drones

MDPI

Review

Recent Research Progress on Ground-to-Air Vision-Based Anti-UAV Detection and Tracking Methodologies: A Review

Arowa Yasmeen * and Ovidiu Daescu

Department of Computer Science, The University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080-3021, USA
* Correspondence: arowa.yasmeen@utdallas.edu

Abstract: Unmanned Aerial Vehicles (UAVs) are increasingly gaining popularity, and their consistent prevalence in various applications such as surveillance, search and rescue, and environmental monitoring requires the development of specialized policies for UAV traffic management. Integrating this novel aerial traffic into existing airspace frameworks presents unique challenges, particularly regarding safety and security. Consequently, there is an urgent need for robust contingency management systems, such as Anti-UAV technologies, to ensure safe air traffic. This survey paper critically examines the recent advancements in ground-to-air vision-based Anti-UAV detection and tracking methodologies, addressing the many challenges inherent in UAV detection and tracking. Our study examines recent UAV detection and tracking algorithms, outlining their operational principles, advantages, and disadvantages. Publicly available datasets specifically designed for Anti-UAV research are also thoroughly reviewed, providing insights into their characteristics and suitability. Furthermore, this survey explores the various Anti-UAV systems being developed and deployed globally, evaluating their effectiveness in facilitating the integration of small UAVs into low-altitude airspace. The study aims to provide researchers with a well-rounded understanding of the field by synthesizing current research trends, identifying key technological gaps, and highlighting promising directions for future research and development in Anti-UAV technologies.

Keywords: anti-UAV; UAV detection; UAV tracking; UAV monitoring

Academic Editor: Xiwang Dong

Received: 5 December 2024
Revised: 11 January 2025
Accepted: 14 January 2025
Published: 15 January 2025

Citation: Yasmeen, A.; Daescu, O. Recent Research Progress on Ground-to-Air Vision-Based Anti-UAV Detection and Tracking Methodologies: A Review. *Drones* **2025**, *9*, 58. https://doi.org/10.3390/drones9010058

1. Introduction

In 2019, a drone attack on Saudi Arabia's largest oil refinery led to an almost 5% decline in global oil supply, highlighting the critical need for effective Anti-Unmanned Aerial Vehicle (Anti-UAV) systems [1]. This incident exemplifies the increased security challenges posed by the growing popularity of Unmanned Aerial Vehicles (UAVs) in both civilian and commercial sectors. As of 2024, the global UAV market has burgeoned into a USD 35.28 billion industry, with projections indicating a compound annual growth rate (CAGR) of 13.9% to reach USD 67.64 billion by 2029 [2]. The Federal Aviation Administration (FAA) forecasts that by 2027, the United States alone will see a recreational UAV fleet of 1.82 million units and a commercial fleet approaching 955,000 [3].

The popularity of UAVs can be attributed to their diverse applications, such as infrastructure inspection, precision agriculture, emergency services, and goods delivery [4]. However, this integration of UAVs into the national airspace presents significant challenges regarding regional and national security, safe airspace operations, and privacy considerations. The current Air Traffic Management (ATM) system is not equipped with the

necessary infrastructure to manage the anticipated scale of UAV operations. Thus, there is a pressing need for the development of a specialized Unmanned Aerial Systems Traffic Management (UTM) system.

Anti-UAV systems, an integral part of the UTM framework, detect and track rogue UAVs. Rogue UAVs are unauthorized or non-cooperative drones that pose substantial threats to public safety, privacy, and national security. These threats range from accidental intrusions into restricted airspace (Figure 1), deliberate acts of espionage [5], to potential disruptions to legitimate UAV traffic (Figure 2). The complexity of the Anti-UAV problem is layered. It must balance security needs with privacy concerns, bystander safety, and the continuity of regular UAV operations.

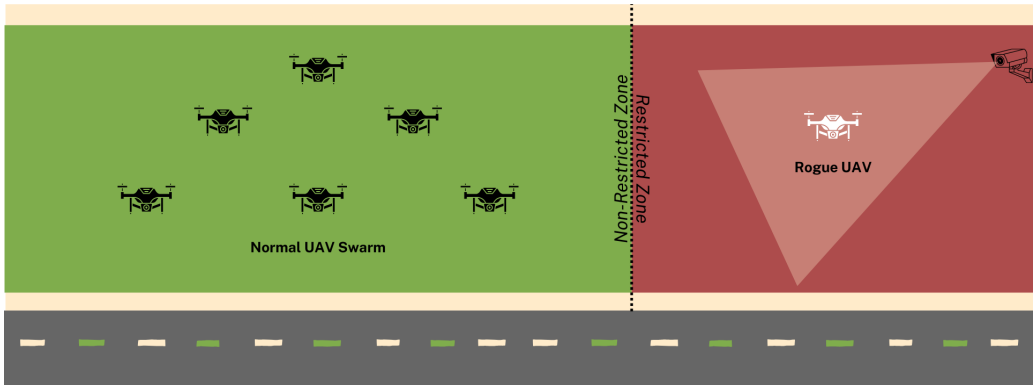

Figure 1. Anti-UAV scenario where a rogue UAV flies into a No Flying Zone.

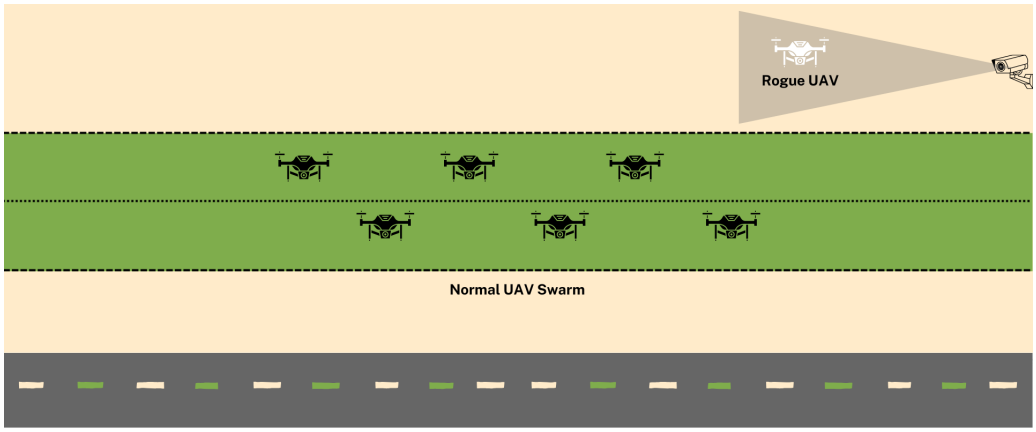

Figure 2. Anti-UAV scenario where a rogue UAV flies beyond the segregated airspace allotted for UAVs.

Existing Anti-UAV methods, which rely on modes such as radar [6,7], radio frequency [8,9], and acoustic sensing [10,11], are limited in effectively detecting and tracking small UAVs that have weak electromagnetic signatures [12]. These systems often struggle with detecting UAVs from a long distance, are susceptible to noise, and require a costly infrastructure [13–15]. In contrast, vision-based Anti-UAV systems offer a promising alternative, providing competitive speed, accuracy, and reliability at relatively lower equipment and computational costs.

This survey paper focuses on recent advancements in ground-to-air [16] vision-based Anti-UAV detection and tracking methodologies. The motivation behind focusing on ground-to-air vision-based systems stems from the anticipated commercialization of UAV traffic, which will prompt the widespread adoption of multi-layered 3D Air Corridors [17] in Class G airspace. These dedicated Air Corridors for UAV operations, illustrated by the green segregated air volumes in Figures 1 and 2, will require stationing ground-to-air Anti-

UAV systems, similar to the existing road traffic monitoring systems, as part of the larger UTM framework. This review paper differs from other existing review literature [18–21], as it focuses on methods and datasets exclusively from the ground-to-air perspective of building a robust Anti-UAV system that would be a sub-component of a larger UTM framework. In addition, this paper discusses both detection and tracking in an Anti-UAV scenario, unlike most of the existing literature [18,19,21] that only focuses on UAV detection.

As such, this paper analyzes state-of-the-art object detection frameworks, including one-stage and two-stage detectors, anchor-based and anchor-free methods, and lightweight architectures optimized for edge computing. Additionally, it reviews tracking methodologies that provide long-term tracking in complex and dynamic environments. Moreover, this paper presents a comprehensive overview of the publicly available UAV detection and tracking datasets. This survey aims to provide researchers and practitioners with a holistic understanding of the current trends, challenges, and technological gaps in Anti-UAV systems.

The main contributions of this paper can be summarized as follows:

- A critical analysis of the most recent vision-based ground-to-air UAV detection and tracking techniques, highlighting their strengths, limitations, and trade-offs.
- A detailed overview of publicly available UAV datasets, presenting their characteristics and how they address (or fail to address) specific research challenges.
- Identification of existing gaps in the literature and areas where further research is needed.

The remainder of this paper is structured as follows: Section 2 examines UAV detection techniques, followed by UAV tracking methodologies in Section 3. Section 4 provides an overview of relevant datasets. Section 5 discusses current challenges and future research directions, and Section 6 concludes the paper.

2. UAV Detection

Detecting Unmanned Aerial Vehicles (UAVs) presents a unique set of challenges, distinguishing it from traditional object detection tasks. These challenges arise due to the inherent characteristics of UAVs and the diverse environments in which they operate. Thus, specialized approaches for effective detection and tracking in Anti-UAV systems are necessary.

One of the primary challenges in UAV detection is the small size of the target objects. This means that UAVs amount to a very small number of pixels in images captured from ground-based cameras, especially at long distances. This small size makes it difficult to distinguish UAVs from other small airborne objects such as birds or even image noise, particularly against complex backgrounds or in unfavorable lighting conditions [22]. The diverse shapes and designs of UAVs make the detection task even more complex. UAVs come in various configurations, including multi-rotor, fixed-wing, and hybrid designs, each with distinct visual characteristics [23]. This variety makes it challenging to develop a single detection model that can effectively detect all types of UAVs. Real-time performance is also critical for effective Anti-UAV systems. The high-speed nature of UAV operations demands detection algorithms to process the input with minimal latency. This need for real-time processing often conflicts with the computational complexity required for accurate detection, especially when dealing with high-resolution images to identify small, distant UAVs. Addressing these challenges requires approaches that balance accuracy, speed, and robustness.

The emergence of deep learning methodologies for computer vision tasks has catalyzed the widespread adoption of either deep convolutional neural networks (DCNNs) or Transformer-based architectures for object detection applications. As such, most UAV detection frameworks utilize different versions of popular DCNN-based or Transformer-

based object detectors. Object detectors can be classified in two main ways: one-stage vs. two-stage detectors and anchor-based vs. anchor-free detectors.

2.1. Two-Stage vs. One-Stage

Two-stage and one-stage detectors each have their own distinctive method of handling the object detection process as illustrated in Figure 3. Two-stage object detectors start by having a region proposal network generate a set of region proposals that may contain objects along with a feature extractor network extracting relevant features. These region proposals are then refined and classified based on the features in the second stage by a different model to produce bounding boxes and class scores. On the other hand, one-stage detectors bypass the region proposal step and directly predict the bounding boxes and class scores in a single pass through the model.

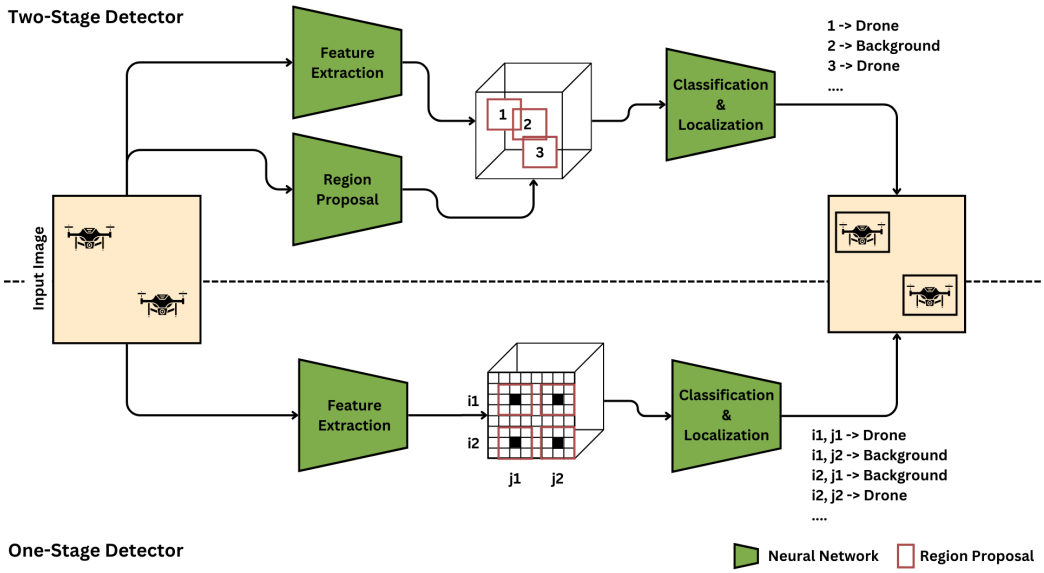

Figure 3. One-stage detector vs. two-stage detector.

Two-stage detectors are known for their higher accuracy, especially for complex tasks and challenging datasets. However, they are slower and more computationally expensive [13]. In contrast, one-stage detectors are faster and more suitable for real-time applications. They are less computationally expensive due to the absence of an intermediate step. However, they come with the trade-off of lower accuracy compared to two-stage detectors. Notably, Region-based Convolutional Neural Network (R-CNN)-based methods [24–26] or methods with a region proposal network included in their architecture are two-stage detectors, while You Only Look Once (YOLO)-based methods [27–29], Single Shot Detector (SSD) [30], RetinaNet [31], and RefineDet [32] are one-stage detectors. Transformer-based methods such as Ghostformer [33] and Detection Transformer (DETR) [34] are two-stage and one-stage detectors, respectively.

2.2. Anchor-Based vs. Anchor-Free

The difference between anchor-based and anchor-free object detectors lies in how they predict bounding boxes for detected objects. Anchor-based detectors use predefined bounding boxes called "anchors" or "priors", which have different scales, aspect ratios, and positions across the image grid [24]. During training, these anchors are matched to ground-truth objects, and the model learns to adjust these anchors to predict the final bounding box. The purpose of anchors is to help the model handle objects of different sizes, shapes, and aspect ratios [24]. Despite the advantages, anchor boxes may lead to inefficiencies due to their large number and because they sometimes require initial design from an expert [35].

On the other hand, anchor-free detectors directly predict the object's center point [36], dimensions, and/or key points to predict the bounding boxes from them [37,38]. Such models typically predict the center of the object, and from there, they calculate the size and position of the bounding box. Some models prefer to predict corner points instead of center points (CornerNet) [37] or key point triplets (CenterNet) [38]. Anchor-free models are more efficient due to their simpler design but may struggle with objects of varying scales and aspect ratios compared to anchor-based models.

Table 1 presents an overview of recent research in the Anti-UAV domain, categorized based on detection approaches. A clear trend can be observed. Most studies use one-stage and anchor-based methods over alternative approaches. This is because of the resource-constrained nature of Anti-UAV systems, which requires detection models to be lightweight. The use of YOLO-based methods across the reviewed works further affirms the validity of this trend. To mitigate the trade-off in accuracy, researchers have introduced modifications to the network architecture to enhance detection performance and reduce network complexity (discussed in Section 2.3).

Table 1. Stage and anchor types used in recent works.

Ref.	Year	Base Model	Stage Type	Anchor Type
[39]	2019	YOLOv3	One-Stage	Anchor-based
[40]	2020	EXTD	One-Stage	Anchor-free
[41]	2020	YOLOv4	One-Stage	Anchor-based
[42]	2021	Faster R-CNN	Two-Stage	Anchor-based
[42]	2021	YOLOv3	One-Stage	Anchor-based
[42]	2021	SSD	One-Stage	Anchor-based
[13]	2021	Faster R-CNN	Two-Stage	Anchor-based
[13]	2021	YOLOv3	One-Stage	Anchor-based
[13]	2021	SSD512	One-Stage	Anchor-based
[13]	2021	DETR	One-Stage	Anchor-free
[43]	2021	YOLOv4	One-Stage	Anchor-based
[44]	2021	YOLOv4	One-Stage	Anchor-based
[45]	2022	Faster R-CNN	Two-Stage	Anchor-based
[45]	2022	Cascade R-CNN	Two-Stage	Anchor-based
[45]	2022	ATSS	N/A	N/A
[45]	2022	YOLOX	One-Stage	Anchor-free
[45]	2022	SSD	One-Stage	Anchor-based
[46]	2022	YOLOv5s	One-Stage	Anchor-based
[47]	2022	YOLOv3	One-Stage	Anchor-based
[48]	2022	YOLOv5	One-Stage	Anchor-based
[49]	2023	YOLOv4	One-Stage	Anchor-based
[50]	2023	YOLOv4	One-Stage	Anchor-based
[51]	2023	YOLOX-nano	One-Stage	Anchor-free
[52]	2024	YOLOv8	One-Stage	Anchor-free
[53]	2024	YOLOv7-tiny	One-Stage	Anchor-based

2.3. Detection Network Architecture

The architectural framework of object detection systems is typically composed of three key components, the backbone, neck, and head, as depicted in Figure 4 [54]. The backbone is the primary feature extractor that derives relevant information from the input image. This component is usually a convolutional neural network (CNN) pre-trained on large-scale image classification datasets. It captures hierarchical features across different scales, with

the initial layers focusing on low-level details such as edges and textures and deeper layers capturing high-level features like object parts and semantic content.

The neck joins the backbone and the head. It is responsible for aggregating and refining the features extracted by the backbone, focusing on enhancing the spatial and semantic information across multiple scales and providing it to the head. The final component, the head, makes predictions based on the features passed through the neck. It consists of one or more task-specific subnetworks dedicated to classification and localization. A post-processing step, such as non-maximum suppression (NMS), is often employed to filter overlapping predictions and retain only the most confident detections.

While many architectures incorporate a neck module, certain detectors, such as YOLOv8, omit its use. Researchers frequently modify the backbone, neck, and head components to improve detection performance and optimize the model for specific tasks. Common backbone architectures identified in the recent literature include variants of DarkNet [55], Visual Geometry Group (VGG) [56], ResNet [57], EfficientNet [58], DenseNet [59], MobileNet [60], and newer designs like Extremely Tiny Face Detector (EXTD) [61] and Mobile Vision Transformer (MobileViT) [62]. Popular neck structures often include Path Aggregation Network (PANet) [63], and Feature Pyramid Network (FPN) [64]. Modifications to the head typically involve adjustments to loss functions, with the type of head determining whether the detector is a one-stage or two-stage model. Specifically, dense predictor heads are used in one-stage models, whereas sparse predictor heads are characteristic of two-stage models [28].

Additionally, data augmentation techniques are extensively employed to enhance the generalization of the models. Techniques commonly referenced in the literature include image cropping, rotation, flipping, blur, hue manipulation, saturation adjustments, exposure manipulation, MixUp [65], Mosaic [28], and CutMix [66].

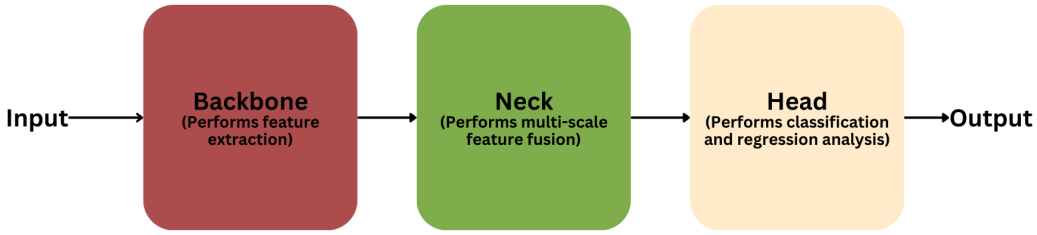

Figure 4. Standard one-stage detection architecture.

2.3.1. Anchor Modification

Optimizing anchor box configurations is one widely adopted strategy to enhance the efficiency and accuracy of object detection networks. Anchor boxes, which represent predefined bounding boxes used to traverse regions of interest (ROIs) in an image, play a critical role in determining the optimal bounding box during object detection. By carefully adjusting the anchor box settings, the model can better learn the spatial characteristics of the target objects, leading to improvements in detection speed and accuracy [67].

Researchers use clustering algorithms, such as k-means and its variant k-means++, to refine anchor box parameters by clustering data samples based on their size and aspect ratios. This method effectively tailors the anchor boxes to the dataset's specific distribution of object sizes, resulting in more precise predictions. For instance, Y. Hu et al. [39], Q. Cheng et al. [50], and C. Bo et al. [53] employ this approach in their respective studies, demonstrating that optimizing anchor box dimensions through clustering leads to significant performance gains. The models exhibit enhanced accuracy in predicting bounding boxes and faster inference times due to better alignment between the anchor boxes and the ground truth in the images.

2.3.2. Backbone Modification

Researchers commonly modify the backbone architecture to improve the feature extraction capability of detection models, particularly in the context of detecting small and fast-moving objects like UAVs. These modifications typically aim to improve model performance or make the models more suitable for specific environments, such as edge computing.

H. Sun et al. [40] propose TIB-Net, a model that uses the EXTD backbone further enhanced with a cyclic pathway and a Spatial Attention Module (SAM). EXTD is a lightweight model, making it highly efficient and suitable for detecting small objects in an edge computing environment. The cyclic pathway allows the model to revisit low-level feature information, crucial for detecting extremely small objects like UAVs. In higher/deeper layers, the model might lose information related to small objects in the image due to subsequent pooling operations. Therefore, by allowing the model to revisit the low-level features, the model can now extract better features. Furthermore, SAM enhances the model's robustness to noise by capturing better spatial information, making TIB-Net particularly effective in handling images with blurring, a common challenge in real-world fast-moving UAV detection.

In a different approach, H. Liu et al. [44], X. Zhou et al. [49], and Q. Cheng et al. [50] modify the YOLOv4 architecture to improve Anti-UAV detection capabilities. H. Liu et al. focus on pruning the convolutional channels and shortcut layers to make the model more lightweight. By reducing the complexity of the model via pruning, they improved its efficiency without sacrificing detection accuracy. X. Zhou et al. expand the backbone modifications by incorporating SAM, allowing the model to capture multi-scale features better. SAM enables the model to better focus on important features while minimizing noise distractions. This enhancement allows the network to identify small UAVs in cluttered environments, where irrelevant background information can lead to false positives. In contrast, Q. Cheng et al. replace the YOLOv4 backbone with MobileViT, a more lightweight transformer-based backbone architecture. This swap ensures the model can still extract local and global information from the input and generate multi-scale feature maps while remaining relatively lightweight.

Similarly, L. Yaowen et al. [46] also modify the backbone of YOLOv5s by integrating the Simple parameter-fee Attention Module (SimAM) and the Ghost module. SimAM improves feature extraction capabilities while the Ghost module uses Ghost convolutions [68] to significantly reduce the number of parameters in the model without compromising performance. This adjustment makes the model suitable for edge computing. B. Liu et al. [48] take a different approach by replacing the YOLOv5s backbone with EfficientLite [69], a backbone known for its lightweight architecture. This change allows the model to achieve faster inference times while maintaining accuracy, making it more promising for real-time UAV detection.

C. Wang et al. [70] propose a novel lightweight UAV swarm detection method based on the YOLOX model. To enhance the YOLO backbone's feature extraction capability, they introduce the Squeeze-and-Excitation (SE) attention mechanism into the Cross-Stage-Partial-connections DarkNet (CSPDarkNet) backbone. The SE module dynamically generates different weight coefficients for each feature channel by leveraging the inter-channel correlations, enhancing the most relevant features for the task while suppressing less important ones.

Lastly, C. Bo et al. [53] modify the YOLOv7-tiny backbone by incorporating an SPPF (Spatial Pyramid Pooling—Fast) module along with SimAM. These modifications reduce feature loss and minimize confusion in complex scenes by improving the model's focus on small target UAV areas.

Researchers modify backbone architectures by integrating attention mechanisms, pruning, or replacing components with lightweight alternatives to better adapt the models to the unique challenges posed by UAV detection. These modifications improve detection performance and enhance model efficiency, making them more suitable for deployment in real-world, resource-constrained environments.

2.3.3. Neck Modification

The neck's role in an object detection network is to perform relevant feature aggregation and fusion. It is often modified to improve the detection of UAVs, which may appear in various scales in the input. Researchers have explored various strategies to modify the neck architecture to enhance the model's ability to capture multi-scale features and handle complex scenes with small or distant targets.

Y. Hu et al. [39] and H. Zhai et al. [47] utilize a slightly modified YOLOv3 framework in their work. They adjust the number of scales in the feature maps from three to four. This modification allows their models to capture more texture and contour information, which is particularly beneficial for small object detection, such as UAVs.

Q. Cheng et al. [50] improve the neck of YOLOv4 by introducing Coordinate Attention (CA) [71] into the PANet structure. This modification allows for better information fusion between low- and high-dimensional features, enhancing the model's ability to extract and fuse information. By using multi-scale attention, the network can better focus on UAVs with small or complex visual patterns by creating direction-aware and position-sensitive attention maps. Similarly, L. Yaowen et al. [46] introduce the SimAM module to the neck of YOLOv5s, improving feature extraction through enhanced attention mechanisms. Additionally, they incorporate the Ghost module to reduce the number of model parameters, which helps maintain a lightweight model architecture without compromising performance. Ghost convolutions are also used by M. Huang et al. [52]. The authors take a novel approach by replacing the Convolution to Feature (C2f) module in the neck with the C3Ghost module, enhancing the feature extraction capability. This modification enables the model to capture richer semantic and contextual information while also reducing computational complexity. Additionally, they include Efficient Multi-scale Attention (EMA) [72], a variant of CA, to further enhance the network's ability to process UAV features across multiple scales, boosting its robustness in detecting UAVs under varying conditions.

Another example of enabling the network to better distinguish important features from less relevant ones is C. Wang et al. [70], who include the Convolutional Block Attention Module (CBAM) [73] in the neck. They add the CBAM to the PANet in the neck. This attention mechanism allows the model to prioritize essential visual information, leading to the more accurate detection of UAVs with cluttered backgrounds.

X. Zhou et al. [49] propose the integration of Spatial Pyramid Pooling (SPP)S [74] and ResNeck modules into the neck of YOLOv4, optimizing the network by compressing it while simultaneously improving detection speed and accuracy. Adding these modules enhances the model's capacity to handle multi-scale feature maps, making it more adept at detecting UAVs, which often appear in varying sizes across frames.

Finally, C. Bo et al. [53] introduce significant changes to the neck of YOLOv7-tiny by replacing the standard FPN architecture with a Get-and-Send module, complemented by InceptionNeXT [75] modules. The InceptionNeXT modules expands the receptive field, facilitating a deeper integration and understanding of the features learned in preceding layers while improving computational efficiency. Meanwhile, the Get-and-Send module tackles the issue of information loss of small UAVs that can occur during cross-layer feature fusion in FPN, as it has many pathways and indirect interactions. This modification combination enables more uniform aggregation and fusion of features from various levels,

boosting the network's information fusion capabilities without adding significant latency. As a result, the model maintains high accuracy and robustness in detecting small UAVs while minimizing computational overhead.

The current trend involves the utilization of various attention-based mechanisms like CBAM, SE, SimAM, CA, and EMA to enhance feature aggregation and fusion for small object detection. Additionally, alternative forms of the traditional convolution operation, such as Ghost convolutions, InceptionNeXT, and depthwise separable convolutions, are being utilized to maintain a lightweight architecture.

2.3.4. Head Modification

The head of an object detection network performs the final classification and localization of the detected objects. To enhance the head's performance, particularly for small UAVs, researchers focus on refining bounding box regression accuracy, improving feature fusion, and introducing new loss functions to more effectively measure prediction errors.

In traditional anchor-based detectors, the predicted bounding boxes are derived by regressing offsets between the ground truth boxes and predefined anchor boxes. Intersection over Union (IoU) loss, a standard bounding box regression loss function, computes the difference between the predicted bounding boxes and ground truth. However, to address the limitations in standard IoU loss, several researchers have proposed alternative approaches. For instance, C. Wang et al. [70] replace IoU loss with Distance-IoU (DIoU) loss to improve regression accuracy. DIoU considers the overlap between bounding boxes and the distance between their center points, resulting in better optimization and faster convergence during training. On the other hand, L. Yaowen et al. [46] make a change to the standard DIoU loss function by replacing it with the α-DIoU loss. This change aims to improve how accurately the model finds bounding boxes, especially when trained on smaller datasets. The α parameter allows the model to give more importance to objects with higher IoU, which helps improve the precision of bounding box regression. By adjusting α, the model becomes better at handling difficult localization tasks, especially when further refinement is needed even when IoU values are already high.

B. Liu et al. [48] introduce Adaptive Spatial Feature Fusion (ASFF) [76] into the head architecture, which facilitates the combination of feature maps at different spatial resolutions. This modification compensates for the accuracy loss that could arise from replacing the model's backbone with the lighter Efficientlite architecture. By leveraging ASFF, the model can selectively fuse features at various scales, ensuring that important spatial details are preserved even with a smaller model size. Furthermore, B. Liu et al. introduce an angular constraint into the original regression loss function, mitigating the mismatch between predicted and actual bounding box orientations. This adjustment not only improves localization accuracy but also accelerates network convergence during training, making the model more efficient.

M. Huang et al. [52] propose replacing the standard detection heads in YOLOv8 with DDetect, a head architecture that utilizes Deformable Convolution v2 (DCNv2) [77]. DCNv2 allows the model to adaptively adjust the shape and size of convolutional kernels based on the input features, providing greater flexibility in handling varying object shapes and sizes. This modification is especially useful for detecting UAVs, which can vary in size and orientation based on their distance from the camera and the scene conditions.

2.4. Critical Analysis

The reviewed works demonstrate a clear trend toward enhancing the detection of UAVs through various architectural modifications. The modifications to the backbone, neck, and head modules of the detectors significantly impact their performance in terms

of accuracy, speed, and model size, allowing researchers to address the unique challenges posed by UAV detection. The primary evaluation metrics for comparing detection networks are Mean Average Precision (mAP) for accuracy and Frames Per Second (FPS) for speed:

$$Precision = \frac{TP}{TP + FN} \tag{1}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{2}$$

Here, TP stands for true positives, FN stands for false negatives, N stands for the number of classes, and AP_i denotes the Average Precision for the i_{th} class. FPS indicates the number of frames a model processes per second, serving as an estimate of the algorithm's speed.

Backbone modifications are central to improving the feature extraction capability of models, particularly for small object detection. For instance, H. Sun et al.'s [40] TIB-Net model employs a lightweight EXTD backbone, enhanced by a cyclic pathway and Spatial Attention Module (SAM), to capture low-level feature information critical for small UAV detection. Despite its small size (697 KB), TIB-Net achieves competitive accuracy (89.2% mAP), closely rivaling larger models like the two-stage detector Cascade R-CNN with ResNet50 (90.1% mAP).

Q. Cheng et al.'s [50] replacement of the YOLOv4 backbone with MobileViT underscores the advantage of using lightweight backbones that maintain high detection accuracy and speed. Their modified YOLOv4-MCA architecture achieves 92.81% mAP at 40 FPS, outperforming the standard YOLOv4 with CSPDarkNet-53 (92.45% mAP, 26 FPS). Similarly, the Efficientlite backbone employed by B. Liu et al. [48] balances accuracy and model size, maintaining strong detection capabilities despite the lightweight design.

In contrast, models using heavier backbones, such as Cascade R-CNN [78] with ResNet50, achieve higher accuracy but at the cost of increased model size and slower speeds. While these models may be suitable for high-precision tasks in well-resourced environments, lightweight backbones like MobileViT and EXTD offer a more practical solution for real-time UAV detection in constrained scenarios and often exhibit similar levels of accuracy if suitable modifications are made. This observation has also been reinforced in the works by B.K. Isaac-Medina et al. [13] and Y. Zeng et al. [42], where they test the performance of two-stage models and one-stage models with different combinations of lightweight and heavy backbones on different datasets. Their experiments prove that while heavier two-stage architectures like Faster R-CNN or Transformer-based detectors like DETR achieve high levels of accuracy, competitive performance can also be achieved by one-stage methods with comparatively lighter backbones such as YOLOv3.

Neck modifications focus on improving feature aggregation and multi-scale adaptability, both of which are crucial for detecting small objects like UAVs. M. Huang et al.'s [52] EDGS-YOLOv8 model combines Ghost convolutions with multi-scale attention mechanisms, achieving an impressive 97.1% mAP at 56.2 FPS. The reduction in computational complexity enabled by Ghost convolutions highlights the trade-off between model efficiency and performance, which then needs to be mitigated by including additional modifications such as an attention mechanism.

Across these works, the integration of attention mechanisms stands out as a key factor in improving both the accuracy and robustness in UAV detection. These mechanisms enable models to focus on critical features and adapt to varying object scales and positions. This is crucial for detecting small UAVs in cluttered or dynamic environments. The popularity of attention-based modifications in recent work reflects a broader trend in computer vision research, emphasizing the importance of feature prioritization and fusion for small object

detection. Furthermore, lightweight architectures consistently demonstrate that model efficiency does not need to come at the expense of accuracy. Models like TIB-Net, YOLOv4-MCA, and EDGS-YOLOv8 prove that with the right combination of backbone, neck, and head modifications, small and efficient models can achieve performance levels comparable to or better than heavier, more resource-intensive architectures.

The reviewed works reveal a strong focus on adapting standard object detection frameworks to the specific challenges of UAV detection. By refining backbone, neck, and head components, researchers have successfully developed models that balance accuracy, speed, and efficiency, making them well suited for real-time applications in UAV detection and tracking. Table A1 offers a comprehensive summary of the reviewed works, facilitating a more effective comparison.

3. UAV Tracking

In conjunction with UAV detection, the tracking of Unmanned Aerial Vehicles (UAVs) is a critical component of Anti-UAV systems. Tracking unauthorized UAVs is paramount for maintaining airspace safety, security, and integrity. UAVs that violate air traffic regulations, such as altitude restrictions or geofencing limitations, present significant risks, and efficient tracking mechanisms enable authorities to identify and respond to such potential threats swiftly. However, despite the need for such tracking systems, there has been little recent work in this domain area.

UAV tracking methodologies can be broadly classified into two main types: Single-Object Tracking (SOT) and Multi-Object Tracking (MOT).

- SOT focuses on tracking the trajectory of a single UAV over time, even as it moves through varying backgrounds or experiences occlusion.
- MOT is concerned with tracking multiple UAVs simultaneously, which adds complexity due to interactions between UAVs and occlusion.

Traditionally, UAV tracking relied on correlation filters and template-matching techniques. For example, methods like Kernelized Correlation Filters (KCFs) [79] and Dual Correlation Filters (DCFs) [79] have been widely used for general object tracking. These techniques are efficient and lightweight, making them suitable for real-time applications. However, they often suffer from low robustness when dealing with fast-moving objects, occlusions, and complex backgrounds, which are common in UAV tracking scenarios.

With the rise of deep learning, more advanced tracking algorithms have been developed, offering improved performance in complex environments. Deep learning-based methods and Siamese networks [80] have shown remarkable success in UAV tracking. These methods can tackle the many challenges in Anti-UAV tracking, such as occlusion, scale variations, and changes in the UAV appearance over time [81].

UAVs often occupy only a few pixels in an image, especially at a distance from the surveillance camera. This makes it difficult for traditional tracking models to maintain accuracy over time. Deep learning-based approaches have significantly improved tracking small objects, but challenges remain, particularly in low-resolution video or when UAVs are far from the camera. Occlusion is a major challenge in UAV tracking, especially in urban or forested environments where buildings, trees, or other objects can temporarily block UAVs. Fully robust occlusion handling remains an open challenge, particularly in real-time scenarios. While real-time tracking is critical in Anti-UAV applications, achieving this without compromising accuracy is difficult. Thus, a potential research scope exists to balance these factors, especially in resource-constrained environments with limited computational power.

3.1. Single-Object Tracking

To tackle such challenging scenarios, F. Cheng et al. [81] and C. Wang et al. [82] propose two different tracking networks. F. Cheng et al. introduce a long-term object tracking method that utilizes a Siamese network, SiamRPN++ [83], and a re-detection module based on YOLOv5. Their proposed method incorporates a hybrid attention mechanism and a hierarchical discriminator to improve feature learning and generate more distinct object representations. Long-term tracking introduces additional challenges, as the UAV may leave and re-enter the camera's field of view, requiring the tracker to relocate the UAV after periods of occlusion. To address the challenges of re-locating the UAV target and updating the template in long-term tracking, the authors employ a hierarchical discriminator to produce response maps for target localization and a reliability criterion to assess the credibility of the produced response maps. When the algorithm detects low-confidence output results, the re-detection module is activated, and the template is adaptively updated.

In contrast, C. Wang et al. employ the ATOM (Accurate Tracking by Overlap Maximization) [84] framework, which focuses on maximizing the overlap between the predicted bounding box and the ground truth by using a combination of an overlap prediction network and a classification component along with the SE attention mechanism to enhance the extraction of features. The strength of ATOM lies in its ability to fine-tune the location of the UAV in each frame. Additionally, the authors integrate an occlusion-sensing module to assess the state of the target, determining whether it is occluded or not. This determination activates a trajectory prediction network that utilizes Long Short-Term Memory (LSTM) to predict the position of the UAV.

3.2. Multi-Object Tracking

As UAV swarms become more common, tracking multiple UAVs simultaneously poses new challenges. MOT methods like ByteTrack and Graph Networks for Multi-Object Tracking (GNMOT) have shown promise in multi-UAV scenarios. ByteTrack focuses on associating every detection box, including low-score ones, with a UAV to recover true objects and filter out false positives. It primarily relies on motion information for associations rather than feature extraction and similarity matching. GNMOT leverages graph neural networks to model the relationships between objects and their interactions over time. This approach focuses on using graph structures to enhance the tracking performance. However, scaling these systems to handle large numbers of UAVs in dense airspace or segregated flight corridors remains a significant research area. The complexity of tracking increases with interactions between UAVs, occlusions, and background clutter, making it difficult for current systems to maintain high accuracy.

3.3. Tracking Methodologies

Siamese network architectures have gained considerable attention in UAV tracking due to their ability to learn a similarity metric between the target object and candidate regions in subsequent frames. SiamFC [80], SiamRPN++ [83], Skimming-Perusal Tracking (SPLT) [85], Long-term Tracking with Meta-Updater (LTMU) [86], Discriminative Model Prediction (DiMP) [87], Transformer Tracking (TransT) [88], and GlobalTrack [89] are all Siamese-based tracking methods that can be used for UAV Tracking. These networks extract feature embeddings from both the initial UAV (from the first frame) and candidate patches in later frames, comparing them using a similarity metric. Curators of the DUT Anti-UAV dataset [45] show that these tracking methods prove to be effective for the UAV tracking objective.

Other methods, such as efficient convolution operators (ECOs), can also be used. ECO is a correlation filter-based tracker that improves efficiency and accuracy using efficient

convolution operators. It combines the strengths of the traditional correlation filters with deep learning features to achieve robust performance.

A novel training strategy is also proposed by N. Jiang et al. [90]. They propose a novel Dual-Flow Semantic Consistency (DSFC) training strategy that includes a Class-Level Semantic Modulation (CSM) module and an Instance-level Semantic Modulation (ISM) module. The CSM module is tasked with finding candidate UAV bounding boxes, while the ISM module focuses on differentiating multiple instances from the complex background. They fine-tune GlobalTrack using the Anti-UAV dataset and their DSFC training strategy and show that DSFC achieves superior performance compared to traditional training. At the same time, the authors' experiments also show that using the DSFC training strategy does not contribute to the inference speed.

Different tracking methodologies are often assessed using two main aspects: precision plots and success plots as defined in [91]. Precision plots are designed to track the center location error of tracked targets. This error is defined as the Euclidean distance between the predicted center of the target and the actual ground truth center. Precision plots display the percentage of frames in which the estimated location falls within a specified threshold distance of the ground truth. Success plots show how much the predicted bounding boxes overlap with the actual bounding boxes. To evaluate a tracker's performance across different frames, we count how many frames are successful—meaning their overlap is above a specific threshold. The success plot displays the ratio of successful frames at thresholds ranging from 0 to 1. Using just one success rate at a single threshold may not provide a fair evaluation of the tracker. Therefore, we calculate the area under the curve (AUC) for each success plot to rank the tracking algorithms. Multiple Object Tracking Accuracy (MOTA) [92] serves as a metric for assessing the effectiveness of object tracking systems over time. It emphasizes the importance of identity tracking, wherein the goal is to consistently recognize individual objects. The concept of identity switches (IDSWs) is significant; these occur when a single ground truth (GT) object is assigned different track predictions across various time frames. The computation of MOTA incorporates temporal dependencies and penalizes modifications in track assignments between consecutive frames. An IDSW is recorded when a ground truth target i is matched with track j in the current frame but was matched to a different track k ($k \neq j$) in the preceding frame:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t} \tag{3}$$

Here, t is the frame index, FN denotes false negatives, FP denotes false positives, $IDSW$ denotes identity switches, and GT refers to the ground truth. A quantitative performance comparison of all the discussed algorithms and methodologies is summarized in Table 2.

The authors of the referenced papers all emphasize the importance of integrating detection with the tracking methodology, highlighting its ability to produce superior results when compared to using standalone trackers for long-term tracking. This prompts an important question regarding the impact of the detection module's performance on the overall effectiveness of the tracker. It is worth noting that the current research on UAV tracking is primarily focused on scenarios with a single-camera view and tracking any UAV within the field of view. However, there is a clear need for further research in the field of UAV tracking with camera hand-off capabilities and the ability to detect a rogue UAV from UAV traffic in a surveillance area equipped with a camera network. This becomes particularly relevant in the context of large-scale UTM systems, where seamless camera hand-off and coordination among multiple cameras are crucial for comprehensive and effective surveillance.

Table 2. Summary of UAV tracking methodologies used in recent works.

Ref.	Techniques Used	Dataset	Best Accuracy Metrics Reported	Comments
[81]	1. Uses SiamRPN++ 2. Includes a re-detection module based on YOLOv5 that uses a hybrid attention mechanism and hierarchical discriminator	Youtube-BoundingBoxes, ImageNet VID, ImageNet, COCO, Anti-UAV	67.7% AUC 88.4% Precision	This method is good for long-term tracking
[82]	1. Uses ATOM 2. Includes an SE attention mechanism 3. Includes an occlusion sensing module	OTB-100	67.7% AUC 87.9% Precision	This method is good for occluded environments.
[82]		GOT-10k	73.1% AUC 59.9% Precision	
[82]		Drones-vs-bird + LaSOT	50.5% AUC 79% Precision	
[70]	Uses ByteTrack	UAVSwarm	UAVSwarm-06 88.3% MOTA UAVSwarm-28 32.5% MOTA UAVSwarm-30 87.2% MOTA UAVSwarm-46 −12% MOTA	The performance of ByteTrack is inconsistent while GNMOT seems to give promising results
[70]	Uses GNMOT	UAVSwarm	UAVSwarm-06 100% MOTA UAVSwarm-28 98.4% MOTA UAVSwarm-30 100% MOTA UAVSwarm-46 99.75 MOTA	
[45]	1. Uses SiamFC with Cascade-RCNN as the built-in detector 2. Cascade-RCNN is used with a ResNet50 backbone	DUT Anti-UAV	61.7% AUC 93.3% Precision	LTMU+Faster-RCNN +VGG16 tracking by detection model gives the best results for DUT Anti-UAV
[45]	1. Uses SiamFC with Faster-RCNN as the built-in detector 2. Cascade-RCNN is used with a VGG16 backbone	DUT Anti-UAV	61.5% AUC 94.3% Precision	
[45]	1. Uses ECO with Faster-RCNN as the built-in detector 2. Cascade-RCNN is used with a VGG16 backbone	DUT Anti-UAV	62% AUC 95.4% Precision	
[45]	1. Uses SPLT with Faster-RCNN as the built-in detector 2. Cascade-RCNN is used with a VGG16 backbone	DUT Anti-UAV	55.3% AUC 87.5% Precision	

Table 2. *Cont.*

Ref.	Techniques Used	Dataset	Best Accuracy Metrics Reported	Comments
[45]	1. Uses ATOM with Faster-RCNN as the built-in detector 2. Cascade-RCNN is used with a ResNet18 backbone	DUT Anti-UAV	63.5% AUC 93.6% Precision	LTMU+Faster-RCNN +VGG16 tracking by detection model gives the best results for DUT Anti-UAV
[45]	1. Uses SiamRPN++ with Faster-RCNN as the built-in detector 2. Cascade-RCNN is used with a VGG16 backbone	DUT Anti-UAV	61.2% AUC 88.1% Precision	
[45]	1. Uses DiMP with Faster-RCNN as the built-in detector 2. Cascade-RCNN is used with a ResNet50 backbone	DUT Anti-UAV	65.7% AUC 94.9% Precision	
[45]	1. Uses TransT with Cascade-RCNN as the built-in detector 2. Cascade-RCNN is used with a ResNet50 backbone	DUT Anti-UAV	62.4% AUC 88.8% Precision	
[45]	1. Uses LTMU with Faster-RCNN as the built-in detector 2. Fascade-RCNN is used with a VGG16 backbone	DUT Anti-UAV	66.4% AUC 96.1% Precision	
[90]	Uses SiamRCNN with DSFC training strategy that CSM and ISM modules	Anti-UAV RGB	67.04% AUC 90.71% Precision	SiamRCNN and GlobalTrack are the best methods.
[90]	Uses GlobalTrack with DSFC training strategy that CSM and ISM modules	Anti-UAV RGB	62.36% AUC 87.65% Precision	
[90]	Uses LTDSE with DSFC training strategy that CSM and ISM modules	Anti-UAV RGB	58.58% AUC 82.56% Precision	
[90]	Uses SiamRPN++LT with DSFC training strategy that CSM and ISM modules	Anti-UAV RGB	57.28% AUC 77.93% Precision	
[90]	Uses Super-DiMP with DSFC training strategy that CSM and ISM modules	Anti-UAV RGB	53.77% AUC 75.29% Precision	

Table 2. *Cont.*

Ref.	Techniques Used	Dataset	Best Accuracy Metrics Reported	Comments
[13]	Uses Tracktor with Faster R-CNN detection module	MAV-VID	95.5% MOTA	Tracktor performs the best consistently but depending on the difficulty of the dataset, different detection modules give different results.
[13]	Uses Tracktor with SSD512 detection module	Drone-vs.-Bird	52.5% MOTA	
[13]	Uses Tracktor with DETR detection module	Anti-UAV RGB	94.0% MOTA	

4. Datasets

The availability of high-quality datasets is critical for advancing research in UAV detection and tracking in the Anti-UAV scenario. For vision-based Anti-UAV systems, datasets must reflect the diversity of environments, UAV types, lighting conditions, and operational contexts that these systems encounter. This section reviews several publicly available datasets used for UAV detection and tracking, offering a detailed analysis of their characteristics, strengths, and limitations. These datasets are also categorized in Table 3 based on their primary focus—detection, tracking, or both. Table 3 also contains the links to where they can be accessed. Table 4 gives a summarized view of the dataset characteristics.

UAV detection and tracking datasets can be broadly classified into visual and non-visual categories. Non-visual systems utilize radar, infrared, laser, or acoustic data, while visual methods exploit image and video data. There have also been multi-modal systems that use multi-sensor data fusion approaches. Non-visual systems are useful when visual surveillance systems are not feasible. Multi-modal systems yield great accuracy but require expensive and extensive hardware capabilities. Non-visual and multi-modal systems usually suffer from limitations such as high prices, susceptibility to noise and frequency interference, poor flexibility, poor concealment, high computational overhead, and inability to detect small UAVs, which are not equipped with any signal transmitters [16,45,49,70]. In contrast, vision-based systems can address such limitations while achieving competitive accuracy, latency speed, and robustness with less expensive equipment.

Table 3. Public datasets available for Anti-UAV objectives.

Dataset	Objective	Link
TIB-Net [40]	Detection	https://github.com/kyn0v/TIB-Net/tree/master (accessed on 4 December 2024)
MAV-VID [93]	Detection, MOT	https://bitbucket.org/alejodosr/mav-vid-dataset/src/master/ (accessed on 4 December 2024)
Anti-UAV [90]	Detection, MOT	https://github.com/ucas-vg/Anti-UAV (accessed on 4 December 2024)
Drone-vs.-Bird [94]	Detection	https://github.com/wosdetc/challenge (accessed on 4 December 2024)
UAVSwarm [70]	Detection, MOT	https://github.com/UAVSwarm/UAVSwarm-dataset (accessed on 4 December 2024)
DUT Anti-UAV [45]	Detection, SOT	https://github.com/wangdongdut/DUT-Anti-UAV (accessed on 4 December 2024)

Table 4. Summarized characteristics of popular datasets.

Dataset	Size	UAV Type	Resolution	Environment	Light Conditions
TIB-Net [40]	2860 images, 694 MB	multi-rotor, fixed-wing	1920 × 1080	Homogeneous, simple backgrounds	Day, nightfall and night
MAV-VID [93]	40,232 images, 64 videos, 11.3 GB	multi-rotor	1920 × 1080	Heterogeneous, Varying background conditions	Day
Anti-UAV(RGB) [90]	93,247 images, 100 videos, 5.25 GB	multi-rotor	1920 × 1080	Heterogeneous, Varying weather conditions	Day, nightfall and night
Drone-vs.-Bird [94]	104,760 images, 77 videos, 7.1 GB	multi-rotor, fixed-wing	varies from 720 × 576 to 3840 × 2160	Heterogeneous, Varying weather and background conditions	Day
UAVSwarm [70]	12,598 images, 72 videos, 1.87 GB	multi-rotor, fixed-wing	varies from 446 × 270 to 1919 × 1079	Heterogeneous, Varying light and background conditions	Day, nightfall and night
DUT Anti-UAV [45]	10,000 images, 20 videos, 8.76 GB	multi-rotor	varies extremely	Heterogeneous, Varying light and background conditions	Day, nightfall and night

4.1. Detailed Analysis of Popular Datasets

4.1.1. TIB-Net

TIB-Net [40] is a dataset primarily designed only for UAV detection. It consists of 2860 high-resolution images (1920 × 1080 pixels) captured from a ground-based camera. This dataset covers a variety of UAV types, including multi-rotor and fixed-wing UAVs, and features day and night conditions, which are crucial for developing robust models that can generalize across different lighting scenarios. One of the strengths of TIB-Net is its focus on the small size of UAVs in distant images (approximately 500 m away), a common challenge in real-world detection systems. Most of the UAVs in the dataset occupy less than 0.1% of the total image area. However, the dataset's relatively small size limits its utility for training deep learning models, especially in comparison to larger, more diverse datasets.

TIB-Net, while useful for small UAV detection, lacks the diversity of environments and UAV types seen in larger datasets. Its primary limitation is the absence of dynamic scenes, such as moving backgrounds or other aerial entities, which would more accurately reflect real-world conditions where UAVs must be detected amidst environmental clutter.

4.1.2. MAV-VID

The Multirotor Aerial Vehicle VID (MAV-VID) dataset [93] stands out for its dynamic content, offering 53 training videos and 11 validation videos captured from a variety of perspectives, including UAV-mounted, ground-based, and handheld devices. The dataset's emphasis on multi-UAV detection and tracking makes it ideal for evaluating algorithms that must generalize across different viewpoints. MAV-VID also includes UAVs of varying sizes, making it useful for developing models that can detect and track drones in diverse settings.

The dataset's heterogeneous nature—comprising both ground-to-air and air-to-air data—introduces challenges for model consistency. Models trained on MAV-VID may struggle to generalize to situations with more consistent, structured camera angles. Moreover, the dataset lacks comprehensive annotation for UAV occlusion scenarios, which are critical for evaluating the robustness of tracking systems in dense environments.

4.1.3. Anti-UAV

The Anti-UAV dataset [90] is a large, multi-modal dataset featuring 100 fully annotated video sequences in both RGB and infrared (IR) spectra, with 186,494 images in total. This dataset focuses on multi-UAV tracking and detection across six different UAV models captured under a range of lighting and background conditions. The inclusion of infrared data makes Anti-UAV highly valuable for developing robust models that can operate in low-visibility conditions, such as nighttime or foggy environments. This makes the dataset particularly valuable for security-focused applications requiring continuous surveillance in various lighting conditions.

4.1.4. Drone-vs.-Bird

The Drone-vs.-Bird dataset [94] is one of the most challenging datasets available due to its focus on the long-range detection of UAVs amidst confounding objects like birds. With 77 videos comprising 104,760 images, this dataset forces models to distinguish between small UAVs and birds, a critical task in scenarios where false positives are a major concern. The small size of UAVs in the dataset (34×23 pixels on average) and the minimal percentage of the image they occupy make it particularly challenging for models, especially those that rely on large bounding boxes or clear object features. The extreme difficulty of detecting UAVs in the Drone-vs.-Bird dataset highlights the limitations of current detection algorithms, particularly in terms of scale variance.

4.1.5. UAVSwarm

The UAVSwarm dataset [70] is specifically designed for multi-UAV tracking and swarm behavior analysis, making it unique in its focus on collective UAV interactions. It contains 12,598 images, each featuring between 3 and 23 UAVs, allowing researchers to study both the detection and tracking of multiple UAVs. This dataset is especially useful for evaluating algorithms in swarm surveillance, where detecting and tracking many objects simultaneously is required.

The inclusion of swarm scenarios makes UAVSwarm an excellent choice for developing MOT algorithms. The diversity of UAV types and the complexity of interactions make it highly suitable for testing the robustness of tracking models in real-world airspace management.

While the dataset offers diverse UAV scenarios, it lacks the variety of environmental conditions (e.g., weather and time of day) seen in other datasets like TIB-Net and Anti-UAV. This may limit the generalizability of models trained on UAVSwarm when applied to more dynamic environments.

4.1.6. DUT Anti-UAV

The DUT Anti-UAV dataset [45] is designed to support both detection and tracking tasks, offering a balanced mix of static images and tracking sequences. The dataset includes 10,000 images for detection and 20 tracking sequences for evaluating single and multi-UAV tracking in complex environments. The dataset consists of mainly small UAVs against cluttered backgrounds, which makes it a particularly challenging evaluation benchmark for detection and tracking tasks. The primary limitation of DUT Anti-UAV is its limited set of tracking scenarios. It lacks diversity in terms of weather conditions, UAV behaviors, and long-term tracking.

The reviewed datasets highlight the diversity of scenarios that ground-to-air vision-based Anti-UAV systems must handle, from small UAV detection to multi-UAV tracking in dynamic environments. However, most existing datasets still face limitations, particularly in terms of environmental diversity and real-world complexity. Many datasets lack comprehensive coverage of adverse conditions like rain, fog, or low-light environments, and few

offer scenarios with high UAV density or diverse UAV behaviors. These gaps highlight the need for future datasets to better reflect real-world challenges, particularly for large-scale UAV detection and long-term Anti-UAV tracking in diverse and complex environments.

5. Challenges and Future Works

Despite significant advancements in ground-to-air vision-based Anti-UAV detection and tracking systems, several challenges remain unresolved. Recent research has made strides in detecting and tracking UAVs under controlled conditions, but real-world applications often present far more complex scenarios. This section highlights the key limitations in existing works and outlines potential directions for future research to overcome these challenges.

The challenges that remain in current Anti-UAV Systems are as follows:

- Detection and Tracking of Small and Fast-moving UAVs: Detecting and tracking small, fast-moving UAVs against cluttered and dynamic backgrounds is a major challenge for vision-based systems. These UAVs appear as only a few pixels in high-resolution images, making it difficult for existing object detection models to identify them accurately. As a result, UAVs are frequently mistaken for birds, clouds, or background noise, leading to false positives and missed detections.
- Real-time Processing Constraints: Real-time detection and tracking are crucial for Anti-UAV systems in scenarios where UAVs pose immediate threats, such as near airports or critical infrastructure. However, achieving high accuracy in real-time conditions is challenging due to the computational complexity of current detection algorithms. Models have made progress in optimizing inference speed but often sacrifice accuracy, especially when detecting small, fast-moving UAVs. Additionally, tracking methods prioritizing high precision may introduce unacceptable latency in time-sensitive applications.
- Multi-UAV Detection and Tracking: The current research primarily focuses on single UAV detection and tracking, which simplifies the problem in controlled or isolated environments. However, real-world commercial scenarios would involve multiple UAVs operating simultaneously in a 3D air volume, introducing complexity due to occlusion, interactions, and merging with background elements. The limitations of existing methods become apparent when faced with the dynamics of UAV swarms or the congested airspace characteristic of urban areas. Existing methods may not be suitable to handle overlapping trajectories and the rapid movements of UAVs of varying sizes and also intent. Little to no work has been performed to identify and track multiple rogues from normal UAV traffic. Consequently, Multi-UAV tracking remains under-explored, offering opportunities for future research. Addressing these challenges will be essential for creating robust Anti-UAV systems with the capability of detecting and tracking multiple rogues or swarms of rogues concurrently.
- Environmental Variability and Robustness: Vision-based systems are impacted by environmental factors such as lighting, weather, and background complexity. Existing models often struggle in low-light conditions, fog, or rain. While efforts have been made to address some limitations, ensuring reliability in diverse conditions remains a challenge for Anti-UAV systems. There is a lack of representative benchmark datasets that truly reflect a commercialized UAV traffic scenario in variable environmental conditions. A potential research direction could be to either collect such a dataset or to artificially generate a dataset using data augmentation techniques, generative AI, etc.
- Integration into the UTM: While Anti-UAV systems are being developed to detect rogue or unauthorized UAVs, there is a significant gap in integrating these systems into broader UTM frameworks. UTM systems are designed to manage authorized UAV traffic in uncontrolled airspace; Anti-UAV systems must operate alongside

these frameworks to ensure that only non-compliant UAVs are targeted. The Federal Aviation Administration's (FAA) proposed UTM system and the European Union's U-space system are both designed with a service-oriented architecture in mind [95,96]. This approach allows for flexibility and innovation, as various components of these frameworks can be developed, managed, and maintained by independent service providers. This service-oriented model not only promotes specialization but also ensures that advancements can be integrated seamlessly, enhancing the overall efficiency and safety of airspace management for Unmanned Aerial Vehicles across both regions. The current literature lacks a holistic approach to seamlessly integrating detection and tracking methodologies within the UTM framework, including real-time coordination between ground stations and flight management systems. This requires more comprehensive solutions as well as better datasets to distinguish between regular and non-conforming UAVs in complex airspace environments.

The advancement of ground-to-air vision-based Anti-UAV systems hinges on several key research priorities. At the forefront is the need to develop more accurate detection models capable of identifying UAVs across diverse sizes and scales. This could be achieved through refined multi-scale detection architectures, coupled with the integration of higher-resolution cameras and super-resolution techniques to enhance image quality and improve detection precision.

Real-time processing remains a critical challenge that demands innovative solutions. To address this, researchers should explore optimization strategies that reduce computational demands while preserving detection accuracy. Promising approaches include model pruning [44] and quantization [97], which have shown success in improving model efficiency. The integration of specialized hardware, particularly edge computing platforms, could further enable real-time processing in resource-constrained environments.

As the field evolves, the focus is shifting toward robust multi-UAV detection and tracking systems. This advancement requires sophisticated data fusion techniques that combine visual and non-visual information. Environmental adaptability is equally crucial—future systems must perform reliably across various conditions such as rain, fog, and low light. This can be achieved through advanced data augmentation techniques that simulate diverse environmental scenarios during model training.

The successful implementation of Anti-UAV systems requires seamless integration with the broader UTM framework. This integration demands the careful consideration of protocols for distinguishing between authorized and unauthorized UAVs while ensuring that countermeasures do not disrupt legitimate UAV operations. Such coordination is essential for maintaining the delicate balance between security and operational efficiency.

As UAV traffic continues to expand, these research directions become increasingly critical. The development of robust, scalable, and real-time solutions will be fundamental in securing both manned and unmanned airspaces, making these priorities not just technical challenges but essential components of future aviation safety.

6. Conclusions

The rapid expansion of UAV technology has created both opportunities and challenges in managing low-altitude airspace. Vision-based Anti-UAV systems provide a promising solution for detecting and tracking rogue UAVs, ensuring the safe integration of UAVs into the airspace. However, significant challenges persist, such as reliably detecting small, fast-moving UAVs, achieving real-time performance, managing multi-UAV scenarios, and ensuring robustness in diverse environmental conditions. The integration of Anti-UAV technologies into UTM frameworks is still in its infancy, limiting their broader applicability. Addressing these challenges is crucial for developing comprehensive, real-time Anti-UAV

systems essential for airspace security. Continued innovation is necessary to meet the increasing demand for UAV detection and tracking in complex real-world environments. The intersection of vision-based technologies with UAV management promises to be a rich area for future exploration, with significant implications for both civil and defense sectors.

Author Contributions: Conceptualization, A.Y; formal analysis, A.Y.; investigation, A.Y.; writing—original draft preparation, A.Y.; writing—review and editing, A.Y. and O.D.; visualization, A.Y.; supervision, O.D.; project administration, O.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

DURC Statement: Current research is limited to the academic field of Computer Vision from an Anti-UAV perspective, which is beneficial in developing Unmanned Aerial Vehicle Traffic Management systems and does not pose a threat to public health or national security. Authors acknowledge the dual-use potential of the research involving UAVs and confirm that all necessary precautions have been taken to prevent potential misuse. As an ethical responsibility, authors strictly adhere to relevant national and international laws about DURC. Authors advocate for responsible deployment, ethical considerations, regulatory compliance, and transparent reporting to mitigate misuse risks and foster beneficial outcomes.

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

Abbreviations

The following abbreviations are used in this manuscript:

AGL	Above Ground Level
ASFF	Adaptive Spatial Feature Fusion
ATC	Air Traffic Control
ATOM	Accurate Tracking by Overlap Maximization
ATM	Air Traffic Management
ATSS	Adaptive Training Sample Selection
AUC	Area Under the Curve
C2f	Convolution to Feature
CA	Coordinate Attention
CBAM	Convolutional Block Attention Module
CSM	Class-Level Semantic Modulation
CSPDarkNet	Cross-Stage-Partial-connections DarkNet
DeepSORT	Simple Online and Real-time Tracking with a Deep Association Metric
DETR	Detection Transformer
DiMP	Discriminative Model Prediction
DSFC	Dual-Flow Semantic Consistency
ECO	Efficient Convolution Operators
EMA	Efficient Multi-scale Attention
EXTD	Extremely Tiny Face Detector
FAA	Federal Aviation Administration
FCOS	Fully Convolutional One-Stage Object Detector
FPN	Feature Pyramid Network
FPS	Frames Per Second
GNMOT	Graph Networks for Multi-Object Tracking
ISM	Instance-Level Semantic Modulation
JDE	Joint Detection and Embedding

LSTM	Long Short-Term Memory
LTMU	Long-term Tracking with Meta-Updater
mAP	Mean Average Precision
MEGA	Memory Enhanced Global-Local Aggregation
MobileViT	Mobile Vision Transformer
MOT	Multi-Object Tracking
MOTA	Multiple Object Tracking Accuracy
PANet	Path Aggregation Network
QDTrack	Quasi-dense Tracking
R-CNN	Region-based Convolutional Neural Network
RF	Radio Frequency
SAM	Spatial Attention Module
SE	Squeeze and Excitation
SimAM	Simple parameter-free Attention Module
SORT	Simple Online Real-time Tracking
SOT	Single-Object Tracking
SPP	Spatial Pyramid Pooling
SPLT	Skimming-Perusal Tracking
SSD	Single Shot Detector
TransT	Transformer Tracking
UAS	Unmanned Aircraft System
UAV	Unmanned Aerial Vehicle
UTM	Unmanned Aircraft System Traffic Management
VGG	Visual Geometry Group
YOLO	You Only Look Once

Appendix A

Table A1. Summary of UAV detection methodologies used in recent works.

Reference	Year	Techniques Used	Dataset(s) Used	Accuracy Metrics Reported	Lightweight
[39]	2019	1. The last four scales of feature maps are adopted instead of the last three in YOLOv3 to predict bounding boxes of objects, which can obtain more texture and contour information 2. K-means clustering is used on the training set to decide the number of the scales 3. The number and size of the anchor boxes are also adjusted using k-means clustering	Not public. Self-collected.	mAP@0.25 37.41% 56.3 FPS	No
[40]	2020	1. Cyclic pathway is added to EXTD, which enhances the capability to extract the effective features of small objects but does not increase the model size much 2. Spatial Attention Module is added to the network backbone to emphasize information of small objects	TIBNet	mAP 89.25%	Yes
[41]	2020	1. Standard YOLOv4 is implemented 2. Dataset is augmented by rotating and flipping collected images	Not public. Self-collected.	mAP@0.75 89.32% 39.64 FPS	No

Table A1. *Cont.*

Reference	Year	Techniques Used	Dataset(s) Used	Accuracy Metrics Reported	Lightweight
[42]	2021	Standard Faster R-CNN is implemented with multiple backbones (VGG-16, ResNet50, DarkNet-53, DenseNet-201)	UAVData	VGG-16 mAP 90.6%, 11 FPS ResNet50 mAP 90.4%, 10 FPS DarkNet-53 mAP 86.3%, 10 FPS DenseNet-201 mAP Failed	No
[42]	2021	Standard YOLOv3 is implemented with multiple backbones (VGG-16, ResNet50, DarkNet-53, DenseNet-201)	UAVData	VGG-16 mAP 90.8%, 70 FPS ResNet50 mAP 90.6%, 86 FPS DarkNet-53 mAP 90.8%, 72 FPS DenseNet-201 mAP 90.7%, 49 FPS	No
[42]	2021	Standard SSD is implemented with multiple backbones (VGG-16, ResNet50, DarkNet-53, DenseNet-201)	UAVData	VGG-16 mAP 74.2%, 42 FPS ResNet50 mAP 75.3%, 22 FPS DarkNet-53 mAP 74.8%, 24 FPS DenseNet-201 mAP 73.5, 14 FPS	Yes (VGG-16) No for rest
[13]	2021	Standard Faster R-CNN is implemented with ResNet-50 backbone and an FPN at the end of each convolutional block	MAV-VID Drone-vs.-Bird Anti-UAV RGB	MAV-VID mAP 97.8%, 18 FPS Drone-vs.-Bird mAP 63.2%, 18 FPS Anti-UAV RGB mAP 98.2%, 18 FPS	No
[13]	2021	Standard YOLOv3 is implemented with DarkNet-53 backbone	MAV-VID Drone-vs.-Bird Anti-UAV RGB	MAV-VID mAP 96.3%, 36 FPS Drone-vs.-Bird mAP 54.6%, 36 FPS Anti-UAV RGB mAP 98.6%, 36 FPS	No
[13]	2021	Standard SSD512 is implemented	MAV-VID Drone-vs.-Bird Anti-UAV RGB	MAV-VID mAP 96.7%, 32.4 FPS Drone-vs.-Bird mAP 62.9%, 32.4 FPS Anti-UAV RGB mAP 97.9%, 32.4 FPS	Yes
[13]	2021	Standard DETR with ResNet-50 backbone	MAV-VID Drone-vs.-Bird Anti-UAV RGB	MAV-VID mAP 97.1%, 21.4 FPS Drone-vs.-Bird mAP 66.7%, 21.4 FPS Anti-UAV RGB mAP 97.8%, 21.4 FPS	No
[43]	2021	1. Standard YOLOv4 is implemented 2. Mosaic data augmentation is applied	Not public. Self-collected.	mAP 74.36% 19.75 FPS	No

Table A1. *Cont.*

Reference	Year	Techniques Used	Dataset(s) Used	Accuracy Metrics Reported	Lightweight
[44]	2021	1. Convolutional channel and shortcut layer of YOLOv4 are pruned to make the model thinner and shallower 2. Dataset is augmented by copy and pasting small drones	Not public. Self-collected.	mAP 92.7% 69 FPS	Yes
[45]	2022	Standard Faster R-CNN is implemented with different backbones	DUT Anti-UAV	ResNet-50 mAP 65.3%, 12.8 FPS ResNet-18 mAP 60.5%, 19.4 FPS VGG-16 mAP 63.3%, 9.3 FPS	No
[45]	2022	Standard Cascade R-CNN is implemented with different backbones	DUT Anti-UAV	ResNet-50 mAP 68.3%, 10.7 FPS ResNet-18 mAP 65.2%, 14.7 FPS VGG-16 mAP 66.7%, 8 FPS	No
[45]	2022	Standard ATSS method is implemented with different backbones	DUT Anti-UAV	ResNet-50 mAP 64.2%, 13.3 FPS ResNet-18 mAP 61%, 20.5 FPS VGG-16 mAP 64.1%, 9.5 FPS	N/A
[45]	2022	Standard YOLOX is implemented with different backbones	DUT Anti-UAV	ResNet-50 mAP 42.7%, 21.7 FPS ResNet-18 mAP 40%, 53.7 FPS VGG-16 mAP 55.1%, 23 FPS	No
[45]	2022	Standard SSD is implemented with different backbones	DUT Anti-UAV	VGG-16 mAP 63.2%, 33.2 FPS	Yes
[46]	2022	1. Background difference is used to extract potential drone targets in high-resolution images to reduce computational overhead 2. Ghost module and SimAM attention mechanism are introduced to reduce the total number of model parameters and improve feature extraction 3. α-DIoU loss is used instead of DIoU loss to improve the accuracy of bounding box regression	Drone-vs.-Bird	mAP 97.6% 13.2 FPS	Yes
[47]	2022	1. The last four scales of feature maps are adopted instead of the last three in YOLOv3 to predict bounding boxes of objects, which can obtain more texture and contour information 2. Data augmentation is performed through changing brightness, and contrast of the images and rotating and flipping the images	Not public. Self-collected.	mAP 25.12% 21 FPS	No

Table A1. *Cont.*

Reference	Year	Techniques Used	Dataset(s) Used	Accuracy Metrics Reported	Lightweight
[48]	2022	1. YOLOv5 backbone is replaced with Efficientlite, to reduce the number of parameters 2. Adaptive spatial feature fusion is injected into the head to improve the accuracy loss caused by the lightweight of the model backbone 3. A constraint of angle is introduced into the original regression loss function to improve the speed of convergence 4. Data augmentation by adding random noise points and binarization	Kaggle Dataset	mAP 94.82%	Yes
[49]	2023	1. Spatial Attention module added to the backbone 2. SPPS and ResNeck modules added to the neck 3. Data are augmented using Mosaic Augmentation	TIBNet	mAP 89.7%	Yes
[50]	2023	1. MobileViT is used as the backbone to reduce network complexity 2. Added Coordinate Attention to PANet of YOLOv4 to obtain better positional information and improve information fusion of high and low-dimensional features 3. K-means++ is used to adjust anchor boxes 4. Data are augmented using Mosaic Augmentation	Not public. Self-collected.	mAP 92.8% 40 FPS	Yes
[51]	2023	1. Depthwise separable convolution is used to simplify and optimize the network 2. Squeeze-and-Excitation (SE) module is introduced into the backbone to improve the model's ability to extract features 3. Convolutional Block Attention Module (CBAM) is added in the feature fusion network to make the network pay more attention to important features and suppress unnecessary features 4. Distance-IoU (DIoU) is used to replace Intersection over Union (IoU) to calculate the regression loss for model optimization 5. Data are augmented using MixUp and Mosaic Augmentation	UAVSwarm	mAP 82.32% 14 FPS	Yes
[52]	2024	1. Ghost convolution is included in the neck to reduce model size 2. Efficient multi-scale attention (EMA) is added to preserve pixel-level attributes and spatial information on the feature map 3. Deformable Convolutional Net v2 (DCNv2) is used in the detection head to improve model robustness	DUT Anti-UAV	mAP 97.1 56.2 FPS	Yes

Table A1. *Cont.*

Reference	Year	Techniques Used	Dataset(s) Used	Accuracy Metrics Reported	Lightweight
[53]	2024	1. Anchor boxes are adjusted using k-means clustering 2. InceptionNeXT module is added to neck to capture more global semantic information 3. SPPFCSPC-SR module is added to the backbone to reduce feature loss, suppress confusion, and make the model pay more attention to small target areas 4. FPN is replaced with Get-and-Send module to improve model's capability to fuse information across different levels	DUT Anti-UAV and Amateur UAV combined	mAP 93.2% 104 FPS	Yes

References

1. Houthi Drone Attacks on 2 Saudi Aramco Oil Facilities Spark Fires. Available online: https://tinyurl.com/UAVattack (accessed on 4 December 2024).
2. Drones Market Size & Share Analysis—Growth Trends & Forecasts (2024–2029). Available online: https://tinyurl.com/dronesfore (accessed on 4 December 2024).
3. Forecast Highlights (2023–2043). Available online: https://tinyurl.com/FAAForecasts (accessed on 4 December 2024).
4. Davies, L.; Vagapov, Y.; Grout, V.; Cunningham, S.; Anuchin, A. Review of air traffic management systems for UAV integration into urban airspace. In Proceedings of the 2021 28th International Workshop on Electric Drives: Improving Reliability of Electric Drives (IWED), Moscow, Russia, 27–29 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
5. Chamola, V.; Kotesh, P.; Agarwal, A.; Gupta, N.; Guizani, M. A comprehensive review of unmanned aerial vehicle attacks and neutralization techniques. *Ad Hoc Netw.* **2021**, *111*, 102324. [CrossRef] [PubMed]
6. Klare, J.; Biallawons, O.; Cerutti-Maori, D. UAV detection with MIMO radar. In Proceedings of the 2017 18th International Radar Symposium (IRS), Prague, Czech Republic, 28–30 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8.
7. Mohajerin, N.; Histon, J.; Dizaji, R.; Waslander, S.L. Feature extraction and radar track classification for detecting UAVs in civillian airspace. In Proceedings of the 2014 IEEE Radar Conference, Cincinnati, OH, USA, 19–23 May 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 674–679.
8. Xiao, Y.; Zhang, X. Micro-UAV detection and identification based on radio frequency signature. In Proceedings of the 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2–4 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1056–1062.
9. Al-Emadi, S.; Al-Senaid, F. Drone detection approach based on radio-frequency using convolutional neural network. In Proceedings of the 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, 2–5 February 2020 IEEE: Piscataway, NJ, USA, 2020; pp. 29–34.
10. Yang, B.; Matson, E.T.; Smith, A.H.; Dietz, J.E.; Gallagher, J.C. UAV detection system with multiple acoustic nodes using machine learning models. In Proceedings of the 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy, 25–27 February 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 493–498.
11. Hauzenberger, L.; Holmberg Ohlsson, E. Drone Detection Using Audio Analysis. Master's Thesis, Lund University, Lund, Sweden, 2015.
12. Behera, D.K.; Raj, A.B. Drone detection and classification using deep learning. In Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 13–15 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1012–1016.
13. Isaac-Medina, B.K.; Poyser, M.; Organisciak, D.; Willcocks, C.G.; Breckon, T.P.; Shum, H.P. Unmanned aerial vehicle visual detection and tracking using deep neural networks: A performance benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, Montreal, QC, Canada, 10–17 October 2021; pp. 1223–1232.
14. Mendis, G.J.; Randeny, T.; Wei, J.; Madanayake, A. Deep learning based doppler radar for micro UAS detection and classification. In Proceedings of the MILCOM 2016—2016 IEEE Military Communications Conference, Baltimore, MD, USA, 1–3 November 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 924–929.
15. Bisio, I.; Garibotto, C.; Lavagetto, F.; Sciarrone, A.; Zappatore, S. Unauthorized amateur UAV detection based on WiFi statistical fingerprint analysis. *IEEE Commun. Mag.* **2018**, *56*, 106–111. [CrossRef]

16. Chu, Z.; Song, T.; Jin, R.; Jiang, T. An Experimental Evaluation Based on New Air-to-Air Multi-UAV Tracking Dataset.In Proceedings of the 2023 IEEE International Conference on Unmanned Systems (ICUS), Hefei, China, 13–15 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 671–676.
17. Muna, S.I.; Mukherjee, S.; Namuduri, K.; Compere, M.; Akbas, M.I.; Molnár, P.; Subramanian, R. Air Corridors: Concept, Design, Simulation, and Rules of Engagement. *Sensors* **2021**, *21*, 7536. [CrossRef] [PubMed]
18. Liu, Z.; An, P.; Yang, Y.; Qiu, S.; Liu, Q.; Xu, X. Vision-Based Drone Detection in Complex Environments: A Survey. *Drones* **2024**, *8*, 643. [CrossRef]
19. 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* **2024**, *24*, 125. [CrossRef]
20. Wang, B.; Li, Q.; Mao, Q.; Wang, J.; Chen, C.L.P.; Shangguan, A.; Zhang, H. A Survey on Vision-Based Anti Unmanned Aerial Vehicles Methods. *Drones* **2024**, *8*, 518. [CrossRef]
21. 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]
22. Coluccia, A.; Fascista, A.; Schumann, A.; Sommer, L.; Dimou, A.; Zarpalas, D.; Akyon, F.C.; Eryuksel, O.; Ozfuttu, K.A.; Altinuc, S.O.; et al. Drone-vs.-bird detection challenge at IEEE AVSS2021. In Proceedings of the 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Washington, DC, USA, 16–19 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
23. Abdullah, Q.A. Classification of the Unmanned Aerial Systems. Pennsylvania State. 2014. Available online: https://www.e-education.psu.edu/geog892/node/5 (accessed on 13 January 2025).
24. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
25. Kaiming, H.; Georgia, G.; Piotr, D.; Ross, G.S. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; Volume 2017, pp. 2961–2969.
26. Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. Grid r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7363–7372.
27. Farhadi, A.; Redmon, J. Yolov3: An incremental improvement. In *Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 1804, pp. 1–6
28. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
29. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
30. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
31. Ross, T.Y.; Dollár, G.K.H.P. Focal loss for dense object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
32. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.
33. Li, S.; Sultonov, F.; Tursunboev, J.; Park, J.H.; Yun, S.; Kang, J.M. Ghostformer: A GhostNet-based two-stage transformer for small object detection. *Sensors* **2022**, *22*, 6939. [CrossRef]
34. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
35. Liu, S.; Zhou, H.; Li, C.; Wang, S. Analysis of anchor-based and anchor-free object detection methods based on deep learning. In Proceedings of the 2020 IEEE International Conference on Mechatronics and Automation (ICMA), Beijing, China, 2–5 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1058–1065.
36. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
37. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
38. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
39. Hu, Y.; Wu, X.; Zheng, G.; Liu, X. Object detection of UAV for anti-UAV based on improved YOLO v3. In Proceedings of the 2019 Chinese Control Conference (CCC), Guangzhou, China, 27–30 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8386–8390.
40. Sun, H.; Yang, J.; Shen, J.; Liang, D.; Ning-Zhong, L.; Zhou, H. TIB-Net: Drone detection network with tiny iterative backbone. *IEEE Access* **2020**, *8*, 130697–130707. [CrossRef]

41. Shi, Q.; Li, J. Objects detection of UAV for anti-UAV based on YOLOv4. In Proceedings of the 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Weihai, China, 14–16 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1048–1052.
42. Zeng, Y.; Duan, Q.; Chen, X.; Peng, D.; Mao, Y.; Yang, K. UAVData: A dataset for unmanned aerial vehicle detection. *Soft Comput.* **2021**, *25*, 5385–5393. [CrossRef]
43. Singha, S.; Aydin, B. Automated drone detection using YOLOv4. *Drones* **2021**, *5*, 95. [CrossRef]
44. Liu, H.; Fan, K.; Ouyang, Q.; Li, N. Real-time small drones detection based on pruned yolov4. *Sensors* **2021**, *21*, 3374. [CrossRef]
45. Zhao, J.; Zhang, J.; Li, D.; Wang, D. Vision-based anti-uav detection and tracking. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25323–25334. [CrossRef]
46. Lv, Y.; Ai, Z.; Chen, M.; Gong, X.; Wang, Y.; Lu, Z. High-resolution drone detection based on background difference and SAG-Yolov5s. *Sensors* **2022**, *22*, 5825. [CrossRef]
47. Zhai, H.; Zhang, Y. Target Detection of Low-Altitude UAV Based on Improved YOLOv3 Network. *J. Robot.* **2022**, *2022*, 4065734. [CrossRef]
48. Liu, B.; Luo, H. An improved Yolov5 for multi-rotor UAV detection. *Electronics* **2022**, *11*, 2330. [CrossRef]
49. Zhou, X.; Yang, G.; Chen, Y.; Gao, C.; Zhao, B.; Li, L.; Chen, B.M. ADMNet: Anti-Drone Real-Time Detection and Monitoring. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 3009–3016.
50. Cheng, Q.; Li, X.; Zhu, B.; Shi, Y.; Xie, B. Drone detection method based on MobileViT and CA-PANet. *Electronics* **2023**, *12*, 223. [CrossRef]
51. Wang, C.; Meng, L.; Gao, Q.; Wang, J.; Wang, T.; Liu, X.; Du, F.; Wang, L.; Wang, E. A lightweight UAV swarm detection method integrated attention mechanism. *Drones* **2022**, *7*, 13. [CrossRef]
52. Huang, M.; Mi, W.; Wang, Y. EDGS-YOLOv8: An Improved YOLOv8 Lightweight UAV Detection Model. *Drones* **2024**, *8*, 337. [CrossRef]
53. Bo, C.; Wei, Y.; Wang, X.; Shi, Z.; Xiao, Y. Vision-based anti-UAV detection based on YOLOv7-GS in complex backgrounds. *Drones* **2024**, *8*, 331. [CrossRef]
54. Terven, J.; Córdova-Esparza, D.M.; Romero-González, J.A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
55. Redmon, J. Darknet: Open Source Neural Networks in C. 2018. Available online: http://pjreddie.com/darknet/ (accessed on 4 December 2024).
56. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
57. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
58. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
59. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
60. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
61. Yoo, Y.; Han, D.; Yun, S. Extd: Extremely tiny face detector via iterative filter reuse. *arXiv* **2019** arXiv:1906.06579.
62. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
63. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
64. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
65. Zhang, H. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
66. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October 2019–2 November 2019; pp. 6023–6032.
67. Christiansen, A. *Anchor Boxes—The Key to Quality Object Detection*; Towards Data Science: San Francisco, CA, USA, 2018.
68. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
69. Guo, J.M.; Yang, J.S.; Seshathiri, S.; Wu, H.W. A light-weight CNN for object detection with sparse model and knowledge distillation. *Electronics* **2022**, *11*, 575. [CrossRef]
70. Wang, C.; Su, Y.; Wang, J.; Wang, T.; Gao, Q. UAVSwarm dataset: An unmanned aerial vehicle swarm dataset for multiple object tracking. *Remote Sens.* **2022**, *14*, 2601. [CrossRef]

71. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
72. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
73. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
74. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
75. Yu, W.; Zhou, P.; Yan, S.; Wang, X. Inceptionnext: When inception meets convnext. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5672–5683.
76. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019** arXiv:1911.09516.
77. Wang, R.; Shivanna, R.; Cheng, D.; Jain, S.; Lin, D.; Hong, L.; Chi, E. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In Proceedings of the Web Conference, Ljubljana, Slovenia, 19–23 April 2021; pp. 1785–1797.
78. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
79. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef]
80. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10, 15–16 October 2016; Proceedings, Part II 14; Springer International Publishing: Cham, Switzerland, 2016; pp. 850–865.
81. Cheng, F.; Liang, Z.; Peng, G.; Liu, S.; Li, S.; Ji, M. An anti-UAV long-term tracking method with hybrid attention mechanism and hierarchical discriminator. *Sensors* **2022**, *22*, 3701. [CrossRef]
82. Wang, C.; Shi, Z.; Meng, L.; Wang, J.; Wang, T.; Gao, Q.; Wang, E. Anti-occlusion UAV tracking algorithm with a low-altitude complex background by integrating attention mechanism. *Drones* **2022**, *6*, 149. [CrossRef]
83. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4282–4291.
84. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. Atom: Accurate tracking by overlap maximization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4660–4669.
85. Yan, B.; Zhao, H.; Wang, D.; Lu, H.; Yang, X. 'skimming-perusal' tracking: A framework for real-time and robust long-term tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2385–2393.
86. Dai, K.; Zhang, Y.; Wang, D.; Li, J.; Lu, H.; Yang, X. High-performance long-term tracking with meta-updater. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6298–6307.
87. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning discriminative model prediction for tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6182–6191.
88. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8126–8135.
89. Huang, L.; Zhao, X.; Huang, K. Globaltrack: A simple and strong baseline for long-term tracking. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 11037–11044. [CrossRef]
90. Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Zhao, J.; Guo, G.; Han, Z. Anti-UAV: A large multi-modal benchmark for UAV tracking. *arXiv* **2021**, arXiv:2101.08466.
91. Wu, Y.; Lim, J.; Yang, M.H. Online object tracking: A benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2411–2418.
92. Bernardin, K.; Elbs, A.; Stiefelhagen, R. Multiple object tracking performance metrics and evaluation in a smart room environment. In Proceedings of the Sixth IEEE International Workshop on Visual Surveillance, Graz, Austria, 13 May 2006; Conjunction with ECCV; Citeseer: University Park, PA, USA, 2006; Volume 90.
93. Rodriguez-Ramos, A.; Rodriguez-Vazquez, J.; Sampedro, C.; Campoy, P. Adaptive inattentional framework for video object detection with reward-conditional training. *IEEE Access* **2020**, *8*, 124451–124466. [CrossRef]
94. Coluccia, A.; Fascista, A.; Schumann, A.; Sommer, L.; Dimou, A.; Zarpalas, D.; Méndez, M.; De la Iglesia, D.; González, I.; Mercier, J.P.; et al. Drone vs. bird detection: Deep learning algorithms and results from a grand challenge. *Sensors* **2021**, *21*, 2824. [CrossRef]

95. Faa.gov. UTM Concept of Operations Version 2.0 (UTM ConOps v2.0) | Federal Aviation Administration. 2022. Available online: https://www.faa.gov/researchdevelopment/trafficmanagement/utm-concept-operations-version-20-utm-conops-v20 (accessed on 13 January 2025).
96. Barrado, C.; Boyero, M.; Brucculeri, L.; Ferrara, G.; Hately, A.; Hullah, P.; Martin-Marrero, D.; Pastor, E.; Rushton, A.P.; Volkert, A. U-space concept of operations: A key enabler for opening airspace to emerging low-altitude operations. *Aerospace* **2020**, *7*, 24. [CrossRef]
97. Wei, L.; Ma, Z.; Yang, C.; Yao, Q. Advances in the Neural Network Quantization: A Comprehensive Review. *Appl. Sci.* **2024**, *14*, 7445. [CrossRef]