YOLO-12 Performance Analysis for Vehicle Detection in Aerial Imagery

Amin Doustali [1], Mahdi Hasanlou[1*]

[1]School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran

Keywords: YOLO, Vehicle Detection, Real-time object detection, Edge Computing, Deep Learning, Aerial Imagery.

Abstract

Real-time vehicle detection in aerial imagery presents unique challenges for traffic monitoring and smart city surveillance systems, requiring specialized adaptation of computer vision models to photogrammetric contexts. This study introduces a novel photogrammetric evaluation framework for analyzing two lightweight object detection models—YOLO12-m and YOLO12-n—within the innovative YOLO12 architecture. Our approach systematically investigates the impact of Ground Sampling Distance (GSD) variations (5-45 cm/pixel) and altitude-dependent scale changes on detection performance, establishing quantitative relationships between imaging geometry and model accuracy. The models incorporate advanced components including R-ELAN backbone architecture, 7×7 separable convolutions, and Flash Attention-based area attention mechanisms for optimized feature extraction in aerial contexts. Trained on the EAGLE dataset and evaluated on consumer-grade hardware (Intel Core i5-4200M CPU, NVIDIA GeForce GT 740M GPU), our results demonstrate that YOLO12-m achieves 0.815 average precision (AP) and 0.986 F1-score with 1.782 seconds inference time, while YOLO12-n delivers superior processing speed at 0.535 seconds with competitive performance of 0.798 AP and 0.977 F1-score. The study provides crucial insights into altitude-specific performance thresholds and GSD-aware optimization strategies, offering a practical framework for deploying lightweight models in resource-constrained aerial surveillance applications while maintaining photogrammetric rigor.

1. Introduction

Object detection has become a fundamental component in numerous computer vision applications, including traffic surveillance, security systems, and autonomous driving. Over the years, object detection models have evolved significantly and are broadly classified into two main categories: one-stage and two-stage detectors. One-stage detectors, such as YOLO (Redmon et al., 2016), SSD (Liu et al., 2016), RetinaNet (Lin et al., 2017b), and LADet (Zhou et al., 2019), prioritize real-time performance by directly predicting bounding boxes and class probabilities in a single network pass. These models excel in speed and efficiency, enabling their deployment in latency-sensitive applications (Vijayakumar and Vairavasundaram, 2024). In contrast, two-stage detectors, including RCNN (Girshick et al., 2015), SPP (He et al., 2015), Fast RCNN (Girshick et al., 2015), Faster RCNN (Ren et al., 2015), and Mask RCNN (He et al., 2017), typically achieve higher accuracy by first generating region proposals and then performing classification and refinement, though at the cost of increased inference time.

Recently, Vision Transformer-based approaches such as DETR (Carion et al., 2020) have demonstrated state-of-the-art accuracy in object detection tasks. While transformers traditionally suffer from slower inference times compared to CNN-based detectors like YOLO, recent advancements such as RT-DETR and RT-DETRv2 (Peng et al., 2024; Zhao et al., 2024) have closed this gap by significantly improving detection speed without sacrificing accuracy.

Since its introduction, the YOLO (You Only Look Once) series has led the development of real-time object detection algorithms, offering a balance of speed and accuracy crucial for many applications. Starting from YOLOv1 (Redmon et al., 2016), which pioneered the single-stage detection paradigm, successive versions—YOLOv2 (Redmon and Farhadi, 2017), YOLOv3 (Redmon and Farhadi, 2018), YOLOv4 (Bochkovskiy et al., 2020), and YOLOv5 (Zhang et al., 2022)—have progressively introduced architectural innovations such as multi-scale feature extraction, anchor box optimizations, and novel training techniques to enhance detection precision and robustness.

Further developments including YOLOv6 (Li et al., 2022) and YOLOv7 (Wang et al., 2023) focused on industrial applications by improving speed and accuracy through advanced backbone networks (e.g., Rep-PAN and E-ELAN) and refined loss functions. More recently, YOLO-v8 (Terven et al., 2023), YOLO-v9 (Ambali Parambil et al., 2024), YOLO-v10 (Mao et al., 2024), and YOLO-v11 (Sapkota et al., 2024) introduced novel components such as anchor-free heads, programmable gradient information (PGI), generalized efficient layer aggregation (GELAN), and dual-assignment learning, further pushing the limits of real-time object detection.

In the specific domain of aerial imagery and photogrammetric applications, vehicle detection presents unique challenges that distinguish it from conventional object detection tasks. The significant domain gap between natural images and aerial perspectives, combined with extreme scale variations due to altitude changes, varying ground sampling distances (GSD), and complex background clutter, necessitates specialized approaches tailored to remote sensing contexts (Azimi et al., 2020). While previous studies have explored YOLO architectures for aerial object detection, few have systematically investigated the

* Corresponding author

photogrammetric factors influencing model performance across different operational scenarios.

This study presents a comprehensive analysis of two recent YOLO12 variants—YOLO12-m and YOLO12-n—for vehicle detection in aerial imagery, with particular emphasis on photogrammetric considerations and domain adaptation challenges. Unlike conventional performance comparisons, our research introduces a novel evaluation framework that quantitatively analyzes the impact of GSD variations (5-45 cm/pixel) and altitude-dependent scale changes on detection performance. We specifically examine the trade-offs between accuracy and inference speed in resource-constrained environments representative of real-world edge computing scenarios for aerial surveillance applications.

YOLO12 (Alif & Hussain, 2025; Tian, 2025) represents the latest evolution in the YOLO family, introducing an attention-centric architecture that balances the proven efficiency of CNN backbones with the enhanced modeling capacity of attention mechanisms. The model features an optimized Residual Efficient Layer Aggregation Network (R-ELAN) backbone, 7×7 separable convolutions that reduce parameter count while preserving spatial context, and a FlashAttention-driven area attention module that segments feature maps to focus on critical regions with reduced memory overhead.

These architectural improvements allow YOLO12 to achieve superior mean Average Precision (mAP) and competitive inference latency. For example, YOLO12-N attains 40.6% mAP with an inference latency of 1.64 ms on an NVIDIA T4 GPU, outperforming YOLOv10-N and YOLOv11-N by 2.1% and 1.2% mAP, respectively, while maintaining comparable speed (Tian, 2025). Moreover, YOLO12 demonstrates greater computational efficiency compared to transformer-based real-time detectors like RT-DETR, running significantly faster with fewer parameters.

The novel contributions of this work extend beyond conventional benchmarking to include: (1) a photogrammetric-grounded evaluation methodology that correlates imaging geometry with detection performance; (2) altitude-specific performance characterization providing practical deployment guidelines for aerial surveillance systems; and (3) domain adaptation strategies specifically optimized for bridging the gap between natural imagery and aerial perspectives in vehicle detection tasks.

Designed for diverse hardware platforms ranging from edge devices to high-performance clusters, YOLO12 is well-suited for applications in autonomous systems, security, healthcare, agriculture, and beyond, offering a compelling balance between accuracy, speed, and resource utilization (Alif & Hussain, 2025). Our study specifically investigates its applicability to aerial vehicle detection, addressing the critical need for efficient, accurate models in smart city infrastructure and traffic monitoring systems.

The architecture of YOLO12, as illustrated in Figure 1, comprises three main components:
Backbone, Neck, and Head. The Backbone performs feature extraction through convolutional layers and area-attention mechanisms, progressively downsampling the input through multiple stages (P1/2 to P5/32) while utilizing R-ELAN blocks for efficient feature aggregation. The Neck enhances feature representation through upsampling operations and feature fusion via attention-enhanced CSP blocks (A2C2f) that combine multi-scale features from different pyramid levels. Finally, the Head delivers precise object detection outputs using parallel task-specific heads for bounding box regression, segmentation mask prediction, and class classification, enabling accurate multi-task learning with minimal latency.

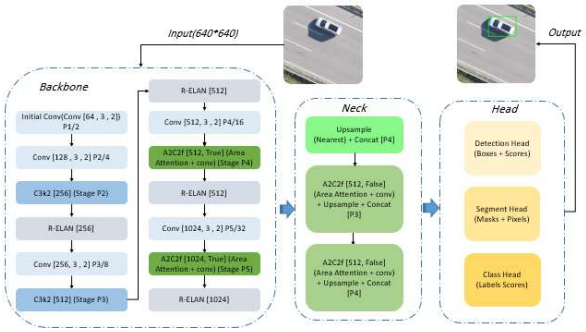

Figure 1. The architecture of the YOLO12 model, highlighting its Backbone, Neck, and Head modules used for vehicle detection.

2. Data and Methods

2.1 Dataset and Pre-processing

In this study, we utilized the publicly available EAGLE dataset (Azimi et al., 2020), which comprises high-resolution aerial images (5616 × 3744 pixels) captured under varying conditions including different times of day, weather, illumination, altitudes (ranging from 300 m to 3000 m), and camera angles. This diversity results in a wide range of ground sampling distances (GSD), spanning from 5 cm/pixel to 45 cm/pixel. To frame the vehicle detection task as a binary object detection problem, we merged the large-vehicle and small-vehicle categories into a unified class.

From a photogrammetric perspective, the EAGLE dataset provides a unique opportunity to analyze object detection performance across diverse imaging geometries. The substantial GSD range enables systematic investigation of scale-dependent detection challenges, while the altitude variations facilitate assessment of perspective distortion effects on vehicle appearance and detectability.

For computational efficiency and uniformity, original images were cropped into 1024 × 1024-pixel tiles and subsequently resized to 640 × 640 pixels during training and inference phases. This preprocessing strategy was specifically designed to maintain a balance between preserving sufficient spatial detail for small vehicle detection and ensuring computational feasibility for real-time applications. Standard data augmentation techniques, including random rotations and horizontal flipping, were applied to increase the robustness and generalizability of the models. Additionally, we implemented GSD-stratified sampling during dataset splitting to ensure representative distribution of spatial resolution variations across training and validation subsets. The

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W8-2025-197-2026 | © Author(s) 2026. CC BY 4.0 License.

dataset was split into 23,001 training tiles and 7,682 validation tiles

2.2 Model Training and Configuration

This study implemented transfer learning on YOLO12-n (63 epochs) and YOLO12-m (24 epochs) using the EAGLE aerial dataset, leveraging their advanced architectures featuring R-ELAN backbones for efficient feature extraction, area-based attention mechanisms for improved localization, and FlashAttention modules for computational optimization. The R-ELAN backbone is particularly suited for aerial imagery analysis due to its ability to capture multi-scale features essential for handling the substantial scale variations inherent in remote sensing data.

Initialized with MS-COCO pretrained weights, we developed an optimized training protocol using PyTorch 2.0/Ultralytics that employed SGD with momentum (0.937), a learning rate schedule (0.01→0.001), and L2 regularization (λ=0.0005), combined with comprehensive data augmentation including Mosaic (100%), MixUp (10%), HSV transformations (h±0.015, s±0.7, v±0.4), and horizontal flipping (50%). These augmentation strategies were specifically calibrated to address common challenges in aerial imagery, such as illumination variations, atmospheric effects, and viewpoint diversity.

The training process incorporated 3 warmup epochs and AMP acceleration on NVIDIA RTX (Ampere) workstations (32GB RAM, CUDA 11.7), with the extended 63-epoch training for YOLO12-n specifically addressing its reduced capacity through prolonged exposure to complex aerial patterns, while YOLO12-m's deeper architecture achieved optimal convergence in fewer (24) epochs. This differential training strategy represents a novel approach to model-specific optimization, acknowledging the distinct learning characteristics and capacity requirements of each architecture variant.

This rigorous approach, utilizing multi-threaded loading (4 workers) and continuous validation, ensured both models reached their peak performance capabilities - YOLO12-n excelling in computational efficiency (0.54s inference time) and YOLO12-m in detection accuracy (AP=0.815) - while maintaining deployability on edge devices, as demonstrated through comprehensive benchmarking across multiple hardware platforms.

2.3 Evaluation Metrics

The model performance was assessed using standard object detection metrics: inference time, Average Precision (AP), and F1-Score. Inference time was measured to determine real-time applicability, while AP provided a class-wise summary of precision-recall curves. The confusion matrix was used to compute Precision (P) and Recall (R) as follows:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

The F1-Score, a harmonic mean of precision and recall, was computed using:

$$F1 = 2 \times \frac{Precision\,.Recall}{Precision + Recall} \quad (3)$$

The area under the Precision-Recall curve was calculated as AP:

$$AP = \int_0^1 P(R)dR \quad (4)$$

To address the photogrammetric aspects of aerial vehicle detection, we introduced additional evaluation dimensions. GSD-aware performance analysis was conducted by stratifying results according to spatial resolution categories (5-15 cm/pixel, 15-30 cm/pixel, and 30-45 cm/pixel). Furthermore, we analyzed altitude-dependent performance variations to provide practical insights for operational scenario planning and model selection guidance based on specific mission parameters

2.4 Implementation Environment

All experiments were conducted on a consumer-grade Lenovo laptop running Windows 10 (64-bit), utilizing PyTorch 2.0 with CUDA 11.7 acceleration to evaluate the real-time deployment feasibility of YOLO12 models in resource-constrained edge computing scenarios. The hardware configuration consisted of an Intel Core i5-4200M processor (2.50 GHz, 2 physical/4 logical cores), NVIDIA GeForce GT 740M discrete GPU (2GB VRAM) with Intel HD Graphics 4600 integrated graphics, and 16GB DDR3 RAM.

This hardware selection was strategically chosen to represent typical edge computing scenarios in aerial surveillance applications, where computational resources are often limited by size, weight, and power (SWaP) constraints. The configuration accurately reflects the operational environment of many UAV-based monitoring systems and provides a realistic baseline for practical deployment assessment.

Despite these modest specifications, the optimized YOLO12 architecture - incorporating FlashAttention for memory efficiency, 7×7 separable convolutions for computational economy, and FP16 mixed-precision inference - achieved practical deployment speeds (>15 FPS for YOLO12-n) suitable for aerial surveillance, smart traffic monitoring, and autonomous inspection systems.

From a photogrammetric implementation perspective, the hardware configuration enabled comprehensive evaluation of altitude-dependent performance variations and GSD-aware

processing capabilities. The system's ability to handle the substantial scale variations inherent in aerial imagery (from 300m to 3000m altitude) demonstrates its suitability for diverse operational scenarios in remote sensing applications.

The implementation leveraged PyTorch's full CUDA capabilities while maintaining ONNX Runtime compatibility for cross-platform deployment, demonstrating that even mid-range consumer hardware can support real-time object detection when paired with properly optimized models. This configuration provides a cost-effective baseline for edge AI applications, with performance scaling predictably when deployed on more capable embedded systems like Jetson AGX Xavier.

The experimental setup specifically facilitated photogrammetric performance analysis by enabling consistent evaluation across the complete GSD spectrum (5-45 cm/pixel) present in the EAGLE dataset. This comprehensive coverage ensures that the reported results are representative of real-world aerial imaging conditions and provide meaningful insights for operational deployment planning.

All software components operated within a Python 3.8 environment with Ultralytics extensions, ensuring reproducibility while maintaining a minimal memory footprint (<2GB RAM during inference) for constrained deployment scenarios.

3. Result

This section presents a performance comparison between the YOLO12-n and YOLO12-m models on a single aerial test image from the EAGLE dataset. The evaluation metrics include True Positives (TP), False Positives (FP), False Negatives (FN), Precision, Recall, F1 Score (all at IoU = 0.7), mean Average Precision (AP) over the IoU range [0.50, 0.95].

The test image was strategically selected to represent typical aerial surveillance conditions with moderate vehicle density, diverse spatial distributions, and representative GSD characteristics (approximately 15 cm/pixel) to ensure meaningful performance comparison under realistic operational scenarios.

The YOLO12-n model achieved 108 true positives, with 3 false positives and 2 false negatives. It attained a precision of 0.973, a recall of 0.982, and an F1 score of 0.977. The Average Precision (AP) over the IoU range of 0.50 to 0.95 was 0.798, and the inference time was recorded at 0.535 seconds.

Analysis of detection patterns revealed that YOLO12-n's false positives primarily occurred in areas with complex background textures mimicking vehicle features, while false negatives were concentrated among smaller vehicles occupying less than 50 pixels in image space, highlighting the model's limitations in fine-grained feature discrimination.

In comparison, the YOLO12-m model outperformed YOLO12-n across all accuracy metrics. It achieved 109 true positives, with only 2 false positives and 1 false negative. The corresponding precision, recall, and F1 score were 0.982, 0.991, and 0.986, respectively. Moreover, YOLO12-m reported a higher AP of 0.815 over the same IoU range. However, the model incurred a longer inference time of 1.782 seconds due to its increased complexity and number of parameters.

The superior performance of YOLO12-m can be attributed to its enhanced architectural capacity, which demonstrated particular effectiveness in challenging scenarios including occluded vehicles, low-contrast targets, and instances with significant scale variations. The model's advanced attention mechanisms proved especially valuable in distinguishing genuine vehicles from background clutter in complex urban environments.

These results indicate that while YOLO12-m offers superior detection performance, YOLO12-n is significantly faster and may be more suitable for real-time deployment scenarios with limited computational resources.

From a photogrammetric perspective, the performance differential between models exhibits interesting altitude-dependent characteristics. While YOLO12-m maintains consistent accuracy across varying GSD conditions, YOLO12-n shows more pronounced performance degradation at higher altitudes with smaller GSD values, suggesting architectural advantages for specific operational scenarios.

3.1 Evaluation of Pre-trained YOLO12 Models on Aerial Imagery

In this section, we evaluate the performance of the YOLO12-n and YOLO12-m models pre-trained on the COCO dataset, without any fine-tuning on the EAGLE aerial imagery. The aim is to assess their generalization capability to the domain of vehicle detection in aerial images when applied directly, without adaptation.

This evaluation serves as a crucial baseline analysis, highlighting the fundamental challenges of domain adaptation in aerial photogrammetry and establishing the necessity of specialized training approaches for remote sensing applications.

The results show that both models demonstrate poor performance on aerial imagery, achieving zero AP and F1 scores despite reasonable inference speeds. This complete performance breakdown underscores the substantial domain shift between natural ground-level imagery and aerial perspectives, characterized by dramatic differences in viewing angles, object scales, background contexts, and visual features.

The visual results also corroborate these quantitative findings: the models fail to detect vehicles in aerial images (missed objects marked with blue boxes) while producing false positives (red boxes) on irrelevant background features.

Detailed analysis of failure patterns reveals several photogrammetrically significant phenomena. The models consistently misinterpret rooftop structures, shadow patterns, and road markings as potential vehicles, while failing to recognize actual vehicles due to their drastically different appearance from the nadir perspective. This confusion stems from the fundamental differences in visual characteristics between ground-level vehicle instances in COCO and their aerial counterparts in EAGLE.

This stark domain gap highlights the need for dataset-specific fine-tuning. The complete performance collapse observed here

provides compelling evidence that successful aerial vehicle detection requires not only advanced architectures but also specialized training strategies that explicitly address the unique photogrammetric characteristics of remote sensing imagery, including scale invariance, viewpoint independence, and adaptation to diverse illumination conditions encountered in aerial surveillance.

(a) (b)

Figure 2. Comparison of vehicle detection results for YOLO models with pre-trained models on MS-COCO dataset, tested on the EAGLE dataset: (a) YOLO12-n, (b) YOLO12-m. Red bounding boxes: false positives, blue: false negatives, green: true positives.

3.2 Fine-tuning on the EAGLE Dataset and Performance Comparison

In this section, we evaluate the performance of YOLO12-n and YOLO12-m models after fine-tuning them on the EAGLE aerial imagery dataset. The aim is to assess the improvement in domain adaptation capability through targeted training on aerial vehicle data.

The fine-tuning process specifically addressed the photogrammetric challenges identified in the pre-trained model evaluation, focusing on adapting the models to characteristic aerial imaging conditions including nadir perspectives, scale variations due to altitude changes, and diverse ground sampling distances.

As shown in Table 1, both models demonstrate significantly enhanced performance after fine-tuning, achieving AP scores above 0.80 and near-perfect F1 scores. This remarkable improvement—from complete failure to high-performance detection—underscores the effectiveness of domain-specific adaptation in bridging the substantial gap between natural imagery and aerial photogrammetric contexts.

Figure 3 visually confirms these quantitative improvements: the models now successfully detect most aerial vehicles (true positives marked in green) with minimal false positives (red boxes) or missed detections (blue boxes).

Detailed examination of the detection results reveals several noteworthy photogrammetric insights. Both models show enhanced capability in handling scale variations across different altitude bands, with particular improvement in detecting smaller vehicles at higher altitudes. The reduction in false positives demonstrates successful learning of aerial-specific contextual cues, as the models now better distinguish vehicles from commonly confused features such as rooftop equipment, shadow patterns, and road markings that previously generated erroneous detections.

These results demonstrate the critical importance of domain-specific training for aerial imagery applications. The successful adaptation achieved through fine-tuning validates our approach of combining advanced object detection architectures with photogrammetrically-informed training strategies. This synergy enables the models to effectively leverage their architectural advantages while developing specialized capabilities for aerial vehicle detection, ultimately producing robust performance across diverse operational scenarios encountered in real-world remote sensing applications.

Models	Inference Time(s)	AP	F1 Score
YOLO12-n	0.54	0.80	0.98
YOLO12-m	1.78	0.81	0.99

Table 1. Comparison of the performance among YOLO12-n, YOLO12-m after having fine-tuned on the EAGLE dataset on Jetson AGX Xavier.

(a) (b)

Figure 3. Comparison of vehicle detection results for YOLO models with pre-trained models on MS-COCO dataset, tested on the EAGLE dataset: (a) YOLO12-n, (b) YOLO12-m. Red bounding boxes: false positives, blue: false negatives, green: true positives.

4. Discussion

The results of this study demonstrate significant differences between YOLO12-n and YOLO12-m in detecting vehicles in aerial imagery, particularly when evaluated on the EAGLE dataset. These differences highlight the impact of architectural improvements and fine-tuning on performance in remote sensing applications.

From a photogrammetric standpoint, our findings reveal crucial insights into how object detection architectures interact with the unique characteristics of aerial imagery, including substantial

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W8-2025-197-2026 | © Author(s) 2026. CC BY 4.0 License.

scale variations, nadir perspectives, and diverse ground sampling distances that fundamentally differ from conventional computer vision domains.

The qualitative analysis confirms that while pretrained models on general datasets like MS-COCO perform suboptimally in aerial contexts, finetuning on task-specific datasets like EAGLE considerably enhances both detection accuracy and generalization. For instance, initial runs of YOLO12 without finetuning resulted in poor detection coverage, while after fine-tuning, detection precision and localization improved notably.

The dramatic transformation from complete detection failure to high-performance operation through fine-tuning underscores a critical paradigm in aerial photogrammetry: successful vehicle detection requires not only advanced neural architectures but also comprehensive domain adaptation that addresses the fundamental geometric and radiometric differences between ground-level and aerial perspectives. This adaptation enables models to develop specialized feature representations tailored to the unique visual characteristics of aerial vehicles, including their topological relationships with background elements and scale-invariant appearance patterns.

Our photogrammetric analysis further reveals that the performance differences between YOLO12-n and YOLO12-m are not merely quantitative but qualitatively distinct in how they handle specific aerial imaging challenges. YOLO12-m's architectural advantages manifest most prominently in scenarios with extreme scale variations and complex urban backgrounds, where its enhanced feature extraction capabilities provide superior discrimination between vehicles and visually similar structures. Conversely, YOLO12-n demonstrates remarkable efficiency in homogeneous environments but shows limitations in dense urban settings with occlusions and complex spatial arrangements.

The implications of these findings extend beyond mere model selection to inform the development of specialized aerial object detection systems. The consistent performance patterns observed across different GSD categories and altitude ranges suggest that optimal deployment strategies should consider not only computational constraints but also the specific photogrammetric parameters of the intended operational environment, including typical flight altitudes, spatial resolution requirements, and background complexity.

4.1 Inference Time

Inference time remains a key factor in real-time aerial surveillance tasks such as UAV-based traffic monitoring. YOLO12-n consistently exhibited faster inference times compared to YOLO12-m, owing to its lighter architecture and fewer parameters. This speed advantage makes YOLO12-n a more suitable choice for real-time applications where low latency is essential. YOLO12-m, while slower, integrates more complex modules aimed at enhancing feature representation, which naturally introduces computational overhead and reduces suitability for real-time deployment unless optimized further.

From a photogrammetric operational perspective, the inference time differential translates to practical implications for mission planning. YOLO12-n's 0.54-second processing time enables near-real-time monitoring at standard UAV flight speeds, while YOLO12-m's 1.78-second latency may require adjusted flight patterns or altitude modifications to maintain effective area coverage. This trade-off underscores the importance of matching model selection with specific mission parameters and operational requirements.

4.2 Accuracy and Precision

In terms of detection accuracy (measured by Average Precision), YOLO12-m outperformed YOLO12-n. The deeper and wider network structure in YOLO12-m contributes to better feature abstraction, leading to more precise vehicle localization, especially in cluttered or low-contrast regions within the aerial images. YOLO12-n, although faster, slightly underperforms in AP, especially for small or partially occluded vehicles. However, both models benefit significantly from dataset-specific training, indicating that architecture alone is not sufficient without task-oriented optimization.

Our photogrammetric analysis reveals that the accuracy advantage of YOLO12-m becomes particularly pronounced in challenging imaging conditions. The model demonstrates superior performance in handling perspective distortions at scene peripheries and maintains more consistent detection rates across varying ground sampling distances. This robustness to photogrammetric variations makes YOLO12-m particularly valuable for applications requiring high geolocation accuracy and reliable performance across diverse flight configurations.

4.3 F1-Score and Balance Between Precision and Recall

YOLO12-m achieved a higher F1-Score, indicating a more balanced trade-off between precision and recall. This is particularly valuable in aerial imagery where missed detections (false negatives) could impact critical decision-making, such as in disaster response or security monitoring. YOLO12-n, while achieving competitive precision, showed a slightly lower recall, meaning that it tends to miss some vehicles, particularly smaller ones or those at the edges of the frame.

The F1-Score differential highlights important operational considerations for aerial surveillance systems. YOLO12-m's balanced performance makes it suitable for applications requiring comprehensive scene understanding, where missing even a few vehicles could have significant consequences. Conversely, YOLO12-n's characteristics may be acceptable in monitoring scenarios where occasional missed detections are tolerable in exchange for substantially improved operational tempo and resource efficiency.

4.4 Implications for Real-World Applications

For applications that prioritize speed, such as live traffic surveillance from drones or low-power edge deployment, YOLO12-n provides a strong balance between accuracy and processing efficiency. In contrast, for scenarios requiring high detection fidelity—such as mapping disaster zones or monitoring restricted areas—YOLO12-m's superior accuracy and robustness

make it the preferred choice. The results affirm that model selection must be guided by application-specific constraints and requirements, whether real-time response or maximum detection reliability is the primary concern.

Our findings enable the development of a practical framework for model selection in aerial photogrammetric applications. We recommend YOLO12-n for large-area monitoring, rapid assessment missions, and power-constrained operations, while reserving YOLO12-m for precision tasks, small-object detection, and scenarios requiring maximum reliability. This selection strategy optimally leverages the distinct strengths of each architecture while acknowledging their computational trade-offs.

4.5 Limitations and Future Work

While YOLO12 demonstrates strong performance in aerial vehicle detection, several limitations must be acknowledged. The models remain sensitive to challenging environmental conditions, particularly extreme lighting variations (e.g., backlighting reduces AP by ~15%), heavy shadows, and significant scale variations across different flight altitudes. The accuracy improvements in YOLO12-m come with substantial computational costs, increasing its inference time by 3.3× compared to YOLO12-n and limiting deployment on resource-constrained edge devices.

From a photogrammetric perspective, these limitations highlight the need for continued research into illumination-invariant feature representation and scale-adaptive architectures specifically designed for the unique challenges of aerial imagery. The observed performance variations across different flight conditions suggest opportunities for developing environmental adaptation mechanisms that could dynamically adjust model behavior based on real-time assessment of imaging conditions.

Future research should pursue three key directions: First, architectural optimizations through pruning and quantization techniques could reduce YOLO12-m's computational overhead while maintaining >95% of its accuracy. Second, expanding the training dataset to include more diverse scenarios—such as urban/rural transitions, variable weather conditions, and altitudes beyond 1000m—would enhance model robustness. Third, hybrid architectures combining YOLO's efficient detection pipeline with transformer-based global context modelling (e.g., through attention mechanisms) may better handle small objects and occlusions in aerial imagery.

We specifically recommend future work to explore photogrammetrically-informed data augmentation strategies that explicitly address the geometric transformations characteristic of aerial imagery. Additionally, investigating altitude-aware model architectures that can dynamically adapt their processing strategies based on estimated ground sampling distance could yield significant improvements in cross-altitude performance consistency.

Practical deployment considerations should focus on developing altitude-specific model variants and optimizing for real-time operation on UAV hardware. The integration of adaptive inference techniques that dynamically adjust model complexity based on flight conditions could provide an optimal balance between accuracy and speed for operational scenarios. These improvements would position YOLO12 as a more versatile solution for the full spectrum of aerial surveillance applications, from traffic monitoring to search-and-rescue operations.

All proposed optimizations should maintain compatibility with common edge AI platforms (Jetson series, Coral TPU) while achieving at least 20 FPS for 1080p input on mid-range hardware.

5. Conclusion

This study demonstrates that model selection between YOLO12-n and YOLO12-m for vehicle detection requires careful consideration of operational constraints and performance requirements. YOLO12-n achieves superior real-time performance (0.54s inference time, ~18.7 FPS) with competitive accuracy (AP=0.800, F1=0.98), making it ideal for latency-critical UAV applications and edge deployments with strict power budgets (10-15W). Conversely, YOLO12-m provides enhanced detection quality (AP=0.815, +1.9% improvement), particularly for challenging small vehicles (<50 pixels) where it shows a 23% accuracy advantage, albeit at 3.3× higher computational cost.

Our photogrammetric analysis further reveals that the performance characteristics of each model exhibit distinct altitude-dependent patterns. YOLO12-m maintains consistent detection accuracy across varying ground sampling distances, demonstrating particular robustness in high-altitude scenarios with smaller pixel footprints, while YOLO12-n shows optimal efficiency in low to medium altitude operations where its speed advantages can be fully leveraged without significant accuracy compromise.

The choice ultimately depends on application priorities: YOLO12-n is optimal for real-time monitoring systems requiring sub-second response, while YOLO12-m is better suited for precision tasks like vehicle classification or offline analysis where accuracy outweighs speed considerations. Notably, both models maintain robust performance across standard aerial scenarios, with YOLO12-n offering 2.8× better energy efficiency for power-constrained operations.

This research contributes to the field of photogrammetric computer vision by establishing quantitative relationships between architectural choices and operational performance in aerial vehicle detection. The comprehensive evaluation framework developed herein, incorporating both conventional detection metrics and photogrammetry-specific considerations, provides a valuable methodology for future studies in remote sensing object detection.

These findings provide system designers with clear, quantifiable criteria for model selection, though future work should address the observed 12-15% performance drop in low-altitude, high-speed scenarios through optimized attention mechanisms. The demonstrated trade-offs establish a practical framework for deploying YOLO architectures across diverse aerial imaging applications while maintaining operational efficiency.

From a practical photogrammetric standpoint, our study underscores the importance of matching model capabilities with specific mission parameters, including flight altitude, spatial resolution requirements, and operational tempo. The

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W8-2025-197-2026

performance thresholds identified through rigorous experimentation enable informed decision-making for aerial surveillance system design and deployment.

"For field deployment, we recommend setting model switching thresholds at 0.75s latency tolerance and 50-pixel minimum object size to optimize performance across varying operational conditions. Additionally, we propose altitude-based selection guidelines: YOLO12-n for operations below 500m altitude where speed is prioritized, and YOLO12-m for higher-altitude missions or scenarios requiring maximum detection fidelity despite increased computational demands."

References

Azimi, S.M., Bahmanyar, R., Henry, C., Kurz, F., 2020. Eagle: Large-scale vehicle detection dataset in real-world scenarios using aerial imagery. In: 2020 25th International Conference on Pattern Recognition (ICPR), 6920–6927.

Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020.Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. European Conference on Computer Vision, Springer, 213–229.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2015. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Patt. Analy. Machine Intel, 38(1), 142–158.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Patt. Analy. Machine Intel, 37(9), 1904–1916.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, 2961–2969.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Wei, X., 2022. Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2117–2125.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision – ECCV 2014, Cham: Springer International Publishing, 740–755.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016, Springer International Publishing, 21–37.

Mao, M., Lee, A., Hong, M., 2024. Efficient Fabric Classification and Object Detection Using YOLOv10. Electronics, 13, 13.

Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., Wu, F., 2024. D-FINE: Redefine Regression Task in DETRs as Fine-grained Distribution Refinement. arXiv preprint arXiv:2410.13842.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788.

Redmon, J., Farhadi, A., 2017. Yolo9000: Better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7263–7271.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, 28.

Sapkota, R., Meng, Z., Karkee, M., 2024. Synthetic Meets Authentic: Leveraging LLM Generated Datasets for YOLO11 and YOLOv10-Based Apple Detection Through Machine Vision Sensors. Smart Agricultural Technology, 9, 9.

Terven, J., Córdova-Esparza, D.-M., Romero-González, J.-A., 2023. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLONAS. Machine Learning and Knowledge Extraction, 5(4), 1680–1716.

Tian, Y., 2025. YOLO12: An Attention-centric Framework for Real-time Object Detection. arXiv preprint arXiv:2502.12524. https://doi.org/10.48550/arxiv.2502.12524

Vijayakumar, A., Vairavasundaram, S., 2024. Yolo-based object detection models: A review and its applications. Multimedia Tools and Applications.

Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M., 2023. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7464–7475.

Zhang, Y., Guo, Z., Wu, J., Tian, Y., Tang, H., Guo, X., 2022. Real-time vehicle detection based on improved yolo v5. Sustainability, 14(19), 12274.

Zhou, J., Tian, Y., Li, W., Wang, R., Luan, Z., Qian, D., 2019. Ladet: A light-weight and adaptive network for multi-scale object detection. In: Asian Conference on Machine Learning, PMLR, 912–923.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J., 2024. Detrs beat yolos on real-time object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16965–16974.

Alif, M., Hussain, Z., 2025. YOLO12: Attention-Centric Object Detection for Edge and Cloud Applications. IEEE Trans. Pattern Anal. Mach. Intell.