In: AIAA SciTech Forum 2023, 23-27 January 2023, National Harbor, Maryland, USA
DOI:10.2514/6.2023-1076

Object Detection for Ground-based Non-cooperative Surveillance in Urban Air Mobility Utilizing Lidar-Camera Fusion

Cheng Huang*, Ivan Petrunin† and Antonios Tsourdos ‡
School of Aerospace, Transport and Manufacturing, Cranfield University, Bedfordshire, MK43 0AL, United Kingdom

Public safety and security are critical components in the Concept of Operations (ConOps) for Urban Air Mobility (UAM). The potential flight conflicts posed to air and ground objects need to be assessed, especially near critical regions and infrastructures, e.g. vertiports. In this sense, all targets, whether cooperative or non-cooperative air and ground targets, should be detected and tracked for conflict and risk assessment. To achieve this goal, ground-based non-cooperative sensors like cameras and lidar are utilized for situational awareness in this paper. In addition, a multi-modal dataset that contains both air and ground objects is constructed in different illumination and foggy weather scenarios. Finally, a lidar-camera fusion framework with multi-resolution voxelization and depth map learning is proposed for data-driven object detection. Experiments on the constructed dataset show the failure of existing lidar-based backbones in learning extremely sparse points, as a comparison, the fusion framework is outstanding in distinguishing air and ground objects, meanwhile, enabling resilient detection in various lighting and clearance conditions.

I. Introduction

The rapid evolving electric vertical takeoff and landing (eVTOL) technique promotes the demand for air-based cargo delivery, air metro, and air taxi services. However, one of the constraints for the wide-ranging deployment of UAM is the public concerns about flight safety and risks posed to people and property on the ground. To achieve safe and robust operations in urban air mobility (UAM) operation environment (UOE), it requires that Unmanned Aerial Systems (UAS) must fly safely and avoid any UAS-to-UAS or UAS-to-ground collisions [1]. In addition, current and future safety risks are supposed to be determined. To estimate the airborne separation conflict, off-nominal trajectory, and potential third-party risk, real-time detection, tracking, and prediction are critical tasks to be performed.

The aforementioned basic tasks of UAM can be fulfilled by the surveillance technique, which is of great importance to operation management, collision avoidance and public safety [2], and also a critical measure to respond to any crisis. In UAM, surveillance is expected to be achieved by distributed systems. The en-route safety is assessed by airborne cooperative and non-cooperative surveillance. But for low-altitude airspace above ground, the participation of ground targets, e.g. pedestrians and ground vehicles, pose challenges to flight operation. As the risk assessment should be performed and safety must be ensured, it is not enough to rely on onboard surveillance alone and is necessary to provide global situational awareness over populated areas, especially near the aerodrome.

The importance of ground-based surveillance is also emphasized in vertiport functional requirements for monitoring objects surrounding the vertiports [3][4]. Especially, Vertiport Operations Area (VOA) and Vertiport Volume (VPV) as well as the vertiport surface are major interested regions. Within VOA, fleet operators coordinate with the Provider of Service to UAM (PSU), whereas flight operations in VPS must follow instructions from vertiport operators. Necessary tasks like negotiation, nominal and off-nominal scenarios motoring, etc. within VPV range are supported by a Vertiport Automation System (VAS), which is also expected to manage vertiport resources, e.g. landing pad availability, hazard identification, and aircraft conformance, etc. much more efficiently with the sensing capabilities.

In detail, the vicinity of the vertiport is vital to be monitored with effective surveillance techniques, because (1). take-off and landing stages are recognized as high-risk phases for a flight [5]; besides, (2). the landing pads or vertiports are mainly located in metropolitan regions for the convenience of passengers' transit and package delivery. As depicted in Fig.1, within the boundaries of VOA and VPV, not only cooperative aircraft, non-cooperative aircraft, and ground targets are possibly involved in the operational environments. Only with the global sensing around the termination, it

*PhD student, Centre for Autonomous and Cyberphysical Systems, Cranfield University
†Senior lecturer, Centre for Autonomous and Cyberphysical Systems, Cranfield University
‡Professor, Centre for Autonomous and Cyberphysical Systems, Cranfield University

Published by [Publisher]. This is the Author Accepted Manuscript issued with: Creative Commons Attribution Non-Commercial License (CC:BY:NC 4.0).
The final published version (version of record) is available online at DOI:10.2514/6.2023-1076. Please refer to any applicable publisher terms of use.

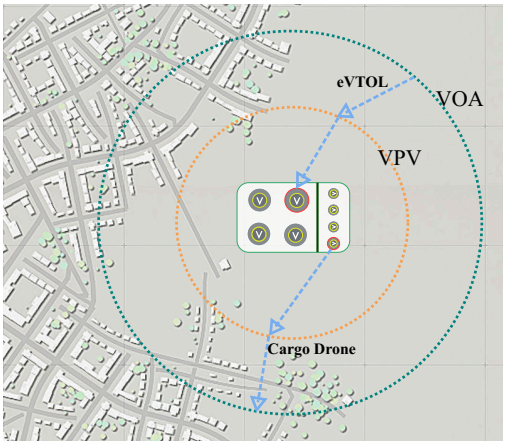


Fig. 1 Operation areas around the vertiport.

becomes promising to cooperate with ground infrastructure such as traffic lights to ensure public safety. The traffic lights could be set to red when the drone is departing or landing in case anyone crosses the vertiport terminal area.

The application scenarios are reviewed above. As for specific surveillance techniques, cooperative sensors can only work well with cooperative targets, whereas ground targets and non-cooperative air objects can not have a response. To achieve high-level safety and security, all air and ground targets near critical areas should be tracked. This function can be ensured by ground-based non-cooperative surveillance. For instance, non-cooperative sensors such as infrared and optical sensing, acoustic detection, and radars are good candidates for heterogeneous measurements.

In this work, we consider complementary lidar and camera signals to achieve resilient non-cooperative object detection in challenging environments since lidar can achieve millimeter-level range accuracy and complement the degradation of the camera in low illumination conditions. Besides, because of lacking the public dataset for aerial and ground object detection in UAM, ground, and air targets are simulated within the VPV region of a vertiport to construct a lidar-camera dataset. After investigating the challenge of employing lidar benchmarks to detect far-distance and small objects in the limited number of sparse points, an effective lidar-camera fusion framework based on 2D convolutions is proposed.

The contributions of this paper can be concluded as follows:

1) A lidar-camera fusion framework is proposed. Due to the difficulty of detecting small objects in the sparse point cloud, we voxelize the point cloud to multiple levels and then project it onto the image frame to obtain depth images. In this way, 2D convolutions can be applied for high-efficiency detection.
2) The spatial encoding based on depth learning enhances the distinction of air and ground objects from the lidar scanning.
3) The performance evaluation is conducted on a range of airborne and ground non-cooperative targets, e.g. vehicles, pedestrians, and drones in the vicinity of vertiports, considering various environmental challenges, e.g. low illumination and high-density fog.

The rest of the paper is organized as follows: Section II discusses various benchmarks about multi-sensor fusion for ground and aerial tasks. The proposed lidar-camera fusion framework is illustrated in Section III. Section IV analyses the results on the collected dataset. Section V concludes the paper.

II. Related Work

Ground-based non-cooperation surveillance is the basic support for monitoring flight trajectories and potential air-air and air-ground risks near important areas. For risk assessment, most of the cases focus on general and large areas instead of specific instances. Even for individual risk assessment, the location of a ground object is required to estimate the individual collision risk of an air crash [6]. There is an Off-Nominal Trajectory and Impact Point Prediction module [7] that can estimate the flight trajectory based on the flight dynamics. Little has been investigated for position prediction of both ground and air objects, even if object detection is the first step for model-free trajectory and interaction prediction. Some fusion-based object detection methods in autonomous driving and aerospace are then presented.

A. Lidar-Camera Fusion for Autonomous Driving

Object detection with multi-sensor fusion is widely applied in autonomous driving, especially with lidar-camera fusion. To locate and classify various ground objects dynamically, point-based fusion methods fuse the point-wise semantics at the feature level; multi-view-based fusion generates proposals from projected bird-eye-view (BEV) and utilizes conventional 2D convolution for detection, in addition, voxel-based fusion algorithms combine voxel-wise features and leverage 3D convolution for feature aggregation [8]. VoxelNet [9] and PointPillars [10] are typical frameworks for lidar-based detection. VoxelNet encodes points in discrete voxels, the issue is that the extracted information would be limited when points in each voxel are extremely sparse. And PointPillars separates the space into pillars with the unlimited spatial extent in the height direction, which would mix ground objects and aerial objects inside the same pillar. Therefore, VoxelNet and PointPillars are not able to distinguish the air and ground vehicles in challenging sparse point clouds. Other lidar-based detection backbones like CenterNet [11] also rely on the standard VoxelNet and PointPillars, as the result, those backbones can not avoid corresponding drawbacks.

Multi-modal fusion frameworks are also developed with standard backbones for lidar detection. Several fusion approaches are concluded in Table 1. Whether point and voxel fusion in MVX-Net [12], or attention mechanism in TransFusion [13] and DeepFusion [14], they are only applicable to ground, near-distance, and large-size objects. The sensors are mounted on the ego-vehicle to capture close-distance objects for autonomous driving tasks, whereas sensors for monitoring the surrounding environments of the vertiport are deployed statically and far from the moving targets, to make current fusion frameworks effective for our task.

Table 1 Comparison of Multi-modal Detection Backbones.

Backbone	Modality	Pros	Cons
VoxelNet [9]	Lidar	Employ 3D convolution	Slow speed; sparse point
PointPillars [10]	Lidar	Employ 2D convolution	Pillar with unlimited height
CenterNet [11]	Lidar	Local peak extraction	Rely on standard lidar backbones
MVX-Net [12]	Lidar+Camera	Point and voxel fusion	Rely on standard lidar backbone
TransFusion [13]	Lidar+Camera	Transformer with attention mechanism	Rely on standard lidar backbone
DeepFusion [14]	Lidar+Camera	Cross-attention	Rely on standard lidar backbones

B. Multi-Sensor Fusion for Air Objects

For drone detection, the common sensors are camera, infrared camera, radio frequency (RF), radar, etc. Single-modality measures use optical and infrared images to detect low-altitude small drones effectively in different scenarios with YOLOv4 [15]. And detection using sparse lidar points is challenging as the number of reflected points from the target is not sufficient to recognize the object [16]. For multi-modality fusion, the fusion of RF data and images performed by conventional 2D convolutional neural network [17] and spatiotemporal information extraction shows the feasibility of data-driven based low-contrast target detection [18]. But little work has focused on lidar-camera for air object detection.

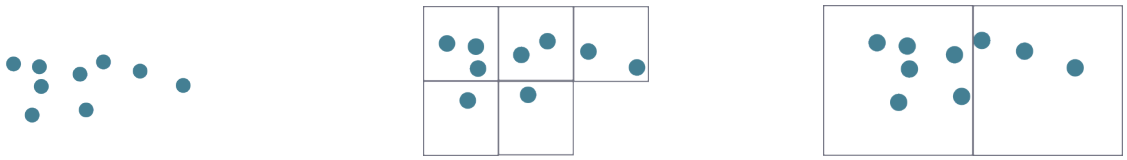
III. Sensor Fusion Framework

The lidar-camera fusion framework is proposed to detect ground and aerial targets in challenging visibility and illumination conditions. The preliminaries about point cloud voxelization and depth map generation are illustrated first, then the detailed architecture is described.

A. Preliminaries

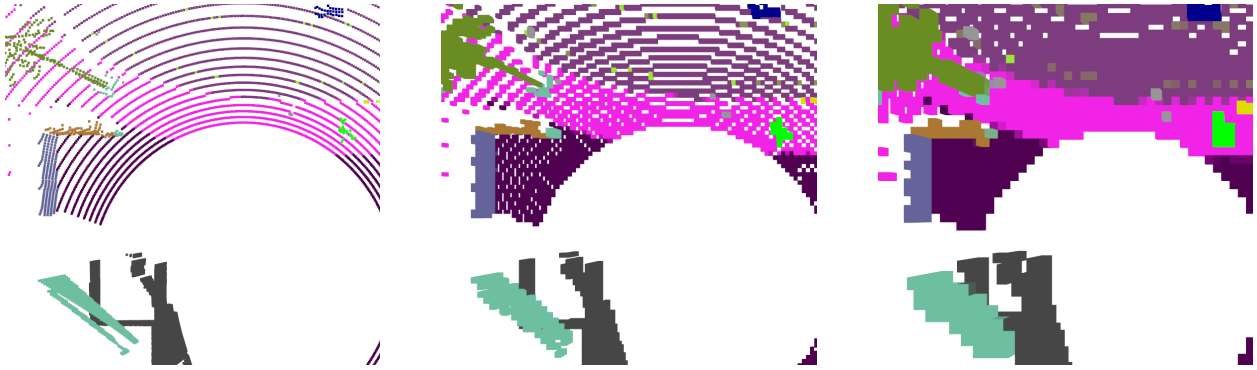
1. Point Cloud Voxelization

As the point cloud is very sparse, especially for far-distance air objects, we voxelize the point cloud to increase its density. Instead of dividing the 3D space into equal grids as in VoxelNet, only valid points are expanded to voxels with a specific size in this work. $p_i = [x_i, y_i, z_i]$ is one point in the lidar frame, and the reflective points from one object can be grouped by a set of points $O_k = \{p_1, p_2, \cdots, p_n\}$. When applying the voxelization, n points belonging to the object



(a) Original points. **(b) Voxelized grids** ($s = 0.5\ m$). **(c) Voxelized grids** ($s = 1.0\ m$).

Fig. 2 Voxelization process.



(a) Original point cloud. **(b) Voxelization** ($s = 0.5\ m$). **(c) Voxelization** ($s = 1.0\ m$).

Fig. 3 Voxelization of the point cloud.

will be represented by the voxel set $V = \{v_1, \cdots, v_m\}$ with a uniform size s [19]. This process can be explained with a simple example shown in Fig. 2 [20]. When the voxel size is set to $s = 0.5\ m$, 8 points can be approximated by 5 voxels, and similarly represented by 2 larger voxels if $s = 1.0\ m$. Even if the number of voxels degrades with the voxel size, it is obvious that valid points become visible and easier to be encoded with rich spatial attributes. The voxelization results with multiple levels are depicted in Fig. 3, we can observe more explicit information with the rising of voxel size from $s = 0.5\ m$ to $1.0\ m$. One point to note is that the voxel size also has its upper limit since one voxel can not exceed the actual size of the object. In this work, we restrict the maximum size to $1.0\ m$.

2. Depth Map Generation

To project the point $p_i = [x_i, y_i, z_i]$ in a lidar frame onto an image frame, the transformation relationship to get target pixel $[u_i, v_i]$ can be written as follows:

$$d_i \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{camera\ intrinsic \quad K} \underbrace{\begin{bmatrix} \mathbf{r_{3\times3}} & \mathbf{t_{3\times1}} \\ \mathbf{o_{1\times3}} & 1 \end{bmatrix}}_{extrinsic} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} = \mathbf{K_{3\times4}} \cdot \mathbf{T_{4\times4}} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} \tag{1}$$

where d_i is the point depth distance in the camera frame, $\mathbf{T_{4\times4}}$ is the relative location and rotation from the lidar to the camera. And the intrinsic parameters of the realistic camera, e.g. focal length f, pixel size (dx, dy) as well as the image center (u_0, v_0), are usually obtained from the manufacturer or calibration. But for the simulated pinhole camera, the intrinsic matrix $\mathbf{K_{3\times4}}$ can be denoted by Eq. (2) and Eq. (3):

$$\mathbf{K_{3\times4}} = \begin{bmatrix} f & 0 & I_w/2 & 0 \\ 0 & f & I_h/2.0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{2}$$

$$f = \frac{I_w}{2 \cdot tan(fov/2)} \tag{3}$$

where I_w and I_h are the width and height of the image, and fov is the field of view angle of the simulated camera.

To project one voxel onto the image frame, we need to project the eight corners $\{(x_1, y_1, z_1), \cdots, (x_8, y_8, z_8)\}$ of this cube with Eq. (1) simultaneously and obtain pixels $\{(u_1, v_1, d_1), \cdots, (u_8, v_8, d_8)\}$. The four corners of the 2D rectangle transformed from the 3D voxel can be represented by Eq. (4):

$$Rect = \begin{pmatrix} [max(\{u_i\}), max(\{v_i\})], & [min(\{u_i\}), max(\{v_i\})] \\ [min(\{u_i\}), min(\{v_i\})], & [max(\{u_i\}), min(\{v_i\})] \end{pmatrix}, i \in \{1, ..., 8\} \tag{4}$$

The depth value of this rectangle in the image frame is calculated from the average depth $d_{avg} = \left(\sum_{i=1}^{8} d_i\right)/8$. Following this procedure, we can transform all voxels into an image to construct the required depth image.

B. Fusion Architecture

It is challenging to detect all non-cooperative targets in the surveillance system with sparse point clouds and various illumination conditions. However, the lidar information can complement the missing feature in images when the illumination and weather become relatively worse. As a result, we introduce a lidar-camera fusion framework for both air and ground object detection, and the general workflow is drawn in Fig. 4.

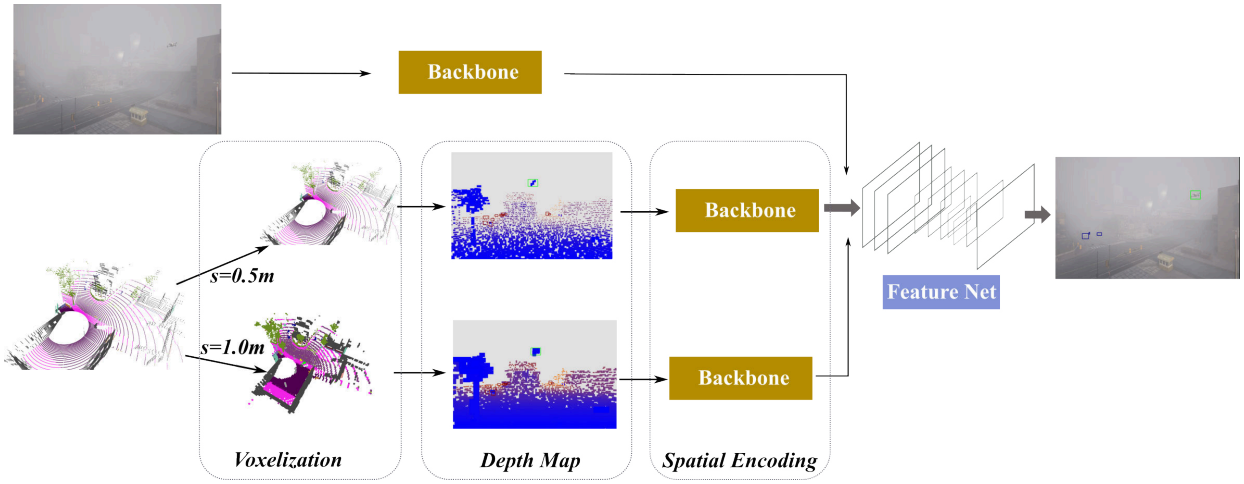


Fig. 4 Lidar-Camera fusion architecture.

For the optical image input, we utilize the image detection backbone to extract visual features. And for the point cloud input, we voxelize the original points with various resolutions. Each voxelized point cloud is projected onto the image frame to get a depth map. The voxel resolution difference will lead to the various density of the depth map. We set the voxel size to be $s = 0.5\ m$ and $s = 1.0\ m$, separately. The standard backbones with 2D convolutions are then applied to multi-resolution depth maps to learn spatial information about altitude and distance, etc. To fuse the features from the optical input and multi-resolution point clouds, another feature network is appended. We select middle components from the standard image detection backbones to be the feature network. Therefore, benefiting from the standard backbone, this fusion architecture can only rely on fast 2D convolutions.

IV. Experiments and Results

In this section, we evaluate the proposed fusion framework on a constructed multi-modal dataset that involves non-cooperative objects, e.g. ground vehicles, pedestrians, and drones with kinds of environmental settings.

A. Dataset

Public datasets, e.g. KITTI [21], Nuscenes [22], and Waymo [23] in Table 2 are only for autonomy research of ground objects such as cars, pedestrians, cyclists, etc. There is no similar open dataset that both contain drone, vehicle, and pedestrian with camera and lidar modalities. To enable non-cooperative surveillance in the urban environment, we

import drone models into Carla Simulator [24] and simulate flights in addition to movements of ground objects. The camera and lidar are deployed near a vertiport. To achieve resilient surveillance in various conditions, the illumination and fog density is set to different values. In particular, the illumination is represented by the solar altitude angle in degrees. An angle value of less than 0 means a dark night, and the brightness increases with the value. In addition, the fog density denotes the fog thickness in meters, and the visibility degrades with the rising value. Finally, we have 48 combinations for generating scenarios, each of which contains 800-frame lidar-camera pairs. We can observe some samples of collected data from Fig. 5 (a) - (j), in kinds of lighting and foggy environments.

Table 2 Comparison of datasets.

Dataset	Frames	Scenarios	Modality	Classes	Night	Foggy
KITTI [21]	15k	multiple	Lidar+Camera	car, pedestrian, cyclist	✗	✗
NuScenes [22]	40k	multiple	Lidar+Camera+Radar	bicycle, pedestrian, bus, etc.	✓	✗
Waymo [23]	198k	multiple	Lidar+Camera	car, pedestrian, cyclist	✓	✗
Carla-UAM	38.4k	urban	Lidar+Camera	drone, pedestrian, ground vehicle	✓	✓

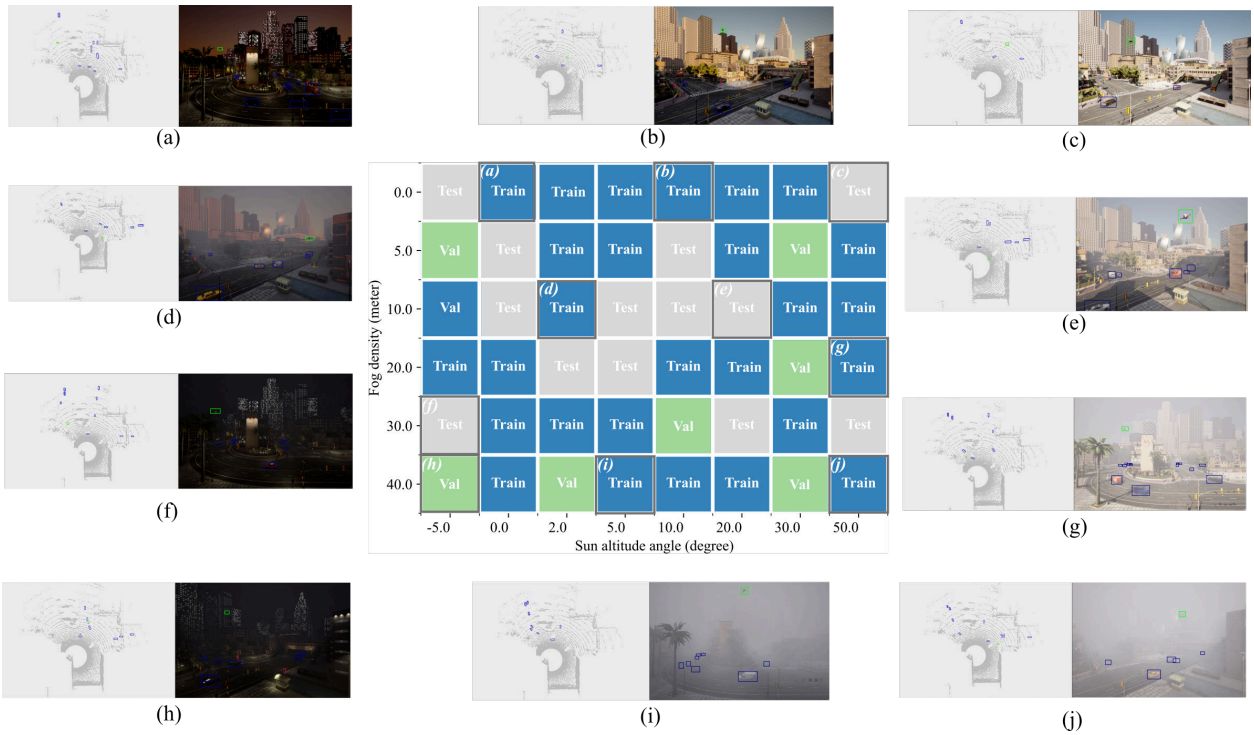


Fig. 5 Samples of Carla-UAM dataset.

Challenging issues in this dataset include:

1) To cover a wide-range view, sensors are mounted on a high-altitude position over the ground, as a result, ground targets are observed relatively small in comparison with capturing from the ego view of ground vehicles. As in Fig. 6(a), the points returned from ground vehicles only start from the 20-meter distance for the simulated 64-beam lidar.
2) Because of the reflective characteristics and remote distance, the average number of points hit on pedestrians is less than 6 as in Fig. 6(b), which makes it challenging to regress accurate bounding boxes.
3) Compared with ground vehicles, drones, of which the average size is 1.5 meters, have a smaller number of lidar points as displayed in Fig. 6(c). In addition, the number of points decreases with the distance increasing. It means that the point cloud is extremely sparse for faraway walkers and drones.

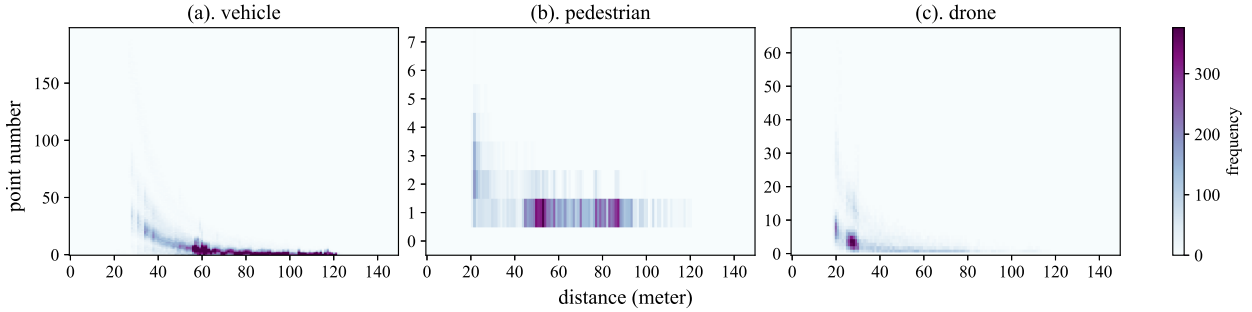


Fig. 6 Number of points belonging to an object (vehicle, pedestrian, drone).

B. Evaluation on Carla-UAM Dataset

We select the YOLOv5 [25] as the backbone of our fusion architecture. All parameters are set to the default YOLOv5-M configurations. The detection results are presented in Table 3. Compared with several approaches, it is evident that our proposed fusion measure surpasses other single-modality and multi-modality methods.

Table 3 mAP(IOU=0.5) comparison for evaluation.

Methods	Input Modality	Vehicle	Pedestrian	Drone	All
YOLOv5 [25]	Camera	0.764	0.15	0.967	0.627
PointPillars [10]	Lidar	0.10	0.0	0.178	0.021
MVX-Net [12]	Lidar+Camera	0.15	0.02	0.10	0.11
Voxel-0.5 (This study)	Lidar	0.709	0.031	0.635	0.458
Voxel-1.0 (This study)	Lidar	0.662	0.006	0.511	0.393
Fusion (This study)	Lidar+Camera	**0.782**	0.080	**0.972**	**0.759**

Typical lidar and camera fusion methods like MVX-Net [12], usually utilize a well-trained detection model to extract features. However, this kind of pipeline can be improvable as this process only works well for close-distance targets with a large number of points. For faraway targets which have a little number of points, the learning process becomes catastrophic, we emphasize this failure according to the training curve in Fig. 7. The training loss contains frequent singularities instead of decreasing steadily. As the consequence, this typical camera-lidar fusion pipeline is not suitable for the surveillance of air and ground objects.

To figure out whether the inside issue is located in the image learning branch or the point learning branch, single-modality methods like camera-based YOLOv5 and lidar-based PointPillars [10] are analyzed. From Fig. 8, although we can observe a reasonably decent learning curve, the actual mean average precision (mAP) of 0.021 reveals the unexpected bad outcome for bounding box regression when IOU(Intersection over Union)=0.5. As a comparison, the average 0.627 precision for image learning with YOLOv5 shows good performance on the Carla-UAM dataset. But it remains an opportunity to improve image-based detection in poor lighting situations. As the consequence, we can conclude that the failed learning in MVX-Net is caused by point cloud learning, which attempts to learn features from raw points.

To tackle issues of point cloud representation and lidar-camera fusion, our deep feature fusion pipeline is evaluated for detecting airborne and ground objects. The point cloud feature learning is refined with the involvement of multi-resolution voxelization, as it is hard to encode features directly from raw sparse points. To figure out the improvement of voxelization, lidar-only data is also trained and evaluated, where Voxel-0.5 means the voxel size is $s = 0.5\ m$ and Voxel-1.0 is $s = 1.0\ m$. When we just train the projected depth maps with the YOLOv5 backbone, the outcome is significant. For all lidar-only approaches, Voxel-0.5 and Voxel-1.0 outperform the PointPillars incredibly, which proves that the learning on projected depth images is better than 3D encoding. One critical issue to mention is that the detection performance for pedestrians is always terrible because of the little number of points belonging to this class. For different voxel sizes, we can know that depth training with voxel size $s = 0.5\ m$ works better than with $s = 1.0\ m$. Ultimately, it shows the effectiveness of learning spatial information when employing the voxelization and depth-generation process for distinguishing ground and air objects.

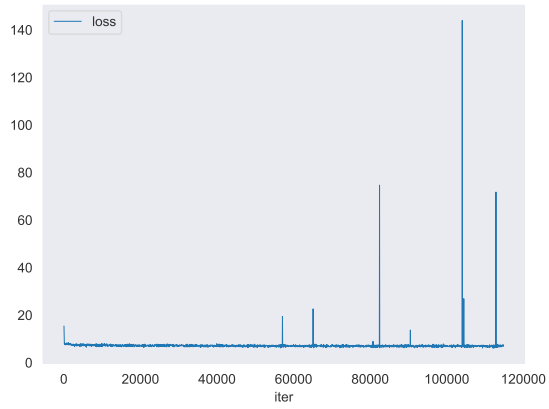


Fig. 7 Failed training of MVX-Net.

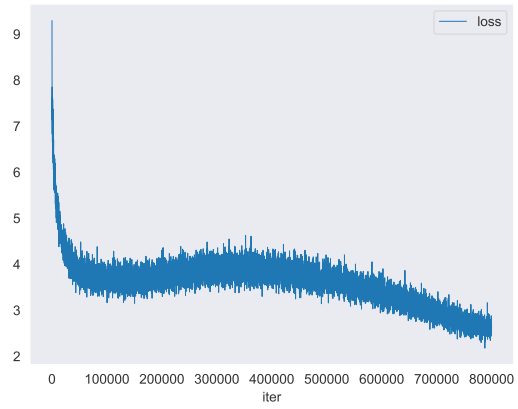


Fig. 8 Failed training of PointPillars.

To analyze the improvement of the fusion framework in specific challenging environmental conditions, we assess the performance when the fog thickness changes at first. As listed in Table 4, the solar altitude angle keeps 30 degrees. There are 800 samples in each test scenario. When applying various approaches, we can find that our fusion measure always outperforms the lidar-only Voxel-0.5 and Voxel-1.0. The performance of the fusion keeps stable with the increasing fog density. We then switch to night, we can also observe similar results as in Table 5.

Table 4 Comparison in Various Fog Density (Day).

Fog	Illumination	Method	Class Precision				mAP(IOU=0.5)			
			Vehicle	Pedestrian	Drone	All	Vehicle	Pedestrian	Drone	All
5	30	Voxel-0.5	0.927	-	0.945	0.936	0.687	-	0.864	0.767
		Voxel-1.0	0.909	-	0.781	0.845	0.681	-	0.632	0.656
		Fusion	**0.947**	-	**0.993**	**0.970**	**0.774**	-	**0.985**	**0.879**
20	30	Voxel-0.5	0.742	-	0.833	0.787	0.589	-	0.47	0.529
		Voxel-1.0	0.68	-	0.664	0.672	0.484	-	0.363	0.423
		Fusion	**0.751**		**0.967**	**0.859**	**0.635**	-	**0.972**	**0.803**
40	30	Voxel-0.5	0.886	-	0.796	0.841	0.748	-	0.498	0.623
		Voxel-1.0	0.696	-	0.557	0.627	0.674	-	0.324	0.499
		Fusion	**0.960**	-	**0.999**	**0.980**	**0.756**	-	**0.975**	**0.865**

Table 5 Comparison in Various Fog Density (Night).

Fog	Illumination	Method	Class Precision				mAP(IOU=0.5)			
			Vehicle	Pedestrian	Drone	All	Vehicle	Pedestrian	Drone	All
5	-5	Voxel-0.5	0.962	1.0	0.749	0.904	0.687	0.0	0.743	0.477
		Voxel-1.0	0.846	1.0	0.651	0.832	0.678	0.0	0.624	0.434
		Fusion	0.962	1.0	**0.853**	**0.971**	**0.880**	**0.04**	**0.767**	**0.563**
40	-5	Voxel-0.5	0.858	0.689	0.828	0.792	0.869	0.065	0.609	0.514
		Voxel-1.0	0.862	0.341	0.836	0.679	0.838	0.020	0.654	0.504
		Fusion	**0.960**	**0.952**	0.751	**0.888**	**0.921**	**0.201**	**0.694**	**0.605**

We then consider changing the illumination condition while keeping a fixed fog density value. With the rising

darkness, we can observe the decreasing performance of fused modalities. For lidar-only comparisons, the detection results are reasonable for various-size targets, which indicates that fusion precision degrades because of the lower light intensity and the lidar-camera system works better under good lighting conditions.

Table 6 Comparison in Various Lighting Settings.

Fog	Illumination	Method	Class Precision			mAP(IOU=0.5)				
			Vehicle	Pedestrian	Drone	All	Vehicle	Pedestrian	Drone	All
40	30	Voxel-0.5	0.886	-	0.796	0.841	0.748	-	0.498	0.623
		Voxel-1.0	0.696	-	0.557	0.627	0.674	-	0.324	0.499
		Fusion	**0.960**	-	**0.999**	**0.980**	**0.756**	-	**0.975**	**0.865**
40	2	Voxel-0.5	0.802	1.0	0.901	0.901	0.797	0.0	0.573	0.456
		Voxel-1.0	0.825	1.0	0.87	0.898	0.76	0.0	0.483	0.414
		Fusion	**0.935**	1.0	**0.990**	**0.975**	**0.854**	0.0	**0.995**	**0.616**
40	-5	Voxel-0.5	0.858	0.689	0.828	0.792	0.869	0.065	0.609	0.514
		Voxel-1.0	0.862	0.341	0.836	0.679	0.838	0.020	0.654	0.504
		Fusion	**0.901**	1.0	**0.987**	**0.963**	0.837	0.0	**0.987**	**0.608**

To sum up, the designed architecture fuses the contextual features from the camera and spatial attributes from the lidar, to achieve better object detection performance in kinds of visibility conditions. Especially, the advantages of camera and lidar are combined to overcome several issues, e.g. weak lighting and sparse point cloud, etc.

V. Conclusion

In this work, the lidar-camera fusion architecture with multi-resolution voxelization is proposed for ground-based non-cooperative surveillance, to achieve robust non-cooperative object detection. In the meanwhile, a multi-modal dataset is constructed for the detection task. Compared with YOLOv5, PoinPillars, and MVX-Net, the kernel issue for small object detection with lidar-based backbones has been analyzed, and it reveals the effectiveness of the voxelization and depth map generation procedure. After evaluating various environmental conditions, our framework shows its feasibility to achieve resilient object detection performance near the vertiport, moreover, the overall 0.759 mAP of which outperforms 0.627 mAP of camera-only detection and other lidar-involved approaches.

This work employs the standard image-based backbones for feature learning. Even though the training is simple and more effective than lidar-based backbones, one limitation is that the prediction is only in the 2D image frame because it is hard to rebuild the 3D shape of small objects in sparse point clouds. Future work includes recovering the accurate 3D position with pairwise cameras which have overlapping observation areas.

Acknowledgments

This research was partially supported by grants from the Funds of China Scholarship Council (202008420248).

References

[1] Roy, S., and Xue, M., "Cyber-Threat Mitigation in an Unmanned Aircraft System (UAS) Enabled Airspace System: A Multi-Scale Dynamical Network Approach," *AIAA AVIATION 2021 FORUM*, 2021, p. 2335.

[2] Stouffer, V. L., Cotton, W. B., DeAngelis, R. A., Devasirvatham, D. M., Irvine, T. B., Jennings, R. E., Lehmer, R. D., Nguyen, T. C., Shaver, M. A., and Bakula, C. J., "Reliable, Secure, and Scalable Communications, Navigation, and Surveillance (CNS) Options for Urban Air Mobility (UAM)," 2020, p. 57.

[3] (NUAIR), N. U. A. I. R. A., "High-Density Automated Vertiport Concept of Operations," Tech. rep., NUAIR, Crown Consulting Inc., Mosalc ATM, Boeing, Deloitte and 5-Alpha, 2021.

[4] Kleinbekman, I. C., Mitici, M. A., and Wei, P., "eVTOL arrival sequencing and scheduling for on-demand urban air mobility," *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, IEEE, 2018, pp. 1–7.

[5] Straubinger, A., Rothfeld, R., Shamiyeh, M., Büchter, K.-D., Kaiser, J., and Plötner, K. O., "An overview of current research and developments in urban air mobility–Setting the scene for UAM introduction," *Journal of Air Transport Management*, Vol. 87, 2020, p. 101852.

[6] Blom, H. A., and Jiang, C., "Safety risk posed to persons on the ground by commercial UAS-based services," *14th USA/Europe Air Traffic Management Seminar*, 2021.

[7] Ancel, E., Capristan, F. M., Foster, J. V., and Condotta, R. C., "Real-time risk assessment framework for unmanned aircraft system (UAS) traffic management (UTM)," *17th aiaa aviation technology, integration, and operations conference*, 2017, p. 3273.

[8] Cui, Y., Chen, R., Chu, W., Chen, L., Tian, D., Li, Y., and Cao, D., "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[9] Zhou, Y., and Tuzel, O., "Voxelnet: End-to-end learning for point cloud based 3d object detection," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[10] Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., and Beijbom, O., "Pointpillars: Fast encoders for object detection from point clouds," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12697–12705.

[11] Yin, T., Zhou, X., and Krahenbuhl, P., "Center-based 3d object detection and tracking," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11784–11793.

[12] Sindagi, V. A., Zhou, Y., and Tuzel, O., "Mvx-net: Multimodal voxelnet for 3d object detection," *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, 2019, pp. 7276–7282.

[13] Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., and Tai, C.-L., "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1090–1099.

[14] Li, Y., Yu, A. W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q. V., et al., "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17182–17191.

[15] Xing, D., Tsoukalas, A., Giakoumidis, N., and Tzes, A., "Computationally Efficient RGB-T UAV Detection and Tracking System," *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, IEEE, 2021, pp. 1410–1415.

[16] Dogru, S., and Marques, L., "Drone Detection Using Sparse Lidar Measurements," *IEEE Robotics and Automation Letters*, 2022.

[17] Aledhari, M., Razzak, R., Parizi, R. M., and Srivastava, G., "Sensor fusion for drone detection," *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, IEEE, 2021, pp. 1–7.

[18] Xie, J., Gao, C., Wu, J., Shi, Z., and Chen, J., "Small Low-Contrast Target Detection: Data-Driven Spatiotemporal Feature Fusion and Implementation," *IEEE Transactions on Cybernetics*, 2021.

[19] Liu, Z., Tang, H., Lin, Y., and Han, S., "Point-voxel cnn for efficient 3d deep learning," *Advances in Neural Information Processing Systems*, Vol. 32, 2019.

[20] Xu, Y., Tong, X., and Stilla, U., "Voxel-based representation of 3D point clouds: Methods, applications, and its potential use in the construction industry," *Automation in Construction*, Vol. 126, 2021, p. 103675.

[21] Geiger, A., Lenz, P., Stiller, C., and Urtasun, R., "Vision meets Robotics: The KITTI Dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[22] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O., "nuScenes: A multimodal dataset for autonomous driving," *CVPR*, 2020.

[23] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., and Anguelov, D., "Scalability in Perception for Autonomous Driving: Waymo Open Dataset," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V., "CARLA: An Open Urban Driving Simulator," *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[25] Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., Michael, K., TaoXie, Fang, J., imyhxy, Lorna, Yifu), Wong, C., V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, UnglvKitDe, Sonck, V., tkianai, yxNONG, Skalski, P., Hogan, A., Nair, D., Strobel, M., and Jain, M., "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation," , Nov. 2022. https://doi.org/10.5281/zenodo.7347926, URL https://doi.org/10.5281/zenodo.7347926.

Cranfield University

CERES https://dspace.lib.cranfield.ac.uk

School of Aerospace, Transport and Manufacturing (SATM) Staff publications (SATM)

2022-01-19

Object detection for ground-based non-cooperative surveillance in urban air mobility utilizing lidar-camera fusion

Huang, Cheng

AIAA

Huang C, Petrunin I, Tsourdos A. (2023) Object detection for ground-based non-cooperative surveillance in urban air mobility utilizing lidar-camera fusion. In: AIAA SciTech Forum 2023, 23-27 January 2023, National Harbor, Maryland, USA. Paper number AIAA 2023-1076
https://doi.org/10.2514/6.2023-1076

Downloaded from Cranfield Library Services E-Repository