applied sciences

MDPI

Article

A Comparative Study of Convolutional Neural Network and Transformer Architectures for Drone Detection in Thermal Images

Gian Gutierrez [1], Juan P. Llerena [1,2], Luis Usero [2] and Miguel A. Patricio [1,*]

[1] Applied Artificial Intelligence Group, Computer Science and Engineering Department, Universidad Carlos III de Madrid (ROR Code 03ths8210), 28270 Colmenarejo, Madrid, Spain; jp.llerena@uah.es (J.P.L.)
[2] Cognitive Science Research Group, Universidad de Alcalá, 28805 Alcalá de Henares, Madrid, Spain; luis.usero@uah.es
* Correspondence: mpatrici@inf.uc3m.es

Abstract: The widespread growth of drone technology is generating new security paradigms, especially with regard to the unauthorized activities of UAVs in restricted or sensitive areas, as well as illegal and illicit activities or attacks. Among the various UAV detection technologies, vision systems in different spectra are postulated as outstanding technologies due to their peculiarities compared to other technologies. However, drone detection in thermal imaging is a challenging task due to specific factors such as thermal noise, temperature variability, or cluttered environments. This study addresses these challenges through a comparative evaluation of contemporary neural network architectures—specifically, convolutional neural networks (CNNs) and transformer-based models—for UAV detection in infrared imagery. The research focuses on real-world conditions and examines the performance of YOLOv9, GELAN, DETR, and ViTDet in different scenarios of the Anti-UAV Challenge 2023 dataset. The results show that YOLOv9 stands out for its real-time detection speed, while GELAN provides the highest accuracy in varying conditions and DETR performs reliably in thermally complex environments. The study contributes to the advancement of state-of-the-art UAV detection techniques and highlights the need for the further development of specialized models for specific detection scenarios.

Keywords: unmanned aerial vehicles (UAVs); convolutional neural networks (CNNs); transformers (TNNs); thermal images

Academic Editors: Jing Jin and Pedro Couto

Received: 22 October 2024
Revised: 21 November 2024
Accepted: 24 December 2024
Published: 27 December 2024

Citation: Gutierrez, G.; Llerena, J.P.; Usero, L.; Patricio, M.A. A Comparative Study of Convolutional Neural Network and Transformer Architectures for Drone Detection in Thermal Images. *Appl. Sci.* **2025**, *15*, 109. https://doi.org/10.3390/app15010109

1. Introduction

In recent years, the use of unidentified Unmanned Aerial Vehicles (UAVs), commonly known as drones, has experienced exponential growth, becoming key tools in various industries such as agriculture [1], filmmaking [2,3], surveillance [4,5], and package delivery [6,7]. However, this widespread adoption has introduced new challenges, particularly in the fields of security and privacy. Drones can be great allies to remote populations by delivering goods [8] or collecting/delivering medical samples [9], but they can also be used for malicious activities such as trespassing [10], unauthorized data collection [11,12], or smuggling [13]. These illegal and/or illicit activities can endanger critical infrastructure such as airports, as well as aviation security as a whole, and ultimately people. These new security/vulnerability paradigms make the detection and neutralization of UAVs a priority for the authorities of all countries in charge of border surveillance, aviation security, or defense.

Detecting these vehicles is a big challenge given the wide variability of these aircraft, which poses challenges to detection systems based on size, speed, altitude, dynamic be-

havior, and even environmental conditions [14,15]. To address the number of challenges in UAV detection, there are four main approaches: radar-based detection [16–18], radio frequency-based detection [19,20], acoustic detection [21,22], and vision-based detection [23–27]. Each of these has advantages and disadvantages, which are discussed in reviews such as Seidaliyeva et al. [14]. Other works such as [28–30] focus on the advantages of sensor fusion to provide robust systems and to compensate for the shortcomings of some methods with the advantages of others, but each of the mentioned approaches includes one or more branches that are under continuous research.

The advances in recent years in Artificial Intelligence (AI) are becoming a transversal field of study for all approaches in detection; however, detection by means of computer vision strategies is arousing great interest in the scientific community with challenges such as the Anti-UAV Challenge 2023 [31–33]. This is due to the advantages that vision systems offer by having a large amount of information for each of the images and the different bands of the electromagnetic spectrum in which these images can be perceived. For example, the infrared thermal images provided by the previous challenge allow to avoid the difficulties of illumination presented by the visible spectrum in low illumination or adverse environmental conditions, as well as other vulnerabilities presented by the other approaches, as in the case of radar, in the case of small objects or low-altitude flights.

Within this context, Artificial Intelligence (AI) and the advances in the last decades in computer vision provide powerful tools for object detection in images. In particular, techniques using convolutional neural networks (CNNs) and/or transformers are continuously demonstrating their potential in the context of object detection in multiple domains. This potential is supported by the ability of AI to process and learn from large volumes of data. Although there are numerous studies in the literature on object detection, segmentation, and tracking systems based on CNNs [34–36] or transformers [37–41], drone detection involves certain complexities inherent to this type of target object and the scenarios in which they are found [39,42,43].

In recent years, particularly in the analysis of complex scenarios, the focus has shifted towards the use of thermal images to enhance visual detection and object tracking systems [44–47]. Specifically, the application of AI in the analysis of thermal or infrared images has become essential for detecting drones under low-visibility conditions, such as at night or during adverse weather situations [48–50]. The use of thermal imaging provides the advantage of identifying objects based on their heat signatures, thus facilitating drone detection regardless of ambient light or unfavorable visual conditions. However, this approach also presents significant challenges, including variability in drone temperatures depending on the time of day or the environment, as well as the necessity to differentiate drones from other hot objects present within the field of view.

This work employs advanced AI techniques for the detection of drones in thermal or infrared images. Considering that the threat posed by unidentified civilian flying objects is increasing daily, it is essential to develop surveillance systems capable of accurate detection and optimal performance under extreme conditions, such as low-visibility or nighttime scenarios.

The main contribution of this paper is a comprehensive comparative analysis of state-of-the-art convolutional neural network (CNN) and transformer-based neural network architectures, adapted for drone detection in thermal infrared images. This study also evaluates the performance of each of the detectors for each of the nine video types derived from the specific conditions of the Anti-UAV Challenge 2023 [31–33]. It is important to note that each of these video types is a challenge as can be seen in Yu Q. et al.'s work [51]. In addition, while the most prominent papers in this challenge focus on the performance of Single-Object-Tracking (SOT), this paper focuses on the specific detection problem that is a

fundamental part of the tracking systems. This will lay the foundation for future research in this area.

This paper is organized as follows. Section 2 reviews the related work on drone detection and thermal imaging using neural networks. Section 3 details the methodology employed in this study, including a description of the dataset and the training process of the selected models. Section 4 presents the results of the experiments, including the performance metrics and a detailed analysis of each model's accuracy and inference speed. Additionally, Section 4 provides an in-depth discussion of the model performance results across different video categories. Finally, the conclusions drawn from this study are presented in Section 5.

2. Related Works

2.1. Cutting-Edge Detection Models

Convolutional neural networks (CNNs) and transformer-based neural networks (TNNs) are the two most recognized architectures in object detection. The former represents a type of artificial neural network specifically designed to process data with a grid-like structure, such as images [52,53]. In contrast to traditional neural networks, CNNs are optimized to capture spatial and hierarchical patterns in images, which renders them particularly suitable for computer vision tasks, including image classification, object detection, and segmentation.

Building upon CNNs, more advanced and complex models have been developed to enhance accuracy and efficiency in object detection. For this study, the most promising models from such architectures have been selected:

- You Only Look Once (YOLO) [54,55] is designed to detect and locate objects in images or videos in real-time. Unlike other methods that perform detection in multiple stages, YOLO integrates object detection and localization into a single neural network. This approach enables faster processing by performing object detection in a single pass through the image via the neural network. Currently, YOLO is considered the state of the art for single-pass detection methods. Due to its capability to effectively detect fast-moving objects by performing a single pass, it is well suited for handling fast-moving drones.
- GELAN (Generative Enhanced Low-light Adversarial Network) uses a multi-layer architecture capable of extracting both high-level semantic features and fine-grained details, which is essential for detecting objects in low-light, low-contrast environments such as thermal imaging. This network uses spatial pyramid clustering and layer aggregation to improve feature representation, enabling it to recognize objects at different scales and resist visual interference such as thermal noise or complex backgrounds. The application of GELAN networks in the context of drone detection in thermal imaging is particularly interesting because of their advanced ability to handle difficult visual conditions as demonstrated in their medical imaging applications [56]. In the specific case of drone detection in thermal imaging, the specific attributes of GELAN can help differentiate drones from other heat-emitting objects, a challenge often amplified in thermal imaging. In addition, the efficient processing of GELAN of multiscale and occluded objects is well suited to the context of drone detection, where drones may vary in size, speed, or visibility.

On the other hand, transformer-based neural networks (TNNs), commonly known as transformers, represent a class of neural network models that have revolutionized the field of machine learning, particularly in natural language processing (NLP) and, more recently, in computer vision [57,58]. In the domain of computer vision, these models have begun to

rival and even surpass traditional convolutional neural networks (CNNs) in several tasks, owing to their capacity to capture global relationships within the data and their scalability. However, their application poses challenges, particularly with respect to the computational and data requirements.

As with CNNs, the models selected for this study are as follows:

- Detection Transformer (DETR) [59] is designed to perform object detection and instance segmentation in a more efficient and accurate manner. Unlike traditional CNN-based methods, DETR employs the transformer architecture, which has proven to be highly effective in natural language processing tasks, to address problems in computer vision. DETR is particularly useful in environments where conditions are highly variable (different temperatures, weather, and times of day). Its ability to capture more abstract and complex patterns in images allows it to enhance drone detection in scenarios with changing thermal conditions or when drones have thermal signatures similar to their surroundings.
- Vision Transformer Backbones for Object Detection (ViTDet) [60] is a variant of Vision Transformers (ViTs) applied to the object detection task. ViTs, in general, have shown outstanding performance in classification tasks, and their ability to model long-range relationships within the image also makes them effective for detection. They can capture long-distance dependencies in the image, which is crucial for correctly interpreting infrared signals, where objects may not be as clearly defined as in visible images.

2.2. Infrared Image: Context and Challenge for Detection

Drone detection in infrared images is a research area that has garnered considerable attention due to increasing concerns regarding security and privacy. This type of imaging facilitates detection under low-visibility conditions, such as at night or during adverse weather, during which traditional optical cameras may be ineffective. However, this approach presents unique challenges that necessitate the application of advanced artificial intelligence and deep learning techniques.

Drones exhibit a thermal signature that can vary significantly depending on factors such as altitude, speed, and ambient temperature. In the following sections, some of the primary challenges are detailed:

- Thermal Noise: In complex environments, other objects, such as animals, vehicles, or even parts of the terrain, can emit thermal radiation, creating "noise" that can interfere with the accurate detection of drones. Models must be able to distinguish between these heat sources and the drone, which is often a small object with low thermal emissions.
- Temperature Variability: The temperature of drones can change during flight due to factors such as altitude and speed, which can complicate detection. Furthermore, ambient temperature, which can vary significantly between day and night or between different seasons, affects the effectiveness of detection algorithms.
- Resolution and Distance: The quality of infrared images is greatly dependent on the resolution of the camera and the distance from the target. At greater distances, the drone's thermal signature may become indistinguishable from the background, posing a significant challenge for detection models.
- Cluttered Environments: In scenarios where many objects are present (e.g., in urban areas), it is crucial for models to accurately identify and track the drone among other elements that may be emitting heat.

In response to these challenges, the Beijing Institute of Electronic Equipment has undertaken significant efforts to advance drone detection technologies by organizing

several workshops during the Conference on Computer Vision and Pattern Recognition (CVPR) in 2020, 2021, and 2023. These workshops have focused on the development of advanced models capable of detecting drones within a database of 600 videos, which have been segmented into frames of infrared images.

Among the most innovative proposals, Xin Yang et al. developed a method for detecting tiny objects in videos, guided by spatio-temporal motion information [61]. This approach, tested on the Anti-UAV 2021 dataset, showed a significant improvement in drone detection in complex scenarios, outperforming other traditional small object detection methods.

On the other hand, Qianjin Yu et al. introduced UTTracker [51], a transformer-based model designed for tracking drones in thermal infrared videos. This model effectively addresses challenges such as variations in target appearance, frequent disappearances, and camera movement, enabling it to achieve competitive performance in the latest 2023 Anti-UAV challenge, where it secured second place overall.

In the same workshop, a model based on YOLOv8 and DINO (Self-Distillation with No Labels) [62] was presented, achieving a significant breakthrough by attaining a score of 69.7% on the complete dataset. Due to these results, the model secured first place in the competition, establishing itself as the state of the art in drone detection using infrared images and representing the most accurate model to date [63]. Consequently, this model is selected as the baseline against which other results will be compared.

3. Methodology

This section describes the dataset utilized for the study, as well as the training processes of the four selected architectures. Finally, the metrics for comparing the various models are delineated.

3.1. Dataset

The Anti-UAV Database is a compilation of thermal images specifically collected for the purpose of detecting drones in various environments. This database was assembled as part of the open challenge proposed by the Beijing Institute of Electronic Equipment during CVPR 2023 [31]. The objective of the challenge was to create a standardized dataset that would facilitate the evaluation and comparison of various drone detection models developed by the participants in the competition.

The database comprises multiple scenarios that simulate real-world situations, consisting of 600 videos corresponding to more than 230,000 thermal images divided into three sets: 70% for training, 15% for validation, and 15% for testing. The images encompass diverse scenarios and lighting conditions, thereby providing a robust foundation for training and evaluating AI models. These images include variations in the size, speed, and shape of the drone. The scenarios incorporate both urban and rural settings, featuring daytime and nighttime conditions, which add complexity to the detection challenge.

The images were captured under various environmental conditions, including differing temperatures, humidity levels, and visibility conditions. This indicates that the model must be capable of detecting drones in scenarios ranging from clear skies to low-visibility conditions. The database comprises eight types of videos, each representing a specific challenge:

- UAV: Images that are not categorized by any particular condition.
- Out of View (OV/VE): The target moves out of the current field of view.
- Occlusion (OC): The target is partially hidden behind another object.
- Fast Motion (FM): The target moves rapidly.
- Scale Variation (SV): The scale of the bounding box varies significantly.

- Thermal and Infrared Crossover (TC/IC): The target temperature is similar to that of another object or the surrounding landscape.
- Dynamic Background Clusters (DBCs): Dynamic background surrounding the target, such as animals or vegetation.
- Low Resolution (LR): The area of the bounding box is very small.
- Target Scale (TS): The target will have a very small or large scale as the frames progress.

Figure 1 illustrates the distribution of data by video type. It can be observed that the maximum concentration of videos is found in the UAV class, as well as in Thermal Crossover (TC) and Fast Motion (FM). Conversely, three groups exhibit minimal representation; occlusion (OC) comprises a total of 275 images, while low resolution (LR) and Out of Vision (OV) contain no images at all. The latter two are excluded; however, in the case of Out of Vision (OV), it does have representation like (VE), and therefore, these cases are considered.

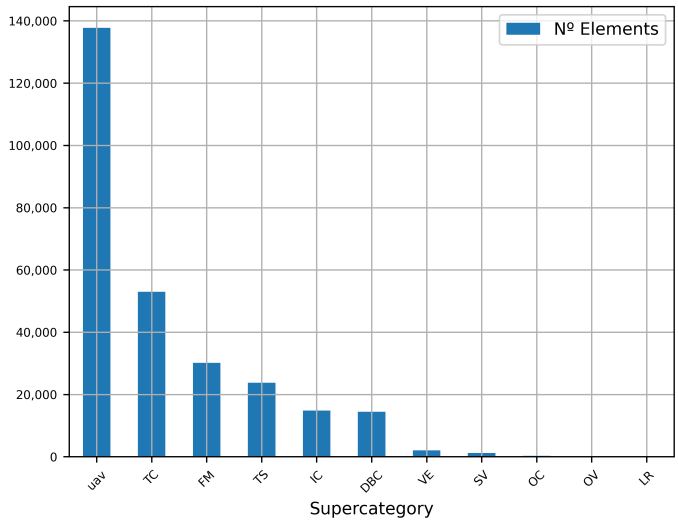


Figure 1. Data distribution by video type [31].

3.2. Models

This section provides a justification for the models selected for evaluation concerning the drone detection problem utilizing thermal imaging. The selected models are as follows:

- **YOLO**: YOLO is chosen because of its ability to perform real-time detection, which is crucial for applications where speed is essential. Additionally, its architecture is well documented and optimized to work with low-resolution images, which is necessary for accurate drone detection in thermal images. It is important to note that, according to most experts, YOLO is considered one of the best image detectors available. Starting from the best-performing models at the CVPR 2023 conference [51] that used YOLOv5 and one of the best-performing models at the same conference [63], which is based on an ensemble model with YOLO-v8, working with this type of model can be considered essential in this type of detection problem. However, recent advances have further improved the performance of YOLO systems with the YOLO-v9 model, which incorporates innovations such as Programmable Gradient Information (PGI) to solve the problem of losing data during information propagation. Therefore, studying the new YOLO model in the context of drone thermal imaging helps to extend the state of the art of drone detection in thermal imaging. This model has different versions [t, s, m, c, e], which have the same input size but with an increasing number of parameters, which influences the detection accuracy but affects the processing speed. For this work, we consider the largest YOLOV9 model corresponding to e, YOLO-v9-e.

- **GELAN**: Since it is a model used in YOLOv9, the goal is to compare its performance with YOLO and evaluate whether its simpler architecture provides improvements in drone detection.
- **DETR**: Although it has not yet been implemented for infrared images, it is considered relevant to include a transformer architecture in a detection comparison, given its ability to achieve more precise detection in situations where spatial and contextual relationships are complex, such as thermal images with multiple heat sources. It is part of the Detectron2 group [64], which is also considered one of the best sets of object detection models.
- **ViTDet**: We are interested in determining whether there is a difference between the transformer-based models applied to this task. Like DETR, it is part of Detectron2, making it a good model for comparison.

3.3. Models Training

To enhance efficiency and reduce training time, transfer learning is employed by utilizing pre-trained models in object detection and adapting them for drone detection. To analyze the performance of the models in the context of the problem, the dataset is divided between testing and validation. These subsets include 40,000 images for training and 5000 images for validation. Initial testing is performed locally, and various batch sizes (2, 4, 8, 16) are evaluated for all models until the size that shows the best final performance is identified. In addition, the effect of using up to eight workers to measure the processing speed is evaluated to avoid GPU overload, which in some cases requires a reduction in the number of workers.

Regarding the optimizer, the two most common options are evaluated: Adam (Adaptive Moment Estimation) and SGD (Stochastic Gradient Descent). Although SGD is more stable than Adam [65] and tends to generalize faster, some work like [66] suggests that hyperparameters may be the reason that adaptive algorithms such as Adam fail to generalize. In our case, Adam yields superior results in terms of accuracy; therefore, it is selected, despite requiring a longer training period.

Once these configurations are determined, server training commences, during which the initial hyperparameters are configured, including the learning rate and batch size, based on the insights obtained from the preliminary tests. This training is conducted using the complete database, which comprises a distribution of 70% images for "train" (162,090), 15% for "validation" (34,733), and the remaining 15% for "test" (34,734).

During training, after each epoch, the results are validated to monitor the model's progress. If overfitting is detected, the hyperparameters are adjusted to mitigate this issue. In instances where performance begins to decline, reverting to previous configurations is considered to prevent losses in training efficiency and prediction accuracy.

3.4. Metrics

For the evaluation of the performance of the trained neural models, this study is based on a set of standards for image segmentation, which is described as follows:

- **Intersection over Union (IoU)**. The IoU is a metric that measures the degree of overlap between the predicted bounding box of the model and the actual bounding box of the object, and is calculated by dividing the overlap area between the two boxes by the overlap area of their union (see Figure 2):

$$\text{IoU} = \frac{\text{Intersection area}}{\text{Union area}} \tag{1}$$

A higher IoU value indicates a better prediction, as the predicted box aligns more closely with the actual box. In video detection problems, where objects can move and change shape, a high IoU is critical to ensure that models not only detect the presence of an object but also accurately localize it in each frame of the video.

- ***F*1-Score** and ***F*1-Confidence**. The *F*1-Score is the harmonic mean between precision and recall. This metric is particularly useful when there is an imbalance between the classes of true positives and false negatives:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

  - **Precision** is the proportion of true positives (correctly detected objects) over the sum of true positives and false positives (incorrect detections).
  - **Recall** measures the proportion of true positives relative to the sum of true positives and false negatives (undetected objects).

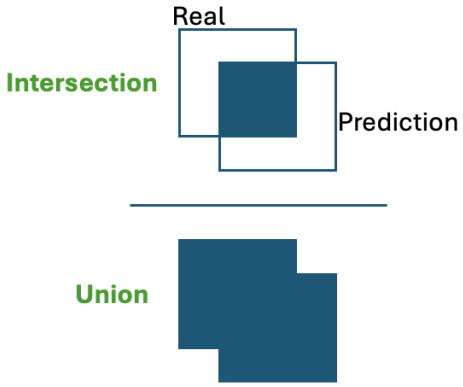


Figure 2. IoU description.

In video object detection, the F1-Score is important, as it balances precision and recall, providing an overall assessment of the model's performance regarding correct detections versus missed detections.

The F1-Confidence is an extension of the F1-Score that incorporates the confidence of the model's predictions. This metric adjusts the F1-Score based on the average confidence of correct detections, allowing for the consideration of both the quantity of correct predictions and the certainty with which the model makes those predictions. It is particularly useful in applications where confidence in the model's decisions is critical, such as aerial surveillance or automated detection systems, where an incorrect prediction made with high confidence can have significant consequences. This metric provides a more detailed evaluation of the performance of the model, particularly in scenarios where the quantity of correct detections is important, and the certainty with which those detections are made is equally crucial. AP measures how precision varies at different levels of recall and is derived from the precision–recall curve.

AP is typically calculated for a single Intersection over Union (IoU) threshold, usually set at 0.5, which is referred to as AP@0.5 or AP@50. Consequently, the AP value varies depending on the IoU threshold that is considered acceptable.

However, for more complex problems and more rigorous evaluations, the mean Average Precision (*mAP*) is utilized. *mAP* averages the *AP* across several IoU thresholds,

typically ranging from 0.5 to 0.95 in increments of 0.05, referred to as *mAP*@0.5:0.95 or *mAP*@50:95. This can be calculated as shown in Equation (3):

$$mAP@50:95 = \frac{AP@50 + AP@55 + \cdots + AP@90 + AP@95}{10} = \frac{\sum_{x=0.5}^{0.95} AP@x}{10} \quad (3)$$

This metric is particularly relevant in video detection tasks, as it evaluates the ability of a model to detect objects with varying degrees of overlap between the predictions and the ground truth annotations.

- **True Negative Rate**. The True Negative Rate (*TNR*) or specificity quantifies the proportion of true negatives (*TNs*) that are accurately identified out of the total number of cases where the target is absent (i.e., the sum of true negatives and false positives):

$$TNR = \frac{TN}{TN + FP} \quad (4)$$

  - **True Negatives (TNs)**: These are cases where there is no drone in the scene, and the model correctly does not detect any target.
  - **False Positives (FPs)**: These are cases where there is no drone in the scene, but the model incorrectly detects something (false positive).

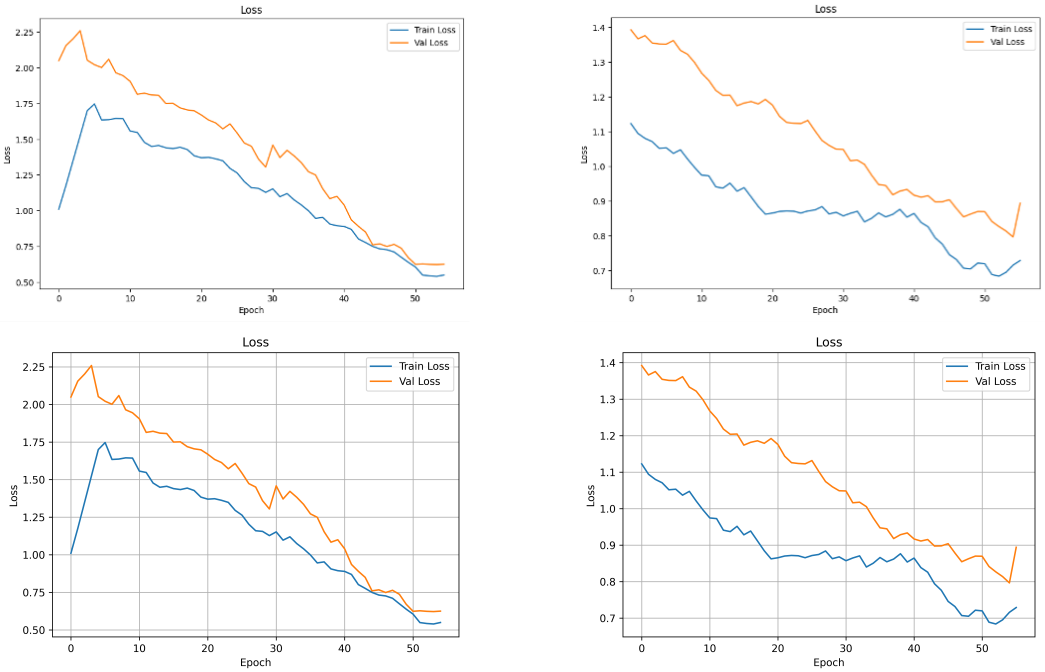
  This metric is essential for evaluating the performance of a model in the absence of a target to detect, which is particularly important in scenarios where false positives are costly or disruptive.
- **Inference Time and FPS**. Inference time refers to the time that a model requires to process a single image from a video and generate a prediction. This metric is vital for real-time applications, such as drone detection, where every millisecond is significant. Related to this is Frames Per Second (FPS), which indicates the number of images a model can process per second. A higher FPS denotes that the model can operate at increased speeds, which is essential in scenarios where immediate detection and reaction are critical, such as in aerial surveillance.
  Both metrics, inference time and FPS, are utilized to assess the efficiency of the model and its suitability for real-time scenarios.

4. Results

This section presents the results obtained from the trained models. During the training process and the evaluation of the models, an environment composed of an NVIDIA RTX 4090 graphics card, an AMD Ryzen 9 7950K processor with 16 cores, and 128 GB of DDR4 RAM is utilized. Table 1 demonstrates the parameters employed during the training phase. Figure 3 illustrates the training and convergence of the four models.

Regarding the YOLO model (Figure 3a), it is trained for a total of 100 epochs, achieving convergence at a loss value of 0.4. The batch size is set to 10, and 8 workers are utilized. This configuration proves to be the most efficient in terms of speed and performance, based on preliminary tests conducted with fewer epochs. Throughout the training process, an iterative optimization procedure is employed, determining that the most appropriate optimizer for this model is Adam. This optimizer adjusts the model's parameter weights to minimize the loss function, thereby facilitating the model's ability to learn more rapidly and efficiently.

(**a**) YOLO (**b**) GELAN



(**c**) DETR (**d**) VitDet

Figure 3. Convergence of the models during training.

Table 1. Parameters used during the training phase.

Model	Epochs	Batch Size	Workers	Optimizer
YOLO	50 + 25 + 25	10	8	Adam
GELAN	75	10	8	Adam
DETR	30 + 25	2	2	AdamW
ViTDet	55	2	2	AdamW

Since YOLO requires 100 epochs to achieve a reasonable point of convergence, GELAN is trained directly for 75 epochs. Beginning from epoch 68, it is observed that the model starts to converge, exhibiting minimal learning loss. In the final 7 epochs, the validation set loss fluctuates by only 0.01, while the training set loss continues to decrease, except in the last epoch, where it is increased by 0.02. This behavior suggests that the model may be beginning to overfit (see Figure 3b). The batch size is set to 10, and 8 workers are utilized, resulting in the optimal combination of speed and efficiency based on the tests conducted during the initial epochs.

The DETR model achieves convergence around Epoch 50, demonstrating signs of overfitting around Epoch 54 (Figure 3c). A batch size of 2 with 2 workers appears to attain the optimal speed–efficiency balance based on tests conducted with fewer epochs. For this model, the optimizer is changed to AdamW, which is more effective for transformer architectures.

Considering that the DETR model requires 55 epochs to achieve convergence, ViTDet is trained for a similar number of epochs. ViTDet exhibits a tendency to converge slightly earlier, suggesting comparable or potentially superior learning efficiency. However, in the final 1-2 epochs, a marginal increase is observed in both the validation loss (0.1) and the training loss (0.01), which may indicate the onset of overfitting during training (see Figure 3d). Regarding hardware configuration, a batch size of 2 and 2 workers provide the optimal balance between speed and efficiency, based on preliminary tests conducted with

fewer epochs. For optimization, the AdamW optimizer is employed, as it is particularly well suited for transformer-based architectures, such as ViTDet.

4.1. Model comparison

Table 2 presents a comparison of the models in terms of speed and inference time, while Table 3 compares them based on accuracy. The YOLO model is distinguished as the most efficient in terms of speed, achieving a performance of 134 FPS and an inference time of 7.5 ms, rendering it the preferred option for real-time applications. It is 3.5 times faster than the transformers and 1.4 times faster than GELAN. In contrast, GELAN, while slower, demonstrates acceptable performance, being 2.5 times faster than DETR and ViTDet. The latter two transformer-based models are considerably slower, which may restrict their applicability in scenarios where speed is critical.

Table 2. Performance of the models in terms of time.

Model	Epochs	FPS	Inference Time	Training Time
YOLO	100	134	7.5 ms	48 h y 5 min
GELAN	75	95	10.5 ms	50 h y 32 min
DETR	55	38	26 ms	93 h y 6 min
ViTDet	55	36	28 ms	97 h y 48 min

Table 3. Accuracy of the models.

Model	AP@50	mAP@50:95	F1-Confidence
YOLO	74.7%	72%	75% at 0.102
GELAN	81%	77%	77% at 0.255
DETR	78.3%	75.7%	81% at 0.063
ViTDet	65%	62%	75% at 0.137
UTTracker	68.8%	—	—
Baseline	69.7%	—	—

Regarding accuracy, GELAN excels with a mAP of 50:95, achieving a score of 77%, closely followed by DETR at 75.7%. Both methods significantly outperform YOLO, which, while exhibiting good accuracy, is more oriented toward applications where speed is a determining factor. Although ViTDet demonstrates lower accuracy, it may be beneficial in contexts where detection under complex conditions is paramount. However, its low speed and accuracy render it less suitable for real-time scenarios.

4.2. Performance Analysis by Video Type

This section analyzes the performance of drone detection models based on various types of videos present in the database. Each video type provides a practical example, showcasing its corresponding "bounding box" along with the detections made by the models in that image. Furthermore, the overall results of the models for that category are presented.

To conduct this analysis, the Average Precision (AP) metric is selected as the primary evaluation tool. AP is chosen for its relevance to the task, its simplicity, and its ability to capture the critical aspects of model performance in drone detection. This selection facilitates a coherent and comparative analysis in accordance with the standards of the computer vision community.

By concentrating on a key metric such as AP, a more comprehensive and detailed analysis of the results can be conducted, thereby avoiding the dispersion that may arise from the inclusion of multiple metrics. This approach facilitates the maintenance of a clear and precise focus on the model's performance regarding accuracy, which is the most critical aspect of this study.

It should be noted that, in the group of videos without targets (VE), where no drones are present, the model's performance is evaluated in terms of its ability to avoid false detections. This performance is measured using the True Negative Rate (TNR) metric, which quantifies the frequency with which the model accurately does not detect a drone in the absence of one. This evaluation is crucial to ensure that the model does not generate false alarms in scenarios in which no drones are present.

4.2.1. UAV

This group encompasses all images that do not belong to a specific category, representing 60% of the database. As illustrated in Table 4, DETR demonstrates higher accuracy in this category, reflecting its capability to effectively manage standard scenarios. The F1-Confidence of DETR is also significantly high, indicating an effective balance between precision and recall in these scenarios. YOLO, although less accurate, maintains high efficiency, rendering it suitable for applications where speed is critical. The detection results for this category of images are presented in Figure 4.

Table 4. Results for video type 'UAV'.

Model	IoU	AP	F1-Confidence
YOLO	84.4%	84.2%	87% at 0.102
GELAN	81.63%	87.2%	91% at 0.323
DETR	95.30%	90%	92% at 0.411
ViTDet	78.6%	70.4%	76% at 0.137

(**a**) Real Bounding-Box (ID 40791) (**b**) Detection (ID 40791)

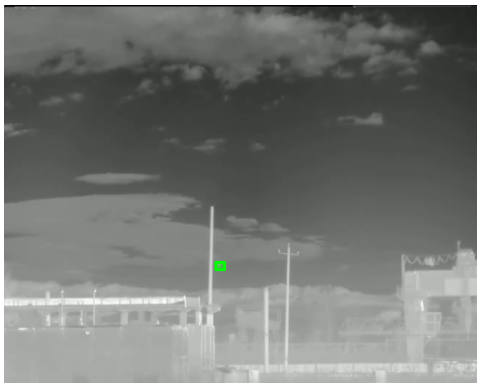
Figure 4. Detection results for "UAV".

4.2.2. Target Scale (TS)

This group comprises images in which the target scale varies from small to large throughout the video, representing 10% of the dataset.

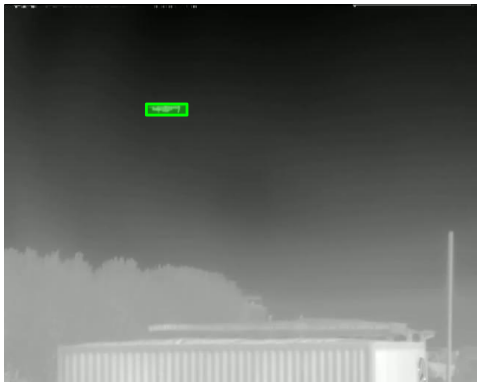
GELAN and DETR exhibit superior performance in this category, with GELAN demonstrating a slight advantage in terms of overall accuracy (see Table 5). The capacity of these models to manage significant variations in target scale indicates that they are more robust in scenarios with diverse target sizes. While YOLO is efficient, it displays slightly lower performance, which may limit its applicability in situations where the target scale

varies considerably. Figure 5 illustrates the results of this category for both small and larger objects.

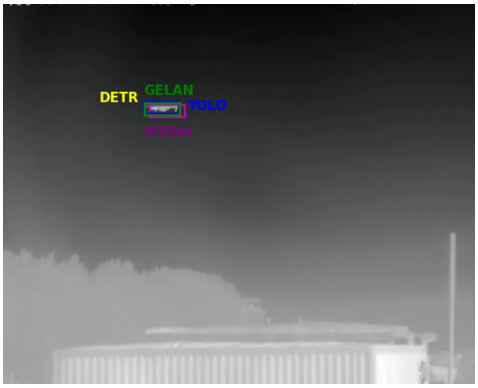


(**a**) Real Bounding-Box (ID = 43,428)

(**b**) Detection (ID = 43,428)



(**c**) Real Bounding-Box (ID = 48,222)



(**d**) Detection (ID = 48,222)

Figure 5. Detection results for target scale (TS).

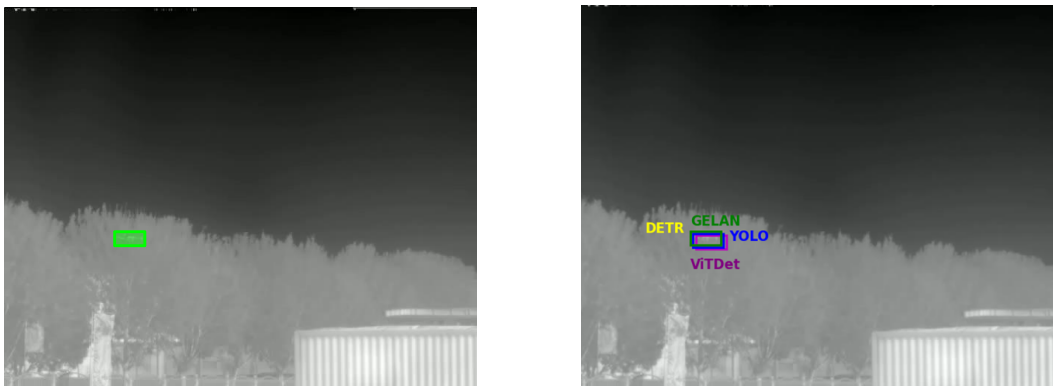
Table 5. Results for video type target scale (TS).

Model	IoU Small	IoU Large	AP	F1-Confidence
YOLO	73.33%	79.5%	73.5%	77% at 0.093
GELAN	76.74%	80.4%	82.5%	84% at 0.426
DETR	80.66%	81%	80.3%	82% at 0.321
ViTDet	62.41%	70.6%	64%	66% at 0.122

4.2.3. Dynamic Background Clusters (DBC)

This type of video features a dynamic background that complicates the detection of drones; in this instance, the wind interacts with the trees. This group contains a total of 14,400 images, thus rendering it the fourth smallest group.

GELAN clearly stands out in this category, exhibiting significantly higher accuracy, which suggests that its architecture is better suited to handling dynamic backgrounds. DETR also demonstrates acceptable performance; however, it is inferior to GELAN (see Table 6). YOLO, due to its emphasis on speed, encounters difficulties in these scenarios as evidenced by its lower accuracy. Figure 6 presents examples of the results in this category.



(**a**) Real Bounding-Box (ID = 49,064) (**b**) Detection (ID = 49,064)

Figure 6. Detection results for Dynamic Background Clusters (DBC).

Table 6. Results for video type Dynamic Background Clusters (DBC).

Model	IoU	AP	F1-Confidence
YOLO	34.33%	38.5%	47% at 0.039
GELAN	52.34%	58%	63% at 0.298
DETR	52.34%	51.2%	58% at 0.175
ViTDet	30.23%	32.1%	43% at 0.038

4.2.4. Thermal and Infrared Crossover (TC)

This group is the second largest, consisting of images in which the background temperature is similar to that of the target, with approximately 53,000 images.

DETR and GELAN demonstrate high effectiveness in this category, as they can differentiate the target despite thermal similarities with the background. This observation reinforces the capability of these models to manage challenging thermal conditions, in which other models, such as YOLO and ViTDet, exhibit increased confusion (see Table 7). Figure 7 presents an example of the results in this category.

(**a**) Real Bounding-Box (ID = 42,390) (**b**) Detection (ID = 42,390)

Figure 7. Detection results for Thermal and Infrared Crossover (TC).

Table 7. Results for video type Thermal and Infrared Crossover (TC).

Model	IoU	AP	F1-Confidence
YOLO	64%	70.6%	74% at 0.046
GELAN	80%	80.7%	81% at 0.107
DETR	81.3%	81%	84% at 0.137
ViTDet	70.4%	68.6%	76% at 0.062

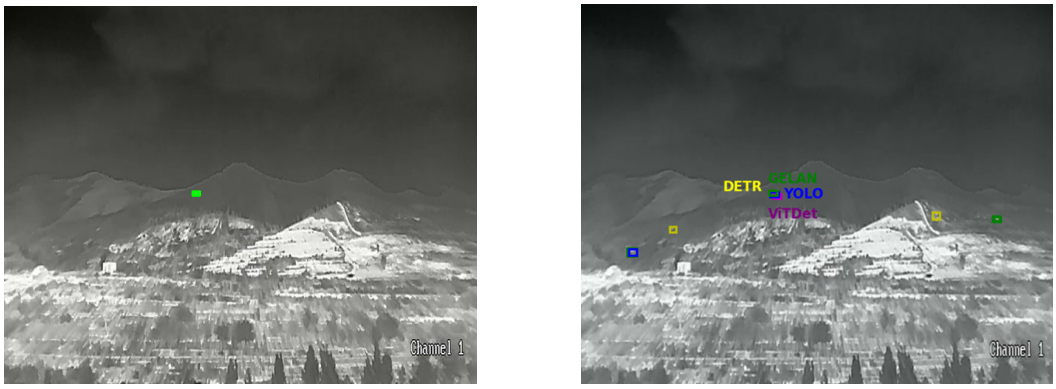
4.2.5. Scale Variation (SV)

This group is among the least represented, comprising only 1214 images, in which the area of the "bounding boxes" varies considerably in size, without a systematic progression from small to large.

Although all models demonstrate solid performance in this category, DETR and GELAN again stand out, with GELAN exhibiting superior detection consistency despite variability in size (Table 8). The small sample size in this group suggests that additional data would be required to fully validate these findings. Figure 8 presents an example of this category.

Table 8. Results for video type Scale Variation (SV).

Model	IoU	AP	F1-Confidence
YOLO	58.44%	83.6%	81% at 0.063
GELAN	68.18%	84.30%	80% at 0.477
DETR	69.3%	80.7%	82% at 0.221
ViTDet	63.6%	68.5%	70% at 0.092



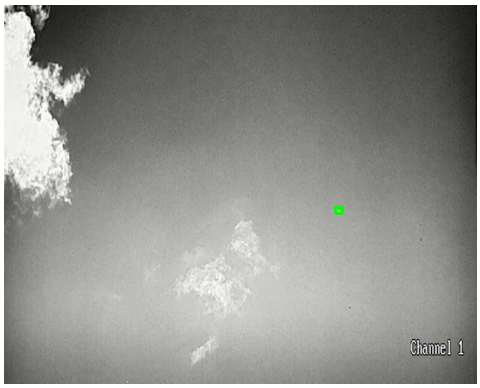
(**a**) Real Bounding-Box (ID = 98,156) (**b**) Detection (ID = 98,156)

Figure 8. Detection results for Scale Variation (SV).

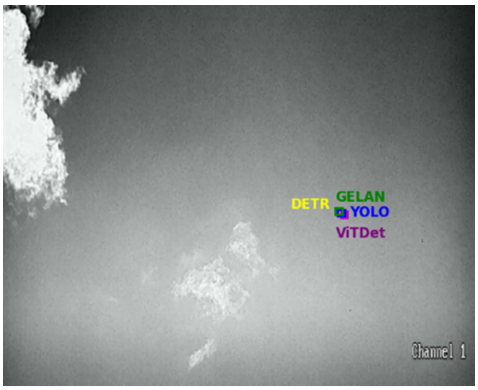
4.2.6. Fast Motion (FM)

This collection of 30,000 images features high-velocity drones.

GELAN demonstrates strong performance in fast-motion scenarios, indicating its robustness in detecting drones under dynamic conditions (Table 9). YOLO also exhibits acceptable performance; however, it is inferior to GELAN, suggesting that its speed may not be sufficient to compensate for the complexity of these scenarios. Refer to Figure 9 for an example of this category.



(**a**) Real Bounding-Box (ID = 105221)



(**b**) Detection (ID = 105221)

Figure 9. Detection results for Fast Motion (FM).

Table 9. Results for video type Fast Motion (FM).

Model	IoU	AP	F1-Confidence
YOLO	78.82%	60.2%	64% at 0.049
GELAN	79%	66.7%	72% at 0.137
DETR	75.6%	59.3%	63% at 0.090
ViTDet	66.9%	50.6%	53% at 0.050

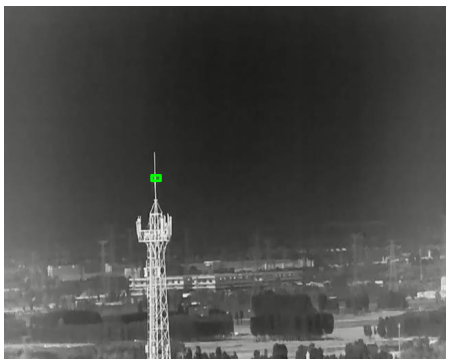
4.2.7. Oclussion (OC)

This is the smallest group, consisting of only 275 images, and it features drones that are partially occluded.

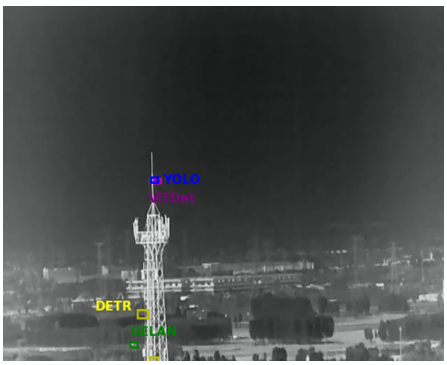
Although it does not achieve correct detection in the sample image (see Figure 10), GELAN demonstrates significantly higher accuracy in this category, suggesting that its architecture is better suited to handle occlusion conditions (Table 10). However, the low representation of this group limits the ability to generalize these results.

Table 10. Results for video type Occlusion (OC).

Model	IoU	AP	F1-Confidence
YOLO	66.55%	3.5%	5% at 0.0001
GELAN	0%	25.6%	32% at 0.107
DETR	0%	10.2%	12% at 0.112
ViTDet	58.33%	8.5%	10% at 0.026



(**a**) Real Bounding-Box (ID = 97,226)



(**b**) Detection (ID = 97,226)

Figure 10. Detection results for Occlusion (OC).

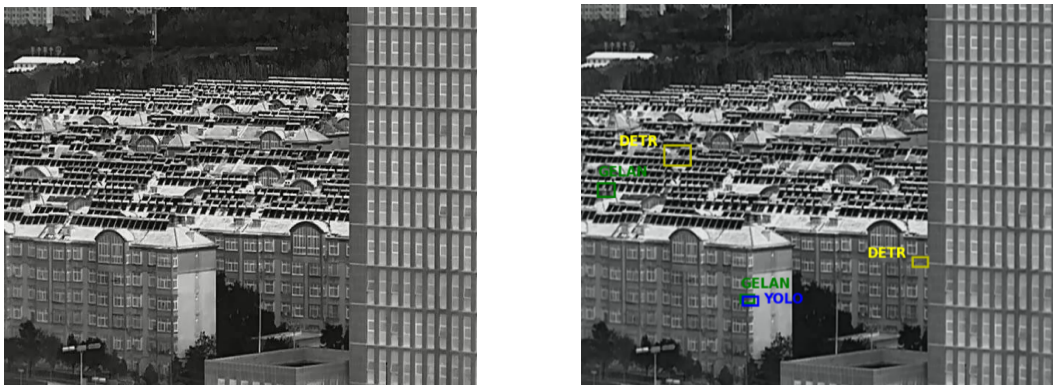
4.2.8. Out of View (VE)

This group comprises images in which the drone is not visible within the scene, presenting a particular challenge for detection models. If a model indicates the presence of a drone in these images, it incurs an error known as a false positive. In this case, the model's efficiency is assessed using the True Negative Rate (TNR), as the emphasis is on the models' ability to refrain from detecting non-existent drones. This group contains a total of 2022 images.

ViTDet is the best-performing model in this group, demonstrating a superior ability to avoid false detections when no target is present (Table 11). This characteristic enhances its reliability in scenarios where minimizing false positives is crucial. In this case, the example illustrates a 0% match if a prediction is made, as the model attempts to compare the predicted bounding box with one that does not exist. Figure 11 presents an example of a result in this category.

Table 11. Results for video type Out Of View (VE).

Model	IoU	TNR	F1-Confidence
YOLO	0%	0.6%	1% at 0.027
GELAN	0%	0.5%	2% at 0.010
DETR	0%	3%	5% at 0.009
ViTDet	100%	10%	12% at 0.415



(**a**) Real Bounding-Box (ID = 39,472) (**b**) Detection (ID = 39,472)

Figure 11. Detection results for Out Of View (VE).

4.3. Discussion

The results obtained in this study reveal a range of behaviors across the evaluated models (YOLO, GELAN, DETR, and ViTDet), highlighting their respective strengths and weaknesses in the detection of drones in thermal images. The main findings are interpreted as follows:

- **YOLO**: This model excels in terms of speed and inference time, being the fastest of the four with 134 FPS and an inference time of 7.5 ms. However, its precision (mAP 50:95 of 72%) is slightly lower than that of GELAN and DETR, indicating that although YOLO is suitable for real-time applications, it may not be the best choice when precision is critical.
- **GELAN**: GELAN stands out for its precision, achieving the highest mAP 50:95 (77%) and an AP 50 of 81%. Its performance is particularly notable in scenarios with dynamic background noise (DBC) and Fast Motion (FM), where other models struggle. Despite

its superior precision, its speed is moderate (95 FPS), making it a viable option in applications where precision is prioritized over speed.

- **DETR**: DETR demonstrates balanced performance with a mAP 50:95 of 75.7% and an AP 50 of 77.3%, positioning itself between YOLO and GELAN in terms of precision. However, its inference time is significantly longer (26 ms), making it less suitable for real-time applications. DETR proves to be particularly effective in detecting drones in complex situations such as Scale Variations (SVs) and thermal-infrared crossovers (TCs).
- **ViTDet**: Despite its promising results in similar tasks in other contexts, ViTDet has the lowest performance in this study, with a mAP 50:95 of 62% and an AP 50 of 65%. Its inference time is also the slowest (28 ms), which, combined with its lower precision, suggests that this model may not be the most suitable option for drone detection in thermal images under the evaluated conditions. In particular, it is the best model to recognize when not to detect a drone in an image, with an accuracy of approximately 9% in such cases (VE), compared to around 1% for the other models.

Taking into account the various scenarios, one may conclude the following:

- **UAV Category**: All models perform well in this category, and DETR achieves the highest precision (90%). This reflects the models' ability to handle standard scenarios without additional complicating factors.
- **Target Scale (TS)** and **Scale Variation (SV)**: Despite having lower precision across all examples, GELAN stands out slightly in these groups, highlighting its ability to manage significant variations in the target size. In particular, all models have similar precision in both groups, except for YOLO, which does not benefit when the Scale Variation is small to large (TS).
- **Dynamic Background Clusters (DBCs)**: Although the GELAN performance in the example is lower, it shows consistent performance for this type of video. However, the difference between models in this category is not as pronounced as in others.
- **Thermal and Infrared Crossover (TC)**: DETR and GELAN are the most effective models, suggesting that they are better at handling images with multiple objects of similar temperature, which tends to confuse YOLO and ViTDet. This is the category in which the difference between the models is the least noticeable.
- **Fast Motion (FM) and occlusion (OC)**: GELAN also performs well in tracking drones in fast-motion scenarios and under occlusion, making it a robust option for more challenging environments.
- **Out of View (VE)**: In this case, ViTDet is clearly the best model, being the only one with a significant percentage, reaching 10%. DETR achieves 3%, ranking second, indicating that transformers seem to better detect the absence of a target.

5. Conclusions

In this study, different neural network architectures were trained and evaluated for drone detection in infrared thermal imagery, focusing on challenging scenarios where detection accuracy is critical. YOLO v9 proved ideal for real-time applications where speed is a priority, while GELAN excelled in accuracy in most categories, with competitive speed. DETR performed well in standard conditions with similar temperatures, but is less suitable for real-time use, and ViTDet showed utility in minimizing false positives.

The study achieved higher average accuracy than state-of-the-art models from the 2023 Anti-UAV Challenge, advancing the field of drone detection using infrared images. Future work could focus on creating models specialized for specific scenarios (e.g., occlusion, high-speed motion, and Thermal Crossover) to improve detection accuracy and efficiency. In addition, improving the efficiency of transformer-based models such as DETR and ViTDet

could make them viable for real-time applications, expanding their utility in complex detection environments.

Finally, this work advances drone detection capabilities with neural networks and outlines a path for the further improvement of detection systems.

Author Contributions: Conceptualization, J.P.L., M.A.P. and L.U.; methodology, J.P.L., M.A.P. and G.G.; software, G.G. and J.P.L.; validation, G.G., J.P.L., M.A.P. and L.U.; formal analysis, G.G., J.P.L., M.A.P. and L.U.; investigation, G.G., J.P.L., M.A.P. and L.U.; resources, M.A.P. and L.U.; writing—original draft preparation, G.G., M.A.P., J.P.L. and L.U.; writing—review and editing, M.A.P., J.P.L. and L.U.; visualization, G.G., M.A.P. and J.P.L.; supervision, J.P.L., M.A.P. and L.U.; project administration, M.A.P. and J.P.L.; funding acquisition, M.A.P. and J.P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the public research projects of the Spanish Ministry of Science and Innovation PID2020-118249RB-C22 and PDC2021-121567-C22—AEI/10.13039/501100011033 and the project under the call PEICTI 2021-2023 with the identifier TED2021-131520B-C22.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, M.; Zhang, R.; Chen, L.; Tang, Q.; Xia, L. Investigation on Advances of Unmanned Aerial Vehicle Application Research in Agriculture and Forestry. *Smart Agric.* **2021**, *3*, 22–37. [CrossRef]
2. Fleureau, J.; Galvane, Q.; Tariolle, F.L.; Guillotel, P. Generic drone control platform for autonomous capture of cinema scenes. In *DroNet 2016: Proceedings of the 2nd Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use*; Association for Computing Machinery: New York, NY, USA, 2016. [CrossRef]
3. Özgen, K. The impact of drones in documentary filmmaking: Renaissance of aerial shot. *AVANCA | CINEMA* **2020**, 559–563. [CrossRef]
4. Sahithya, N.; Reddy, D.S.; Chandu, V.; Yashaswini, G.S.; Reddy, A.M.; Sukrutha, G. Security Drone for Surveillance in Military. *Int. J. Res. Appl. Sci. Eng. Technol.* **2024**, *12*, 1269–1275. [CrossRef]
5. Das, M.S.; Kumar, G.R.; Ram Kumar, R.P. An Insight on Drone Applications in Surveillance Domain. In *Big Data, Machine Learning, and Applications* ; Lecture Notes in Electrical Engineering (LNEE, Volume 1053); Springer: Berlin/Heidelberg, Germany, 2024. [CrossRef]
6. Benarbia, T.; Kyamakya, K. A literature review of drone-based package delivery logistics systems and their implementation feasibility. *Sustainability* **2022**, *14*, 360. [CrossRef]
7. Betti Sorbelli, F. UAV-Based Delivery Systems: A Systematic Review, Current Trends, and Research Challenges. *ACM J. Auton. Transp. Syst.* **2024**, *1*, 1–40. [CrossRef]
8. Garg, V.; Niranjan, S.; Prybutok, V.; Pohlen, T.; Gligor, D. Drones in last-mile delivery: A systematic review on Efficiency, Accessibility, and Sustainability. *Transp. Res. Part D Transp. Environ.* **2023**, *123*, 103831. [CrossRef]
9. Flemons, K.; Baylis, B.; Khan, A.Z.; Kirkpatrick, A.W.; Whitehead, K.; Moeini, S.; Schreiber, A.; Lapointe, S.; Ashoori, S.; Arif, M.; et al. The use of drones for the delivery of diagnostic test kits and medical supplies to remote First Nations communities during Covid-19. *Am. J. Infect. Control* **2022**, *50*, 849–856. [CrossRef]
10. Famili, A.; Stavrou, A.; Wang, H.; Park, J.M.; Gerdes, R. Securing your airspace: Detection of drones trespassing protected areas. *Sensors* **2024**, *24*, 2028. [CrossRef] [PubMed]
11. Lykou, G.; Moustakas, D.; Gritzalis, D. Defending airports from uas: A survey on cyber- attacks and counter-drone sensing technologies. *Sensors* **2020**, *20*, 3537. [CrossRef]
12. Mekdad, Y.; Aris, A.; Babun, L.; Fergougui, A.E.; Conti, M.; Lazzeretti, R.; Uluagac, A.S. A survey on security and privacy issues of UAVs. *Comput. Netw.* **2023**, *224*, 109626. [CrossRef]
13. Krame, G.; Vivoda, V.; Davies, A. Narco drones: Tracing the evolution of cartel aerial tactics in Mexico's low-intensity conflicts. *Small Wars Insur.* **2023**, *34*, 1095–1129. [CrossRef]
14. Seidaliyeva, U.; Ilipbayeva, L.; Taissariyeva, K.; Smailov, N.; Matson, E.T. Advances and challenges in drone detection and classification techniques: A state-of-the-art review. *Sensors* **2023**, *24*, 125. [CrossRef] [PubMed]

15. Taha, B.; Shoufan, A. Machine Learning-Based Drone Detection and Classification: State-of-the-Art in Research. *IEEE Access* **2019**, *7*, 138669–138682. [CrossRef]
16. Batool, S.; Frezza, F.; Mangini, F.; Simeoni, P. Introduction to Radar Scattering Application in Remote Sensing and Diagnostics: Review. *Atmosphere* **2020**, *11*, 517. [CrossRef]
17. Li, S.; Chai, Y.; Guo, M.; Liu, Y. Research on detection method of UAV based on micro-Doppler effect. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 3118–3122.
18. Coluccia, A.; Parisi, G.; Fascista, A. Detection and Classification of Multirotor Drones in Radar Sensor Networks: A Review. *Sensors* **2020**, *20*, 4172. [CrossRef] [PubMed]
19. Mandal, S.; Satija, U. Time–Frequency Multiscale Convolutional Neural Network for RF-Based Drone Detection and Identification. *IEEE Sens. Lett.* **2023**, *7*, 1–4. [CrossRef]
20. Aouladhadj, D.; Kpre, E.; Deniau, V.; Kharchouf, A.; Gransart, C.; Gaquière, C. Drone Detection and Tracking Using RF Identification Signals. *Sensors* **2023**, *23*, 7650. [CrossRef]
21. Sun, Y.; Li, J.; Wang, L.; Xv, J.; Liu, Y. Deep Learning-based drone acoustic event detection system for microphone arrays. *Multimed. Tools Appl.* **2024**, *83*, 47865–47887. [CrossRef]
22. Fang, J.; Li, Y.; Ji, P.N.; Wang, T. Drone Detection and Localization Using Enhanced Fiber-Optic Acoustic Sensor and Distributed Acoustic Sensing Technology. *J. Light. Technol.* **2023**, *41*, 822–831. [CrossRef]
23. Wang, B.; Li, Q.; Mao, Q.; Wang, J.; Chen, C.P.; Shangguan, A.; Zhang, H. A Survey on Vision-Based Anti Unmanned Aerial Vehicles Methods. *Drones* **2024**, *8*, 518. [CrossRef]
24. Aydin, B.; Singha, S. Drone Detection Using YOLOv5. *Eng* **2023**, *4*, 416–433. [CrossRef]
25. Coluccia, A.; Fascista, A.; Sommer, L.; Schumann, A.; Dimou, A.; Zarpalas, D. The Drone-vs-Bird Detection Grand Challenge at ICASSP 2023: A Review of Methods and Results. *IEEE Open J. Signal Process.* **2024**, *5*, 766–779. [CrossRef]
26. Rizzoli, G.; Barbato, F.; Caligiuri, M.; Zanuttigh, P. SynDrone—Multi-Modal UAV Dataset for Urban Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Paris, France, 2–6 October 2023; pp. 2210–2220.
27. Steininger, D.; Widhalm, V.; Simon, J.; Kriegler, A.; Sulzbachner, C. The Aircraft Context Dataset: Understanding and Optimizing Data Variability in Aerial Domains. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 3823–3832.
28. Lee, H.; Han, S.; Byeon, J.I.; Han, S.; Myung, R.; Joung, J.; Choi, J. CNN-Based UAV Detection and Classification Using Sensor Fusion. *IEEE Access* **2023**, *11*, 68791–68808. [CrossRef]
29. Mehta, V.; Dadboud, F.; Bolic, M.; Mantegh, I. A Deep Learning Approach for Drone Detection and Classification Using Radar and Camera Sensor Fusion. In Proceedings of the 2023 IEEE Sensors Applications Symposium (SAS), Ottawa, ON, Canada, 18–20 July 2023 ; pp. 1–6. [CrossRef]
30. Dudczyk, J.; Czyba, R.; Skrzypczyk, K. Multi-Sensory Data Fusion in Terms of UAV Detection in 3D Space. *Sensors* **2022**, *22*, 4323. [CrossRef]
31. CVPR 2023 Anti-UAV Challenge Dataset. In Proceedings of the The 3rd Anti-UAV Workshop & Challenge, Vancouver, BC, Canada, 18–22 June 2023.
32. Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Ye, Q.; Jiao, J.; Han, Z.; et al. Anti-UAV: A large-scale benchmark for vision-based UAV tracking. *IEEE Trans. Multimed.* **2021**, *25*, 486–500. [CrossRef]
33. Zhao, J.; Wang, G.; Li, J.; Jin, L.; Fan, N.; Wang, M.; Wang, X.; Yong, T.; Deng, Y.; Guo, Y.; et al. The 2nd anti-UAV workshop & challenge: Methods and results. *arXiv* **2021**, arXiv:2108.09909.
34. Huang, S.; Jiang, Y.; Jiang, Y. Design of Target Detection and Tracking System for Sports Video. *IEEE Access* **2024** . [CrossRef]
35. Zhu, H.; Wei, H.; Li, B.; Yuan, X.; Kehtarnavaz, N. A review of video object detection: Datasets, metrics and methods. *Appl. Sci.* **2020**, *10*, 7834. [CrossRef]
36. Jiao, L.; Zhang, R.; Liu, F.; Yang, S.; Hou, B.; Li, L.; Tang, X. New Generation Deep Learning for Video Object Detection: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 3195–3215. [CrossRef]
37. Mayer, C.; Danelljan, M.; Bhat, G.; Paul, M.; Paudel, D.P.; Yu, F.; Van Gool, L. Transforming Model Prediction for Tracking. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; Volume 2022. [CrossRef]
38. Kugarajeevan, J.; Kokul, T.; Ramanan, A.; Fernando, S. Transformers in Single Object Tracking: An Experimental Survey. *IEEE Access* **2023**, *11*, 80297–80326. [CrossRef]
39. Al-lQubaydhi, N.; Alenezi, A.; Alanazi, T.; Senyor, A.; Alanezi, N.; Alotaibi, B.; Alotaibi, M.; Razaque, A.; Hariri, S. Deep learning for unmanned aerial vehicles detection: A review. *Comput. Sci. Rev.* **2024**, *51*, 100614. [CrossRef]
40. Chen, X.; Yan, B.; Zhu, J.; Lu, H.; Ruan, X.; Wang, D. High-Performance Transformer Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8507–8523. [CrossRef] [PubMed]

41. Wang, N.; Zhou, W.; Wang, J.; Li, H. Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021. [CrossRef]
42. Barhate, M.M.; Inamdar, C.S.; Ingale, C.D.; Inamdar, Y.S.; Humne, S.S.; Mahendrakumar, H.I.; Hulenwar, H.P. Drone Detection Through CCTV. *Int. J. Res. Appl. Sci. Eng. Technol.* **2024**, *12*, 57107. [CrossRef]
43. Bhagat, P.N.; Dasarwar, H.V.; Sayyad, M.R.; Shelake, S.D. Drone Detection and Identification Using Artificial Intelligence. *Int. J. Adv. Res. Sci. Commun. Technol.* **2024** . [CrossRef]
44. Munir, F.; Azam, S.; Rafique, M.A.; Sheri, A.M.; Jeon, M.; Pedrycz, W. Exploring thermal images for object detection in underexposure regions for autonomous driving. *Appl. Soft Comput.* **2022**, *121*, 108793 . [CrossRef]
45. Kristo, M.; Ivasic-Kos, M.; Pobar, M. Thermal Object Detection in Difficult Weather Conditions Using YOLO. *IEEE Access* **2020**, *8*, 125459–125476. [CrossRef]
46. Batchuluun, G.; Kang, J.K.; Nguyen, D.T.; Pham, T.D.; Arsalan, M.; Park, K.R. Deep Learning-Based Thermal Image Reconstruction and Object Detection. *IEEE Access* **2021**, *9*, 5951–5971. [CrossRef]
47. Eltahan, M.; Elsayed, K. Enhancing Autonomous Driving By Exploiting Thermal Object Detection Through Feature Fusion. *Int. J. Intell. Transp. Syst. Res.* **2024**, *22*, 146–158. [CrossRef]
48. Jiang, C.; Ren, H.; Ye, X.; Zhu, J.; Zeng, H.; Nan, Y.; Sun, M.; Ren, X.; Huo, H. Object detection from UAV thermal infrared images and videos using YOLO models. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102912. [CrossRef]
49. Mebtouche, N.E.D.; Baha, N. Robust UAV detection based on saliency cues and magnified features on thermal images. *Multimed. Tools Appl.* **2023**, *82*, 20039–20058. [CrossRef]
50. Zou, X.; Peng, T.; Zhou, Y. UAV-Based Human Detection With Visible-Thermal Fused YOLOv5 Network. *IEEE Trans. Ind. Inform.* **2024**, *20*, 3814–3823. [CrossRef]
51. Yu, Q.; Ma, Y.; He, J.; Yang, D.; Zhang, T. A unified transformer based tracker for anti-uav tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3036–3046. [CrossRef]
52. Milosevic, N. *Introduction to Convolutional Neural Networks*; Apress: Berkeley, CA, USA, 2020 . [CrossRef]
53. Saxena, A. An Introduction to Convolutional Neural Networks. *Int. J. Res. Appl. Sci. Eng. Technol.* **2022**, *10*, 943–947. [CrossRef]
54. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv* **2024**, arXiv:2402.13616.
55. Xu, W.; Zhu, D.; Deng, R.; Yung, K.; Ip, A.W.H. Violence-YOLO: Enhanced GELAN Algorithm for Violence Detection. *Appl. Sci.* **2024**, *14*, 6712. [CrossRef]
56. Balakrishnan, T.; Sengar, S.S. RepVGG-GELAN: Enhanced GELAN with VGG-STYLE ConvNets for Brain Tumour Detection. *arXiv* **2024**, arXiv:2405.03541.
57. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. [CrossRef] [PubMed]
58. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [CrossRef]
59. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* ; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany , 2020; Volume 12346 LNCS. [CrossRef]
60. Li, Y.; Mao, H.; Girshick, R.; He, K. Exploring Plain Vision Transformer Backbones for Object Detection. In *Computer Vision—ECCV 2022, Proceedings of the 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX* ; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2022; Volume 13669 LNCS. [CrossRef]
61. Yang, X.; Wang, G.; Hu, W.; Gao, J.; Lin, S.; Li, L.; Gao, K.; Wang, Y. Video Tiny-Object Detection Guided by the Spatial-Temporal Motion Information. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Vancouver, BC, Canada, 17–24 June 2023; Volume 2023. [CrossRef]
62. Caron, M.; Touvron, H.; Misra, I.; Jegou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. [CrossRef]
63. Tang, Z.; Gao, Y.; Xun, Z.; Peng, F.; Sun, Y.; Liu, S.; Li, B. Strong Detector with Simple Tracker. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023 ; Volume 2023. [CrossRef]
64. Biró, A.; Jánosi-Rancz, K.T.; Szilágyi, L.; Cuesta-Vargas, A.I.; Martín-Martín, J.; Szilágyi, S.M. Visual Object Detection with DETR to Support Video-Diagnosis Using Conference Tools. *Appl. Sci.* **2022**, *12*, 5977. [CrossRef]

65. Hardt, M.; Recht, B.; Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In Proceedings of the International Conference on Machine Learning (PMLR 48), New York, NY, USA, 20–22 June 2016; pp. 1225–1234.
66. Choi, D. On empirical comparisons of optimizers for deep learning. *arXiv* **2019**, arXiv:1910.05446.