Hierarchical neural network model for identifying similar objects in drone images

Dmytro Borovyk[1,*], Oleksander Barmak[1], Pawel Komada[2] and Sergii Babichev[3]

[1]Khmelnytskyi National University, 11, Institutes str., Khmelnytskyi, 29016, Ukraine
[2]Politechnika Lubelska, 38 D, Nadbystrzycka, 29016, Lublin, 20 – 618, Poland
[3]Jan Evangelista Purkyne University in Usti nad Labem, Pasteurova, 15, 400 96, Usti nad Labem, Czech Republic

Abstract

Precise and timely object detection in UAV imagery plays a vital role in modern situational awareness systems, yet deep learning models often struggle with inter-class ambiguity among visually similar objects. The problem addressed in this study is the inefficiency of standard multiclass detection models when applied to complex aerial scenes requiring fine-grained distinction. In this work, we propose a hierarchical deep learning model that restructures the detection task into a multi-level classification cascade. This architecture employs Faster R-CNN for initial object proposals, YOLO for detailed feature extraction, and the FT-Transformer for classifying combined feature vectors, allowing for targeted optimization at each level. Experiments on a dataset of over 8,000 annotated images demonstrate the approach's effectiveness. The hierarchical model achieved an overall F1 score of 94.9%, significantly outperforming the baseline non-hierarchical model's score of 92.46%. The significant conclusion of this study is that a cascaded, modular framework effectively reduces ambiguity and enhances scalability, providing a highly accurate solution for real-time operational situational awareness.

Keywords

Unmanned aerial vehicles (UAVs), similar objects, object recognition, Faster R-CNN, YOLOv11, FT-Transformer, classification, deep learning

1. Introduction

Modern situational awareness (SA) systems are essential for effective decision-making in diverse scenarios, many of which directly involve the protection of human life and property across fields such as healthcare, energy, communications, agriculture, transportation, and law enforcement. Formally, SA begins with the perception of environmental elements [1], meaning that one of the system's core capabilities is the rapid, accurate, and autonomous detection of relevant objects in UAV imagery [2]. Owing to their mobility and relatively low cost, unmanned aerial vehicles (UAVs) have become a primary source of remote sensing data [3]. Yet, acquiring UAV imagery is only the first step—its true value emerges through automated analysis, requiring systems that can process large volumes of visual information in real time [4, 5]. Such systems must reliably recognize objects critical to the task at hand, an area where deep learning models have demonstrated state-of-the-art performance in computer vision.

This study is motivated by the practical need to enhance the efficiency of object detection in UAV imagery. Leveraging deep learning for automatic detection not only reduces operators' cognitive burden but also accelerates real-time decision-making. Furthermore, these technologies lay the groundwork for fully autonomous systems capable of functioning under demanding conditions.

The central challenge addressed here is the development of a model that can recognize objects in UAV imagery with both high accuracy and speed. This task holds dual significance: theoretically, it advances

ExplAI-2025: Advanced AI in Explainability and Ethics for the Sustainable Development Goals, November 07, 2025, Khmelnytskyi, Ukraine

*Corresponding author.
dborovyk86@gmail.com (D. Borovyk); barmako@khmnu.edu.ua (O. Barmak); p.komada@pollub.pl (P. Komada); sergii.babichev@ujep.cz (S. Babichev)
0009-0001-5337-3519 (D. Borovyk); 0000-0003-0739-9678 (O. Barmak); 0000-0002-9032-9285 (P. Komada); 0000-0001-6797-1467 (S. Babichev)

the design and optimization of deep learning architectures for computer vision, while practically, it supports the creation of next-generation SA systems.

The article is structured as follows. The Related Works section reviews recent studies on military target detection in images using deep learning. The Materials and Methods section introduces a sequential classification approach of object classification and describes each level classification. The Results and Discussion section presents experimental findings that validate the effectiveness of the proposed method and compares it against existing solutions.

2. Related works

Among contemporary deep learning architectures for object classification, the most widely recognized are R-CNN, Fast R-CNN, Faster R-CNN, You Only Look Once (YOLO), Single Shot MultiBox Detector (SSD), MobileNet, and SqueezeNet [6, 7]. These approaches rely on convolutional neural networks (CNNs), which are capable of automatically extracting visual features and thus enable reliable recognition of objects in complex environments. To effectively apply CNNs in aerial image analysis, three critical aspects must be taken into account: the diversity and quality of training data, the optimization of network design (including depth, activation functions, and regularization), and the availability of sufficient computational resources.

Beyond conventional CNNs, researchers have explored alternative strategies. In [8], a multi-stage method is proposed that combines CNN and DNN models with communication signal analysis between UAVs and controllers to detect flight types. CNNs have also been applied successfully to remote sensing tasks such as land cover mapping [9], flood monitoring, agricultural crop classification using CNNs, LSTMs, and transformers, and vegetation detection using U-Net [10]. Multimodal methods integrating stereo imagery, LiDAR, radar, and audio sensing are discussed in [11]. A systematic comparison of YOLO models in [12] highlights their superior accuracy but also their reliance on larger datasets and stronger hardware.

Despite their strong accuracy, CNN-based detectors often face challenges in terms of inference speed, which limits their real-time applicability. Fast R-CNN [13] accelerates detection by generating regions of interest (ROIs) from internal feature maps, but still cannot achieve true real-time performance. Faster R-CNN [14], with its Region Proposal Network (RPN), improves efficiency and accuracy, yet remains computationally intensive. Conversely, YOLO processes the entire image in a single step by dividing it into a grid and predicting bounding boxes and object categories simultaneously—allowing considerably higher speed while maintaining competitive accuracy.

Recent improvements in object detection have focused on the problem of identifying small-scale targets in UAV imagery. For instance, HSP-YOLOv8 [15] enhances performance for small objects through an added prediction head and SPD-Conv module, resulting in an 11% accuracy improvement compared with YOLOv8s on the VisDrone2019 dataset [16]. A tailored YOLOv7 [17] addresses UAV-specific challenges such as variable scales, dense clusters, and uneven target distribution, achieving higher detection accuracy with reduced computational costs. Similarly, UN-YOLOv5s [18] introduces MASD and MCF mechanisms, boosting mAP by 8.4% on VisDrone2019.

Other studies emphasize environmental influences on UAV detection. For example, [18] investigates how background complexity and atmospheric effects, such as rainfall, reduce recognition accuracy, while [19, 20] explore deep learning solutions for real-time detection, localization, and segmentation in UAV video streams.

More recently, transformer-based architectures have attracted significant attention. Vision Transformer (ViT) [21] leverages self-attention mechanisms applied to image patches to learn rich feature representations. Data-efficient transformers (DeiT) [22] reduce training requirements for smaller datasets, while Perceiver [23] can integrate multimodal data such as feature vectors, making it suitable for classification under limited input conditions. TabTransformer [24] is adapted for tabular data, effectively encoding categorical and numerical attributes. Hybrid designs, such as Swin Transformer [25] and ConvNeXt, combine convolutional layers with attention mechanisms to balance speed and accuracy.

These approaches demonstrate the flexibility of transformers in capturing long-range dependencies and modeling complex interactions, particularly valuable in UAV imagery analysis and multimodal data fusion [26].

Building on this analysis, we propose a hybrid architecture combining YOLOv11 [27] for detection, Faster R-CNN for feature vector extraction, and FT-Transformer [28] for classification. Two research hypotheses are introduced: (1) a multi-level framework with separately trained models on specialized datasets increases efficiency and accuracy; (2) transformer-based architectures, adapted for structured tabular data, can effectively classify CNN-derived feature vectors.

Hierarchical classification distributes recognition tasks across levels, each handling a limited set of classes, thereby reducing ambiguity, enhancing scalability, and enabling the system to expand through additional levels when necessary. The proposed architecture—YOLOv11 for detection, Faster R-CNN for structured feature extraction, and FT-Transformer for classification—helps prevent information overload and improves recognition accuracy.

3. Materials and methods

3.1. Description of proposed approach

The essence of the proposed approach to improving the efficiency of object classification in UAV imagery lies in constructing a multi-level structure that classifies objects step by step, gradually refining their classification into specific classes. At each level, classification is carried out within a limited number of classes, which makes it possible to extract features characteristic of specific objects more accurately than in the case of a single large multiclass model.

A key feature of the proposed approach is that two deep learning models are used at each level instead of one. The first model is responsible for extracting object features and constructing their vector representation based on a certain principle. The second model, in turn, is used directly for classifying the detected objects based on the results of the first model.

A schematic representation of the proposed approach is shown in Figure 1. This approach is scalable and flexible, as it allows new classification levels to be easily added without the need to retrain the entire system.

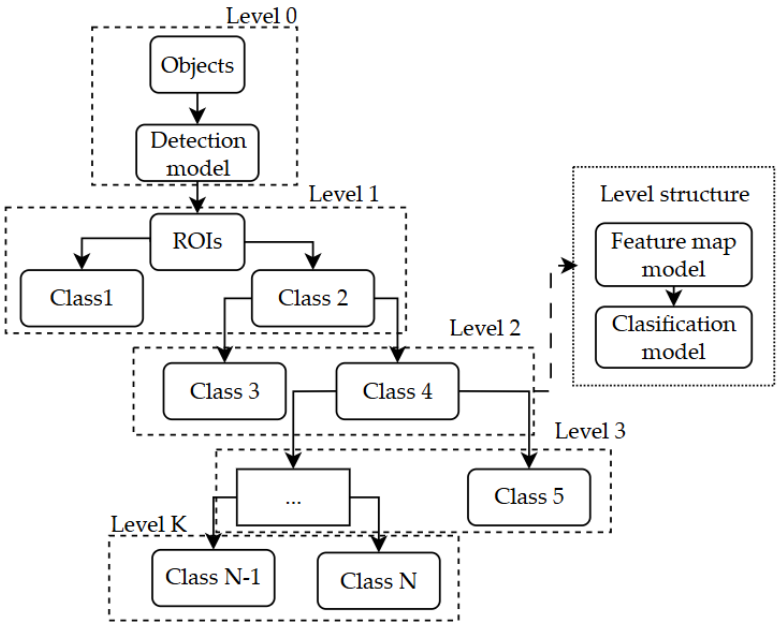


Figure 1: A separate possible architecture variant for cascading multi-level step-by-step object classification in UAV images.

At each classification level, two separate deep learning models are used: the first model (Feature map model) extracts object features and constructs their feature vector according to a certain principle, while the second model (Classification model) performs the direct classification of the detected objects based on the results of the first model.

3.2. Cascaded multi-level model for stepwise classification of objects in UAV imagery

In automated tasks of target detection and classification from UAV images, structured multi-level processing plays a crucial role, as it allows for the step-by-step refinement of object classes while maintaining high accuracy at each stage. This approach prevents excessive feature dispersion when a single model is applied to multiclass classification and enables effective system expansion through the addition of new levels.

The processing of UAV-acquired images can be formalized as a set-theoretic mathematical model that describes the sequential transformation of input data into structured output object labels.

The essence of this set-theoretic mathematical model is as follows. Let the input image be denoted as element x from the set of all possible UAV images X. The goal of processing is to construct an ordered set of object regions with corresponding classes.

At the first stage, object detection is performed. This is formalized as a function D that maps an element x from the set X to a subset $\{r_1, r_2, \ldots, r_k\}$ of the set R, where R is the set of possible object regions in an image.

The next step is constructing feature vectors $\{f_1, f_2, \ldots, f_k\}$ for each detected region using a function F, which maps each element of the set R to a vector in the space $\mathbb{R}^n$. The obtained vectors are then passed to classifier C, which is a deep learning model trained on feature vectors produced by the Feature map model, and which classifies each vector into one of the classes considered at the corresponding level.

Suppose that at a given classification level the recognized classes are $A_1, A_2, \ldots, A_s$. Then this transformation can be expressed as:

$$\{r_1, r_2, \ldots, r_k\} \xrightarrow{F} \{f_1, f_2, \ldots, f_k\} \xrightarrow{C} \{c_1, c_2, \ldots, c_k\}, \tag{1}$$

where each c_i belongs to the set of classes $\{A_1, A_2, \ldots, A_s\}$.

Further refinement is performed if certain classes at level k have subsequent classification levels. For example, if a class A_γ has a further classification stage, then the refinement of each corresponding object r_i, such that $c_i = A_\gamma$, can be expressed as:

$$c_i \xrightarrow{g} c_i', \tag{2}$$

where each c_i' belongs to the set of classes at that next classification level.

Generalizing this, a global function Φ can be defined that maps each input image to a set of ordered pairs consisting of an object's coordinates and its final class label. In other words, function Φ maps the set X to a subset of the Cartesian product $R \times C$, where C is the set of all final classes:

$$\Phi(x) = \{(r_1, c_1), (r_2, c_2), \ldots, (r_k, c_k)\}, \tag{3}$$

where each c_i is the final classification label for the corresponding region r_i.

The generalization of this process for an arbitrary number of cascade levels n allows it to be described as a hierarchical composition of functions. Let for each level $i \in \{1, \ldots, n\}$, F_i be the feature extraction function and C_i the classifier at that level. Then the generalized mapping of function Φ can be represented as:

$$\Phi(x) = \{(r_i, c_i) \mid r_i \in D(x), c_i = \Gamma(r_i)\}, \tag{4}$$

where $\Gamma : R \to C$ is the cascade classification function, defined recursively as:

$$\Gamma(r) = \begin{cases} C_1(F_1(r)), & \text{if } C_1 \in C_{last}; \\ C_2(F_2(r)), & \text{if } C_1 \notin C_{last} \text{ and } C_2 \in C_{last}; \\ C_n(F_n(r)), & \text{if } C_{n-1} \notin C_{last} \text{ and } C_n \in C_{last}, \end{cases} \tag{5}$$

where C_{last} represents the set of final classes.

Thus, the cascaded multi-level model enables consistent stepwise refinement of class assignments for each object, ensuring high classification accuracy even with a large number of classes. Moreover, it is easily scalable: to extend the system, it is sufficient to add a new layer with corresponding functions F_{n+1} and C_{n+1}, without modifying the previous levels.

3.3. Method for selecting a deep learning model for target feature vector extraction

At the core of the cascade multi-level step-by-step classification architecture for objects in UAV images lies the sequential extraction of features from object regions previously identified by detection methods. The effectiveness of feature vector extraction largely determines the success of subsequent classification, since at each processing level the model must highlight the characteristics that allow distinguishing targets of different types. The feature vector formed for each object represents spatial, contextual, textural, and morphological information captured in the image.

All input data, when passed through the convolutional block, are represented in the form of feature vectors. Feature vectors are extracted from each convolutional layer and form the set of vectors F in the space $\mathbb{R}^n$. For feature extraction and vector construction, we use only the convolutional block Feature map model from the architecture shown in Figure 1.

In the proposed architecture (Figure 1), each cascade level is responsible for extracting features relevant only to a specific subtask. This separation prevents a single model from being overloaded with too many target classes, which often leads to dispersion in the feature vector space and reduced classification quality. In contrast, highly specialized models focusing on a small number of classes can form more expressive feature vectors with a higher inter-class distance.

At each cascade layer, its own optimized deep learning model is applied for feature extraction. The model selection depends not only on the level of classification detail but also on the size of objects, their typical positions in images, and computational resource constraints. The most important factor is the alignment of the receptive field scale with the expected object sizes at the corresponding level.

It should be noted that due to the cascade approach to classifying detected targets, each classification layer operates independently. Therefore, different layers may use different models for object feature extraction and different sequences of feature vector construction, which will then be used for direct classification.

Moreover, to construct the feature vector, features from either a single level or combinations of feature vectors from different levels can be used (Figure 2). This provides a clear separation between classes even in cases of high object density or complex background, which is critically important for UAV combat applications in real-time conditions.

Thus, the task is to obtain the optimal feature vector F for object r_i by concatenating vectors extracted from specific convolutional layers. This task can be formalized as follows:

$$\Phi^* = \arg \max_{concat \subset S} \left(\lim_{i,j \to \infty} \min_{c_i, c_j} |f_i - f_j| \right), \tag{6}$$

where S is the set of objects, c_i is the class of object i, and f_i is the constructed feature vector of object i.

4. Results and discussion

4.1. Dataset

To train the models responsible for feature extraction at different classification levels, the overall dataset was divided into subsets aligned with the subclasses corresponding to each recognition stage. As a basis, we used the publicly available VisDrone2019 dataset [16], which is widely adopted in aerial object detection research. It includes more than 8,000 UAV-captured images with detailed annotations of object classes and bounding boxes, covering categories such as buses, trucks, cars, vans, and others. The dataset provides high-resolution imagery with precise bounding box labels for each object.

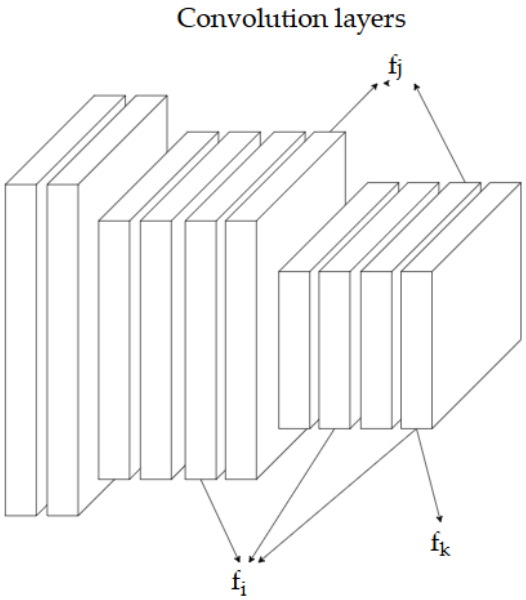


Figure 2: General structure of feature vector concatenation.

For each classification level, the prepared subsets were randomly split three times at the sequence level, following an 80% training and 20% testing ratio.

4.2. Experiment results

To evaluate the effectiveness of the proposed method, a series of experiments was conducted and the outcomes compared. Within the study, a sequential classification system was implemented for recognizing objects in UAV-acquired images and videos. The primary objective of the experiments was to measure the system's performance using Precision, Recall, and F1-score metrics. Additionally, the results of the developed approach were benchmarked against existing solutions to the same problem.

To demonstrate the proposed approach, the following example is provided. Drawing on empirical observations, a three-level target classification sequence for UAV imagery is suggested (Figure 3).

Each model in the cascaded classification system was initially trained on the datasets corresponding to the classification level it was intended to handle.

Table 1
Calculation of the average metrics to all levels in percents.

Level	DS	Precision, %	Recall, %	F1-score, %	mAP@.50, %	mAP@.50:.95, %
1	Train	95.5	94.5	94.4	94.1	84.2
	Test	94.0	93.8	93.5		
2	Train	93.6	96.1	95.7		
	Test	92.7	94.8	94.2		
3	Train	95.3	96.9	95.7		
	Test	94.5	95.9	93.9		
Avg	Train	94.8	95.8	95.2		
	Test	93.7	94.8	93.8		

Table 1 above also presents the values of mAP@.50:.95 and mAP@.50 metrics, which were used to evaluate the effectiveness of the entire cascaded classification system and to mitigate the impact of cumulative error propagation through the cascade architecture on other accuracy metrics.

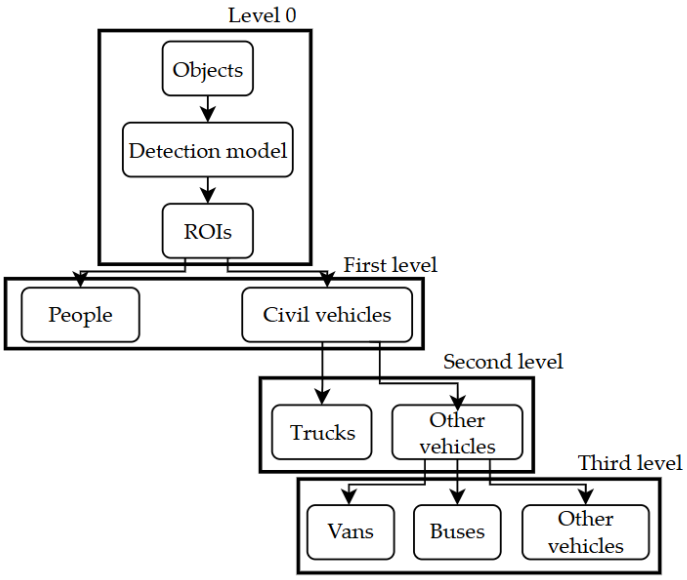


Figure 3: Proposed test classification structure.

Figure 4 illustrates the process of sequential object classification within aerial imagery, highlighting how a hierarchical recognition framework refines object detection across multiple levels. The central image, captured from a drone or elevated viewpoint, shows a roadway scene populated with various entities such as vehicles, trucks, vans, and pedestrians. Each object in the scene is enclosed within a bounding box, with distinct colors representing the assigned class labels.

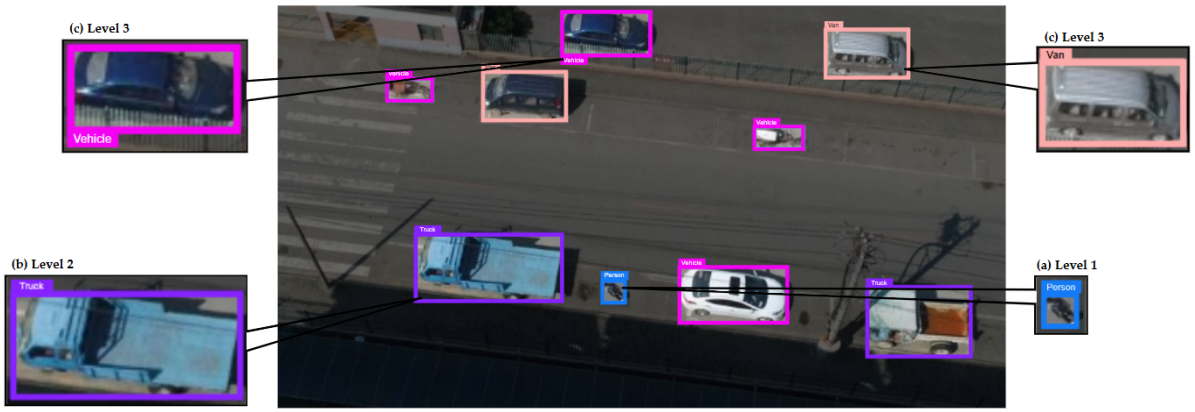


Figure 4: Example of classification results.

To better demonstrate the classification hierarchy, selected objects are enlarged and annotated at three distinct levels. At Level 1, the system identifies relatively simple objects with clear, well-defined shapes, such as a pedestrian crossing the street. This level demonstrates the model's ability to detect smaller and more isolated targets, where precision in locating a human figure is of particular importance for safety-critical applications such as traffic monitoring or autonomous navigation.

Moving to Level 2, the classification system begins to distinguish larger and more complex objects, such as trucks. Here, the bounding box captures the full extent of the vehicle, emphasizing not only its presence but also its categorization as a specific type of road user. This stage shows the system's capacity to handle variations in object scale, perspective, and partial occlusion while still maintaining accurate labeling.

Finally, at Level 3, the recognition framework demonstrates its ability to provide even finer distinctions within the general "vehicle" category. For example, two separate vehicles are refined into the

subcategories of “Van” and “Vehicle,” showing the granularity achievable in higher-level classification. These refinements are crucial in real-world applications where decision-making depends on differentiating between types of vehicles—for instance, distinguishing commercial vans from private cars in traffic analytics, or differentiating emergency vehicles from standard ones.

Overall, the figure emphasizes the progressive nature of hierarchical object classification. Rather than providing a single-level detection, the system incrementally enhances recognition from general object identification (e.g., “person” or “vehicle”) toward more detailed and context-sensitive categorization (e.g., “van” or “truck”). This layered approach reflects a more human-like perception process, where understanding a scene often begins with broad identification before narrowing to specific details. Such a framework is particularly well-suited for surveillance, intelligent transportation systems, and smart city applications, where both accuracy and contextual understanding of different object classes are required.

The confusion matrices in Figure 5 offer further insights. For each classification level, two matrices are displayed, reflecting the use of two distinct deep learning models at every stage.

Additionally, to directly assess the quality of training and classification, experiments were performed across all three levels using the COCO dataset [29], which was not involved in the training process (Table 2). This dataset is commonly used for training deep learning models to recognize basic object categories.

Table 2
Test results on COCO dataset.

Level	Precision, %	Recall, %	F1-score, %	mAP@.50, %	mAP@.50:.95, %
1	94.2	93.4	94.0	93.1	82.2
2	93.8	95.8	94.2		
3	94.8	96.5	95.2		
Avg	94.2	95.2	94.4		

When testing the proposed approach on an independent dataset COCO [29], the obtained results were slightly lower compared to those achieved on the FECL dataset [30]. This difference can be attributed to several factors, including variations in image resolution, object scale, environmental conditions, and annotation style between datasets. Such performance degradation is a common phenomenon in machine learning, as models often demonstrate higher accuracy on data they were trained on, while generalization to new and unseen data introduces additional challenges. Nevertheless, the approach maintained a stable detection capability, showing that it is not overfitted exclusively to the training data and can still effectively recognize objects across different environments. The slight decrease in performance highlights the importance of evaluating models under diverse real-world scenarios to ensure robustness and adaptability.

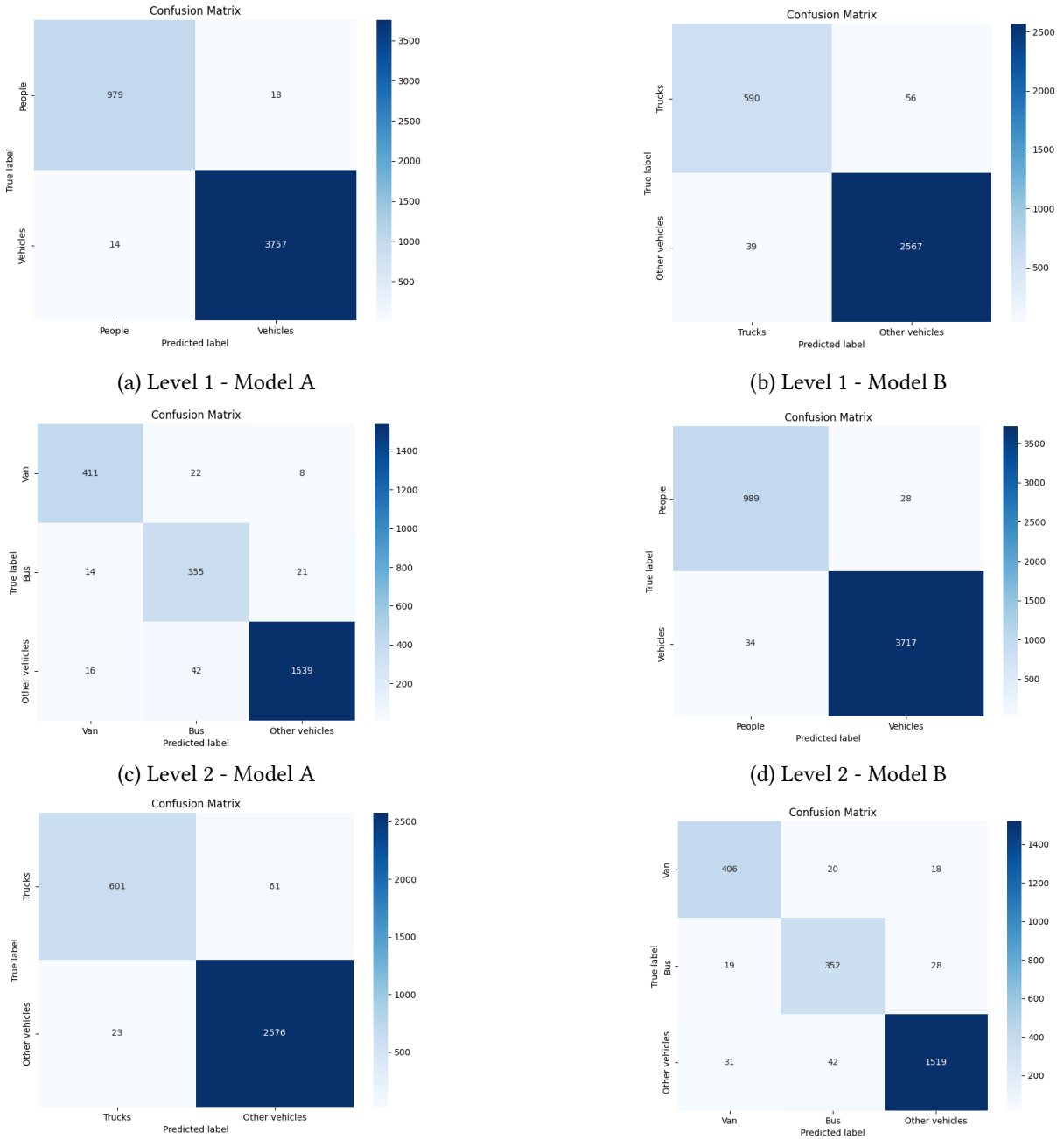
In addition, the proposed approach was also applied to the task of detecting military targets. Although military target recognition represents a distinct research domain with its own challenges—such as camouflage, irregular object shapes, and diverse environmental conditions—the model demonstrated promising results. This outcome suggests that the developed method is not limited to civilian traffic analysis but can be effectively transferred to other application areas. The ability to adapt to such a specialized context highlights the robustness and versatility of the approach, opening possibilities for its further use in defense-related surveillance and reconnaissance tasks.

Figure 6 describes classification sequence structure for detection and classification of military targets.

The Figure 7 shows how a hierarchical recognition framework refines object detection across multiple levels for military targets detection. Each object in the scene is enclosed within a bounding box, with distinct colors representing the assigned class labels.

Table 3 represents metric results for military targets classification task.

As illustrated in Table 3, the proposed approach demonstrates its suitability for different tasks beyond the initial training scenario. The performance metrics presented in the table indicate that the method



(a) Level 1 - Model A

(b) Level 1 - Model B

(c) Level 2 - Model A

(d) Level 2 - Model B

(e) Level 3 - Model A

(f) Level 3 - Model B

Figure 5: Confusion matrices by levels, showing performance for the two models used at each stage of the hierarchy: Level 1 (a, b), Level 2 (c, d), and Level 3 (e, f).

achieves consistently good results, even when applied to datasets with varying characteristics and object types. Such outcomes confirm that the approach is not narrowly tailored to a single domain but possesses the flexibility to be adapted to multiple application areas. This adaptability underscores its potential for broader use in both civilian and military contexts, where reliable object detection under diverse conditions is essential.

In order to evaluate the effectiveness of the proposed approach, a comparative analysis with existing methods was conducted. The results of this comparison are summarized in Table 4, where the performance of our model is presented alongside that of previously reported approaches. As shown, the proposed method achieves competitive results, in several cases outperforming the existing techniques, which highlights its advantages in terms of accuracy and adaptability. This comparison further validates

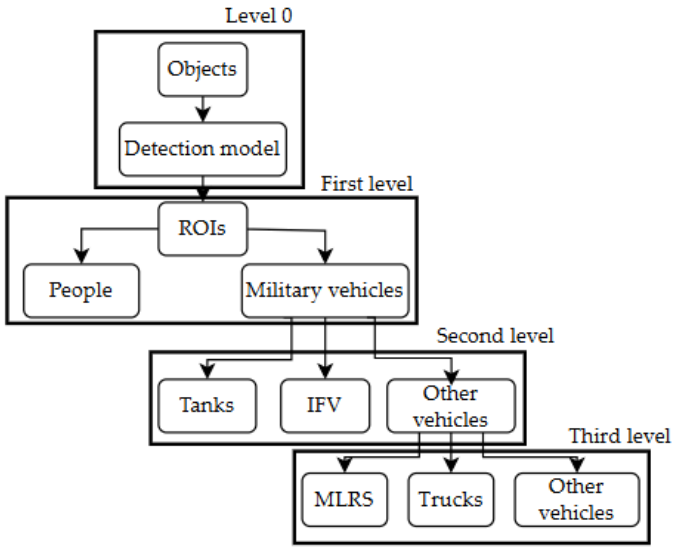
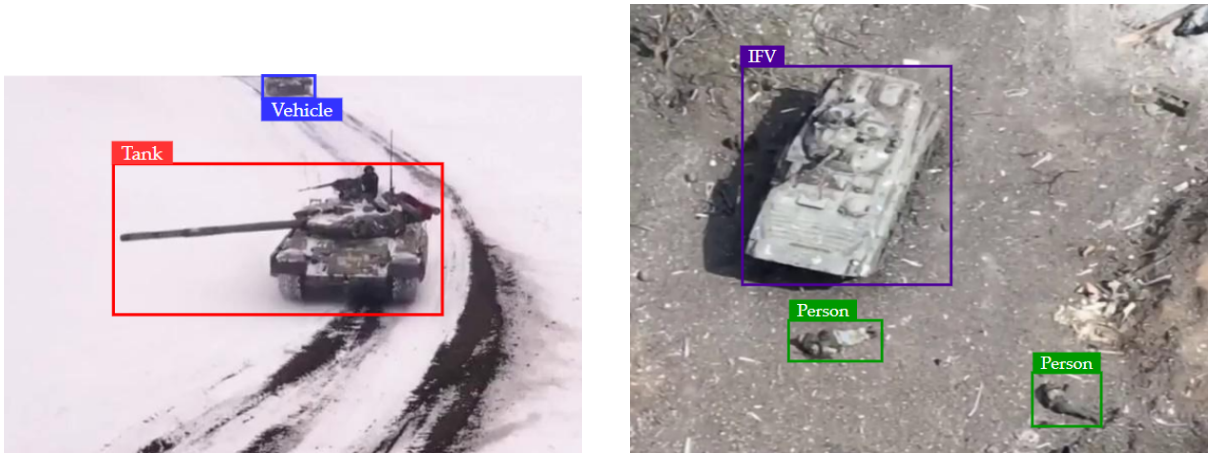


Figure 6: Example of the three-level stepwise object classification architecture in UAV imagery for military targets.



(a) Tank detection

(b) IFV and Person detection

Figure 7: Examples of classification of military targets.

the robustness of the approach and its potential for broader application in diverse object detection tasks.

5. Conclusions

In this study, we developed a method aimed at improving the accuracy of military object classification using aerial images and video materials obtained from unmanned aerial vehicles (UAVs) in real time under conditions of limited computational resources. Alongside the method, a complete system was designed to implement it effectively. The proposed system relies on a multi-level architecture that integrates modern deep learning approaches for both detection and classification. In particular, YOLOv11 and Faster R-CNN were employed for the detection of objects and the extraction of their feature vectors, while the FT-Transformer model was used to perform classification based directly on these extracted features. Such a design enables not only the detection of objects but also their accurate categorization, even in complex operational scenarios.

An extensive experimental evaluation was conducted. Within this experiment, a three-level classification pipeline was implemented. This multi-level approach reflects the complexity of real-world military

Table 3
Calculation of the average metrics to all levels for military target detection in percents.

Level	DS	Precision, %	Recall, %	F1-score, %	mAP@.50, %	mAP@.50:.95, %
1	Train	94.5	95.1	94.9	94.5	86.5
	Test	93.6	93.9	93.6		
2	Train	95.3	96.3	95.6		
	Test	93.7	95.1	95.1		
3	Train	95.1	95.7	95.7		
	Test	94.8	94.5	93.8		

Table 4
Comparation with existing approaches.

Approach	Precision, %	Recall, %	F1-score, %	mAP@.50, %	mAP@.50:.95, %
Proposed approach	**94.5**	**95.1**	**94.9**	**94.5**	**86.5**
Existing method 1 [31]	91.4	91.0	91.4	90.8	80.2
Existing method 2 [32]	94.1	93.7	93.9	92.1	83.2

recognition tasks, where different levels of granularity are required depending on the operational context. The experimental results clearly demonstrate the effectiveness of the proposed method. Across all classification levels, the system achieved high performance, with Precision, Recall, and F1-score exceeding 94%. Importantly, the architecture was optimized to ensure fast data processing, allowing the system to operate in real-time conditions—an essential requirement for time-sensitive military applications. The analysis also revealed that the approach remains stable when applied to large-scale and diverse datasets, ensuring robustness and adaptability under varying circumstances.

In addition to the internal evaluation, a comparative analysis with state-of-the-art methods for object detection and classification was carried out. This comparison confirmed the competitiveness of the proposed solution, particularly in terms of recognition accuracy and processing speed. As highlighted in our results, the method not only excelled on the VisDrone dataset but also demonstrated strong generalization on the COCO dataset and specific military targets, maintaining high mean Average Precision (mAP) scores (Table 4). The balance between efficiency and accuracy positions the system as a strong alternative to existing methods, with distinct advantages for real-world scenarios. Overall, the developed system can be regarded as an effective and practical tool for automatic detection and classification of military objects. It has potential applications in real-time battlefield monitoring, operational situational awareness, and decision support, where both accuracy and speed are critical. Beyond its immediate application, the system's modular architecture also allows for future extensions. Further research may focus on enhancing resilience to variable imaging conditions such as weather or illumination changes, integrating the system with other artificial intelligence technologies, and adapting the classification framework to new categories of emerging military equipment. Such improvements would expand the scope of applicability and further strengthen the role of AI-driven methods in modern defense and security contexts.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o and Grammarly in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. R. Endsley, Designing for Situation Awareness, CRC Press, Boca Raton, 2016. doi:10.1201/b11371.

[2] H. Liu, Y. Yu, S. Liu, W. Wang, A military object detection model of UAV reconnaissance image and feature visualization, Applied Sciences 12 (2022) 12236. doi:10.3390/app122312236.

[3] G. Tang, J. Ni, Y. Zhao, Y. Gu, W. Cao, A survey of object detection for UAVs based on deep learning, Remote Sensing 16 (2023) 149. doi:10.3390/rs16010149.

[4] S. Liu, H. He, Z. Zhang, Y. Zhou, LI-YOLO: An object detection algorithm for UAV aerial images in low-illumination scenes, Drones 8 (2024) 653. doi:10.3390/drones8110653.

[5] X. Zhao, Y. Chen, YOLO-DroneMS: Multi-scale object detection network for unmanned aerial vehicle (UAV) images, Drones 8 (2024) 609. doi:10.3390/drones8110609.

[6] D. Yan, G. Li, X. Li, et al., An improved faster R-CNN method to detect tailings ponds from high-resolution remote sensing images, Remote Sensing 13 (2021) 2052. doi:10.3390/rs13112052.

[7] J. Zhang, Y. Tang, J. Qian, et al., HR-YOLOv8: A crop growth status object detection method based on YOLOv8, Electronics 13 (2024) 1620. doi:10.3390/electronics13091620.

[8] T. Lu, L. Wan, S. Qi, M. Gao, Land cover classification of UAV remote sensing based on transformer–CNN hybrid architecture, Sensors 23 (2023) 5288. doi:10.3390/s23115288.

[9] H. Munawar, F. Ullah, S. Qayyum, A. Heravi, Application of deep learning on UAV-based aerial images for flood detection, Smart Cities 4 (2021) 1220–1243. doi:10.3390/smartcities4030065.

[10] I. Teixeira, R. Morais, J. Sousa, A. Cunha, Deep learning models for the classification of crops in aerial imagery: a review, Agriculture 13 (2023) 965. doi:10.3390/agriculture13050965.

[11] R. Pierdicca, L. Nepi, A. Mancini, E. Malinverni, M. Balestra, UAV4TREE: deep learning based system for automatic classification of tree species using RGB optical images obtained by an unmanned aerial vehicle, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences X-1/W1 (2023) 1089–1096. doi:10.5194/isprs-annals-X-1-W1-2023-1089-2023.

[12] A. Munir, A. Siddiqui, S. Anwar, et al., Impact of adverse weather and image distortions on vision-based UAV detection: a performance evaluation of deep learning models, Drones 8 (2024) 638. doi:10.3390/drones8110638.

[13] Z. Liu, P. An, Y. Yang, et al., Vision-based drone detection in complex environments: a survey, Drones 8 (2024) 643. doi:10.3390/drones8110643.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020). URL: https://arxiv.org/abs/2010.11929.

[15] H. Zhang, W. Sun, C. Sun, R. He, Y. Zhang, HSP-YOLOv8: UAV aerial photography small target detection algorithm, Drones 8 (2024) 453. doi:10.3390/drones8090453.

[16] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, H. Ling, Detection and tracking meet drones challenge, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 7380–7399. doi:10.1109/TPAMI.2021.3119563.

[17] Z. Bai, X. Pei, Z. Qiao, G. Wu, Y. Bai, Improved YOLOv7 target detection algorithm based on UAV aerial photography, Drones 8 (2024) 104. doi:10.3390/drones8030104.

[18] J. Guo, X. Liu, L. Bi, H. Liu, H. Lou, UN-YOLOv5s: a UAV-based aerial photography detection algorithm, Sensors 23 (2023) 5907. doi:10.3390/s23135907.

[19] X. Wei, L. Yin, L. Zhang, F. Wu, DV-DETR: improved UAV aerial small target detection algorithm based on RT-DETR, Sensors 24 (2024) 7376. doi:10.3390/s24227376.

[20] X. Luo, Y. Wu, F. Wang, Target detection method of UAV aerial imagery based on improved YOLOv5, Remote Sensing 14 (2022) 5063. doi:10.3390/rs14195063.

[21] M. Rahman, M. Sejan, M. Aziz, et al., A comprehensive survey of unmanned aerial vehicles detection and classification using machine learning approach: challenges, solutions, and future directions, Remote Sensing 16 (2024) 879. doi:10.3390/rs16050879.

[22] Y. Mo, J. Huang, G. Qian, Deep learning approach to UAV detection and classification by using compressively sensed RF signal, Sensors 22 (2022) 3072. doi:10.3390/s22083072.

[23] H. Touvron, M. Cord, M. Douze, et al., Training data-efficient image transformers & distillation through attention, in: Proceedings of the International Conference on Machine Learning (ICML), 2021, pp. 1–22. URL: https://arxiv.org/abs/2012.12877.

[24] A. Jaegle, F. Gimeno, A. Brock, et al., Perceiver: General perception with iterative attention, in: Proceedings of the International Conference on Machine Learning (ICML), 2021, pp. 1–43. URL: https://arxiv.org/abs/2103.03206.

[25] X. Huang, A. Khetan, M. Cvitkovic, Z. Karnin, TabTransformer: tabular data modeling using contextual embeddings, arXiv preprint arXiv:2012.06678 (2020). URL: https://arxiv.org/abs/2012.06678.

[26] S. Svystun, O. Melnychenko, P. Radiuk, O. Savenko, A. Sachenko, A. Lysyi, Thermal and RGB images work better together in wind turbine damage detection, International Journal of Computing 23 (2024) 526–535. doi:10.47839/ijc.23.4.3752.

[27] Z. Liu, Y. Lin, Y. Cao, et al., Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the International Conference on Computer Vision (ICCV), 2021, pp. 1–14. URL: https://arxiv.org/abs/2103.14030.

[28] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 548–558. doi:10.1109/ICCV48922.2021.00061.

[29] T. Lin, M. Maire, S. Belongie, et al., Microsoft COCO: Common objects in context, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2014, pp. 740–755. doi:10.1007/978-3-319-10602-1_48.

[30] MVDdetection, Fecl dataset, 2025. URL: https://universe.roboflow.com/mvddetection/mv_detection-fecl.

[31] X. Zhao, W. Zhang, Y. Xia, et al., G-YOLO: A lightweight infrared aerial remote sensing target detection model for UAVs based on YOLOv8, Drones 8 (2024) 495. doi:10.3390/drones8090495.

[32] X. Du, L. Song, Y. Lv, S. Qiu, A lightweight military target detection algorithm based on improved YOLOv5, Electronics 11 (2022) 3263. doi:10.3390/electronics11203263.