electronics

MDPI

Article

Multi-Scale Fusion Uncrewed Aerial Vehicle Detection Based on RT-DETR

Minling Zhu * and En Kong

Computer School, Beijing Information Science and Technology University, Beijing 100101, China; kongen@bistu.edu.cn
* Correspondence: zhuminling@bistu.edu.cn

Abstract: With the rapid development of science and technology, uncrewed aerial vehicle (UAV) technology has shown a wide range of application prospects in various fields. The accuracy and real-time performance of UAV target detection play a vital role in ensuring safety and improving the work efficiency of UAVs. Aimed at the challenges faced by the current UAV detection field, this paper proposes the Gathering Cascaded Dilated DETR (GCD-DETR) model, which aims to improve the accuracy and efficiency of UAV target detection. The main innovations of this paper are as follows: (1) The Dilated Re-param Block is creatively applied to the dilatation-wise Residual module, which uses the large kernel convolution and the parallel small kernel convolution together and fuses the feature maps generated by multi-scale perception, greatly improving the feature extraction ability, thereby improving the accuracy of UAV detection. (2) The Gather-and-Distribute mechanism is introduced to effectively enhance the ability of multi-scale feature fusion so that the model can make full use of the feature information extracted from the backbone network and further improve the detection performance. (3) The Cascaded Group Attention mechanism is innovatively introduced, which not only saves the computational cost but also improves the diversity of attention by dividing the attention head in different ways, thus enhancing the ability of the model to process complex scenes. In order to verify the effectiveness of the proposed model, this paper conducts experiments on multiple UAV datasets of complex scenes. The experimental results show that the accuracy of the improved RT-DETR model proposed in this paper on the two UAV datasets reaches 0.956 and 0.978, respectively, which is 2% and 1.1% higher than that of the original RT-DETR model. At the same time, the FPS of the model is also improved by 10 frames per second, which achieves an effective balance between accuracy and speed.

Keywords: UAV detection; DETR; attention mechanism; multi-scale fusion

check for updates

Citation: Zhu, M.; Kong, E. Multi-Scale Fusion Uncrewed Aerial Vehicle Detection Based on RT-DETR. *Electronics* **2024**, *13*, 1489. https://doi.org/10.3390/electronics13081489

Academic Editor: Eva Cernadas

Received: 3 March 2024
Revised: 9 April 2024
Accepted: 11 April 2024
Published: 14 April 2024

1. Introduction

With the development of uncrewed aerial vehicle (UAV) technology and its wide application, how to effectively monitor and control the flight of UAVs has become an important topic. The flight of UAVs may affect airspace security, civil aviation, military, government, public facilities, personal privacy, etc. And it may even be used for illegal or malicious purposes [1]. In military terms, UAV detection can help the military find and lock the enemy's UAV or other targets [2] for accurate attack or interception. It can also help the military protect its own UAV from being found or interfered with by the enemy and improve the effectiveness and security of surveillance [3]. In areas such as airports, drone detection can help airports prevent drones from entering no-fly areas and causing flight delays or hazards. UAV detection can also help the airport to monitor and manage UAV activities around the airport and maintain the order and safety of the airspace [4]. In areas with high confidentiality, such as government agencies or military bases, UAV detection can help government agencies prevent UAV snooping or threats to important people or occasions and protect the interests and security of the country [5]. Therefore, UAV detection is a key technology that can help us find, locate, track, and manage UAVs to protect the interests and security of society and the country [6].

Traditional UAV detection methods mainly include methods based on radar, acoustic waves, electromagnetic, optical, infrared, and other sensors, but they all have some shortcomings. For example, the disadvantage of radar is that it is easily affected by electromagnetic interference or reflection, and the disadvantage of an acoustic wave is that it has a small detection range and is easily affected by environmental noise or wind speed [7]. The disadvantage of electromagnetic interference or reflection is that its detection range is limited by signal strength and frequency, and it is vulnerable to encryption or spoofing [8]. The disadvantage of infrared is that its detection is affected by ambient temperature and humidity, and it requires complex temperature calibration and analysis [9]. Therefore, optical image-based detection methods can be used, and UAV detection methods using image detection have many advantages that make them highly favored in various application scenarios. For example, it can automate the monitoring process of the UAV and also achieve accurate target detection and positioning in complex environments. Image-based detection methods can be adjusted and optimized according to different environments and weather conditions to ensure that UAVs can be effectively detected in various complex environments [10]. This adaptability makes the image detection system widely applicable in diverse application scenarios.

By using advanced object detection algorithms, we will be able to detect and identify various types of UAVs more accurately, including small UAVs and high-speed vehicles. This will not only help improve public safety but also promote the sustainable development and application of UAV technology. UAV detection can help regulators detect and respond to potential UAV threats in time to protect people's lives and property. In the field of commercial applications, it can provide UAV operators with more reliable monitoring and management solutions to help them better plan flight paths and avoid collisions and unexpected events between UAVs.

With the continuous development of machine learning and computer vision algorithms, image-based detection techniques are constantly improving. Object detection models such as Faster R-CNN [11], You Only Look Once (YOLO) family [12], SSD [13], Mask R-CNN [14], RetinaNet [15], and EfficientDet [16] have been widely used and studied in various object detection. However, these object detection models are rarely used in UAV detection, which has only been commonly used in the last two years and is still limited by factors such as data scarcity, complex environments, diversity of target categories, computational resource requirements, and legal and privacy issues. UAV detection involves various UAV morphologies, complex environmental conditions, and challenges such as data acquisition and privacy protection. More research and resource investment are needed to realize its wide application in practice.

In recent years, some works have been applied to UAV detection. The Facebook AI research team proposed Detection Transformer (DETR) in 2020 [17]. The model uses the Transformer architecture for object detection and realizes object detection and recognition in an end-to-end manner without the use of traditional techniques such as prior boxes and non-maximum suppression. The proposed DETR model has attracted extensive attention and achieved remarkable results in the field of object detection. In 2023, Tushar Sangam et al. improved DETR and Dogfight [18] and applied them to UAV detection. They proposed a simple and effective improved DETR framework TransVisDrone [19], providing an end-to-end solution with higher computational efficiency. The CSPDarkNet-53 network is used to learn the spatial features related to the target. Then, the VideoSwin model is used to understand the spatio-temporal dependencies of the UAV motion to improve the detection ability of the UAV in challenging scenes.

At present, there is a lot of work on designing efficient DETR-based models that are applied to other object detection tasks. However, the high computational cost of these schemes limits the practical application of DETR, which cannot make full use of its advantages, such as non-maximum suppression (NMS). In 2023, Wenyu Lv proposed the real-time detection transformer (RT-DETR) [20], a real-time end-to-end object detector. In particular, they designed an efficient hybrid encoder to process multi-scale features efficiently by decoupling intra-scale interactions and cross-scale fusion. RT-DETR outperforms comparably sized state-of-the-art YOLO detectors in both speed and accuracy. At the same time, Xinyu Liu et al. [21] proposed a new Attention mechanism called Cascaded Group Attention to solve the problems of computational efficiency and attention diversity in vision transformers. By providing different input data segmentation for each attention head, the Cascaded Group Attention reduces computational redundancy and improves attention diversity. Haoran Wei et al. proposed the DWRSeg network [22] in 2023 to address the efficiency of capturing multi-scale context information in real-time semantic segmentation. They designed the dilatation-wise Residual module, which employs a well-designed two-step feature extraction method aimed at capturing multi-scale information efficiently. The module effectively obtains multi-scale context information through region residualization and semantic residualization. Chengcheng Wang et al. [23] improved feature fusion by globally integrating features from different levels using convolution and self-attention operations. They injected global information into features at various levels, thereby enhancing information fusion capability. We are inspired by these works to propose the Gathering Cascaded Dilated DETR (GCD-DETR) model for UAV detection, and we primarily make the following contributions:

(1) First, we propose the DWR-DRB Module, which applies Dilated Re-param Block in the Dilatation-Wise Residual module, uses large kernel convolution with parallel small kernel convolution, and fuses multi-scale perceptual wild generated feature maps. The feature extraction ability is greatly improved.

(2) Secondly, Cascaded Group Attention (CGA) and the Gather-and-Distribute Mechanism (GD) are applied to the RT-DETR model. The model provides complete feature segmentation to each detection head, and the attention calculation is explicitly decomposed to each detection head. Moreover, multi-scale fusion is carried out to save the computational cost and improve the attention to the target feature region.

(3) We design new real-time Transformer models that strike a good balance between efficiency and accuracy. The model has shown good detection ability in a variety of comparative experiments.

The remainder of this paper is organized as follows: In Section 2, we will first review the current state of the art in UAV detection technology to provide a deeper understanding of the background and motivation of the research. In Section 3, we present the details of our proposed novel UAV detection method, including the details of the adopted GCT-DETR model and the modules in it. In Section 4, we will present and analyze the experimental results to verify the effectiveness of our model. Finally, we summarize and discuss the results of this study and suggest future research directions and improvements in Section 5.

2. Related Work

2.1. Drone Detection

UAV dataset has challenges such as high-altitude perspective, low resolution, motion blur, illumination change, and occlusion. Therefore, there are few existing public and recognized datasets that are faced with the problems of insufficient data volume and low data quality. In recent years, deep learning has been widely used in the field of UAV detection, but it uses non-uniform data sets. Therefore, there is still a lot of room for the development of UAV detection, and it is necessary to gradually improve the quality of various environmental datasets and use more efficient and accurate models.

In 2020, Ulzhalgas et al. embarked on the UAV detection challenge by dissecting it into two distinctive subtasks: moving object detection and object classification. Their approach was innovative as it leveraged background subtraction for moving object detection while relying on the robust feature extraction capabilities of convolutional neural networks (CNNs) for object classification [24]. This division of labor ensured a comprehensive and efficient method for identifying UAVs amidst complex visual backgrounds.

Aamish Sharjeel introduced a groundbreaking UAV detection methodology in 2021, merging Continuous Outlier Representation with Online Low-rank Approximation (COROLA) alongside CNNs. The brilliance of this approach lay in COROLA's adeptness at pinpointing small moving objects within scenes, complemented by CNNs' prowess in accurately classifying UAVs across diverse and intricate backgrounds [25]. This amalgamation not only fortified the detection system's resilience but also significantly elevated its efficacy. In the same year, Muhammad et al. proposed "Dogfight" [18], a novel approach diverging from conventional region proposal-based methods. Instead, they adopted a two-stage segmentation technique grounded on spatio-temporal attention cues. Their method intricately incorporated pyramid pooling to capture detailed contextual information within convolutional feature maps, followed by pixel and channel-level attention mechanisms to precisely localize UAVs. This sophisticated strategy underscored a paradigm shift in UAV detection methodologies, prioritizing accuracy and adaptability. Yaowen Lv et al. introduced a novel detection paradigm in 2022, intertwining background difference analysis with the lightweight SAG-YOLOv5s network. By exploiting background difference, their method effectively isolated potential UAV targets within high-resolution images while concurrently minimizing computational overhead by eliminating extraneous background elements [26]. This innovative fusion of techniques showcased a leap forward in optimizing detection efficiency while conserving computational resources. Yuliang Zhao's 2023 proposal, the information enhancement model TGC-YOLOv5, marked a significant advancement in UAV detection methodologies. By integrating Transformer encoder modules and Global Attention Mechanisms (GAMs) into YOLOv5, the model exhibited a twofold increase in detection accuracy. This augmentation facilitated enhanced focus on the regions of interest while mitigating information diffusion across layers, thus enhancing the model's overall effectiveness [27]. Jun-Hwa Kim's 2023 contribution revolutionized UAV detection by integrating multi-scale image fusion layers and P2 layers into the YOLO-V8 medium model. This integration aimed at bolstering the model's adaptability to diverse UAV scales, thereby fortifying its robustness in detection scenarios [28]. This strategic enhancement underscored a concerted effort towards ensuring comprehensive and accurate UAV detection across varying environmental conditions. Qianqing Cheng's 2023 innovation, the CA-PANet multi-scale attention module, heralded a breakthrough in feature fusion for UAV detection. Leveraging improved MobileViT as a feature extraction network, the introduction of coordinate attention within PANet facilitated enhanced fusion of low-dimensional and high-dimensional features. This not only enriched location information capture but also significantly augmented detection accuracy, highlighting a pivotal advancement in UAV detection methodologies [29].

2.2. Detection Transformer

Conditional DETR makes an innovative improvement to solve the problem of slow convergence speed of DETR. In particular, this method increases the number of queries from 100 to 300 and optimizes the classification loss by adopting Focal loss to improve the performance of the model. The key contribution of conditional DETR is the proposal of the conditional attention mechanism. By decoupling content attention and location attention, they implement a redesign of self-attention and cross-attention inputs. The original method is to add query and query_pos and input them into the linear layer of the attention structure. At the same time, Conditional DETR modifies it so that query and query_pos go through different linear layers, respectively, and then aggregate the results, thereby improving the effect of the attention mechanism of the model [30]. Deformable DETR proposes Multi-scale Deformable Attention (MSDA) to replace Self-attention in the Encoder and Cross-attention in the Decoder. The model of DETR multi-scale feature detection is designed, which not only gives DETR the advantage of multi-scale but also reduces the amount of calculation. In addition, it also proposes the idea of a two-stage DETR, which uses the encoder output features to initialize the decoder query and its corresponding position [31]. Sparse DETR offers an effective encoder token sparsification method for end-to-end object detectors, by which the attention complexity in the encoder is reduced. This efficiency allows Deformable DETR to stack more encoder layers, thus improving performance with the same amount of computation [32]. The end-to-end object detection algorithm DETR does not require hand-crafted post-processing (NMS), but it requires longer training to converge. It is found that one-to-one label matching makes DETR lack supervision signal in the training process (because the number of positive object Queries is small), so it needs to extend the training time to achieve good results.

Group DETR provides a new label assignment strategy for the DETR family of algorithms: group-wise One-to-Many label assignment. The algorithm cleverly decouples the "one-to-many allocation" problem into the "one-to-one allocation of multiple groups" problem. It accelerates the convergence of DETR series algorithms, removes redundant predictions while ensuring the support of multiple positive queries, and realizes end-to-end detection [33]. In 2023, Decoupled DETR proposes the Task-aware Query Generation Module: this module is responsible for initializing queries to match different visual regions, thus providing more suitable features for classification and localization tasks. They also propose a Disentangled Feature Learning Process: in this process, the classification and localization tasks are spatially separated, allowing task-aware queries to be matched to different visual regions. It solves the problem of space misalignment encountered in traditional DETR training [34].

3. Proposed Methods

Figure 1 shows the network structure of GCD-DETR designed in this paper. In the backbone part, we first propose the Dilation-wise Residual and Dilated Re-param Block (DWR-DRB) module for feature extraction, which can reduce the difficulty of extracting multi-scale context information and is an efficient multi-scale feature extraction method. Next, we introduce the Cascaded Group Attention module (CGA). Cascaded Group Attention assigns different weights to the feature maps based on the relevance of different positions in the input image. It can help the model better understand the features in the image, thus improving the detection performance. In the Neck part, we use a novel information interaction and fusion Mechanism: Gather-and-Distribute mechanism (GD). The mechanism obtains global information by fusing features at different levels globally and injects global information into features at different levels to achieve efficient information interaction and fusion. It improves the detection ability of the model for objects of different sizes.

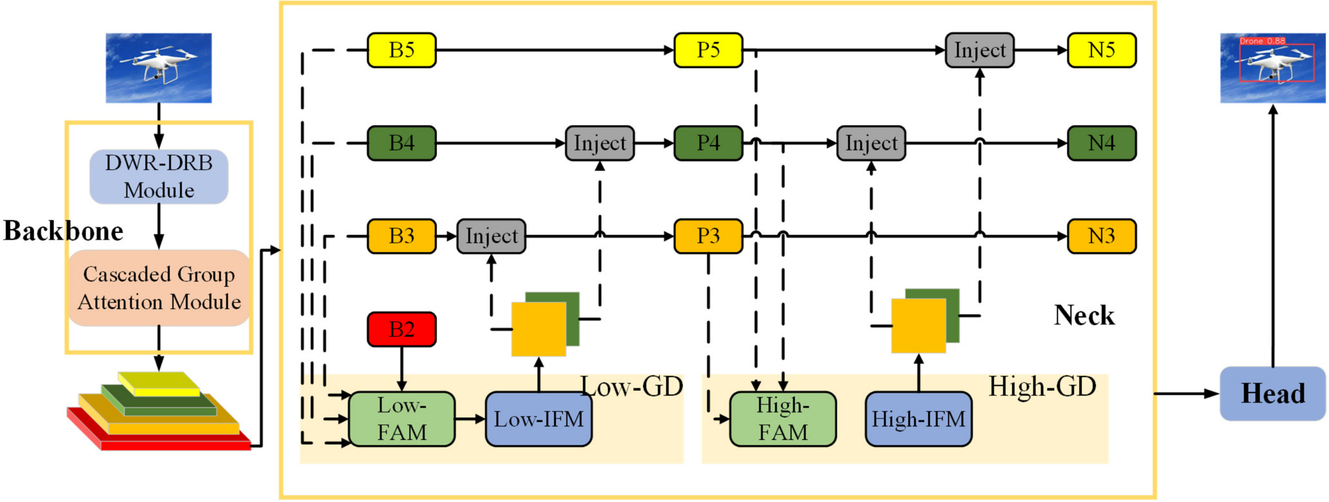


Figure 1. The network structure of GCD-DETR.

3.1. Dilation-Wise Residual and Dilated Re-Param Block Module

In this section, we propose the Dilation-wise Residual and Dilated Re-param Block (DWR-DRB). We incorporate the Dilated Re-param Block into the dilatation-wise Residual module and utilize a combination of large kernel convolutions and parallel small kernel convolutions. Furthermore, we fuse multi-scale receptive field-generated feature maps, thereby significantly enhancing the feature extraction capability.

3.1.1. Dilated Re-Param Block

In convolutional neural networks (CNNs), combining large kernel convolutions with parallel small kernel convolutions helps capture features at various scales. Their outputs are summed after two respective batch normalization (BN) layers [35]. The structural re-parameterization method [36] can be employed to integrate BN layers with convolutional layers, and after training, they can be merged effectively to incorporate small kernel convolutions into large kernel convolutions for inference.

The main idea of Dilated Re-param Block (DRB) structure is to use the combination of large kernel convolution and dilated convolution to improve the performance of convolutional neural network. By using large kernel convolutions and multiple dilated convolutional layers in parallel, the "Dilated Re-param Block" structure is able to capture both local fine features and widely distributed sparse features. This combination enables the model to perceive the structural information of the input data more comprehensively. The whole module is converted into a single non-dilated convolutional layer in the inference phase. This step consists of converting each dilated convolutional layer into an equivalent non-dilated convolutional layer and merging their output feature maps. The performance of the structure can be flexibly controlled by adjusting the large kernel size K, the dilation rate r, and the small kernel size k, as shown in Figure 2, $K = 9$, $r = (1,2,3,4)$, and $k = (5,5,3,3)$ [37].

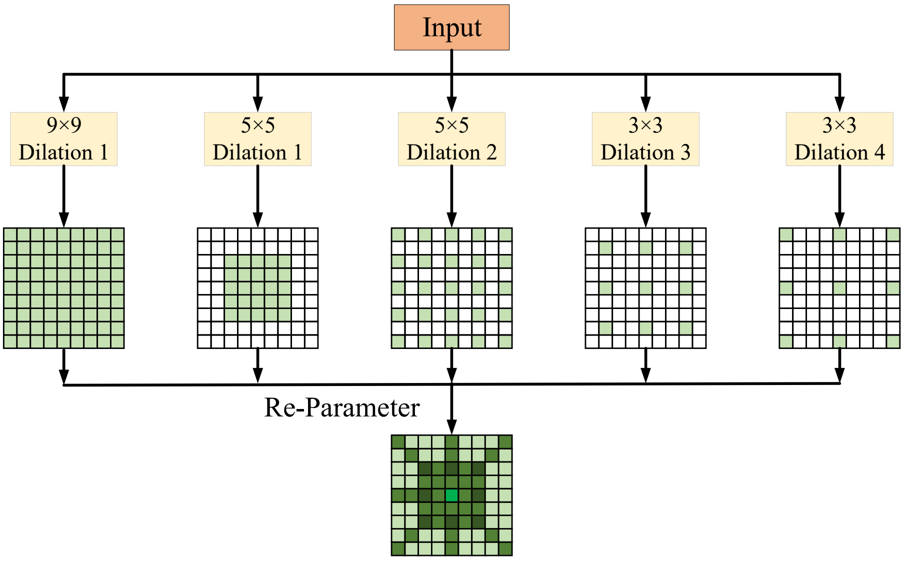


Figure 2. Dilated Re-param Block.

3.1.2. Dilation-Wise Residual and Dilated Re-Param Block Module

In this section, we introduce the Dilation-wise Residual and Dilated Re-param Block (DWR-DRB) module, designed to efficiently acquire multi-scale context information, as illustrated in Figure 3. This module effectively extracts and fuses feature maps generated from multiple receptive fields through a two-step multi-scale context information acquisition method within the dilation-wise Residual (DWR) [22] in combination with the previously mentioned Dilated Re-param Block.

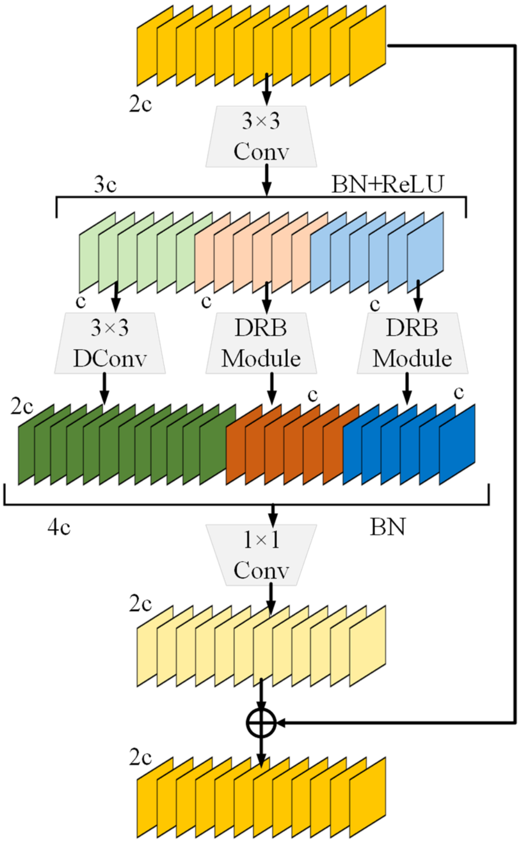


Figure 3. DWR-DRB module.

In particular, the first step involves generating region residual features from input features through a regular 3 × 3 convolution combined with batch normalization (BN) and ReLU activation. The second step employs multi-rate depth-wise deformable convolution (DConv) and Dilated Re-param Block (DRB) modules to perform morphological filtering on region features of different sizes, referred to as semantic residualization. This method not only extracts multi-scale contextual information but also refines features and effectively controls redundant receptive fields through the generation of region residual features and reverse matching of receptive fields. Consequently, the model maintains simplicity while dealing with complex semantic information and achieves significant improvements in feature representation and model performance. Furthermore, aggregating multiple output feature maps, employing batch normalization, and merging feature maps using pointwise convolution enhance the model's perception of multi-scale information, thereby improving feature representation and performance.

3.2. Cascaded Group Attention

Cascaded Group Attention is based on the concept of group attention, dividing the image into multiple groups or regions and focusing on features within each group. Unlike traditional global attention mechanisms, Cascaded Group Attention achieves more intricate feature focus by cascading multiple layers of attention.

In Cascaded Group Attention, the input image is first divided into groups, where each group may contain specific semantic information or adjacent pixels in space. Subsequently, a local attention mechanism is applied to each group, allowing the network to concentrate more on the features within each group. This progressive focusing process enables the network to refine feature representation at multiple levels, thereby enhancing the model's perceptual ability and accuracy in feature representation [21]. This attention mechanism can be described as follows:

$$\widetilde{X}_{ij} = Attn(X_{ij}W_{ij}^{Q}, X_{ij}W_{ij}^{K}, X_{ij}W_{ij}^{V}) \tag{1}$$

$$\widetilde{X}_{i+1} = Concat[\widetilde{X}_{ij}]_{j} = {}_{1:h}W_{i}^{P} \tag{2}$$

For the j-th attention head, it computes self-attention on the j-th split X_{ij} of the input feature X_i. The input feature X_i is divided into h different splits, with each split corresponding to one attention head. This partitioning is achieved by projection layers $W_{ij}^{Q}, W_{ij}^{K}, W_{ij}^{V}$, which split the input feature X_i into different subspaces. The purpose of splitting the input features into different subspaces is to compute the self-attention on each subspace. W_i^P is the linear layer. These projection layers map the input feature into different subspaces, enabling self-attention computation in each subspace.

During the computation of the attention map for each head, a cascaded approach is employed, where the output of each head is added to the subsequent heads' inputs. This cascading method facilitates gradual improvement in feature representation, enabling the model to better capture the structure and relationships within the data, as shown in Figure 4. This entails aggregating the output of each attention head with that of the following heads, enabling the iterative enhancement of feature representation:

$$X'_{ij} = X_{ij} + \widetilde{X}_{i(j-1)}, \quad 1 < j \leq h \tag{3}$$

the output X'_{ij} of each head is the addition of its input split X_{ij} and the output $X\ _{i(j-1)}$ of the previous head $(j-1)$, calculated by Equation (2). $X\ _{i(j-1)}$ replaces X_{ij} as the new input feature for computing self-attention in the j-th head. Additionally, after Q projection, there is a flag indicating the inclusion of an interaction layer, enabling the self-attention mechanism to simultaneously capture both local and global features.

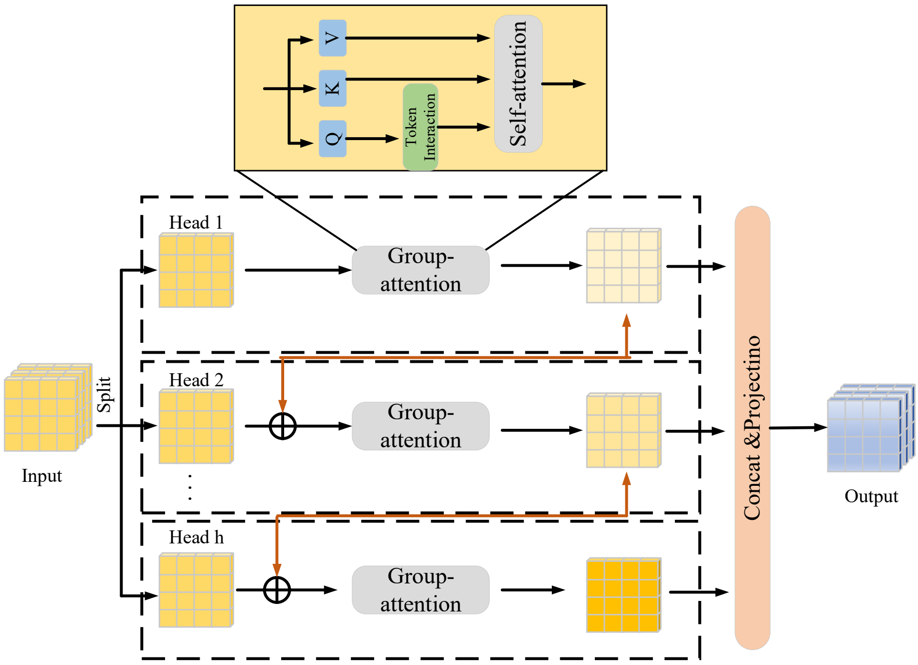


Figure 4. Cascaded Group Attention module.

The introduction of the Cascaded Group Attention module allows the model to effectively focus on specific regions or features, thereby enhancing feature representation. By iteratively refining feature representation at multiple levels, this module improves the model's perceptual ability and accuracy. Additionally, its adaptability to various datasets and scenarios further highlights its versatility and effectiveness in enhancing feature understanding and interpretation.

3.3. Gather-and-Distribute Mechanism

When detecting UAV targets, targets of different sizes are often generated due to the distance between them. In order to improve the detection ability of these targets, we use the low-stage gather-and-distribute branch (Low-GD) and the high-stage gather-and-distribute branch (High-GD) [23]. The core idea of this method is to use different feature extraction and fusion strategies for different sizes of objects to better adapt to various sizes of target objects. Two key modules are included in both branch networks: feature alignment module (FAM) and information injection module (IFM), as shown in Figures 5 and 6. These modules are designed to efficiently extract and fuse feature maps from the backbone network in order to better capture various size features of the target object. The inputs of these two branch networks are the feature maps B2, B3, B4, and B5 output by the backbone network, where $B_i \in R^{N \times C_{Bi} \times R_{Bi}}$. Here, the batch size is denoted by N, the channel by C, and the feature map size by $R = H \times W$, where H and W denote the height and width of the feature map, respectively. And the dimensions of $R_{B2}, R_{B3}, R_{B4}, R_{B5}$ are R, $\frac{1}{2}R$, $\frac{1}{4}R$, and $\frac{1}{8}R$, respectively.

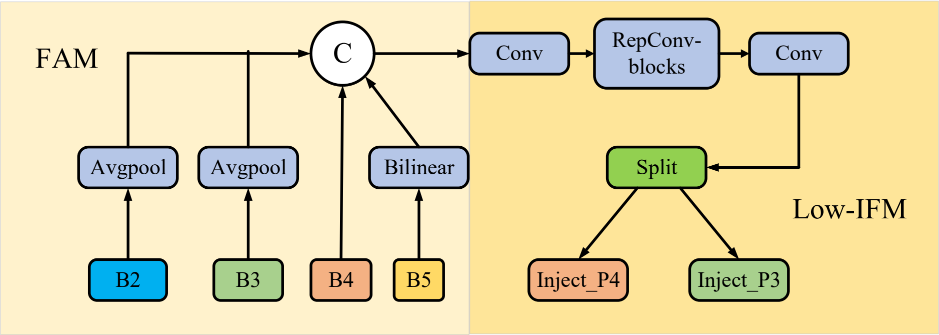


Figure 5. Low-stage gather-and-distribute branch.

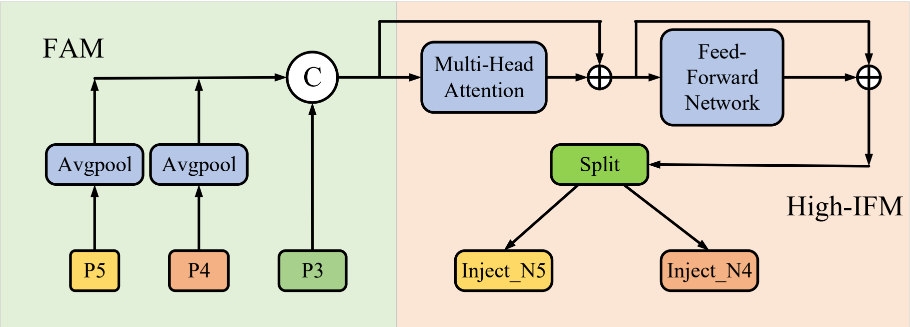


Figure 6. High-stage gather-and-distribute branch.

3.3.1. Feature Alignment Module

The main function of FAM module is to align the feature maps of different levels to a uniform size and then merge these feature maps by the concatenation operation on the channel. This reduces information loss and enhances the ability of the model to detect objects of different sizes without significantly increasing the latency.

In particular, the FAM module will first adjust the input feature maps to the same spatial resolution by average pooling operation, and then concatenate them in the channel dimension. In this process, the feature map is resized to the smallest feature size within the group in order to control the computation latency while preserving low-level information. As shown in Figure 5, if the feature maps of B2, B3, B4, and B5 correspond to different dimensions, the FAM module will unify them to the dimensions of B4 and then concatenate them on the channel. In Figure 6, the input feature maps are P3, P4, and P5, and the FAM module will unify them to the dimensions of P3. There are several benefits to resizing the feature map to the dimensions of B4. Firstly, B4 provides a balance point where it is neither the largest feature map (as in B2) nor the smallest feature map (as in B5), which means that it is able to preserve sufficient details while avoiding the computational stress caused by processing overly large feature maps. Secondly, feature alignment using B4 as a benchmark can better preserve the information of medium-size objects, which is especially important for object detection, as it ensures that the model can effectively detect objects of various sizes. Finally, selecting B4 as the benchmark for alignment can simplify the process of information fusion. By resizing all feature maps to the size of B4, the concatenation can be performed directly on the channel, which reduces the average pooling operation, which helps to reduce the latency and makes the model more suitable for real-time application scenarios. Therefore, B4 is chosen as the benchmark for feature alignment in order to find a compromise between preserving critical information and controlling computational cost to improve the performance and efficiency of the model.

3.3.2. Information Fusion Module

The Information Fusion Module (IFM) is designed to improve the ability of multi-scale feature fusion. As shown in Figure 5, first, the low-stage IFM (Low-IFM) receives the aligned feature maps from the FAM module. These feature maps have been unified in spatial resolution for further processing. The aligned feature map goes through a multi-layer Rep-Block structure, which is a combination of a series of convolutional layers and activation functions to extract and enhance feature information. The feature maps processed by Rep-Block will be split into two parts in the channel dimension, which can provide more specialized information for feature maps at different scales. The segmented feature maps are regarded as global information, and they will be used to inject features at different levels to achieve effective information interaction and fusion. IFM is designed to reduce information loss and enhance the model's ability to detect objects of different sizes without significantly increasing latency. This mechanism obtains global information by fusing features at different levels globally and injects global information into features at different levels to achieve efficient information interaction and fusion. The advantage of this procedure is that it allows the model to make better use of the features extracted from backbone and can be easily integrated into existing similar network structures. Through its design, IFM improves the overall performance of the model, making it more accurate and efficient when dealing with objects of different sizes.

The High-stage Information Fusion Module (High-IFM) is designed to improve the accuracy of object detection while maintaining low latency. As shown in Figure 6, High-FAM first receives feature maps from different layers of the network and aligns them to a uniform spatial resolution. This step is carried out by FAM, which ensures that the feature maps have the same dimensions before the subsequent fusion. The aligned feature maps are then passed to High-IFM, where transformer modules are used for processing. Each transformer module consists of a multi-head attention module and a feedforward network module. These operations allow the model to combine features at a higher level, which are typically more abstract but contain more semantic information. These modules work together to capture long-distance dependencies between features. The High-IFM processed feature maps are channel-simplified by the Conv 1 × 1 operation, which helps to reduce the computational complexity and maintain the efficiency of the model.

Feature segmentation and fusion: The reduced feature map is segmented in the channel dimension and fused with the horizontal features of the current stage. This step ensures that the features of different levels can be effectively combined, thus improving the model's ability to detect objects of different sizes.

3.3.3. Information Injection Module

To effectively utilize global information in images and inject it into different levels of feature representations, we employ the information injection module for information fusion, as illustrated in Figure 7.

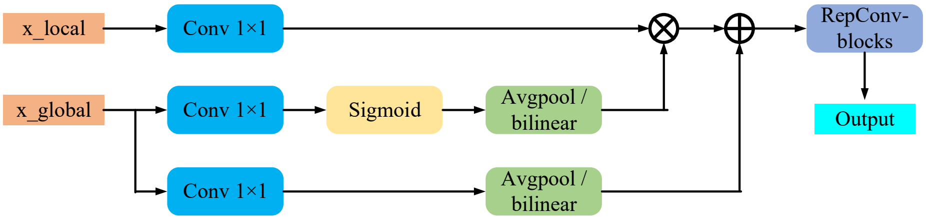


Figure 7. Information injection module.

The information injection module is responsible for injecting global information into different levels of features to enhance the ability of the model to detect objects. The information injection module first receives global information from the Information Fusion Module, which contains features fused from different levels of the network. The module uses the attention mechanism to weight the received global information. This step highlights the key information by calculating the importance of each feature while suppressing the unimportant information. The weighted global information is subsequently injected into the local features at the current level. This process is accomplished by specific operations such as addition or concatenation, which enables the effective combination of global information with local features. By injecting global information, local features are enhanced, allowing the model to better understand and recognize objects in the image. Finally, the information injection module outputs feature representations that fuse global and local information, and these features will be used in subsequent object detection layers.

4. Experiments

4.1. Datasets and Implementation Details

We utilized two UAV datasets in our experiments. We first utilized the Rotor UAV dataset proposed by DASMEHDIXTR et al., which consists of 1360 images of drones. All images are labeled with the class “drone” and include various complex backgrounds and drone models. We randomly selected 1000 images as the training set, 200 images as the validation set, and 160 images as the test set. The dataset we used can be found at https://www.kaggle.com/datasets/dasmehdixtr/drone-dataset-uav/data (accessed on 1 February 2024).

We also utilized an open-source, available military UAV dataset on Roborflow. This dataset comprises multiple environments, including sky, city, countryside, and coastline. It encompasses UAV images captured under different weather conditions and at various times, covering a wide range of UAV use scenarios. The dataset can be found at https://universe.roboflow.com/military-drone/dronemil-u8fqk (accessed on 7 February 2024). It consists of 5238 training images, 1345 validation images, and 678 test images.

For training and evaluation, we conducted numerous experiments using these two UAV datasets. The model architecture was implemented using PyTorch 1.11.0 and Timm 0.5.4. The model was trained from scratch for 200 epochs on 2 Nvidia V100 GPUs using the AdamW optimizer and cosine learning rate scheduler. The size of all the image is 640 × 640. We used a batch size of 16. The input images were resized and randomly cropped to a size of 640 × 640. The initial learning rate is 1×10^{-4}, and the weight decay is 2.5×10^{-2}.

4.2. Comparision Results

4.2.1. Comparison with Prior Works

To evaluate the performance of our UAV detection model, we conducted a series of comparative experiments. Firstly, we compared the number of parameters and Floating

Point Operations Per Second (FLOPs) across various models to evaluate their efficiency. Subsequently, we evaluated the models using metrics such as Recall, AP@50, and AP@50:95. Recall measures the proportion of correctly detected objects among all labeled objects. AP@50 represents the Mean Average Precision for each class when the Intersection over Union (IOU) threshold is set to 0.5. AP@50:95 calculates the average AP over different IOU thresholds ranging from 0.5 to 0.95 with a step size of 0.05.

Our model was compared against several established models, including different versions of YOLOv7 [12], YOLOv8, RT-DETR [20], and the latest Gold-YOLO [23]. The comparison results on Rotor UAV dataset are presented in Table 1.

Table 1. Comparative experiments with prior works on Rotor UAV dataset.

Model	Input Size	Backbone	Neck	Layers	Parameters	GFLOPs	Recall	AP@50	AP@50:95
YOLOv7 [12]	640	CBS + ELAN	SPPSCP + E-ELAN	415	37,196,556	105.1	0.814	0.858	0.476
YOLOv7x [12]	640	CBS + ELAN	SPPSCP + E-ELAN	467	70,815,092	188.9	0.837	0.883	0.53
YOLOv7-w6 [12]	1280	CBS + ELAN	SPPSCP + E-ELAN	477	80,944,472	102.4	0.824	0.924	0.57
YOLOv7-d6 [12]	1280	CBS + ELAN	SPPSCP + E-ELAN	733	152,886,360	198.3	0.878	0.934	0.588
YOLOv8s	640	C2F + SPPF	C2F	168	11,125,971	28.4	0.878	0.941	0.626
YOLOv8m	640	C2F + SPPF	C2F	295	25,856,899	79.1	0.864	0.948	0.631
YOLOv8n	640	C2F + SPPF	C2F	225	3,157,200	8.9	0.90	0.949	0.626
Gold-YOLO-s [23]	640	Efficient-Rep	Gather-and-Distribute	/	21.5M	46.0	0.670	0.928	0.582
Gold-YOLO-m [23]	640	Efficient-Rep	Gather-and-Distribute	/	41.3M	87.5	0.693	0.934	0.604
Gold-YOLO-n [23]	640	Efficient-Rep	Gather-and-Distribute	/	5.6M	12.1	0.671	0.919	0.580
RT-DETR-r18 [20]	640	ResNet 18	AIFI + CCFM	299	19,873,044	56.9	0.941	0.936	0.621
RT-DETR-r34 [20]	640	ResNet 34	AIFI + CCFM	387	31,106,233	88.8	0.896	0.933	0.602
RT-DETR-r50 [20]	640	ResNet 50	AIFI + CCFM	629	42,782,275	134.4	0.878	0.905	0.581
GCD-DETR (Ours)	640	DWR-DRB + CGB	Gather-and-Distribute	494	23,262,488	61.0	0.93	0.956	0.624

We list the network structures used in the backbone and neck parts of all models. The backbone of Yolov7 uses CBS (Conv + BN + SiLU) and ELAN modules, which are composed of multiple CBS modules. The neck part is mainly composed of information fusion module SPPCSP and ELAN module. The backbone and neck of YOLOv8 are mainly composed of C2f (CSPLayer2Conv) module and SPPF (Spatial Pyramid Pooling Fast). C2f has more skip connections and additional split operations. Gold-YOLO is mainly composed of Efficient Repblock and Gather-and-Distribute (GD). The GD mechanism significantly enhances the information fusion ability of the neck part and improves the detection ability of the model for objects of different sizes. RT-DETR is mainly composed of Attention-based Intro-scale Feature Interaction (AIFI) module and the CNN based Cross-scale Feature-fusion Module (CCFM).

The backbone of our model mainly consists of the Dilation-wise Residual and Dilated Re-param Block Module (DWR-DRB) module and Cascaded Group Attention (CGB) module, and the neck is mainly composed of GD mechanism. Despite variations in the parameters of these aforementioned methods, the accuracy of the largest model from each of these methods is consistently lower than that of our proposed GCD-DETR model.

On the rotor UAV dataset, our model achieves a 2% higher AP@50 compared to the highest accuracies of RT-DETR prior to improvement and a 1% higher accuracy than YOLOv8n. On the military UAV dataset, as shown in Table 2, our model exhibits a 1% higher AP@50 than YOLOv8s and a 0.5% higher AP@50 than Gold-YOLO -s. These results demonstrate that our model can achieve high accuracy with a parameter count similar to other models. Our model showcases its lightweight nature by achieving higher accuracy while having significantly fewer GFLOPs compared to YOLOv7-d6. It indicates that our model can achieve precision while being lightweight and can be more easily deployed on UAV equipment. The effectiveness of our model can be attributed to the multi-dimensional feature extraction of the DWR-DRB module and the Gather-and-Distribute mechanism, which efficiently combines features from different maps. Additionally, the skip connection helps reduce computational requirements, further contributing to high precision with minimal computation.

Table 2. Comparative experiments with prior works on military UAV dataset.

Model	Input Size	Year	Layers	Parameters	GFLOPs	Recall	AP@50	AP@50:95
YOLOv7 [12]	640	2022	415	37,196,556	105.1	0.914	0.955	0.624
YOLOv7x [12]	640	2022	467	70,815,092	188.9	0.919	0.958	0.631
YOLOv7-w6 [12]	1280	2022	477	80,944,472	102.4	0.92	0.954	0.633
YOLOv7-d6 [12]	1280	2022	733	152,886,360	198.3	0.922	0.964	0.657
YOLOv8s	640	2023	168	11,125,971	28.4	0.934	0.968	0.687
YOLOv8m	640	2023	295	25,856,899	79.1	0.946	0.959	0.658
YOLOv8n	640	2023	225	3,157,200	8.9	0.957	0.962	0.676
Gold-YOLO-s [23]	640	2023	/	21.5M	46.0	0.897	0.973	0.680
Gold-YOLO-m [23]	640	2023	/	41.3M	87.5	0.944	0.953	0.636
Gold-YOLO-n [23]	640	2023	/	5.6M	12.1	0.950	0.958	0.675
RT-DETR-r18 [20]	640	2023	299	19,873,044	56.9	0.954	0.953	0.692
RT-DETR-r34 [20]	640	2023	387	31,106,233	88.8	0.925	0.977	0.639
RT-DETR-r50 [20]	640	2023	629	42,782,275	134.4	0.927	0.967	0.657
GCD-DETR (Ours)	640	/	494	23,262,488	61.0	0.966	0.978	0.711

The DWR-DRB module has significant advantages in dealing with multi-scale information. Through deep-separated dilated convolution and a two-step residual feature extraction method, it can effectively extract the features of small objects and perform well in real-time semantic segmentation tasks. This allows the DWR-DRB module to outperform traditional backbone networks such as Gold-YOLO and RT-DETR in terms of accuracy and efficiency, especially in scenarios where large amounts of detail and dynamic range need to be processed. In addition, the design of the DWR-DRB module also considers the optimization of computing resources so that it can maintain high performance even in resource-constrained environments. The advantage of the DWR-DRB module and the CGB module is its advanced multi-scale feature extraction ability, especially for small object detection and real-time semantic segmentation tasks, which provides more accurate feature extraction than YOLOv7 and YOLOv8 through deep separation dilated convolution and refined receptive field design, thereby improving the overall network performance.

4.2.2. Comparison of Evaluation Metrics

We conducted a comparison between the training process curves of the original RT-DETR and our model in the military UAV detection task. Throughout the training process, we recorded the precision and recall values of each model on the training set, as well as AP@0.5 and AP@0.5:0.95. We plotted corresponding curves to observe their training progress in Figure 8.

Our model demonstrates superior performance in terms of precision. The precision curve of our model maintains a high level of stability during training and converges to a high precision level. Conversely, while the precision curve for the original RT-DETR starts with an increasing pattern, it experiences considerable fluctuations during training, ultimately settling at a precision slightly inferior to that of our model.

Regarding recall, our model also surpasses the original RT-DETR. The recall curve of our model exhibits early-stage growth, maintains a stable upward trend throughout training, and finally converges to a high level. However, the recall curve of the original RT-DETR demonstrates a slower growth rate and noticeable fluctuations in the later stages of training. Overall, our model showcases higher precision and recall during training, along with better stability and faster convergence speed.

In addition, we compare the AP@0.5 and AP@0.5:0.95 curves between our model and the original RT-DETR for object detection. These metrics evaluate the detection accuracy and robustness of the models at different confidence thresholds. In terms of AP@0.5, our model outperforms the original RT-DETR at lower confidence thresholds. The curve shows a faster upward trend and eventually reaches a relatively high average precision. Conversely, the AP@0.5 curve of the original RT-DETR exhibits slower growth, and the detection performance at lower confidence thresholds is slightly lower than that of our model.

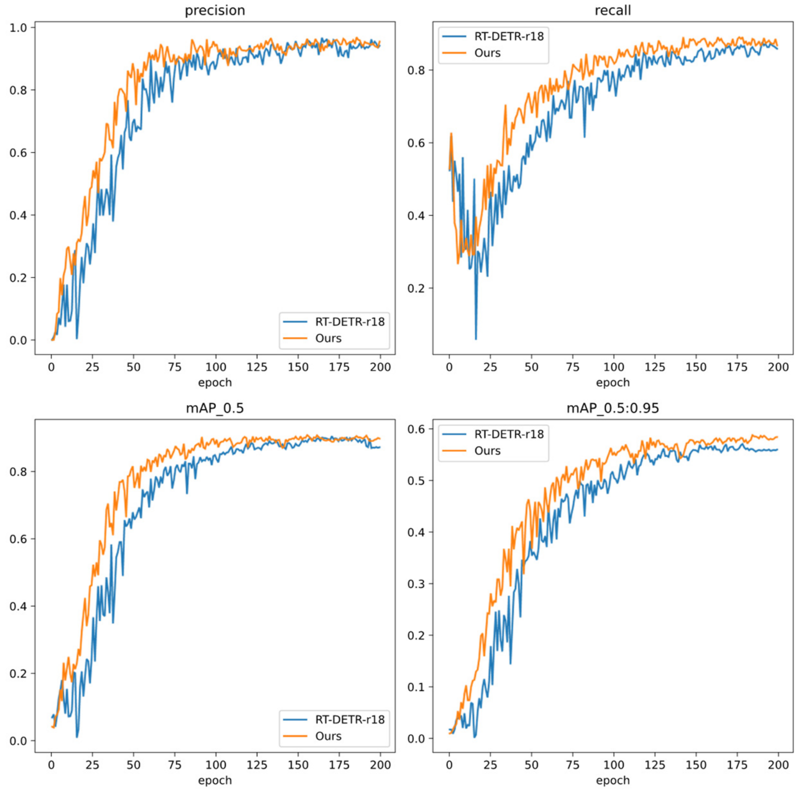
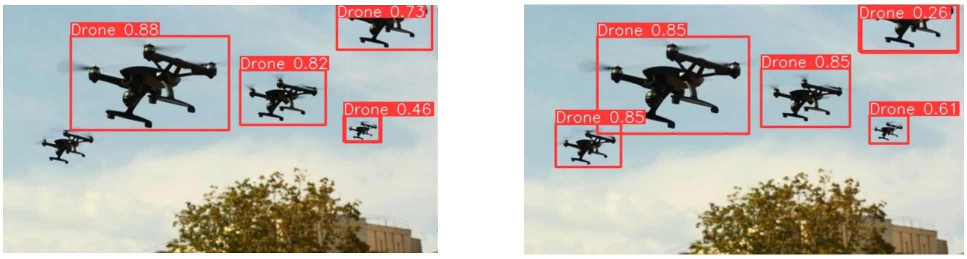


Figure 8. Comparison of evaluation metrics between GCD-DETR and RT-DETR-r18.

When considering the AP@0.5:0.95 curve, the performance gap between our model and the original DETR becomes more evident at higher confidence thresholds. Our model's curves maintain high stability and exhibit high average accuracy across a range of confidence thresholds from 0.5 to 0.95. However, the AP@0.5:0.95 curve of the original RT-DETR performs poorly at high confidence thresholds, with significantly lower average accuracy compared to our model. Therefore, in a comprehensive sense, our model demonstrates better detection performance not only at low confidence thresholds but also at high confidence thresholds, showcasing better overall robustness. Overall, our model demonstrates superior precision, recall, and detection accuracy during training when compared to the original RT-DETR.

4.2.3. Comparison of Detection

We conducted a comparison of the images detected by the original RT-DETR-r18 model and our model. Figure 9a illustrates that the RT-DETR-r18 model fails to detect multiple targets, while Figure 9b demonstrates that our GCD-DETR successfully detects all targets. This discrepancy may arise from the fact that when neighboring targets are in close proximity, the larger target can obstruct the detection of the smaller target, resulting in missed detections. However, our model overcomes this limitation by incorporating an attention module and performing multi-scale feature fusion. As a result, our model accurately identifies objects of various sizes, hence achieving the detection of all objects in the given scenario.

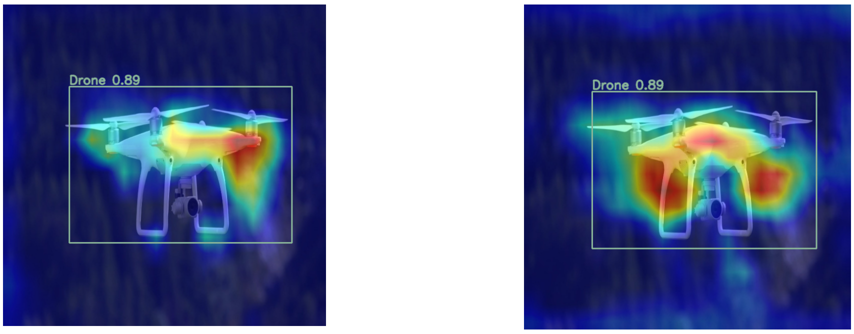


(**a**) (**b**)

Figure 9. Comparison of detection results between RT-DETR-r18 and GCD-DETR: (**a**) RT-DETR-r18 and (**b**) GCD-DETR.

4.2.4. Comparison of Heatmap

The heatmap is a visualization technique utilized in object detection to display the intensity distribution of objects detected by a model in an input image. Heatmaps are commonly employed to indicate the location and confidence of the detected target, with brighter areas representing higher confidence in the detection. In Figure 10, we compare the heatmaps generated by the RT-DETR-r18 model and our proposed model.



(**a**) (**b**)

Figure 10. Comparison results of heat map visualization: (**a**) RT-DETR-r18 and (**b**) GCD-DETR.

The first heatmap corresponds to the RT-DETR-r18 model, revealing that the deeper regions of intensity are concentrated only in a specific area of the UAV, while other regions show lower levels of focus. Conversely, the second heatmap corresponds to our model, where darker-colored areas are concentrated on the body and support parts of the UAV, encompassing almost the entire UAV. Additionally, a high level of attention is observed towards the overall structure of the UAV in our model's heatmap, indicating a more confident detection of the UAV target. These findings demonstrate the effectiveness of our model in detecting UAVs compared to the RT-DETR-r18 model.

4.3. Training Metrics

During the training process of object detection models, loss functions are the key indicators used to assess the accuracy of model predictions and guide model optimization. The Generalized Intersection over Union Loss (giou_loss) is a metric for evaluating the localization accuracy of object detection models. It measures not only the overlap between the predicted and actual bounding boxes but also includes a penalty term that considers the area of the smallest enclosing box containing both bounding boxes. The lower the giou_loss, the more accurate the model is at localizing targets. Classification Loss (cls_loss) is used to measure the model's performance in recognizing and classifying targets. It calculates the difference between the model's predicted output and the actual target values. In object detection, this often involves classification problems, and the lower the Classification Loss, the more accurate the model is at classification tasks. The l1 loss is a method for measuring the difference between predicted values and actual values. It works by calculating the average absolute difference between them. This method performs well when dealing with outlier data because it does not allow individual extreme values to overly influence the overall loss. The lower the l1_loss, the better the model's performance.

Figure 11 shows the changes in the various metrics of our model during the training process. We note that the GCD-DETR model not only exhibits smooth performance across all evaluation indicators but also consistency. The steady decline in giou_loss indicates a continuous improvement in the model's accuracy in target localization. The reduction in Classification Loss reflects an enhanced ability of the model to distinguish between different categories of targets. The gradual decline in the l1_loss demonstrates a better balance between precision and recall. These smooth curves of the indicators show that the model exhibits stability and reliability during training, with no overfitting or underfitting issues. Additionally, our model performs exceptionally well in AP@0.5 and AP@0.5:0.95, meaning it maintains high-level performance in object detection tasks of varying difficulty levels. These results indicate that the GCD-DETR model excels not only in single tasks but also has strong adaptability and robustness when dealing with diverse and complex object detection scenarios.

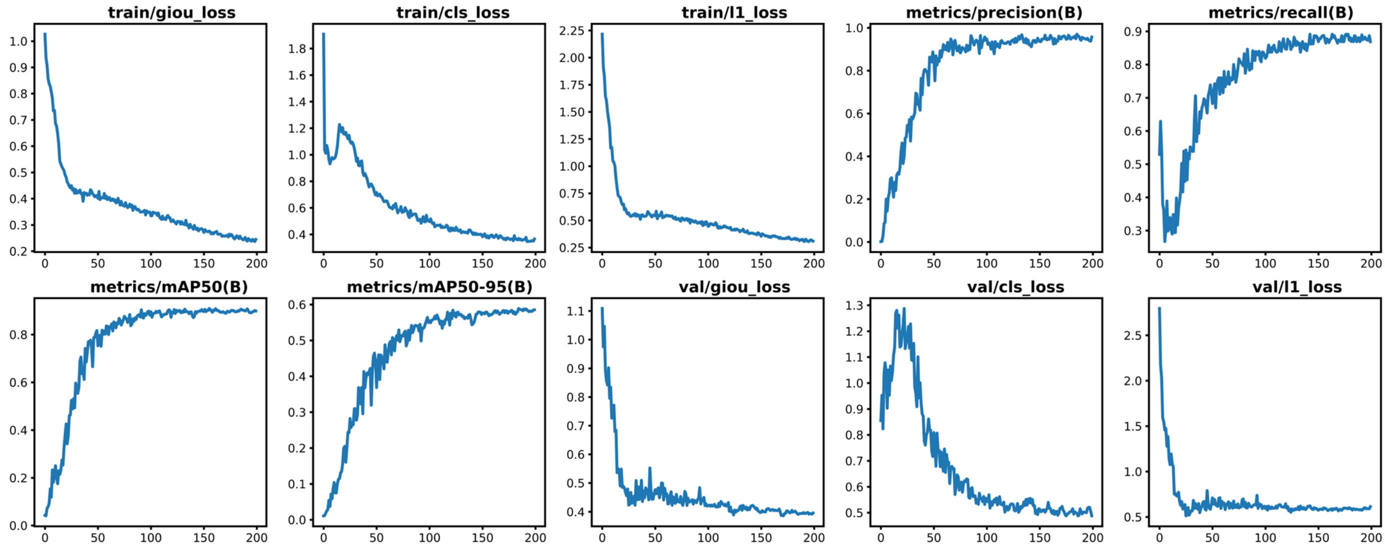


Figure 11. Training metrics over 200 epochs.

4.4. Ablation Results

In this section, we remove important design elements in our designed model for ablation experiments. To amplify the difference and reduce the training time, all models are trained 200 epochs.

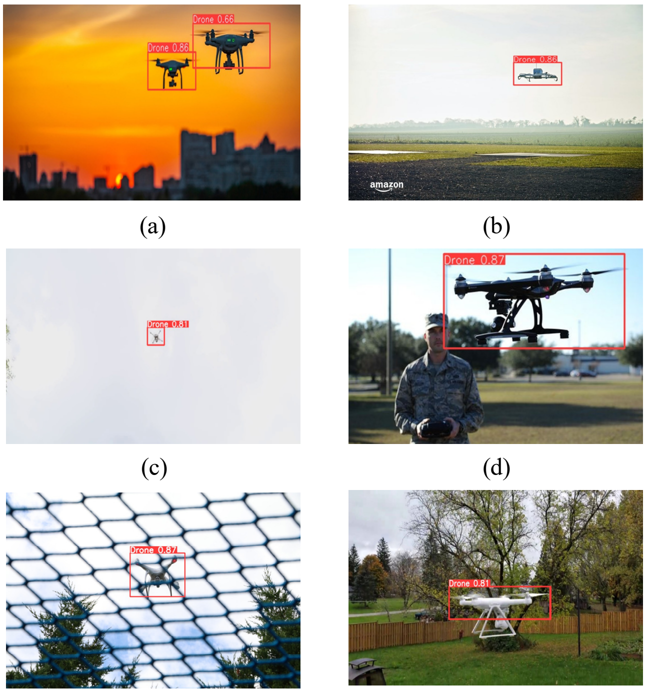
First, we just use the original RT-DETR model for testing; then, we add our Cascaded Group Attention (CGA), Dilation-wise Residual and Dilated Re-param Block Module (DWR-DRB), and Gather-and-Distribute Mechanism (GD) in turn. The results are shown in Table 3. It is shown that after adding Cascaded Group Attention, AP@50 improved by 1.3%, and the FPS reached 52.5 frames per second with almost no increase in GFLOPs. It shows that Cascaded Group Attention reduces the amount of computation with almost no increase in cost, and the AP@50 is improved by 1% when only adding the DWR-DRB module. In the case of adding only Gather-and-Distribute mechanism, the AP@50 is improved by 1.2%. With all modules added, our final model improves the AP@50 by 3% over the original RT-DETR-r18 model and achieves 41.9 frames per second at almost no additional computational cost, which is about 10 frames per second higher than RT-DETR-r18. It shows that our model improves the inference speed while improving the accuracy and can be better applied to UAV detection.

Table 3. Ablation experiments.

RT-DETR	DWR-DRB	CGA	GD	Layers	Parameters	GFLOPs	Recall	AP@50	AP@50:95	FPS
✓				299	19,873,044	56.9	0.941	0.937	0.621	31.7
✓		✓		330	19,703,192	57.0	0.94	0.95	0.629	52.5
✓	✓			328	21,048,084	57.9	0.953	0.947	0.625	29.8
✓			✓	434	22,257,300	59.9	0.92	0.949	0.624	27.9
✓		✓	✓	463	23,432,340	60.9	0.899	0.952	0.635	25.0
✓	✓	✓	✓	494	23,262,488	61.0	0.93	0.956	0.624	41.9

4.5. Detection Results

We have used our model to detect UAV pictures with different scenes and sizes, and the detection results are shown in Figure 12. Figure 12a shows that in the dusk scene, our model can also accurately identify when there is less light and the UAV is dark. Figure 12b shows that the sky color of the background and the color of the UAV are both white, indicating that our model can still achieve accurate recognition when the color of the UAV is similar to the sky background. In Figure 12c, it can be seen that the target UAV is very small, and its color is almost integrated into the background sky. However, our model can still recognize it, indicating that our model also has a good effect on detecting small targets. Figure 12d is a picture of the UAV at a close distance, and the UAV takes up a large portion of the picture, indicating that when using the UAV to detect the UAV, the recognition of a nearby target can achieve a high accuracy rate. Figure 12e,f shows UAV detection under a complex background. It can be seen that in the complex background, the UAV is easier to integrate with the environment, and at the same time, it is more difficult to detect, but our model can still achieve a high accuracy rate, indicating the effectiveness of our model in the face of complex background detection. Figure 13 shows the detection results of more complex scenes, which can be seen in the woods or on the road. Our model can detect drones even when there are occlusions or complex backgrounds.



(a) (b)

(c) (d)

(e) (f)

Figure 12. Detection results on rotor UAV dataset.(**a**) drones at dusk (**b**)drone is similar in color to the sky (**c**) small target drone (**d**) large target drone (**e**) drone in complex background (**f**) drone in a forest scene.

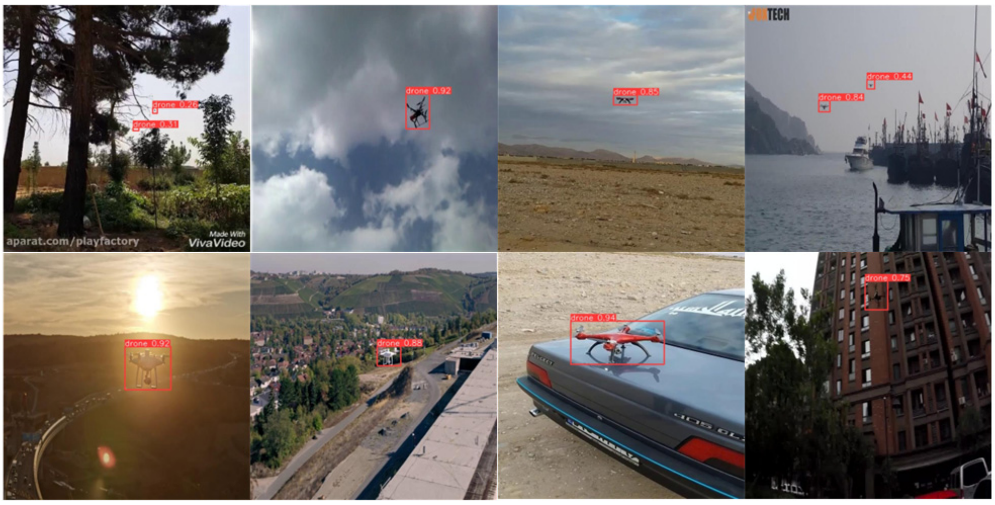


Figure 13. Detection results on military UAV dataset.

5. Conclusions

The appearance of uncrewed aerial vehicles in images or videos is diverse and variable, with their scale, angle, appearance, and other characteristics changing based on distance and viewpoint. This variability greatly increases the complexity and challenges of UAV detection algorithms. In complex backgrounds like the sky, trees, or buildings, UAV detection algorithms must possess excellent object detection capability to exclude interference information. Moreover, UAVs may also face occlusion from other objects, further complicating detection.

Additionally, UAVs differ in morphology and appearance due to varying models, manufacturers, and purposes. It requires detection algorithms to have strong adaptability and accurately identify UAVs in different situations. Real-time detection and tracking of UAVs are often required in applications like military surveillance and border monitoring. As a result, detection algorithms must exhibit efficient computing performance and fast response speeds. However, obtaining a representative UAV image dataset is challenging due to the diverse operating environments and the substantial workload associated with data collection and annotation.

To address these challenges, this paper proposes an improved transformer model called GCD-DETR and conducts extensive experiments on two public datasets. The GCD-DETR model introduces the DWR-DRB module, leveraging the Cascaded Group Attention and Gather-and-Distribute mechanism to strike a balance between efficiency and accuracy. In particular, the DWR-DRB module enhances adaptability to changes in UAV morphology and appearance via a two-step multi-scale context information acquisition method. Cascaded Group Attention assists the model in focusing on UAV targets and eliminating interference information in complex backgrounds. The Gather-and-Distribute mechanism further enhances detection accuracy through global information interaction and fusion. The experimental results demonstrate significant performance improvements in the GCD-DETR model in UAV detection tasks, particularly when dealing with complex backgrounds and occlusions. The successful application of this model offers substantial support for the intelligent development of UAVs, especially in areas such as military surveillance and border monitoring.

However, practical applications of UAV detection technology still face limitations and challenges, including computing resource constraints and real-time requirements. Future research aims to improve the robustness and efficiency of UAV detection technology to address even more complex and variable application scenarios. Additionally, with the continuous development and popularization of UAV technology, the application of UAV detection technology will witness further expansion and development opportunities.

Author Contributions: Methodology, M.Z. and E.K.; Formal analysis, E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (No. 62076028), Subject Research of Beijing Information Science and Technology University (No. 5027023400), Qiyuan Innovation Foundation and sub-themes: (S20210201067) (No. 9072323404).

Data Availability Statement: Data are contained within the article.

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

References

1. Kaleem, Z.; Rehmani, M.H. Amateur drone monitoring: State-of-the-art architectures key enabling technologies and future research directions. *IEEE Wirel. Commun.* **2018**, *25*, 150–159. [CrossRef]
2. Rossiter, A. Military technology and revolutions in warfare: Priming the drone debate. *Def. Secur. Anal.* **2023**, *39*, 253–255. [CrossRef]
3. Emimi, M.; Khaleel, M.; Alkrash, A. The current opportunities and challenges in drone technology. *Int. J. Electr. Eng. Sustain.* **2023**, *1*, 74–89.
4. McFarland, M. Airports Scramble to Handle Drone Incidents. Available online: https://edition.cnn.com/2019/03/05/tech/airports-drones/index.html (accessed on 5 March 2019).
5. Raivi, A.M.; Huda, S.A.; Alam, M.M.; Moh, S. Drone Routing for Drone-Based Delivery Systems: A Review of Trajectory Planning, Charging, and Security. *Sensors* **2023**, *23*, 1463. [CrossRef] [PubMed]
6. Taha, B.; Shoufan, A. Machine learning-based drone detection and classification: State-of-the-art in research. *IEEE Access* **2019**, *7*, 138669–138682. [CrossRef]
7. Ahmad, B.I.; Harman, S.; Godsill, S. A Bayesian track management scheme for improved multi-target tracking and classification in drone surveillance radar. *IET Radar Sonar Navig.* **2024**, *18*, 137–146. [CrossRef]
8. Zhang, H.; Li, T.; Li, Y.; Li, J.; Dobre, O.A.; Wen, Z. RF-based drone classification under complex electromagnetic environments using deep learning. *IEEE Sens. J.* **2023**, *23*, 6099–6108. [CrossRef]
9. Han, Z.; Zhang, C.; Feng, H.; Yue, M.; Quan, K. PFFNET: A Fast Progressive Feature Fusion Network for Detecting Drones in Infrared Images. *Drones* **2023**, *7*, 424. [CrossRef]
10. Valaboju, R.; Harshitha, C.; Kallam, A.R.; Babu, B.S. Drone Detection and Classification using Computer Vision. In Proceedings of the 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 11–13 April 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1320–1328.
11. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
12. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y. 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.
13. 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: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
14. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
15. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
16. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
17. 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.
18. Ashraf, M.W.; Sultani, W.; Shah, M. Dogfight: Detecting dronesfrom drones videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7067–7076.
19. Sangam, T.; Dave, I.R.; Sultani, W.; Shah, M. Transvisdrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 6006–6013.
20. Lv, W.; Xu, S.; Zhao, Y.; Wang, G.; Wei, J.; Cui, C.; Du, Y.; Dang, Q.; Liu, Y. Detrs beat yolos on real-time object detection. *arXiv* **2023**, arXiv:2304.08069.
21. Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; Yuan, Y. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.

22. Wei, H.; Liu, X.; Xu, S.; Dai, Z.; Dai, Y.; Xu, X. DWRSeg: Rethinking Efficient Acquisition of Multi-scale Contextual Information for Real-time Semantic Segmentation. *arXiv* **2022**, arXiv:2212.01173.
23. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. In Proceedings of the 37th Conference on Neural Information Processing Systems, Virtual, 10–16 December 2024; Volume 36.
24. Seidaliyeva, U.; Akhmetov, D.; Ilipbayeva, L.; Matson, E.T. Real-time and accurate drone detection in a video with a static background. *Sensors* **2020**, *20*, 3856. [CrossRef] [PubMed]
25. Sharjeel, A.; Naqvi, S.A.Z.; Ahsan, M. Real time drone detection by moving camera using COROLA and CNN algorithm. *J. Chin. Inst. Eng.* **2021**, *44*, 128–137. [CrossRef]
26. 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] [PubMed]
27. Zhao, Y.; Ju, Z.; Sun, T.; Dong, F.; Li, J.; Yang, R.; Fu, Q.; Lian, C.; Shan, P. TGC-YOLOv5: An Enhanced YOLOv5 Drone Detection Model Based on Transformer, GAM & CA Attention Mechanism. *Drones* **2023**, *7*, 446. [CrossRef]
28. Kim, J.H.; Kim, N.; Won, C.S. High-Speed Drone Detection Based On Yolo-V8. 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. 1–2.
29. Cheng, Q.; Li, X.; Zhu, B.; Shi, Y.; Xie, B. Drone detection method based on MobileViT and CA-PANet. *Electronics* **2023**, *12*, 223. [CrossRef]
30. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional detr for fast training convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3651–3660.
31. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
32. Roh, B.; Shin, J.; Shin, W.; Kim, S. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv* **2021**, arXiv:2111.14330.
33. Chen, Q.; Chen, X.; Wang, J.; Zhang, S.; Yao, K.; Feng, H.; Han, J.; Ding, E.; Zeng, G.; Wang, J. Group detr: Fast detr training with group-wise one-to-many assignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 6633–6642.
34. Zhang, M.; Song, G.; Liu, Y.; Li, H. Decoupled detr: Spatially disentangling localization and classification for improved end-to-end object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 6601–6610.
35. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 11963–11975.
36. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13733–13742.
37. Ding, X.; Zhang, Y.; Ge, Y.; Zhao, S.; Song, L.; Yue, X.; Shan, Y. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. *arXiv* **2023**, arXiv:2311.15599.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.