2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SRCF-UAV: Sparse Radar-Camera Fusion for 3D UAV Detection

Yiming Zhao[1], Zijun Gong[1], Yang Yang[2], and Ying Cui[1,*]

Abstract— **With the rapid development of the low-altitude economy, accurate detection and localization of UAVs have become increasingly important. Conventional radar and visual detection methods have low accuracy, whereas current radar-camera fusion methods are computationally intensive. To overcome these issues, we propose a novel 3D UAV detection approach based on sparse radar-camera fusion, called SRCF-UAV, to achieve high-precision, low-complexity UAV detection in diverse scenarios. Specifically, we first propose an improved query initialization method that incorporates locations from 2D image proposals and radar point clouds. Then, we propose a query update method that sparsely fuses radar and image queries based on features, velocity, and spatial distance. Furthermore, we develop a radar-camera multimodal data collection platform based on real-time kinematic positioning (RTK) and collect a dataset of centimeter-level precision, comprising over 20,000 UAV instances that cover various scenarios, UAV models, and lighting conditions. Finally, extensive experiments on this dataset demonstrate that the proposed approach can achieve an average precision of up to 91.65% and an inference latency as low as 17 ms, validating its effectiveness and efficiency. The dataset and code will be publicly available to support further research.**

I. Introduction

With the rapid growth in civilian and industrial applications, unmanned aerial vehicles (UAVs) have been widely used in various fields such as environmental monitoring, precision agriculture, public safety, and disaster management. Despite their significant benefits, UAVs present critical challenges, including threats to public safety, privacy breaches, and unauthorized intrusions into restricted zones. These challenges demonstrate the importance of developing effective UAV detection and estimation systems, especially for low-altitude, small UAVs, which can hardly be achieved by traditional surveillance technologies.

Object detection can be divided into 2D and 3D approaches. 2D-based methods [1]–[3] only provide object positions in the image plane, whereas 3D-based methods [4]–[21] obtain real-world object locations, sizes, yaws, and velocities. 3D-based methods have received increasing attention due to their dominating effectiveness over the 2D counterparts. Existing 3D object detection methods are developed mainly for autonomous driving [4]–[14], [20], [21] and rarely for UAVs [15]–[19]. 3D UAV detection is fundamentally different from 3D detection of ground objects in autonomous driving for two main reasons. First, ground targets are mostly large, slow moving and have small location variations, whereas UAV targets are typically small, fast moving, and have high location uncertainty. In addition, ground targets, e.g., cars and trucks, have large radar cross sections (RCSs), whereas UAVs often have small RCSs. Therefore, 3D UAV detection is more a challenging sensing and estimation problem, and existing 3D object detection methods for autonomous driving can hardly be transferred to UAV detection with consistent effectiveness.

Most existing 3D object detection methods exploit radar and vision modalities and can be classified into three categories: radar-based, camera-based, and radar-camera fusion-based methods. Specifically, radar-based 3D object detection methods provide estimations of ranges, yaws, and Doppler velocities [15], [16], [20], [21]. However, the accuracy for radar-based 3D UAV detection is usually low, due to UAVs' small RCSs and high speeds. Image-based 3D object detection methods, including BEVFormer, BEVDepth, SparseBEV, Sparse4D, and RayFormer [4]–[8], provide locations, sizes, yaws, and velocities. Nevertheless, their effectiveness is highly dependent on lighting conditions and objective sizes and distances, significantly limiting their 3D detection accuracy for flying UAVs of small sizes and at far distances. Exisiting radar–camera fusion-based 3D object detection methods address the above limitations of single modalities to a certain extent [9]–[14]. Most of these fusion-based methods [9]–[12], [14] project image and radar point cloud into the BEV space and use transformer models to extract dense features, resulting in intensive computation and slow inference. Recently, sparse fusion [13] strategies have been proposed to reduce the computation cost. However, the sparse method has yet to deeply fuse radar and camera features to achieve satisfactory accuracy for 3D UAV detection.

Additionally, in contrast to widely accessible large-scale multimodal datasets for autonomous driving scenarios, such as nuScenes [22], there are very few multimodal datasets specifically designed for UAV applications, to date, there exists only one public radar–camera dataset for 3D UAV, MMAUD [23]. Unfortunately, MMAUD lacks extrinsic calibration parameters and covers limited illumination scenarios, restricting its practical value for developing effective 3D UAV detection methods based on radar–camera fusion.

This paper attempts to bridge these gaps. The main contributions of our work are summarized as follows. (i) First, we propose a novel 3D UAV detection approach based on sparse radar-camera fusion, called SRCF-UAV, to achieve high-precision, low-complexity UAV detection in diverse scenarios. Specifically, we propose an improved query initialization method that incorporates locations from 2D image proposals and radar point clouds. We also propose

[1]Thrust of Internet of Things, Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China.
[2]HKUST Shanghai Center, Shanghai 200232; Peng Cheng Laboratory, Shenzhen 518055; and Terminus Group, Beijing 100027, China.
*Corresponding author: Ying Cui (e-mail: yingcui@hkust-gz.edu.cn).

a query update method that sparsely fuses radar and image queries based on features, velocity, and spatial distance. The proposed methods improve the convergence speed and reduce the computational cost for query refinement and enhance the final query qualities. (ii) Then, we develop a radar-camera multimodal data collection platform based on real-time kinematic positioning (RTK) and collect a dataset of centimeter-level precision that covers various scenarios. In particular, the dataset comprises over 20,000 UAV instances covering various UAV models, environments, and lighting conditions, providing a vital foundation for the advancement of 3D UAV detection methods based on radar–camera fusion. (iii) Finally, we conduct extensive experiments to validate the promising advantages of the proposed SRCF-UAV for 3D UAV detection in accuracy and inference time over the start-of-the-art radar-camera fusion-based methods. Specifically, our gains in accuracy and inference time are up to 91.65% and decrease to 17ms, respectively.

II. Related Work

A. 3D Object Detection in Autonomous Driving Scenarios

3D object detection in autonomous driving scenarios is highly challenging due to object diversities, wide location ranges, and complex conditions. Most research adopts radar-based, image-based, and radar-camera fusion detection methods.

Radar-based methods typically use voxel or grid to represent radar point clouds [20], [21]. SMURF [20] detects 3D objects using a single 4D imaging radar by voxelizing point clouds and projecting them to BEV pseudo-images. Scheiner et al. [21] instead convert radar point clouds into 2D temporal grids with Doppler and positional information for detection.

Image-based 3D object detection methods achieve superior performance by using rich texture information. BEVFormer [4] learns a unified spatial-temporal BEV representation with a deformable-attention transformer and achieves image-based 3D detection and map segmentation. SparseBEV [6] replaces dense BEV grids with a fully sparse query design using scale-adaptive attention and adaptive spatio-temporal sampling, achieving 23.5 FPS with higher accuracy. Sparse4D [7] extends this idea by assigning multiple 4D keypoints per anchor to reduce depth ambiguity, surpassing earlier sparse detectors on nuScenes [22]. RayFormer [8] initializes object queries along camera rays and designs ray-centric feature sampling to reduce ambiguity in multi-camera 3D detection.

To address the limitations of single-modality detection, recent work has explored radar–camera fusion. Early approaches fuse radar points and images in the image branch, such as RRPN [24], which projects radar onto the image plane for region proposals, and CenterFusion [25], which applies frustum-based association. Dense BEV-based methods then enabled deeper cross-modal interaction: CRN [11] constructs a unified BEV with cross-modal attention, RCM [10] aligns features with a radar-guided BEV encoder and grid refinement, HyDRa [12] enhances occupancy prediction, and RCBEVDet [9] aligns dual-stream BEV features via deformable attention. RaCFormer [14] corrects depth before fusion to reduce misalignment. More recently, sparse-query paradigms have emerged. SpaRC [13] leverages sparse frustum fusion with range-adaptive radar aggregation for efficient high-accuracy detection.

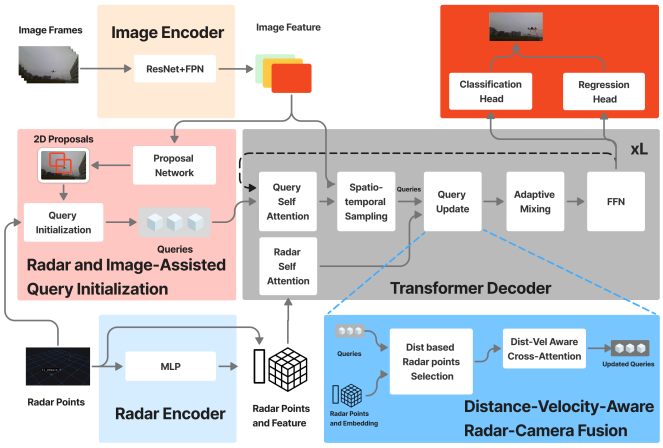

Fig. 1. **SRCF-UAV system overview.** Image and radar features are extracted for radar–image-assisted query initialization. A transformer decoder refines queries via spatial-temporal sampling and distance–velocity-aware radar–camera fusion. Detection heads output object classes and 3D parameters.

B. 3D Object Detection in UAV Scenarios

3D object detection for UAV scenarios is far less explored than for the autonomous driving scenario. [15] is the first work to use micro-Doppler signatures (MDS) for small UAV detection. It applies an LSTM network for target detection and an angle-of-arrival (AOA) method for localization. [16] introduces an integrated detection and tracking strategy to accumulate non-localized trajectories from radar frames. [18] achieves 3D detection, localization, and tracking of malicious MAVs using panoramic stereo camera networks. [17] proposes a two-stage radar–vision fusion framework, which combines radar and visual detections to suppress clutter and accelerate processing and then fuses both for more accurate tracking.

III. SRCF-UAV

We propose a novel 3D UAV detection approach based on sparse radar-camera fusion, called SRCF-UAV, to achieve high-precision, low-complexity UAV detection in diverse scenarios. As Fig.1 illustrates, SRCF-UAV includes five modules: Image Encoder, Radar Encoder, Radar and Image-Assisted Query Initialization, Transformer Decoder (which includes a key component, namely Distance-Velocity-Aware Radar-Camera Fusion) and Detection. Specifically, SRCF-UAV takes image frames and radar points as input and produces object locations, sizes, yaws, and velocities as output. We train the entire neural network from end to end with the same loss function as in SparseBEV [6].

A. Image Encoder and Radar Encoder

First of all, we extract image and radar features using Image Encoder and Radar Encoder, respectively. In Image Encoder, four consecutive image frames are processed by a ResNet backbone with FPN, producing multi-scale feature maps, as in most 3D object detection works [4], [6], [7]. In Radar Encoder, the stack of four frames of radar point clouds is processed by a lightweight MLP, embedding each radar point into a 256-dimensional vector to capture the radar context feature in a sparse way. A radar point can be represented by a 5D vector:

$$\mathbf{p}_r = (x,\ y,\ z,\ rcs,\ v_r)$$

Unlike most dense BEV representations [4], [10], [12], this sparse representation maintains the inherent sparsity of the radar points and is well suited for sparse fusion with image features.

B. Radar and Image Assisted Query Initialization

In the Radar and Image Assisted Query Initialization, we initialize queries using 2D proposals and the stacked radar point cloud and produce reliable queries. Following most query-based 3D object detection works [6], [7], [26]–[28], we model each object with a query and its corresponding context features, together referred to as a query. Specifically, a query for an object can be represented by a 10D vector:

$$\mathbf{p}_q = (x,\ y,\ z,\ w,\ l,\ h,\ \sin\alpha,\ \cos\alpha,\ v_x,\ v_y)$$

containing the object's 3D coordinates (x, y, z), width w, length l, height h, the sine and cosine components of yaw $(\sin\alpha, \cos\alpha)$, and 2D velocity (v_x, v_y) in the x and y directions.

Most existing work initializes queries by generating (x, y) in the BEV view uniformly and setting z at a fixed height value. This query initialization method is reasonable for 3D object detection on the ground, where objects are abundant and of similar altitudes. However, they are ill-suited for 3D UAV detection, where objects are highly sparse and have diverse altitudes. Recently, Rayformer [8] proposed a query initialization method based on 2D image proposals that utilizes only 2D coordinates in the world coordinate system. This query initialization method influences the effectiveness of capturing height information. In summary, existing query initialization methods may not fully increase the chance of covering UAVs in the initial step, reducing the query refinement efficiency.

To increase query refinement efficiency, we propose an improved query initialization method that incorporates locations from 2D image proposals and radar point clouds. As Fig.2 shows, initial queries consist of radar point cloud-based initial queries, 2D image proposal-based initial queries, and uniformly sampled initial queries from the BEV view to increase the chance of covering UAVs in the initial step.

For radar point cloud-based initial queries, we first select high-confidence radar points, filtered by a reasonable velocity range and the camera Field of View (FoV). Then, we set their 3D coordinates $(x,\ y,\ z)$ as the 3D coordinates of the radar point cloud-based initial queries. For 2D image proposal-based initial queries, we extract reliable 2D proposals by a proposal network [29] and project their centers in the 2D space to rays in the 3D space using the intrinsic and extrinsic properties of the camera. Denote the camera's intrinsic matrix by:

$$\mathbf{T} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where f_x, f_y represent the focal lengths in pixels, and c_x, c_y represent the projection of the optical center in the image plane. Specifically, for each 2D proposal center, we convert its image coordinates (u, v) to camera coordinates (x_n, y_n) using the intrinsic matrix of the camera T, where

$$x_n = \frac{u - c_x}{f_x}, \quad y_n = \frac{v - c_y}{f_y}. \tag{2}$$

Then, for the corresponding ray, the direction vector in the camera coordinate system is given by:

$$\mathbf{d}_{\text{cam}} = \begin{bmatrix} x_n & y_n & 1 \end{bmatrix}, \tag{3}$$

and the starting point is the optical center of the camera $(0, 0, 0)$. Transforming the camera coordinate system to the world coordinate system, the ray's direction becomes:

$$\mathbf{d}_{\text{world}} = \mathbf{R} \cdot \mathbf{d}_{\text{cam}}, \tag{4}$$

where $\mathbf{R}$ denotes the camera-to-world rotation matrix. Then, for each 2D proposal, we uniformly sample points along each ray in the world coordinate system and set their 3D coordinates as the 3D coordinates of the 2D image proposal-based initial queries. Unlike Rayformer [8] which utilizes only 2D coordinates in the world coordinate system to form 2D image proposed-based initial queries, here we use 3D coordinates to capture height information. For uniformly sampled initial queries, we generate (x, y) in the BEV view uniformly and associate them with a fixed height value as in most existing works [4], [7], forming the 3D coordinates of the corresponding initial queries.

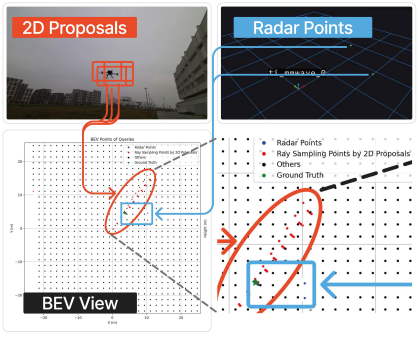

Fig. 2. Query initialization in BEV view. Top left: 2D proposals; top right: radar points; bottom: initial query coordinates from three ways.

Compared to existing query initialization method [8], the proposed query initialization method offers three key advantages: (1) it introduces extra radar information, potentially

increasing the overall query quality, especially in visually degraded scenarios; (2) it utilizes height information in forming 2D image proposal-based initial queries, facilitating the searching for UAVs at various heights; and (3) it uniformly utilizes the 3D coordinates offered by radar point clouds and 2D image proposals, increasing the convergence speed of query refinement. The more effective proposed query initialization method is expected to enhance the performance of query-based 3D object detection.

C. Transformer Decoder

Starting from the initialized queries, Transformer Decoder iteratively updates queries based on image and radar features and produces improved queries after a fixed number of iterations L. It consists of five components: Self-Attention, Spatial-Temporal Sampling [6], Distance-Velocity-Aware Radar-Camera Fusion, Adaptive Mixing [6], and Feed-forward Neural Network (FFN), connected in succession. The main difference from existing works lies in Distance-Velocity-Aware Radar-Camera Fusion, which will be illustrated in detail.

First, Self-Attention, including Query Self-Attention and Radar Self-Attention, refines the context information of queries and radar features, respectively, by multi-head attention, and produces enhanced queries and radar features. Then, Spatial-temporal Sampling [6] efficiently samples spatial-temporal features from multi-frame image features produced by Image Encoder and produces updated queries based on the samples and enhanced queries.

Next, Distance-Velocity-Aware Radar-Camera Fusion sparsely fuses radar features and queries (which have incorporated image and radar features) and produces updated queries. Existing BEV-based radar-camera fusion methods [9], [10], [12] typically project radar points onto the BEV space by CNN or Transformer, requiring intensive computations and introducing unnecessary spatial mappings. Existing query-based radar-camera fusion methods [13] alleviate this problem to some extent, but do not explicitly utilize the velocity information from the radar points, limiting its ability to detect UAVs.

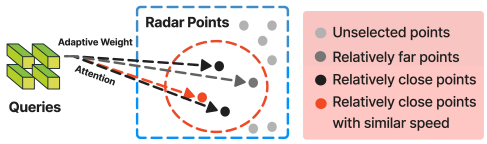

Fig. 3. Adaptive weighting of query–radar point pairs. Red circles: K nearest points; color intensity indicates assigned weight.

To address this issue, we propose a query update method that fuses radar and image queries based on feature similarity, velocity, and spatial distance, as Fig.3 illustrates. Specifically, for each query, we first select the K closest radar points to its 3D coordinates and perform multi-head attention of the query and the selected K radar points. The attention weight between query $\mathbf{p}_q$ with feature vector $\mathbf{f}_q$ and each of its K closest radar points, $\mathbf{P}_r = (\mathbf{p}_{r,k})_{k=1,\cdots,K}$, with feature matrix $\mathbf{F}_r = (\mathbf{f}_{r,k})_{k=1,\cdots,K}$ is calculated with a multi-factor softmax gating mechanism that captures feature similarity, 3D spatial distance, and velocity difference:

$$\begin{aligned}&\mathrm{Attn}(\mathbf{p}_q, \mathbf{P}_r) = \\ &\underset{k=1,\cdots,K}{\mathrm{softmax}}\Big(\frac{\mathbf{f_q}\,\mathbf{F}_r^T}{\sqrt{d}} - \alpha\,\|\mathbf{l}_q - \mathbf{l}_{r,k}\|_2 - \beta\,\|\mathbf{v}_q - \mathbf{v}_{r,k}\|_2\Big)\,\mathbf{F}_r\,.\end{aligned} \tag{5}$$

Here, d is the feature dimension, $\mathbf{l}_q$ and $\mathbf{l}_{r,k}$ represent the 3D coordinates of query $\mathbf{p}_q$ and the k-th closest radar point $\mathbf{p}_{r,k}$, respectively, $\mathbf{v}_q$ and $\mathbf{v}_{r,k}$ represent the radial velocities of query $\mathbf{p}_q$ and the k-th closest radar point $\mathbf{p}_{r,k}$, respectively. The hyperparameters α and β control the influences of distance and velocity. In (5), $\frac{\mathbf{f_q}\,\mathbf{F}_r^T}{\sqrt{d}}$ represents feature similarity, and $\alpha\,\|\mathbf{l}_q - \mathbf{l}_{r,k}\|_2$ and $\beta\,\|\mathbf{v}_q - \mathbf{v}_{r,k}\|_2$ penalize large 3D spatial gaps and velocity gaps, respectively. Thus, a larger distance or velocity gap directly lowers the corresponding softmax attention weight.

The proposed fusion method, encoding the influence of distance and velocity in radar-camera fusion, has three primary benefits: (1) it does not rely on computation-intensive neural networks to introduce additional spatial encoding, significantly reducing the computation cost; (2) it effectively utilizes UAVs' velocity information in fusing radar features and queries, which is extremely critical for UAV detection given their small RCSs; and (3) it selects K closest radar points for each query, reducing the computation cost in attention weight calculation.

We then employ Adaptive Mixing [6] to enhance query features at the channel level and the point level. Finally, FFN, a point-wise two-layer MLP, further enhance query features. The output of FFN is fed back and input to Transformer Decoder in the next iteration until termination. FFN's output in the last iteration is forwarded to Detection.

D. Detection

Finally, Detection uses queries to determine object classes and estimate their 3D coordinates, sizes, yaw angles, and velocities. Following the most advanced 3D detectors [4], [6], we use two lightweight heads: one classification head (a few conv and full connected layers) for object classification and one parallel regression head for 3D parameter estimation.

IV. Radar-Camera Multimodal Data Collection Platform

To support robust evaluation and development of 3D UAV detection in diverse and challenging scenarios, we constructed a high-precision multimodal data collection platform that integrates millimeter-wave radar, camera, and RTK-assisted UAV tracking system. The tracking system can acquire the UAV's real-time position, velocity, and pose, serving as ground truth. By combining time-synchronized radar measurements and visual data, our system enables large-scale and reliable dataset construction for advancing multimodal 3D UAV detection.

A. Experimental Platform

We developed a two-stage data collection platform comprising *(i)* a radar-camera module and *(ii)* an RTK-assisted UAV Tracking module, as illustrated in Fig. 4. The radar-camera module integrates a TI *IWR6843AOPEVM* 60 GHz mmWave radar evaluation board and an *Azure Kinect DK* RGB camera, mounted on a tripod. The radar features a 4Rx $\times$ 3Tx MIMO array with a $120° \times 120°$ (azimuth $\times$ elevation) field of view, streaming 3D point clouds in real time via the TI mmWave SDK [30]. The camera captures RGB frames at 1920×1080 px and 30 fps, with a $90° \times 59°$ field of view [31]. Both sensors are connected to a Linux-based ROS 2 system [32] to ensure temporal synchronization. High-precision ground truth is obtained using two multirotor UAVs—*DJI Mavic 3/4 Enterprise* and *DJI Matrice 350 RTK*—in conjunction with a *DJI D-RTK 3* base station.[1] With RTK enabled, the Mavic 3E and Matrice 350 RTK achieve a precision of 1 cm horizontally and 1.5 cm vertically [33].

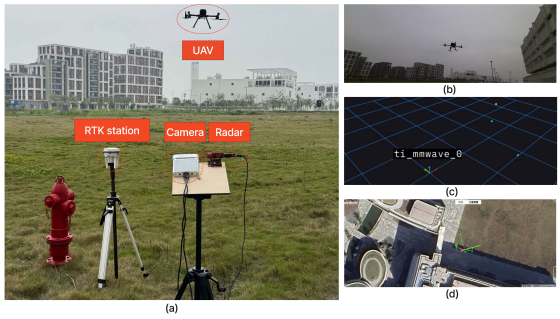

Fig. 4. (a): Data collection platform. (b): Collected image. (c): Radar point cloud. (d): Parsed ground truth from RTK.

B. Data Collection

Our collected dataset consists of about 120,000 time-synchronized radar-camera samples acquired at 30 fps. Recordings span 2 outdoor scenes (an open playground and an rooftop with strong building reflections), 4 lighting scenarios (sunny-daytime, cloudy-daytime, dusk, and night.), 2 pitch angles ($0°$ and $30°$).[2] UAV velocity covers $0-15\,\mathrm{m/s}$ with yaw angles covering $-180°$ to $180°$. During the experiment, the maximum horizontal and vertical distances between the UAVs and the data collection platform are 35 m and 15 m, respectively. To ensure centimeter-level precision, ground truth poses were retained only when DJI flight logs reported *GPSLevel*≥ 5; under these conditions a *Mavic 3E* yields 5 valid pose–frame per second, whereas a *Matrice 350 RTK* logs at 10 Hz. Due to this limitation, we sample key frames at 5 Hz. After quality control, we collected a dataset with about 20000 samples.

C. Data Synchronization

Data synchronization includes both temporal and spatial synchronization. For temporal synchronization, the data acquisition platform operates under ROS, which assigns each radar point cloud and image frame a Coordinated Universal Time (UTC) timestamp; each DJI UAV stores its flight log with matching UTC timestamps; and all devices are synchronized via the Network Time Protocol (NTP). To further refine time synchronization, a corner reflector marker is introduced in every recording session, allowing manual temporal alignment of the sensors. For spatial alignment, intrinsic parameters of camera are obtained from the official SDK [31], while extrinsic parameters of camera and radar are estimated via OpenCV's `solvePnP` algorithm with several pairs of targets. DJI logs 3D coordinates, which are converted to the universal transverse mercator (UTM) coordinate system; the UTM coordinate origin is defined as the location of the data-collection platform. The rigid transformation from the radar and camera frames to the UTM coordinate is then derived through the previous spatial alignment method, ensuring all radar points, camera data, and UAV poses are consistently represented within a unified coordinate system.

V. Experiments

A. Datasets

All experiments are conducted on our collected multi-modal UAV dataset, which contains approximately 20,000 samples.[3] Each sample consists of a key image frame together with three historical image frames, and a radar point cloud formed by stacking four consecutive radar sweeps. Samples with any missing modality are excluded. The dataset is split into training, validation, and test subsets and each subset contains samples from every environment, lighting condition, and UAV.

B. Evaluation Metrics

We adopt most of the evaluation metrics from the nuScenes dataset [22], including mean Average Precision (mAP) and four true positive metrics: ATE (translation error), ASE (scale error), AOE (orientation error), AVE (velocity error). Besides, we introduce AHE (Average Height Error) to evaluate the mean altitude difference between predicted boxes and ground truth.

C. Implementation Details

During the training process, data augmentation is applied only to images. The model is trained using AdamW for 36 epochs with a total batch size of 16 on two GPUs. The learning rate follows a cosine schedule. We evaluate both ResNet-50 and ResNet-101 variants [34] for baselines and proposed models. Specifically, SparseBEV-tiny and SparseBEV-small [6] are adopted as image-based baselines, while our models with ResNet-50 and ResNet-101 backbones are denoted as SRCF-UAV-tiny and SRCF-UAV-small, respectively. All ablation studies and module comparisons are conducted based on SRCF-UAV-tiny. The ResNet-50 based models use input images resized to 320×832, while the

[1] Our dataset covers most DJI enterprise UAVs with RTK capability [33].
[2] Rainy-day scenarios are excluded due to UAV flight safety requirements.
[3] No public radar–camera UAV dataset available for 3D object detection.

TABLE I

3D UAV DETECTION PERFORMANCE COMPARISONS. ↑ INDICATES HIGHER IS BETTER; ↓ INDICATES LOWER IS BETTER. † DENOTES METHODS THAT LEVERAGE MULTI-FRAME INPUTS. TOP: RESNET-50 BACKBONE; BOTTOM: RESNET-101 BACKBONE.

Methods	Modality	Backbone	mAP(%)↑	ATE↓	ASE↓	AOE↓	AVE↓
CenterFusion	Img+Rad	R-50	17.66	0.977	0.460	1.800	1.270
BEVFormer-tiny	Img	R-50	70.10	0.468	**0.057**	**0.898**	2.680
SparseBEV-tiny†	Img	R-50	74.18	0.356	0.317	1.400	2.000
RCM-tiny	Img+Rad	R-50	74.98	0.408	0.282	1.420	2.910
Racformer-tiny†	Img+Rad	R-50	76.32	0.333	0.310	1.240	1.454
SRCF-UAV-tiny†	Img+Rad	R-50	**83.20**	**0.317**	0.298	1.090	**1.030**
BEVFormer-small	Img	R-101	75.67	0.374	**0.063**	**0.880**	2.120
SparseBEV-small†	Img	R-101	76.35	0.336	0.319	1.340	1.920
RCM-small	Img+Rad	R-101	76.56	0.367	0.277	1.350	2.600
Racformer-small†	Img+Rad	R-101	80.22	0.300	0.342	1.270	1.102
SRCF-UAV-small†	Img+Rad	R-101	**91.65**	**0.213**	0.318	1.030	**0.770**

ResNet-101 models use input images of size 640×1664. The transformer decoder is set to $L = 5$ layers. In the radar–camera fusion module, the number of nearest radar points is set to $K = 32$, and the distance and velocity weighting factors are set to $\alpha = 0.5$ and $\beta = 0.5$. For fairness, some LiDAR-dependent methods [7], [11] are excluded from comparison. As the codes of SpaRC [13] and RayFormer [8] are not public, we re-implemented key modules for comparison.

D. Experimental Results

a) Comparison with the State-of-the-art Methods: Table I compares 3D UAV detection results on our dataset: our radar–image fusion models, SRCF-UAV-tiny and SRCF-UAV-small, achieve the highest mAP scores—83.20 % and 91.65 %, respectively—outperforming all image-based and prior radar-camera fusion based methods. By leveraging the distance and velocity of radar point cloud, SRCF-UAV sharply reduces localization error (ATE) and velocity error (AVE) than all baselines, confirming the benefit of proposed initialization method and radar–camera fusion in challenging UAV scenarios.

Among image-based baselines, BEVFormer [4] stands out: its high-capacity backbone and dense BEV grid capture the limited size variation of UAVs, leading to the lowest ASE and a slight AOE edge; yaw estimation depends mainly on image features, and radar offers little extra help here. In contrast, CenterFusion [25] performs worst because UAVs' low RCS yields very sparse radar point cloud, making its frustum-to-point matching unreliable. Finally, both AOE and AVE are higher on our UAV dataset than on driving datasets such as nuScenes [22], mainly because UAV yaw is less visually salient, especially under poor lighting or fast maneuvers, and UAVs are small and fast, making velocity estimation more difficult.

b) Height Estimation: Table II further compares height estimation. SRCF-UAV attains the lowest mean height error (0.212 m), improving upon RCM [10] and SparseBEV [6] by 13% and surpassing BEVFormer [4] by over 31%, confirming its effectiveness in accurate UAV height prediction.

c) Inference Time: As shown in Table III, under identical hardware and input settings, our sparse fusion design attains the shortest forward-pass latency among comparable methods, realizing real-time 3D UAV detection. Although SRCF-UAV processes images with four consecutive frames together with the aligned radar point cloud, the inference time is still lower than other single-frame detectors. Measured latency is the average forward time over 500 samples under NVIDIA RTX 4090, excluding disk I/O and post-processing, to ensure fair comparison.

TABLE II

MEAN ALTITUDE ESTIMATION ERROR (AHE, IN METERS).

Method	mAHE (m)↓
BEVFormer	0.308
SparseBEV	0.243
RCM	0.245
SRCF-UAV	**0.212**

TABLE III

PER-SAMPLE INFERENCE LATENCY ON AN RTX 4090 (BATCH = 1, SINGLE VIEW, SAME INPUT RESOLUTION).

Methods	Image Frames	Latency (ms)
BEVFormer-tiny	1	21
RCM-tiny	1	30
SparseBEV-tiny	4	19
RaCFormer-tiny	4	32
SRCF-UAV-tiny	4	**17**

E. Ablation Studies

Table IV presents a comprehensive summary of module ablations. The complete SRCF-UAV achieves the highest accuracy of 83.2% mAP with the lowest ATE (0.32) and AVE (1.03). Removing Radar+Image Assisted Initialization reduces mAP to 78.8%, while excluding the Distance–Velocity-Aware Fusion leads to the largest drop (75.5% mAP, 1.27 AVE), confirming the necessity of integrating Doppler and spatial cues. Omitting radar self-attention also causes performance degradation.

Table V further provides external comparisons to validate our design. For drone type, the larger M350 RTK achieves 95.4% mAP with lower localization and yaw errors than

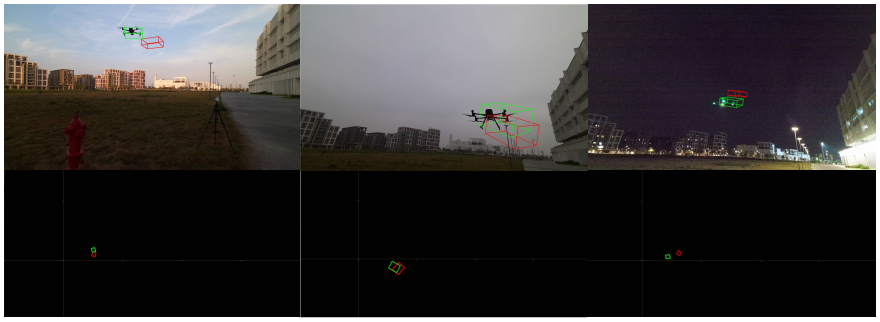

Fig. 5. Qualitative results across sunny, cloudy, and night. Top: 2D image view; bottom: BEV view. Green boxes: ground truth; red boxes: prediction.

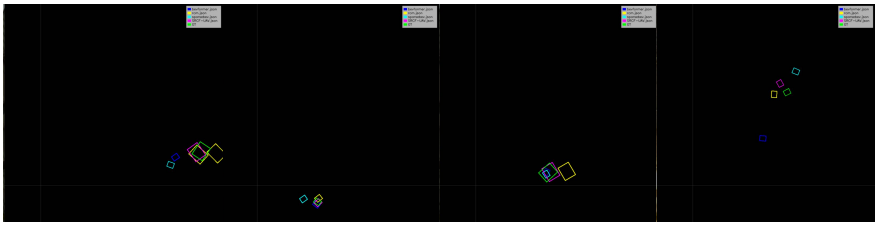

Fig. 6. **Qualitative comparison of BEV detection results.** Green boxes: ground truth; magenta: SRCF-UAV (ours); blue: BEVFormer; yellow: RCM; indigo: SparseBEV.

TABLE IV

ABLATION ON SRCF-UAV MODULES.

Modules	**mAP (%)↑**	**ATE↓**	**AVE↓**
Full model	**83.2**	**0.32**	**1.03**
w/o Radar+Image Init	78.8	0.37	1.16
w/o Dist–Vel–Aware Fusion	75.5	0.41	1.27
w/o radar self-attention	80.8	0.34	1.07

TABLE V

COMPARISONS ACROSS UAV SIZE AND LIGHTING CONDITIONS.

Setting	**mAP (%)↑**	**ATE↓**	**AVE↓**
Drone Type			
M350 RTK (large)	95.4	0.20	0.83
Mavic 3E (small)	80.6	0.30	0.84
Lighting Condition			
Daytime	90.2	0.22	0.66
Night	68.3	0.55	1.98

the smaller Mavic 3E, benefiting from a larger radar cross section. Under different lighting, all metrics degrade at night, with mAP dropping by over 20% and AVE tripling, highlighting the difficulty of low-light UAV detection.

Table VI presents the ablation results on query initialization and radar–camera fusion strategies. For query initialization, uniform sampling yields 78.8% mAP, while RayFormer-style initialization slightly improves to 79.1%. Leveraging 2D image proposals (79.5%) and radar points (80.8%) provides stronger 3D priors, and combining both achieves the best performance at 83.2%. For fusion strategies, conventional cross-attention reaches 76.5%, while range-adaptive aggregation (RAR [13]) increases accuracy to 80.5%. Our proposed Distance–Velocity-Aware Fusion further boosts

TABLE VI

ABLATION ON QUERY INITIALIZATION AND FUSION METHOD.

Methods	**mAP (%) ↑**
Query Initialization	
Uniform initialization	78.8
RayFormer initialization	79.1
2D image proposal initialization	79.5
Radar-based initialization	80.8
Radar+Image assisted initialization	**83.2**
Radar–Camera Fusion	
Conventional cross attention	76.5
RAR	80.5
Distance–Velocity-Aware Fusion	**83.2**

performance to 83.2%, showing that incorporating Doppler velocity improves radar–camera association.

F. Qualitative Results

Fig. 5 presents the qualitative detection outcomes of SRCF-UAV, with the top row showing 3D detection results projected in the 2D image view and the bottom row in the BEV. These examples demonstrate robustness across diverse scenarios, including sunny daytime, cloudy daytime, and night conditions. The UAVs on the left and right are the smaller Mavic 3E models, while the center shows the larger M350 RTK UAV. From the BEV results, it can be observed that the detection accuracy declines in the night scenario, which can be attributed to weak lighting conditions.

As shown in Fig. 6, we qualitatively compare the BEV detection outputs of four detection methods. Green boxes indicate ground truth annotations, while magenta, blue, yellow, and indigo boxes correspond to SRCF-UAV, BEVFormer [4], RCM [10], and SparseBEV [6], respectively. In general, SRCF-UAV's predictions tend to align more closely with

the ground truth. Moreover, in low-light scenes where image quality degrades, the radar-assisted methods, SRCF-UAV and RCM, maintain higher localization accuracy, highlighting the benefit of incorporating radar measurements.

VI. Conclusions

We proposed a novel 3D UAV detection approach based on sparse radar–camera fusion, SRCF-UAV, which effectively utilizes radar and camera data for reliable and efficient UAV detection. We first initialized queries using 2D image proposals and radar points and then fused radar and image features via distance-velocity-guided cross-attention. The proposed approach achieves high precision with significantly lower computational cost than dense BEV methods. Extensive experiments confirmed the effectiveness of SRCF-UAV in various scenarios. In addition, we produced a centimeter-level radar–camera UAV dataset with over 20,000 instances, providing a valuable foundation for future research.

Acknowledgments

This work was supported in part by the National Science and Technology Major Project of China on Mobile Information Networks under Grant 2024ZD1300400, the National Key Research and Development Program of China under Grant 2024YFE0200603, the National Natural Science Foundation of China under Grants 62371412 and 62571467, and Guangdong S&T Program under Grant 2024B0101020004.

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, Jun 2016, pp. 779–788.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, Jun 2014, pp. 580–587.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," May 2020, arXiv:2005.12872.

[4] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 2020–2036, Mar 2025.

[5] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proc. AAAI*, Jun 2023, pp. 1477–1485.

[6] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang, "SparseBEV: High-performance sparse 3D object detection from multi-camera videos," in *Proc. ICCV*, Oct 2023, pp. 18 534–18 544.

[7] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su, "Sparse4d: Multi-view 3d object detection with sparse spatio-temporal fusion," Feb 2023, arXiv:2211.10581.

[8] X. Chu, J. Deng, G. You, Y. Duan, Y. Li, and Y. Zhang, "RayFormer: Improving query-based multi-camera 3D object detection via ray-centric strategies," in *Proc. ACM MM*, Oct 2024, pp. 4620–4629.

[9] Z. Lin, Z. Liu, Z. Xia, X. Wang, Y. Wang, S. Qi, Y. Dong, N. Dong, L. Zhang, and C. Zhu, "RCBEVDet: Radar-camera fusion in bird's eye view for 3D object detection," in *Proc. CVPR*, Jun 2024, pp. 14 928–14 937.

[10] J. Kim, M. Seong, G. Bang, D. Kum, and J. W. Choi, "RCM-Fusion: Radar–camera multi-level fusion for 3d object detection," in *Proc. IEEE ICRA*, May 2024, pp. 18 236–18 242.

[11] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, "CRN: Camera radar net for accurate, robust, efficient 3D perception," in *Proc. ICCV*, Oct 2023, pp. 17 569–17 580.

[12] P. Wolters, J. Gilg, T. Teepe, F. Herzog, A. Laouichi, M. Hofmann, and G. Rigoll, "Unleashing HyDRa: Hybrid fusion, depth consistency and radar for unified 3D perception," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2025, pp. 7467–7474.

[13] P. Wolters, J. Gilg, T. Teepe, F. Herzog, F. Fent, and G. Rigoll, "Sparc: Sparse radar–camera fusion for 3d object detection," Nov 2025, arXiv:2411.19860.

[14] X. Chu, J. Deng, G. You, Y. Duan, H. Li, and Y. Zhang, "RaCFormer: Towards high-quality 3D object detection via query-based radar-camera fusion," in *Proc. CVPR*, Jun 2025, pp. 17 081–17 091.

[15] Y. Sun, S. Abeywickrama, L. Jayasinghe, C. Yuen, J. Chen, and M. Zhang, "Micro-Doppler Signature-Based Detection, Classification, and Localization of Small UAV With Long Short-Term Memory Neural Network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6285–6300, Aug 2021.

[16] X. Fang, M. He, D. Huang, Z. Zhang, L. Ge, and G. Xiao, "JTEA: A Joint Trajectory Tracking and Estimation Approach for Low-Observable Micro-UAV Monitoring With 4-D Radar," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024.

[17] F. Vitiello, F. Causa, R. Opromolla, and G. Fasano, "Radar/visual fusion with fuse-before-track strategy for low altitude non-cooperative sense and avoid," *Aerosp. Sci. Technol.*, vol. 146, p. 108946, Mar 2024.

[18] Y. Zheng, C. Zheng, X. Zhang, F. Chen, Z. Chen, and S. Zhao, "Detection, Localization, and Tracking of Multiple MAVs With Panoramic Stereo Camera Networks," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 2, pp. 1226–1243, Apr 2023.

[19] G. Wu, F. Zhou, K. Kit Wong, and X.-Y. Li, "A Vehicle-Mounted Radar-Vision System for Precisely Positioning Clustering UAVs," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 10, pp. 2688–2703, Oct 2024.

[20] J. Liu, Q. Zhao, W. Xiong, T. Huang, Q.-L. Han, and B. Zhu, "Smurf: Spatial multi-representation fusion for 3d object detection with 4d imaging radar," Jun 2023, arXiv:2307.10784.

[21] N. Scheiner, F. Kraus, F. Wei, B. Phan, F. Mannan, N. Appenrodt, W. Ritter, J. Dickmann, K. Dietmayer, B. Sick, and F. Heide, "Seeing around street corners: Non-line-of-sight detection and tracking In-the-wild using doppler radar," in *Proc. CVPR*, Jun 2020, pp. 2065–2074.

[22] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A Multimodal Dataset for Autonomous Driving," in *Proc. CVPR*, Jun 2020, pp. 11 618–11 628.

[23] 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 *Proc. IEEE ICRA*, May 2024, pp. 2745–2751.

[24] R. Nabati and H. Qi, "RRPN: Radar region proposal network for object detection in autonomous vehicles," in *Proc. IEEE ICIP*, Sep 2019, pp. 3093–3097.

[25] ——, "CenterFusion: Center-based radar and camera fusion for 3D object detection," in *Proc. WACV*, Jan 2021, pp. 1526–1535.

[26] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. CoRL*, Jan 2022, pp. 180–191.

[27] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," in *Proc. ECCV*, Oct 2022, pp. 531–548.

[28] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "PETRv2: A unified framework for 3D perception from multi-camera images," in *Proc. ICCV*, Oct 2023, pp. 3239–3249.

[29] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High Quality Object Detection and Instance Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, pp. 1483–1498, Jun 2019.

[30] Texas Instruments, "TI mmWave SDK," [Online]. Available: https://www.ti.com/tool/MMWAVE-SDK, 2025.

[31] Microsoft Corporation, "Azure Kinect DK Hardware Specification," [Online]. Available: https://learn.microsoft.com/en-us/previous-versions/azure/kinect-dk/hardware-specification, 2025.

[32] Open Source Robotics Foundation, "Robot Operating System (ROS)," [Online]. Available: https://www.ros.org/, 2025.

[33] DJI, "DJI D-RTK 3 High-Precision GNSS System," [Online]. Available: https://enterprise.dji.com/cn/d-rtk-3, 2025.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun 2016, pp. 770–778.