SAGE first establishes reliable global correspondence in the low frequency branch and then concentrates high frequency refinement on residual local errors. These results demonstrate that frequency specific branch assignment provides more effective and coherent alignment than applying a single mechanism to both frequency components.

4.6 Downstream Task Evaluation

To evaluate the downstream benefits of alignment and fusion, we train and test YOLO11m [35] on the unaligned DroneVehicle dataset using fusion results from different methods. Recall, precision, and mAP@0.5 are used to assess

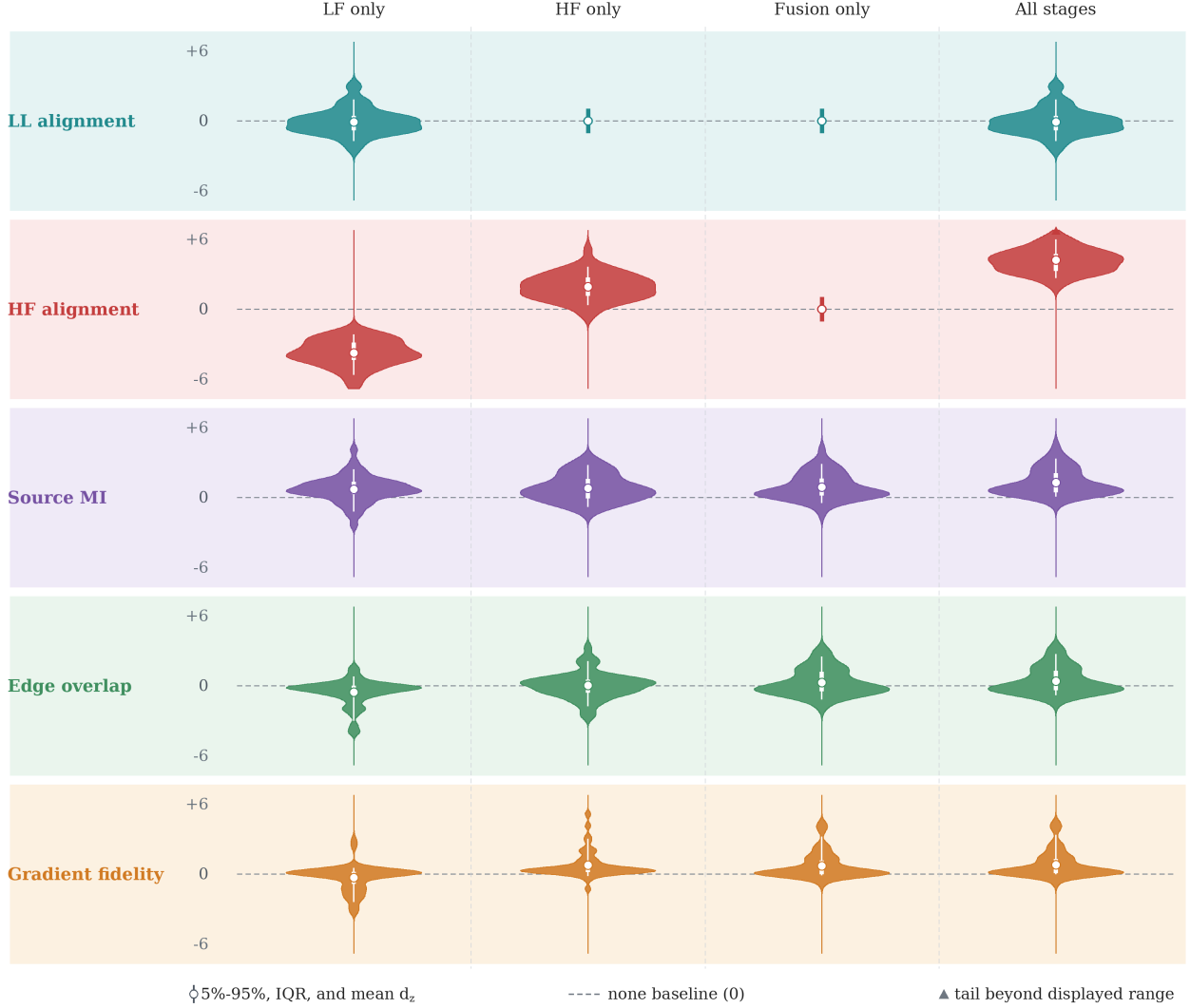


Figure 5: Effects of guidance propagation to low-frequency alignment, high-frequency alignment, fusion, or all stages.

target coverage, prediction reliability, and detection accuracy, respectively. As reported in Table 3, SAGE achieves the best performance across all metrics. The simultaneous improvements in recall and precision indicate that SAGE recovers more targets without sacrificing prediction reliability. Qualitative results in the supplementary material further show that SAGE improves the detection of small and densely distributed vehicles and reduces missed and inaccurate detections, demonstrating its effectiveness in preserving structural correspondence and complementary multimodal information for downstream detection.

4.7 Complexity Comparison

Table 4 compares the computational complexity and inference efficiency of SAGE with representative methods. SAGE requires 91.45G FLOPs and 3.53M parameters, maintaining moderate computational cost among the compared approaches. Despite jointly performing frequency equalization, hierarchical alignment, and subband fusion, SAGE achieves a runtime of 0.02s and 56.50 FPS, demonstrating competitive efficiency. These results indicate a favorable balance between model complexity and inference performance.

5 Conclusion

We presented SAGE, a unified frequency-domain framework for weakly registered RGB-T alignment and fusion. SAGE preserves source content and derives structural and gain guidance, HFCA couples global geometry with local

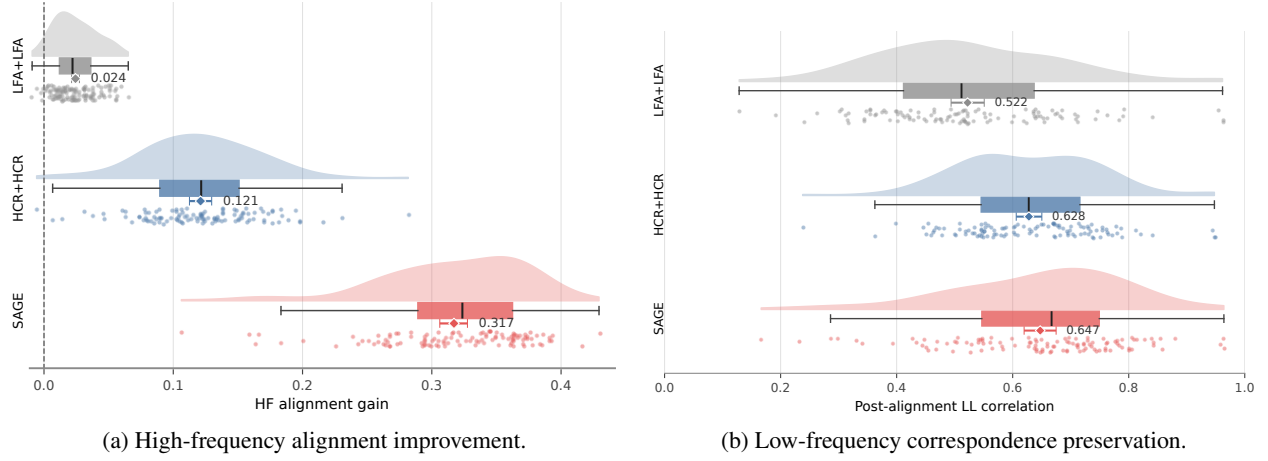


Figure 6: Comparison of collaboration strategies in (a) HF alignment gain and (b) post-alignment LL correlation.

Methods	FLOPs (G)	Parameters (M)	Time (s)	FPS
SuperFusion	16.36	1.96	0.03	38.69
ReCoNet	15.33	0.21	0.33	3.05
MURF	100.10	4.08	0.12	8.13
UMF-CMGR	131.38	14.44	0.09	10.69
IMF	123.30	15.70	0.06	17.07
AU-Net	14.04	2.27	0.02	60.93
C2RF	165.12	2.70	0.06	16.69
FusionRegister	49.79	0.77	0.08	12.93
SAGE	91.45	3.53	0.02	56.50

Table 4: Model complexity and inference efficiency.

residual refinement, and GSF integrates propagated cues and alignment reliability for subband fusion. Extensive experiments demonstrate robust performance across diverse scenes and deformation conditions. Guidance propagation and downstream detection results further confirm that SAGE maintains structural correspondence and task-relevant cross-modal information. Overall, SAGE effectively coordinates source preservation, geometric alignment, and frequency-specific fusion within a unified framework.

References

- [1] Timing Li, Bing Cao, Pengfei Zhu, and Kewen Li. Multimodal image registration for low altitude platforms: Methods, challenges, and future trends. *Tsinghua Science and Technology*, 2026.
- [2] Timing Li, Bing Cao, Jiahe Feng, Haifang Cao, Qinghua Hu, and Pengfei Zhu. Hyperbolic cycle alignment for infrared-visible image fusion. *IEEE Transactions on Image Processing*, 2026.
- [3] Haodong Zhu, Wenhao Dong, Linlin Yang, Hong Li, Yuguang Yang, Yangyang Ren, Qingcheng Zhu, Zichao Feng, Changbai Li, Shaohui Lin, et al. Wavemamba: Wavelet-driven mamba fusion for rgb-infrared object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11219–11229, 2025.
- [4] Zhinan Gao, Dongdong Li, Yangliu Kuai, Rui Chen, and Gongjian Wen. Visible-infrared image alignment for uavs: Benchmark and new baseline. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.

- [5] Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17882–17891, 2025.
- [6] Haifeng Zhong, Fan Tang, Zhuo Chen, Hyung Jin Chang, and Yixing Gao. Amdanet: Attention-driven multi-perspective discrepancy alignment for rgb-infrared image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10645–10655, 2025.
- [7] Timing Li, Bing Cao, Pengfei Zhu, Bin Xiao, and Qinghua Hu. Bi-directional self-registration for misaligned infrared-visible image fusion. *arXiv preprint arXiv:2505.06920*, 2025.
- [8] Xingyu Hu, Junjun Jiang, Chenyang Wang, Kui Jiang, Xianming Liu, and Jiayi Ma. Balancing task-invariant interaction and task-specific adaptation for unified image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11262–11272, 2025.
- [9] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*, 2022.
- [10] Di Wang, Jinyuan Liu, Long Ma, Risheng Liu, and Xin Fan. Improving misaligned multi-modality image fusion with one-stage progressive dense registration. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):10944–10958, 2024.
- [11] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19679–19688, 2022.
- [12] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):12148–12166, 2023.
- [13] Zhiying Jiang, Zengxi Zhang, and Jinyuan Liu. Harmonized domain enabled alternate search for infrared and visible image alignment. *IEEE Transactions on Image Processing*, 2025.
- [14] Ming Lu, Min Jiang, Xuefeng Tao, and Jun Kong. Au-net: Adaptive unified network for joint multi-modal image registration and fusion. *IEEE Transactions on Image Processing*, 2025.
- [15] Xingyuan Li, Haoyuan Xu, Xingyue Zhu, Jun Ma, Yang Zou, Zhiying Jiang, and Jinyuan Liu. Uncertainty-aware spatial-frequency registration and fusion for infrared and visible images. *arXiv preprint arXiv:2605.13049*, 2026.
- [16] Linfeng Tang, Qinglong Yan, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *International journal of computer vision*, 133(8):5262–5280, 2025.
- [17] Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3673–3690, 2025.
- [18] Song Wang, Xie Han, Liqun Kuang, Boying Wang, Zhongyu Chen, Zherui Qiao, Fan Yang, Xiaoxia Liu, Bingyu Zhang, and Zhixun Wang. The source image is the best attention for infrared and visible image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13513–13522, 2025.
- [19] Weichen Dai, Hexing Wu, Xiaoyang Weng, Yuxin Zheng, Yuhang Ming, and Wanzeng Kong. Multi-modal synergistic implicit image enhancement for efficient optical flow estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2173–2182, 2025.
- [20] Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7457–7468, 2025.
- [21] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2226–2235, 2025.
- [22] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25912–25921, 2024.
- [23] Yiming Sun, Xin Li, Pengfei Zhu, Qinghua Hu, Dongwei Ren, Huiying Xu, and Xinzhong Zhu. Task-gated multi-expert collaboration network for degraded multi-modal image fusion. In *International Conference on Machine Learning*, pages 57571–57586. PMLR, 2025.

- [24] Lin Guo, Xiaoqing Luo, Wei Xie, Zhancheng Zhang, Hui Li, Rui Wang, Zhenhua Feng, and Xiaoning Song. Revisiting generative infrared and visible image fusion based on human cognitive laws. *Advances in Neural Information Processing Systems*, 38:96322–96352, 2026.
- [25] Tianwei Guan, Haozhen Wei, Yuhao Zhou, Jun Ma, Zecheng Xu, Zhiying Jiang, Jinyuan Liu, and Xingyuan Li. Domain adaptation guided infrared and visible image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4376–4384, 2026.
- [26] Jianghan Xia, Hong Song, Jinfu Li, Yucong Lin, Shihan Ma, Jingfan Fan, Danni Ai, Tianyu Fu, Deqiang Xiao, and Jian Yang. Regionfuse: Region-adaptive pixel distribution learning for infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19539–19548, 2026.
- [27] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14816–14826, 2021.
- [28] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023.
- [29] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE TCSVT*, 32(10):6700–6713, 2022.
- [30] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.
- [31] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. FusionDn: A unified densely connected network for image fusion. In *proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.
- [32] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.
- [33] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European conference on computer Vision*, pages 539–555. Springer, 2022.
- [34] Congcong Bian, Haolong Ma, Hui Li, Zhongwei Shen, Xiaoqing Luo, Xiaoning Song, and Xiao-jun Wu. Fusionregister: Every infrared and visible image fusion deserves registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41551–41561, 2026.
- [35] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023.