Multimodal Large Models-Driven Precise Perception in Complex Low-Altitude UAV Environments: A Survey on Adaptive Edge Intelligence and Swarm Collaboration

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

(1.Beijing Jiazhixing Commercial Co., Ltd, Beijing 100124； 2. CITIC Consumer Finance Co., Ltd,Beijing 100000； 3.Fortune Pass (Beijing) Information Technology Co., Ltd，Beijing 100016)

ABSTRACT Unmanned aerial vehicles (UAVs) equipped with long-endurance remote sensing capabilities have revolutionized applications in economic development, national defense, emergency response, and disaster monitoring. However, traditional centralized processing paradigms suffer from high latency, resource inefficiency, and poor adaptability to dynamic low-altitude environments. This survey reviews advancements in multimodal large models (MLMs) for precise detection and perception in complex low-altitude scenarios, emphasizing three core challenges: enhancing intelligent terminal perception for adaptive learning, bolstering multi-UAV collaborative coverage through federated evolution, and achieving high-fidelity 3D perception via multimodal fusion. We synthesize recent developments in self-evolving online learning frameworks, asynchronous distributed federated optimization, and Transformer-based MLMs tailored for heterogeneous sensor data (e.g., LiDAR and multi-view cameras). Key contributions include a taxonomy of adaptive algorithms mitigating catastrophic forgetting and data heterogeneity, alongside benchmarks for edge deployment in resource-constrained UAV systems. By highlighting gaps in unsupervised multimodal alignment and real-time scalability, this work outlines future directions toward autonomous, resilient UAV swarms, fostering innovations in edge intelligence for spatial information technologies.

Keywords Multimodal large models, UAV perception, low-altitude environments, federated learning, self-evolving adaptation.

I. Introduction

Unmanned aerial vehicles (UAVs) serve as pivotal platforms for multi-scale, multi-perspective observation of terrestrial features, underpinning critical domains such as economic growth, national security, disaster mitigation, and resource surveying [1]. The integration of long-endurance remote sensing—spanning satellites and UAVs—has enabled unprecedented data acquisition, yet the prevailing "edge collection-centralized processing" workflow introduces bottlenecks: prolonged transmission delays, excessive bandwidth demands, and sluggish model updates ill-suited to the volatility of low-altitude operations [2]. As environments grow increasingly complex—with dynamic occlusions, heterogeneous threats (e.g., smoke, drones), and sparse annotations—there is an urgent need to shift computation to the edge, empowering UAVs with autonomous, self-evolving intelligence [3].

This paradigm shift aligns with global initiatives accelerating edge-native capabilities. The U.S. DARPA Blackjack program envisions low-Earth orbit constellations with in-orbit collaboration and decision-making [4], while the EU's Future Sky program and Japan's 2035 drone swarm roadmap prioritize collaborative autonomy [5]. In parallel, multimodal large models (MLMs), building on Transformer architectures, have surged as enablers for fusing diverse sensor streams (e.g., visual, LiDAR, spectral data), surpassing unimodal limits in 3D scene understanding [6]. Yet, deploying MLMs on UAVs confronts tripartite hurdles: (i) high adaptability—online learning must counter catastrophic forgetting in streaming, unlabeled data [7]; (ii) high collaboration—federated systems require robust handling of data heterogeneity and topological flux in multi-UAV swarms [8]; and (iii) high perception fidelity—cross-modal alignment demands efficient fusion without exhaustive annotations [9].

This survey provides a comprehensive synthesis of MLM-driven technologies for precise low-altitude perception, distilling theoretical foundations, algorithmic innovations, and empirical validations. Unlike prior reviews focused on generic multimodal fusion [10] or federated learning in static networks [11], we center on UAV-specific edge constraints, offering a structured lens through self-adaptive terminal enhancement, multi-agent coverage amplification, and MLM-centric 3D reconstruction. Section 2 delineates intelligent terminal perception advancements, including self-supervised online paradigms. Section 3 explores multi-UAV federated evolution for coverage. Section 4 delves into Transformer-based MLMs for multimodal sensing. Section 5 benchmarks performance metrics and deployment challenges, concluding with prospective trajectories toward fully autonomous aerial ecosystems.

By bridging these silos, this work not only illuminates the trajectory from isolated UAV sensing to symbiotic swarms but also equips researchers with actionable insights to propel edge intelligence beyond current frontiers.

II. Intelligent Terminal Perception Advancements

Intelligent terminal perception forms the cornerstone of edge-enabled UAV systems, enabling real-time adaptation to dynamic, unlabeled data streams in complex low-altitude environments. Traditional offline learning paradigms falter under the deluge of heterogeneous aerial imagery—characterized by scale variations, spectral complexities, and sparse annotations—leading to inefficiencies in resource-constrained platforms [12]. Recent advancements pivot toward self-supervised online learning frameworks that facilitate incremental, autonomous model evolution, mitigating catastrophic forgetting while enhancing cross-scene generalization [7]. This section delineates key progress in self-supervised paradigms, adaptive self-evolving algorithms, dataset augmentation strategies, and domain-specific adaptations tailored for UAV remote sensing.

2.1 Cross-Scene Self-Supervised Online Learning

Self-supervised online learning has emerged as a pivotal technique for UAVs, leveraging unlabeled data flows to extract transferable knowledge representations without exhaustive human intervention. By distilling intrinsic supervisory signals from data perturbations—such as rotations, spectral shifts, or temporal inconsistencies—these methods bootstrap robust features for downstream tasks like object classification, detection, and tracking in aerial scenes [3].

A foundational approach involves contrastive learning augmented with domain-invariant alignments. For instance, RS-FewShotSSL employs a deep self-supervised learner to classify remote sensing scenes under few-shot constraints, achieving superior performance on datasets with fewer than 20 labeled samples per class by aligning multi-level semantic hierarchies [13]. This is particularly salient for UAVs, where cross-scene transitions (e.g., from urban to rural terrains) induce distribution shifts; here, adversarial feature alignment and knowledge distillation reduce underlying discrepancies, elevating migration efficiency by up to 15% in benchmarks [14]. Similarly, FastSiam tailors efficient self-supervised pretraining for multispectral UAV imagery, utilizing momentum encoders to capture spectral-spatial correlations, outperforming supervised baselines in low-data regimes [15].

Mitigating catastrophic forgetting remains central, as sequential task learning often erodes prior knowledge. Techniques like hidden knowledge representation architectures, informed by contrastive-adversarial objectives, capture domain-invariant invariants, fostering high-reliability paradigms for streaming aerial data [16]. Empirical validations on SoundingEarth—a crowdsourced audiovisual dataset—demonstrate that such integrations yield 20-30% gains in cross-modal transfer for remote sensing tasks [17]. These paradigms underscore a shift from static pretraining to continual, edge-deployable learning, primed for UAVs navigating occluded or adversarial low-altitude vistas.

2.2 Adaptive Self-Evolving Learning Algorithms

To address the volatility of multi-scene dynamics—encompassing unordered, non-stationary inputs—adaptive self-evolving algorithms dynamically refine model architectures and parameters, emulating biological evolution via meta-learning and gradient-based exploration [18]. These methods, rooted in evolutionary computation, iteratively generate and prune network variants, optimizing for UAV-specific constraints like computational latency and energy budgets [19].

Evolutionary strategies, such as those in niche adaptive elite evolutionary algorithms (NAEEA), adapt swarm intelligence for clustering in aerial unmanned sensor networks, reducing energy overhead by 25% through fitness-guided mutations [20]. For drone perception, reinforcement learning-infused adaptations enable pathfinding under incomplete information; adaptive differential evolution (IADE) dynamically tunes mutation and crossover rates based on iteration progress and fitness landscapes, solving single-UAV multitasking with 18% improved convergence [21]. In multi-drone pursuits, unseen algorithm zoos—incorporating greedy and collaborative agents—facilitate teaming via proxy predictors, enhancing adaptability in simulated low-altitude chases [22].

Catastrophic forgetting is curtailed through progressive incremental learning, where variational inference selects evolution strategies on-the-fly, ensuring stability amid scene flux [23]. For aerial imagery, class-incremental detectors like those using knowledge inheritance modules preserve prior task proficiency, with distillation losses yielding 10-15% retention in incremental remote sensing object detection [16]. These self-evolving mechanisms not only bolster UAV autonomy but also pave the way for federated extensions, as explored in subsequent sections.

2.3 Dataset Construction and Augmentation for Remote Sensing Imagery

The scarcity of annotated aerial datasets hampers UAV perception; thus, strategic augmentation expands corpora while preserving semantic fidelity, simulating diverse flight conditions like varying altitudes or weather perturbations [24]. Geometric transformations—rotations, flips, scaling—and advanced synthesis via generative models form the bedrock, with deep learning-driven augmentations further enriching multispectral UAV inputs [25].

Recent pipelines, such as those for thermal aerial enhancement, introduce synthetic drone classes in urban scenes, augmenting baselines like HIT-UAV-TL to boost detection recall by 22% [26]. YOLOv9 variants, paired with transfer learning, apply adjustable augmentations to UAV vehicle detection datasets, mitigating overfitting in sparse regimes [27]. Composite augmentations, blending real and synthesized imagery, address data sparsity in semantic segmentation; by overlaying perturbations like fog or shadows, these yield 15% accuracy uplifts on sparse remote sensing benchmarks [28].

For maritime surveillance, augmentation via Stable Diffusion generates float-object variants, enhancing SAR detection algorithms with minimal annotation costs [29]. UAV-specific methods, including mosaic blending and mixup, simulate construction site variabilities, expanding datasets tenfold while curbing class imbalances [30]. These techniques, integrated with self-supervised loops, ensure scalable, robust training for edge terminals.

2.4 Domain-Specific Online Self-Evolving Learning for Remote Sensing

Tailoring online paradigms to remote sensing idiosyncrasies—such as high-resolution scale variances and spectral intricacies—demands domain-dedicated frameworks that fuse incremental learning with knowledge alignment [31]. These evolve models via lightweight, continual updates, alleviating forgetting through prototype storage and distillation [32].

In aerial contexts, HDCPAA employs few-shot incremental learning with prototypes to sustain classification amid evolving classes, achieving 12% gains over vanilla continual learners on remote sensing benchmarks [32]. Knowledge distillation variants, like those in SIL-LAND, distill segmentation heads for incremental aerial land-use mapping, preserving 90% prior performance via uncertainty-aware replays [33]. For UAVs, ER-PASS integrates experience replay with submodular selection, countering domain shifts in continual segmentation and yielding state-of-the-art forgetting mitigation [34].

Innovations in variational self-adaptation further enable semi-feedback loops, where minimal human oversight guides evolution, as in WVA for online tracking control [35]. Collectively, these advancements forge resilient, UAV-native perception engines, bridging to collaborative paradigms in Section 3.

III. Multi-UAV Collaborative Coverage Enhancement

Multi-UAV collaborative coverage enhancement addresses the imperative for swarm intelligence in low-altitude operations, where individual drones contend with limited sensing footprints, topological instabilities, and heterogeneous data distributions. Traditional centralized coordination falters under communication bottlenecks and single-point failures, necessitating distributed paradigms that amplify collective perception without raw data exchange [36]. Federated learning (FL) emerges as a linchpin, enabling model aggregation across UAVs while preserving privacy and mitigating latency in dynamic topologies [8]. This section surveys asynchronous distributed FL frameworks, generalized topological architectures, federated AutoML integrations, and knowledge migration techniques, underscoring their role in self-evolving edge ecosystems for UAV swarms.

3.1 Asynchronous Distributed Optimization in Federated Learning

Asynchronous distributed optimization decouples UAV updates from rigid synchronization, accommodating erratic flight patterns and intermittent links prevalent in low-altitude swarms [37]. By allowing local iterations to proceed independently, these methods curtail convergence delays and model drift, pivotal for real-time tasks like threat detection amid class imbalances [38].

Core advancements leverage variance-reduced stochastic gradients with event-driven communication, where UAVs upload parameters only upon significant deviations, slashing overhead by 40% in simulated mesh networks [39]. For instance, FedAvg extensions incorporate global gradient estimates to rectify local biases, as in FedProx variants tailored for UAV power constraints, yielding 25% faster convergence under non-IID data [40]. In multi-target tracking scenarios, dual-decomposition algorithms distribute Lagrangian relaxations across UAVs, optimizing trajectories asynchronously while enforcing coverage constraints [41].

Empirical benchmarks on UAV swarm datasets reveal robustness: asynchronous FL mitigates stragglers in heterogeneous fleets, with 15-20% gains in accuracy for intrusion detection over synchronous baselines [42]. These optimizations form the bedrock for scalable, resilient collaboration, transitioning to topological generalizations.

3.2 Generalized Topology Structures for Multi-UAV High-Performance Computing

UAV swarms often manifest fluid topologies—star, mesh, or ad-hoc—demanding FL frameworks that generalize across structures to sustain coverage in contested environments [43]. Generalized architectures abstract cloud-edge-end hierarchies, enabling seamless transitions via multiplier-based consensus protocols [44].

Recent surveys highlight hybrid topologies fusing NOMA-assisted FL with UAV relays, enhancing spectral efficiency and coverage by 30% in dense deployments [45]. For instance, decentralized FL over mesh networks employs gossip protocols for parameter dissemination, achieving near-centralized performance with 50% reduced bandwidth in 5G-enabled swarms [46]. In energy-constrained settings, distributed task assignment via auction-based mechanisms optimizes node selection, extending swarm endurance by 18% while preserving global optimality [47].

Blockchain-augmented topologies further secure FL aggregates, countering Byzantine faults in adversarial low-altitude ops [48]. Validations on real-world UAV fleets demonstrate 20% uplifts in collaborative mapping fidelity, underscoring the efficacy of these structures in amplifying perceptual breadth [49].

3.3 Federated Auto Machine Learning for Edge Computing

Federated AutoML automates hyperparameter tuning and architecture search across UAV edges, democratizing advanced ML amid resource scarcity [50]. By federating neural architecture search (NAS) with surrogate models, these paradigms evolve lightweight detectors in-situ, bypassing exhaustive grid searches [51].

Pioneering works integrate differentiable NAS into FL rounds, where UAVs collaboratively refine supernets via reinforcement learning proxies, converging 2x faster than local AutoML [52]. For drone inspection, cloud-edge-end FL clusters clients by data similarity, yielding personalized models with 12% accuracy boosts on heterogeneous sensors [53]. Adaptive frameworks like EdgeFed dynamically prune search spaces, curbing energy draw by 35% in IoT-drone hybrids [54].

Challenges in non-convex landscapes are met with uncertainty-weighted proxies, as in multi-robot SLAM extensions, enhancing generalization across swarm variants [49]. These integrations empower self-orchestrating UAVs, bridging to knowledge transfer mechanisms.

3.4 Federated Knowledge Migration for Edge Computing

Knowledge migration in FL facilitates cross-UAV expertise sharing, vital for domain shifts in evolving low-altitude threats [55]. Transfer learning-infused FL, such as KIBTL, distills pre-trained encoders via proxy datasets, accelerating convergence by 40% without data leakage [56].

In UAV networks, CORAL-aligned migrations harmonize feature statistics across clients, bolstering personalization in task offloading [57]. For multi-task swarms, attention-gated transfers balance related objectives, as in UAV-assisted FL where shared backbones yield 15% gains in federated IDS [58]. Privacy-preserving variants employ homomorphic encryption for gradient exchanges, safeguarding migrations in contested airspace [59].

Simulations on federated UAV benchmarks affirm efficacy: knowledge-distilled DQN variants optimize load balancing, reducing energy deviations by 22% [59]. These techniques culminate in holistic self-evolution, priming swarms for multimodal perception in Section 4.

IV. Transformer-Based Multimodal Large Models for Multimodal Sensing

Transformer-based multimodal large models (MLMs) represent a paradigm shift in UAV sensing, leveraging self-attention mechanisms to fuse heterogeneous inputs—such as multi-view cameras, LiDAR point clouds, and spectral imagery—into coherent 3D representations for low-altitude navigation [60]. Unlike convolutional backbones, Transformers excel in capturing long-range dependencies and cross-modal interactions, enabling precise perception amid occlusions, varying altitudes, and dynamic threats [61]. This section explores MLM architectures for 3D perception, parameter-efficient fine-tuning (PEFT) adaptations, and asynchronous distributed alignment for multi-UAV knowledge sharing, drawing on recent benchmarks and empirical advancements.

4.1 Multimodal Large Models for 3D Perception in Complex Low-Altitude Environments

At the core of low-altitude 3D perception lies the unification of diverse sensor modalities via Transformer encoders, which tokenize inputs into sequences for parallel processing, mitigating the pitfalls of sequential fusion in resource-limited UAVs [62]. Modality-specific tokenizers—e.g., patch-based for images and voxel encoders for LiDAR—feed into shared Transformer backbones, facilitating intra- and inter-modal learning through dynamic set attention and cross-attention blocks [63].

Pioneering frameworks like UAV3D benchmark Transformer-driven collaborative 3D detection, aggregating multi-UAV views to achieve 25% mAP gains over unimodal baselines on sparse aerial datasets [64]. The RA3T model exemplifies region-aligned adaptations, employing 3D sparse convolutions with Transformer decoders for self-supervised sim-to-real transfer, reducing domain gaps in urban low-altitude scenes by 18% [65]. For bird's-eye-view (BEV) mapping, BEVFusion variants extend Transformers with geometric projections, fusing camera-LiDAR tokens in 2D/3D spaces to enhance segmentation fidelity, as validated on nuScenes-UAV extensions with 15% IoU improvements [66].

Cross-modal interactions are amplified via alternating partitions: perspective-aligned for semantic bridging and geometric for depth-aware fusion, circumventing projection ambiguities through pre-computable offsets [67]. These models adapt to tasks like 3D object detection and BEV segmentation, with LSS-based enhancements yielding real-time inference under 50ms on edge hardware [68]. Benchmarks underscore Transformers' superiority, outperforming CNNs by 10-20% in multimodal UAV tracking amid wind perturbations [69].

4.2 Parameter-Efficient Fine-Tuning for Multimodal Model Adaptation

Deploying MLMs on UAVs demands PEFT to curb parameter explosion, enabling task-specific tuning with minimal overhead—critical for memory-constrained flights [70]. Techniques like adapters and low-rank adaptations (LoRA) insert lightweight modules into frozen backbones, preserving pre-trained knowledge while aligning to aerial domains [71].

The Position Insertion Module (PIN) innovates by injecting learnable spatial embeddings post-visual encoder, optimized via negative log-likelihood on synthetic bounding-box prompts, unlocking localization in vision-language models (VLMs) without altering core parameters

[72]. Empirical studies on MLLMs reveal PEFT variants like QLoRA yielding 12% accuracy boosts for UAV crack segmentation, fine-tuning only 0.1% of weights on multimodal asphalt datasets [73]. Aurora's prefix-tuning for large-scale multimodal foundations achieves 1.8% gains on video QA benchmarks with 0.05% tunable parameters, adaptable to UAV trajectory prediction [74].

For remote sensing, adaptive PEFT selects high-quality multimodal subsets via uncertainty sampling, accelerating convergence by 30% in land-cover mapping [75]. Surveys highlight prompt-based PEFT's efficacy in VLMs, with BitFit and sparse updates mitigating forgetting in continual aerial adaptation [76]. These methods democratize MLM deployment, transitioning to distributed paradigms.

4.3 Asynchronous Distributed Computing for Multi-UAV Knowledge Alignment

In multi-UAV swarms, asynchronous distributed computing ensures robust knowledge alignment across modalities, countering temporal misalignments and topological drifts via federated updates [77]. Tokenizers map inputs to shared embeddings, followed by CLIP-ViT encoders for contrastive alignment, with cross-attention fusing features in a unified space [78].

Frameworks like LLVM-Drone integrate LLMs for vision missions, employing homomorphic encryption in async FL to synchronize gradients without data exposure, enhancing swarm perception by 20% in collaborative localization [79]. IRADA's reward aggregation distributes task allocation, aligning multimodal states via submodular proxies for persistent monitoring [80]. For embodied VL, OODA-guided interactions facilitate human-swarm alignment, with DRL-infused async updates optimizing multimodal rewards in low-altitude pursuits [81].

On-device pipelining accelerates inference, compressing gradients for 2x throughput in edge MLMs, as in temporal attack mitigations preserving fusion integrity [82]. These alignments culminate in resilient, scalable sensing, as benchmarked on UAVScenes with 15% cross-UAV transfer gains [83].

V. Benchmarks, Performance Metrics, Deployment Challenges, and Future Directions

This section synthesizes empirical evaluations across the surveyed paradigms, benchmarking intelligent terminal perception, multi-UAV collaboration, and multimodal large models (MLMs) for low-altitude UAV sensing. We delineate key datasets and metrics, highlighting trade-offs in accuracy, latency, and resource utilization. Subsequently, deployment challenges in edge-constrained environments are dissected, followed by prospective trajectories toward fully autonomous aerial ecosystems, informed by emerging trends in adaptive AI and swarm orchestration.

5.1 Benchmarks and Performance Metrics

Benchmarking UAV perception requires multimodal datasets that capture low-altitude complexities—such as dynamic occlusions, spectral variances, and sparse annotations—while metrics must balance fidelity (e.g., mAP, IoU) with operational viability (e.g., inference latency, energy draw) [84]. Recent datasets like UAVScenes provide a large-scale multimodal corpus for 2D/3D tasks, including semantic segmentation and novel view synthesis, with baselines showing Transformer-based models achieving 45% mIoU on urban aerial scenes under varying altitudes [85]. Similarly, UEMM-Air evaluates multi-modal environmental perception, reporting 28% gains in cross-task generalization for federated setups, using metrics like task-averaged accuracy and transfer efficiency [86].

For federated learning in UAV swarms, performance hinges on communication efficiency and convergence speed. The AERPAW platform benchmarks anomaly detection, where async FL reduces training latency by 35% compared to centralized baselines, measured via rounds-to-convergence and per-round energy (e.g., 12 mJ per UAV iteration) [87]. In multi-task scenarios, task attention mechanisms in UAV-enabled FL yield 22% uplifts in global loss minimization, with key metrics including client drift variance and spectral efficiency under non-IID distributions [88]. Edge-specific evaluations, such as those on battery-constrained IoT-UAV networks, quantify trade-offs: FedProx variants cut energy by 40% while maintaining 92% accuracy in intrusion detection, benchmarked on simulated swarms with 50-node topologies [89].

Multimodal MLMs are assessed via 3D perception fidelity and fusion robustness. ATR-UMMIR, a benchmark for image registration under complex conditions, reports alignment errors below 2 pixels for RGB-TIR pairs, with Transformer decoders outperforming CNNs by 15% in perceptual hashing metrics [90]. Kust4K extends this to urban traffic segmentation, where BEV fusion achieves 62% mIoU, emphasizing cross-modal IoU and depth estimation RMSE (under 0.5m) for low-altitude viability [91]. Holistic metrics, like those in RGBTDronePerson, integrate detection latency (<100ms) and energy-normalized F1-scores, revealing 18% efficiency gains from PEFT-tuned VLMs [92].

Benchmark Dataset	Modalities	Key Tasks	Primary Metrics	Baseline Performance
UAVScenes [85]	RGB, Depth, LiDAR	Segmentation, Localization	mIoU, RMSE, Latency	45% mIoU (Transformer)
UEMM-Air [86]	RGB-TIR, Spectral	Environmental Perception	Task-Avg. Acc., Transfer Eff.	28% Gain (Federated)
ATR-UMMIR [90]	RGB-TIR	Registration, Fusion	Alignment Error, Hashing	<2px Error (Decoder)
Kust4K [91]	RGB-TIR	Segmentation	mIoU, Depth RMSE	62% mIoU (BEV Fusion)
AERPAW [87]	Network Logs	Anomaly Detection	Rounds-to-Conv., Energy/mJ	35% Latency Red. (Async FL)

These benchmarks underscore synergies: self-evolving paradigms boost adaptability (e.g., 20% forgetting reduction), while federated MLMs enhance scalability, though at 10-15% accuracy costs in heterogeneous swarms [93].

5.2 Deployment Challenges

Deploying these technologies on UAVs confronts multifaceted hurdles, from computational austerity to real-time exigencies in contested low-altitude regimes [94]. Resource constraints—limited payloads (e.g., <500g compute modules) and power budgets (10-50W)—amplify MLM inference overheads; full Transformer stacks exceed 100 GFLOPs, necessitating 4x quantization for <200ms latency, yet degrading precision by 8% in fusion tasks [95]. Edge federation exacerbates this: async updates in swarms induce model drift (up to 12% variance under link failures), demanding robust aggregation like variance-reduced gradients [96].

Privacy and security pose acute risks; federated gradients leak via inversion attacks, with UAV telemetry amplifying exposure in multi-agent settings [97]. Multimodal alignment falters under sensor asynchrony (e.g., 50ms LiDAR-camera offsets), yielding 15% fusion errors in dynamic scenes, while environmental factors like wind shear (>10m/s) inflate localization RMSE beyond 1m [98]. Scalability bottlenecks emerge in swarms: topological flux in 20+ UAVs spikes communication by 30%, mitigated imperfectly by NOMA relays [99]. Human-UAV interfaces lag, with VLMs struggling on ambiguous prompts (e.g., 25% misinterpretation in tactical commands), underscoring needs for embodied fine-tuning [100].

5.3 Prospective Trajectories Toward Fully Autonomous Aerial Ecosystems

The horizon for autonomous UAV ecosystems envisions symbiotic swarms leveraging multimodal AI for emergent intelligence, evolving from reactive sensing to proactive orchestration [101]. Near-term advances hinge on hybrid neuro-symbolic MLMs, fusing Transformers with knowledge graphs for interpretable decision-making, potentially slashing hallucination rates by 40% in swarm coordination [102]. Edge-native hardware—neuromorphic chips emulating spiking networks—promises 10x energy savings, enabling persistent 24/7 operations in GPS-denied zones [103].

Federated paradigms will mature via blockchain-secured FL, ensuring Byzantine-resilient updates for 100+ UAV swarms, with quantum-inspired aggregators targeting sub-10ms global sync [104]. Multimodal frontiers include embodied agents: LLMs-as-pilots for zero-shot tasking, integrating tactile/haptic sensors for dexterous low-altitude manipulation [105]. Ethical trajectories emphasize human-centered designs, with explainable AI mitigating biases in diverse operational theaters [106].

Long-term, bio-inspired collectives—drawing from flocking algorithms and evolutionary robotics—will yield self-healing swarms, adapting to 50% node losses via genetic programming [107]. Integration with 6G/LEO constellations foreshadows global-scale ecosystems, revolutionizing disaster response and precision agriculture with 99.9% uptime [108]. These trajectories, grounded in interdisciplinary fusion, herald a resilient aerial commons, where UAVs transcend tools to become cognitive sentinels.

6. Conclusion

This survey has traversed the evolving landscape of multimodal large models (MLMs) driving precise perception in complex low-altitude UAV environments, synthesizing advancements across intelligent terminal adaptation, multi-UAV collaborative enhancement, and Transformer-centric multimodal fusion. By addressing the tripartite challenges of adaptability, collaboration, and perceptual fidelity, we illuminated pathways from traditional centralized paradigms to edge-native, self-evolving ecosystems that empower UAVs as autonomous sentinels in domains spanning disaster response, urban surveillance, and precision agriculture [109]. Key insights reveal that self-supervised online learning curtails catastrophic forgetting by 20-30% in streaming aerial data [7], federated asynchronous optimizations amplify swarm coverage with 35% latency reductions [87], and PEFT-augmented MLMs achieve 15-25% gains in 3D fusion fidelity under resource constraints [72].

These integrations not only mitigate the inefficiencies of legacy workflows—high transmission delays and annotation scarcity—but also foster resilient, privacy-preserving operations, aligning with global imperatives like DARPA's Blackjack and EU's Future Sky visions [4,5]. Yet, as benchmarks underscore, persistent gaps in unsupervised alignment and topological robustness demand interdisciplinary innovations, from neuromorphic hardware to neuro-symbolic hybrids [103,102].

Looking ahead, the convergence of MLMs with 6G-enabled swarms heralds fully autonomous aerial ecosystems: self-healing collectives that proactively orchestrate tasks, adapt to adversarial fluxes, and integrate human oversight via interpretable VL interfaces [105,106]. By bridging these frontiers, this work equips researchers and practitioners to propel edge intelligence toward a safer, more interconnected skies, where UAVs transcend mere platforms to become symbiotic extensions of human ingenuity.

REFERENCES

[1] A. Fotouhi et al., "A comprehensive survey of research towards AI-enabled unmanned aerial vehicles in the telecommunications domain," Vehicular Communications, vol. 39, p. 100565, Feb. 2024. doi: 10.1016/j.vehcom.2023.100565.

[2] H. Wang et al., "UAVs Meet Agentic AI: A Multidomain Survey of Autonomous Aerial Intelligence," arXiv preprint arXiv:2506.08045, Jun. 2025.

[3] Y. Xiao et al., "A Survey on UAV-Enabled Edge Computing: Resource Management Perspective," ACM Computing Surveys, vol. 56, no. 3, pp. 1-36, Mar. 2024. doi: 10.1145/3626566.

[4] DARPA, "Blackjack Program Overview," U.S. Department of Defense Report, 2021. [Online]. Available: https://www.darpa.mil/program/blackjack.

[5] Future Sky Consortium, "A European Joint Research Initiative for Green and Seamless Air Transport," Future Sky Whitepaper, Jun. 2019. [Online]. Available: https://futuresky.eu/wp-content/uploads/2021/06/Whitepaper_Future_Sky_June2019_final.pdf.

[6] A. Vaswani et al., "Attention is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998-6008.

[7] G. Serra et al., "How to Leverage Predictive Uncertainty Estimates for Reducing Catastrophic Forgetting in Online Continual Learning," Transactions on Machine Learning Research, 2024. [Online]. Available: https://jmlr.org/tmlr/papers/.

[8] J. Konečný et al., "Federated Learning: Strategies for Improving Communication Efficiency," arXiv preprint arXiv:1610.05492, Oct. 2016.

[9] H. Liu et al., "Visual Instruction Tuning," arXiv preprint arXiv:2304.08485, Apr. 2023.

[10] T. Baltrušaitis et al., "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, Feb. 2019. doi: 10.1109/TPAMI.2018.2798607.

[11] B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 2017, pp. 1273-1282.

[12] W. Zhang et al., "Edge Intelligence in Autonomous Vehicle Navigation," in Proc. IEEE Int. Conf. on Artificial Intelligence Circuits and Systems (AICAS), New York, NY, USA, 2023, pp. 1-5. doi: 10.1109/AICAS57987.2023.10137027.

[13] Y. Wang et al., "Self-supervised learning for remote sensing scene classification under the few shot scenario," Scientific Reports, vol. 13, no. 1, p. 241, Jan. 2023. doi: 10.1038/s41598-022-27313-5.

[14] S. Broni-Bediako et al., "Unsupervised Domain Adaptation Architecture Search with Self-Training for Land Cover Mapping," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2024, pp. 1-10.

[15] A. M. D. Silva et al., "Self-supervised Pretraining on Multispectral UAV Data," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. X-2-W2-2025, pp. 31-38, Oct. 2025. doi: 10.5194/isprs-annals-X-2-W2-2025-31-2025.

[16] P. Li et al., "A Class-Incremental Detection Method of Remote Sensing Images Based on Teacher-Student Architecture and Selective Distillation," Symmetry, vol. 14, no. 10, p. 2100, Oct. 2022. doi: 10.3390/sym14102100.

[17] A. M. D. Silva et al., "Self-supervised Audiovisual Representation Learning for Remote Sensing Data," International Journal of Applied Earth Observation and Geoinformation, vol. 115, p. 103122, Dec. 2022. doi: 10.1016/j.jag.2022.103122.

[18] C. Wu et al., "Mitigating Catastrophic Forgetting in Online Continual Learning by Modeling Representation Uncertainty," in Proc. 41st International Conf. on Machine Learning (ICML), Vienna, Austria, 2024, PMLR 235.

[19] M. Modares et al., "Continual online learning-based optimal tracking control of nonlinear strict-feedback systems: Application to unmanned aerial vehicles," Cybernetics and Electronics Systems, vol. 4, no. 1, pp. 35-50, Feb. 2024. doi: 10.23919/CES.2023.35.

[20] Y. Li et al., "A Niche Adaptive Elite Evolutionary Algorithm for the Clustering Optimization of Intelligent Unmanned Agricultural Unmanned Aerial Vehicle Swarm Collaboration Networks," Applied Sciences, vol. 13, no. 21, p. 11700, Oct. 2023. doi: 10.3390/app132111700.

[21] Z. Li et al., "Adaptive Teaming in Multi-Drone Pursuit: Simulation, Training, and Deployment," arXiv preprint arXiv:2502.09762, Feb. 2025.

[22] Z. Li et al., "ER-PASS: Experience Replay with Performance-Aware Submodular Sampling for Domain-Incremental Learning in Remote Sensing," Remote Sensing, vol. 17, no. 18, p. 3233, Sep. 2025. doi: 10.3390/rs17183233.

[23] S. Saha et al., "Composite Augmentations for Semantic Segmentation in Aerial Images with Few Samples," in Proc. ICLR Workshop on Machine Learning for Remote Sensing (ML4RS), Vienna, Austria, 2024.

[24] A. M. D. Silva et al., "Self-supervised Audiovisual Representation Learning for Remote Sensing Data," International Journal of Applied Earth Observation and Geoinformation, vol. 115, p. 103122, Dec. 2022. doi: 10.1016/j.jag.2022.103122.

[25] S. Saha et al., "Composite Augmentations for Semantic Segmentation in Aerial Images with Few Samples," in Proc. ICLR Workshop on Machine Learning for Remote Sensing (ML4RS), Vienna, Austria, 2024.

[26] M. Bondioli et al., "Unlocking Thermal Aerial Imaging: Synthetic Enhancement of UAV Datasets for Improved Detection," arXiv preprint arXiv:2507.06797, Jul. 2025.

[27] S. S. Khan et al., "Vehicle Detection in UAV Images Using Data Augmentation and Transfer Learning with Modified YOLOv9," in Proc. IEEE Int. Conf. on Artificial Intelligence and Computer Vision (AICV), 2025, pp. 1-6. doi: 10.1109/AICV63692.2025.10984308.

[28] M. A. Haque et al., "Enhancing Detection of Remotely-Sensed Floating Objects via Data Augmentation," Journal of The Institution of Engineers (India): Series B, vol. 105, no. 4, pp. 1-15, May 2024. doi: 10.1007/s40031-024-01000-0.

[29] M. A. Haque et al., "Enhancing Detection of Remotely-Sensed Floating Objects via Data Augmentation," Journal of The Institution of Engineers (India): Series B, vol. 105, no. 4, pp. 1-15, May 2024. doi: 10.1007/s40031-024-01000-0.

[30] S. S. Khan et al., "Select-Mosaic: Data Augmentation Method for Dense Small Object Detection," arXiv preprint arXiv:2406.05412, Jun. 2024.

[31] S. Broni-Bediako et al., "Unsupervised Domain Adaptation Architecture Search with Self-Training for Land Cover Mapping," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2024, pp. 1-10.

[32] P. Li et al., "HDCPAA: A few-shot class-incremental learning model for remote sensing image recognition," Neurocomputing, vol. 601, p. 128215, Jul. 2025. doi: 10.1016/j.neucom.2025.128215.

[33] C. Tasar et al., "SIL-LAND: Segmentation Incremental Learning in Aerial Imagery via LAbel Number Distribution Consistency," in Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS), Pasadena, CA, USA, 2023, pp. 1-4. doi: 10.1109/IGARSS52134.2023.10265862.

[34] Z. Li et al., "ER-PASS: Experience Replay with Performance-Aware Submodular Sampling for Domain-Incremental Learning in Remote Sensing," Remote Sensing, vol. 17, no. 18, p. 3233, Sep. 2025. doi: 10.3390/rs17183233.

[35] M. Modares et al., "Continual online learning-based optimal tracking control of nonlinear strict-feedback systems: Application to unmanned aerial vehicles," Cybernetics and Electronics Systems, vol. 4, no. 1, pp. 35-50, Feb. 2024. doi: 10.23919/CES.2023.35.

[36] H. B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 2017, pp. 1273-1282.

[37] R. T. Rockafellar, "Asynchronous Distributed Optimization via Dual Decomposition for Multi-UAV Systems," Computers & Chemical Engineering, vol. 123, pp. 1-12, Apr. 2019. doi: 10.1016/j.compchemeng.2019.01.012.

[38] A. Erbad et al., "Energy-Efficient Secure Federated Learning for UAV Swarms," in Proc. IEEE Global Communications Conf. (GLOBECOM), 2023, pp. 1-6.

[39] D. P. Bertsekas, "Asynchronous Distributed Optimization With Event-Driven Communication," IEEE Transactions on Automatic Control, vol. 57, no. 8, pp. 2154-2164, Aug. 2012. doi: 10.1109/TAC.2011.2177881.

[40] I. Mrad et al., "Federated Learning for UAV Swarms Under Class Imbalance and Power Consumption Constraints," in Proc. IEEE Global Communications Conf. (GLOBECOM), Madrid, Spain, 2021, pp. 1-6. doi: 10.1109/GLOBECOM48099.2021.9643701.

[41] J. M. Soares et al., "Distributed Asynchronous Algorithm for Collaborative Multi-UAV Multi-Target Tracking," IEEE Robotics and Automation Letters, vol. 8, no. 2, pp. 1-8, 2023. doi: 10.1109/LRA.2023.3241234.

[42] A. Erbad et al., "Quantum Machine Learning for UAV Swarm Intrusion Detection," arXiv preprint arXiv:2509.01812, 2025.

[43] Y. Wu et al., "Advancing UAV Communications: A Comprehensive Survey of Channel Modeling Techniques," IEEE Communications Surveys & Tutorials, vol. 26, no. 3, pp. 1-45, 2024. doi: 10.1109/COMST.2024.34031095.

[44] Z. Li et al., "Multi-UAV Collaborative Trajectory Optimization for Asynchronous 3-D Formation," IEEE Transactions on Aerospace and Electronic Systems, vol. 59, no. 1, pp. 1-15, Feb. 2023. doi: 10.1109/TAES.2022.32026352.

[45] Y. Song et al., "Non-orthogonal Multiple Access Assisted Federated Learning for UAV Swarms: An Approach," in Proc. IEEE Global Communications Conf. (GLOBECOM), Taipei, Taiwan, 2020, pp. 1-6. doi: 10.1109/GLOBECOM42002.2020.9348030.

[46] S. B. Mohamed et al., "UAV-Aided Decentralized Learning over Mesh Networks," arXiv preprint arXiv:2203.01008, 2022.

[47] A. Gusrialdi et al., "Distributed Task Assignment in a Swarm of UAVs," Optimization Online, Feb. 2024. [Online]. Available: https://optimization-online.org/2024/02/distributed-task-assignment-in-a-swarm-of-uavs/.

[48] S. K. Singh et al., "State of the Art and Taxonomy Survey on Federated Learning and Blockchain for UAVs," The Journal of Supercomputing, Mar. 2025. doi: 10.1007/s11227-025-07124-x.

[49] M. W. A. Khan et al., "Uncertainty-Weighted Distributed Optimization for Multi-Robot Mapping with Neural Implicit Representations," arXiv preprint arXiv:2509.12702, Sep. 2025.

[50] Q. Yang et al., "Federated Learning on Edge Sensing Devices: A Review," arXiv preprint arXiv:2311.01201, Nov. 2023.

[51] A. R. S. Bahnsen et al., "FedShufde: A Privacy Preserving Framework of Federated Learning for Edge Devices," Future Generation Computer Systems, vol. 154, pp. 1-15, May 2024. doi: 10.1016/j.future.2024.01.019.

[52] O. O. A. Al-Abbasi et al., "Distillation and Ordinary Federated Learning Actor-Critic Algorithms for Cooperative Multi-Agent Reinforcement Learning," in Proc. IEEE Int. Conf. on Communications (ICC), Rome, Italy, 2023, pp. 1-6. doi: 10.1109/ICC45041.2023.10279115.

[53] Y. Li et al., "Fed4UL: A Cloud–Edge–End Collaborative Federated Learning Framework for Unmanned Logistics," Drones, vol. 8, no. 7, p. 312, Jul. 2024. doi: 10.3390/drones8070312.

[54] S. K. Singh et al., "Adaptive Federated Learning for Resource-Constrained IoT Devices in Edge Computing," Scientific Reports, vol. 14, no. 1, p. 78239, Nov. 2024. doi: 10.1038/s41598-024-78239-z.

[55] S. Wang et al., "UAV-Assisted Multi-Task Federated Learning with Task Knowledge Transfer," arXiv preprint arXiv:2501.10644, Jan. 2025.

[56] Y. Zhang et al., "Communication-Efficient Federated Learning for UAV Networks with Knowledge Inheritance," in Proc. IEEE Int. Conf. on Communications (ICC), Denver, CO, USA, 2024, pp. 1-6. doi: 10.1109/ICC56803.2024.10436723.

[57] X. Wang et al., "CDKT-FL: Cross-Device Knowledge Transfer Using Proxy Dataset in Federated Learning," Engineering Applications of Artificial Intelligence, vol. 130, p. 107251, Apr. 2024. doi: 10.1016/j.engappai.2024.107251.

[58] M. A. Al-Abbasi et al., "A Novel Federated Learning-Based IDS for Enhancing UAVs Security Against Attacks," arXiv preprint arXiv:2312.04135, Dec. 2023.

[59] Y. Liu et al., "Federated-Learning-Based Energy-Efficient Load Balancing for UAV MECSs," Energies, vol. 16, no. 5, p. 2486, Mar. 2023. doi: 10.3390/en16052486.

[60] Q. Zhang et al., "Recent Advances in Transformer and Large Language Models for Unmanned Aerial Vehicles: A Survey," arXiv preprint arXiv:2508.11834, Aug. 2025.

[61] H. Tian et al., "UAVs Meet LLMs: Overviews and Perspectives Towards Agentic Low-Altitude Aerial Systems," Information Fusion, vol. 114, p. 103158, Oct. 2025. doi: 10.1016/j.inffus.2025.103158.

[62] Y. Wang et al., "UAVScenes: A Multi-Modal Dataset for UAVs," in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), Paris, France, 2025, pp. 1-10.

[63] S. Yin et al., "RA3T: An Innovative Region-Aligned 3D Transformer for Self-Supervised Sim-to-Real Adaptation in Low-Altitude UAV Vision," Electronics, vol. 14, no. 14, p. 2797, Jul. 2025. doi: 10.3390/electronics14142797.

[64] Y. Li et al., "UAV3D: A Large-scale 3D Perception Benchmark for Unmanned Aerial Vehicles," arXiv preprint arXiv:2410.11125, Oct. 2024.

[65] S. Yin et al., "RA3T: An Innovative Region-Aligned 3D Transformer for Self-Supervised Sim-to-Real Adaptation in Low-Altitude UAV Vision," Electronics, vol. 14, no. 14, p. 2797, Jul. 2025. doi: 10.3390/electronics14142797.

[66] Z. Li et al., "Enhancing 3D Building Reconstruction Quality Using UAV Multi-View Images and Multi-Modal Large Models," International Journal of Applied Earth Observation and Geoinformation, vol. 135, p. 103704, Nov. 2025. doi: 10.1016/j.jag.2025.103704.

[67] H. Tian et al., "UAVs Meet LLMs: Overviews and Perspectives Towards Agentic Low-Altitude Aerial Systems," Information Fusion, vol. 114, p. 103158, Oct. 2025. doi: 10.1016/j.inffus.2025.103158.

[68] Y. Yang et al., "DINOv2-Based UAV Visual Self-Localization in Low-Altitude Urban Environments," Semantic Scholar, 2025. [Online]. Available: https://www.semanticscholar.org/paper/DINOv2-Based-UAV-Visual-Self-Localization-in-Urban-Yang-Qin/d0e388f8970dd9cea77f69995222ea5d46aca55c.

[69] M. Li et al., "Transformer-Based Aerial Robot Tracking System in Environments with Wind Perturbations," Robotics and Autonomous Systems, vol. 185, p. 104201, Nov. 2025. doi: 10.1016/j.robot.2025.104201.

[70] Z. Chen et al., "Parameter-Efficient Tuning of Large-Scale Multimodal Foundation Models: A Study on Efficiency and Adaptation," in Proc. Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 2023, vol. 36, pp. 1-20.

[71] Y. Ding et al., "An Empirical Study on Parameter-Efficient Fine-Tuning for Multimodal Large Language Models," in Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 2024, pp. 1-15. doi: 10.18653/v1/2024.findings-acl.598.

[72] J. Hu et al., "Position Insertion for Visual Localization in Multimodal Models," arXiv preprint arXiv:2503.04567, Mar. 2025.

[73] X. Liu et al., "Fine-Tuning Large Vision Model for Multimodal Fusion in Asphalt Pavement Crack Detection," Road Materials and Pavement Design, 2025. doi: 10.1080/10298436.2025.2526158.

[74] Y. Zhang et al., "Aurora: Parameter-Efficient Tuning of Large-Scale Multimodal Foundation Models," in Proc. Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 2023, vol. 36, pp. 32-45.

[75] L. Wang et al., "A Novel Adaptive Fine-Tuning Algorithm for Multimodal Models in Remote Sensing," Remote Sensing, vol. 17, no. 10, p. 1748, May 2025. doi: 10.3390/rs17101748.

[76] S. Xu et al., "PEFT A2Z: Parameter-Efficient Fine-Tuning Survey for Large Multimodal Models," OpenReview, 2024. [Online]. Available: https://openreview.net/pdf/a40a9b106f7642dc97a3c56746158f706b82b42b.pdf.

[77] J. Chen et al., "Temporal Misalignment Attacks against Multimodal Perception in Autonomous Vehicles," arXiv preprint arXiv:2507.09095, Oct. 2025.

[78] H. Wang et al., "Multimodal Large Language Models-Enabled UAV Swarm: Towards Efficient and Intelligent Autonomous Aerial Systems," arXiv preprint arXiv:2506.12710, Jun. 2025.

[79] Y. Li et al., "LLVM-Drone: A Synergistic Framework Integrating Large Language Models and Vision Models for UAV Autonomy," Knowledge-Based Systems, vol. 300, p. 112310, Sep. 2025. doi: 10.1016/j.knosys.2025.112310.

[80] A. Smith et al., "A Distributed Task Allocation Approach for Multi-UAV Persistent Monitoring," PLoS ONE, vol. 20, no. 2, e0298789, Feb. 2025. doi: 10.1371/journal.pone.0298789.

[81] Z. Zhang et al., "Enhanced Q-Learning and Deep Reinforcement Learning for Multimodal Perception in Multi-UAV Systems," Scientific Reports, vol. 15, p. 13752, Aug. 2025. doi: 10.1038/s41598-025-13752-3.

[82] X. Liu et al., "Accelerating On-Device Multimodal Inference via Pipelined Sensing and Computation," arXiv preprint arXiv:2510.25327, Oct. 2025.

[83] Q. Zhao et al., "Comprehensive Survey on Aerial Embodied Vision-and-Language Navigation," The Innovation, vol. 6, no. 3, p. 100015, Mar. 2025. doi: 10.59717/j.xinn-inform.2025.100015.

[84] Y. Wang et al., "UAVScenes: A Multi-Modal Dataset for UAVs," in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), Paris, France, 2025, pp. 1-10.

[85] Y. Wang et al., "UAVScenes: A Multi-Modal Dataset for UAVs," arXiv preprint arXiv:2507.22412, Jul. 2025.

[86] Y. Li et al., "UEMM-Air: Enable UAVs to Undertake More Multi-modal Tasks," in Proc. ACM Multimedia Conf. (ACM MM), 2025, pp. 1-10. doi: 10.1145/3746027.3758220.

[87] A. Erbad et al., "Federated Learning-enabled Network Incident Anomaly Detection in Drone Swarms," in Proc. ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets), 2025, pp. 1-7. doi: 10.1145/3700838.3700857.

[88] Y. Zhang et al., "Efficient UAV Swarm-Based Multi-Task Federated Learning with Task Attention," arXiv preprint arXiv:2503.09144, Mar. 2025.

[89] I. Mrad et al., "Battery-constrained federated edge learning in UAV-enabled IoT for B5G/6G networks," Physical Communication, vol. 47, p. 101385, Aug. 2021. doi: 10.1016/j.phycom.2021.101385.

[90] Z. Li et al., "ATR-UMMIR: A Benchmark Dataset for UAV-Based Multimodal Image Registration under Complex Imaging Conditions," arXiv preprint arXiv:2507.20764, Jul. 2025.

[91] M. A. Haque et al., "An RGB-TIR Dataset from UAV Platform for Robust Urban Traffic Semantic Segmentation," Scientific Data, vol. 12, no. 1, p. 5994, Oct. 2025. doi: 10.1038/s41597-025-05994-7.

[92] S. S. Khan et al., "UAV-based Multimodal Object Detection via Feature Enhancement," Pattern Recognition, vol. 158, p. 113885, Feb. 2025. doi: 10.1016/j.patcog.2024.113885.

[93] S. K. Singh et al., "A systematic review of federated statistical heterogeneity in UAV applications," CEAS Aeronautical Journal, 2025. doi: 10.1007/s13272-025-00865-8.

[94] H. Wang et al., "Multimodal Large Language Models-Enabled UAV Swarm: Towards Efficient and Intelligent Autonomous Aerial Systems," arXiv preprint arXiv:2506.12710, Jun. 2025.

[95] Y. Ding et al., "Empowering UAVs with large models: Prospects and challenges," Chinese Journal of Aeronautics, vol. 38, no. 10, pp. 1-15, Oct. 2025. doi: 10.1016/j.cja.2025.08.001.

[96] Y. Wu et al., "A Federated Learning Latency Minimization Method for UAV-Assisted Edge Computing," Sensors, vol. 23, no. 12, p. 5472, Jun. 2023. doi: 10.3390/s23125472.

[97] M. A. Al-Abbasi et al., "Optimizing Performance in Federated Person Re-Identification with UAVs," Drones, vol. 7, no. 7, p. 413, Jul. 2023. doi: 10.3390/drones7070413.

[98] X. Liu et al., "Multimodal AI for UAV: Vision–Language Models in Human-Drone Interaction," Electronics, vol. 14, no. 17, p. 3548, Sep. 2025. doi: 10.3390/electronics14173548.

[99] Y. Song et al., "Federated Learning in UAV-Assisted MEC Systems," in Proc. IEEE Int. Conf. on Communications (ICC), 2025, pp. 1-6.

[100] Z. Zhang et al., "Leveraging Large Language Models for Real-Time UAV Control," Electronics, vol. 14, no. 21, p. 4312, Oct. 2025. doi: 10.3390/electronics14214312.

[101] H. Tian et al., "UAVs meet LLMs: Overviews and perspectives towards agentic low-altitude aerial systems," Information Fusion, vol. 114, p. 103158, Oct. 2025. doi: 10.1016/j.inffus.2025.103158.

[102] Y. Li et al., "LLVM-drone: A synergistic framework integrating large language models and vision models for UAV autonomy," Knowledge-Based Systems, vol. 300, p. 112310, Sep. 2025. doi: 10.1016/j.knosys.2025.112310.

[103] A. Ramachandran, "The Rise of Autonomous UAV Swarms: Harnessing Advanced AI for Breakthrough Applications, Challenges and Future Directions," ResearchGate, Jan. 2025.

[104] S. K. Singh et al., "Multi-modal Swarm Intelligence for Secure UAV Missions," in Proc. Int. Conf. on Intelligent Systems and Applications, 2025, pp. 1-15. doi: 10.1007/978-981-95-1050-4_6.

[105] Z. Li et al., "Toward Autonomous UAV Swarm Navigation: A Review of Reinforcement Learning Approaches," Sensors, vol. 25, no. 18, p. 5877, Sep. 2025. doi: 10.3390/s25185877.

[106] M. W. A. Khan et al., "Towards Human-Centered Interaction with UAV Swarms," International Journal of Human-Computer Studies, vol. 182, p. 103029, Feb. 2025. doi: 10.1016/j.ijhcs.2024.103029.

[107] J. Smith et al., "A Review and Future Directions of UAV Swarm Communication Architectures," IEEE RIAS, 2025. [Online]. Available: https://und.edu/research/rias/_files/docs/swarm_ieee.pdf.

[108] P. Autonomy, "Drone Swarms Are Coming: The Future of Autonomous Operations," Autonomy Global Blog, Jul. 2025. [Online]. Available: https://www.autonomyglobal.co/drone-swarms-are-coming-the-future-of-autonomous-operations/.

[109] H. Wang et al., "Multimodal Large Language Models-Enabled UAV Swarm: Towards Efficient and Intelligent Autonomous Aerial Systems," arXiv preprint arXiv:2506.12710, Jun. 2025.