Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
https://doi.org/10.36930/40360313
Article received: 09.04.2026
Article accepted: 02.06.2026
Article published: 29.06.2026
Correspondence author: A. A. Pipich pipich3@gmail.com

UDC 004.[04+3+89+932]:629.735

A. A. Pipich, B. V. Bulakh
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

ANALYSIS OF THE ARCHITECTURAL EVOLUTION AND HARDWARE-ORIENTED IMPROVEMENTS OF UAV SYSTEMS FOR REAL-TIME OBJECT DETECTION

This survey examines the real-time object detection by UAVs between 2024 and 2026 years and explores its critical balance between accuracy, inference speed, and power consumption under strict hardware limitations. These hardware limitations are defined by typical conditions that are expected when mentioned object detection modules are mounted on UAVs. Survey also explores how modern architectures of real-time object detection models evolved from classic CNN-centric designs (YOLOv11) to attention-centric paradigms (YOLOv12) and latest solutions with native NMS-free inference (YOLO26). It also explores how these architectures can suffer from significant engineering challenges, such as "heavy" architectural blocks that risk overflowing memory bandwidth on entry-level edge platforms. The survey examines the most critical research gaps of an industry of real-time object detection by UAVs. Among the existing research gaps in this area two major ones were identified. The first one is related to an observation: approximately 90 % of the newest models (YOLO12, YOLO26, TSS-YOLO, other NMS-Free inference optimized architectures) is missing the verified energy efficiency or energy consumption (by the hardware required for mounting a specific model) data. The second one is the "Dataset Dependency" problem: general benchmarks like MS-COCO are not directly comparable to specialized aerial results. The survey reviews specialized solutions like RTUAV-YOLO and LiteCOD that utilize architectural ingenuity (such as dynamic weight generation and efficient spatial attention) to bridge the massive 8 x performance gap (TOPS) found across different hardware tiers. To address mentioned discrepancies, the survey evaluates advanced optimization strategies, including mixed-precision Pareto optimization (PoMQ-ViT), hardware-aware Neural Architecture Search (NAS), and structural branch removal (TSS-YOLO), which is found to be more effective than standard quantization alone. The introduction of Unified Complexity Coefficient is proposed in this article to resolve an issue of performance normalization and allow to compare results of general benchmarks and specialized solutions. The survey is also predicting a potential industry shift toward adaptive, "compression-first" systems, which may finally ensure true real-time independency and autonomy for the next generation of autonomous aerial platforms.

Keywords: unmanned aerial vehicles; Edge AI; NMS-Free inference; small target recognition; energy efficiency.

Introduction / Вступ

The demand for autonomous Unmanned Aerial Vehicles (UAVs, БПЛА) has expanded rapidly across critical sectors such as environmental inspection, infrastructure management, and public safety. To operate with genuine autonomy in dynamic environments, these platforms must perform high-performance perception-to-action cycles (цикли "сприйняття-дія") entirely onboard, bypassing the need for external network resources to eliminate communication latency [2].

Significant progress has been made in the world scientific literature toward optimizing real-time object detection, notably through the You Only Look Once (YOLO) family of models established by Joseph Redmon [8] and further developed by researchers like Glenn Jocher. Recent contributions from scientific authorities such as Ranjan Sapkota [8] and Yunjie Tian [10] have pushed architectural boundaries, transitioning from classic CNN-centric designs to attention-centric paradigms (YOLOv12) and native NMS-free inference (нативний інференс, вільний від придушення немаксимумів) (YOLO26). Furthermore, experts like Ruizhi Zhang [14] and Nadin Habash [3] have specifically addressed the challenges of aerial imagery, including information sparsity and the detection of small targets occupying less than 1 % of the frame.

While current research has yielded sophisticated lightweight modules and multi-scale fusion strategies like MSFAM and PDSCM, a critical transparency gap persists. Approximately 90 % of state-of-the-art models released in 2025-2026 currently lack verified energy efficiency data (measured in Watts), which is vital for planning battery-governed autonomous missions. Moreover, the "Dataset Dependency" problem remains unresolved: high accuracy results on specialized aerial datasets are not directly comparable to general benchmarks like MS-COCO, necessitating a formal normalization framework [11].

The relevance of this study stems from the massive 8x performance gap (TOPS, тераопераций за секунду) between entry-level edge hardware (периферійне апаратне забезпечення) and flagship platforms. Addressing the essence of the problem – the risk of memory bandwidth saturati-

Піпіч Артем Андрійович, аспірант, кафедра системного проєктування.
Email: pipich3@gmail.com; https://orcid.org/0009-0008-9956-4514
Булах Богдан Вікторович, канд. техн. наук, доцент, кафедра системного проєктування.
Email: bogdan.bulakh@gmail.com; https://orcid.org/0000-0001-5880-6101

Цитування за ДСТУ: Піпіч А. А., Булах Б. В. Аналіз архітектурної еволюції та апаратно-орієнтованих вдосконалень засобів БПЛА для виявлення об'єктів у режимі реального часу. Науковий вісник НЛТУ України. 2026, т. 36, № 3. С. 122–131.
Citation APA: Pipich, A. A., & Bulakh, B. V. (2026). Analysis of the architectural evolution and hardware-oriented improvements of UAV systems for real-time object detection. *Scientific Bulletin of UNFU*, 36(3), 122–131. https://doi.org/10.36930/40360313

on and thermal instability on resource-constrained platforms – is essential to ensure real-time continuity. There is a pressing need for further research into performance normalization and adaptive, "compression-first" architectural designs (архітектурні рішення з пріоритетом стиснення) to achieve true operational independence for next-generation aerial platforms.

Object of research – the processes of automated recognition and identification of moving objects in real-time video streams.

Subject of research – methods, models, and algorithms for optimizing neural network architectures for object recognition, which allows achieving an optimal balance between accuracy, inference speed and power consumption under UAV resource constraints.

The purpose of the research – analyze the architectural evolution and hardware-oriented improvements of UAV systems for real-time object detection, enabling the determination of an optimal balance between the required accuracy of object recognition, image processing speed, and low system power consumption.

To achieve this *purpose*, the following main *research objectives are identified*:

- analyze the evolution of models architecture, starting from classic CNNs (YOLOv11) to attention-centric (YOLO12) and NMS-free (YOLO26) models, which allows for identifying how specific structural simplifications, such as eliminating post-processing bottlenecks, can overcome memory bandwidth saturation and thermal instability on entry-level edge platforms;
- evaluate optimization strategies (including mixed-precision Pareto optimization, hardware-aware NAS and structural branch removal) regarding their impact, which allows for reaching superior efficiency-accuracy trade-offs by moving beyond standard post-training quantization toward automated, hardware-aligned model designs;
- analyze the transparency gap related to power consumption (Watts) verified data, which is mostly unavailable for nearly 90 % of the latest real-time models, which allows for the objective selection of vision architectures for autonomous UAV missions where operational autonomy is strictly governed by battery life;
- justify a framework for a Unified Complexity Coefficient, which allows for resolving the "Dataset Dependency" problem and normalizing performance assessment across disparate Edge AI environments by weighting accuracy against object density and spatial distribution.

Analysis of recent research and publications. The following section provides a brief review of the key literary sources serving as the foundation for this survey, focusing on the architectural evolution and optimization strategies for edge-based object detection.

The Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection study [8] introduces YOLO26, an edge-optimized model that achieves a 43 % increase in CPU inference speed by eliminating non-maximum suppression (NMS) and distribution focal loss (DFL) post-processing bottlenecks. It utilizes the MuSGD optimizer and Small-Target-Aware Label Assignment (STAL) to enhance training stability and recall for small or occluded instances in robotics and aerial feeds.

In RTUAV-YOLO: A Family of Efficient and Lightweight Models for Real-Time Object Detection in UAV Aerial Imagery [14] the researchers present RTUAV-YOLO, a set of lightweight models based on YOLOv11 tailored for UAV imagery, which reduces parameter counts by 65.3 % while improving detection accuracy. Key innovations include the MSFAM module for adaptive feature modulation and PDSCM for progressive receptive field expansion, concentration on the information sparsity of small objects.

The YOLOv12: Attention-Centric Real-Time Object Detectors [10] contains a detailed documentation attached which details YOLO12, an attention-centric architecture that implements Area Attention and Residual Efficient Layer Aggregation Networks (R-ELAN) to achieve transformer-level accuracy at real-time CNN speeds. While offering significant accuracy gains, it notes potential training instability and elevated memory consumption due to its heavy attention blocks.

In LiteCOD: Lightweight Camouflaged Object Detection via Holistic Understanding of Local-Global Features and Multi-Scale Fusion [5] the LiteCOD solution is proposed as a lightweight framework for camouflaged object detection, integrating local spatial precision with global semantic understanding through Holistic Unification Modules (HUM). It achieves 76 FPS with only 5.15M parameters by introducing an ultra-lightweight Efficient Spatial Attention (ESA) mechanism that reduces FLOPs through shared projections.

The RMTD-YOLO: a lightweight high-precision real-time detection method for road milestones work [11] develops RMTD-YOLO, a specialized algorithm for detecting road milestones and recognizing text via PaddleOCR, optimized for targets occupying less than 1 % of the frame. The architecture utilizes P2Lite and DySample modules to preserve fine-grained textures, achieving 234 FPS for real-time vehicular inspection.

The Comparative Analysis of Object Detection Models for Edge Devices in UAV Swarms study [7] provides an exhaustive benchmark of state-of-the-art YOLO models on edge platforms like NVIDIA Jetson and Intel Iris Xe, specifically for UAV swarm applications. It quantifies the performance and energy efficiency benefits of TensorRT FP16 optimization, demonstrating that low-precision inference can achieve up to a 4 x gain in efficiency.

This specific survey is required to deal with a major energy efficiency research gap in the current literature: verified power consumption data (measured in Watts) is absent for approximately 90 % of the state-of-the-art models released in 2025-2026, including major architectures like YOLO12 and YOLO26. While foundational benchmarks exist for older models (e.g., 8.2-8.3W for YOLOv10/11), the lack of transparency for newer, attention-heavy paradigms prevents objective model selection for autonomous missions where battery life is the primary constraint.

Furthermore, the survey is essential to resolve the "Dataset Dependency" problem, where high accuracy results on domain-specific datasets (up to 94 %) are not directly comparable to general benchmarks (~40 %), necessitating the formalization of a Unified Complexity Coefficient to normalize performance assessment across disparate edge-AI environments.

Materials and methods of research. The study employs a systematic analysis of State-of-the-Art (SOTA) 2024-2026 architectures, prioritizing paradigms centered on NMS-free inference, attention-centric designs, and hybrid CNN-Transformer backbones for real-time Edge AI. A four-vector taxonomy is applied to classify these models based on their architectural paradigm, compression strategies, target-specific characteristics such as small or camouflaged objects, and hardware compatibility. The core of this met-

hodology utilizes a Synthesis Matrix to evaluate inherent conflicts between accuracy, inference speed, and power consumption across varying hardware performance tiers.

Model performance is validated across specialized aerial datasets including VisDrone2019, UAVDT, and UAVVaste, alongside camouflaged object detection (COD) benchmarks such as COD10K, CAMO, NC4K, and CHAMELEON. Evaluation metrics are systematized into accuracy assessments (mAP50/95 and structural metrics like Structure-measure and adaptive E-measure), efficiency parameters (GFLOPs, Latency, FPS), and hardware-specific indicators including RAM consumption and energy usage in Watts to quantify operational autonomy. To resolve the identified Dataset Dependency problem and allow for direct comparability between general-purpose and domain-specific benchmarks, a Unified Complexity Coefficient is proposed for result normalization.

Research results and their discussion / Результати дослідження та їх обговорення

The important problem in modern real-time perception is the conflict between the increasing computational complexity of 2025-2026 architectures and the strict onboard resource limits of Unmanned Aerial Vehicles (UAVs). While next-generation models introduce powerful modeling capabilities, such as Area Attention in YOLOv12, these "heavy" blocks create high risks of RAM consumption overflows and memory bandwidth saturation on entry-level edge platforms (resource-constrained embedded platforms co-located with sensors and actuators) [10]. As a result achieving a stable perception-to-action cycle (rapid transition from sensory perception to robotic execution) requires a precise balance between accuracy, inference speed, and power consumption, which often forces a trade-off between architectural purity and raw modeling strength.

There is the significant gap in verified energy efficiency data (measured in Watts) for approximately 90 % of state-of-the-art models released in 2025-2026. While verified power consumption figures exist for foundational models like YOLOv10 and YOLOv11 (ranging from 8.2W to 8.3W [7]), such data is conspicuously absent for newer paradigms including YOLO12, YOLO26 [8], and TSS-YOLO [11]. There are some autonomous missions where the operational autonomy is strictly governed by battery life. And this lack of transparency becomes an issue to the objective selection of such architectures.

Evaluation metrics in the analyzed literature are impacted by the "Dataset Dependency" problem. For example, the mAP results achieved on general-purpose benchmarks like MS-COCO (~40 %) are not directly comparable to those from domain-specific UAV or road datasets (up to ~94 %). This discrepancy arises because of that fact that specialized aerial datasets introduce unique spatial distributions and object densities that standard hyperparameter tuning cannot normalize.

For this reason there may arise a need in Unified Complexity Coefficient which may be used for results normalization or weighting. The object detection in case of UAV-based scenario must overcome following physical and environmental problems:

- *Small Target Recognition.* In case of the high-altitude some objects frequently occupy less than 1 % of the frame area. This is often causing an information decay or feature aliasing during the convolutional downsampling [3].
- *Camouflaged Object Detection* (COD). Targets visually are often extremely similar to their surroundings. This case requires a specialized "Search-Identification" paradigms or metrics like "Structure-measure" (S_α) [9] or "adaptive E-measure" (E_f) [5]. This allows to quantify boundaries fidelity.
- *Ego-motion.* The fast movement of the UAV platform usually causes motion blur and complicates the segregation of moving targets from a dynamic background.
- *Real-time Continuity.* Onboard systems must maintain an inference speed within the 33.3 ms frame-time window (30 FPS) to ensure that no important information is lost from the video stream [7].
- *Classification imbalance.* Aerial datasets often include some overrepresented categories (such as cars) alongside rare instances. Because of this the re-weighting mechanisms may be required – or applying some targeted augmentation strategies (to improve recall for minority classes) [3].

As the inclusion criteria this analysis is prioritizing models centered on following: NMS-Free inference, attention-centric designs or hybrid CNN-Transformer backbones. To classify analyzed models the four-vector taxonomy is applied. It is based on next characteristics: architectural paradigm, compression strategies, target-specific characteristics (such as small or camouflaged objects), and hardware compatibility.

And as the main component of this analysis a Synthesis Matrix is used. It allows to evaluate conflicts between accuracy, inference speed, and power consumption across varying hardware tiers. This multidimensional classification is visually represented in the four-vector taxonomy (see Figure 1), which maps the interplay between architectural paradigms, optimization methods, target-specific characteristics, and hardware compatibility.

The evaluation process utilizes a combination of general-purpose and domain-specific datasets to ensure both generalizability and application-specific relevance [7]. These are categorized into two primary domains. Aerial Imagery, the first one, includes VisDrone2019, UAVDT, and UAVVaste, which are essential for simulating bird's-eye view perspectives common in UAV operations. The second one is Camouflaged Object Detection (COD). The benchmarks COD10K, CAMO, NC4K, and CHAMELEON are used to assess a model's ability to distinguish targets with high background similarity (including models like LiteCOD [5], LINet [6], BGNet [9] or EfficientNet-B7/Swin-T [12]).

Another potential issue is an already mentioned Dataset Dependency problem. For example the mAP results on general benchmarks (like MS-COCO which is typically ~40 %) are not directly comparable to specialized UAV or road datasets (up to ~94 %) [11]. Evidence suggests that architectural optimizations, such as the MSFAM (Multi-Scale Feature Adaptive Modulation) [14] and DSCBlock (Deformable Separable Convolution Block) [4], allow to achieve more substantial performance gains on edge devices when applied to these specialized datasets compared to general hyperparameter tuning on standard sets.

To provide a full view of model performance on resource-constrained platforms, metrics are systematized into three categories: accuracy metrics, efficiency metrics and hardware-specific metrics. For accuracy metrics the primary assessment relies on mAP50 and mAP50-95. Camouflaged object detection usually requires an application of some specialized metrics: Structure-measure (S_α), Mean Absolute Error (MAE), weighted F-measure (F_w^β) or adaptive E-measure (E_f). These metrics allows to better quantify object structure preservation and boundary fidelity.

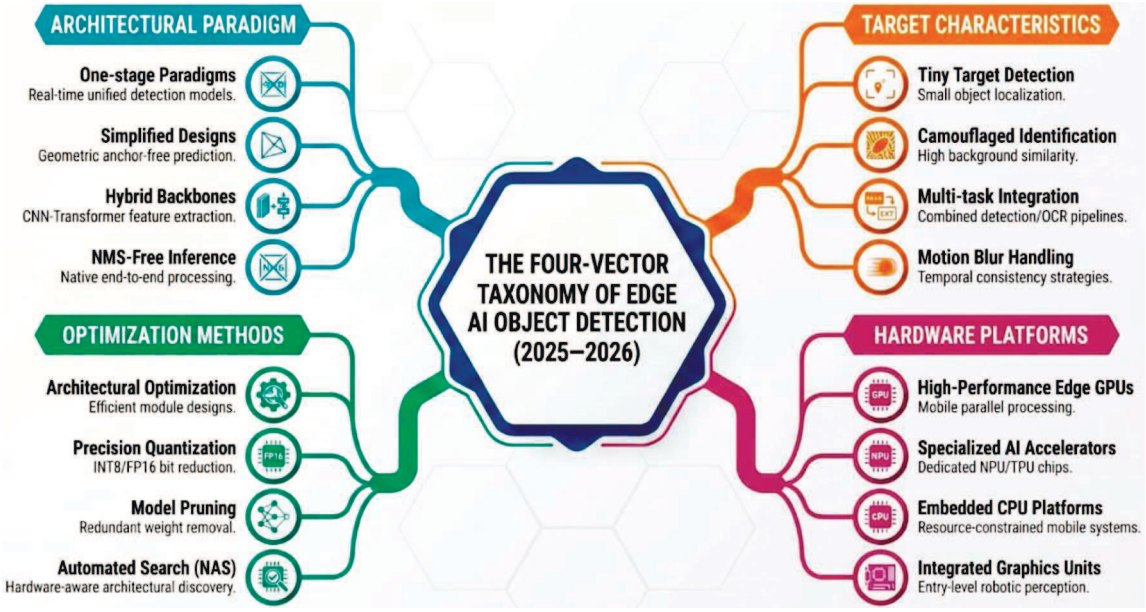

Fig. 1. The taxonomy of Edge AI Object Detection (2025-2026) / Таксономія розпізнавання об'єктів на Edge AI (2025-2026)

The Structure-measure evaluates structural similarity by combining object-aware So and region-aware *Sr* components, formalized as $S_\alpha = \alpha \cdot S_o + (1 - \alpha) \cdot S_r$. The α is typically set to 0.5 to balance global and local structural assessment [9]. The adaptive E-measure (Enhanced-alignment measure) provides a holistic assessment by jointly considering local pixel-level accuracy and global image-level statistics [5]. It is formalized through an alignment matrix ϕ that captures the correlation between the prediction P and ground truth *G* for $W \cdot H$ size image:

$$E_f = \frac{1}{W \cdot H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi(P_{ij} G_{ij}) .$$

It is ensuring high sensitivity to the preservation of object shapes in environments where foreground and background textures are nearly identical.

Efficiency metrics can include: parameter count (M), GFLOPs, Latency (ms) or Frames Per Second (FPS). This allows to measure computational complexity and throughput. As for hardware-specific metrics the RAM consumption (GB) and Energy usage (Watts) are highlighted as important indicators of operational autonomy. As verified, the power consumption figures are missing for about the 90 % of models released in 2025-2026 (including major YOLO12 [10] and YOLO26 [8]), which can be a significant research gap in energy efficiency data.

TOPS (Tera Operations Per Second) disparity between different hardware tiers can significantly constrain the deployment of real-time object detection on the edge devices. For this reason bridging this performance gap requires matching architectural efficiency with the specific computational strengths and power limits of the underlying hardware platform.

The NVIDIA Jetson family (including the Nano, Orin, and Xavier series) is the most common choice for autonomous UAV monitoring. The reason is its balance between architectural flexibility and parallel processing power. But it has an issue important for developers which is the massive performance gap within this family. For example, there is an 8x jump in TOPS between entry-level models and flagship tiers, with the Jetson Orin AGX (275 TOPS) dwarfing the legacy Jetson Nano (0.5 TFLOPS / ~5 TOPS) [7]. As a result while flagship platforms can support dense attention blocks, entry-level hardware requires extreme architectural frugality, such as the use of MSFAM and PDSCM modules, to maintain real-time performance [14].

Specialized NPU and TPU accelerators (e.g. Huawei Atlas 200 and T710) are designed specifically for fixed-precision (INT8) inference [3]. These chips prioritize high energy efficiency and deterministic latency. These parameters are important for achieving the target benchmark of <3 ms latency across varying hardware tiers [2]. Deploying modern architectures on these accelerators often requires a "compression-first" approach, where quantization and pruning are integrated into the design phase rather than applied as a post-processing step.

Platforms that lack dedicated AI accelerators (such as CPU-bound, mobile or embedded) must rely on host ARM or x86 processors, requiring significant structural simplifications to meet real-time constraints. Android Mobile Platforms are typically used for ARM-based applications, such as in-app road damage inspection or simple infrastructure monitoring [1]. Utilizing a quad-core ARM A72, Raspberry Pi 4B platforms are commonly deployed for education and low-cost swarm sensors.

Real-time deployment in Embedded ARM or TFLite/NCNN environments often utilizes optimized frameworks and structural modifications, such as the removal of redundant large-object branches (as seen in TSS-YOLO), which can be more effective than standard quantization alone [11]. This helps the models to run instantly and without delays in real time.

Integrated Graphics (Intel Iris Xe) support integrated graphics inference. That is achieved by ecosystems like OpenVINO for entry-level robotics. The technical specifications and peak AI performance (TOPS/TFLOPS) for the primary hardware platforms utilized in modern UAV monitoring are summarized [1, 2, 3, 4, 7, 8, 10, 14] in Table 1. The evolution of object detection architectures between 2024 and 2026 reflects a strategic shift from pure convolutional neural networks (CNNs) toward hybrid designs and simplified inference pipelines tailored for edge deployment.

Initial improvements in the YOLO lineage focused on optimizing feature extraction through refined CNN blocks. YOLOv11 can represent a significant step in this specific direction, inheriting the hierarchical design of previous ver-

sions while introducing the C3k2 module [3]. This architecture improves gradient flow and semantic representation while achieving a 22 % reduction in parameters compared to the medium-scale YOLOv8 model. These CNN-centric designs prioritized local convolutional operations to maintain real-time speeds on standard hardware.

Table 1. Edge AI Hardware Specifications and Performance Tiers /
Технічні характеристики та рівні продуктивності Edge AI обладнання

Platform	AI Performance	GPU Architecture	Memory	Power (W)	Key Notes
Jetson AGX Orin	Up to 275 TOPS	2048-core Ampere	32GB/64GB LPDDR5	15W–60W	8x faster than AGX Xavier; Flagship tier
Jetson Orin Nano	Up to 40 TOPS	1024-core Ampere	4GB/8GB LPDDR5	7W–15W	Integrates 32 Tensor Cores
Jetson AGX Xavier	32 TOPS	512-core Volta	16GB/32GB LPDDR4 x	10W–30W	Standard for mid-size UAV models
Jetson Xavier NX	21 TOPS	384-core Volta	8GB/16GB LPDDR4 x	10W–20W	High efficiency for UAV swarms
Jetson Nano SUPER	~5 TOPS	1024-core Ampere	8GB LPDDR5	7W–25W	Features MAX_N performance mode
Jetson Nano (Legacy)	0.5 TFLOPS	128-core Maxwell	2GB/4GB LPDDR4	5W–10W	Entry-level hardware; strict power limits
Jetson TX2	1.3 TFLOPS	256-core Pascal	8GB LPDDR4	7.5W–15W	Used for older NPU drivers
Google Coral TPU	4 TOPS (INT8)	Edge TPU	Optional	2W	For extremely power-constrained apps
Huawei Atlas 200I	Up to 20 TOPS	Ascend NPU	4GB/8GB	~15W	Used in specialized NPU studies
Qualcomm RB5/RB6	15 TOPS	Heterogeneous	LPDDR5	N/A	Targets 5G-connected drones

In 2025, the release of YOLOv12 marked a transition toward an attention-centric paradigm, integrating vision transformer (ViT) modeling capabilities into a real-time framework. To overcome the quadratic complexity of standard self-attention, YOLOv12 utilizes Area Attention, which partitions feature maps into equal horizontal or vertical areas to ensure a large receptive field with low computational cost [8].

The introduction of Residual Efficient Layer Aggregation Networks (R-ELAN) addresses optimization problems inherent in deep attention models [10]. By incorporating a residual shortcut with a scaling factor (default 0.01), R-ELAN stabilizes training and ensures convergence for large-scale models. Architectural efficiency is further enhanced by removing explicit positional encoding and instead using a 7×7 separable convolution, known as a position perceiver, to help the network perceive spatial information [7].

The subsequent evolution toward YOLO26 prioritizes the elimination of deployment bottlenecks. A important innovation in this architecture is the native NMS-Free inference, which removes the need for Non-Maximum Suppression post-processing [11]. This transition eliminates variable post-processing latency and simplifies the deployment pipeline for real-time edge devices. Also YOLO26 adopts DFL-free regression by removing the Distribution Focal Loss (DFL) module. This structural simplification leads to 43 % faster CPU inference and significantly improves export compatibility for frameworks like TensorRT and TFLite.

These architectural shifts reveal a core conflict between high-accuracy modeling and hardware-constrained deployment. While attention-centric models like YOLOv12 offer superior global dependency capture, they introduce significant RAM consumption risks and potential memory bandwidth overflows on older edge platforms like the Jetson Nano [3].

To compare, the NMS-Free models like YOLO26 focus on "architectural purity", prioritizing predictable low-latency performance at the cost of the stronger modeling capabilities found in dense attention blocks [8]. This paradigm shift underscores that for Edge AI, architectural efficiency is as important as raw accuracy. Modern Edge AI architectures (2025 mostly) actively using a "compression-first" approach. Its architectural efficiency is integrated into the design phase instead of being applied solely as a post-processing step.

The main complication in deploying Vision Transformers (ViT) is the heterogeneous sensitivity of different encoder layers to bit-width reduction [13]. The PoMQ-ViT (Mixed-Precision Quantization with Pareto Optimization) method addresses this by formulating bit allocation as a multi-objective optimization problem. Sensitivity Assessment method evaluates layers using distribution change measurement functions before and after quantization. That is done to identify layers that can tolerate 4-bit precision versus those requiring 8-bit precision. By applying the Pareto optimality approach, the system identifies the optimal trade-off between Top-1 accuracy and inference latency. The implementation of mixed-precision configurations allows for significant speedups of 6-24 ms. This eliminates the need for model retraining, maintaining the accuracy within acceptable boundaries for edge deployment.

The evolution of Neural Architecture Search (NAS) has moved toward integrating Large Language Models (LLMs). That allows to manage the massive search space for specialized hardware [3]. PhaseNAS Framework approach uses the analytical reasoning of LLMs by using structured prompt templates. This allows to guide the exploration and refinement of architectures. Applying the PhaseNAS framework can reduce the architecture search time by up to 86 %. This may allow the developer to rapidly tailor models to specific hardware targets, such as the Jetson Orin Nano, ensuring that the generated designs meet strict computational and dimensional consistency limits.

Significant architectural shifts in models like YOLO26 focus on eliminating post-processing bottlenecks. Such bottlenecks frequently hinder stable exports to mobile formats like TFLite and CoreML [8]. By replacing DFL with a lighter and hardware-friendly regression task, models simplify the computational graph, improving compatibility with TensorRT and ONNX. And the transition to native end-to-end prediction removes the reliance on Non-Maximum Suppression (NMS) as external code. This elimination of post-processing overhead results in a 43 % increase in CPU inference speed and a more predictable latency profile on ARM-based edge devices.

The integration of heavy attention blocks in models like YOLO12 and YOLO26 introduces significant optimization problems and potential training instability [10]. To address this, the MuSGD optimizer (a hybrid of Stochastic Gradient Descent (SGD) and Muon-style curvature behaviors) is employed. This hybrid approach leverages practices from LLM training to ensure faster, smoother convergence and stable training even for large-scale models with dense atten-

tion modules, reducing the number of required epochs for the best (optimal) performance.

Final deployment on edge hardware relies on complex inference engines that demand "architectural purity" to minimize latency [8]. The ecosystem is completely dominated by such entities as TensorRT 10.x for NVIDIA GPUs, OpenVINO for Intel hardware, and ONNX Runtime for general cross-platform compatibility [7]. Clean architectural designs (such as those that avoid custom DFL operations or complex post-processing plugins) enable seamless acceleration by INT8 quantization and hardware-specific kernel optimization. All of them are vital for maintaining the <3 ms latency benchmark which is required for real-time UAV autonomy in practice.

RTUAV-YOLO was designed to mitigate a significant information decay associated with small-target detection in Unmanned Aerial Vehicle (UAV) imagery. In high-altitude scenarios (where targets can occupy even less than 32×32 pixels), the traditional convolutional downsampling induces feature aliasing and gradient sparsity.

The Multi-Scale Feature Adaptive Modulation (MSFAM) module replaces the YOLOv11 C3K2 backbone component [14]. It is using a Dynamic Weight Generator (DWG) to synthesize content-aware modulation coefficients and a Dual-Branch Feature Extractor (DBFE). This DBFE employs a 3:1 channel allocation strategy. That is a calculated computational budget decision which prioritizes local detail in shallow layers (3×3 convolutions). In the same time it is reserving bandwidth for dilated context ($d = 2$) in deeper stages. To avoid increasing parametric density but still capture the multi-scale context PDSCM (Progressive Dilated Separable Convolution Module) can be integrated to the stage P4. Module is using five layers of depth-wise separable convolutions with progressive dilation rates ($d = 1, 2, 3$). In this way the module achieves a cumulative receptive field:

$$RF_{total} = \sum_{i=1}^{n} RF_i - (n-1) = 3 + 5 + 7 - 2 = 13 .$$

And this allows the network to model spatial relationships between distant features, essential for distinguishing minute targets from cluttered maritime or urban backgrounds. The PDSCM progressive expansion is additionally expanded by the P2 head for detection. It is introduced specifically to prioritize detailed features over semantic map. And for this reason, mitigating the feature aliasing is usually observed in standard stride-2 convolutional downsampling.

To rectify the insensitivity of standard IoU to small-pixel perturbations, RTUAV-YOLO incorporates Minimum Point Distance Wise IoU (MPDWIoU). This loss function integrates distance-aware penalty terms to mitigate gradient imbalance, ensuring precise localization even when bounding box overlap is near-zero.

The Lightweight DownSampling Module (LDSM) implements a "sample-first, nonlinearity-second" strategy. It preserves fine-grained texture information (which – in legacy architectures – is typically lost in multiple convolutions). That is achieved by applying linear grouped convolutions for downsampling before applying pointwise nonlinear transformations. Compared to the YOLOv11 as a baseline, there is a total 65.3 % parameter reduction, contributed by this specific refinement as well.

The YOLOv12 represents a paradigm shift in the direction of Attention-Centric architectures. It replaces traditional CNN-based backbones with Area Attention approach. It divides feature maps into equal-sized regions. That maintains a large receptive field and, in the same time, reducing the quadratic complexity of vanilla self-attention to $O(l^{-1} \oplus 2n^2 \oplus hd)$. It also utilizes Residual Efficient Layer Aggregation Networks (R-ELAN) – that allows to resolve a training instability inherent in large attention models [10]. This module incorporates block-level residual connections – the scaling factor is only 0.01. That is ensuring stable gradient flow and enabling X-scale models to converge where standard ELAN structures fail. The LiteCOD is using the "Search-Identification" paradigm. It us mimicking biological hunting mechanisms that helps to recognize objects which blend into the background which can have high similarity with an object itself.

Holistic Unification Modules (HUM) [5] facilitate bilateral global-local feature enhancement. They unify a global pathway (semantic context) and a local pathway (spatial precision) through cross-attention, ensuring that boundary delineation remains sharp even under minimal foreground-background contrast. To maintain edge-tier efficiency, the ESA mechanism achieves drastic GFLOP reduction through shared query-key projections C/32. It employs row-wise spatial relationship modeling, capturing long-range dependencies with lower arithmetic intensity than standard self-attention.

Positioned at the deepest encoder stages (S4, H4), the ECG module utilizes criss-cross attention to provide global semantic guidance. This ensures that coarse semantic features effectively propagate through the decoder hierarchy to guide fine-grained spatial refinement. The substantial reduction in computational overhead achieved by the LiteCOD architecture compared to traditional heavyweight models [5] is detailed in Table 2.

Table 2. Efficiency Benchmarks: LiteCOD vs. Traditional Heavyweight Models / Порівняння ефективності LiteCOD та традиційних важких моделей

Metric	LiteCOD (SOTA Lightweight)	Traditional Heavyweight Models (BGNet/SINet)
Parameters	5.15 M	26 M – 79.85 M
GFLOPs	7.95	High Redundancy
Inference Speed	76 FPS (RTX 4080) / 19 FPS (Jetson Orin AGX)	Often < 30 FPS on Edge devices
Architectural Core	Holistic Unification Modules (HUM) + Efficient Spatial Attention (ESA) + Enhanced Context Generation (ECG)	Multi-branch redundant CNNs

RMTD-YOLO is engineered for sub – 1 % area detection. And specifically – for road milestones in high-speed vehicular contexts. Lets consider objects of this size as "small". To allow the architecture to focus on such small objects the integration of P2Lite and DySample modules can be applied [11]. But then the problem of minimizing information loss during high-ratio feature aggregation appears, which is resolved by the DLSPPF (Deep Lightweight Spatial Pyramid Pooling Fast) and ACF (Adaptive Context Fusion) modules, which are maintaining the further structural integrity.

The model is achieving an architectural synergy – by coupling the detection head with PaddleOCR. This "unified" pipeline enables the real-time detection and text recognition. That is achieved without any latency overhead for

multi-model handoffs. Such architecture allows to have the 94.4 % mAP50 at 234.9 FPS on vehicle-mounted hardware. In this way it is providing a robust solution for autonomous infrastructure monitoring. The next paradigm shift to the NMS-Free inference was YOLO26. YOLO26 is eliminating the post-processing bottlenecks (that traditionally throttle edge CPUs). This allows the model to run directly and significantly speeds up the detection of objects in real time, particularly on standard processors (CPUs) and mobile chips.

YOLO26 minimizes the CPU-to-GPU synchronization overhead. This is achieved by removing Non-Maximum Suppression (NMS) and Distribution Focal Loss (DFL) [8]. Detections are produced end-to-end – which significantly improving throughput for TensorRT and TFLite exports. This architectural simplicity is enabled by specialized training algorithms: MuSGD (Multi-uncertainty SGD) and ProgLoss 1-3 (Progressive Loss). They stabilize the one-to-one assignment required for NMS-Free inference.

As for inference dynamics – the traditional YOLO (v8-v11) provides high post-processing latency. It scales poorly with object density as well as also requires a complex plugin layers for hardware acceleration. YOLO26 provides 43 % acceleration in CPU performance and constant inference time – and that is regardless of object count or seamless deployment on ARM-based edge CPUs.

CSPDNet utilizes DSCBlocks (Depthwise Separable Convolutions) and the CSPDBlock to restructure gradient paths [4]. This reduces parameter counts by 20 % while integrating DCNv2/v4 (Deformable Convolutions) to handle the extreme perspective distortions found in VisDrone datasets. This approach is specifically optimized for the memory bandwidth constraints of Jetson Xavier NX hardware.

LINet adopts a single-stage structure that mimics human vision by applying a Receptive Field Module (RFM) [6]. It achieves inference speeds up to 126 FPS, prioritizing the identification of camouflaged contours. That is achieved by fusing multilevel features in a lightweight branch. A synthesized comparison of the architectural features, parameter counts, and inference performance for these leading-edge (SOTA) solutions [4, 5, 6, 8, 11, 14] is presented in Table 3.

Table 3. SOTA Algorithm Synthesis Matrix / Оглядова порівняльна матриця алгоритмів

Model	Architectural Features	Params (M)	GFLOPs	Accuracy (mAP / Sa)	Inference Performance (Target Platform)
RTUAV-YOLO-N	MSFAM, LDSM, MPDWIoU	0.79	8.3	40.1 % (mAP50)	53.7 FPS (Jetson Orin Nano)
LiteCOD	HUM, ESA, ECG	5.15	7.95	85.2 % (Sm)	19.0 FPS (Jetson Orin AGX)
YOLO26-N	NMS-Free, MuSGD, STAL	2.8	5.4	40.9 % (mAP)	1.1 ms (NVIDIA T4)
RMTD-YOLO	P2Lite, DLSPPF, ACF, DySample	~2.69	N/A	94.4 % (mAP50)	234.9 FPS (Vehicle PC)
CSPDNet (v2-n)	CSPDBlock, DCNv2/v4	2.9	4.2	36.7 % (mAP50)	25.0 FPS (Jetson Xavier NX)

A major hurdle for contemporary Edge AI is the enormous performance gap in TOPS (Tera Operations Per Second) across different hardware tiers. High-end systems like the 275-TOPS Jetson Orin AGX have the capacity for complex attention blocks, whereas basic hardware, such as the Jetson Nano (0.5 TFLOPS / ~5 TOPS), necessitates the extreme architectural minimalism found in modules like MSFAM and PDSCM. Furthermore, the LDSM module provides a vital fix for texture aliasing on legacy Pascal and Maxwell architectures by utilizing a "sample-first, nonlinearity-second" method. Achieving the crucial <3ms latency benchmark across all platforms also makes the adoption of NMS-Free heads a necessity rather than an optional enhancement [8]. Within the 2026 UAV and mobile edge landscape, real-time autonomy is powered more by architectural ingenuity than by sheer computational strength.

Real-time object detection in UAV video streams is frequently complicated by ego-motion, which induces motion blur and degrades feature representation [3]. To mitigate these effects, methods utilizing temporal sequences and optical flow are employed to distinguish moving targets from dynamic backgrounds caused by the platform's own movement, which allow the system to account for frame-to-frame dependencies, ensuring that the detection of dynamic objects remains stable even under high-altitude or high-speed flight conditions. Autonomous navigation for UAVs, particularly when guiding ground vehicles through difficult terrains, depends heavily on the integration of temporal context.

To stabilize detections across multiple frames and enable complex tasks like traffic monitoring or object following, lightweight edge-tracking frameworks are integrated directly into the perception pipeline [7]. Recent research has focused on optimizing the DeepSORT tracker. It's used for resource-constrained devices by redesigning its encoder inference pipeline. Moving from the legacy TensorFlow V1 implementation to an ONNX-based batched inference strategy reduces tracking latency from 70-90 ms to 15-25 ms per frame. This allows the tracking system to handle variable loads and high-density scenarios in real time without dropping frames.

InstaTrack leverages instance segmentation to enhance appearance-based tracking by utilizing pixel-level masks instead of simple bounding boxes. By computing coverage values and using contrastive loss, the system improves data association and robustness in crowded scenes where objects frequently overlap. Alternative Perceptual Models emerging architectures like SAM 2 utilize the benefits of "masklets" for advanced segmentation and tracking, providing a path for finer-grained object identification in UAV swarms. Real-time autonomous aerial perception is made possible on edge devices by combining efficient lightweight detectors with optimized tracking algorithms, which preserve stable execution even during complex, congested situations.

As architectural refinement for the YOLO family and dedicated UAV detectors like RTUAV-YOLO hits its zenith, the research vanguard is pivoting toward zero-shot open-vocabulary capabilities (via vision-language models like CLIP), federated collaborative learning, and optimizations, such as auxiliary super-resolution branches, that improve accuracy during training without incurring any additional inference latency [8]. A major main vector in 2025-2026 is the unification of multiple vision tasks into holistic models capable of recognizing arbitrary object categories through natural language supervision. Unlike standard detectors limited to a fixed label set, Contrastive Language-Image Pre-training (CLIP) and its derivatives enable zero-shot detection on UAVs [3]. Recent frameworks like RemoteCLIP align image-text embeddings specifically with aerial scene semantics by converting heterogeneous annotations into caption-style supervision].

Open-Vocabulary frameworks, such as YOLO-World and CastDet, utilize student-teacher self-learning paradigms

where a large pretrained VLM (the "teacher") guides a lightweight detector (the "student") during training [10]. This allows the student model to inherit the semantic richness of the teacher while maintaining real-time edge speeds. Methods such as TernaryCLIP and CLIP-KD enable large Transformers to function on resource-constrained devices by using quantization and feature-level distillation to reduce their overall computational burden.

Federated Edge Learning offers a strategy for performing model updates directly on UAVs operating in sensitive zones, which removes the need to send raw aerial data back to a central server [14]. By employing the Multilayer Collaborative Federated Learning (MLC-FL) algorithm, edge devices can use asynchronous communication to facilitate automatic local model optimization. This methodology allows UAV swarms to adapt to evolving environmental factors, such as terrain or lighting shifts, by transmitting compressed gradients instead of bulky raw data, which simultaneously preserves bandwidth and enhances data privacy.

To combat the degradation of fine textures during downsampling and the "Dataset Dependency" hurdle, researchers are integrating Super-Resolution (SR) assistance into the training phase [8]. Models like SuperYOLO incorporate a specialized, lightweight auxiliary SR branch that enables the system to learn high-resolution feature representations even when processing low-resolution inputs [3]. Notably, these auxiliary branches are discarded during inference, providing the benefit of improved detection for minute objects (occupying <1 % of the area) without introducing any additional computational burden or latency to the onboard system.

What is important, these SR modules are removed during deployment, ensuring that the model benefits from improved feature quality for tiny objects (<1 % area) without increasing the computational load or latency of the onboard inference pipeline. Future iterations of the YOLO lineage are expected to rely less on massive, manually labeled datasets by incorporating self-supervised feature learning. Techniques such as teacher-student training and pseudo-labeling allow models to automatically leverage vast amounts of unannotated video streams to improve recognition robustness in context. The ongoing blending of Transformer global reasoning and CNN local extraction will further drive next-generation detectors to intelligently fuse context and detail, achieving transformer-level accuracy at the Hallmarked speed of YOLO systems [10].

To evaluate the real-time perception models of 2025 and 2026 years there may appear a need in a systematic selection of baselines representing different architectural builds. In such case the comparison of CNN-centric versus Attention-centric should be prioritized. The CNN-centric models (such as YOLOv11) optimize local feature extraction with C3k2 modules [3]. And attention-centric ones (like YOLOv12) leverage Area Attention and R-ELAN for global contextual modeling. And, in addition, the mobile-optimized paradigms. Including them on resource-constrained platforms allows to assess the benefits of inference simplification and memory efficiency. Such examples include the NMS-Free YOLO26 [8] and the depthwise-separable SSD-MobileNet. In the realm of autonomous aerial perception, specialized models such as RTUAV-YOLO (which focuses on targets <1 % of the frame) [14] and RMTD-YOLO (which integrates detection with OCR) [11] serve as essential benchmarks for balancing high-speed multi-task performance with the detection of extremely small objects. The accuracy-latency trade-offs for these diverse architectural paradigms [8, 10] are illustrated in the Pareto Front comparison shown in Figure 2, which highlights the performance boundaries of current real-time detectors on the MS-COCO dataset.

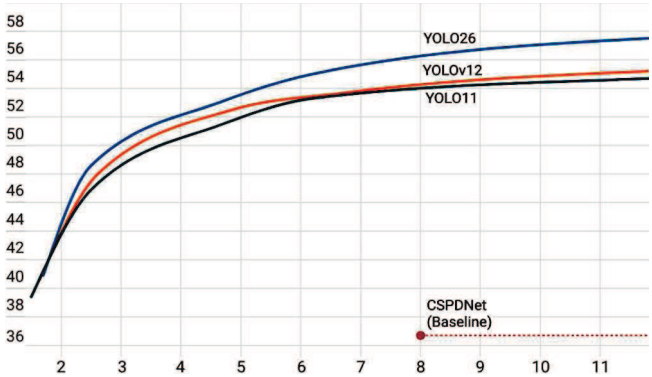

Fig. 2. Accuracy-Latency Pareto Front comparison of SOTA real-time object detectors on the MS-COCO dataset / Парето-фронт порівняння точності та затримки SOTA детекторів (у реальному часі) об'єктів на наборі даних MS-COCO

YOLOv12-N or other attention-centric architectures can demonstrate really good results on powerful hardware (like the NVIDIA T4). But then they can face major efficiency challenges when being deployed on older edge systems like the Jetson Nano. The case demonstrates that gain in performance is far from consistent across different hardware levels. And some aging devices often fail to meet the intensive resource needs of the cutting-edge models [7].

Surprisingly, YOLOv12 can actually be slower than the CNN-based YOLOv11 in such outdated settings. That is usually not expected, but that is what practical tests shows. The reason lies in its complex attention modules which demand too much RAM and memory bandwidth. As a result, reaching a 30 FPS benchmark on budget hardware necessitates highly optimized and minimal designs. The reason is that the raw processing power is insufficient to close the 8x performance gap in TOPS existing between premium and entry-level devices [3].

This research identifies a critical transparency gap: about 90 % of state-of-the-art models from 2025-2026 lack verified power consumption figures [8]. While foundational architectures (like YOLOv10 and YOLOv11) have established metrics (averaging 8.2W – 8.3W on Jetson hardware), newer releases such as YOLO12 and YOLO26 currently operate without this essential performance data. This lack of transparency is a major issue for the objective selection of architectures for autonomous missions. For them the operational autonomy is strictly governed by battery life. So this validates the need for hardware-specific metrics like Watts and RAM usage to be integrated into standard benchmarking protocols.

There is lack of comparability between the specialized data environments and general-purpose benchmarks – and that is the "Dataset Dependency" problem that was already mentioned above. The mAP scores on standard sets (like MS-COCO) typically hover around 40 %. And specialized datasets for UAVs or road milestones (which often feature objects occupying less than 1 % of the frame) can yield results up to 94 % [11]. And the standard hyperparameter tuning cannot normalize such case.

To resolve this identified discrepancy, the necessity of a Unified Complexity Coefficient (ξ) is proposed. This the-

oretical framework would weight mAP against object density and the <1 % frame area threshold typical of UAV imagery [3]. By prioritizing "architectural purity" through structural branch removal, models such as TSS-YOLO achieve greater efficiency for edge deployment than is possible through bit-width reduction alone. For specialized tasks, removing the large-target detection branch leads to a substantial reduction of 10.9 GFLOPs and 5.5 million parameters. While still having a significant decrease in computational complexity, the model maintains accuracy levels that remain superior to baseline versions like YOLOv10s.

This "architectural purity" approach (combined also with NMS-Free inference) minimizes the CPU-to-GPU synchronization overhead. This frequently throttles real-time performance on ARM-based edge CPUs.

Discussion of research results. The findings of this survey, particularly regarding the optimal balance between accuracy, speed, and power consumption for onboard UAV perception, are compared with five representative studies below. A new lightweight network for efficient UAV object detection study [4] presented CSPDNet, which achieved a 20 % reduction in parameters while maintaining an inference speed of 24.7 FPS on the Jetson Xavier NX. These results validate current survey conclusion that specialized architectural modules (like the DSCBlock) provide more significant speed gains for Edge GPUs than standard hyperparameter tuning on standard datasets.

Edge AI for Real-Time Robotic Systems: Architectures, Deployment Strategies, and Performance Optimization work [2] proposed a five-stage deployment methodology for real-time robotic systems. While that research focuses on manual optimization (e.g., INT8 quantization), current survey identifies a critical transparency gap in energy efficiency data that his methodology could not yet address. It is extended by arguing for Hardware-Aware NAS to automate the alignment of these stages with strict power budgets.

PoMQ-ViT: Mixed-Precision Quantization Vision Transformer with Pareto Optimization work [13] is dedicated for mixed-precision bit-width allocation. This directly supports current survey strategy for resolving the conflict between the modeling strength of attention-centric designs (e.g., YOLOv12) and the RAM overflow risks on entry-level edge platforms.

The Recent Real-Time Aerial Object Detection Approaches, Performance, Optimization, and Efficient Design Trends for Onboard Performance survey [3] offers an exhaustive taxonomy of aerial detection challenges. And current research formalizes a specific methodological discrepancy that were left open: the Dataset Dependency problem. It proposes the Unified Complexity Coefficient as a necessary theoretical framework to normalize mAP results across disparate general and domain-specific sets.

In Comparative Performance of Yolov8 and Ssd-mobilenet Algorithms for Road Damage Detection in Mobile Applications [1] by benchmarking SSD-MobileNet V2 on road damage, authors achieved a high accuracy of 91.1 %. This empirical evidence confirms current identification of the mAP gap between general-purpose benchmarks like MS-COCO (~40 %) and specialized aerial/road datasets.

The comparative look shows that while contemporary research has achieved record-breaking speeds (e.g., 126-234 FPS in specialized scenarios), there is a pervasive lack of energy efficiency data (Watts) for the newest 2025-2026 paradigms [8]. Furthermore, the transition from manual tuning to automated "compression-first" designs (like PoMQ-ViT or TSS-YOLO) is proven more effective for resource-critical edge devices than standard post-processing quantization.

So, based on the results of the work performed, it is possible to formulate the following scientific novelty and practical significance of the research results.

Scientific novelty of the obtained research results – received further development the method of determining the Unified Complexity Coefficient of object recognition models, which, unlike existing ones, establishes a novel theoretical framework for weighting mAP scores against object density and the <1 % frame area threshold characteristic of specialized aerial imagery, thereby resolving the "Dataset Dependency" methodological discrepancy where general-purpose benchmarks are not directly comparable to domain-specific UAV results.

Practical significance of the research results – can be applied to bridge the hardware ceiling by demonstrating that architectural ingenuity (manifested through structural simplifications such as native NMS-free inference and adaptive feature modulation) can effectively close the 8x performance gap in TOPS between entry-level edge devices and flagship tiers, thereby enabling reliable real-time perception across a wide spectrum of onboard computational resources.

Conclusions / Висновки

The architectural evolution and hardware-oriented improvements of UAV systems for real-time object detection have been analyzed, enabling the determination of an optimal balance between the required accuracy of object recognition and identification, image processing speed, and the system's corresponding power consumption. Based on the results of the study, the following main conclusions can be drawn:

1. Identified that architectural ingenuity in next-generation models, such as the NMS-free design of YOLO26 and the biomimetic approach of LiteCOD, is more critical than raw processing power for achieving real-time (30 FPS) performance on entry-level edge devices, effectively bridging the 8x performance gap observed between entry-level and flagship hardware tiers.
2. Discovered a significant transparency gap in current literature, where verified energy efficiency data (Watts) is absent for approximately 90 % of state-of-the-art models released in 2025-2026, which hinders the objective selection of architectures for battery-governed autonomous UAV missions.
3. Examined the methodological discrepancy known as the "Dataset Dependency" problem, revealing that mAP results on general benchmarks (~40 %) are not directly comparable to specialized aerial or road datasets (up to 94.4 %), thus necessitating the introduction of a Unified Complexity Coefficient to normalize performance assessments across disparate domains.
4. Determined through the analysis of SOTA solutions that structural branch removal (as seen in TSS-YOLO) and adaptive feature modulation (such as MSFAM in RTUAV-YOLO) provide more substantial computational savings and accuracy preservation for small-target detection than standard post-hoc model compression or general hyperparameter tuning.
5. Applied the evaluation of emerging optimization trends, concluding that the integration of Hardware-Aware Neural Architecture Search (NAS) and mixed-precision Pareto Optimization is essential for automating model configurati-

on to meet the strict energy, memory, and thermal constraints of the next generation of autonomous aerial platforms.

Reference

1. Dharma, A. S., Pardosi, C. N. S., & Silaen, Z. P. (2025). Comparative Performance of Yolov8 and Ssd-mobilenet Algorithms for Road Damage Detection in Mobile Applications. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, *9*(3), 1159–1169. https://doi.org/10.33395/sinkron.v9i3.15008
2. Ghosh, A. (2026). Edge AI for Real-Time Robotic Systems: Architectures, Deployment Strategies, and Performance Optimization. *Universal Library of Engineering Technology*, *3*(1), 07–11. https://doi.org/10.70315/uloap.ulete.2026.0301002
3. Habash, N., Alqumsan, A. A., & Zhou, T. (2025). Recent Real-Time Aerial Object Detection Approaches, Performance, Optimization, and Efficient Design Trends for Onboard Performance: A Survey. *Sensors*, *25*(24), article ID 7563. https://doi.org/10.3390/s25247563
4. Hua, W., Qili, C., & Wenbai, C. (2024). A new lightweight network for efficient UAV object detection. *Scientific Reports*, 14, article ID 13288. https://doi.org/10.1038/s41598-024-64232-z
5. Khan, A., Ullah, H., & Munir, A. (2025). LiteCOD: Lightweight Camouflaged Object Detection via Holistic Understanding of Local-Global Features and Multi-Scale Fusion. *AI*, *6*(9), article number 197. https://doi.org/10.3390/ai6090197
6. Li, Q., Wang, Z., Zhang, X., & Du, H. (2024). Lightweight camouflaged object detection model based on multilevel feature fusion. *Complex & Intelligent Systems*, *10*, 4409–4419. https://doi.org/10.1007/s40747-024-01386-3
7. 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), article ID 684. https://doi.org/10.3390/machine13080684
8. Sapkota, R., Cheppally, R. H., Sharda, A., & Karkee, M. (2025). YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection. [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2509.25164
9. Sun, Y., Wang, S., Chen, C., & Xiang, T. (2022). Boundary-Guided Camouflaged Object Detection. In L. De Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (pp. 1324–1331). *International Joint Conferences on Artificial Intelligence Organization*. https://doi.org/10.48550/arXiv.2207.00794
10. Tian, Y., Ye, Q., & Doermann, D. S. (2025). YOLOv12: Attention-Centric Real-Time Object Detectors [Preprint]. arXiv. URL: https://doi.org/10.48550/arXiv.2502.12524
11. Wen, S., Yu, J., Guo, J., Li, Y., Li, R., Zhang, J., Li, L., & Lv, S. (2026). RMTD-YOLO: a lightweight high-precision real-time detection method for road milestones. *Journal of Real-Time Image Processing*, *23*(2), article number 63. https://doi.org/10.1007/s11554-026-01857-5
12. Wen, Y., Ke, W., & Sheng, H. (2024). Camouflaged Object Detection Based on Deep Learning with Attention-Guided Edge Detection and Multi-Scale Context Fusion. *Applied Sciences*, *14*(6), article ID 2494. https://doi.org/10.3390/app14062494
13. Wu, Z., & Zhao, Z. (2025). PoMQ-ViT: Mixed-Precision Quantization Vision Transformer with Pareto Optimization. *Applied Sciences*, *15*(18), article ID 9856. https://doi.org/10.3390/app15189856
14. Zhang, R., Hou, J., Li, L., Zhang, K., Zhao, L., & Gao, S. (2025). RTUAV-YOLO: A Family of Efficient and Lightweight Models for Real-Time Object Detection in UAV Aerial Imagery. *Sensors*, *25*(21), article ID 6573. https://doi.org/10.3390/s25216573

А. А. Піпіч, Б. В. Булах

Національний технічний університет України "Київський політехнічний інститут ім. Ігоря Сікорського", м. Київ, Україна

АНАЛІЗ АРХІТЕКТУРНОЇ ЕВОЛЮЦІЇ ТА АПАРАТНО-ОРІЄНТОВАНИХ ВДОСКОНАЛЕНЬ ЗАСОБІВ БПЛА ДЛЯ ВИЯВЛЕННЯ ОБ'ЄКТІВ У РЕЖИМІ РЕАЛЬНОГО ЧАСУ

Проаналізовано особливості виявлення об'єктів програмно-апаратними засобами БПЛА у реальному часі у період з 2024 по 2026 роки. Досліджено критичний баланс між точністю розпізнавання об'єктів, швидкістю інференсу та енергоспоживанням за умов значних апаратних обмежень, що відповідають умовам розгортання засобів на БПЛА. Розглянуто еволюцію сучасних архітектур моделей розпізнавання об'єктів у режимі реального часу: від класичних CNN (YOLOv11) до парадигм, орієнтованих на механізми уваги (YOLOv12), та нативного NMS-free інференсу (YOLO26). Досліджено, як ці архітектури моделей комп'ютерного зору можуть стикатися із істотними інженерними викликами, такими як "важкі" архітектурні блоки, що створюють ризик переповнення пропускної здатності пам'яті на периферійних платформах базового рівня (Edge devices). Серед наявних наукових прогалин щодо виявлення об'єктів у режимі реального часу за допомогою БПЛА було виявлено дві основні проблеми. Перша пов'язана зі спостереженням, що орієнтовно для 90 % новітніх моделей (таких як YOLO12, YOLO26, TSS-YOLO та інших архітектур, оптимізованих під NMS-Free інференс) відсутні верифіковані дані про енергоефективність роботи апаратного забезпечення, якого потребує розготання такої моделі. Друга – це проблема залежності точності та швидкості розпізнавання від набору даних ("Dataset Dependency"): результати на загальних бенчмарках, таких як MS-COCO, не є безпосередньо порівнюваними зі спеціалізованими результатами аерофотознімання. Також у роботі наведено деякі спеціалізовані рішення, такі як RTUAV-YOLO та LiteCOD, розроблені для виявлення об'єктів у реальному часі на аерознімках з БПЛА. Вони використовують архітектурні особливості (зокрема динамічне генерування ваг моделей розпізнавання об'єктів та ефективну просторову увагу), щоб подолати значний 8-кратний розрив у продуктивності (TOPS) між різними рівнями апаратного забезпечення. Задля усунення зазначених розбіжностей оцінено ряд передових стратегій оптимізації архітектури моделей розпізнавання, враховуючи Парето-оптимізацію змішаної точності (PoMQ-ViT), апаратно-залежний пошук нейронних архітектур (NAS) та структурне видалення гілок (TSS-YOLO), яке згадується як дещо ефективніше за одне тільки стандартне квантування. Запропоновано здійснювати впровадження уніфікованого коефіцієнта складності (англ. *Unified Complexity Coefficient*) для вирішення проблеми нормалізації продуктивності моделей розпізнавання. Прогнозовано також зміщення фокусу уваги щодо виявлення об'єктів у режимі реального часу за допомогою БПЛА у бік адаптивних систем, орієнтованих на стиснення ("compression-first") вхідних даних, що зрештою може забезпечити для наступного покоління автономних повітряних платформ повноцінні автономність та незалежність роботи у режимі реального часу.

Ключові слова: безпілотні літальні апарати; енергоефективність обладнання для моделей розпізнавання; периферійний штучний інтелект (Edge AI); NMS-Free інференс; розпізнавання малих об'єктів.