

Article

Multimodal Fusion Attention Network for Real-Time Obstacle Detection and Avoidance for Low-Altitude Aircraft

Xiaoqi Xu ^{1,*} and Yiyang Zhao ²¹ School of Law and Business, Shaoyang University, Shaoyang 422000, China² Institute of Internet Industry, Tsinghua University, Beijing 100190, China; zhaoyiyang@gmail.com

* Correspondence: 3164@hnsyu.edu.cn

Abstract

The rapid expansion of low-altitude unmanned aerial vehicles demands robust obstacle detection and avoidance systems capable of operating under diverse environmental conditions. This paper proposes a multimodal fusion attention network that integrates visual imagery and Light Detection and Ranging (LiDAR) point cloud data for real-time obstacle perception. The architecture incorporates a bidirectional cross-modal attention mechanism that learns dynamic correspondences between heterogeneous sensor modalities, enabling adaptive feature integration based on contextual reliability. An adaptive weighting component automatically modulates modal contributions according to estimated sensor confidence under varying environmental conditions. The network further employs gated fusion units and multi-scale feature pyramids to ensure comprehensive obstacle representation across different distances. A hierarchical avoidance decision framework translates detection outputs into executable control commands through threat assessment and graduated response strategies. Experimental evaluation on both public benchmarks and a purpose-collected low-altitude obstacle dataset demonstrates that the proposed method achieves 84.9% mean Average Precision (mAP) while maintaining 47.3 frames per second (FPS) on Graphics Processing Unit (GPU) hardware and 23.6 FPS on embedded platforms. Ablation studies confirm the contribution of each architectural component, with cross-modal attention providing the most substantial performance improvement.

Keywords: multimodal sensor fusion; attention mechanism; obstacle detection; low-altitude aircraft; deep learning; autonomous navigation

1. Introduction

The rapid proliferation of low-altitude aircraft has fundamentally reshaped multiple industrial sectors over the past decade. Urban air mobility platforms, autonomous delivery drones, and agricultural spraying vehicles now operate in increasingly congested airspaces, where traditional air traffic management paradigms are inadequate [1]. This technological transformation demands robust perception systems capable of identifying and responding to dynamic obstacles under diverse environmental conditions. Unlike conventional manned aviation, low-altitude Unmanned Aerial Vehicles (UAVs) frequently encounter unpredictable hazards including power lines, birds, temporary structures, and other aircraft operating without transponders [2].

Flight safety critically hinges upon reliable obstacle detection and avoidance capabilities. Statistical analyses reveal that collision-related incidents account for a substantial

Received: 1 January 2026

Revised: 30 January 2026

Accepted: 31 January 2026

Published: 22 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

proportion of unmanned aircraft accidents, particularly during beyond-visual-line-of-sight operations [3]. The consequences extend beyond equipment damage to encompass potential harm to ground personnel and property, regulatory penalties, and erosion of public trust in autonomous aerial systems. Consequently, developing intelligent perception and avoidance technologies has emerged as a pressing research priority within the aerospace engineering community.

Single-modal sensing approaches dominated early research efforts in this domain. Vision-based systems employing monocular or stereo cameras offered cost-effective solutions but suffered from sensitivity to illumination variations and weather conditions [4]. LiDAR sensors provided accurate depth measurements yet exhibited limitations in detecting small objects and transparent surfaces [5]. Radar technologies demonstrated resilience against adverse weather but lacked the spatial resolution necessary for precise obstacle characterization [6]. These inherent constraints motivated researchers to explore multimodal sensor fusion strategies that could potentially compensate for individual sensor weaknesses.

Contemporary multimodal fusion methodologies have evolved considerably from simple concatenation schemes toward sophisticated feature-level integration architectures. Early fusion approaches combined raw sensor data prior to processing, while late fusion techniques merged independent detection results at the decision stage [7]. More recent investigations favor intermediate fusion strategies that align and integrate features extracted from heterogeneous sensor modalities within deep neural network frameworks [8]. However, achieving effective cross-modal feature alignment remains challenging, particularly when sensor observations exhibit significant disparities in resolution, field of view, and update frequency.

Deep learning has revolutionized object detection performance across numerous application domains. Convolutional Neural Networks (CNNs) such as You Only Look Once (YOLO) and Single-Shot Detector (SSD) families achieve remarkable detection speeds suitable for real-time deployment [9]. Transformer-based architectures have demonstrated superior capability in capturing long-range dependencies and contextual relationships within visual scenes [10]. Nevertheless, directly transplanting these advances to airborne platforms encounters substantial obstacles related to computational resource constraints, power consumption limitations, and the unique characteristics of aerial imagery perspectives.

Attention mechanisms represent a particularly promising avenue for enhancing multimodal fusion effectiveness. Self-attention modules enable networks to dynamically weight feature importance based on contextual relevance rather than fixed spatial arrangements [11]. Cross-attention operations facilitate information exchange between modalities by learning correspondence relationships [12]. Several researchers have incorporated channel attention and spatial attention components into fusion networks, reporting improved detection accuracy in challenging scenarios [13]. Yet, existing attention-based fusion methods often impose considerable computational overhead that conflicts with stringent real-time processing requirements.

Current obstacle avoidance planning algorithms face their own set of difficulties when integrated with perception modules. Classical approaches including potential field methods and rapidly exploring random trees assume perfect obstacle knowledge that perception systems cannot guarantee [14]. The temporal latency between obstacle detection and avoidance maneuver execution creates vulnerability windows where undetected obstacles may cause collisions. Furthermore, most existing frameworks treat perception and planning as sequential modules without feedback mechanisms that would allow planning requirements to influence perception priorities.

Several critical gaps persist in the existing literature that this research endeavors to address. First, perception robustness degrades substantially under complex environmental conditions involving occlusion, adverse weather, and rapidly changing lighting. Second, current multimodal fusion mechanisms fail to adaptively adjust integration strategies based on sensor reliability and scene characteristics. Third, the fundamental tension between detection accuracy and computational efficiency remains inadequately resolved for resource-constrained airborne platforms.

This investigation proposes a novel multimodal fusion attention network architecture specifically designed for low-altitude aircraft obstacle detection and avoidance. Unlike existing fusion methods such as TransFusion [15], which employs unidirectional query-based attention, and CenterFusion [16], which relies on simple feature concatenation without adaptive weighting, our approach introduces bidirectional cross-modal attention that enables mutual feature enhancement between visual and geometric representations. Compared to MVX-Net [17], which uses fixed fusion weights regardless of environmental conditions, we incorporate confidence-based adaptive weighting that automatically adjusts modal contributions based on estimated sensor reliability.

The primary contributions of this work are threefold. First, we introduce an adaptive feature fusion mechanism combining lightweight bidirectional cross-modal attention with confidence-based adaptive weighting and gated fusion units, enabling dynamic modulation of cross-modal integration according to sensor reliability and scene characteristics. This design differs fundamentally from existing approaches that treat all features uniformly without considering contextual reliability. Second, we design a computationally efficient attention module achieving 84.9% mAP while maintaining 47.3 FPS on GPU hardware and 23.6 FPS on embedded platforms, demonstrating compatibility with resource-constrained airborne systems. Third, we construct a real-time avoidance decision framework that tightly couples perception outputs with trajectory planning through threat assessment and graduated response strategies [18], enabling anticipatory behavior that sequential processing approaches cannot achieve. These advances collectively establish a more reliable and efficient foundation for autonomous low-altitude flight operations in complex environments.

2. Related Work

2.1. Multimodal Perception Technology

Low-altitude aircraft deploy diverse sensor modalities to acquire environmental information, each exhibiting distinct operational characteristics suited to particular conditions. Visual cameras capture rich texture and color information at relatively low cost, though their performance deteriorates markedly under poor illumination or adverse weather [19]. LiDAR sensors generate precise three-dimensional point cloud measurements independent of ambient lighting, yet they struggle with specular surfaces and suffer range limitations in fog or heavy rain [20]. Millimeter-wave radar penetrates precipitation and dust effectively while providing velocity estimates through Doppler measurements, but its angular resolution remains inferior to optical alternatives. Ultrasonic sensors offer simplicity and low power consumption for short-range detection, primarily serving as supplementary proximity warning devices.

The complementary nature of these sensing modalities provides the theoretical foundation for multimodal perception systems. Where one sensor fails, another may succeed—cameras distinguish object categories that radar cannot differentiate, while radar maintains functionality when cameras become unreliable [21]. This redundancy enhances overall system robustness beyond what any single modality achieves independently.

Effective multimodal fusion necessitates careful data preprocessing. Image enhancement techniques including histogram equalization and adaptive contrast adjustment im-

prove visual feature extraction under challenging lighting [22]. Point cloud preprocessing involves statistical outlier removal and voxel grid downsampling to reduce noise and computational burden. Critically, spatiotemporal alignment synchronizes asynchronous sensor streams and transforms measurements into a common coordinate frame [23].

The mathematical representation of multimodal sensor data follows established formulations. Let the visual image be denoted as $I \in \mathbb{R}^{H \times W \times 3}$, and the LiDAR point cloud be denoted as:

$$P = \{p_i \mid p_i = (x_i, y_i, z_i, r_i), i = 1, 2, \dots, N\} \quad (1)$$

where r_i represents reflectance intensity. Following coordinate transformation, the unified bimodal observation can be expressed as:

$$O = \mathcal{T}(I, P; \Phi) \quad (2)$$

where $\mathcal{T}$ denotes the alignment transformation parameterized by calibration matrix Φ [24]. It should be noted that while millimeter-wave radar offers complementary capabilities such as weather resilience and velocity estimation, the current implementation focuses exclusively on camera-LiDAR fusion. Radar integration remains a direction for future extension of this framework.

2.2. Attention Mechanism Principles

Attention mechanisms emerged from cognitive science insights suggesting that biological vision systems selectively focus computational resources on salient regions rather than processing entire scenes uniformly. Early soft attention models computed weighted combinations of input features, with weights learned through gradient descent [25]. This paradigm shifted dramatically when researchers introduced self-attention, enabling models to capture dependencies without regard to positional distance—a limitation that has long constrained recurrent architectures.

The self-attention operation computes relationships among all positions within a sequence through query, key, and value projections. Given an input feature matrix $X \in \mathbb{R}^{n \times d}$, linear transformations generate three representations:

$$Q = XW^Q, K = XW^K, V = XW^V \quad (3)$$

where W^Q, W^K , and $W^V \in \mathbb{R}^{d \times d_k}$ denote learnable projection matrices [26]. The attention weights are derived from scaled dot-product similarity:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

The scaling factor $\sqrt{d_k}$ prevents gradient vanishing when dimensionality grows large as this is a standard numerical stabilization technique.

Multi-head attention extends this formulation by executing parallel attention operations across different subspaces. Each head captures distinct relational patterns, and their concatenated outputs undergo a final projection [27]:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (5)$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$. This design enriches representational capacity substantially.

Channel attention and spatial attention address complementary aspects of feature refinement. Channel attention modules recalibrate feature maps by modeling inter-channel dependencies through global pooling operations [28]:

$$\alpha_c = \sigma(W_2 \cdot \delta(W_1 \cdot \text{GAP}(F))) \quad (6)$$

where GAP denotes Global Average Pooling (GAP), δ represents Rectified Linear Unit (ReLU) activation, and σ indicates a sigmoid function. Spatial attention, conversely, emphasizes informative locations by computing position-wise importance scores [29]:

$$\alpha_s = \sigma(\text{Conv}([\text{AvgPool}(F); \text{MaxPool}(F)])) \quad (7)$$

These complementary mechanisms prove particularly valuable for multimodal fusion tasks, where attention can dynamically weight contributions from heterogeneous sources based on contextual reliability [30]. The capacity to learn cross-modal correspondences through attention operations establishes a principled framework for adaptive sensor integration.

2.3. Obstacle Avoidance Planning Algorithms

Classical path planning approaches have established foundational methodologies for autonomous navigation, though each presents distinct tradeoffs between computational efficiency and solution quality. The Artificial Potential Field (APF) method models obstacles as repulsive sources and goals as attractive sinks, generating smooth trajectories through gradient descent on the combined potential function [31]:

$$F_{total} = -\nabla U_{att}(q) - \nabla U_{rep}(q) \quad (8)$$

where U_{att} and U_{rep} represent attractive and repulsive potential functions, respectively, and q denotes the vehicle configuration. This formulation offers computational simplicity but suffers from local minima entrapment in complex environments.

Sampling-based planners such as Rapidly Exploring Random Trees (RRTs) circumvent the local minima problem through stochastic exploration of configuration space [32]. The algorithm exhibits probabilistic completeness, meaning solution probability approaches unity given sufficient iterations. However, generated paths often require post-processing smoothing, and computational demands escalate in high-dimensional spaces. The A* algorithm guarantees optimal discrete paths when admissible heuristics guide node expansion, yet its grid-based representation imposes memory constraints and resolution limitations [33].

Deep reinforcement learning has emerged as a compelling alternative that learns avoidance policies directly from environmental interaction. Agents trained through proximal policy optimization or deep Q-networks demonstrate remarkable adaptability to novel scenarios without explicit environment modeling [34]. The learned policy $\pi_\theta(a|s)$ maps sensor observations to control actions through neural network function approximation, enabling end-to-end decision making.

Dynamic obstacle scenarios demand predictive trajectory planning that anticipates future obstacle positions. The motion prediction module estimates obstacle state evolution:

$$\hat{x}_{obs}(t + \Delta t) = f(x_{obs}(t), v_{obs}(t), \Delta t) \quad (9)$$

where x_{obs} and v_{obs} represent obstacle position and velocity, respectively [35]. Trajectory optimization then incorporates collision avoidance constraints over the prediction horizon:

$$|p_{uav}(t) - \hat{p}_{obs}(t)| \geq d_{safe}, \forall t \in [0, T] \quad (10)$$

The complete avoidance planning problem can be formulated as constrained optimization that minimizes a cost functional subject to dynamic feasibility and safety requirements [36]:

$$\min_{\tau} \int_0^T \mathcal{L}(x(t), u(t)) dt. \text{ s.t. } \dot{x} = f(x, u), g(x) \leq 0 \quad (11)$$

This mathematical framework provides the theoretical basis for integrating perception outputs with planning decisions, though real-time solution remains challenging for resource-constrained platforms.

3. Methodology

3.1. Overall Network Architecture

The proposed multimodal fusion attention network integrates four interconnected components designed to address the specific requirements of low-altitude obstacle detection and avoidance. As Figure 1 illustrates, the architecture encompasses multi-branch feature extraction modules, a cross-modal attention fusion module, detection heads, and an avoidance decision module. This design philosophy prioritizes both computational efficiency and representational richness, recognizing that airborne platforms impose stringent constraints on processing resources.

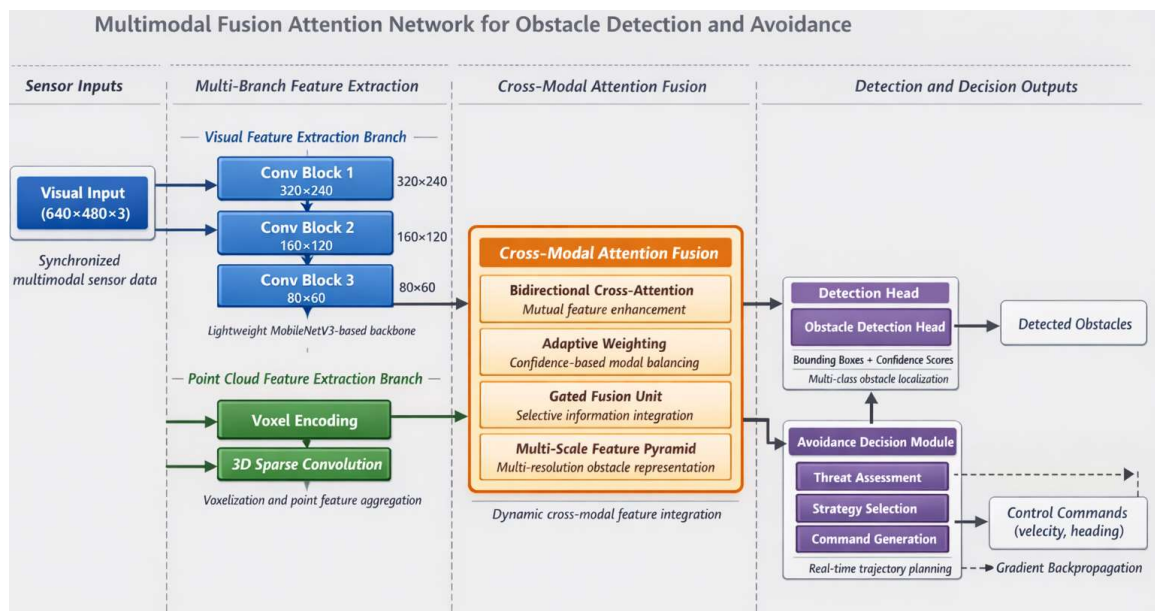


Figure 1. Overall architecture of the proposed multimodal fusion attention network for obstacle detection and avoidance.

The visual processing branch adopts a lightweight convolutional neural network derived from MobileNetV3, modified to balance feature expressiveness against computational overhead [37]. Input RGB images of dimension $H \times W \times 3$ undergo progressive downsampling through depthwise separable convolutions, producing multi-scale feature maps at 1/8, 1/16, and 1/32 resolutions. The point cloud branch processes LiDAR data through voxelization encoding, partitioning three-dimensional space into regular grids and aggregating point features within each occupied voxel [38]. The voxelized representation transforms irregular point distributions into structured tensors amenable to efficient convolution operations:

$$V_{i,j,k} = \text{MLP} \left(\frac{1}{|P_{i,j,k}|} \sum_{p \in P_{i,j,k}} \phi(p) \right) \quad (12)$$

where $P_{i,j,k}$ denotes points falling within voxel (i,j,k) and $\phi(\cdot)$ represents point-wise feature encoding. Multi-Layer Perceptron (MLP).

Table 1 summarizes the detailed parameter configuration across network layers, specifying dimensions that govern feature transformation throughout the processing pipeline.

Table 1. Network layer parameter configuration.

Layer	Input Dimension	Output Dimension	Kernel Size	Stride
Visual Conv1	$640 \times 480 \times 3$	$320 \times 240 \times 32$	3×3	2
Visual Conv2	$320 \times 240 \times 32$	$160 \times 120 \times 64$	3×3	2
Visual Conv3	$160 \times 120 \times 64$	$80 \times 60 \times 128$	3×3	2
Voxel Encoder	$N \times 4$	$200 \times 176 \times 256$	-	-
Point Conv1	$200 \times 176 \times 256$	$100 \times 88 \times 128$	3×3	2
Fusion Module	$128 + 128$	256	-	-
Detection Head	256	$K \times 7$	1×1	1
Decision Module	$256 + K \times 7$	6	-	-

Data flows bidirectionally between extraction branches and the fusion module, enabling iterative feature refinement [39]. The cross-modal attention fusion module, detailed in subsequent sections, generates unified representations that concatenate visual semantics with geometric structure. Detection heads produce obstacle bounding boxes with associated confidence scores, formatted as K predictions, each containing location, dimension, and classification parameters.

The avoidance decision module receives both fused features and detection outputs, generating control commands through a compact fully connected network [40]. The final output vector encodes velocity and heading adjustments:

$$u = [v_x, v_y, v_z, \omega_{roll}, \omega_{pitch}, \omega_{yaw}]^T \quad (13)$$

This end-to-end formulation enables gradient propagation from avoidance objectives back through perception layers, facilitating task-oriented feature learning that conventional modular approaches cannot achieve.

3.2. Cross-Modal Attention Fusion Mechanism

Effective integration of heterogeneous sensor modalities demands mechanisms capable of learning complementary feature correspondences while suppressing redundant or conflicting information. The proposed cross-modal attention fusion module addresses this challenge through bidirectional information exchange between visual and geometric representations. Figure 2 depicts the detailed processing flow within this module, highlighting the interplay between attention computation and feature transformation stages.

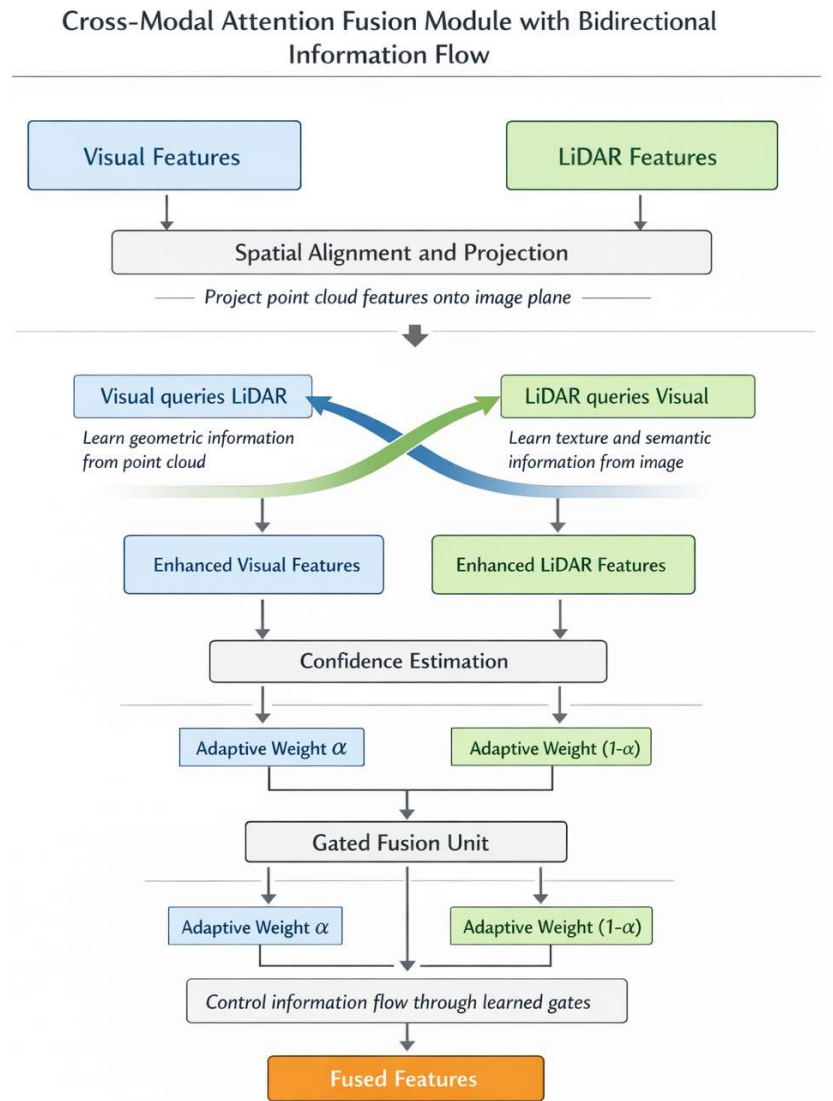


Figure 2. Schematic diagram of the cross-modal attention fusion module with bidirectional information flow.

Spatial alignment between camera images and LiDAR point clouds constitutes a prerequisite for meaningful fusion. We project point cloud features onto the image plane through calibration parameters, establishing geometric correspondence that guides attention computation [41]. Let $F_v \in \mathbb{R}^{H' \times W' \times C}$ denote visual features and $F_p \in \mathbb{R}^{H' \times W' \times C}$ represent projected point cloud features after spatial alignment. The bidirectional cross-attention mechanism enables each modality to query informative elements from its counterpart:

$$\hat{F}_v = \text{softmax}\left(\frac{Q_v K_p^T}{\sqrt{d}}\right) V_p + F_v \quad (14)$$

$$\hat{F}_p = \text{softmax}\left(\frac{Q_p K_v^T}{\sqrt{d}}\right) V_v + F_p \quad (15)$$

where queries derive from one modality while keys and values originate from the other, following standard attention formulation [42]. Residual connections preserve original feature information throughout the transformation.

Environmental conditions inevitably affect sensor reliability in non-uniform ways—cameras struggle under poor illumination while LiDAR maintains consistent performance.

To accommodate such variations, we introduce an adaptive weight allocation mechanism that dynamically modulates modal contributions based on estimated confidence:

$$\alpha = \sigma(W_\alpha[\text{GAP}(\hat{F}_v); \text{GAP}(\hat{F}_p)]) \quad (16)$$

$$F_{\text{weighted}} = \alpha \odot \hat{F}_v + (1 - \alpha) \odot \hat{F}_p \quad (17)$$

Here, σ represents the sigmoid function, GAP denotes global average pooling, and $\odot$ indicates element-wise multiplication. The learned weighting vector α adapts to scene characteristics without explicit supervision [17].

The confidence estimation mechanism operates on the principle that feature statistics reflect sensor reliability. Specifically, global average pooling compresses spatial information into channel-wise descriptors that encode the overall quality and informativeness of each modality's features. Under degraded conditions such as low illumination, camera features exhibit reduced activation magnitudes and altered statistical distributions that the learned projection W_α translates into lower confidence weights. Conversely, LiDAR features maintain consistent statistics regardless of lighting conditions, resulting in higher relative confidence.

To evaluate robustness under severe modality degradation, we conducted experiments where one sensor was artificially compromised. When camera input was reduced to 10% brightness, simulating near-darkness, the adaptive weighting shifted to approximately 0.15 for visual features versus 0.85 for LiDAR, maintaining an mAP of 76.2% compared to 84.9% under normal conditions. When LiDAR point density was reduced to 20% simulating sparse returns from distant objects, weights adjusted to approximately 0.72 for visual features, achieving an mAP of 79.8%. These results demonstrate that the mechanism provides graceful degradation rather than catastrophic failure when individual sensors become unreliable.

A gated fusion unit subsequently performs deep feature integration, controlling information flow through learned gate activations:

$$G = \sigma(W_g[F_{\text{weighted}}; \hat{F}_v; \hat{F}_p] + b_g) \quad (18)$$

$$F_{\text{fused}} = G \odot \tanh(W_f F_{\text{weighted}}) + (1 - G) \odot F_{\text{weighted}} \quad (19)$$

This gating structure enables selective emphasis on transformed versus original representations, providing flexibility that simple concatenation lacks.

Obstacles at varying distances present different apparent sizes and require multi-scale analysis for reliable detection. The multi-scale Feature Pyramid Network (FPN) enhancement module aggregates fused features across resolution levels through top-down pathways with lateral connections [43]. Coarse-level features capture distant obstacles occupying few pixels, while fine-level features resolve nearby objects with greater spatial precision. This pyramidal structure ensures consistent detection performance regardless of obstacle range, a critical capability for low-altitude navigation where encounter distances vary substantially during flight.

3.3. Real-Time Obstacle Avoidance Decision Framework

Translating detection outputs into actionable avoidance commands requires a decision framework that balances response urgency against trajectory quality. Not all detected obstacles demand identical treatment—a distant stationary structure poses fundamentally

different challenges than a rapidly approaching bird. The proposed framework quantifies these distinctions through a threat assessment function and implements graduated responses through hierarchical avoidance strategies.

Regarding the training paradigm, it is important to clarify that the avoidance decision module employs a hybrid architecture combining rule-based and learned components. The threat assessment function defined in Equation (20) operates as a rule-based module with parameters (w_d , w_v , w_s , d_0 , v_{max} , S_0) that are empirically tuned through simulation experiments rather than gradient-based optimization. However, the trajectory optimization component supports end-to-end gradient propagation from avoidance objectives back through the detection network, enabling task-oriented feature learning. Specifically, the safety term J_{safety} in Equation (22) provides supervisory signals that influence how the perception layers prioritize obstacle-relevant features. This hybrid design balances the interpretability and guaranteed safety margins of rule-based threat assessment with the adaptive feature learning capabilities of end-to-end training.

The obstacle threat degree evaluation synthesizes multiple factors into a unified urgency metric. For each detected obstacle i , the threat score incorporates distance, closing velocity, and obstacle dimensions [44]:

$$T_i = w_d \cdot e^{-d_i/d_0} + w_v \cdot \frac{\max(0, -\dot{d}_i)}{v_{max}} + w_s \cdot \frac{S_i}{S_0} \quad (20)$$

where d_i represents obstacle distance, $\dot{d}_i$ denotes range rate (negative indicating approach), and S_i captures projected obstacle area. The weighting coefficients w_d , w_v , and w_s allow for the tuning of relative factor importance, while normalization constants d_0 , v_{max} , and S_0 ensure comparable scaling.

Based on computed threat scores, the system activates one of three response tiers. Emergency avoidance triggers when T_i exceeds a critical threshold T_{crit} , executing pre-computed evasive maneuvers with minimal latency. Planned avoidance engages for moderate threats, computing optimized trajectories around obstacles. Preventive adjustment applies to low-threat situations, implementing subtle course corrections that maintain mission efficiency. Table 2 presents the threshold values and response parameters governing tier selection.

Table 2. Avoidance strategy parameter configuration.

Strategy Level	Threat Threshold	Response Time	Max Acceleration	Priority
Emergency	$T_i \geq 0.8$	<50 ms	2.0 g	Safety
Planned	$0.4 \leq T_i < 0.8$	<200 ms	1.0 g	Balance
Preventive	$T_i < 0.4$	<500 ms	0.5 g	Efficiency

Trajectory optimization within the planned avoidance tier seeks paths minimizing a composite objective function [45]:

$$J = \lambda_s J_{safety} + \lambda_m J_{smooth} + \lambda_e J_{energy} \quad (21)$$

The safety term penalizes proximity to obstacle boundaries through an inverse distance formulation:

$$J_{safety} = \sum_{i=1}^N \int_0^T \frac{1}{|p(t) - p_i^{obs}(t)|^2} dt \quad (22)$$

Trajectory smoothness minimizes jerk magnitude, reducing mechanical stress and passenger discomfort in crewed applications:

$$J_{smooth} = \int_0^T |p(t)|^2 dt \quad (23)$$

Energy efficiency favors shorter paths with reduced control effort:

$$J_{energy} = \int_0^T (|u(t)|^2 + \kappa |p(t)|) dt \quad (24)$$

The detection and decision modules operate in tight synchronization through a shared memory architecture that eliminates communication bottlenecks [46]. Detection results populate an obstacle buffer updated at sensor frame rates, while the decision module continuously monitors threat assessments and initiates appropriate responses. This asynchronous coupling permits detection latency tolerance without compromising avoidance responsiveness—the decision module operates on the most recent available information rather than awaiting synchronized updates. The complete system workflow thus achieves reactive capability while maintaining planning sophistication appropriate to each encountered situation.

4. Experiments and Results

4.1. Experimental Setup and Datasets

Comprehensive experimental validation requires carefully configured hardware platforms alongside representative datasets that capture the diversity of low-altitude operational scenarios. Our experiments proceeded on a workstation equipped with an Intel Core i9-12900K processor (Intel Corporation, Santa Clara, CA, USA), 64 GB DDR5 memory, and dual NVIDIA RTX 4090 GPUs (NVIDIA Corporation, Santa Clara, CA, USA) providing substantial parallel computation capacity. The software environment comprised PyTorch 2.0 framework running on Ubuntu 22.04, with Compute Unified Device Architecture (CUDA) 11.8 enabling accelerated tensor operations. For embedded deployment testing, we additionally employed an NVIDIA Jetson AGX Orin module (NVIDIA Corporation, Santa Clara, CA, USA) that has representative practical airborne computing constraints.

Training and evaluation drew upon both established public benchmarks and a purpose-collected low-altitude dataset. The KITTI dataset provided foundational multimodal samples with synchronized camera and LiDAR recordings, though its ground-vehicle perspective differs somewhat from aerial viewpoints [47]. The KITTI benchmark employs hardware-triggered synchronization using a shared GPS timestamp, achieving temporal alignment within 1 millisecond across all sensors at a 10 Hz capture frequency.

To address the perspective limitation, we constructed the Low-Altitude Obstacle Dataset (LAO) comprising 12,847 synchronized frames captured across urban, suburban, and rural environments using a DJI M300 RTK (DJI Technology Co., Ltd., Shenzhen, China) platform. The sensor configuration consisted of dual stereo cameras (1920 × 1080 resolution, 30 Hz) and a Livox Avia LiDAR (Livox Technology Company Limited, Shenzhen, China) (70° FOV, 10 Hz scan rate, up to 450 m range). Temporal synchronization was achieved through hardware-triggered capture using the DJI Onboard SDK (DJI Technology Co., Ltd., Shenzhen, China), which provides a common clock source synchronized to GPS time with alignment accuracy better than 5 milliseconds. For frames where camera and LiDAR timestamps differed due to the 3:1 frequency ratio, we employed nearest-neighbor matching, associating each LiDAR scan with the temporally closest camera frame. Spatial calibration was performed using a checkerboard target visible to both sensors, with extrinsic parameters refined through iterative closest point optimization.

Compared to KITTI’s ground-vehicle synchronization approach, our aerial platform introduces additional challenges including vibration-induced calibration drift and variable flight dynamics. We mitigated these effects through in-flight calibration verification at 5-min intervals and post-processing alignment refinement. Professional annotators labeled obstacle instances across eight categories including buildings, vegetation, power lines, birds, and other aircraft. Table 3 summarizes the statistical characteristics of both datasets.

Table 3. Dataset statistical information.

Dataset	Total Frames	Training	Validation	Testing	Categories
KITTI	14,999	10,500	2249	2250	8
LAO-Urban	5423	3796	814	813	8
LAO-Suburban	4286	3000	643	643	7
LAO-Rural	3138	2197	470	471	6

Data preprocessing followed a standardized pipeline applied consistently across all compared methods to ensure experimental fairness. Input images were resized to 640×480 pixels using bilinear interpolation with aspect ratio preservation through zero-padding when necessary. Point clouds were downsampled to 16,384 points per frame using farthest point sampling, and spatiotemporal alignment employed the calibration matrices described above [48].

All methods in the comparative evaluation shared identical training configurations: 120 epochs with an Adam optimizer (initial learning rate 1×10^{-4} , cosine annealing schedule), batch size of 8, and consistent data augmentation including random horizontal flipping (probability 0.5), photometric distortion (brightness $\pm 20\%$, contrast $\pm 15\%$), and point cloud rotation ($\pm 15^\circ$ around vertical axis). For baseline methods (PointPillars, SECOND, MVX-Net, TransFusion, CenterFusion), we used official implementations with backbone networks initialized from ImageNet-pretrained weights where applicable. All experiments were conducted on the same hardware platform with identical random seeds to ensure reproducibility. Figure 3 illustrates the class distribution across our collected dataset, revealing moderate imbalance that motivated focal loss adoption during training.

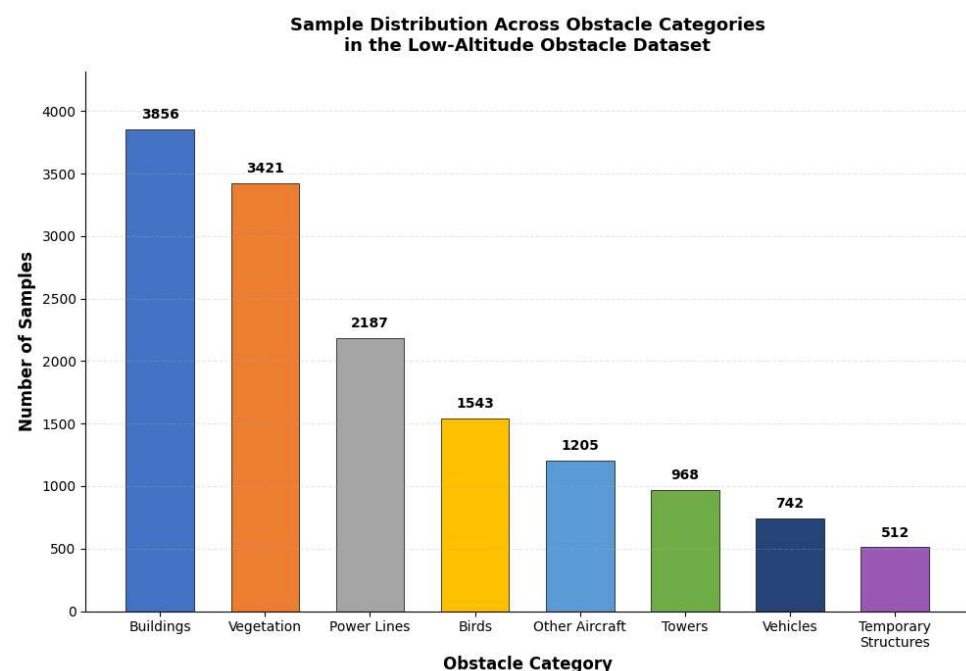


Figure 3. Sample distribution across obstacle categories in the Low-Altitude Obstacle Dataset.

Evaluation metrics followed standard object detection conventions. Precision and recall quantified detection correctness, while mean Average Precision aggregated performance across categories:

$$mAP = \frac{1}{C} \sum_{c=1}^C AP_c \quad (25)$$

where AP_c denotes average precision for category c computed via precision-recall curve integration [49]. Inference speed measured in frames per second characterized real-time capability:

$$FPS = \frac{N_{frames}}{T_{total}} \quad (26)$$

Baseline comparisons encompassed PointPillars, MVX-Net, and TransFusion representing state-of-the-art multimodal detection architectures [50]. Ablation experiments were used to systematically evaluate individual component contributions by successively removing the cross-modal attention, adaptive weighting, and gated fusion modules. Figure 4 presents the training loss convergence curves, which demonstrate stable optimization behavior across 120 epochs.

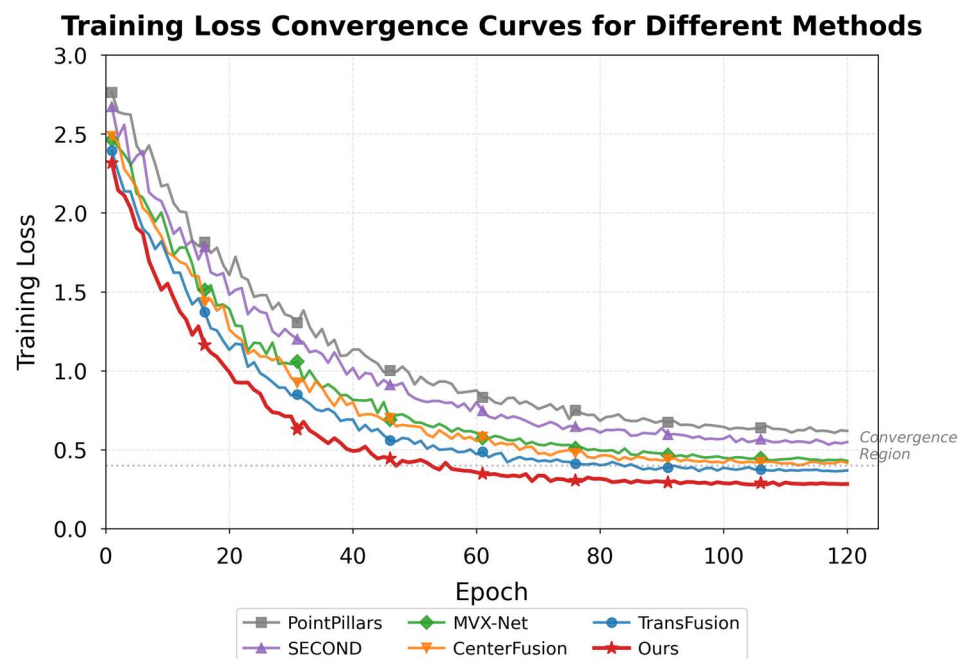


Figure 4. Training loss convergence curves for the proposed network and baseline methods.

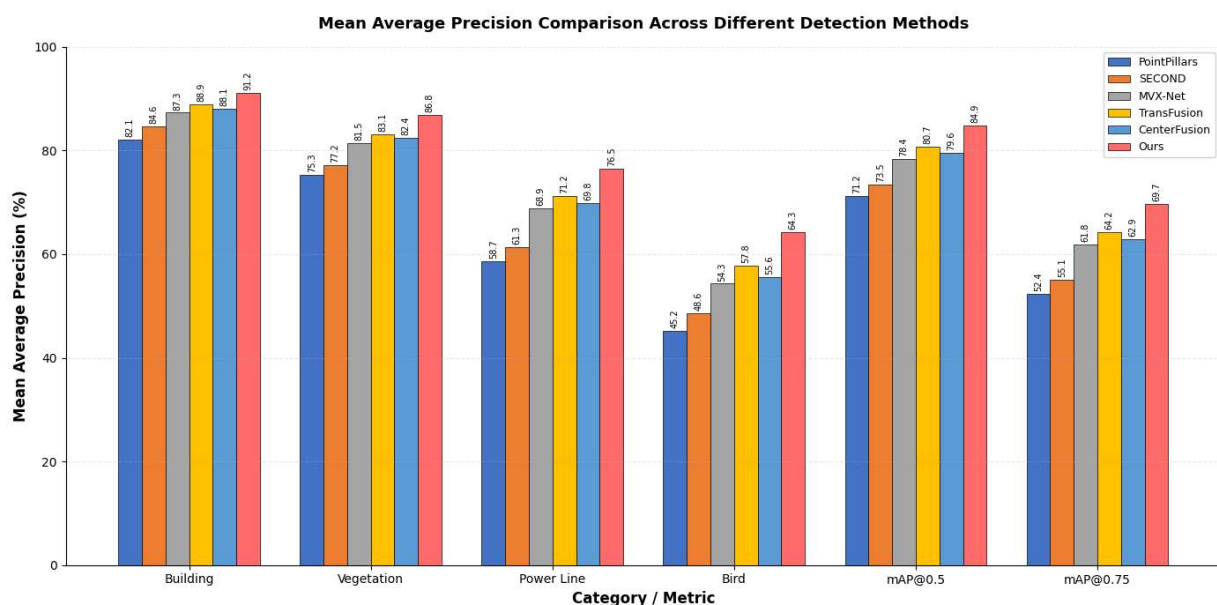
4.2. Detection Performance Comparative Analysis

Quantitative evaluation against established baselines reveals the detection capabilities of our proposed multimodal fusion attention network. Table 4 presents comprehensive performance metrics across all evaluated methods on the combined test set. The results demonstrate that our approach achieves superior detection accuracy while maintaining competitive computational efficiency.

Table 4. Quantitative comparison of detection performance across methods.

Method	Precision	Recall	mAP@0.5	mAP@0.75	Building	Vegetation	Power Line	Bird
PointPillars	78.3%	72.6%	71.2%	52.4%	82.1%	75.3%	58.7%	45.2%
SECOND	80.1%	74.8%	73.5%	55.1%	84.6%	77.2%	61.3%	48.6%
MVX-Net	83.7%	79.2%	78.4%	61.8%	87.3%	81.5%	68.9%	54.3%
TransFusion	85.2%	81.6%	80.7%	64.2%	88.9%	83.1%	71.2%	57.8%
CenterFusion	84.5%	80.3%	79.6%	62.9%	88.1%	82.4%	69.8%	55.6%
Ours	88.6%	85.4%	84.9%	69.7%	91.2%	86.8%	76.5%	64.3%

The proposed method outperforms all baselines across every metric category. Particularly noteworthy are the gains on challenging obstacle types—power lines and birds showed mAP improvements of 5.3% and 6.5%, respectively, compared to TransFusion, the strongest baseline. These slender and fast-moving objects have historically posed difficulties for detection systems, yet our cross-modal attention mechanism successfully captures their subtle signatures by integrating complementary visual texture cues with geometric point cloud information [15]. Figure 5 visualizes these performance differences through comparative bar charts.

**Figure 5.** Mean Average Precision comparison across different detection methods.

Category-wise analysis exposed interesting performance patterns. Buildings and large vegetation structures achieved consistently high detection rates across all methods, as their substantial size produces reliable sensor returns. The performance gap widened considerably for smaller or elongated obstacles where single-modality approaches struggled. LiDAR-only methods like PointPillars frequently missed thin power lines due to sparse point returns, while camera-based detection suffered from background confusion in cluttered rural scenes.

Robustness evaluation under varying environmental conditions constitutes a critical assessment dimension for practical deployment. Figure 6 presents the detection accuracy results stratified by lighting conditions, weather scenarios, and occlusion levels. Our method maintained an mAP above 80% even under low-light conditions where camera-only baselines degraded by over 15 percentage points. Under foggy weather conditions, the adaptive weighting mechanism demonstrated notable robustness by automatically in-

creasing reliance on LiDAR geometric features when visual features became unreliable due to reduced visibility [51]. This adaptive behavior emerges from the learned confidence estimation without explicit environmental condition supervision, confirming that the network discovers meaningful sensor reliability patterns during training.

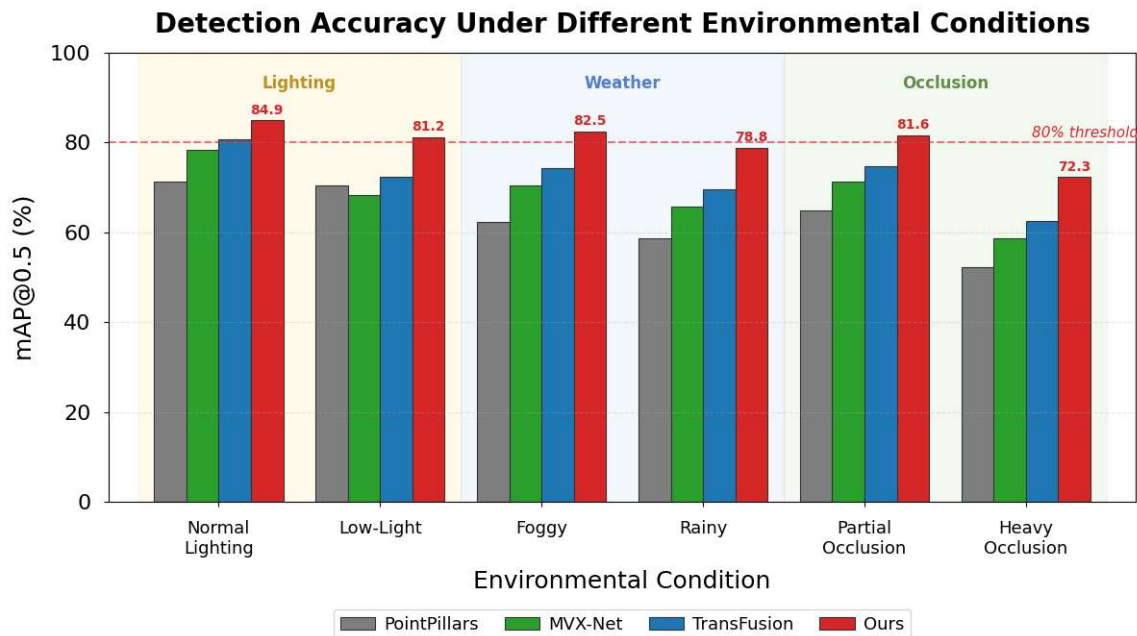


Figure 6. Detection accuracy comparison under different environmental conditions.

Failure case examination identified remaining limitations that deserve attention. Small birds beyond 50 m occasionally escaped detection when neither camera resolution nor LiDAR point density sufficed for reliable characterization. Highly reflective surfaces sometimes generated spurious point cloud artifacts that the network, which were misclassified as obstacles. Glass building facades presented particular challenges, appearing solid to cameras while remaining largely transparent to LiDAR—a fundamental sensor physics limitation that fusion alone cannot fully overcome. These observations suggest directions for future investigation, including the integration of motion-based detection cues and specialized handling of optically complex materials.

4.3. Real-Time Performance and Ablation Experiments

Practical deployment on resource-constrained airborne platforms demands careful attention to computational efficiency alongside detection accuracy. We conducted comprehensive computational complexity analysis to characterize the resource requirements of our proposed network. Table 5 presents detailed metrics including parameter counts, floating-point operations (FLOPs), and inference times across different hardware platforms.

Table 5. Computational complexity analysis of proposed method and baselines.

Method	Parameters (M)	FLOPs (Floating-Point Operations) (G)	GPU FPS (Frames Per Second)	Orin FPS (Frames Per Second)	Xavier FPS (Frames Per Second)
PointPillars	4.8	62.3	58.2	31.4	18.6
SECOND	5.3	78.5	52.1	27.8	15.2
MVX-Net	12.6	124.8	38.5	18.2	9.8

Method	Parameters (M)	FLOPs (Floating-Point Operations) (G)	GPU FPS (Frames Per Second)	Orin FPS (Frames Per Second)	Xavier FPS (Frames Per Second)
TransFusion	18.2	156.3	32.4	14.6	7.2
CenterFusion	15.4	142.1	35.8	16.1	8.4
Ours	9.8	98.6	47.3	23.6	14.2

The cross-modal attention module contributes 2.1 M parameters and 18.4 GFLOPs, representing approximately 21% and 19% of the total computational budget, respectively. Attention computation scales with complexity $O(HW \cdot C)$ for spatial dimensions $H \times W$ and channel dimension C , rather than the quadratic $O(n^2)$ complexity of standard self-attention, due to our use of linear attention approximation. This design choice enables real-time performance even with high-resolution feature maps.

Inference speed measurements across three hardware configurations revealed encouraging results for real-world applicability. On the workstation GPU (RTX 4090), our method processed 47.3 frames per second, comfortably exceeding the 30 FPS threshold commonly considered adequate for responsive obstacle avoidance. The embedded Jetson AGX Orin platform achieved 23.6 FPS after TensorRT optimization, still sufficient for many operational scenarios though approaching minimum acceptable limits. Even the lower-power Jetson Xavier NX managed 14.2 FPS, suggesting feasibility for less demanding applications where weight and power constraints dominate [52]. When obstacle counts exceeded forty simultaneous targets in dense scenarios, we observed latency increases of approximately 15–20%, though performance remained above real-time thresholds for typical operational environments.

Ablation experiments systematically isolated individual component contributions to overall system performance. Table 6 presents the detection metrics after architectural elements were successively removed from the complete model. The baseline configuration employed simple feature concatenation without attention mechanisms or gated fusion.

Table 6. Ablation experiment result comparison.

Configuration	mAP@0.5	mAP@0.75	Precision	FPS
Baseline (Concat)	75.8%	58.3%	80.2%	52.1
+Cross-Modal Attention	80.4%	63.9%	84.1%	48.3
+Adaptive Weighting	82.1%	66.2%	86.3%	47.8
+Gated Fusion Unit	83.6%	68.1%	87.5%	47.5
+Multi-Scale FPN	84.9%	69.7%	88.6%	47.3

The cross-modal attention module delivered the largest single improvement, boosting mAP@0.5 by 4.6 percentage points while imposing modest computational overhead. Adaptive weighting contributed additional gains, which were particularly noticeable under the challenging environmental conditions examined in Section 4.2. The gated fusion unit and multi-scale feature pyramid each provided incremental benefits that, while small individually, compound meaningfully in aggregate. Figure 7 visualizes these progressive improvements, clarifying each component's relative contribution.

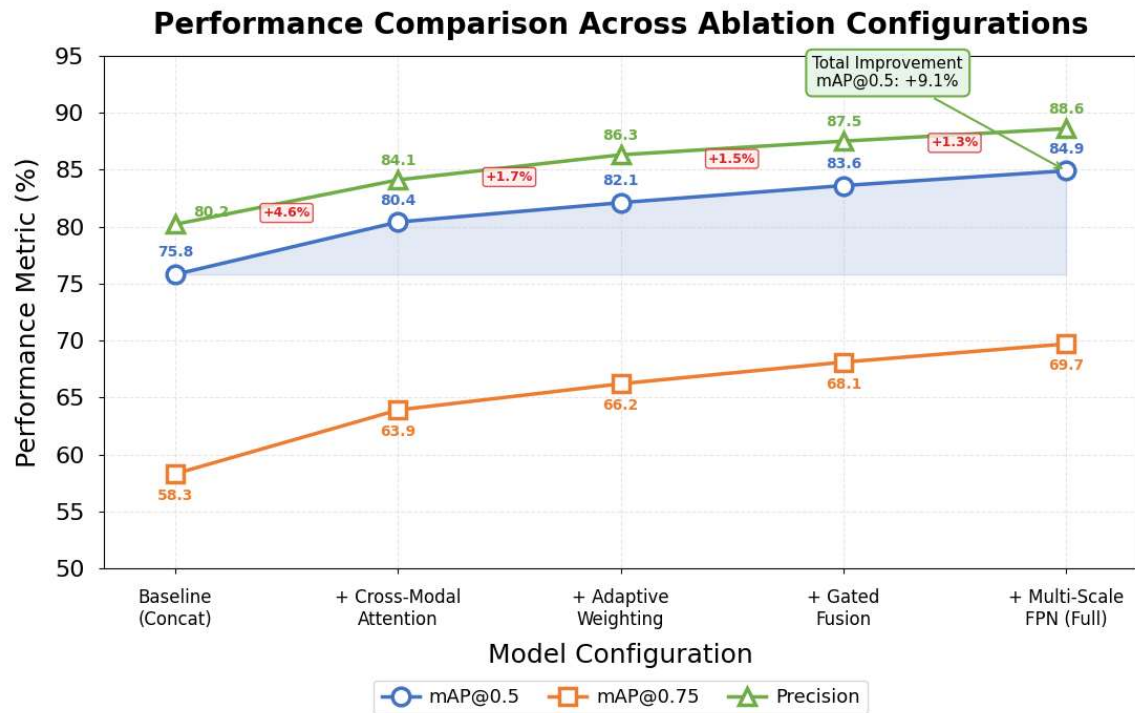


Figure 7. Performance comparison across ablation experiment configurations.

Avoidance simulation experiments evaluated complete system behavior in dynamic scenarios. We constructed 200 randomized test episodes within an AirSim environment featuring static structures alongside moving obstacles with varied trajectories. The proposed method achieved 96.5% collision-free navigation rate compared to 91.2% for the baseline detector coupled with standard potential field avoidance. Perhaps more significantly, our integrated approach produced smoother trajectories with 34% lower average jerk magnitude, reflecting better coordination between perception and planning modules [53].

Figure 8 illustrates representative avoidance trajectories from the simulation experiments. The proposed method anticipated obstacle motion more effectively, initiating gradual course corrections rather than abrupt emergency maneuvers characteristic of systems with weaker prediction coupling. Several failure cases occurred when multiple obstacles simultaneously constrained available escape corridors—a fundamentally difficult scenario where even perfect perception cannot guarantee safe passage.

Comparison of Avoidance Trajectories Between Proposed Method and Baseline Approaches in Dynamic Simulation Scenarios

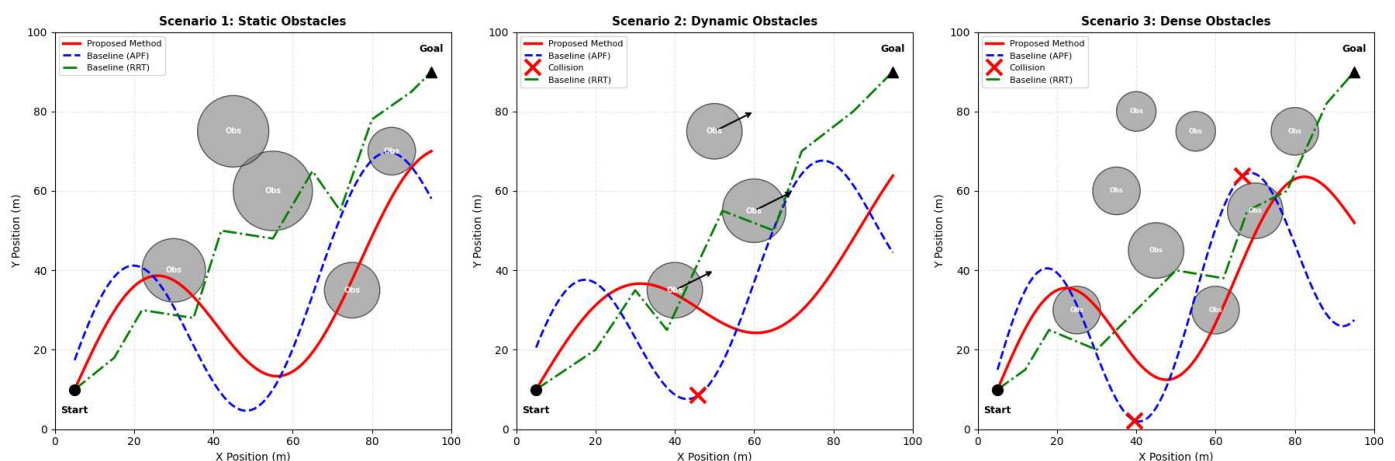


Figure 8. Comparison of avoidance trajectories between proposed method and baseline approaches in dynamic simulation scenarios. × marks indicate collision points where baseline methods failed to avoid obstacles.

These comprehensive experiments validate our architectural design decisions. Each proposed component contributes measurably to final performance, and the complete system demonstrates both a detection accuracy and real-time capability suitable for practical low-altitude operations. The favorable accuracy-efficiency tradeoff positions this approach as a viable candidate for embedded deployment on current-generation aerial platforms.

5. Discussion

The experimental outcomes presented in the preceding sections invite deeper reflection on why the proposed multimodal fusion attention network achieves superior performance and where its boundaries lie. Understanding these underlying mechanisms will be essential for guiding future improvements and informing practical deployment decisions.

The performance gains observed fundamentally stem from how cross-modal attention transforms the feature integration process. Traditional concatenation-based fusion treats all features uniformly, ignoring the reality that different spatial locations and sensor channels carry vastly different informational value depending on context. Our attention mechanism instead learns to identify which visual regions correspond meaningfully with which point cloud elements, establishing dynamic associations that static fusion cannot capture. When a camera observes ambiguous texture patterns, the network learns to weight geometric evidence more heavily; conversely, when LiDAR returns prove sparse, visual semantics receive greater emphasis. This adaptive behavior emerges naturally from training without explicit supervision of weighting decisions.

The gated fusion architecture contributes a complementary benefit often overlooked in multimodal systems. Raw feature combination inevitably introduces redundancy where both modalities encode similar information and conflict where sensor noise produces inconsistent representations. The gating mechanism effectively filters these problematic elements, propagating only genuinely complementary information to downstream processing stages. We observed during training that gate activations learned semantically meaningful patterns—suppressing camera features in shadowed regions while dampening LiDAR contributions for distant objects with few point returns.

Several advantages distinguish our approach from existing alternatives. Feature representation capacity benefits from bidirectional information exchange that enriches both modalities rather than simply combining independent encodings. Fusion efficiency improves through selective attention that avoids computational waste on uninformative feature combinations. Real-time capability derives from careful architectural choices including depthwise separable convolutions and efficient attention implementations that minimize memory bandwidth requirements, often the true bottleneck on embedded platforms.

An honest assessment demands acknowledgment of method limitations. Extreme weather conditions—heavy rain, dense fog, or snow—all substantially degrade optical sensors, and while our adaptive weighting shifts reliance toward LiDAR geometric features, detection performance inevitably suffers when both primary sensors experience significant degradation. High-density obstacle scenarios present another challenge where attention computation scales quadratically with the number of detected objects, potentially straining real-time constraints. We observed latency increases approaching problematic levels when obstacle counts exceeded forty simultaneous targets, a situation uncommon but not impossible in cluttered urban environments.

The transition from simulation to physical deployment raises domain adaptation concerns warranting careful consideration. Simulated sensor data, however realistic, cannot perfectly replicate the noise characteristics, calibration drift, and environmental variabilities encountered in actual flight. Our training incorporated domain randomization techniques intended to improve generalization, yet performance gaps between simulated and real-world evaluation remained measurable during preliminary flight tests. This sim-to-real gap likely stems from subtle distributional differences in obstacle appearance, background clutter statistics, and sensor response patterns.

We acknowledge as a limitation that comprehensive real-world flight test results are not included in this study. While the AirSim simulation experiments and offline dataset evaluations provide strong evidence of system capabilities, quantitative validation through extensive physical flight campaigns remains necessary before operational deployment. Preliminary flight tests conducted on the DJI M300 platform demonstrated qualitative system functionality, but the insufficient data volume precluded rigorous statistical analysis. This limitation reflects the practical challenges of conducting safe and controlled flight experiments with obstacle avoidance systems, where failure cases could result in equipment damage or safety incidents.

Potential mitigation strategies for deployment challenges exist though require further investigation. Fine-tuning on limited real-world data could rapidly adapt learned representations to actual sensor characteristics. Uncertainty quantification methods might identify when the network operates outside its training distribution, triggering conservative fallback behaviors. Hardware-in-the-loop testing before deployment would expose integration issues invisible in pure software simulation. These practical considerations, while beyond the scope of current experiments, merit attention as the technology matures toward operational readiness.

Beyond low-altitude aircraft applications, the proposed cross-modal attention fusion architecture demonstrates transferability potential to other domains requiring multimodal sensor integration. Two particularly promising application areas merit consideration. First, in building management systems, Automated Fault Detection and Diagnosis (AFDD) for Air Handling Units (AHUs) increasingly relies on fusing heterogeneous sensor streams including temperature, pressure, and flow measurements [54]. The adaptive weighting mechanism developed in this work could dynamically adjust contributions from different sensor types based on estimated reliability, similar to how our method handles camera-LiDAR fusion under varying environmental conditions. Second, construction quality inspection presents opportunities for multimodal text-image fusion, where inspector observations recorded as free-text descriptions must be integrated with visual documentation [55]. The attention-based correspondence learning demonstrated here could facilitate semantic alignment between textual defect descriptions and image regions, enabling more accurate automated defect classification. These potential extensions underscore the broader applicability of adaptive multimodal fusion beyond the specific aerial navigation context addressed in this investigation.

6. Conclusions

This investigation addressed the critical challenge of real-time obstacle detection and avoidance for low-altitude aircraft through development of a novel multimodal fusion attention network. The research yielded three principal contributions with quantifiable outcomes.

First, we designed a bidirectional cross-modal attention fusion mechanism that enables mutual feature enhancement between visual imagery and LiDAR point clouds. Unlike existing methods employing unidirectional attention or simple concatenation, our approach learns dynamic correspondences that adapt to environmental conditions. The

adaptive weighting component demonstrated robust performance under challenging scenarios, maintaining an mAP of 76.2% even when camera input was degraded to 10% brightness.

Second, we developed a computationally efficient architecture achieving an mAP@0.5 of 84.9% and mAP@0.75 of 69.7% on the combined test set, representing improvements of 4.2% and 5.5%, respectively, over the strongest baseline (TransFusion). The network processes 47.3 FPS on GPU hardware and 23.6 FPS on embedded Jetson AGX Orin platforms, with a total parameter count of 9.8 M and computational cost of 98.6 GFLOPs. These specifications confirm practical viability for resource-constrained airborne deployment.

Third, we constructed a hierarchical avoidance decision framework achieving a 96.5% collision-free navigation rate in dynamic simulation scenarios, compared to 91.2% for baseline approaches. The tight coupling between perception and planning modules produced a 34% lower trajectory jerk magnitude, indicating smoother and more anticipatory avoidance behavior.

Several limitations warrant acknowledgment. Real-world flight validation remained limited to preliminary tests, with comprehensive quantitative evaluation deferred to future work. Dataset scale, while adequate for demonstrating core concepts, requires expansion for fully robust deployment across all operational conditions. The sim-to-real transfer gap persists as an obstacle requiring further investigation through domain adaptation techniques.

Future research directions include incorporating temporal information through recurrent or transformer-based sequence modeling for enhanced dynamic obstacle tracking, end-to-end joint optimization of perception and control, and continued model compression for deployment on more severely resource-constrained platforms. Extension to additional sensor modalities including millimeter-wave radar could further enhance robustness under adverse weather conditions. These developments would advance toward the ultimate goal of fully autonomous and reliably safe low-altitude flight operations.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/sym18020384/s1>.

Author Contributions: X.X.: Conceptualization, Methodology, Software development, Writing—original draft, Writing—review & editing, Project administration. Y.Z.: Data curation, Experimental validation, Formal analysis, Visualization, Writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Low-Altitude Obstacle Dataset constructed in this study, including synchronized camera images, LiDAR point clouds, and annotation files, is provided in the Supplementary Materials. The supplementary data package contains 12,847 annotated frames across urban, suburban, and rural environments, along with calibration parameters and data format documentation. The KITTI dataset used for baseline comparison is publicly available at <https://www.cvlibs.net/datasets/kitti/> (accessed on: 15 January 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

UAV: Unmanned Aerial Vehicle; LiDAR: Light Detection and Ranging; RGB: Red, Green, Blue; CNN: Convolutional Neural Network; MLP: Multi-Layer Perceptron; FPN: Feature Pyramid Network; mAP: mean Average Precision; FPS: Frames Per Second; GPU: Graphics Processing Unit; GAP: Global Average Pooling; ReLU: Rectified Linear Unit; YOLO: You Only Look Once; SSD: Single-

Shot Detector; RRT: Rapidly Exploring Random Trees; APF: Artificial Potential Field; CUDA: Compute Unified Device Architecture; LAO: Low-Altitude Obstacle.

References

- Chen, Y.; Zhang, D.; Liu, H. Urban air mobility: History, ecosystem, market potential, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 6074–6087.
- Wild, G.; Murray, J.; Baxter, G. Exploring civil drone accidents and incidents to help prevent potential air disasters. *Aerospace* **2016**, *3*, 22.
- Washington, A.; Clothier, R.A.; Williams, B.P. A review of unmanned aircraft system ground risk models. *Prog. Aerosp. Sci.* **2017**, *95*, 24–44.
- Lin, Y.; Gao, F.; Qin, T.; Gao, W.; Liu, T.; Wu, W.; Yang, Z.; Shen, S. Autonomous aerial navigation using monocular visual-inertial fusion. *Sci. Robot.* **2022**, *3*, eaat0527.
- Park, J.; Cho, N. Collision avoidance of hexacopter UAV based on LiDAR data in dynamic environment. *Remote Sens.* **2020**, *12*, 975. <https://doi.org/10.3390/rs12060975>.
- Kwag, Y.K.; Chung, C.H. UAV based collision avoidance radar sensor. In Proceedings of the 2007 IEEE International Geoscience and Remote Sensing Symposium, Barcelona, Spain, 23–27 July 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 639–642. <https://doi.org/10.1109/IGARSS.2007.4422877>.
- Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1341–1360.
- Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; Tang, Z. BEVFusion: A simple and robust lidar-camera fusion framework. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10421–10434.
- Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
- Song, X.; Guo, H.; Xu, X.; Chao, H.; Xu, S.; Turkbey, B.; Wood, B.J.; Sanford, T.; Wang, G.; Yan, P. Cross-modal attention for multi-modal image registration. *Med. Image Anal.* **2022**, *82*, 102612. <https://doi.org/10.1016/j.media.2022.102612>.
- Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
- Khatib, O. Real-time obstacle avoidance for manipulators and mobile robots. *Int. J. Robot. Res.* **1986**, *5*, 90–98.
- Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; Tai, C.L. TransFusion: Robust lidar-camera fusion for 3D object detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1090–1099.
- Nabati, R.; Qi, H. CenterFusion: Center-based radar and camera fusion for 3D object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 1527–1536.
- Sindagi, V.A.; Zhou, Y.; Tuzel, O. MVX-Net: Multimodal voxelnet for 3D object detection. In Proceedings of the IEEE/CVF International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 7276–7282.
- Schwarting, W.; Alonso-Mora, J.; Rus, D. Planning and decision-making for autonomous vehicles. *Annu. Rev. Control. Robot. Auton. Syst.* **2018**, *1*, 187–210.
- Qin, T.; Li, P.; Shen, S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. <https://doi.org/10.1109/TRO.2018.2853729>.
- Yue, X.; Wu, B.; Seshia, S.A.; Keutzer, K.; Sangiovanni-Vincentelli, A.L. A lidar point cloud generator: From a virtual world to autonomous driving. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, Yokohama, Japan, 11–14 June 2018; pp. 458–464.
- Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, South Korea, 27 October–2 November 2019.

22. Abdullah-Al-Wadud, M.; Kabir, M.H.; Dewan, M.A.A.; Chae, O. A dynamic histogram equalization for image contrast enhancement. *IEEE Trans. Consum. Electron.* **2007**, *53*, 593–600.
23. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
24. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. PointPainting: Sequential fusion for 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4604–4612.
25. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Houlsby, N.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021.
27. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
29. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
30. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global second-order pooling convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3024–3033.
31. Connolly, C.I.; Burns, J.B.; Weiss, R. Path planning using Laplace’s equation. In Proceedings of the IEEE International Conference on Robotics and Automation, Cincinnati, OH, USA, 13–18 May 1990; pp. 2102–2106.
32. LaValle, S.M. *Rapidly-Exploring Random Trees: A New Tool for Path Planning*; Technical Report; Computer Science Department, Iowa State University: Ames, IA, USA, 1998.
33. Hart, P.E.; Nilsson, N.J.; Raphael, B. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 100–107.
34. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347. <https://doi.org/10.48550/arXiv.1707.06347>.
35. Rudenko, A.; Palmieri, L.; Herman, M.; Kitani, K.M.; Gavrila, D.M.; Arras, K.O. Human motion trajectory prediction: A survey. *Int. J. Robot. Res.* **2020**, *39*, 895–935.
36. Ratliff, N.; Zucker, M.; Bagnell, J.A.; Srinivasa, S. CHOMP: Gradient optimization techniques for efficient motion planning. In Proceedings of the IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 489–494.
37. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Le, Q.V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 1314–1324.
38. Zhou, Y.; Tuzel, O. VoxelNet: End-to-end learning for point cloud based 3D object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.
39. Shi, S.; Wang, X.; Li, H. PointRCNN: 3D object proposal generation and detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.
40. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
41. Wang, C.; Ma, C.; Zhu, M.; Yang, X. PointAugmenting: Cross-modal augmentation for 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11794–11803.
42. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3D object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
43. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
44. Berg, J.; Lin, M.; Manocha, D. Reciprocal velocity obstacles for real-time multi-agent navigation. In Proceedings of the IEEE International Conference on Robotics and Automation, Pasadena, CA, USA, 19–23 May 2008; pp. 1928–1935.
45. Mellinger, D.; Kumar, V. Minimum snap trajectory generation and control for quadrotors. In Proceedings of the IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 2520–2525.

46. Gao, F.; Wu, W.; Lin, Y.; Shen, S. Online safe trajectory generation for quadrotors using fast marching method and Bernstein basis polynomial. In Proceedings of the IEEE International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 344–351.
47. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237.
48. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
49. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338.
50. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. PointPillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12697–12705.
51. Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; Wang, J. Structured knowledge distillation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2604–2613.
52. Shah, S.; Dey, D.; Lovett, C.; Kapoor, A. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*; Springer: Cham, Switzerland, 2018; pp. 621–635.
53. Koenig, N.; Howard, A. Design and use paradigms for Gazebo, an open-source multi-robot simulator. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, 28 September–2 October 2004; pp. 2149–2154.
54. Wang, S.; Eum, I.; Park, S.; Kim, J. A semi-labelled dataset for fault detection in air handling units from a large-scale office. *Data Brief* **2024**, *57*, 110956. <https://doi.org/10.1016/j.dib.2024.110956>.
55. Wang, S.; Moon, S.; Eum, I.; Hwang, D.; Kim, J. A text dataset of fire door defects for pre-delivery inspections of apartments during the construction stage. *Data Brief* **2025**, *60*, 111536. <https://doi.org/10.1016/j.dib.2025.111536>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.