AerialFusion: Co-Motion-Driven Unified Registration and Fusion on Multi-modal Data Streams from Aerial View

Junhui Qiu [1,3*], **Xiang Xiang** [1,2,3*†], **Hongyun Wang** [1,3], **Jiaqi Gui** [1,3]

[1] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
[2] School of Computer Science and Technology, Huazhong University of Science and Technology, China
[3] HUST AI and Visual Learning Lab (HAIV Lab), Huazhong University of Science and Technology (HUST), China

Abstract

Aerial multi-modal visual streams registration and fusion can generate more comprehensive scene information representations for UAVs' cross-modal perception. However, current challenges lie primarily in the essential difficulty of joint spatiotemporal representation learning from dynamic background and moving targets, and a critical shortage exists in large-scale, well-annotated multi-modal visual streams benchmark for UAV platforms. In this paper, we propose ***AerialFusion***, a co-motion-driven unified UAVs visual streams registration and fusion that fully mines modality-invariant common features based on motion-aware, enabling spatiotemporally coherent registration and fusion. Specifically, 1) a Skewed Motion Distribution Field Co-Motion-Driven Image Registration, 2) a Co-Motion Generative Fusion, 3) a Streams-based Unified Learning. Furthermore, we introduce ***EUM3D***, a registration and fusion benchmark for UAVs cross-modal perception. This benchmark contains 60 synchronized visible-infrared visual streams, or 122k spatially and temporally aligned pairs, most of which were taken at low-light scenes. And EUM3D provides pixel-level alignment guarantees via perspective-transform ground-truth. Extensive experiments reveal that ***AerialFusion*** surpasses current focus on image and static background fusion methods in aerial sequence scenarios, addressing spatiotemporal mismatches while suppressing cross-modal interference.

Code — https://github.com/HAIV-Lab/AerialFusion

Introduction

Multi-modal fusion aims to address the limited robustness of single-modal representations by integrating complementary information from diverse sensors (Wu et al. 2025). Typically, visible and infrared image fusion is an effective way: visible light provides detailed texture and color cues (Zhao et al. 2025), while infrared provides reliable thermal and contour features (Jie et al. 2024). Thus, it establishes a visual foundation for cross-modal sensing in intelligent systems.

With the growing demand for aerial applications, which can generate substantial momentum for civilian economies and military technological advancement through downstream applications (Xing et al. 2025), fusion technologies have expanded their operational domains from terrestrial to aerial platforms (Gong et al. 2025). When applied to real-world aerial tasks, these systems face challenges such as visual streams containing dynamic background and moving targets, as opposed to static images (Zhang et al. 2025c); parallax effects arising from divergent optical axes (Zhang et al. 2025a), and data-driven algorithmic framework lacking corresponding source data provisioning.

In early research, researchers traditionally treated parallax effects, registration, and modality reinforcement fusion as separate research domains. Firstly, feature extrac-

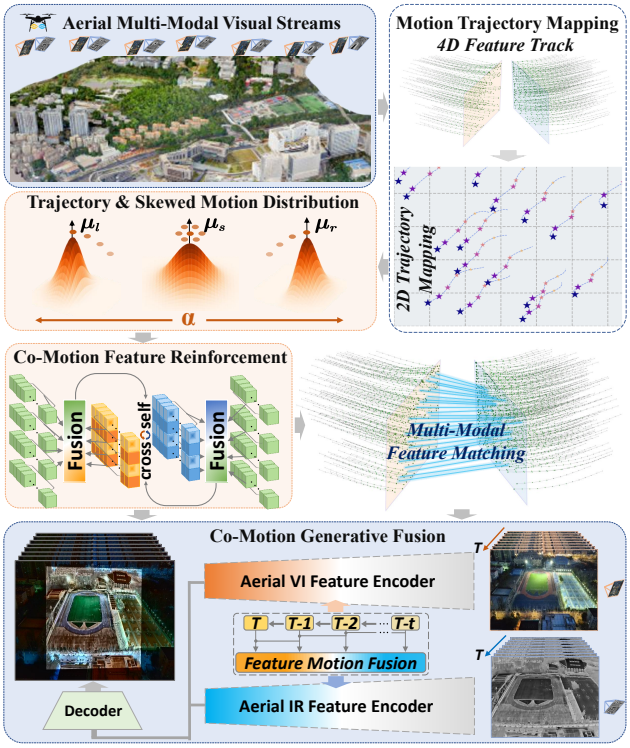

Figure 1: The framework. We map modality-invariant motion features into skewed distribution probability descriptor, then enhance motion characterization by fusing them with static descriptors to enable cross-modal communication.

*Equal contribution; co-first author.
†Correspondence to xex@hust.edu.cn; also w/ Peng Cheng Lab.

tion and descriptor matching, and transformation as the registration pipeline. The traditional methods SIFT (Lowe 2004), ORB (Rublee et al. 2011) and the learnable methods SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018), ALIKED (Zhao et al. 2023), XFeat (Potje et al. 2024), RDD (Chen et al. 2025). The learnable matchers SuperGlue (Sarlin et al. 2020), LightGlue (Lindenberger, Sarlin, and Pollefeys 2023) and SemaGlue (Zhang et al. 2025b). Then, image fusion has evolved, progressing from based on GANs methods FusionGAN (Ma et al. 2019), U2Fusion (Xu et al. 2020), AT-GAN (Rao et al. 2023), to Transformer methods SwinFusion (Ma et al. 2022), to Diffusion methods Diff-IF (Yi et al. 2024), to Mamba methods FusionMamba (Xie et al. 2024b), STFMamba (Zhao, Jiang, and Huang 2025), achieving continuous improvement in fusion performance.

Thus, registration and fusion have been coupled, and giving rise to a series of end-to-end methods. MURF (Xu, Yuan, and Ma 2023) achieves high-quality novel view synthesis from sparse inputs via multi-view geometric constraints and depth consistency optimization. C2RF (Tang et al. 2025b) achieves image registration and fusion through mutually-enhanced commonality mining and contrastive learning.

Recently, the continuous surveillance scenarios Unified motheds were explored, and the scene background remains nearly static. RCVS (Xie et al. 2024a) achieves stable video sequence registration through a spatiotemporal calibration module, followed by a fast fusion to deliver stable fused video streams. VideoFusion (Tang et al. 2025a) harnesses cross-modal complementarity and temporal dynamics to generate coherent videos from potentially degraded inputs, modality-guided adaptive fusion.

In this paper, we focus on an aerial scene that has a dynamic background and moving targets. As show in Fig.1, we propose AerialFusion, a co-motion-driven unified registration and fusion for embodied UAVs visual streams, which consists of four essential parts: (a) **Skewed Motion Distribution Field (SMDF)** that probabilistic features motion and reinforces feature descriptors, (b) **Co-Motion Generative Fusion** that through multi-modal motion feature communication constructs fusion, (c) **Streams-based Unified Learning** that learning continuous feature communication and smooth transformation, (d) **Embodied UAV Motion Multi-Modal Dataset (EUM3D)** that provides aerial scene dataset support. The main contributions of our work are:

- We propose AerialFusion, a co-motion-driven unified registration and fusion for aerial-to-ground visual streams, by constructing a skewed motion distribution field to extract modality-invariant motion features, achieve stable registration and fusion.
- We propose a stream-based unified learning struction, for continuous multi-modal visual streams in dynamic scenes, we learn spatiotemporally smooth transformations and fusion to mitigate discrete data artifacts.
- We construct EUM3D dataset, a high-quality spatiotemporal aligned embodied UAVs visual streams dataset, which contains a large number of day-night visible-infrared visual stream pairs capturing identical scenes.

Related Works

End-to-end Registration and Fusion

To address the inherent parallax in dual-modal imaging systems, recent work integrates registration as a prerequisite, combining it with fusion to form a unified framework. This approach leads to learning strategies where registration and fusion mutually enhance each other. UMF-CMGR (Wang et al. 2022) through convert visible images to pseudo-infrared, to achieve their fusion, but numerous bottlenecks arise in real-world. SuperFusion (Tang et al. 2022a) couples high-level vision tasks with registration and fusion in a unified framework, due to the optical-flow registration for parallax, the performance degrades significantly under large-angle parallax. MURF (Xu, Yuan, and Ma 2023) enables high-quality sparse-view synthesis through multiview geometry and depth optimization, but global optimization fails to achieve precise artifact removal in semantically rich regions. Thus, SemLA (Xie et al. 2023) and RA-MMIR (Qiu et al. 2024) represent semantically rich regions through feature representations for registration and fusion. RA-MMIR constructs a robust and adaptive attention module based on semantic features, guiding focus to extract and match features in semantically rich regions. C2RF (Tang et al. 2025b) enables multi-modal registration and fusion via commonality mining and contrastive learning.

For dynamic scenes, the stable registration and fusion of surveillance video have also been investigated. RCVS (Xie et al. 2024a) stabilizes video registration via robust matching and spatiotemporal calibration, coupled with lightweight fusion for steady output. VideoFusion (Tang et al. 2025a) leverages cross-modal dynamics and temporal coherence through differential reinforcement, adaptive fusion, and bi-temporal attention. However, they optimize registration and fusion of static image and static background video, ignoring the need for dynamic background and moving targets processing in real-world embodied intelligence applications.

Multi-Modal Aerial Dataset

Novel UAV datasets spanning multi-modal tracking Anti-UAV (Jiang et al. 2021) and VTUAV (Zhang et al. 2022), cross-modal generation AVIID (Han et al. 2023), and multi-spectral analysis MUST have significantly propelled drone vision research. The Anti-UAV dataset has 418 visible-infrared video pairs introducing state accuracy for drone tracking under occlusion and fast motion. Its successor VTUAV includes 500 videos and 1.7M frames, enabling short and long-term tracking and segmentation via hierarchical multi-modal fusion. AVIID advances cross-modal generation with 127 visible-infrared video pairs, while MUST (Qin et al. 2025) pioneers multispectral tracking (250 videos, 8-band 390–950nm) using a spectral-temporal framework.

However, the aerial datasets for visual stream registration and fusion tasks remain severely limited. The primary challenge lies in the strict requirements for precise temporal consistency and spatial alignment in data-driven registration and fusion pipelines.

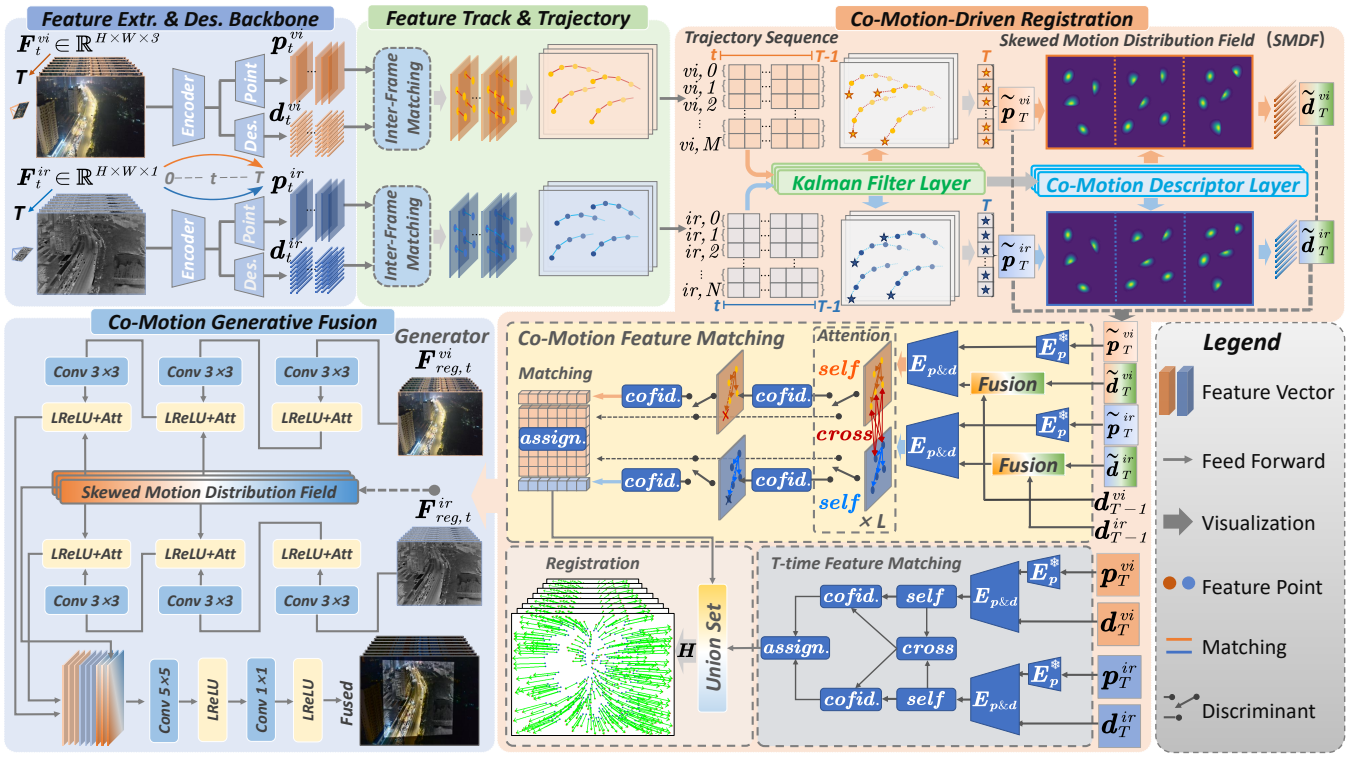

Figure 2: The framework of AerialFusion. Our method comprises four main components: Feature Extraction & Description Backbone, Feature Motion & Trajectory, Co-Motion-Driven Frame Registration, Co-Motion Generative Fusion.

Methodology

Given aerial multi-modal visual streams V_{UAV}^{vi} and V_{UAV}^{ir}, the scene contains both dynamic background and moving targets. Our work is to generate a stable fusion stream V_{UAV}^{fus}.

The AerialFusion workflow is shown in Fig.2. Initially, we extract keypoints and descriptions from frame sequences. Subsequently, getting the motion trajectories through inter-frame matching. Then, we build the Skewed Motion Distribution Field (SMDF) module to reinforce the feature. Ultimately, the SMDF module leads Co-Motion Fusion.

Feature Extraction and Trajectory

This section introduces the feature extraction and description, feature motion tracking, and trajectory generation.

Feature Extraction Backbone. Acquiring the frame sequences $\{\mathcal{F}_t^{vi}, \mathcal{F}_t^{ir}\}_{t=0}^{T}$ from $\{V_{UAV}^{vi}, V_{UAV}^{ir}\}$, where $\mathcal{F}_t^{vi} \in \mathbb{R}^{H\times W\times 3}$ and $\mathcal{F}_t^{ir} \in \mathbb{R}^{H\times W\times 1}$, where H, W are the width and height. We use VGG (DeTone, Malisiewicz, and Rabinovich 2018) for the feature extract and descriptor sequences $\{\boldsymbol{p}_t^{vi}, \boldsymbol{d}_t^{vi}\}_{t=0}^{T}$ and $\{\boldsymbol{p}_t^{ir}, \boldsymbol{d}_t^{ir}\}_{t=0}^{T}$.

Feature Motion Track & Trajectory. Through matching we find inter-frame feature correspondence, the feature trajectory set $\boldsymbol{\psi}^{vi} = \{\boldsymbol{\psi}_k^{vi}\}_{k=1}^{K}$ and $\boldsymbol{\psi}^{ir} = \{\boldsymbol{\psi}_q^{ir}\}_{q=1}^{Q}$ comprises inter-frame features maintain bijective correspondence $\forall t \in [t, T-1]$, where $\boldsymbol{\psi} = \{p_t, p_{t+1}, \cdots, p_{T-1}\}$.

Co-Motion-Driven Frame Registration

To achieve smooth transitions of keypoint positions at the current moment, we use the Kalman filter-based prediction (van der Zee et al. 2025). The expression is given by:

$$\tilde{\boldsymbol{p}}_T = \boldsymbol{\eta}\boldsymbol{\Lambda}^{T-1}\hat{\mathbf{I}}_t + \boldsymbol{\eta}\sum_{n=t}^{T-1}\boldsymbol{\Lambda}^{T-n}\mathbf{K}_n\left(\boldsymbol{\psi}_n - \boldsymbol{\eta}\boldsymbol{\Lambda}\hat{\mathbf{I}}_n\right) \quad (1)$$

where $\tilde{\boldsymbol{p}}_T$ is predicted features at T, $\boldsymbol{\psi}_n$ is measurement vector at t, $\boldsymbol{\eta}$ is observation matrix, $\boldsymbol{\Lambda}$ is state transition matrix, $\hat{\mathbf{I}}_t$ is initial estimate at t, $\mathbf{K}_n$ is Kalman gain matrix at n.

Skewed Motion Distribution Field Module. The $\boldsymbol{\psi}^{vi}$ and $\boldsymbol{\psi}^{ir}$ deliver an important inherent proper is motion tendency. The motion tendency (motion or stasis) of features directly affects the stability of registration.

In SMDF module, we build Skewed Normal Distribution (SND) to mathematically model both feature movement and stillness. Firstly, quantify the motion tendency per keypoint:

$$\nabla\alpha_n^m = \sum_{\zeta=1}^{t}\pi_\zeta\nabla\left(p_{n,T-\zeta}^m, \tilde{p}_{n,T}^m\right) \quad (2)$$

where $\nabla\alpha_n^m \in \mathbb{R}^{1\times 2}$, and $\{m, n\} \in \{vi, ir\} \times \{K, Q\}$, $\nabla(\cdot,\cdot)$ is matrix subtraction, $p_{n,T-1}^m$ is the position of p_n^m at time $T-1$, $\tilde{p}_{n,T}^m$ is the predicted features at time T, $\sum_{\zeta=1}^{t}\pi_\zeta = 1$ the weight values exhibit exponential growth.

Then, building a probabilistic model based on SND for feature motion tendency. For k and q features motion ten-

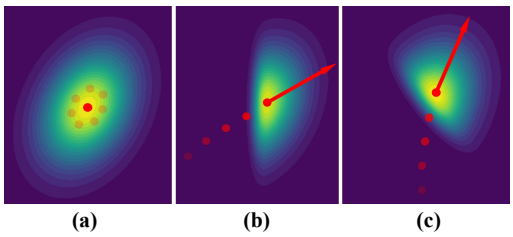

Figure 3: The futures SMDF integrates spatiotemporal feature effect display. • : feature locations over the interval $[t, T-1]$. (a) implies that features remain quasi-stationary over the interval. (b) and (c) mean feature stay motion state.

dencies of $\nabla\boldsymbol{\alpha}^{vi}$ and $\nabla\boldsymbol{\alpha}^{ir}$, we set each SND mutually independent. The SND formula for feature motion tendency:

$$P\left(\boldsymbol{F}_T^m\right) = \Phi\left(\boldsymbol{F}_T^m; \tilde{\boldsymbol{p}}_T^m, \boldsymbol{\varepsilon}\right) \cdot \phi\left[\nabla\boldsymbol{\alpha}^m \boldsymbol{\gamma}\left(\boldsymbol{F}_T^m - \tilde{\boldsymbol{p}}_T^m\right)\right] \quad (3)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are univariate and bivariate normal distribution, $\boldsymbol{F}_T^m$ as bivariate random variable, $\tilde{\boldsymbol{p}}_T^m$ is predicted keypoint position vector at time T, $\nabla\boldsymbol{\alpha}^m$ as the bivariate skewness parameter vector. For possible values of $\nabla\boldsymbol{\alpha}^m$, we conduct a plausibility analysis on $P\left(\boldsymbol{F}_T^m\right)$:

$$P\left(\mathcal{F}_{n,T}^m\right) \propto \begin{cases} \Phi\left(\mathcal{F}_{n,T}^m; \tilde{p}_{n,T}^m, \varepsilon_n\right) & , \nabla\alpha_n^m \to [0,0] \\ \Phi\left(\cdot\right) \cdot \phi\left[\nabla\alpha_n^m \Omega_n\right] & , \nabla\alpha_n^m \nrightarrow [0,0] \end{cases} \quad (4)$$

where the Ω_n is parameters computation $\gamma_n\left(\mathcal{F}_{n,T}^m - \tilde{p}_{n,T}^m\right)$.

In the physical world, the $\nabla\alpha_n^m \nrightarrow [0,0]$ means feature stay motion state over the interval $[t, T-1]$, and satisfying $sup_{\tau\in[t,T-1]} \left\|p_{n,\tau}^m - \tilde{p}_{n,T}^m\right\|_2 > \epsilon$. Similarly, $\nabla\alpha_n^m \to [0,0]$ means feature remains quasi-stationary. The $P\left(\mathcal{F}_{n,T}^m\right)$ serves as a feature spatiotemporal integrated probability function, with performance demonstrated in Fig.3.

Modality-invariant Motion Feature Reinforcement. We construct modality-invariant descriptors by uniformly sampling local neighborhoods of $P(\mathcal{F}_{n,T}^m; \tilde{p}_{n,T}^m, \nabla\alpha_{n,T}^m)$:

$$\tilde{d}_{n,T}^m \leftarrow \left\{ P\left(p; \tilde{p}_{n,T}^m, \nabla\alpha_{n,T}^m\right) \mid \left\|p - \tilde{p}_{n,T}^m\right\|_2 \leqslant R \right\} \quad (5)$$

where $\tilde{d}_{n,T}^m \in \mathbb{R}^{16\times16}$, $p = (x, y) \in [0, W] \times [0, H]$ is the coordinate, R is the radius of circular sampling region. Then, we linearly combine $\tilde{d}_{n,T}^m$ with $d_{n,T-1}^m$ to obtain modality-invariant motion feature reinforcement descriptors $\hat{d}_{n,T}^m$:

$$\hat{d}_{n,T}^m \leftarrow d_{n,T-1}^m + \mathbf{Flat}\left(\tilde{d}_{n,T}^m\right) \quad (6)$$

where $\mathbf{Flat}(\cdot)$ is feature flattening operation. Similarly to LightGlue (Lindenberger, Sarlin, and Pollefeys 2023), we encode $\hat{\boldsymbol{d}}_T^m$ and $\boldsymbol{p}_T^m$ through MLP, and construct $\mathcal{G}_{\text{self}}^m, \mathcal{G}_{\text{cross}}^m$:

$$\mathcal{T}_{\text{self}}^{\{vi,ir\}} = \mathcal{G}_{\text{self}}^m\left(\hat{\boldsymbol{d}}_T^m, \mathbf{MLP}\left(\tilde{\boldsymbol{p}}_T^m, \hat{\boldsymbol{d}}_T^m\right)\right) \quad (7)$$

$$\mathcal{T}_{\text{cross}}^{\{vi,ir\}} = \mathcal{G}_{\text{cross}}^m\left(\mathcal{T}_{\text{self}}^{vi}, \mathcal{T}_{\text{self}}^{ir}\right) \quad (8)$$

The assignment matrix $\mathbf{P}$ as:

$$\mathbf{P}_{ij} = \prod_m \sigma^m \cdot \underset{m\in vi}{\text{Soft}\max}\left(\mathcal{S}_{mj}\right) \cdot \underset{m\in ir}{\text{Soft}\max}\left(\mathcal{S}_{im}\right) \quad (9)$$

where $\mathcal{S}_{ij} = \text{Liner}\left(\mathcal{T}_{\text{cross}}^{vi}\right)^\top \cdot \text{Liner}\left(\mathcal{T}_{\text{cross}}^{ir}\right)$, and $\sigma^m = \text{Sigmoid}\left(\text{Liner}\left(\mathcal{T}_{\text{cross}}^m\right)\right)$.

Figure 4: Discriminator of Co-Motion Generative Fusion.

Co-Motion Generative Fusion

We use SMDF to modulate multi-modal fusion weights to suppress artifacts, including visible motion blur or infrared thermal trailing.

In Fig.2, the generator contains two branches for extracting visible and infrared features. Firstly, the registered frames $\boldsymbol{F}_{reg,t}^m \in \mathbb{R}^{H\times W\times C}$ undergo feature integration and obtain correlation coefficients to fuse infrared and visible frames through 3×3 convolution, getting $\left\{Q_m^\ell, K_m^\ell, V_m^\ell\right\} \in \mathbb{R}^{HW\times C'}$, and attention layer. The formulation of attention is as follows:

$$\mathbf{Att}_m^\ell = \text{Softmax}\left(\boldsymbol{E}_m^\ell / \sqrt{C_{\boldsymbol{Q}_m^\ell}}\right) \quad (10)$$

where $\{m, C\} \in \{vi, ir\} \times \{3, 1\}$, $\eta \in \{Q, K, V\}$, ℓ is operation count, $\boldsymbol{E}_m^\ell$ is the energy function. Then, the SMDF's feature $P\left(\boldsymbol{F}_T^m\right)$ interacts synergistically with $\mathbf{Att}_m^\ell$:

$$\mathbf{Att}_m^\ell \leftarrow \mathbf{Att}_m^\ell \oplus P\left(\boldsymbol{F}_T^m\right) \quad (11)$$

$$\mathcal{T}_f = \text{Concat}\left[\mathbf{Att}_{vi}^\ell, \mathbf{Att}_{ir}^\ell, P\left(\boldsymbol{F}_T^m\right)\right] \quad (12)$$

where $\oplus$ is the weighted feature fusion, $\text{Concat}(\cdot)$ is the channel-wise concatenation. As illustrated in Fig.4, we introduce the detailed workflow of the discriminator.

In Fig.4, the discriminator of Co-Motion Generation Fusion first goes through two 3×3 and one 5×5 convolutions, then through three FC Layers and $\text{Softmax}(\cdot)$, finally, training under the supervision of the loss functions.

Stream-based Learning

Unlike conventional unordered training paradigms, AerialFusion adopts a continuous-frame learning in visual streams, enabling spatiotemporal registration and fusion learning.

Registration Loss Function. For continuous frame registration, we use a smooth transformation $\mathbf{H}_{GT}^t$ as ground-truth. This offers two key advantages: 1) correcting feature matching errors in spatial transforms, 2) avoiding the requirement for extensive truth depth. The loss of registration $\mathcal{L}_{reg}$:

$$\begin{aligned} \mathcal{L}_{reg} = & \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \left[\frac{1}{\mathcal{M}_t} \sum_{\mathcal{M}_t} \log \mathbf{P}_{ij}^t \right. \\ & \left. + \sum_{m\in\{vi,ir\}} \frac{1}{2|\mathcal{F}_t^m|} \sum_{\mathcal{M}_t} \log\left(1 - \sigma_t^m\right) \right] \end{aligned} \quad (13)$$

Figure 5: Visualization of various scenarios in our EUM3D dataset.

where $\mathcal{M}_t = \left\{(k,q) \,|\, ||\boldsymbol{p}_{t,k}^{vi} \cdot \mathbf{H}_{GT}^{t} - \boldsymbol{p}_{t,q}^{ir}||_2 \leqslant \epsilon, \epsilon = 6\right\}$, $t_2 - t_1$ is the each batch size, other parameters are defined by the following Eq.9.

Fusion Loss Function. The generator loss directly measures how closely the synthesized samples approximate the desired data characteristics. The generator loss $\mathcal{L}_g$:

$$\mathcal{L}_g = \mathcal{L}_{SMDF} + \mathcal{L}_{SSIM} + \mathcal{L}_{content} + \mathcal{L}_{adv} \quad (14)$$

The SMDF loss $\mathcal{L}_{SMDF}$ is seted as:

$$\mathcal{L}_{SMDF} = \lambda \cdot G_{\mathcal{T}_f, \mathbf{Att}_{vi}} + (1-\lambda) \cdot G_{\mathcal{T}_f, \mathbf{Att}_{ir}} \quad (15)$$

where $G_{\mathcal{F}_1,\mathcal{F}_2} = ||\nabla\mathcal{F}_1 - \nabla\mathcal{F}_2||_2$, the ∇ is denoted to calculate the image gradient by Laplace filter.

The structure loss $\mathcal{L}_{SSIM}$ and content loss $\mathcal{L}_{content}$ are the same as (Rao et al. 2023). Adverse loss $\mathcal{L}_{adv}$:

$$\mathcal{L}_{adv} = ||D\left(G_{\mathcal{F}_T^{vi}, \mathcal{F}_T^{ir}}\right) - \left(P_T^{vi} - P_T^{ir}\right)||_2 \quad (16)$$

where the discriminator $D(\cdot)$ is a binary classifier: the P_T^{vi} and P_T^{ir}, where each element represents the likelihood of visible or infrared domains.

For discriminator loss $\mathcal{L}_d$, which aims at accurate classification of infrared and visible frames, it can be expressed as $\mathcal{L}_d = \mathcal{L}_{d_1} + \mathcal{L}_{d_2}$:

$$\mathcal{L}_{d_1} = \left(D\left(\mathcal{F}_T^{vi}\right)[1] - c\right)^2 + \left(D\left(\mathcal{F}_T^{ir}\right)[0] - c\right)^2 \quad (17)$$

$$\mathcal{L}_{d_2} = \left(D\left(G_{\mathcal{F}}\right)[1] - d\right)^2 + \left(D\left(G_{\mathcal{F}}\right)[0] - d\right)^2 \quad (18)$$

where $D\left(\mathcal{F}_T^{vi}\right)[1]$ and $D\left(\mathcal{F}_T^{ir}\right)[0]$ stand for the prediction result of classification to the infrared and visible frames. During training, we design the parameters $c \to 1$ and $d \to 0$.

Embodied UAV Motion Multi-Modal Dataset

As show in Tab.1, we conducted a multi-attribute comparison between our EUM3D and other multi-modal datasets, including aerial and common scenes. For the common scene, MSRS (Tang et al. 2022b), M^{3}FD (Liu et al. 2022) and KAIST (Choi et al. 2018) provide numerous non-sequential image pairs. Recently, the M3SVD (Tang et al. 2025a) and HDO (Xie et al. 2024a) for continuous surveillance scenarios have been proposed, but there remains a lack of effective dataset for embodied motion. For the aerial scene, there remains a critical scarcity of aerial scene datasets with precise spatiotemporal annotations tailored for fusion research.

Therefore, we present an embodied UAV motion multi-modal dataset (EUM3D), which captures synchronized visible and infrared spectra via a dual-modal aerial platform.

Dataset	Med.Aer	Video	Challenging			
			MotB.	LowL.	OverE.	Occ.
Com Scene						
MSRS	—	—	✗	✓	✓	✗
M^{3}FD	—	—	✗	✓	✓	✓
LLVIP	—	15	✓	✓	✓	✗
TNO	—	3	✗	✓	✗	✓
INO	—	15	✗	✓	✓	✗
KAIST	—	—	✗	✓	✗	✗
HDO	—	24	✓	✓	✗	✗
M3SVD	—	220	✓	✓	✓	✓
Aerial Scene						
Anti-UAV	**YES**	418	✗	✓	✓	✗
VTUAV	NO	500	✗	✓	✗	✓
AVIID	**YES**	127	✓	✓	✓	✗
MUST	NO	250	✓	✗	✗	✓
EUM3D(ours)	**YES**	60	✓	✓	✓	✓

Table 1: Comparison of different aligned Multi-Modal common and aerial scene datasets.

Raw Data of EUM3D. The acquisition setup comprises: 1) a visible sensor (3840×2160 resolution, @30fps), and 2) an uncooled infrared sensor (8∼14μm spectral response, 640×512 resolution, ≈@30fps). The preplaned flight-path data collection setup comprises: 1) the flight altitude is primarily maintained 100∼500m, 2) the flight speed is primarily maintained 5∼15m/s, 3) the flight-path is primarily configured within the HUST area and the wider Wuhan region.

Ground-Truth of EUM3D. Firstly, we performed temporal calibration to achieve less 35ms error. Subsequently, we performed feature matching and transformation preprocessing to obtain the initial transformation $\mathbf{h}_{GT}^{t}$. Finally, for the $\mathbf{h}_{GT}^{t}$ still introduces pixel-level errors, we developed an adjustable transformation tool, that takes frame pairs $\left\{\mathcal{F}_t^{vi}, \mathcal{F}_t^{ir}\right\}$ and initial transformation $\mathbf{h}_{GT}^{t}$ as input to generate initial fusion, in which the transformation can be dynamically adjusted to achieve optimal transformation $\mathbf{H}_{GT}^{t}$.

Experiments

Relative Pose Estimation

Dataset. We evaluate the quality of correspondences estimated on MegaDepth-1500 (Li and Snavely 2018) for single-modal and EUM3D for multi-modal. MegaDepth-1500 contains 1500 photos, including camera poses, sparse point clouds, and rendered depth maps. For multi-modal experiment, we use EUM3D, the dataset includes 60 synchronized visible-infrared visual streams.

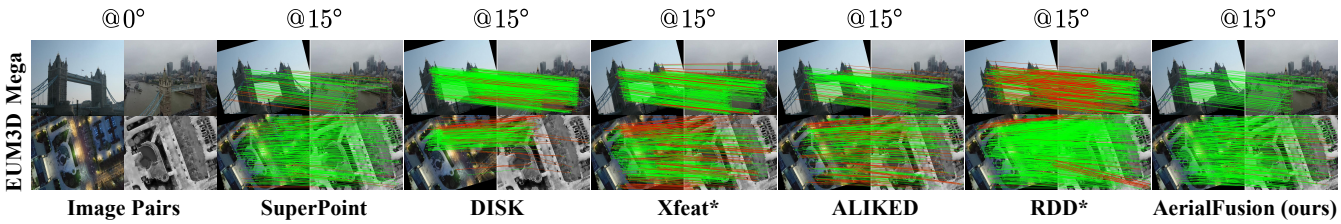

Figure 6: Visualization of elative pose estimation in MegaDepth-1500 and EUM3D dataset. – are inliers, – are outliers.

Method		MegaDepth-1500		EUM3D dataset	
		@5°	@15°	@5°	@15°
Dense					
DKM		60.42	82.46	45.53	62.32
RoMa		62.56	81.68	52.38	61.43
Semi-Dense					
LoFTR		52.69	75.35	43.64	59.53
ASpanFormer		55.32	81.34	46.84	61.34
Xfeat*		38.64	59.64	26.94	49.93
RDD*		51.54	75.46	40.35	65.78
Sparse with Learned Matcher					
SuperPoint	SuperGlue	49.73	80.60	56.63	58.25
SuperPoint	LightGlue	49.92	80.10	56.88	58.00
DISK	LightGlue	55.95	58.78	45.95	48.78
Dedode-V2-G	LightGlue	44.10	76.50	33.54	56.27
RDD	LightGlue	52.32	81.80	61.23	62.34
ALIKED	SemaGlue	53.56	80.49	48.99	62.79
AerialFusion (ours)		**63.08**	**82.80**	**62.62**	**69.97**

Table 2: Relative pose estimation on MageDepth and EUM3D datasets. AUC at @5° and @15° is reported.

Metrics and Comparing Methods. We report the AUC of recovered pose under thresholds of (5° and 15°). We use RANSAC (Fischler and Bolles 1981) to estimate the essential matrix. We compared AerialFusion against the state-of-the-art detector/descriptor methods on the MegaDepth and EUM3D datasets. For all dense and sparse methods, we use a resolution is set to 640×480 use the top 2048 features.
Results are presented in Tab.2 and visually in Fig.7. AerialFusion demonstrates superior performance in the dense methods (Edstedt et al. 2023, 2024; Sun et al. 2021; Chen et al. 2022), and sparse methods setting compared to the state-of-the-art learned methods (Tyszkiewicz, Fua, and Trulls 2020; Edstedt, Bökman, and Zhao 2024) combined with LightGlue, especially in continuous dynamic scenes.

Multi-modal Visual Streams Registration

Dataset. We conduct multi-modal visual streams registration experiments on spatiotemporally continuous datasets, primarily including: the $\text{LLVIP}_{\text{raw}}$ (Jia et al. 2021) for common scenes, the VTUAV dataset for low aerial scenes, and the EUM3D dataset for medium aerial scenes. The $\text{LLVIP}_{\text{raw}}$ dataset includes 15 visible-infrared video pairs without spatiotemporal registration, predominantly captured in nighttime street and sidewalk scenes. The VTUAV dataset is a multi-modal aerial tracking benchmark featuring 500 synchronized visible-infrared video pairs captured by drones. The EUM3D dataset was collected at medium aerial using planning flight paths, containing numerous day-night visible-infrared videos of identical scenes.
Metrics and Comparing Methods. We evaluate multi-modal visual streams registration performance using four quantitative metrics (Tang et al. 2025b): mean squared error (MSE), mean edge error (MEE), cross-correlation (CC), and structural similarity index (SSIM). For the metrics, we conducted comprehensive comparative experiments on the prevailing semi-dense and sparse methods. By establishing feature correspondences under large disparity conditions, we computed frame perspective transformations and derived quantitative metric results from the warped images.
Results are reported in Tab.3. AerialFusion performs better overall, particularly showing outstanding results on the more challenging nighttime multi-modal scenes.

Multi-modal Visual Streams Fusion

Dataset. Similar to multi-modal visual streams registration experiments, we conduct the fusion experiments on the $\text{LLVIP}_{\text{raw}}$, VTUAV, and EUM3D datasets.
Metrics and Comparing Methods. The evaluation employs four quantitative metrics to assess multi-modal fusion (Tang et al. 2025b,a): entropy (EN) quantifies information richness, standard deviation (SD) evaluates contrast and dynamic range, edge-based similarity ($Q_{AB/F}$) specifically evaluates edge feature transfer in fusion, and quality of structural similarity (Q_{SSIM}) combines structural and spectral preservation analysis. Based on the aforementioned metrics, we primarily compared state-of-the-art registration and fusion methods, including SuperFusion, UMF-CMGR, CrossRAFT (Zhou, Tan, and Yan 2022), MuRF, TIMFusion (Liu et al. 2024), and C2RF. All experiments are conducted using raw data captured without disparity preprocessing.
Results. The quantitative results are shown in Tab.3; we outperform most current popular end-to-end methods on the majority of metrics, and also achieve optimal performance in spatiotemporal stability. The qualitative results are shown in Fig.7, for registration, AerialFusion effectively minimizes artifacts in continuous scenes, for fusion, we achieve superior preservation of multi-modal textures and contours, providing a robust visual foundation.

Ablation Study

Dataset. To validate the contribution of each component, we conduct ablation studies by progressively removing/modifying 1) SMDF module, 2) streams-based learning registration

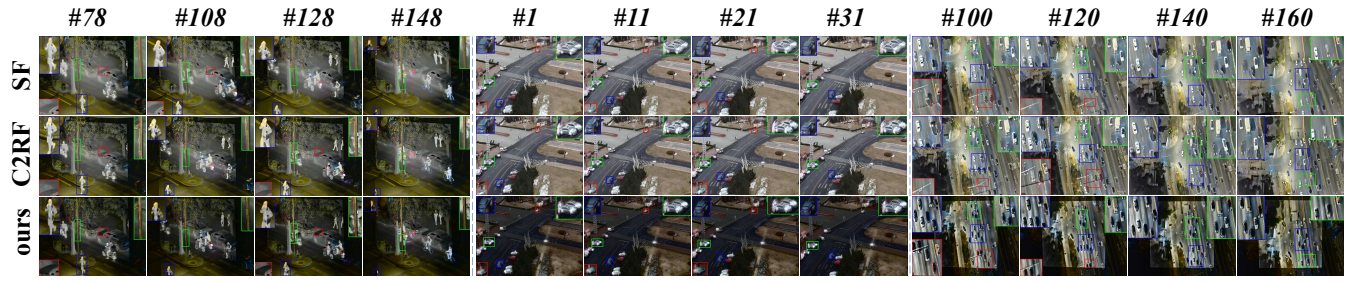

Figure 7: Visualization of multi-modal visual streams registration and fusion in LLVIP, VTUAV and EUM3D datasets.

Method		LLVIP dataset				VTUAV dataset				EUM3D dataset			
Registration		**MSE↓**	**MEE↓**	**CC↑**	**SSIM↑**	**MSE↓**	**MEE↓**	**CC↑**	**SSIM↑**	**MSE↓**	**MEE↓**	**CC↑**	**SSIM↑**
Semi-Dense													
LoFTR		117.62	97.80	0.14	0.40	101.57	62.23	0.28	0.09	22.01	95.72	0.52	0.16
ASpanFormer		426.63	96.27	0.15	0.38	67.72	86.17	0.45	0.11	267.29	144.37	0.59	0.18
Xfeat*		422.63	101.17	0.21	0.41	90.29	76.59	0.31	0.18	10.00	43.92	**0.64**	**0.20**
RDD*		167.78	136.23	0.56	0.42	84.65	71.81	0.34	0.15	17.21	66.88	0.57	0.18
Sparse with Learned Matcher													
SuperPoint	SuperGlue	176.86	172.18	0.28	0.44	63.75	54.10	0.28	0.11	13.88	53.95	0.60	0.18
SuperPoint	LightGlue	**88.28**	140.06	0.30	0.48	64.89	55.05	0.34	0.13	10.40	44.06	**0.64**	0.18
DeDoDe-V2-G	LightGlue	119.57	143.63	0.57	0.52	60.12	51.44	0.39	0.15	14.27	55.26	0.57	0.18
RDD	LightGlue	126.74	103.5	0.53	0.47	61.48	52.66	0.45	0.18	11.30	46.69	**0.64**	0.17
ALIKED	SemaGlue	134.78	105.64	0.52	0.43	59.25	**50.26**	0.42	0.16	12.34	48.23	0.62	0.19
AerialFusion (ours)		110.83	**92.78**	**0.62**	**0.57**	**56.43**	57.87	**0.56**	**0.22**	**9.98**	**43.86**	**0.64**	**0.20**
Unified Registration & Fusion		**EN↑**	**SD↑**	**$Q_{AB/F}$↑**	**Q_{SSIM}↑**	**EN↑**	**SD↑**	**$Q_{AB/F}$↑**	**Q_{SSIM}↑**	**EN↑**	**SD↑**	**$Q_{AB/F}$↑**	**Q_{SSIM}↑**
SuperFusion (SF)		7.305	41.058	0.358	0.456	7.608	45.492	0.673	0.646	7.676	54.941	**0.700**	0.678
UMF-CMGR		6.721	39.281	0.381	0.314	6.539	41.042	0.663	0.535	7.795	64.330	0.678	0.671
CrossRAFT		6.821	39.279	0.421	0.552	7.593	45.382	0.671	0.643	7.812	64.521	0.682	0.675
MURF		7.156	40.256	0.468	0.432	7.551	45.127	0.666	0.638	7.752	64.215	0.673	0.666
TIMFusion		7.367	42.731	0.536	0.638	7.577	**46.315**	0.669	0.641	7.281	41.218	0.512	0.635
C2RF		7.232	41.076	0.679	0.649	7.622	45.571	0.675	0.648	7.782	57.812	0.625	0.545
AerialFusion (ours)		**7.658**	**45.931**	**0.681**	**0.653**	**7.653**	45.921	**0.671**	**0.655**	**7.840**	**64.648**	0.686	**0.681**

Table 3: Multi-modal visual streams registration and unified registration & fusion on LLVIP, VTUAV and EUM3D datasets.

	Dataset	EUM3D dataset			
registration	**Metrics**	**MSE↓**	**MEE↓**	**CC↑**	**SSIM↑**
Modules	*w/o SMDF*	10.23	44.06	0.60	0.18
	w/o $\mathcal{L}_{reg}$	10.11	44.01	0.62	0.19
	w/o $\mathcal{L}_{SMDF}$	9.99	43.93	0.62	**0.20**
	AerialFusion	**9.98**	**43.86**	**0.64**	**0.20**
fusion	**Metrics**	**EN↑**	**SD↑**	**$Q_{AB/F}$↑**	**Q_{SSIM}↑**
Modules	*w/o SMDF*	5.48	55.95	0.43	0.46
	w/o $\mathcal{L}_{reg}$	6.99	63.36	0.58	0.59
	w/o $\mathcal{L}_{SMDF}$	7.72	63.98	0.62	0.61
	AerialFusion	**7.84**	**64.65**	**0.69**	**0.68**

Table 4: Ablation analyzes on EUM3D dataset.

loss $\mathcal{L}_{reg}$ module, 3) fusion loss $\mathcal{L}_{SMDF}$ module, and measuring the performance drop on EUM3D dataset.

Metrics and Comparing Methods. We conduct separate ablation analyses for the registration and fusion stages, with the metrics of each stage aligned with the experimental indicators from the preceding subsection.

Results. As shown in Tab.4, the full model achieves the best registration and fusion on EUM3D, outperforming all ablated variants by w/o SMDF, $\mathcal{L}_{reg}$, and $\mathcal{L}_{SMDF}$ on average. This validates the necessity of all proposed components.

Conclusion

In this paper, we propose AerialFusion, which leverages the co-motion of dual-spectral sensors to construct SMDF, that unifies the description of dynamic/static scene features into a modality-invariant representation via a SND, and enhances the descriptor through modality-invariant motion features, achieving the best spatiotemporally-unified registration and fusion in dynamic scenes with moving targets.

Acknowledgments

This work was supported by the HUST Interdisciplinary Research Support Program (2025JCYJ077), the project of Peng Cheng Lab (PCL2025AS214), the 2026 Optics-Valley Excellence Project funded by Nat'l Graduate College for Elite Engineers of HUST, and School of Computer Science and Technology, School of Artificial Intelligence and Automation, Hopcroft Center for Computing Science, and AI Institute. The contribution of Zichen Qiu from Wannan Medical College to the data processing of this paper is appreciated.

References

Chen, G.; Fu, T.; Chen, H.; Teng, W.; Xiao, H.; and Zhao, Y. 2025. RDD: Robust Feature Detector and Descriptor using Deformable Transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6394–6403.

Chen, H.; Luo, Z.; Zhou, L.; Tian, Y.; Zhen, M.; Fang, T.; Mckinnon, D.; Tsin, Y.; and Quan, L. 2022. Aspanformer: Detector-free image matching with adaptive span transformer. In *European conference on computer vision*, 20–36. Springer.

Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J. S.; An, K.; and Kweon, I. S. 2018. KAIST multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3): 934–948.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236.

Edstedt, J.; Athanasiadis, I.; Wadenbäck, M.; and Felsberg, M. 2023. DKM: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17765–17775.

Edstedt, J.; Bökman, G.; and Zhao, Z. 2024. Dedode v2: Analyzing and improving the dedode keypoint detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4245–4253.

Edstedt, J.; Sun, Q.; Bökman, G.; Wadenbäck, M.; and Felsberg, M. 2024. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19790–19800.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395.

Gong, X.; Luo, Y.; Chen, W.; Chang, Y.; Wan, Y.; Ma, A.; and Zhong, Y. 2025. BASHVS: A Multispectral and SAR Image Fusion Method Based on Bidirectional Aggregation of Saliency in Human Visual System. *IEEE Transactions on Geoscience and Remote Sensing*, 1–1.

Han, Z.; Zhang, Z.; Zhang, S.; Zhang, G.; and Mei, S. 2023. Aerial visible-to-infrared image translation: Dataset, evaluation, and baseline. *Journal of remote sensing*, 3: 0096.

Jia, X.; Zhu, C.; Li, M.; Tang, W.; and Zhou, W. 2021. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3496–3504.

Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Guo, G.; Ye, Q.; Jiao, J.; et al. 2021. Anti-UAV: A large-scale benchmark for vision-based UAV tracking. *IEEE Transactions on Multimedia*, 25: 486–500.

Jie, Y.; Xu, Y.; Li, X.; and Tan, H. 2024. TSJNet: A multimodality target and semantic awareness joint-driven image fusion network. *arXiv preprint arXiv:2402.01212*.

Li, Z.; and Snavely, N. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2041–2050.

Lindenberger, P.; Sarlin, P.-E.; and Pollefeys, M. 2023. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, 17627–17638.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5802–5811.

Liu, R.; Liu, Z.; Liu, J.; Fan, X.; and Luo, Z. 2024. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6594–6609.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2): 91–110.

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7): 1200–1217.

Ma, J.; Yu, W.; Liang, P.; Li, C.; and Jiang, J. 2019. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48: 11–26.

Potje, G.; Cadar, F.; Araujo, A.; Martins, R.; and Nascimento, E. R. 2024. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2682–2691.

Qin, H.; Xu, T.; Li, T.; Chen, Z.; Feng, T.; and Li, J. 2025. MUST: The First Dataset and Unified Framework for Multispectral UAV Single Object Tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16882–16891.

Qiu, J.; Li, H.; Cao, H.; Zhai, X.; Liu, X.; Sang, M.; Yu, K.; Sun, Y.; Yang, Y.; and Tan, P. 2024. RA-MMIR: Multimodal image registration by robust adaptive variation attention gauge field. *Information Fusion*, 105: 102215.

Rao, Y.; Wu, D.; Han, M.; Wang, T.; Yang, Y.; Lei, T.; Zhou, C.; Bai, H.; and Xing, L. 2023. AT-GAN: A generative adversarial network with attention and transition for infrared and visible image fusion. *Information Fusion*, 92: 336–349.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, 2564–2571. Ieee.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938–4947.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8922–8931.

Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; and Ma, J. 2022a. SuperFusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12): 2121–2137.

Tang, L.; Wang, Y.; Gong, M.; Li, Z.; Deng, Y.; Yi, X.; Li, C.; Xu, H.; Zhang, H.; and Ma, J. 2025a. VideoFusion: A Spatio-Temporal Collaborative Network for Multimodal Video Fusion and Restoration. *arXiv preprint arXiv:2503.23359*.

Tang, L.; Yan, Q.; Xiang, X.; Fang, L.; and Ma, J. 2025b. C2RF: Bridging Multi-modal Image Registration and Fusion via Commonality Mining and Contrastive Learning. *International Journal of Computer Vision*, 1–19.

Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; and Ma, J. 2022b. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83: 79–92.

Tyszkiewicz, M.; Fua, P.; and Trulls, E. 2020. Disk: Learning local features with policy gradient. *Advances in neural information processing systems*, 33: 14254–14265.

van der Zee, T. J.; Tecchio, P.; Hahn, D.; and Raiteri, B. J. 2025. UltraTimTrack: a Kalman-filter-based algorithm to track muscle fascicles in ultrasound image sequences. *PeerJ Computer Science*, 11: e2636.

Wang, D.; Liu, J.; Fan, X.; and Liu, R. 2022. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*.

Wu, G.; Liu, H.; Fu, H.; Peng, Y.; Liu, J.; Fan, X.; and Liu, R. 2025. Every SAM Drop Counts: Embracing Semantic Priors for Multi-Modality Image Fusion and Beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17882–17891.

Xie, H.; Sang, M.; Zhang, Y.; Yang, Y.; Zhao, S.; and Zhong, J. 2024a. Rcvs: A unified registration and fusion framework for video streams. *IEEE Transactions on Multimedia*.

Xie, H.; Zhang, Y.; Qiu, J.; Zhai, X.; Liu, X.; Yang, Y.; Zhao, S.; Luo, Y.; and Zhong, J. 2023. Semantics lead all: Towards unified image registration and fusion from a semantic perspective. *Information Fusion*, 98: 101835.

Xie, X.; Cui, Y.; Tan, T.; Zheng, X.; and Yu, Z. 2024b. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1): 37.

Xing, G.; Wang, M.; Wang, F.; Sun, F.; and Li, H. 2025. Lightweight Edge-Aware Mamba-Fusion Network for Weakly Supervised Salient Object Detection in Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–13.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 502–518.

Xu, H.; Yuan, J.; and Ma, J. 2023. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(10): 12148–12166.

Yi, X.; Tang, L.; Zhang, H.; Xu, H.; and Ma, J. 2024. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110: 102450.

Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; and Ruan, X. 2022. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8886–8895.

Zhang, S.; Li, Z.; Zhang, K.; Lu, Y.; Deng, Y.; Tang, L.; Jiang, X.; and Ma, J. 2025a. Deep Learning Reforms Image Matching: A Survey and Outlook. *arXiv preprint arXiv:2506.04619*.

Zhang, S.; Zhu, Z.; Li, Z.; Lu, T.; and Ma, J. 2025b. Matching while perceiving: Enhance image feature matching with applicable semantic amalgamation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10094–10102.

Zhang, T.; Kong, F.; Deng, D.; Tang, X.; Wu, X.; Xu, C.; Zhu, L.; Liu, J.; Ai, B.; Han, Z.; and Deng, R. H. 2025c. Moving Target Defense Meets Artificial-Intelligence-Driven Network: A Comprehensive Survey. *IEEE Internet of Things Journal*, 12(10): 13384–13397.

Zhao, M.; Jiang, X.; and Huang, B. 2025. STFMamba: Spatiotemporal satellite image fusion network based on visual state space model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228: 288–304.

Zhao, W.; Cui, H.; Wang, H.; He, Y.; and Lu, H. 2025. FreeFusion: Infrared and Visible Image Fusion via Cross Reconstruction Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, X.; Wu, X.; Chen, W.; Chen, P. C.; Xu, Q.; and Li, Z. 2023. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–16.

Zhou, S.; Tan, W.; and Yan, B. 2022. Promoting single-modal optical flow network for diverse cross-modal flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3562–3570.