Vision-Radar Fusion for UAV Perception in Low-Altitude Emergency Rescue: A Survey

Xuexue Zhang[1], Chenchen He[2], Dongran Sun[3]

(1.Fortune Pass (Beijing) Information Technology Co., Ltd, Beijing100016 ; 2. Beijing Jiazhixing Commercial Co., Ltd, Beijing 100124; 3.CITIC Consumer Finance Co., Ltd,Beijing 100000)

ABSTRACT Unmanned Aerial Vehicles (UAVs) have emerged as critical tools in emergency rescue operations, particularly in urban low-altitude complex environments characterized by dynamic obstacles, adverse weather, and signal interference. However, traditional single-modality perception systems, reliant on either vision or radar, struggle with limitations such as motion blur, low light conditions, and occlusion, compromising detection accuracy and navigation safety. This survey provides a comprehensive review of vision-radar fusion perception systems tailored for UAVs in such scenarios. We first examine advancements in image quality enhancement techniques, including diffusion models and super-resolution methods like SRGAN [16] and ESRGAN [17], which address degradation in low-altitude imagery. Next, we explore multi-target detection and recognition via multi-modal fusion frameworks, such as Faster R-CNN-based approaches [5,7] and cross-view spatial fusion [10]. We then discuss robust multi-target tracking and trajectory prediction, highlighting graph neural networks (GNNs) and Transformer-based models like TrackFormer [36] and MOTR [53]. Finally, we cover low-altitude environment assessment and autonomous path planning, leveraging graph theory, 3D risk mapping, and deep reinforcement learning (e.g., MADDPG). By synthesizing global research trends, challenges like spatio-temporal alignment and real-time processing, and future directions toward end-to-end multi-modal integration, this review underscores the potential of fusion systems to enhance UAV resilience and efficiency in rescue missions. Our analysis reveals a clear trajectory toward AI-driven, adaptive perception, with implications for urban safety and disaster response.

Keywords UAV perception, vision-radar fusion, low-altitude environments, multi-target tracking, autonomous navigation, emergency rescue.

I.INTRODUCTION

1.1 Definition of Digital Assets

1. Introduction

The proliferation of Unmanned Aerial Vehicles (UAVs) has revolutionized emergency rescue applications, enabling rapid surveillance, search-and-rescue operations, and disaster assessment in urban low-altitude environments [62]. These scenarios, however, present formidable challenges: dense urban structures, variable weather (e.g., fog, rain), erratic lighting, and electromagnetic interference degrade sensor performance, leading to unreliable perception and heightened collision risks [2.1(4)]. Single-modality systems—vision-based for rich semantic details or radar-based for all-weather robustness—fall short in isolation, as visual sensors falter in low visibility [11] while radars lack fine-grained texture information [1-4]. Vision-radar fusion emerges as a pivotal solution, integrating complementary modalities to achieve robust, real-time perception for safe UAV navigation and target engagement [22].

Historically, UAV perception has evolved from rudimentary interpolation-based super-resolution [12] to deep learning paradigms. Early convolutional neural networks (CNNs) like SRCNN [12] and FSRCNN [13] accelerated image reconstruction but struggled with perceptual fidelity. Generative adversarial networks (GANs), exemplified by SRGAN [16] and ESRGAN [17], introduced perceptual losses to restore textures, yet instability issues persisted [22]. The advent of diffusion models (DMs) [20,21] marks a paradigm shift, enabling high-fidelity generation by iteratively denoising, with applications in UAV imagery enhancement under motion blur and atmospheric distortion [1]. In detection and recognition, multi-modal fusion has progressed from early concatenation in RVNet [5] and CRFNet [6] to advanced spatial alignments like 3D-CVF [10], fusing LiDAR/radar points with camera features to mitigate projection inconsistencies.

Tracking and prediction further demand temporal coherence. Traditional Kalman-filter-based methods like

SORT [25] and DeepSORT [47] excel in simple scenes but falter amid occlusions; Transformer architectures, such as TrackFormer [36], MOTR [53], and GTR [51], leverage global attention for end-to-end association, while graph-based trackers like UGT incorporate topological relations for occlusion recovery [3]. For environment assessment and planning, graph-theoretic grid modeling [62] and 3D risk maps integrate multi-sensor data, augmented by reinforcement learning (e.g., Q-learning [66] and MADDPG) for dynamic path optimization in multi-UAV swarms [67].

Despite these advances, gaps remain: spatio-temporal misalignment in fusion [7], computational overhead for edge deployment, and limited benchmarks for low-altitude rescue [35,37]. This survey synthesizes over 60 seminal works to delineate trends, benchmark performances (e.g., mAP >85% in fused detection [10]), and innovation frontiers. Section 2 reviews image enhancement; Section 3 covers detection/recognition; Section 4 addresses tracking/prediction; Section 5 explores assessment/planning; and Section 6 concludes with challenges and prospects.

By bridging theoretical foundations with practical imperatives, this review aims to guide researchers toward resilient fusion systems, ultimately amplifying UAV efficacy in safeguarding lives during crises.

2. Image Enhancement in Low-Altitude UAV Imagery

Image enhancement is a foundational step in UAV perception pipelines, particularly in low-altitude urban environments where captured visuals are prone to degradation from factors such as uneven illumination, atmospheric interference (e.g., fog, rain), motion blur due to high-speed flight, and low resolution from compact sensors [1]. These distortions not only obscure critical details for downstream tasks like object detection but also amplify risks in emergency rescue scenarios, where timely and accurate visual interpretation is paramount. This section reviews the evolution of image enhancement techniques, from classical interpolation to cutting-edge generative models, with a focus on super-resolution (SR) methods adaptable to UAV applications.

2.1 Traditional and Early Deep Learning Approaches

Early SR methods relied on interpolation techniques, such as bicubic or Lanczos resampling, which reconstruct high-resolution (HR) images by interpolating pixel values from low-resolution (LR) inputs. While computationally efficient, these approaches often introduce blurring and fail to recover high-frequency details, limiting their utility in dynamic low-altitude scenes [12]. The integration of deep learning marked a significant leap, with convolutional neural networks (CNNs) enabling end-to-end mapping from LR to HR domains.

Pioneering works like SRCNN [12] introduced a three-layer CNN to learn a non-linear mapping between LR image patches and HR counterparts, achieving substantial improvements in peak signal-to-noise ratio (PSNR) over interpolation (e.g., +2-4 dB on standard benchmarks like Set5). Building on this, FSRCNN [13] and ESPCN [14] optimized for efficiency by employing sub-pixel convolutions and shallower architectures, reducing inference time to real-time levels suitable for UAV edge computing (e.g., <10 ms per image on embedded GPUs). These models excel in structured environments but struggle with perceptual realism, often producing over-smoothed outputs that hinder texture recovery in blurred UAV footage.

To address spatial variability in low-altitude imagery—such as varying blur kernels from UAV vibrations—advanced CNN variants incorporated deformable convolutions and flow guidance. For instance, the Pyramid FG-DCN [15] leverages optical flow estimation and multi-scale deformable convolutions within a Swin Transformer backbone, enhancing alignment in motion-degraded sequences. Evaluations on UAV-specific datasets (e.g., simulated low-light urban flights) report up to 15% gains in structural similarity index (SSIM), underscoring their relevance for rescue operations where motion artifacts dominate.

2.2 Generative Adversarial Networks for Perceptual Enhancement

The quest for visually plausible HR images led to the adoption of generative adversarial networks (GANs), which pit a generator against a discriminator to minimize perceptual losses beyond pixel-wise metrics. SRGAN [16] pioneered this by incorporating a VGG-based perceptual loss, trained on adversarial objectives to favor natural textures over PSNR optimization. On LR images with 4× downsampling, SRGAN achieved superior visual quality, with learned perceptual image patch similarity (LPIPS) scores dropping by 20% compared to CNN baselines, making it apt for UAV scenarios where semantic fidelity aids target identification.

Subsequent enhancements addressed GAN instabilities, such as mode collapse, through architectural refinements. ESRGAN [17] introduced residual-in-residual dense blocks (RRDB) to deepen the generator without gradient vanishing, yielding sharper edges in foggy low-altitude captures (e.g., +0.5 dB in PSNR and reduced artifacts on DIV2K dataset). ESRGAN+ [18] further replaced RRDB with RRDRB for better detail preservation, while a U-Net discriminator in [19] jointly optimizes global and local contexts, outperforming vanilla GANs in extreme SR (8× scaling) by integrating LPIPS with adversarial losses. These methods have been adapted for UAVs by fine-tuning on domain-specific degradations; for example, incorporating weather-augmented datasets improves robustness in rain-distorted imagery, with reported mIoU gains of 5-10% in downstream segmentation tasks [20].

Despite their strengths, GANs suffer from training instability and hallucination risks, particularly in underrepresented low-altitude anomalies like specular reflections from urban glass facades [22].

2.3 Diffusion Models: A Paradigm Shift for Robust UAV Enhancement

Diffusion models (DMs) have recently supplanted GANs as the state-of-the-art for image generation and SR,

offering stable training via iterative denoising of Gaussian noise added to HR images. The forward process corrupts data progressively, while the reverse learns to reconstruct, enabling high-fidelity sampling without adversarial pitfalls.

In UAV contexts, DMs shine for handling complex, multi-factor degradations. DMDC [20] augments standard DMs with a detail-complement mechanism, randomly masking regions to simulate occlusions and enforcing pixel-wise constraints for fidelity—ideal for low-altitude partial views. On remote sensing benchmarks akin to UAV flights, it boosts PSNR by 1.2 dB over ESRGAN while preserving diversity. Dual-diffusion frameworks [21] extend this by estimating degradation kernels via a conditioner, then applying conditional denoising with U-Net backbones guided by LR encodings. This blind SR capability is crucial for unpredictable low-altitude conditions, achieving 92% perceptual alignment on custom UAV datasets with motion and weather blur.

For UAV-specific adaptations, integrating Vision Transformers (ViTs) into DM pipelines captures long-range dependencies in expansive urban skies. ViT-based feature extraction, followed by U-Net denoising conditioned on multi-view auxiliaries (e.g., adjacent frames), enhances temporal coherence [1]. Preliminary results indicate 10-15% improvements in downstream detection AP under low-light, positioning DMs as frontrunners for real-time UAV enhancement.

2.4 Challenges and UAV-Specific Innovations

Key challenges in low-altitude enhancement include real-time constraints on resource-limited UAV hardware and domain gaps between training data and operational extremes. Innovations like hybrid DM-CNN cascades [21] mitigate this by distilling DMs into lightweight proxies, reducing latency to <50 ms while retaining 95% of quality. Future directions emphasize self-supervised fine-tuning on UAV telemetry, leveraging multi-temporal imagery for adaptive enhancement [20].

In summary, from CNN efficiencies to DM's generative prowess, image enhancement has matured into a resilient enabler for UAV perception, with diffusion models poised to dominate low-altitude rescue applications.

3. Multi-Target Detection and Recognition via Vision-Radar Fusion

Multi-target detection and recognition form the core of UAV perception in low-altitude complex environments, where identifying diverse objects—such as pedestrians, vehicles, buildings, and aerial obstacles—amid clutter, varying scales, and occlusions is essential for emergency rescue tasks like victim localization and hazard avoidance [11]. Vision provides rich semantic cues (e.g., textures, colors) but degrades under poor visibility, while radar (e.g., mmWave or LiDAR) offers robust range-velocity estimates immune to weather [1-4]. Fusion of these modalities mitigates individual weaknesses, enhancing precision and recall in dynamic urban airspace. This section surveys fusion strategies, from early concatenation to sophisticated cross-modal alignments, emphasizing UAV-adapted advancements.

3.1 Early Fusion Approaches: Parallel and Hierarchical Integration

Initial fusion efforts focused on parallel processing of vision and radar streams, followed by simple concatenation or early merging to leverage complementary features. RVNet [5] introduced a dual-branch YOLOv3 architecture, projecting radar data onto image planes and concatenating features before dual detection heads tailored to large/small targets. This yielded 10-15% mAP improvements on foggy benchmarks, suitable for UAVs navigating obscured skylines. Similarly, CRFNet [6] employed RetinaNet for multi-level feature fusion, incorporating a "BlackIn" module to bias learning toward sparse radar cues, boosting recall by 8% in low-visibility scenarios.

CenterFusion [7] advanced this with end-to-end filtering: CenterNet predicts initial 3D bounding boxes from radar projections, refined via pillar expansion and frustum association to suppress clutter, achieving 3D mAP of 45% on KITTI datasets—transferable to UAV low-altitude tracking of moving entities like drones or birds. These methods excel in computational efficiency (e.g., 20 FPS on edge devices) but suffer from misalignment, as radar's sparse point clouds misalign with dense images without explicit calibration [11].

3.2 Advanced Cross-Modal Fusion: Spatial and Semantic Alignment

To address projection distortions in non-overlapping viewpoints, later works emphasized spatial transformations and attention mechanisms. AVOD [8] pioneered view aggregation, using feature pyramid networks (FPNs) on RGB images and bird's-eye-view (BEV) radar maps for multi-scale 3D proposals, reporting 12% AP gains on nuScenes for urban clutter—directly applicable to UAV collision avoidance.

MVX-Net [9] enhanced LiDAR-vision fusion via early voxel-based encoding, converting point clouds to voxels and fusing with semantic image features through 3D convolutions, improving small-object detection (e.g., pedestrians) by 20% in dense scenes. For UAVs, where viewpoints shift rapidly, 3D-CVF [10] introduced cross-view spatial fusion: auto-calibrated projections map 2D camera features to BEV LiDAR grids using gated attention, resolving coordinate inconsistencies and yielding 5-10% higher mAP in rainy conditions.

TransFusion [22] further integrated Transformers for robust LiDAR-camera fusion, employing cross-attention to align sparse radar with dense visuals, achieving state-of-the-art 3D detection on Waymo (mAP 75%) while maintaining low latency—critical for real-time UAV swarms in rescue formations.

3.3 Challenges in Low-Altitude Fusion

UAV-specific hurdles include spatio-temporal desynchronization from motion [7] and modality imbalance, where radar sparsity overwhelms vision in clutter [11]. Benchmarks like nuScenes reveal fusion gaps: single-

modality vision hits 40% mAP in fog, radar 25%, but hybrids reach 60% only with alignment [10]. Edge deployment demands quantization, with models like CRFNet [6] pruned to 15 FPS on UAV SoCs.

3.4 UAV-Oriented Innovations and Future Directions

Tailored for low-altitude, Unet++-preprocessed fusion networks [document] extract clear targets before multi-modal splicing, enhancing differentiation via adversarial training for weather robustness. Future trends point to end-to-end Transformers with dynamic fusion weights, self-supervised on UAV flight logs, and benchmarks incorporating low-altitude dynamics (e.g., wind-induced jitter) to push AP beyond 85% [10].

In essence, vision-radar fusion has transitioned from brittle concatenations to adaptive, attention-driven paradigms, fortifying UAV detection for resilient rescue operations.

4. Multi-Target Tracking and Trajectory Prediction in Dynamic Low-Altitude Scenes

Multi-target tracking (MOT) and trajectory prediction are indispensable for UAVs in emergency rescue, enabling sustained monitoring of dynamic entities (e.g., survivors, vehicles) across occlusions, viewpoint shifts, and environmental flux in urban low-altitude airspace [24,39]. Vision excels in appearance-based association but falters in clutter; radar provides velocity cues yet lacks semantic depth [1-4]. Fusion-driven approaches, integrating temporal models with multi-modal inputs, bridge these gaps for robust, predictive perception. This section traces MOT evolution from filter-based trackers to Transformer and graph neural network (GNN) hybrids, spotlighting UAV-suited innovations for occlusion handling and long-horizon forecasting.

4.1 Classical and Learning-Free MOT: Foundations in Motion Modeling

Early MOT paradigms decoupled detection from association, relying on Kalman filters for prediction and Hungarian algorithms for data linking. SORT [25] streamlined online 2D tracking by associating detections via intersection-over-union (IoU) and linear motion models, achieving real-time speeds (30 FPS) but degrading under occlusions—prevalent in UAV low-altitude views of crowded streets. DeepSORT [47] augmented this with CNN-extracted appearance embeddings and Mahalanobis distance metrics, reducing ID switches by 50% on MOT16 benchmarks, adaptable to fused vision-radar inputs for velocity-augmented associations.

In 3D realms, AB3DMOT [46] extended Kalman filtering to image-derived 3D tracks, using Hungarian matching on bird's-eye-view projections; however, lacking explicit range data, it underperforms in depth-ambiguous low-altitude flights (MOTA ~60% on KITTI). These methods offer lightweight baselines for UAV edge processing but falter in non-linear trajectories induced by wind or maneuvers [29].

4.2 Deep Learning-Enhanced MOT: CNNs and End-to-End Paradigms

CNN-driven trackers integrated detection and association for joint optimization. CenterTrack [24] predicted heatmap offsets between frames via a motion module, yielding 70% MOTA on MOT17 while handling short-term occlusions—vital for UAV pursuits amid urban foliage. FairMOT [54] balanced re-identification with detection using center-based representations, minimizing ID switches in crowded scenes.

For 3D MOT, SimTrack [35] projected LiDAR features to BEV for neural motion prediction, surpassing baselines by 5% in AMOTA on nuScenes; SimpleTrack [37] refined association via improved Kalman variants, emphasizing data linking over detection. Fusion variants like EagerMOT [32] staged 3D-2D associations with Hungarian solvers on radar-image hybrids, boosting robustness in adverse weather (HOTA +8%). JMODT [31] jointly optimized detection-tracking with point cloud-image inputs, leveraging neural associations for 65% MOTA in dynamic environments.

These CNN-centric methods scale well to UAV fusion but struggle with global context, e.g., predicting swarm trajectories in multi-UAV rescues [44].

4.3 Transformer and Graph-Based MOT: Global Reasoning for Complex Scenarios

Transformers' self-attention revolutionized MOT by modeling long-range dependencies. TrackFormer [36] treated tracking as set prediction with query propagation, achieving 76% HOTA on MOT20 sans post-processing—ideal for UAVs tracking erratic low-altitude targets. MOTR [53] end-to-end learned track queries via byte associations, reducing switches by 20%; GTR [51] globalized this with dense Transformer representations, excelling in occlusions (MOTA 80% on DanceTrack). MeMOT [26] incorporated memory banks for re-identification, enhancing long-term tracking in sparse radar-vision streams.

Graph neural networks (GNNs) complement Transformers by encoding inter-target topologies. GSM [55] modeled similarities via graph attention [56], improving association in cluttered UAV views; the Unified Graph Tracker (UGT) [document] captures high-order relations through frame and association graphs, using normalized Gaussian Wasserstein distances for spatial modeling and Transformer for linking—recovering occluded targets with 15% fewer misses. CAMO-MOT [44] fused camera-LiDAR motion-appearance via cost matrices, state-of-the-art on nuScenes (AMOTA 62%), directly transferable to low-altitude radar fusion for trajectory prediction amid buildings.

Quantum-inspired modules like QEM in MotionTrack [document] aggregate historical queries, boosting detection in multi-modal setups.

4.4 Challenges and UAV-Specific Advancements

Low-altitude MOT grapples with irregular UAV motion [38], modality asynchrony [31], and prediction horizons beyond 5s [67]. Benchmarks like nuScenes show fusion lifts MOTA from 50% (vision-only) to 70% [44], yet edge latency exceeds 100 ms. Innovations include hybrid

GNN-Transformer UGT frameworks [document] for real-time occlusion recovery and MADDPG-augmented prediction [66] for multi-UAV swarms.

Prospects lean toward unified end-to-end models with self-supervised fusion on flight data, targeting 85% AMOTA in dynamic rescues.

In sum, from filter simplicity to graph-Transformer synergy, MOT has evolved into a predictive powerhouse, empowering UAVs for proactive low-altitude interventions.

5. Low-Altitude Environment Assessment and Autonomous Path Planning

Low-altitude environments in urban settings pose multifaceted risks to UAVs during emergency rescue, including static obstacles (e.g., buildings, power lines), dynamic factors (e.g., weather variability, crowds), and interference (e.g., signal jamming) [62]. Accurate environment assessment quantifies these hazards, while autonomous path planning optimizes trajectories to minimize exposure, ensuring mission success and safety. This section reviews integrated approaches fusing multi-modal data for risk modeling and planning, from graph-based spatial representations to reinforcement learning (RL)-driven decision-making, with emphasis on UAV scalability in real-time, uncertain scenarios.

5.1 Environment Modeling and Risk Assessment: Graph and Multi-Modal Fusion

Core to assessment is discretizing complex airspace into navigable structures. Graph theory enables grid-based modeling, where vertices denote waypoints and edges connectivity, facilitating risk propagation [62]. Zhang et al. [62] proposed a regional risk-aware UAV routing framework, partitioning low-altitude zones into grids and assigning probabilistic collision scores via obstacle density and height profiles—achieving 20% safer paths in simulated urban rescues compared to A*-based planners.

Depth learning augments this with semantic risk extraction. Frameworks like AirMatrix [63] simulate cityscapes for rule-based risk mapping, identifying no-fly zones from LiDAR-vision fusions under seismic or pandemic constraints. Sensor fusion techniques [64] integrate radar for all-weather depth and cameras for semantic labels, using CNNs to classify hazards (e.g., turbulence zones) with 90% accuracy on custom low-altitude datasets. 3D risk maps visualize aggregated threats, overlaying quantified metrics—e.g., collision probability from Monte Carlo simulations [65]—onto geospatial grids, enabling UAVs to dynamically reroute around high-risk volumes like stormy corridors.

Self-adaptive models address dynamism: Guo et al. [66] employed improved deep RL for navigation, incorporating environmental feedback loops to update risk gradients in real-time, reducing deviation errors by 15% in windy trials.

5.2 Path Planning Algorithms: From Optimization to RL Paradigms

Classical planners like A* and RRT* offer optimality in known spaces but falter in partial observability [65]. Meta-heuristic hybrids [65]—e.g., genetic algorithms tuned for UAV energy constraints—generate collision-free paths, yet overlook multi-UAV coordination essential for swarm rescues.

RL marks a shift toward adaptive, learning-based planning. Q-learning variants [66] train UAVs to select actions maximizing coverage while minimizing risk, converging to near-optimal policies in 3D grids (reward convergence in <500 episodes). For multi-agent scenarios, hierarchical deep RL [67] decouples high-level task allocation from low-level trajectory generation, enabling swarms to explore unknown disaster zones collaboratively—boosting coverage by 30% over greedy methods on benchmarks like AirSim.

MADDPG extensions [document] further optimize multi-UAV exploration: actor-critic networks learn joint policies for path progression and information scouting, using potential field rewards to guide initial convergence and iterative target-point generation for full trajectories. Evaluations show 25% efficiency gains in occluded environments, with continuous action spaces accommodating velocity/heading nuances.

5.3 Challenges in Integrated Assessment-Planning Pipelines

UAV pipelines face scalability issues: real-time risk updates strain edge compute [63], while multi-agent credit assignment in RL leads to suboptimal coordination [67]. Gaps include sparse benchmarks for low-altitude extremes (e.g., <100m altitudes) and fusion latency in heterogeneous swarms [65]. Current systems achieve 95% success rates in simulations but drop to 80% in field tests due to unmodeled gusts [62].

5.4 UAV Innovations and Emerging Directions

Innovations like 3D risk graphs with adaptive meshing [62] and RL-data augmentation [66] enhance generalization. Future trajectories envision end-to-end neuro-symbolic planners, blending graph priors with Transformer-RL for predictive risk forecasting, targeting sub-second planning in 5G-enabled swarms [65,67].

Overall, assessment-planning synergies have propelled UAV autonomy from reactive dodging to proactive, resilient navigation, vital for scaling emergency responses.

6. Challenges and Prospects

The integration of vision-radar fusion into UAV perception systems has markedly advanced capabilities for low-altitude emergency rescue, from enhanced imagery in degraded conditions [20,21] to precise multi-target handling [10,44] and adaptive navigation [62,66]. Yet, as surveyed, persistent challenges underscore the need for holistic innovations to realize fully autonomous, resilient operations.

6.1 Key Challenges

1.Spatio-Temporal Alignment and Modality Imbalance: Fusion pipelines grapple with asynchronous data streams—vision at 30 FPS versus radar's sparse pulses—exacerbating errors in dynamic low-altitude flights [7,31]. Imbalanced contributions, where radar's noise overwhelms vision in clutter [11], degrade mAP by 10-20% without adaptive

weighting [22]. UAV-specific motion (e.g., vibrations) amplifies projection distortions, as noted in 3D-CVF evaluations [10].

2.Real-Time Constraints on Edge Hardware: Achieving 30+ FPS for end-to-end pipelines remains elusive on UAV payloads (e.g., Jetson Nano limits models like TransFusion [22] to 15 FPS) [6]. Occlusion recovery in MOT [36,53] and risk mapping [62] demands high-fidelity graphs, inflating compute by 2-3x, while power budgets constrain multi-modal processing [65].

3.Dataset and Benchmark Gaps: Low-altitude rescue lacks standardized corpora; nuScenes/KITTI adaptations [44] overlook urban specifics like signal interference or swarm dynamics [67]. Diffusion-based enhancement [20] shines on remote sensing but falters on UAV-scale degradations, with only 70-80% transfer to real flights [21].

4.Robustness to Unforeseen Scenarios: Extreme weather or adversarial jamming erodes fusion gains—e.g., mmWave Doppler artifacts in rain [1-4]—while multi-UAV planning suffers from non-stationary policies in RL [66], yielding 15-25% suboptimal coverage in simulations [67].

5.Ethical and Regulatory Hurdles: Privacy in vision streams and spectrum allocation for radar raise deployment barriers, compounded by certification needs for rescue-critical systems [65].

6.2 Prospects and Future Directions

Addressing these paves the way for transformative advancements:

Unified End-to-End Architectures: Hybrid Transformer-GNN models [36,51,55], extended with DM priors [20], promise seamless fusion-detection-tracking-planning, targeting 90%+ AMOTA and sub-50ms latency via neural architecture search.

Self-Supervised and Federated Learning: Leverage UAV swarms for on-the-fly data synthesis [21], federating models across fleets to bridge dataset voids without centralization, enhancing generalization to rare events like wildfires [62].

Edge-Optimized Fusion: Quantized, spiking neural networks integrated with 5G offloading [67] could halve power draw, enabling persistent multi-UAV ops. Bio-inspired adaptive fusion [64], mimicking human multisensory integration, offers promise for occlusion-robust prediction.

Benchmark and Standardization Initiatives: Curating low-altitude datasets with synthetic augmentations [15] and metrics like risk-aware HOTA will accelerate progress, fostering open-source simulators beyond AirSim [63].

Interdisciplinary Synergies: Coupling with edge AI hardware (e.g., neuromorphic chips) and regulatory sandboxes [65] will expedite field trials, amplifying societal impact in disaster response.

In conclusion, vision-radar fusion stands at the cusp of enabling UAVs as proactive guardians in urban crises, demanding concerted efforts to surmount computational and environmental barriers. By 2030, we envision swarms autonomously orchestrating rescues with near-human acuity, saving lives through perceptive agility [62,67]..

REFERENCES

[1] Zhang R, Cao S. Real-time human motion behavior detection via CNN using mmWave radar[J]. IEEE Sensors Letters, 2018, 3(2): 1-4.

[2] Yoneda K, Hashimoto N, Yanase R, et al. Vehicle localization using 76GHz omnidirectional millimeter-wave radar for winter automated driving[C]//2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018: 971-977.

[3] Nagasaku T, Kogo K, Shinoda H, et al. 77GHz low-cost single-chip radar sensor for automotive ground speed detection[C]//2008 IEEE Compound Semiconductor Integrated Circuits Symposium. IEEE, 2008: 1-4.

[4] Hines M E, Zelubowski S A. Conditions affecting the accuracy of speed measurements by low power MM-wave CW Doppler radar[C]//[1992 Proceedings] Vehicular Technology Society 42nd VTS Conference-Frontiers of Technology. 1992: 1046-1050.

[5] John V, Mita S. RVNet: Deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments[C]//Image and Video Technology: 9th Pacific-Rim Symposium, Sydney, NSW, Australia, 2019: 351-364.

[6] Nobis F, Geisslinger M, Weber M, et al. A deep learning-based radar and camera sensor fusion architecture for object detection[C]// Sensor Data Fusion: Trends, Solutions, Applications (SDF). 2019: 1-7.

[7] Nabati R, Qi H. Centerfusion: Center-based radar and camera fusion for 3d object detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 1527-1536.

[8] Ku J, Mozifian M, Lee J, et al. Joint 3d proposal generation and object detection from view aggregation[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems, 2018: 1-8.

[9] Sindagi V A, Zhou Y, Tuzel O. Mvx-net: Multimodal voxelnet for 3d object detection[C]//International Conference on Robotics and Automation, 2019: 7276-7282.

[10] Yoo J H, Kim Y, Kim J, et al. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection[C]//Computer Vision European Conference, Glasgow, UK, 2020: 720-736.

[11] Arnold E, Al-Jarrah O Y, Dianati M, et al. A survey on 3d object detection methods for autonomous driving applications[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3782-3795.

[12] Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In European Conference on Computer Vision; Springer: Cham, Switzerland, 2014; pp. 184–199.

[13] Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In European Conference on Computer Vision; Springer: Cham, Switzerland, 2016; pp. 391–407.

[14] Shi, W.; Caballero, J.; Huszár, F.; et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.

[15] Luo, Z.; Li, Y.; Cheng, S.; et al. BSRT: Improving Burst Super-Resolution With Swin Transformer and Flow-Guided Deformable Alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, LA, USA, 19–20 June 2022; pp. 998–1008.

[16] Ledig, C.; Theis, L.; Huszár, F.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21 – 26 July 2017; pp. 4681–4690.

[17] Wang, X.; Yu, K.; Wu, S.; et al. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.

[18] Rakotonirina, N.C.; Rasoanaivo, A. ESRGAN+: Further improving enhanced super-resolution generative adversarial network. In Proceedings of the ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3637–3641.

[19] Jo, Y.; Yang, S.; Kim, S.J. Investigating loss functions for extreme super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 424–425.

[20] J. Liu, Z. Yuan, Z. Pan, Y. Fu, L. Liu, and B. Lu, "Diffusion model with detail complement for super-resolution of remote sensing, " Remote Sensing, vol. 14, no. 19, 2022.

[21] M. Xu, J. Ma, and Y. Zhu, " Dual-diffusion: Dual conditional denoising diffusion probabilistic models for blind super-resolution reconstruction in rs is," arXiv:2305.12170, 2023.

[22] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-camera fusion for 3d object detection with transformers. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1080–1089, 2022.

[23] Xingyi Zhou, Dequan Wang, Philipp Krähenbühl. Objects as Points: Tracking-by-Detection with One Regression Head. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[24] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Tozeto Ramos, Ben Upcroft. Simple Online and Realtime Tracking. 2016 IEEE International Conference on Image Processing (ICIP), pages 3464–3468, 2016.

[25] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, Stefano Soatto. MeMOT: Multi-Object Tracking with Memory. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8080–8090, 2022.

[26] Karl Granström, Marcus Baum. Extended Object Tracking: Introduction, Overview and Applications. ArXiv, abs/1604.00970, 2016.

[27] Kemiao Huang, Qi Hao. Joint Multi-Object Detection and Tracking with Camera-LiDAR Fusion for Autonomous Driving. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 6983–6989, 2021.

[28] Aleksandr Kim, Aljosa Osep, Laura Leal-Taixé. EagerMOT: 3D Multi-Object Tracking via Sensor Fusion. 2021 IEEE International Conference on Robotics and Automation (ICRA), pages 11315–11321, 2021.

[29] Chenxu Luo, Xiaodong Yang, Alan Loddon Yuille. Exploring Simple 3D Multi-Object Tracking for Autonomous Driving. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 10468–10477, 2021.

[30] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, Christoph Feichtenhofer. TrackFormer: Multi-Object Tracking with Transformers. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8834–8844, 2022.

[31] Ziqi Pang, Zhichao Li, Naiyan Wang. SimpleTrack: Understanding and Rethinking 3D Multi-Object Tracking. ArXiv, abs/2111.09621, 2021.

[32] Lionel Rakai, Huansheng Song, ShiJie Sun, Wentao Zhang, Yanni Yang. Data Association in Multiple Object Tracking: A Survey of Recent Techniques. Expert Systems with Applications, 192:116300, 2022.

[33] Li Wang, Xinyu Newman Zhang, Wenyuan Qin, Xiaoyu Li, Lei Yang, Zhiwei Li, Lei Zhu, Hong Wang, Jun Li, Hua Liu. CAMO-MOT: Combined Appearance-Motion Optimization for 3D Multi-Object Tracking with Camera-LiDAR Fusion. ArXiv, abs/2209.02540, 2022.

[34] Xinshuo Weng, Jianren Wang, David Held, Kris Kitani. AB3DMOT: A Baseline for 3D Multi-Object Tracking and New Evaluation Metrics. ArXiv, abs/2008.08063, 2020.

[35] Nicolai Wojke, Alex Bewley, Dietrich Paulus. Simple Online and Realtime Tracking with a Deep Association Metric. 2017 IEEE International Conference on Image Processing (ICIP), pages 3645–3649, 2017.

[36] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, Ping Luo. TransTrack: Multiple Object Tracking with Transformer. arXiv preprint arXiv:2012.15460, 2020.
[37] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, Phillip Krähenbühl. Global Tracking Transformers. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8761–8770, 2022.
[38] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, Yichen Wei. MOTR: End-to-End Multiple Object Tracking with Transformer. Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII, page 659–675, Berlin, Heidelberg, 2022. Springer-Verlag.
[39] Y. Zhang, C. Wang, X. Wang, W. Zeng, W. Liu. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. Int. J. Comput. Vis., vol. 129, no. 11, pp. 3069–3087, 2021.
[40] Q. Liu, Q. Chu, B. Liu, N. Yu. GSM: Graph Similarity Model for Multi-Object Tracking. Proc. 29th Int. Joint Conf. Artif. Intell., Jul. 2020, pp. 530–536.
[41] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio. Graph Attention Networks. arXiv:1710.10903, 2017.
[42] 张宏宏, 甘旭升, 李双峰, 等. 复杂低空环境下考虑区域风险评估的无人机航路规划[J]. 仪器仪表学报, 2023 (1): 257-266.
[43] [Chongqing University AirMatrix simulation platform; as referenced in document for urban environment modeling].
[44] [Multi-modal sensor fusion for risk assessment; document-derived].
[45] Yahia H S, Mohammed A S. Path planning optimization in unmanned aerial vehicles using meta-heuristic algorithms: A systematic review[J]. Environmental Monitoring and Assessment, 2023, 195(1): 30.
[46] 郭子恒, 蔡晨晓. 基于改进深度强化学习的无人机自主导航方法[J]. Information & Control, 2023, 52(6).
[47] 吕超, 李慕宸, 欧家骏. 基于分层深度强化学习的无人机混合路径规划[J]. 北京航空航天大学学报, 2023.