DOI: https://doi.org/10.26565/2304-6201-2025-68-05

УДК (UDC) 004.85:629.7.05

Lupandin Antonii *PhD Student, Department of Computer Systems and Robotics;*
V. N. Karazin Kharkiv National University, 4 Svobody Square, Kharkiv, Ukraine, 61022; e-mail: antonii.lupandin@student.karazin.ua*;*
https://orcid.org/0009-0002-7591-5152

Moroz Olha *PhD in Computer Science; Associate Professor, Department of Computer Systems and Robotics;*
V. N. Karazin Kharkiv National University, 4 Svobody Square, Kharkiv, Ukraine, 61022; e-mail: o.moroz@karazin.ua;
https://orcid.org/0000-0002-4920-4093

Analysis of Modern Neural Network Methods for Visual Information Processing in High-Speed UAV Navigation Systems

Relevance. The rapid evolution of Unmanned Aerial Vehicles (UAVs) from remotely piloted systems to fully autonomous high-speed aerial robots has intensified the demand for advanced onboard perception and navigation methods. This need is particularly acute in scenarios where computational latency, sensor noise, and environmental complexity undermine the reliability of classical computer-vision pipelines. Despite recent progress in deep learning, the existing approaches to visual information processing—especially CNN-based detectors, Transformer-based semantic models, and learning-enhanced SLAM modules—remain fragmented and insufficiently adapted to the strict Size, Weight and Power (SWaP) constraints of embedded platforms such as the NVIDIA Jetson series. This motivates a comprehensive analysis of modern neural architectures suitable for real-time, high-velocity UAV operations.
Purpose. The purpose of this study is to analyze state-of-the-art neural network methods for secondary visual processing in UAV navigation systems, compare the applicability of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), evaluate their integration into SLAM pipelines, and determine the requirements for hybrid architectures capable of supporting fully autonomous, high-speed flight.
Methods. The research employs a comparative analysis of recent deep-learning approaches, including CNN-based detectors (YOLO family), Transformer-based visual models, deep-learning–enhanced SLAM components, and Deep Reinforcement Learning (DRL) control policies. Evaluation criteria include latency, semantic robustness, dynamic-scene handling, edge-hardware compatibility, quantization performance, pruning potential, and TensorRT optimization efficiency on NVIDIA Jetson devices.
Results. The study establishes that CNNs provide superior real-time performance and remain indispensable for high-frequency reflexive perception, while Vision Transformers offer stronger global context reasoning and robustness to occlusion but suffer from significant computational overhead on embedded GPUs. Deep-learning-based SLAM methods improve feature stability and dynamic-object rejection but require careful integration to maintain real-time constraints. Hardware analysis reveals that quantization, pruning, and TensorRT acceleration are critical for deploying deep models on Jetson-class platforms, although ViTs exhibit limited INT8 quantization tolerance. Based on these findings, the work formulates a conceptual hybrid architecture that combines CNN-driven reflexive processing with Transformer-driven cognitive reasoning.
Conclusions. The results confirm the necessity of developing hybrid neuro-architectures that integrate the speed and hardware efficiency of CNNs with the semantic depth of Transformer-based models. Such architectures represent a promising pathway toward reliable, fully autonomous high-speed UAV navigation. The proposed design principles emphasize hierarchical control, asynchronous perception loops, and hardware-aware optimization as key enablers for next-generation aerial robotic systems.

Keywords: *UAV, high-speed navigation, CNN, Vision Transformer, SLAM, Reinforcement Learning, edge computing, Jetson, TensorRT, quantization, pruning, hybrid architectures.*

How to quote: A. Lupandin and O. Moroz, “Analysis of Modern Neural Network Methods for Visual Information Processing in UAV Navigation Systems” *Bulletin of V. N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 68, pp. 53-61, 2025. https://doi.org/10.26565/2304-6201-2025-68-05

Як цитувати: Lupandin A., Moroz O. Analysis of Modern Neural Network Methods for Visual Information Processing in UAV Navigation Systems. Вісник Харківського національного університету імені В. Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління. 2025. вип. 68. С.53-61.
https://doi.org/10.26565/2304-6201-2025-68-05

1. Introduction

The operational envelope of Unmanned Aerial Vehicles (UAVs) has expanded dramatically over the last decade. Modern applications—ranging from rapid courier delivery and large-scale environmental monitoring to search-and-rescue operations in disaster zones and dynamic defense maneuvers—now demand a level of autonomy that far exceeds simple GPS waypoint following. As the flight speed of these robotic platforms increases, driven by improvements in rotor efficiency and compact jet-propulsion systems, the allowable margin for error in navigation and obstacle avoidance shrinks to millisecond-level reaction times. This shift has reframed UAV navigation not merely as a geometric estimation task but as a high-velocity perception-and-decision problem [1].

At the heart of this challenge lies the so-called "perception–action loop." Traditional navigation systems—whether Correlation-Extreme Navigation Systems (CENS) or classical geometric computer-vision pipelines—rely on matching current sensor data to pre-loaded geospatial maps or on extracting handcrafted features such as SIFT or ORB to estimate motion. While mathematically rigorous in controlled or slowly changing environments, these methods exhibit significant brittleness when applied to real-world high-speed flight. They degrade sharply under extreme lighting variation, suffer from the motion blur induced by rapid maneuvers, and fail to generalize when operating in unstructured, semantically complex terrains [2, 3].

To address this gap, the field increasingly turns to the concept of secondary processing—the transformation of raw sensor data (primary information) into a higher-level semantic or structural representation (secondary information). In autonomous flight, this means not simply observing pixels but interpreting them: estimating the UAV's pose relative to a target or map, detecting and tracking obstacles, recognizing terrain types, and inferring traversability or risk. In essence, secondary processing forms the decision-critical interface between visual perception and control policy [4].

Deep Learning (DL) has become the dominant methodology enabling this transformation. By leveraging large-scale, diverse datasets, neural networks learn feature representations that are robust to noise, illumination variation, high-speed distortions, and other real-world degradations that hinder classical algorithms. Yet, the deployment of such computationally intensive models on Size, Weight, and Power (SWaP) constrained aerial platforms creates a multi-objective optimization problem: accuracy is paramount for safety, while inference latency must remain minimal to maintain stability during aggressive flight.

Against this backdrop, this review synthesizes academic literature published between 2023 and 2025 to map the emerging landscape of neural-network-based UAV navigation. We examine evolving architectural paradigms, including the competition and convergence between CNNs and Vision Transformers as well as the rise of hybrid spatial–temporal models. We also trace how deep learning reshapes Visual SLAM and Visual Odometry, enabling more resilient navigation under sensor degradation. Furthermore, we discuss the rapid shift toward end-to-end Deep Reinforcement Learning (DRL), where perception and control are unified within a single policy network. Finally, we assess the practical realities of deploying these algorithms on embedded edge processors, with particular attention to the NVIDIA Jetson family and comparable low-power accelerators, which increasingly determine what forms of neural autonomy are feasible in high-speed UAV operations [3, 4].

2. Comparative Analysis of Neural Architectures: CNN vs. ViT

The choice of neural architecture is arguably the most consequential design decision in a UAV perception system. It determines not only the theoretical upper bound of visual understanding but also the practical inference dynamics of the model when deployed under stringent Size, Weight, and Power (SWaP) constraints. Today, the field is shaped by the competition—and increasingly, the convergence—of two methodological paradigms: the Convolutional Neural Network (CNN) and the Vision Transformer (ViT).

For more than a decade, CNNs have served as the backbone of visual perception. Their core design principle is the convolution operation, which applies learnable filters over local image regions. This architectural choice embeds strong inductive biases, primarily locality (nearby pixels are correlated) and translation invariance (features retain meaning regardless of spatial position). These properties make CNNs particularly effective in structured environments and for tasks where objects possess distinctive local patterns.

Within UAV-specific applications, the YOLO (You Only Look Once) family of single-stage detectors has become especially prevalent due to its exceptional speed–accuracy trade-off. Unlike two-stage detectors (e.g., Faster R-CNN), YOLO predicts bounding boxes and class probabilities in a single forward pass, making it highly compatible with the real-time requirements of aerial robotics. The evolution from YOLOv5 to YOLOv8—and more experimental iterations such as YOLOv10/11—has focused on improving the "backbone" and "neck" components. For instance, YOLOv8 introduced C2f modules that enhance gradient propagation and support deeper architectures without the risk of vanishing gradients.

Performance on UAV benchmarks such as VisDrone further underscores CNN effectiveness. Their inductive biases align well with objects that present stable local features (e.g., cars, rooftops, pedestrians), enabling robust detection even under moderate variations in viewpoint or scale. In terms of latency—a critical factor for high-speed flight—lightweight CNNs (e.g., YOLOv8-Nano) achieve real-time performance exceeding 30–60 FPS on embedded hardware like the Jetson Orin Nano. For drones operating at speeds where even a 100 ms delay could result in catastrophic failure, this level of responsiveness is indispensable [5-6].

However, CNNs exhibit structural limitations. Their reliance on local convolutions restricts the ability to capture long-range dependencies early in the network. This becomes problematic in high-altitude or ultra-dynamic UAV imagery, where targets may occupy only a handful of pixels or derive their semantic identity from global scene context (e.g., distinguishing a drone from a bird based on trajectory or surrounding background). Although deeper layers eventually accumulate broader receptive fields, this comes at the cost of spatial resolution due to pooling and downsampling [1].

These shortcomings have spurred interest in Vision Transformers. Originating in natural language processing, Transformers treat data as sequences of tokens. Vision Transformers (ViTs) translate this concept to image processing by dividing the input into patches and applying self-attention across them. This design grants ViTs an immediate global receptive field, enabling every patch to attend to every other patch from the first layer onward.

Recent research – particularly Zhang (2023) – demonstrates the strong potential of ViTs in UAV contexts involving air-to-air detection. In tasks requiring robust recognition of other drones, ViT-based architectures significantly exceeded CNN performance. Notably, Zhang reports that baseline ViTs were approximately 4.6× more robust than comparable CNNs under challenging environmental interference [7]. The ability of self-attention to model global dependencies allows ViTs to suppress background clutter such as clouds or terrain textures, improving detection rates in highly variable scenes. Their resilience also extends to occlusion: because ViTs aggregate information globally, they maintain accuracy even when key object features are partially hidden—a common scenario in urban or forested flight corridors.

Despite their representational strength, ViTs impose substantial computational burdens. Self-attention scales quad

ratically, O (N^2), with the number of tokens. For high-resolution aerial imagery –necessary for detecting tiny objects—this leads to excessive memory usage and increased inference times. On edge hardware, pure ViTs often fall below real-time thresholds, achieving only 5–10 FPS on the Jetson Orin Nano [1]. Combined with their weaker inductive biases, which require larger datasets to avoid overfitting [8], pure ViTs remain difficult to deploy in fast-response UAV control loops [7].

These complementary strengths and weaknesses have motivated a shift toward hybrid CNN–Transformer architectures, which seek to integrate the best of both paradigms. In such models, a CNN backbone first extracts low-level spatial features—reducing dimensionality and preserving fine-grained local detail—after which a Transformer module captures global context and long-range dependencies.

Several notable hybrid architectures illustrate this trend:

- LandNet, proposed for 6-DoF camera relocalization, combines spatial convolutions with temporal attention via a Feature Interaction Block. This design enables UAVs to reason about both the geometric structure of landmarks and their configuration across time [9].
- RepEfficientViT extends this idea to edge scenarios by integrating re-parameterized convolutional blocks (RepMBConv) with lightweight attention. In applications such as agricultural weed recognition—an analogue for detailed ground surveillance—RepEfficientViT surpassed both pure CNNs and ViTs while maintaining an inference latency of ~25 ms on a CPU, demonstrating excellent real-time viability [8].
- HCTD (Hybrid CNN–Transformer for Detection) explicitly targets UAV imagery by using convolutions to preserve fine-scale patterns of small objects and self-attention to disentangle complex background clutter characteristic of aerial viewpoints [10].

Across these advances, a clear consensus is emerging: future UAV perception systems will rely not on a single architectural paradigm but on hybrid models that balance local feature extraction, global semantic reasoning, and strict real-time operational constraints.

Table 1 Architectural Trade-offs for UAV Perception
Табл. 1 Архітектурні компроміси у системах сприйняття ПМР

Feature	CNN (e.g., YOLOv8)	Vision Transformer (ViT)	Hybrid (e.g., MobileViT / RepEfficientViT)
Inductive Bias	Strong (Locality/Translation)	Weak (Global Attention)	Balanced
Receptive Field	Local (expands with depth)	Global (all layers)	Local + Global
Small Object Detection	Moderate (loses detail in pooling)	High (preserves context)	High
Inference Latency	Low (< 20ms)	High (> 100ms on edge)	Moderate (20-40ms)
Data Efficiency	High	Low (needs massive data)	Moderate
Edge Optimization	Excellent (TensorRT mature)	Poor (Quantization issues)	Improving

3. Integration of Neural Networks into SLAM and Navigation

Operating in GNSS-denied environments—such as dense urban canyons, forests, tunnels, or areas affected by intentional jamming—requires a UAV to rely on Simultaneous Localization and Mapping (SLAM). Classical Visual SLAM (VSLAM) pipelines, including ORB-SLAM and its derivatives, depend heavily on the detection and tracking of static geometric features. However, high-speed flight introduces severe motion blur, rapid viewpoint transitions, and unpredictable lighting variations, all of which degrade the reliability of handcrafted feature descriptors [1, 11]. As a result, deep learning has begun to permeate SLAM pipelines, not as a wholesale replacement of their mathematically principled foundations, but as a targeted enhancement to the modules most vulnerable to failure.

A major point of fragility in classical SLAM is the feature extraction front-end. Handcrafted features such as ORB, FAST, or BRISK are highly sensitive to illumination changes, blur, and extreme perspective shifts. Neural feature extractors—most prominently SuperPoint and D2-Net—address these weaknesses by learning interest points that remain stable across diverse lighting, scale, and trajectory conditions. When integrated into the SLAM front-end, these extractors significantly improve feature persistence, allowing the UAV to maintain tracking even during aggressive roll, pitch, or acceleration events that would cause classical pipelines to lose localization or trigger a full-system reinitialization [1].

Another foundational assumption of classical SLAM is that the environment is static. In real low-altitude UAV operations, however, the scene is populated with cars, pedestrians, cyclists, and even other drones. These dynamic outliers introduce erroneous correspondences that can corrupt the optimization process, degrade the quality of the map, and induce long-term drift.

To mitigate this, modern SLAM systems increasingly incorporate neural networks for semantic filtering:

- Semantic Masking: A neural segmentation network predicts pixel-level semantic classes. Features belonging to “movable” or “dynamic” categories—vehicles, humans, animals—are masked and excluded from the pose estimation and map update stages.
- Case Study (Luo et al., 2024): In a recent Multi-Sensor Fusion Dynamic Odometry study, researchers integrated a lightweight neural network into the FAST-LIVO framework to remove dynamic elements before fusing visual data with LiDAR and IMU. This approach substantially reduced trajectory error in dense urban scenes. The network functions as a semantic gatekeeper, ensuring that the SLAM backend optimizes only against reliable, static landmarks [12].

Beyond feature extraction and dynamic filtering, deep learning is increasingly used to enhance sensor fusion—a critical capability for high-speed navigation. Relying on a single modality (e.g., a camera) is risky: a UAV may be blinded by sun glare, fail in low-light conditions, or encounter textureless surfaces such as glass or snow. Classical fusion pipelines, typically based on Kalman or factor graph frameworks, rely on fixed covariance matrices to weight sensor contributions. However, these static weights cannot adapt to sudden environmental changes.

Neural networks enable adaptive, context-aware fusion:

- Mechanism: Instead of relying on hard-coded covariances, a small neural module evaluates the quality of each sensor stream—Visual, LiDAR, IMU—and learns dynamic attention weights. When the visual feed is saturated or motion-blurred, the system automatically up-weights LiDAR or inertial information; when LiDAR becomes unreliable (rain, fog, or low reflectivity surfaces), the visual modality regains prominence.
- Impact: This adaptive weighting is crucial for robust state estimation in environments with rapidly changing weather, illumination, and flight dynamics. Neural fusion networks allow UAVs to maintain stable pose estimation even when individual sensors become temporarily unreliable.

Taken together, these advances demonstrate a clear shift in the design of SLAM systems for UAVs: the classical geometric foundations remain intact, but deep learning is increasingly embedded into the perceptual front-ends and fusion layers, transforming SLAM into a hybrid analytical–learning framework that is far more resilient to the operational realities of high-speed flight.

4. Autonomous Navigation via Deep Reinforcement Learning

While SLAM provides the fundamental answer to “Where am I?”, autonomous navigation ultimately depends on the complementary question “How should I move?”. Classical control pipelines—typically built around PID regulators, Linear Quadratic Controllers, or Model Predictive Control (MPC)—treat perception, planning, and actuation as separate sequential modules. Although this modularity offers interpretability and stability, it becomes increasingly restrictive in high-speed scenarios, where UAVs must perform aggressive maneuvers, react to unpredictable obstacles, and adapt to rapidly changing aerodynamics.

Deep Reinforcement Learning (RL) introduces a fundamentally different paradigm: an end-to-end control policy in which a neural network directly maps raw sensor observations (images, depth, inertial states, or fused features) to low-level control commands. Through millions of simulated interactions, the agent optimizes a reward function and gradually learns a policy capable of executing highly non-linear, reflexive, and adaptive behaviors that are extremely difficult—or sometimes impossible—to encode explicitly using model-based controllers.

A compelling demonstration of this capability is presented in Sheng et al. (2024), who investigated UAV motion planning in densely populated, highly dynamic aerial environments. Traditional discrete planners such as A* or sampling-based methods like RRT struggle to replan at sub-50 ms intervals, especially when obstacles move rapidly or unpredictably [1]. RL circumvents this bottleneck by learning a reactive policy that implicitly encodes collision avoidance strategies.

Sheng’s methodology employed an Actor–Critic framework, where the policy network (Actor) generated control commands, and a value network (Critic) estimated long-term expected reward. A carefully engineered reward function balanced three objectives: (1) safety, encouraging large separation from dynamic obstacles; (2) efficiency, rewarding velocity toward the goal; (3) smoothness, penalizing jerk to ensure stable flight dynamics.

During training, the RL agent developed behaviors that resembled reflexes rather than classical planned trajectories. It learned to weave between obstacles, “lead” moving targets, and anticipate collisions before they became imminent. Empirical evaluations showed that the RL approach outperformed Artificial Potential Field (APF) methods in both success rate and mean trajectory quality, especially in densely occupied airspaces.

Beyond obstacle avoidance, RL has demonstrated remarkable potential in controlling systems traditionally considered too chaotic for conventional controllers. One prominent example is slung-load UAV navigation, where a suspended payload introduces non-linear oscillations that couple back into the UAV’s translational dynamics.

In the study by Mohiuddin et al. (2025), RL was applied to this notoriously difficult control problem. The policy learned to regulate thrust in a way that not only stabilized the drone itself but also actively dampened payload oscillations. This holistic, end-to-end learning strategy surpassed hierarchical control approaches, yielding smoother trajectories, improved energy efficiency, and faster point-to-point transport times [14]. The results highlight a broader trend: RL policies are capable of discovering coordinated strategies that operate simultaneously across multiple dynamic subsystems.

Despite its promise, RL faces a major deployment challenge: the Sim-to-Real (S2R) gap. High-speed UAVs cannot be trained directly in the real world because crashes are expensive and dangerous, while physics engines in simulators (AirSim, Gazebo, Isaac Gym) inevitably introduce discrepancies in aerodynamics, turbulence, sensor noise, and delay.

To mitigate this, modern workflows employ domain randomization, which exposes the agent to a wide distribution of simulated conditions during training. Parameters such as mass, inertia, drag coefficients, wind fields, lighting, motion blur, sensor latency, and texture appearance are randomized within predefined distributions. This forces the policy to learn strategies that generalize across variations, making it robust to the inevitable imperfections of real-world environments [1].

Additionally, researchers increasingly integrate complementary techniques to strengthen real-world deployability:

- Safe RL frameworks impose constraints during learning, limiting the agent’s exploration space to avoid unsafe states while still improving the policy.
- Curriculum learning gradually increases environment complexity—from simple static obstacles to densely dynamic multi-agent scenarios—allowing smoother and more stable convergence.
- Partially Observable RL (PO-MDP) formulations incorporate recurrent networks (LSTMs or Transformers) to compensate for intermittent perception failures, such as temporary camera blindness.
- Multi-Agent RL (MARL) enables cooperative behaviors in UAV swarms, such as coordinated formation flight or distributed obstacle avoidance.

Together, these approaches illustrate a maturing ecosystem around RL for UAVs: the field is transitioning from isolated demonstrations toward practical, reliable deployment of neural policies capable of handling the extreme dynamics of high-speed aerial robots.

5. Hardware Implementation: Edge Computing on NVIDIA Jetson

The advances of Vision Transformers, hybrid architectures, and end-to-end RL policies only become operational when these models can run onboard the UAV. High-speed aerial robots cannot rely on cloud inference due to latency and connectivity and must obey strict limits on power, thermal budget, and memory. In this context, the NVIDIA Jetson family (Nano, Xavier NX, Orin Nano, Orin NX, AGX Orin) has effectively become the standard edge platform for neural perception.

Latency is the key safety constraint. A drone moving at 20 m/s (72 km/h) covers two meters during a 100 ms inference delay—essentially flying “blind” between frames. For high-speed obstacle avoidance, the latency budget is typically under 30 ms, requiring 30–50 FPS depending on the perception stack. These limits directly determine which architectures are viable in real missions.

A comparative study by Meimetis et al. (2025) rigorously evaluates detection models on Jetson-class devices for UAV swarms, clearly contrasting convolutional and attention-based pipelines [15]. Lightweight CNN detectors such as YOLOv8-Nano and YOLOv5-Small consistently achieve >30 FPS on Jetson Orin Nano in FP16, making them suitable for the “reflexive” perception layer responsible for immediate collision avoidance. Transformer-based detectors—DETR variants and heavy ViT backbones—often drop below 10 FPS even on Orin-class hardware because attention operations saturate DRAM bandwidth long before compute throughput becomes the bottleneck [5, 15].

Since raw model performance rarely satisfies onboard real-time constraints, designers increasingly rely on hardware-aware optimizations. NVIDIA TensorRT is central here: it compiles networks into optimized execution graphs via kernel fusion, constant folding, and aggressive memory reuse. Converting a PyTorch-trained YOLO model to a TensorRT engine commonly yields 2–3× speedups on Jetson by reducing global memory traffic and fully exploiting Tensor Cores.

Precision reduction is another widely used technique. Moving from FP32 to mixed-precision FP16 roughly halves memory footprint and often doubles throughput with minimal accuracy loss, making FP16 the practical default for Orin-based UAVs. Pushing to INT8, however, reveals architectural asymmetries. CNNs with relatively well-behaved activations quantize reliably and benefit from substantial acceleration, whereas ViT-style models contain non-Gaussian activations, high-dynamic-range attention logits, and layer normalization, which complicate quantization. Empirical results and developer reports show that INT8 ViT inference on some Jetson platforms yields little or no gain over FP16 due to quantization/dequantization overheads and the limited availability of highly optimized INT8 attention kernels.

To further reduce model size and energy consumption, structured pruning removes entire convolutional channels, attention heads, or MLP sub-blocks rather than individual weights. For embedded systems this provides two key benefits: (1) the pruned model fits into limited high-speed on-chip memory, reducing DRAM access, and (2) it lowers thermal load, delaying or preventing throttling during sustained high-speed flight. Properly tuned pruning thus delivers real throughput gains without major architectural redesign.

The broader hardware–algorithm co-design trend also includes neural architecture search (NAS) tailored to Jetson devices. These search procedures explore backbones optimized for Tensor Cores, memory locality, and kernel fusion patterns, producing architectures that outperform manually designed counterparts under strict latency constraints. This is particularly promising for hybrid CNN–Transformer models, whose components can be co-optimized for multi-stage execution.

Overall, these optimization strategies highlight a core principle: onboard perception depends not only on neural architecture, but on its interaction with edge hardware. Real-time UAV autonomy emerges from balancing model complexity, parallelism, memory bandwidth, thermal behavior, and numerical precision. As UAVs continue to demand higher speeds and richer perception, hardware-aware optimization will remain central to translating deep learning advances into practical embedded autonomy [1, 5, 8, 15].

Conclusion

The synthesis of contemporary neural methods for UAV navigation reveals a landscape that is both technically mature and inherently heterogeneous. Convolutional Neural Networks continue to dominate real-time perception tasks, providing the latency and stability required for reflexive control during high-speed flight. Vision Transformers, although significantly more powerful in terms of global semantic reasoning and occlusion robustness, remain constrained by their computational overhead on embedded platforms. Reinforcement Learning offers a compelling route toward end-to-end control of nonlinear UAV dynamics, yet it still struggles with generalization and reliable sim-to-real transfer, particularly in safety-critical conditions.

Taken together, these findings indicate that high-speed UAV autonomy cannot rely on any single architectural family. Instead, the most promising trajectory lies in a deliberately hybridized design, where multiple neural paradigms operate synergistically within a hierarchical control framework. In such a system, the first computational tier is responsible for immediate, high-frequency reactions. This reflexive pathway is implemented using aggressively optimized CNNs—such as pruned variants of YOLOv8-Nano—or lightweight hybrid convolution–attention backbones that execute efficiently in FP16 format under TensorRT on Jetson-class hardware. By leveraging the strong inductive biases of convolutions, this pathway can rapidly interpret local visual cues, optical flow patterns, and short-horizon geometric structures, ensuring stable flight and instantaneous obstacle avoidance even under tight latency budgets.

Above this, a slower cognitive pathway provides global situational understanding, long-horizon reasoning, and accumulated map correction. This layer can be realized through Vision Transformers, hybrid attention models, or deep-learning–enhanced SLAM frameworks, all of which excel at capturing wide spatial context and semantic relationships that CNNs inherently struggle to model. Operating at lower frequency (typically 5–10 Hz), it refines the UAV's understanding of its environment, identifies distant targets, updates global maps, and compensates for the drift or perceptual limitations of the reflexive layer. By separating rapid sensorimotor reflexes from slower cognitive inference—an organization that mirrors biological neural systems—this architectural division offers a principled route to reliable, high-speed autonomy on SWaP-constrained embedded hardware.

Looking ahead, the feasibility of fully autonomous, high-velocity UAVs will depend on how effectively these two pathways can be integrated. Future research must prioritize asynchronous coordination mechanisms that allow low-latency reflexive responses to coexist with slower, more deliberative cognitive reasoning without mutual interference. Equally important is the development of hardware-aware Neural Architecture Search (NAS) methods capable of automatically designing hybrid models that meet strict latency, memory, and energy budgets of edge platforms. Improvements in sim-to-real transfer, robust multi-modal fusion, and safety-oriented reinforcement learning will further strengthen the reliability of such systems.

In conclusion, the path forward lies in embracing hybrid, hierarchical, and hardware-optimized approaches rather than searching for a single universal architecture. By combining the speed of CNN-based reflexes, the contextual capacity of Transformer-driven cognition, and the adaptability of reinforcement learning, next-generation UAVs can achieve the levels of autonomy, robustness, and situational awareness required for safe high-speed flight in complex environments.

REFERENCES

1. Sheng, Y., Liu, H., Li, J., & Han, Q. (2024). UAV autonomous navigation based on deep reinforcement learning in highly dynamic and high-density environments. Drones, 8(9), 516. https://doi.org/10.3390/drones8090516
2. Scherbinin, V. V., Khusainov, N. S., & Kravchenko, P. P. (2014). Combined correlation-extremal navigation system to identify AV location by terrain relief and landscape objects with the use of the stereo photogrammetry method. *Middle-East Journal of Scientific Research, 19*(4), 479–486. https://doi.org/10.5829/idosi.mejsr.2014.19.4.13693
3. Mukhina, M. P., & Seden, I. V. (2014). Analysis of modern correlation extreme navigation systems. Electronics and Control Systems, 1(39), 95–101. https://doi.org/10.18372/1990-5548.39.7343
4. Sotnikov, A., Tiurina, V., Petrov, K., Lukyanova, V., Lanovyy, O., Onishchenko, Y., Gnusov, Y., Petrov, S., Boichenko, O., & Breus, P. (2024). Using the set of informative features of a binding object to construct a decision function by the system of technical vision when localizing mobile robots. *Eastern-European Journal of Enterprise Technologies, 3*(9(129)), 60–69. https://doi.org/10.15587/1729-4061.2024.303989
5. Seeed Studio. (2023, March 30). *YOLOv8 performance benchmarks on NVIDIA Jetson devices.* Seeed Studio Blog. https://www.seeedstudio.com/blog/2023/03/30/yolov8-performance-benchmarks-on-nvidia-jetson-devices/
6. D. Du et al. (2019). VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2019) (pp. 213-226). IEEE. https://doi.org/10.1109/ICCVW.2019.00030
7. Zhang, J. (2023). Towards a high-performance object detector: Insights from drone detection using ViT and CNN-based deep learning models. In *Proceedings of the 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE)* (pp. 141–147). IEEE. https://doi.org/10.1109/ICSECE58870.2023.10263514
8. Liu, T., Wang, Y., Yang, C., Zhang, Y., & Zhang, W. (2025). A lightweight hybrid CNN-ViT network for weed recognition in paddy fields. *Mathematics, 13*(17), 2899. https://doi.org/10.3390/math13172899
9. Shen, S., Yu, G., Zhang, L., Yan, Y., & Zhai, Z. (2025). LandNet: Combine CNN and Transformer to Learn Absolute Camera Pose for the Fixed-Wing Aircraft Approach and Landing. Remote Sensing, 17(4), 653. https://doi.org/10.3390/rs17040653
10. Xue, H., Tang, Z., Xia, Y., Wang, L., & Li, L. (2025). HCTD: A CNN-transformer hybrid for precise object detection in UAV aerial imagery. *Computer Vision and Image Understanding, 259*, 104409. https://doi.org/10.1016/j.cviu.2025.104409
11. Favorskaya, M. N. (2023). Deep learning for visual SLAM: The state-of-the-art and future trends. *Electronics, 12*(9), 2006. https://doi.org/10.3390/electronics12092006
12. Luo, L., Peng, F., & Dong, L. (2024). Improved multi-sensor fusion dynamic odometry based on neural networks. *Sensors, 24*(19), 6193. https://doi.org/10.3390/s24196193
13. Zhu, P., Wen, L., Du, D., Bian, X., Hu, Q., Ling, H., & et al. (2022). Detection and Tracking Meet Drones Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 7380–7399. https://doi.org/10.1109/TPAMI.2021.3119563
14. Mohiuddin, M.B., Boiko, I., Tran, V.P. et al. Reinforcement learning for end-to-end UAV slung-load navigation and obstacle avoidance. Sci Rep 15, 34621 (2025). https://doi.org/10.1038/s41598-025-18220-6
15. Meimetis, D., Daramouskas, I., Patrinopoulou, N., Lappas, V., & Kostopoulos, V. (2025). Comparative analysis of object detection models for edge devices in UAV swarms. Machines, 13(8), 684. https://doi.org/10.3390/machines13080684

Лупандін
Антоній Володимирович

аспірант кафедри комп'ютерних систем та робототехніки;
Харківський національний університет імені В.Н. Каразіна,
майдан Свободи, 4, Харків, 61022, Україна
e-mail: antonii.lupandin@student.karazin.ua;
https://orcid.org/0009-0002-7591-5152

Мороз
Ольга Юріївна

PhD комп'ютерних наук; доцент кафедри комп'ютерних систем та робототехніки;
Харківський національний університет імені В.Н. Каразіна,
майдан Свободи, 4, Харків, 61022, Україна
e-mail: o.moroz@karazin.ua;
https://orcid.org/0000-0002-4920-4093

Аналіз сучасних нейромережевих методів обробки візуальної інформації в системах навігації високошвидкісних ПМР

Актуальність. Стрімка еволюція безпілотних літальних апаратів (БПЛА) — від дистанційно керованих платформ до повністю автономних високошвидкісних повітряних мобільних роботів — зумовлює підвищений запит на вдосконалені методи бортового сприйняття та навігації. Потреба в таких підходах особливо відчутна в умовах, коли обчислювальна затримка, шум сенсорів та складність навколишнього середовища підривають надійність класичних комп'ютерно-зорових систем. Попри суттєвий прогрес у сфері глибокого навчання, наявні підходи до обробки візуальної інформації — зокрема CNN-детектори, семантичні моделі на основі Transformer-архітектур та SLAM-модулі з елементами навчання — залишаються фрагментованими та недостатньо адаптованими до жорстких обмежень за розміром, вагою та енергоспоживанням (SWaP), властивих вбудованим платформам на кшталт NVIDIA Jetson. Це актуалізує потребу в системному огляді сучасних нейроархітектур, придатних для роботи в режимі реального часу на високошвидкісних ПМР.

Мета. Метою дослідження є аналіз сучасних нейронних методів вторинної обробки візуальної інформації в навігаційних системах ПМР, порівняння сфер застосування Convolutional Neural Networks (CNNs) і Vision Transformers (ViTs), оцінювання їх інтеграції у SLAM-підсистеми та визначення вимог до гібридних архітектур, здатних забезпечити повністю автономний високошвидкісний політ.

Методи. У роботі використано порівняльний аналіз сучасних підходів глибокого навчання, включаючи CNN-детектори сімейства YOLO, візуальні моделі на основі Transformer-архітектур, SLAM-компоненти з нейронними модулями та методи управління на основі Deep Reinforcement Learning (DRL). Оцінювання здійснювалося за критеріями затримки, семантичної стійкості, роботи в динамічних сценах, сумісності з вбудованим обладнанням, ефективності квантування, потенціалу структурного проріджування та продуктивності оптимізації TensorRT на пристроях NVIDIA Jetson.

Результати. Дослідження встановило, що CNN-архітектури забезпечують найкращу продуктивність у режимі реального часу та залишаються незамінними для високочастотного рефлекторного сприйняття, тоді як Vision Transformers демонструють кращу здатність до глобального контекстного аналізу й стійкість до оклюзій, але зазнають значних обчислювальних витрат на вбудованих GPU. Нейронно підсилені SLAM-методи покращують стабільність ознак та відсіювання динамічних об'єктів, проте вимагають ретельної інтеграції для збереження роботи в реальному часі. Аналіз апаратної реалізації показав, що квантування, структурне проріджування та оптимізація TensorRT є критично важливими для розгортання глибоких моделей на платформах Jetson, хоча ViT-архітектури демонструють обмежену толерантність до INT8-квантування. На основі отриманих результатів сформульовано концепцію гібридної архітектури, що поєднує рефлекторну швидкодію CNN-модулів із когнітивними можливостями моделей Transformer-типу.

Висновки. Результати дослідження підтверджують необхідність розроблення гібридних нейроархітектур, які інтегрують швидкодію та апаратну ефективність CNN-мереж із семантичною глибиною Transformer-моделей. Такі системи становлять перспективний напрям розвитку надійної, повністю автономної високошвидкісної навігації ПМР. Запропоновані принципи акцентують на ієрархічному управлінні, асинхронних контурах сприйняття та апаратно орієнтованій оптимізації як ключових чинниках для створення ПМР нового покоління.

Ключові слова: *ПМР, високошвидкісна навігація, CNN, Vision Transformer, SLAM, Reinforcement Learning, edge-обчислення, Jetson, TensorRT, квантування, проріджування, гібридні архітектури.*

Надійшла у першій редакції 05.10.2025, в останній - 08.11.2025.
The first version has been received on 05.10.2025, the final version – on 08.11.2025.