

UAV-MM3D: A Large-Scale Synthetic Benchmark for 3D Perception of Unmanned Aerial Vehicles with Multi-Modal Data

Longkun Zou^{1,*}, Jiale Wang^{2,*,†}, Rongqin Liang¹, Hai Wu¹, Ke Chen^{1,✉}, Yaowei Wang¹

Abstract—Accurate perception of UAVs in complex low-altitude environments is critical for airspace security and related intelligent systems. Developing reliable solutions requires large-scale, accurately annotated, and multimodal data. However, real-world UAV data collection faces inherent constraints due to airspace regulations, privacy concerns, and environmental variability, while manual annotation of 3D poses and cross-modal correspondences is time-consuming and costly. To overcome these challenges, we introduce UAV-MM3D, a high-fidelity multimodal synthetic dataset for low-altitude UAV perception and motion understanding. It comprises 400K synchronized frames across diverse scenes (urban areas, suburbs, forests, coastal regions) and weather conditions (clear, cloudy, rainy, foggy), featuring multiple UAV models (micro, small, medium-sized) and five modalities — RGB, IR, LiDAR, Radar, and DVS (Dynamic Vision Sensor). Each frame provides 2D/3D bounding boxes, 6-DoF poses, and instance-level annotations, enabling core tasks related to UAVs such as 3D detection, pose estimation, target tracking, and short-term trajectory forecasting. We further propose LGFusionNet, a LiDAR-guided multimodal fusion baseline, and a dedicated UAV trajectory prediction baseline to facilitate benchmarking. With its controllable simulation environment, comprehensive scenario coverage, and rich annotations, UAV3D offers a public benchmark for advancing 3D perception of UAVs.

Index Terms—Unmanned Aerial Vehicle (UAV), 3D Perception, Multi-Modal Dataset, Synthetic Benchmark

1 INTRODUCTION

The rapid proliferation of unmanned aerial vehicles (UAVs) in consumer, commercial, and industrial applications has significantly expanded low-altitude airspace activities. However, this growth also introduces severe challenges to low-altitude security, critical infrastructure protection, and public privacy [1], [2]. In complex low-altitude environments—characterized by diverse urban, suburban, forest, and coastal scenes, adverse weather conditions (e.g., rain, fog, strong sunlight), and fast-moving, small-sized aerial targets—accurate and real-time UAV perception (including detection, 3D pose estimation, tracking, and trajectory forecasting) plays a central role in low-altitude airspace management and security systems [3], [4]. Despite substantial progress in aerial perception, the development of robust and generalizable UAV perception models is fundamentally constrained by the lack of large-scale, multimodal datasets.

Existing UAV-related datasets exhibit several limitations that hinder further advancements. **First**, real-world UAV data collection is inherently constrained by strict airspace regulations, limited access to sensitive regions, and strong legal and privacy concerns regarding aerial imaging [5]. Meanwhile, environmental stochasticity (e.g., abrupt weather changes) makes it difficult to systematically capture diverse real-world conditions. **Second**, multimodal sensing remains insufficient: most datasets focus on RGB imagery

or RGB–thermal pairs [6], [7], and only a few include LiDAR [8], [9]. Crucial modalities such as radar—important for long-range detection under occlusion—and Dynamic Vision Sensors (DVS)—essential for capturing high-speed motion—are rarely integrated, limiting robustness in complex scenarios [10], [11]. **Third**, annotation richness is limited: most datasets provide only 2D bounding boxes without accurate 3D bounding boxes, 6-degree-of-freedom (6-DoF) pose annotations, or instance-level identity information [12]. This restricts the development of multi-task methods for 3D detection, pose estimation, target tracking, and trajectory prediction. **Finally**, benchmark baselines for multimodal UAV perception are scarce, making it difficult to evaluate cross-modal alignment strategies and integrated perception models.

To overcome these limitations, we introduce **UAV-MM3D**, a large-scale, high-fidelity multimodal synthetic dataset tailored for low-altitude UAV perception and airborne motion understanding. Compared with real-world collections, simulation enables scalable, safe, and controllable data generation with comprehensive coverage of environmental conditions, sensor configurations, and UAV dynamics [13], [14], [15]. UAV-MM3D incorporates five complementary modalities—RGB, infrared (IR), LiDAR, radar, and DVS—captured synchronously to leverage their complementary strengths: LiDAR provides accurate 3D geometric priors; radar offers robust long-range sensing under occlusion or adverse weather; DVS captures high-speed motion without motion blur; and RGB/IR ensure reliable perception across day–night transitions. The dataset contains 400,000 multi-sensor frames across diverse scenes, UAV types (micro, small, medium), weather conditions,

¹ Pengcheng Laboratory.

² University of Southern California.

* Equal contribution.

† The work was done during internship at the Pengcheng Laboratory.
Corresponding author: Ke Chen (Email: chen02@pcl.ac.cn)

and flight behaviors. Each frame is annotated with 2D and 3D bounding boxes, 6-DoF poses, instance IDs, and object categories, enabling research on 2D/3D detection, 6-DoF pose estimation, target tracking, and short-term trajectory prediction.

To further support benchmarking and research, we introduce two dedicated baseline models: **LGFusionNet**, a LiDAR-guided multimodal fusion network designed to align heterogeneous sensor features using LiDAR geometric priors; and a lightweight trajectory prediction baseline tailored to high-speed UAV motion patterns. Together with the dataset, these baselines provide a unified evaluation framework to standardize low-altitude UAV perception research. The main contributions of this work are summarized as follows:

- 1) We present **UAV-MM3D**, a large-scale multimodal synthetic dataset containing 400K synchronized frames across five sensing modalities with comprehensive annotations. It fills the current gap in multimodal, multi-scenario, and multi-task UAV datasets.
- 2) UAV-MM3D provides diverse environments, weather conditions, UAV models, and precise 6-DoF annotations, supporting four essential low-altitude UAV perception tasks: 2D/3D detection, 6-DoF pose estimation, multi-object tracking, and trajectory prediction.
- 3) We introduce **LGFusionNet**, a LiDAR-guided multimodal fusion baseline, along with a trajectory prediction baseline, offering standardized benchmarks for evaluating multimodal UAV perception algorithms.

2 RELATED WORK

Multimodal UAV datasets. The increasing deployment of UAVs in safety-critical low-altitude scenarios has encouraged the development of multimodal UAV datasets that combine heterogeneous sensing modalities. Existing datasets such as Anti-UAV-RGBT [6], Anti-UAV600 [12], and CHASE_DB [22] provide RGB-thermal pairs for robust tracking or detection, especially under illumination changes. DroneRGBT [7] extends this direction by offering aligned RGB-thermal sequences for UAV tracking. MMDrone [8] introduces RGB-event modality pairing to address motion blur and high-speed flight. However, these datasets primarily focus on single-object tracking and offer only 2D annotations. In contrast, our proposed dataset offers aligned multimodal streams with complete 3D annotations suitable for 3D detection, 6-DoF pose estimation, and trajectory prediction of multiple aerial targets.

Multimodal small object detection. Small object detection benefits greatly from multimodal fusion, especially under adverse conditions. Datasets such as KAIST Multi-spectral [23], and LLVIP [24] enable RGB-thermal fusion for small pedestrian detection under low-light settings. The DENSE dataset [25] provides a rich combination of lidar, RGB, radar, and thermal sensors for all-weather perception, inspiring robust small-object perception via cross-modality complementarity. Similarly, RODNet [26] introduces radar-camera fusion for detecting small and heavily occluded road objects. Nonetheless, most existing datasets are limited to ground-level scenes and lack low-altitude

aerial targets. Our dataset fills this gap by offering multimodal imagery and synchronized 3D annotations for small flying objects, enabling more generalizable research on multimodal small-object detection in the sky.

Multimodal autonomous driving perception datasets.

Multimodal perception in autonomous driving has been extensively studied with datasets such as nuScenes [27], Waymo [28], and KITTI [29], which provide rich sensor suites including cameras, lidar, and radar. Recent datasets such as Argoverse 2 [30] and ONCE [31] scale up real-world diversity and offer dense 3D annotations for autonomous systems. Although these datasets demonstrate the effectiveness of multimodal fusion for 3D detection, tracking, and 6-DoF reasoning, they are constrained to road scenes and ground vehicles. The sensing geometry and target characteristics differ fundamentally from low-altitude aerial objects such as drones. Our dataset extends multimodal 3D perception research from road-plane targets to airborne objects with high maneuverability, richer orientation changes, and larger pose degrees of freedom.

6-DoF pose estimation datasets. 6-DoF object pose estimation has predominantly focused on static, ground-based objects with datasets such as LineMOD [32], YCB-Video [33], and T-LESS [34]. Objectron [35] and DexYCB [36] further provide category-level annotations and hand-object interactions. For aerial vehicles, MAV6D [21] is the most relevant dataset, offering 6-DoF pose labels for micro aerial vehicles. However, it is limited to small indoor environments and offers relatively small data volume. Our UAV-MMSim dataset provides large-scale outdoor multimodal data with accurate 3D bounding boxes and full 6-DoF poses, enabling more realistic studies in UAV 3D detection and tracking.

3 THE UAV-MM3D DATASET

To support research on multimodal UAV and anti-UAV perception, we summarize representative UAV-related datasets in Table 1. Existing datasets typically focus on limited sensing modalities, provide only 2D annotations, or lack sufficient environmental and viewpoint diversity. In contrast, our UAV-MM3D dataset is designed to deliver richer multimodal inputs, high-quality 3D annotations, and broader scenario coverage, providing a more comprehensive foundation for developing robust UAV perception and counter-UAV algorithms.

3.1 Data Collection

Despite the availability of various real-world UAV datasets, they typically fail to capture the full spectrum of environmental and weather diversity required for robust perception. To overcome this limitation, we present UAV-MM3D, a large-scale multi-modal, multi-scenario, and multi-weather synthetic dataset tailored for comprehensive model development, evaluation, and pretraining.

Collection Framework. Our multi-modal data collection framework adopts a modular architecture, as illustrated in Fig. 1. It consists of two main components: (1) an Unreal Engine 4 (UE4)-based simulation server that integrates the CARLA [37] platform, and (2) a Python client featuring two controllers and one processor—the Weather Controller,

TABLE 1
Comparison of recent low-altitude aerial (LAA) / anti-UAV datasets. Modalities: which sensors are included.

Dataset	Modalities	3D Boxes	2D Boxes	Indoor/Outdoor	Class	Frames	Scenes	Day & Weather	Data Type
Anti-UAV-RGBT [6]	RGB, IR	✗	✓	Outdoor	6	~297,000	3	2	Real
DUT Anti-UAV [16]	RGB, IR	✗	✓	Outdoor	35	~10,000	–	2	Real
Anti-UAV410 [17]	IR	✗	✓	Outdoor	1	150,000	6	4	Real
MMAUD [18]	RGB, Audio, Radar, LiDAR	✗	✓	Outdoor	5	~50,000	1	1	Real
M3D [19]	RGB, Radar, LiDAR	✗	✓	Outdoor	10	83,999	8	1	Real/Sim
DrIFT [20]	RGB, Radar	✗	✓	Outdoor	7	47,991	3	4	Real/Sim
MAV6D [21]	RGB	✓	✗	Indoor	2	33,489	1	1	Real
UAV-MM3D (Ours)	RGB, IR, DVS, Radar, LiDAR	✓	✓	Outdoor	7	400,000	8	8	Sim

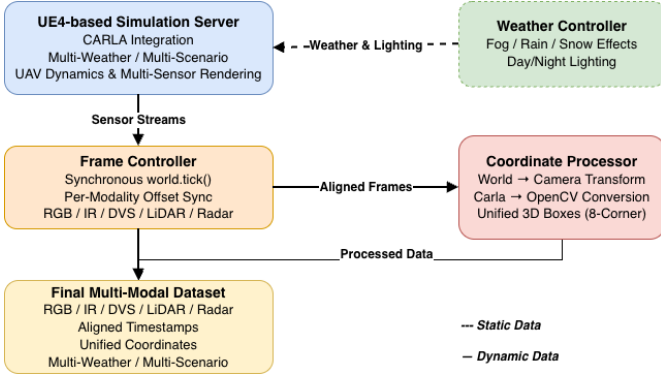


Fig. 1. UAV-MM3D data collection framework. UE4–CARLA simulation server provides multi-weather UAV scenes, while the Python client consists of a Weather Controller, Frame Controller, and Coordinate Processor. It outputs a fully aligned multi-modal dataset across diverse conditions.

Frame Controller, and Coordinate Processor. The Weather Controller configures environmental parameters such as lighting and precipitation, the Frame Controller ensures temporal synchronization and data alignment across modalities, and the Coordinate Processor converts sensor outputs into a unified coordinate space for further processing.

Sensor Settings. Our sensor configuration is designed to enable comprehensive ground-based observation of UAVs in diverse environmental conditions. The perception platform integrates multiple complementary modalities to achieve both geometric precision and visual robustness. We set up three different vision sensors: an RGB camera, an infrared (IR) camera, and a Dynamic Vision Sensor (DVS), each operating at a resolution of 1280×720 with a field of view of 90° (FoV). These vision sensors capture rich appearance and motion cues under varying illumination and dynamic flight patterns.

To enhance spatial awareness, 256-channel LiDAR and millimeter-wave radar are jointly placed. The LiDAR provides dense 3D measurements with a 120° horizontal and $\pm 20^\circ$ vertical FoV, operating at 15 Hz and generating up to 2.5 million points per second. The radar, with a FoV of $120^\circ \times 40^\circ \times 40^\circ$ and an effective range of 30 m, outputs the velocity, azimuth, and altitude for each detected object at 0.5 million points per second. Weather-dependent noise models are applied to both sensors to emulate realistic degradation in rain, fog, and snow.

Scene Setting. We selected four representative simulated environments from the CARLA simulator—Town01,

TABLE 2
Physical configurations of the seven UAV platforms in UAV-MM3D. Lightweight drones with incomplete public documentation use baseline parameters recorded in the dataset metadata.

UAV Model	Mass	Dimensions (mm)	Max Speed
DJI Avata 2	0.377 kg	185×212×64	27 m/s
DJI Mavic Mini	0.249 kg	159×202×55	13 m/s
DJI Phantom 4	1.38 kg	289×289×196	20 m/s
Matrice 210 RTK V2	4.69 kg	883×886×398	15 m/s
Matrice 600 Pro	10.5 kg	1668×1518×727	18 m/s
Matrice 300 RTK	3.6 kg	810×670×430	23 m/s
K3 Mini (E99)	0.12 kg	130×70×55 (folded)	5 m/s

Town03HD, Town07HD and Town10HD—covering urban and suburban layouts with varying building densities. Data from each scene is from multiple ground viewpoints to ensure coverage diversity and cross-view generalization.

To evaluate robustness under environmental variations, data is collected under eight distinct weather conditions, including clear day/night, rain day/night, fog day/night, and snow day/night. Each weather setting adjusts physical parameters such as cloudiness, precipitation, fog density, and sun altitude angle within the CARLA [37] physics-based atmosphere model, ensuring realistic visual and sensor-domain perturbations.

Drone Setting. Within each scene–weather combination, we simulate 7 UAVs of different physical configurations, including DJI Avata 2, DJI Mavic Mini, DJI Phantom 4, M210 RTK, Matrice 600 Pro, Matrice 300 RTK, and K3 Mini Drone (Fig. 2). Each UAV is modeled according to its speed, mass and dimensions in the real-world, as shown in Table 2. In this way, we are able to make physically accurate flight dynamics such as acceleration, deceleration, turning, and roll tilting. This design enables the dataset to capture realistic flight postures and multi-UAV interactions across complex weather and illumination conditions.

Trajectory Generation. To simulate realistic UAV motion within each scene, we design a procedural waypoint generator implemented in Unreal Engine. For each scenario, up to seven UAVs are initialized with distinct physical and kinematic properties, and a unique set of waypoints is generated using an elliptical trajectory model. The trajectory is defined along the forward direction of a scene anchor (denoted as *Locator*), forming a flattened oval path that gradually moves away from the ground sensors. Each trajectory consists of 15 to 25 discrete waypoints, evenly distributed along the

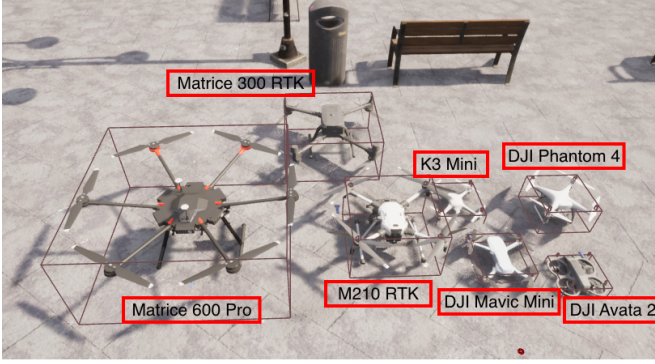


Fig. 2. Visualization of the seven UAV platforms simulated in UAV-MM3D, including DJI Avata 2, DJI Mavic Mini, DJI Phantom 4, M210 RTK, Matrice 600 Pro, Matrice 300 RTK, and the K3 Mini drone.

parametric curve:

$$x = a \cos(\theta), \quad y = b \sin(\theta), \quad z = h_0 + \Delta h(\theta), \quad (1)$$

where a and b control the longitudinal and lateral span of the ellipse, and $\Delta h(\theta)$ introduces stochastic height variation sampled within a predefined bound. The altitude and velocity profiles are randomized across UAVs to mimic diverse acceleration, deceleration, and turning behaviors.

3.2 Data Processing

All collected data are processed through a unified multi-threaded pipeline to ensure spatial and temporal consistency across all modalities. The pipeline performs synchronization and coordinate transformation.

Spatio-temporal Alignment. Images, LiDAR point clouds, radar data, and DVS event streams are aligned using their simulation timestamps. Before each iteration of data capture, a world-tick function is invoked to synchronize all sensors within the same simulation frame, ensuring that every modality is updated under an identical timestamp. Frames are retained only when all modalities share a common timestamp, achieving pixel-level temporal consistency across sensors.

Coordinate Transformation. As summarized in Table 3, UAV-MM3D adopts six distinct coordinate frames to ensure consistent spatial representation across simulators and sensors. The CARLA world coordinate follows a left-handed ESU convention, whereas all physical sensors and image frames are defined in right-handed systems (e.g., FLU for LiDAR, RDF for cameras). All transformations between simulator, sensor, and image coordinates are stored in each sequence to maintain reproducible geometry and pixel-level alignment. After all, annotated 3D bounding boxes and projected multi-modal data are expressed in the camera-centric OpenCV coordinate system (RDF, right-down-forward), which serves as the unified reference frame for training, visualization, and evaluation.

3.3 Statistical Analysis

In total, UAV-MM3D comprises approximately 1.1K simulated clips, corresponding to about 341K frames captured at a frame rate of 15 Hz. Each clip spans roughly 20 seconds of continuous synchronized recording across all modalities,

TABLE 3
Different Coordinate Systems in UAV-MM3D.

Name	Category	Type	Origin
World	Global Ref.	ESU (L)	Fixed reference point
Ego	UAV Model	FLU (R)	Drone geometric center
Sensor-Cam	Image Sensor	RDF (R)	Camera optical center
Sensor-LiDAR	Range Sensor	FLU (R)	LiDAR center
Sensor-Radar	Motion Sensor	FRD (R)	Radar center
Simulator	CARLA Engine	ESU (L)	World origin
Rendered Img	OpenCV Frame	RDF (R, 2D)	Top-left of image

Note: ESU: East-South-Up, FLU: Forward-Left-Up, FRD: Forward-Right-Down, RDF: Right-Down-Forward. Handedness: R (Right-handed), L (Left-handed).

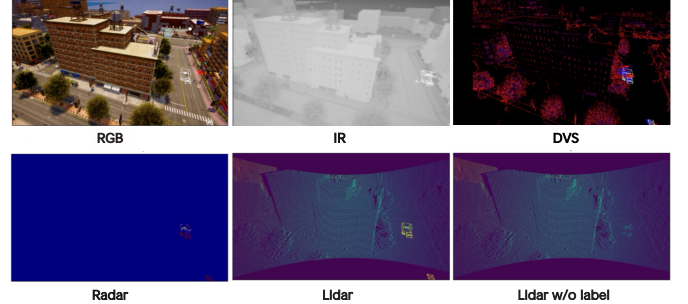


Fig. 3. Examples of multi-modal sensor outputs in UAV-MM3D. Each synchronized frame includes RGB imagery, infrared thermal data, and DVS event streams. LiDAR point clouds are reprojected onto the RGB view for 2D alignment, while millimeter-wave radar measurements are rendered as velocity heatmaps. All modalities are temporally aligned under a unified simulation timestamp.

including RGB, infrared, DVS, LiDAR, and millimeter-wave radar, as illustrated in Fig. 3. Together, these clips yield over 1.7 million sensor images with 3D point clouds, forming a large-scale dataset for perception research under complex aerial-ground scenarios.

Each simulated scene involves **one to seven UAVs** with heterogeneous physical models and flight dynamics, producing substantial variation in scale, appearance, and occlusion across viewpoints. As shown in Fig. 4(a), UAV types are well balanced across the dataset: the Matrice-300-RTK appears most frequently due to its representative size and dynamics, while extremely large or small platforms naturally occur less often. Fig. 4(b) further reports the distribution of UAV-sensor distances, where most cases fall within 20–60 meters, accompanied by both close-range and long-range examples that together form a realistic and diverse depth profile. Finally, Fig. 4(c) shows the orientation statistics, which exhibit broad roll and pitch variation but a narrow yaw range. This reflects realistic multirotor flight dynamics—real UAVs primarily rely on roll-based banking rather than large heading changes during lateral motion.

Data Categorization. To enable comprehensive evaluation under varying perception challenges, all annotated samples in UAV-MM3D are divided into three difficulty levels: *easy*, *moderate*, and *hard*. The difficulty score is defined as the ratio between the UAV’s distance and its physical size, scaled by a weather-dependent weighting factor:

$$S = \frac{d}{s} \times w_{\text{weather}}, \quad (2)$$

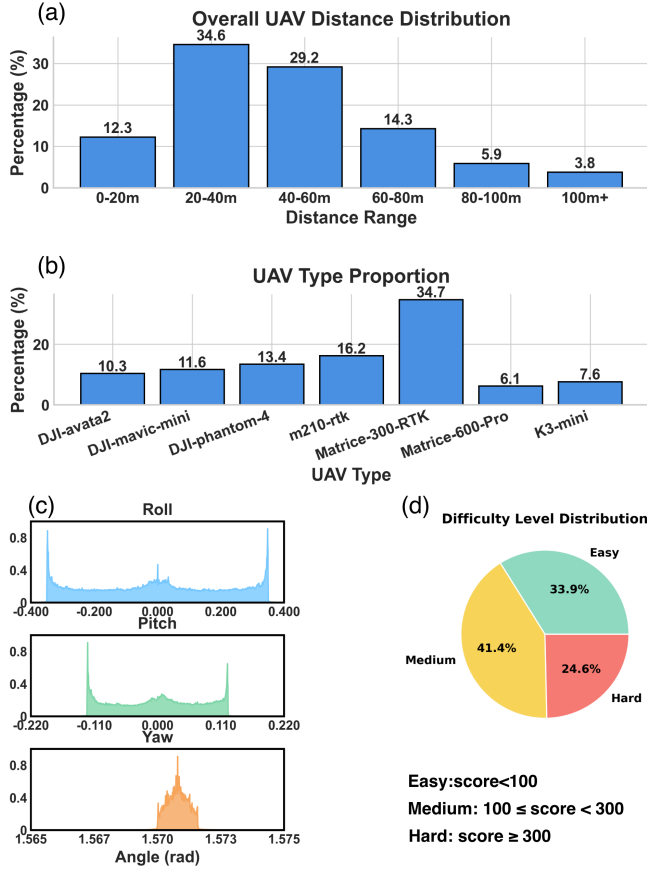


Fig. 4. Distribution statistics of UAV-MM3D. (a) Overall UAV-sensor distance distribution. (b) Proportion of UAV types across all simulated scenes. (c) Orientation statistics (roll, pitch, yaw) computed from 3D bounding boxes. (d) Difficulty distribution estimated from UAV size, sensing distance, and weather conditions.

where d denotes the camera-to-UAV distance, s represents the UAV's maximum physical dimension, and w_{weather} corresponds to the visibility degradation (e.g., 1.0 for clear, 2.5 for fog). This stratified organization ensures balanced coverage across different visibility and scale conditions, facilitating robust model benchmarking under diverse sensing scenarios, as illustrated in Fig. 4(d).

Scene Statistics. The UAV-MM3D dataset encompasses a broad spectrum of simulation conditions, integrating four representative CARLA towns with distinctive urban and suburban topologies and eight physically based weather configurations that modulate illumination, precipitation, and atmospheric visibility, as illustrated in Fig. 5. These environments jointly provide diverse visual and geometric contexts that reflect varying levels of structural complexity, traffic density, and environmental dynamics.

4 APPLICATIONS OF UAV-MM3D

UAV-MM3D provides a unified and comprehensive evaluation environment for a wide range of Anti-UAV perception tasks, including 2D detection, 3D detection, 6-DoF pose estimation, multi-object tracking, and short-term trajectory prediction. Among these tasks, 6-DoF pose estimation plays

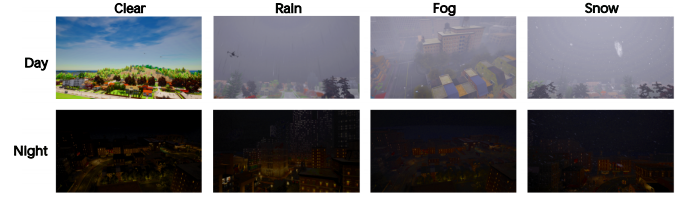


Fig. 5. Illustration of the eight physically based weather conditions in UAV-MM3D. Each environment is rendered under both day and night illumination, covering clear, rain, fog, and snow settings. These combinations introduce diverse lighting, atmospheric, and visibility variations that enhance the realism and challenge of UAV perception across different urban and suburban scenes.

a central role because accurate UAV localization and orientation are essential prerequisites for downstream tracking and interception strategies. In this section, we present the multimodal 6-DoF pose estimation task supported by UAV-MM3D, followed by the proposed LiDAR-guided baseline LGFusionNet. We further introduce three multimodal 3D detection baselines and a series of ablation studies.

4.1 UAVs Multimodal 6-DoF Pose Estimation

4.1.1 Task Definition

The objective of multimodal 6-DoF pose estimation is to recover the full 3D position (x, y, z) and 3D rotation (θ, ϕ, ψ) of UAVs from synchronized multimodal sensor streams. Compared to ground vehicles or large-scale objects, UAVs in low-altitude airspace exhibit several unique challenges: (1) extremely fast motion and abrupt trajectory changes; (2) small physical size and low pixel footprint at medium-to-long distances; (3) severe viewpoint variation due to free-space 3D maneuvering; and (4) strong modality inconsistency across RGB, IR, Radar, and LiDAR caused by illumination changes or sparse geometry. UAV-MM3D offers temporally aligned multimodal data, allowing systematic exploration of cross-modal feature correspondence and uncertainty reduction in pose estimation.

4.1.2 LGFusionNet Architecture

To exploit the complementary strengths of different modalities, we introduce **LGFusionNet** — a LiDAR-guided multimodal 6-DoF pose estimation network (architecture shown in Fig. 6).

Two-Branch Multi-Sensor Feature Encoding. LGFusionNet adopts a dual-branch structure to handle heterogeneous modalities:

- **RGB Branch:** Processes appearance-rich inputs (RGB imagery + LiDAR-projected depth + Radar-projected velocity) via a ResNet-50 backbone. LiDAR/Radar projections are mapped to the RGB image plane using RGB camera extrinsic parameters, providing geometric/motion priors for texture-poor scenarios. The branch feature is

$$F_{\text{RGB-branch}} = \text{Concat}(F_{\text{RGB}}, F_{\text{LiDAR(RGB)}}, F_{\text{Radar(RGB)}}), \quad (3)$$

where F_{RGB} , $F_{\text{LiDAR(RGB)}}$, $F_{\text{Radar(RGB)}}$ denote features of RGB imagery, LiDAR-projected depth, and Radar-projected velocity (RGB branch), respectively.

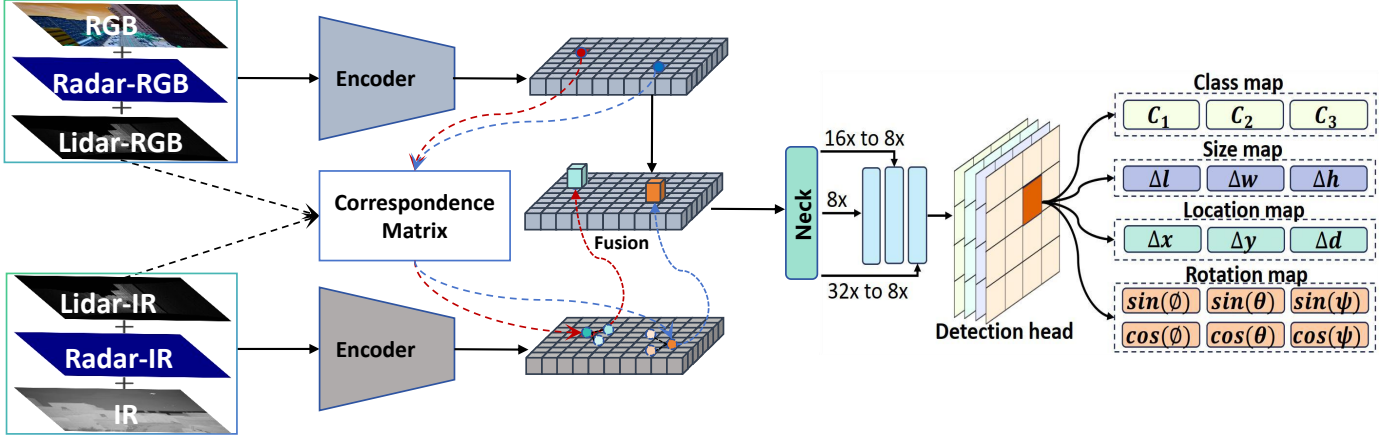


Fig. 6. Overall architecture of the proposed LGFusionNet. The network adopts a dual-branch multi-sensor feature encoding design, where the RGB and IR branches independently encode their stacked inputs (image, LiDAR-projected depth, and Radar-projected velocity) using two separate ResNet-50 backbones with identical architecture. LiDAR points are projected onto both RGB and IR views to construct cross-view geometric correspondences, which are then used by the LiDAR-guided cross-branch spatial alignment module. This module samples IR-branch features according to the RGB-IR projection pairs and aggregates them via KNN-based neighborhood fusion to obtain aligned IR descriptors. The aligned IR-branch features are then fused with RGB-branch features in the RGB coordinate space through channel-wise concatenation followed by convolution, producing a spatially consistent multimodal representation. This fused representation goes through a neck to the prediction head, supporting both 6-DoF UAV pose estimation and object classification as downstream tasks. Specifically, rotations are predicted using Direction-Angle Parameterization (DAP) for stable and continuous angular regression.

- **IR Branch:** Processes illumination-invariant inputs (IR imagery + LiDAR-projected depth + Radar-projected velocity) via the same ResNet-50 backbone. Projections use IR camera extrinsic parameters, and the branch feature is

$$F_{\text{IR-branch}} = \text{Concat}(F_{\text{IR}}, F_{\text{LiDAR(IR)}}, F_{\text{Radar(IR)}}), \quad (4)$$

where $F_{\text{IR}}, F_{\text{LiDAR(IR)}}, F_{\text{Radar(IR)}}$ correspond to IR branch features.

LiDAR-Guided Cross-Branch Spatial Alignment. To establish geometric consistency between the two multimodal branches rather than raw RGB/IR images, LGFusionNet performs LiDAR-guided alignment between $F_{\text{RGB-branch}}$ and $F_{\text{IR-branch}}$. Since each branch internally incorporates image (RGB or IR), LiDAR-projected depth, and Radar-projected velocity cues, LiDAR points serve as a stable geometric anchor to bridge the heterogeneous representation spaces of the two branches.

- **LiDAR-Based Cross-View Point Correspondence Construction.** LiDAR points are defined in world coordinates and can be accurately projected to both RGB and IR cameras using their intrinsic and extrinsic parameters. For a LiDAR point $P_k = (X_k, Y_k, Z_k)^\top$, the projection onto camera $c \in \{\text{RGB}, \text{IR}\}$ is

$$p_{c,k} \sim K_c[R_c | t_c], \quad (5)$$

where $p_{c,k} = (u_{c,k}, v_{c,k})^\top$ denotes the pixel coordinates of P_k in camera c 's image plane, K_c is the intrinsic matrix of camera c , and $[R_c | t_c]$ represents its extrinsic parameters that transform 3D points from world to camera coordinates. By projecting the same LiDAR point to RGB and IR cameras, we obtain correspondence pairs $(\pi_{\text{RGB}}(P_k), \pi_{\text{IR}}(P_k))$, providing sparse but geometrically reliable anchors across the RGB-branch and IR-branch feature spaces.

- **Projection-Guided IR-Branch Feature Sampling and KNN Aggregation.** Let $F_{\text{RGB-branch}}, F_{\text{IR-branch}} \in \mathbb{R}^{H \times W \times D}$ denote the two branch feature maps at the same resolution. For each pixel (i, j) in $F_{\text{RGB-branch}}$, we collect the corresponding IR projections whose RGB projections fall in the local neighborhood $\mathcal{N}((i, j); r)$:

$$S_{i,j} = \{\pi_{\text{IR}}(P_k) \mid \pi_{\text{RGB}}(P_k) \in \mathcal{N}((i, j); r)\}.$$

Features of the IR-branch are retrieved at the projected positions and aggregated using K -nearest neighbors:

$$\mathcal{I}_{i,j} \leftarrow \text{KNN}(S_{i,j}, K), \quad (6)$$

$$\hat{f}_{\text{IR-branch}}(i, j) = \begin{cases} \mathcal{A}(F_{\text{IR-branch}}, \mathcal{I}_{i,j}), & S_{i,j} \neq \emptyset, \\ \mathbf{0}, & \text{otherwise,} \end{cases} \quad (7)$$

where $\hat{f}_{\text{IR-branch}}(i, j) \in \mathbb{R}^D$ denotes the aligned IR feature at pixel (i, j) , and $\mathcal{I}_{i,j}$ is the set of K nearest neighbors in the IR-branch feature map corresponding to LiDAR points projected within the neighborhood of RGB-branch pixel (i, j) . Here, $\mathcal{A}(\cdot, \cdot)$ represents the aggregation operator (e.g., averaging) over the selected features.

- **Cross-Branch Feature Fusion on RGB-Branch Coordinate Space.** The aligned IR-branch feature is fused with the RGB-branch feature at each pixel by channel-wise concatenation followed by a convolution:

$$f_{\text{cat}}(i, j) = \text{Concat}(f_{\text{RGB-branch}}(i, j), \hat{f}_{\text{IR-branch}}(i, j)), \quad (8)$$

$$f_{\text{fused}}(i, j) = \text{Conv}(f_{\text{cat}}(i, j)), \quad (9)$$

where $f_{\text{cat}}(i, j)$ denotes the concatenated RGB-IR feature vector at pixel (i, j) prior to convolutional fusion. This convolutional operation produces the cross-branch fused feature map $F_{\text{fused}} \in \mathbb{R}^{H \times W \times D'}$, which enhances RGB-branch perception with illumination-robust IR

Algorithm 1 LiDAR-Guided RGB–IR Branch Alignment and Feature Fusion**Require:**

- 1: LiDAR points $\{P_k\}$;
- 2: intrinsics/extrinsics $K_c, [R_c|t_c]$;
- 3: feature maps $F_{\text{RGB-branch}}, F_{\text{IR-branch}}$;
- 4: radius r ;
- 5: neighbors K

Ensure: Fused feature map F_{fused}

```

6: for each LiDAR point  $P_k$  do
7:    $p_{\text{RGB}} \leftarrow \text{Project}(P_k, K_{\text{RGB}}, [R_{\text{RGB}}|t_{\text{RGB}}])$ 
8:    $p_{\text{IR}} \leftarrow \text{Project}(P_k, K_{\text{IR}}, [R_{\text{IR}}|t_{\text{IR}}])$ 
9:   store correspondence pair  $(p_{\text{RGB}}, p_{\text{IR}})$ 
10: end for
11: for each pixel  $(i, j)$  in  $F_{\text{RGB-branch}}$  do
12:    $S_{i,j} \leftarrow \{p_{\text{IR}} \mid p_{\text{RGB}} \in \mathcal{N}((i, j); r)\}$ 
13:   if  $S_{i,j} = \emptyset$  then
14:      $\hat{f}_{\text{IR-branch}}(i, j) \leftarrow 0$ 
15:   else
16:      $\mathcal{I}_{i,j} \leftarrow \text{KNN}(S_{i,j}, K)$ 
17:      $\hat{f}_{\text{IR-branch}}(i, j) \leftarrow \text{Aggregate}(F_{\text{IR-branch}}, \mathcal{I}_{i,j})$ 
18:   end if
19:    $f_{\text{cat}}(i, j) \leftarrow \text{Concat}(f_{\text{RGB-branch}}(i, j), \hat{f}_{\text{IR-branch}}(i, j))$ 
20:    $f_{\text{fused}}(i, j) \leftarrow \text{Conv}(f_{\text{cat}}(i, j))$ 
21: end for
22: return  $F_{\text{fused}}$ 

```

cues while preserving geometric correctness guaranteed by LiDAR projections. The overall LiDAR-guided cross-branch alignment and feature fusion pipeline is summarized in Algorithm 1.

Remarks. This LiDAR-guided procedure ensures that cross-branch fusion is performed between $F_{\text{RGB-branch}}$ and $F_{\text{IR-branch}}$ at geometrically corresponding spatial locations, rather than relying on appearance similarity alone. Such spatially enforced alignment is crucial given the view-point differences between the RGB and IR cameras and the heterogeneous sensing characteristics of the associated LiDAR/Radar projections within each branch. In addition, the KNN-based aggregation on the IR feature map improves robustness to the sparsity of LiDAR projections and alleviates discretization artifacts near depth discontinuities, ensuring stable multimodal fusion in long-range or low-visibility scenarios.

4.1.3 Comparative Baselines

To comprehensively assess the contribution of LiDAR-guided cross-modal alignment in LGFusionNet, we design comparative baselines based on two representative technical paradigms. Each paradigm is evaluated under three modal configurations (RGB-only, IR-only, Score-Level Fusion) to isolate the impact of fusion strategies, with all baselines implemented within the same unified framework:

- **YOLO-6D** [38]: A geometric-based paradigm that inherits the core idea of Keypoint+PnP—extracting 2D keypoints from input data, followed by the PnP (Perspective-n-Point) algorithm for pose estimation. It provides a lower bound for geometry-driven pose inference, with three modal settings:

- RGB-only: Uses only RGB inputs (texture-rich but illumination-sensitive);
- IR-only: Uses only IR inputs (illumination-robust but lacking fine-grained appearance cues);
- Score-Level Fusion (Late Fusion): Decision-level fusion of independent detectors trained on four unaligned modalities—RGB, IR, LiDAR projection, and Radar projection—via weighted averaging. Modalities are not spatially aligned prior to fusion, relying solely on post-hoc score merging.

- **Center-based Regression** [39]: A modern one-stage paradigm that models pose estimation as center-point heatmap regression, representing state-of-the-art monocular pose approaches. It also adopts three modal configurations:

- RGB-only: RGB inputs as the sole input;
- IR-only: IR inputs as the sole input;
- Score-Level Fusion (Late Fusion): Decision-level weighted fusion of regression results from four unaligned modalities (RGB, IR, LiDAR projection, Radar projection), with no prior spatial alignment between heterogeneous sensors.

For all Late Fusion variants in the baselines, weighted averaging is applied to the final detection confidence scores from each modality branch:

$$S_{\text{final}} = \alpha_1 S_{\text{RGB}} + \alpha_2 S_{\text{IR}} + \alpha_3 S_{\text{LiDAR-Proj}} + \alpha_4 S_{\text{Radar-Proj}}, \quad (10)$$

where S_{RGB} , S_{IR} , $S_{\text{LiDAR-Proj}}$, and $S_{\text{Radar-Proj}}$ denote the modality-specific confidence scores for object existence/classification. The coefficients satisfy $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$, and are fixed as $\alpha_1 = 0.3$, $\alpha_2 = 0.3$, $\alpha_3 = 0.2$, $\alpha_4 = 0.2$. The fused score S_{final} is used for detection ranking and NMS, while regression outputs (center, rotation, and size) are taken from the modality branch with the highest confidence to avoid instability caused by averaging heterogeneous regression predictions.

For comparison, LGFusionNet’s own Late Fusion variant adopts a distinct design: it fuses results from two dedicated branches (RGB-branch and IR-branch) rather than directly merging four independent modalities. Each branch inherently integrates its core modality (RGB/IR) with LiDAR projection and Radar projection as auxiliary geometric cues. This variant serves as an additional ablation to verify whether LiDAR-guided spatial alignment (the core innovation) provides incremental value beyond the structured branch design.

4.1.4 Pose Decoding and Metrics

All methods predict full 6-DoF pose parameters—including 3D translation (x, y, z) , rotation (θ, ϕ, ψ) , size (w, h, l) —together with 2D and 3D bounding boxes. Performance is evaluated using standard metrics:

- Rotation error (degrees), Position error (meters), Size error (meters);
- 2D Average Precision (2D AP), 3D Average Precision (3D AP).

4.1.5 Results

Although the UAV-MM3D dataset includes a DVS (event camera) modality, it is excluded from our main fusion

pipeline. In rain/fog and nighttime scenes, illumination flickering and reflections often trigger dense pseudo-events, yielding unstable temporal statistics and weak geometric correspondence with frame-based modalities. Preliminary experiments showed that even after event denoising and temporal resampling, incorporating DVS degraded pose supervision and overall accuracy. To ensure fair comparison and focus on the effect of LiDAR-guided fusion, LGFusionNet is evaluated using RGB-IR-LiDAR-Radar inputs. Nonetheless, the high dynamic range and motion sensitivity of DVS remain valuable, and developing robust event-frame alignment under adverse weather will be explored in future work.

Implementation details. All detectors were trained on 8×RTX 4090 GPUs using the Adam optimizer [40] with an initial learning rate of 0.001 and cosine decay. LGFusionNet was trained for 50 epochs on UAV-MM3D with a batch size of 4 per GPU. For the LiDAR-guided alignment module, we set the projection neighborhood radius to $r = 4$ pixels and used $K = 5$ nearest neighbors for local feature aggregation. Both RGB and IR branches employ ResNet-50 backbones with identical initialization and synchronized multimodal augmentations (random horizontal flip, brightness jitter, and depth/velocity dropout).

Evaluation results. Table 4 summarizes the comparative results across different technical paradigms and modal configurations. Across both YOLO-6D [38] and Center-based Regression paradigms, a consistent trend emerges: IR-only unimodal models outperform their RGB-only counterparts, thanks to IR’s inherent robustness to low-illumination conditions in low-altitude UAV scenes. Score-Level Fusion (Late Fusion) further improves upon single-modality performance for all baselines, as it leverages complementary appearance and thermal cues. However, Late Fusion’s gains remain limited—its lack of spatial alignment between RGB and IR modalities leads to suboptimal feature interaction, especially under fast UAV motion and viewpoint variations.

Among the baseline paradigms, Center-based Regression [39] consistently outperforms YOLO-6D [38] across all modal configurations. This is attributed to its end-to-end regression design, which better captures fine-grained pose correlations compared to YOLO-6D’s two-stage geometric inference (prone to keypoint detection errors, inheriting the limitation of the Keypoint+PnP approach). Even so, the best baseline (Center-based Regression [39] with Late Fusion) still struggles to mitigate cross-modal misalignment, resulting in relatively high rotation error (17.84°) and position error (4.92 m).

In stark contrast, LGFusionNet demonstrates superior performance across all metrics, with its core LiDAR-guided Fusion configuration achieving the best results. Compared to the top-performing baseline, LGFusionNet reduces position error by more than 25% (from 4.92 m to 3.64 m) and rotation error by over 40% (from 17.84° to 10.57°), while boosting 3D AP by more than 50% (from 17.38 to 26.72). Notably, LGFusionNet’s own Late Fusion variant also outperforms all baseline methods, confirming the effectiveness of its dual-branch encoding architecture. The additional gains from LiDAR-guided spatial alignment further validate that geometry-driven cross-modal correspondence is critical for resolving ambiguities in complex low-altitude scenes—this

is the key advantage that distinguishes LGFusionNet from traditional fusion approaches.

4.1.6 Ablation Studies

We further analyze the impact of key components in our framework.

Effect of Input Resolution. Since the perceived object scale in monocular 3D detection is strongly influenced by image resolution, we replace the original focal-length ablation with a resolution-based study. We evaluate three commonly used input resolutions $\{960 \times 540, 1060 \times 600, 1280 \times 720\}$ while keeping all camera intrinsics unchanged. As shown in Table 5, input resolution has a **significant impact** on pose estimation performance: with the gradual increase of resolution, all evaluation metrics are continuously optimized. Specifically, compared with the lowest resolution (960×540), the optimal resolution (1280×720) reduces rotation error by 3.85° ($14.42^\circ \rightarrow 10.57^\circ$), position error by 0.84 m ($4.48 \text{ m} \rightarrow 3.64 \text{ m}$), and size error by 0.06 ($0.41 \rightarrow 0.35$), while boosting 3D AP by 6.57 ($20.15 \rightarrow 26.72$) — a relative improvement of over 32%. This is attributed to higher resolution providing clearer object boundaries, richer texture details, and more precise geometric cues, which are critical for accurate 6D pose inference. The consistent performance gain across resolution scales confirms that our model can effectively leverage the additional information from high-resolution inputs, and the 1280×720 configuration achieves the best balance between precision and computational efficiency.

Effect of LiDAR-Guided Cross-Branch Fusion. To investigate the contribution of each component in LGFusionNet, we conduct ablation experiments by progressively enhancing the fusion strategy and introducing the LiDAR-guided correspondence module. As shown in Table 6, the RGB-branch and IR-branch variants exhibit complementary strengths: the RGB branch achieves better spatial precision under normal lighting, while the IR branch is more robust in low-illumination or high-glare conditions.

We further explore two fusion paradigms: *Late Fusion* (score-level merging of independent branch outputs) and *Mid-Attention Fusion* (feature-level interaction via cross-attention before decoding). The results show that Mid-Attention Fusion outperforms Late Fusion, as feature-level interaction captures finer-grained cross-modal correlations compared to post-hoc score averaging. However, both fusion strategies without LiDAR guidance still suffer from heterogeneous modality misalignment, leading to suboptimal performance.

In contrast, the full model with LiDAR-guided correspondence achieves the best performance across all metrics. This demonstrates that LiDAR serves as a reliable geometric anchor to align cross-sensor features, and when combined with mid-level attention fusion, it maximizes the complementary value of RGB and IR modalities.

4.2 UAVs Target Tracking

Reliable UAV target tracking is essential for accurate low-altitude perception and serves as a critical foundation for downstream tasks such as trajectory prediction and intent understanding. Our dataset offers large-scale, high-fidelity,

TABLE 4

Comparison of multimodal and unimodal 6-DoF pose estimation methods on the UAV-MM3D dataset. LGFusionNet achieves the best overall performance by leveraging LiDAR-guided cross-modal alignment. Center&DAP Encoding denotes the combination of Center Encoding (for target center localization) and DAP Encoding (for rotation angle parameterization), addressing spatial positioning and angular regression in 6-DoF pose estimation respectively. Note: † indicates Score-Level Fusion of four unaligned modalities (RGB, IR, LiDAR projection, Radar projection); ‡ indicates Score-Level Fusion of two branches (RGB-branch and IR-branch), each integrating RGB/IR with LiDAR/Radar projections.

Method	Modality Fusion	Target Encoding	Backbone	Rot. Err↓	Pos. Err↓	Size Err↓	2D AP↑	3D AP↑
YOLO-6D [38]	RGB-only	Keypoint Encoding	ResNet50	26.47	7.21	0.43	36.44	11.78
	IR-only	Keypoint Encoding	ResNet50	24.15	6.83	0.41	38.92	12.26
	Late Fusion†	Keypoint Encoding	ResNet50	22.73	5.95	0.40	40.51	13.89
Center-based Re-gression [39]	RGB-only	Center Encoding	ResNet50	21.65	6.48	0.42	41.28	13.07
	IR-only	Center Encoding	ResNet50	19.29	5.76	0.40	42.15	14.64
	Late Fusion†	Center Encoding	ResNet50	17.84	4.92	0.39	43.67	17.38
LGFusionNet (Ours)	Late Fusion‡	Center&DAP Encoding	ResNet50	12.48	3.95	0.37	43.89	23.11
	LiDAR-guided Fusion	Center&DAP Encoding	ResNet50	10.57	3.64	0.35	45.35	26.72

TABLE 5

Ablation study on input image resolution. Higher resolution provides richer geometric and texture cues, leading to continuous performance improvement.

Resolution	Rot. Error↓	Pos. Error↓	Size Error↓	3D AP↑
960×540	14.42	4.48	0.41	20.15
1060×600	12.98	3.92	0.39	23.32
1280×720	10.57	3.64	0.35	26.72

TABLE 6

Ablation of LGFusionNet components. RGB-branch = RGB branch only; IR-branch = IR branch only; Dual-LateFusion = dual branches with Late Fusion (score-level); Dual-MidAttn = dual branches with Mid-Attention Fusion (feature-level); Full (LGF) = full model with LiDAR-guided mid-attention fusion.

Model Variant	Rot. Err↓	Pos. Err↓	Size Err↓	3D AP↑
RGB-branch	18.92	4.78	0.42	14.86
IR-branch	15.64	4.22	0.41	18.93
Dual-LateFusion	12.48	3.95	0.37	23.11
Dual-MidAttn	11.75	3.82	0.37	24.68
Full (LGF)	10.57	3.64	0.35	26.72

and multi-modal observations, providing comprehensive support for advancing UAV tracking research under challenging conditions, including appearance variation, and scale changes.

TABLE 7

UAV target tracking performance on UAV-MM3D dataset.

Method	AUC ↑	P ↑	P_{norm} ↑
OTrack [41]	21.8	51.9	41.3
DropTrack [42]	28.8	53.7	41.3
ARTrack [43]	29.8	58.0	42.9
ODTrack [44]	30.7	58.1	44.9

Metrics. We report the following widely used metrics for the target tracking task. Area Under the Curve (AUC) measures the area under the success plot, which represents the fraction of frames whose Intersection-over-Union (IoU) exceeds varying thresholds as a function of the overlap threshold. Precision (P) evaluates the percentage of frames in which the predicted bounding-box center lies within a

specified distance (typically 20 pixels) from the ground-truth center. Normalized Precision (P_{norm}) normalizes the center error by the size of the ground-truth bounding box, providing a scale-invariant measurement. Higher values of AUC, P , and P_{norm} indicate better tracking performance.

Baselines. We evaluate our dataset using four representative and widely adopted target tracking methods: OTrack [41], DropTrack [42], ARTrack [43], and ODTrack [44]. These methods cover diverse and strong tracking paradigms and are chosen to provide a comprehensive and fair evaluation of tracking performance on our dataset. All methods are implemented using their official configurations and trained under the same experimental protocol for fair comparison.

Results. Table 7 reports the tracking performance of different baselines on our dataset. Existing methods achieve reasonable results in AUC and precision, but all exhibit noticeable performance degradation compared to standard benchmarks, indicating the increased difficulty of low-altitude UAV tracking. The results demonstrate that our dataset is challenging and well-suited for evaluating robust UAV target tracking methods.

Future work will explore multi-modal fusion of complementary sensing cues, including RGB, infrared, LiDAR, and radar data, to enable more accurate and robust low-altitude UAV target tracking.

4.3 UAVs Trajectory Prediction

Accurate prediction of low-altitude UAVs trajectories is essential for fine-grained aerial perception and intelligent airspace management. Our dataset provides a comprehensive foundation for advancing UAV trajectory prediction research, particularly for exploring multi-modal fusion and modeling of dynamic flight behaviors.

Approach. To validate the effectiveness of our dataset, we introduce a baseline model, UAVsTrajectory, as illustrated in Fig. 7. The proposed model adopts an adaptive gated weighting mechanism to fuse RGB appearance features with historical flight trajectories. Specifically, the appearance modality encodes UAV posture and motion orientation, while the trajectory modality captures temporal dynamics. Their complementary fusion enables the model

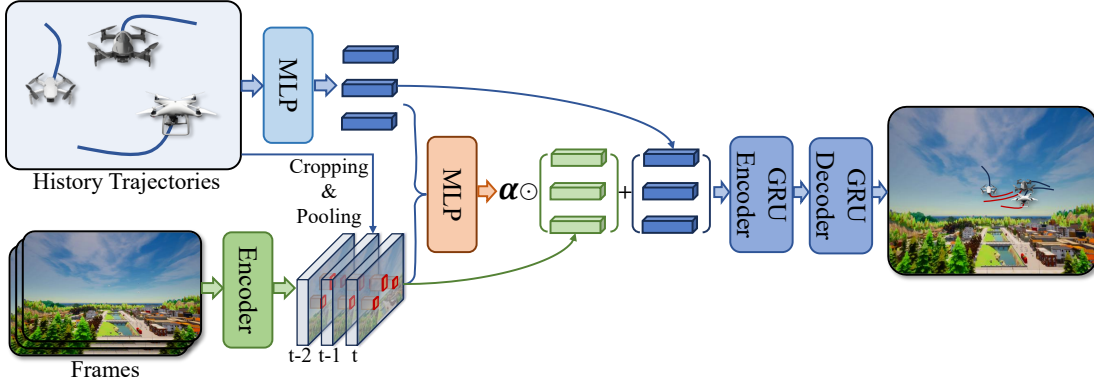


Fig. 7. The framework of the proposed UAVsTrajectory. Our framework takes historical 3D UAV trajectories and corresponding RGB images as inputs. The historical trajectories and image frames are first encoded separately. Guided by the 2D bounding boxes, target regions are cropped from the images and pooled to obtain compact appearance representations. An adaptive fusion module then learns modality-wise weighting coefficients to integrate trajectory and appearance features. Finally, the fused representation is decoded to predict future UAV trajectories.

TABLE 8

Trajectory prediction performance (ADE / FDE in meters) on UAV-MM3D dataset. We evaluate the proposed method over prediction horizons of 1 s, 3 s, and 5 s. ADE measures the average per-frame displacement error over the prediction interval, while FDE denotes the displacement error at the final predicted timestep. We compare our approach with a traditional Kalman filter-based method and an LSTM-based learning model.

Method	ADE@1s ↓	FDE@1s ↓	ADE@3s ↓	FDE@3s ↓	ADE@5s ↓	FDE@5s ↓
Kalman Filter [45]	5.73	9.85	17.10	33.50	29.27	58.96
LSTM [46]	2.45	4.97	8.86	17.78	14.93	27.84
Ours	2.37	4.32	8.58	17.29	14.25	26.94

to generate more accurate and stable future trajectory predictions.

Metrics. We follow common practices in trajectory prediction and report Average Displacement Error (ADE) and Final Displacement Error (FDE) as our evaluation metrics. All models are evaluated under a consistent setting—observing 1 second and predicting the subsequent 1, 3, and 5 seconds—to ensure a fair comparison.

Baselines. We compare our approach with two representative methods: a Kalman filter-based predictor [45] and an LSTM-based neural predictor [46].

Results. As summarized in Table 8, UAVsTrajectory achieves the lowest ADE and FDE across all comparisons, demonstrating the benefit of fusing complementary modalities for trajectory prediction. The Kalman filter-based method, constrained by its non-data-driven nature, struggles with complex and nonlinear UAV motion patterns. The LSTM-based model learns motion dynamics from data but lacks appearance cues, leading to inaccurate predictions when UAV posture changes significantly.

Future work will focus on incorporating additional sensing modalities (e.g., radar, infrared) to further enhance the robustness and generalization of trajectory prediction in diverse and dynamic environments.

5 CONCLUSION

We present UAV-MM3D, a comprehensive multimodal dataset designed for low-altitude UAV perception, tracking, and short-term trajectory prediction in complex low-altitude environments. To meet the growing demand for high-quality annotated data, the dataset integrates RGB, infrared, depth, acoustic, LiDAR, Radar, and DVS modalities,

covering diverse scenes, weather conditions, and UAV models. It provides rich annotations including 2D/3D bounding boxes, 6-DoF poses, and instance identities, enabling unified benchmarking for key tasks such as detection, tracking, trajectory prediction, and multimodal fusion. We establish robust baselines, including our proposed LGFusionNet, a LiDAR-guided multimodal fusion network, alongside dedicated UAV target tracking and trajectory prediction baselines. Experimental results demonstrate that LGFusionNet substantially improves detection robustness under low illumination, noisy environments, and long-range scenarios, while the UAV target tracking and trajectory prediction baselines highlight the utility of the dataset for proactive low-altitude airspace management. We anticipate that UAV-MM3D will serve as a foundational resource for advancing low-altitude situational awareness and reinforcing the technical infrastructure for low-altitude security and airspace management.

REFERENCES

- [1] K. P. Valavanis and G. Vachtsevanos, *Handbook of Unmanned Aerial Vehicles*. Springer, 2015.
- [2] B. Rao, J. Lopez, and T. Walker, "Drone threats: The technical landscape," *IEEE Security & Privacy*, 2020.
- [3] C. Jiang, H. Wang, Y. Li, S. Gong, and X. Chen, "Anti-uav: A survey of uav detection, tracking, and fight back," *IEEE Access*, 2021.
- [4] M. Guo, X. Huang, J. Zhang, and Q. Chen, "Multimodal sensor fusion for object detection: A survey," *Pattern Recognition*, 2022.
- [5] R. Clothier, D. Greer, D. Jenkins, and S. Harper, "Regulatory challenges of low-altitude uav operations," *Aerospace*, 2020.
- [6] N. Jiang, K. Wang, X. Peng, X. Yu, Q. Wang, J. Xing, G. Li, G. Guo, Q. Ye, J. Jiao *et al.*, "Anti-uav: A large-scale benchmark for vision-based uav tracking," *IEEE Transactions on Multimedia*, vol. 25, pp. 486–500, 2021.

- [7] X. Li, Q. Liang, Y. Liu, X. Wei, Y. Bai, and Y. Zheng, "Dronergbt: A benchmark for rgbt aerial tracking," *IEEE Transactions on Instrumentation and Measurement*, 2022.
- [8] J. Wang, W. Luo, Z. Cao, G. Zhang, Z. Li, and H. Chen, "Mmdrone: Multimodal uav tracking with rgb and event cameras," in *ICRA*, 2023.
- [9] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *FSR*, 2017.
- [10] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [11] X. Ma, H. Chen, Y. Tian, Z. Zhang, and Y. Gao, "mmwave radar for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [12] X. Zhu, T. Xu, J. Zhao, J.-W. Liu, K. Wang, G. Wang, J. Li, Z. Zhang, Q. Wang, L. Jin *et al.*, "Evidential detection and tracking collaboration: New problem, benchmark and algorithm for robust anti-uav system," *CoRR*, 2023.
- [13] K. Rematas, Y. Yang, H. Liu, D. Zhang, and V. Ferrari, "Synthcity: A large-scale synthetic dataset for street scene understanding," in *ICCV*, 2019.
- [14] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Deepgtav: A large-scale video dataset via game simulation," in *CVPR*, 2016.
- [15] B. Ravichandran, A. Kar, and G. Shakhnarovich, "Sim4cv: A photo-realistic simulator for computer vision applications," *International Journal of Computer Vision*, 2021.
- [16] J. Zhao, J. Zhang, D. Li, and D. Wang, "Vision-based anti-uav detection and tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 25 323–25 334, 2022.
- [17] B. Huang, J. Li, J. Chen, G. Wang, J. Zhao, and T. Xu, "Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2852–2865, 2023.
- [18] S. Yuan, Y. Yang, T. H. Nguyen, T.-M. Nguyen, J. Yang, F. Liu, J. Li, H. Wang, and L. Xie, "Mmaud: A comprehensive multi-modal anti-uav dataset for modern miniature drone threats," in *ICRA*. IEEE, 2024, pp. 2745–2751.
- [19] Y. Zhang, J. Deng, P. Liu, W. Li, and S. Zhao, "Domain adaptive detection of mavs: A benchmark and noise suppression network," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 1764–1779, 2024.
- [20] F. Dadboud, H. Azad, V. Mehta, M. Bolic, and I. Mantegh, "Drift: Autonomous drone dataset with integrated real and synthetic data, flexible views, and transformed domains," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 6900–6910.
- [21] Y. Zheng, C. Zheng, J. Shen, P. Liu, and S. Zhao, "Keypoint-guided efficient pose estimation and domain adaptation for micro aerial vehicles," *IEEE Transactions on Robotics*, vol. 40, pp. 2967–2983, 2024.
- [22] W. Zhao, X. Xing, C. Li, Z. Wu, W. Feng, J. Tang, and L. Zhang, "Chase: A multimodal dataset for uav tracking with rgb and thermal imagery," in *ICIP*, 2021.
- [23] S. Hwang, J. Park, N. A. Kim, Y. Kim, and I. S. K. Kim, "Multi-spectral pedestrian detection benchmark dataset," in *CVPR*, 2015.
- [24] W. Jiang, W. Zhang, H. Pan, K. Chen, and R. Li, "Llvp: A visible-infrared paired dataset for low-light vision," *IEEE Transactions on Image Processing*, 2021.
- [25] S. Meyer, G. Kusch, M. Gavrilescu, F. Espeter, R. Koester, M. Rottmann, J. Behley, and C. Stachniss, "Dense: Diverse environmental sensor dataset," in *CVPR*, 2019.
- [26] Z. Xu, J. Gao, S. Zhang, L. Li, C. Liu, C. Wang, and M. Liu, "Rodnet: Radar object detection network for autonomous driving," in *ECCV*, 2020.
- [27] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020, pp. 11 621–11 631.
- [28] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020, pp. 2446–2454.
- [29] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*. IEEE, 2012, pp. 3354–3361.
- [30] B. Wilson, H. Caesar, A. Ranjan, V. Guizilini, R. Ambrus, S. Chen, and R. Urtasun, "Argoverse 2: Multi-sensor dataset for autonomous driving," in *NeurIPS Datasets Track*, 2021.
- [31] J. Mao, M. Niu, C. Jiang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu *et al.*, "One million scenes for autonomous driving: Once dataset," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.
- [32] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Computer Vision – ACCV 2012*, K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 548–562.
- [33] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.
- [34] T. Hodaň, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.
- [35] A. Ahmadyan, L. Zhang, A. Ablavatski, J. Wei, and M. Grundmann, "Objectron: A large scale dataset of object-centric videos in the wild with pose annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7822–7831.
- [36] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield *et al.*, "Dexycb: A benchmark for capturing hand grasping of objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9044–9053.
- [37] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.
- [38] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6d object pose prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 292–301.
- [39] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.
- [40] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [41] B. Ye, H. Chang, B. Ma, J. Wu, and Y. Gong, "Joint feature learning and relation modeling for tracking: A one-stream framework," in *ECCV*, 2022.
- [42] Q. Wu, T. Yang, Z. Liu, D. Zhang, J. Hu, and F. Qiao, "Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks," in *CVPR*, 2023.
- [43] X. Wei, Y. Bai, Y. Zheng, T. Zhang, and H. Wu, "Autoregressive visual tracking," in *CVPR*, 2023.
- [44] Y. Zheng, B. Zhong, Q. Liang, X. Sun, and C. Wang, "Odtrack: Online dense temporal token learning for visual tracking," in *CVPR*, 2024.
- [45] D. Simon, "Kalman filtering," *Embedded systems programming*, vol. 14, no. 6, pp. 72–79, 2001.
- [46] A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.