remote sensing

MDPI

Article

MCG-RTDETR: Multi-Convolution and Context-Guided Network with Cascaded Group Attention for Object Detection in Unmanned Aerial Vehicle Imagery

Chushi Yu and Yoan Shin *

School of Electronic Engineering, Soongsil University, Seoul 06978, Republic of Korea; csyu@soongsil.ac.kr
* Correspondence: yashin@ssu.ac.kr

Abstract: In recent years, object detection in unmanned aerial vehicle (UAV) imagery has been a prominent and crucial task, with advancements in drone and remote sensing technologies. However, detecting targets in UAV images pose challenges such as complex background, severe occlusion, dense small targets, and lighting conditions. Despite the notable progress of object detection algorithms based on deep learning, they still struggle with missed detections and false alarms. In this work, we introduce an MCG-RTDETR approach based on the real-time detection transformer (RT-DETR) with dual and deformable convolution modules, a cascaded group attention module, a context-guided feature fusion structure with context-guided downsampling, and a more flexible prediction head for precise object detection in UAV imagery. Experimental outcomes on the VisDrone2019 dataset illustrate that our approach achieves the highest AP of 29.7% and AP_{50} of 58.2%, surpassing several cutting-edge algorithms. Visual results further validate the model's robustness and capability in complex environments.

Keywords: UAV images; object detection; MCG-RTDETR; real-time detection transformer; deep learning

Citation: Yu, C.; Shin, Y. MCG-RTDETR: Multi-Convolution and Context-Guided Network with Cascaded Group Attention for Object Detection in Unmanned Aerial Vehicle Imagery. *Remote Sens.* **2024**, *16*, 3169. https://doi.org/10.3390/rs16173169

Academic Editors: Wei Li, Haiyong Gan, Heng-Chao Li and Wenshuai Hu

Received: 12 July 2024
Revised: 16 August 2024
Accepted: 19 August 2024
Published: 27 August 2024

1. Introduction

The rapid development of electronic information engineering and communication technologies have propelled unmanned aerial vehicles (UAV) systems to the forefront of international research in remote sensing [1]. Combining drones, the global positioning system (GPS), and remote sensing techniques, researchers can obtain high-quality, low-altitude images through UAV platforms, thereby reducing information loss due to climate and day–night conditions. These drone images are crucial for subsequent military and civil missions, including natural disaster rescue, the Internet of Things (IoT), smart cities, and traffic monitoring. Nonetheless, owing to factors like lighting conditions, shooting angles, and background complexities, the intelligent interpretation of drone data is more comparative and challenging.

Deep learning (DL) is a branch of machine learning methods that simulates the neural network structure of the human brain, employing multi-layer neural networks to automatically learn features and patterns from massive data. Convolutional neural networks (CNNs), a representative algorithm of deep learning, employ feedforward neural network architecture with deep structures and convolution operations, inspired by the visual perception mechanism of biological systems. With the further improvement of computing devices and theory, CNNs have seen rapidly development and are now extensively applied in fields such as computer vision and natural language processing. Taking UAV-based images as an example, computer vision can be subdivided into main tasks such as image classification retrieval [2], object detection [3], image segmentation, image recognition, tracking [4], etc. Among them, object detection plays a connecting role between raw data processing and practical applications, aiming to precisely locate and categorize specific objects within images or videos.

Over recent decades, a multitude of object detection methods have emerged, spanning template-based, feature-based, and DL-based approaches. These DL-based methods not only overcome the inherent limitations of traditional approaches but also leverage rich features and advanced scene understanding, leading to enhanced accuracy and reliability in object positioning. In general, DL-based algorithms typically fall into two categories: two-stage approaches like the faster region-based CNN (Faster R-CNN) [5] and one-stage methods such as the RetinaNet [6] and You Only Look Once (YOLO) series [7,8]. The Faster R-CNN [5] serves as the foundational model upon which many two-stage detectors have been built. In contrast, one-stage detectors directly predict object class probabilities and position coordinates, omitting a separate region proposal stage. In addition to the structural differences, the RetinaNet [6] attempt to introduces focal loss to tackle the imbalance of categories and special emphasize hard examples. YOLO [7], representing a prominent one-stage detection algorithm, excels in speed, especially for small object detection, making it highly impactful and enjoys widespread adoption. Two-stage detectors deliver higher accuracy at the expense of real-time responsiveness, whereas one-stage detectors, despite their advantages in end-to-end performance, tend to exhibit lower accuracy in localizing and recognizing small targets.

The detection transformer (DETR) is a novel algorithm introduced in 2020 by researchers at Facebook AI Research [9]. Unlike the traditional two-stage pipeline, the DETR replaces it with a transformer, providing the advantages of an end-to-end architecture and global context modeling. Leveraging the self-attention mechanisms, transformer-based methods can understand contextual features and their interrelationships well, which has become a recent breakthrough compared to CNN-based detectors. However, the unaddressed challenge of high computational costs associated with the DETR hampers their practical utility, hindering the full exploitation of benefits such as eliminating the non-maximum suppression (NMS) process and other post-processing steps. In addition, the real-time detection transformer (RT-DETR) was proposed to eliminate the inference delay caused by NMS, and it outperforms YOLO-based detection methods of the equivalent scale with regard to accuracy and speed [10]. The creators introduce a newly devised hybrid encoder efficiently to process multi-scale features. Additionally, they propose an intersection over union (IoU)-aware query selection strategy to initialize object or position queries from the encoder. This detector is conducive to practical deployment while satisfying the real-time requirements, because it eliminates unnecessary retraining process and can effectively reduce inference cost.

In recent years, a multitude of diverse models and structures have surfaced to advance performance in semantic segmentation and object detection, focusing on different perspective like lightweight and attention mechanism. The innovation of DualConv, proposed by [11], introduces dual convolutional kernels designed to create a slim and efficient deep neural network. This approach can be integrated into a wide range of CNN architectures through adjustments to their design, resulting in noteworthy reductions in computational costs and parameter counts. Deformable convolutional networks version 2 (DCNv2) enhances object detection and semantic segmentation by introducing deformable convolutions, which dynamically adjust kernel shapes and positions [12]. DCNv2 expands its capability by integrating offset learning across multiple convolutional layers, enabling precise control over feature-level sampling. Within its deformable convolution blocks or modules, every instance undergoes personalized offset adjustments and modulation, adapting to feature characteristics. This flexibility enables the architecture to dynamically reshape spatial distributions and modulate the influence of individual instances. The context-guided network (CGNet) is a lightweight neural network tailored for efficient semantic segmentation tasks, particularly in resource-constrained environments. The context-guided (CG) block, introduced in [13], emulates the human visual system by leveraging contextual relationships to interpret scenes for semantic segmentation. The CG block plays a crucial role in capturing local features, surrounding context, and global context, en-

abling precise boundary detection through refined feature extraction and seamless context fusion, which integrates the information to enhance accuracy.

CNN architectures typically demand significant memory and computational resources, rendering them impractical for embedded systems constrained by hardware limitations. We introduce a new method called MCG-RTDETR built upon RT-DETR, which collaborates multi-convolution and context-guided downsampling with a cascaded group attention module to achieve precise object detection in UAV images, all while maintaining a balance between detection accuracy and computational efficiency. Firstly, we introduce the dual convolutional filter and deformable convolution into the backbone to extract features. This advancement empowers the model to proficiently identify and dissect the intricate details present in small targets. It facilitates the integration of intricate local elements with broader contextual information, thereby seamlessly incorporating detailed local aspects with overarching global structures, thus understanding the image more accurately and comprehensively. Unlike previous research, we adjust the structure of the neck by using context-guided downsampling blocks instead of the traditional neck and detection head for small object detecting, aiming to alleviate the effects of varying target scales. The applicability of our proposed scheme is validated using the VisDrone2019 dataset [14]. Our enhanced model MCG-RTDETR exhibited robust performance in UAV imagery despite the complex remote sensing environment. Meanwhile, ours illustrates an improvement of average precision (*AP*) of approximately 4.7% to 6.8% with the original RT-DETR model.

The main contributions of this work are as follows:

- We integrated dual convolution and deformable convolutions into the backbone part of original RT-DETR. These convolution operations better capture complex feature information and geometric deformations in various scenarios and object sizes.
- We incorporated a cascaded group attention module into the encoder part to focus on critical feature regions while suppressing non-relevant background information. We replaced the traditional downsampling operation with context-guided downsampling to preserve contextual information of the targets.
- To tackle challenges posed by varying scales and dense scenes, we specifically optimized the structure of the neck to fuse features better, and the detection heads include adjusting output layers suitable for small objects.
- Through the aforementioned enhancements, we conducted rigorous experimental validations on the VisDrone2019 dataset. The results demonstrate significant improvements in both quantitative and qualitative evaluation metrics. These performance improvements not only affirm the efficacy of our method but also showcase its potential in practical scenarios.

The remaining structure of this work is as follows: Related work in the existing literature is summarized in Section 2. In Section 3, we describe our approach's architecture and working mechanism comprehensively. Section 4 outlines the experimental setup and implementation details and provides ablation and comparison experiments using the VisDrone2019 dataset to verify our approach. Section 5 address a discussion of our proposed scheme and explores future avenues for research. Finally, Section 6 offers this study's conclusion.

2. Related Work

2.1. General Object Detection

Recently, notable advancements have emerged in object recognition and detection algorithms, which are driven by the swift progress in artificial intelligence techniques. Considering the detection and recognition processes, detectors can be broadly categorized into two main types: two-stage and one-stage detectors. Classical examples of two-stages include the Faster R-CNN [5] and Cascade R-CNN [15], while the SSD [16] and the YOLO [7] series exemplify one-stages algorithms. Two-stage detection algorithms handle the object's classification and predicted bounding boxes' regression as separate steps, whereas one-stage detectors execute these tasks simultaneously, offering higher efficiency and lower computational

requirements. The architecture of an object detector generally comprises three key components: the Backbone network for capturing features, the Neck for fusing features, and the Head for managing classification and regression tasks. For instance, in the Faster R-CNN, the residual neural network (ResNet) [17] and the visual geometry group (VGG) [18] network are usually used as backbone networks for initial feature extraction. Faster R-CNN then utilizes the region proposal network (RPN) to generate region of interest (ROI) proposals using predefined bounding boxes or anchors. These feature maps are resized, classified, and normalized, and detection is ultimately performed via the NMS operation. Despite their high detection accuracy, two-stage detectors are burdened with significant computational overhead and operation latency, which can hinder their applications in scenarios requiring immediate responsiveness. The training process demands substantial memory, especially with high-resolution images, and the accuracy of region proposals from the RPN directly impacts performance. These limitations have spurred the development of faster, more efficient one-stage detectors such as YOLOv5 [19] and YOLOv8 [20]. The neck structure bridges the backbone and head parts, refining and fusing features through well-established pathways, including bottom-up and top-down approaches like the feature pyramid network (FPN) [21] and path aggregation network (PANet) [22]. Yu et al. [23] introduced the weighted bidirectional FPN (BiFPN) [24] into the YOLOv5 model, enhancing feature fusion and effectively addressing the problem of varying ship scales in synthetic aperture radar (SAR) datasets. In addition, deformable ConvNets v2 [12] introduced deformable operations to convolutional neural networks, including deformable ROI pooling and deformable receptive fields, to better adapt to irregular object shapes and poses. These advancements have significantly boosted capabilities in domains like object localization and semantic segmentation.

2.2. Object Detection in UAV Images

When developing DL-based object detection approaches tailored for UAV scenarios, traditional computer vision techniques are predominantly utilized. However, small UAV platforms and specific imaging conditions present unique challenges, including diverse perspectives, complex backgrounds, varying scales and orientations, and difficulties in detecting small objects. Researchers have focused on addressing these challenges through sub-tasks such as static or dynamic object detection, detection in images or videos, and single or multiple object. Due to varying altitudes and sizes of objects on the ground in UAV imagery, some approaches have tackled scale diversity by employing multi-scale features, as seen in [16,25,26]. Others utilized dilated or deformable convolution kernels to handle this issue, such as those described in [27–29]. The flight altitude of drones inevitably results in diverse scales, small target sizes, and dense object arrangements, limiting the extractable feature information. Various approaches have emerged to optimize small object detection, encompassing the Cascade network [28], FSSSD [30], UAV-YOLO [31], depthwise-separable attention-guided network (DAGN) [32], and HRDNet [33]. In the realm of UAV remote sensing, achieving real-time processing and the interpretation of high-quality images is crucial. YOLO-based models generally meet the needs of detecting objects immediately. SlimYOLOv3 [34] is a streamlined version of YOLOv3 that achieves real-time object detection by optimizing the trade-off among parameters, memory consumption, and inference speed. Furthermore, an enhanced YOLOv8-based UAV object detection method proposed by Wang et al. [35], introducing a small target detection structure (STC) to enable the integration of deep and shallow features and obtaining better semantic capture and detection accuracy.

3. Proposed Method

3.1. Overall Framework

Figure 1 illustrates the framework of the MCG-RTDETR algorithm proposed in this study. Our approach builds upon the RT-DETR, one of cutting-edge end-to-end object detectors recognized for balancing speed and accuracy in various tasks. We selected the RT-DETR-r18 as the baseline to develop our network. Additionally, there are several versions

of the RT-DETR model, including the RT-DETR-r34, RT-DETR-r50, and RT-DETR-x. As depicted in Figure 1, the architecture comprises a backbone part, an efficient hybrid encoder, a decoder, and prediction heads.

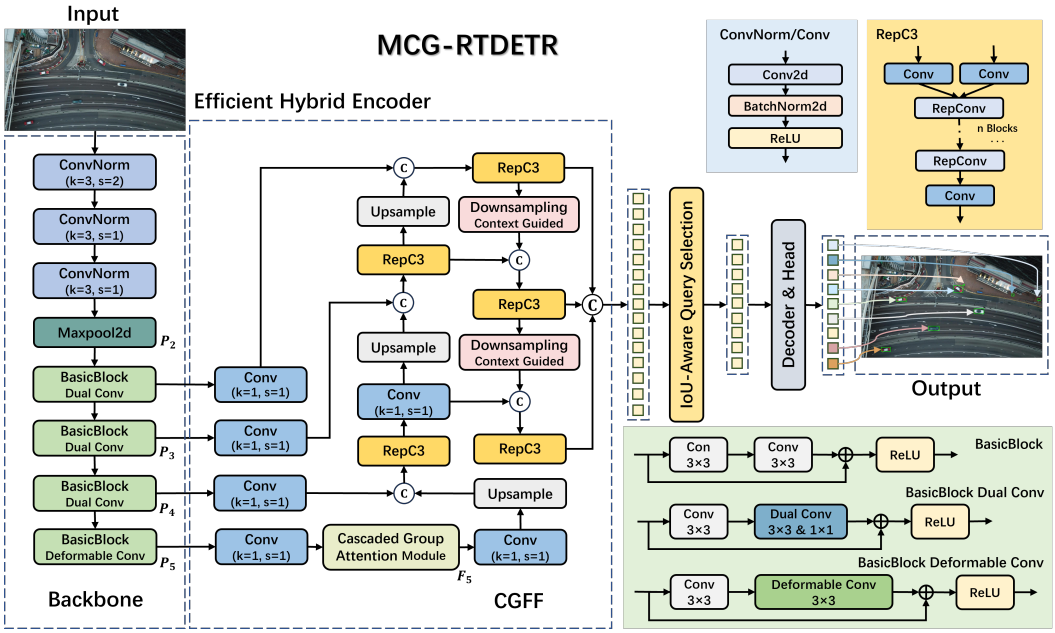

Figure 1. The architecture of the proposed MCG-RTDETR.

Firstly, the backbone network captures key information from input UAV images, producing multi-scale feature maps from the last four stages $\{P_2, P_3, P_4, P_5\}$. Secondly, these feature maps are fused through the efficient hybrid encoder that combines intra-scale feature interaction with cross-scale feature fusion modules. Subsequently, a fixed number of image features is selected by the *IoU*-aware query selection mechanism to act as starting queries for this mentioned decoder. Utilizing auxiliary heads, the decoder progressively refines these queries, producing bounding boxes with confidence scores. The central innovations of our approach include the cascaded group attention module and context-guided downsampling, which are designed to enhance detection precision and maintain contextual integrity, respectively. The innovative improvements will be detailed in the following sections.

3.2. Improvement of Feature Extractor

We employ the dual convolutional filter in the backbone instead the original operation. The DualConv integrates convolution kernels of 3×3 and 1×1, allowing for the concurrent processing of input channels and efficient filter arrangement using grouped convolution, as shown in Figure 2. In this setup, M represents the input channel count, N denotes the output channels and convolution filters, and G represents the group count within dual convolution. Next, we consider dividing the N filters into G groups, where each group processes this complete feature map. The 3×3 and 1×1 convolutional kernels concurrently handle $\frac{M}{G}$ input channels, while the remaining $(M - \frac{M}{G})$ channels exclusively by a convolution kernel of 1×1. The summed results, as represented by the $\oplus$ sign in Figure 2, integrate the outputs of these processes. The grouped architecture can enhance the sparsity of the block diagonal, facilitating the structured learning of a highly correlated filter without requiring a shifted arrangement. The DualConv method reduces parameters within the backbone by utilizing grouped convolutions, fostering information exchange between different layers. It preserves input data and enables optimal cross-channel transmission using M times 1×1 convolution. Consequently, the channel shuffle operations are unnecessary for constructing the DualConv filter.

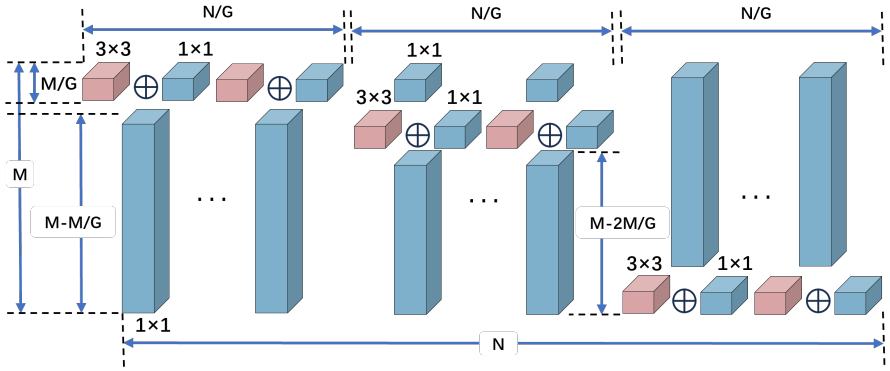

Figure 2. The structure of the dual convolutional filter. *M* is the input channel count, *N* denotes the number of output channels and convolution filters, and *G* is the group count within dual convolution.

The inclusion of deformable convolution layers from DCNv2 [12] at the backbone's final stage enables better semantic representation and localization as shown in Figure 3. Conventional convolution-based networks struggle with geometric transformations, as their convolution and pooling layers are too rigid, which hinders their ability to perform well in various viewpoint and object sizes from drone images. To enhance the adaptability of extracted features, we incorporated a deformable convolution layer into ResNet18, as it dynamically adjusts the receptive field. The original convolutional kernel has a regular rectangular shape, and the deformable convolutional kernel adds offset to each sample point to make an irregular arrangement. The offsets are generated by applying another convolutional layer to the consistent input map.

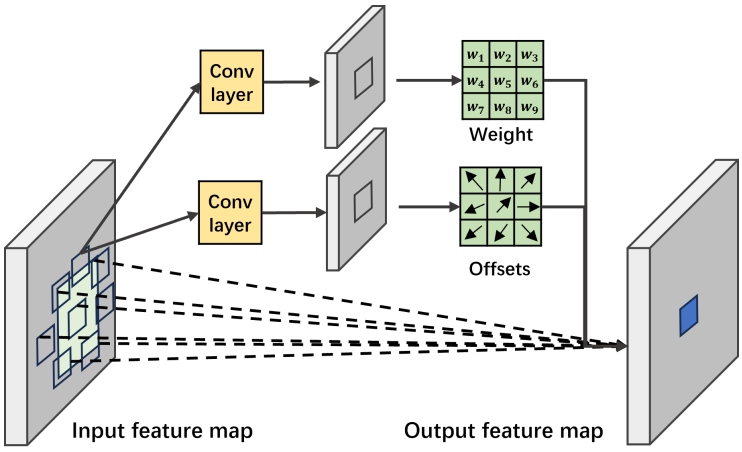

Figure 3. The structure of the 3 × 3 deformable convolution.

3.3. Improvement of Efficient Hybrid Encoder

In this proposed MCG-RTDETR, P_2, P_3, P_4, and P_5 from the backbone part are fed into the modified encoder. According to the [10], the attention-based intra-scale feature interaction (AIFI) and CNN-based cross-scale feature fusion (CCFF) are two essential blocks of the original encoder. AIFI employs a multi-head attention structure [36], which increases computational complexity and model parameters, potentially degrading performance. To address this, we integrated a cascaded group attention module (CGAM) instead of an AIFI module, and we applied it to P_5. The CGAM is a pivotal component to focus on relevant feature regions while filtering out irrelevant background noise, which is particularly beneficial for UAV imagery, where objects are often small and located in cluttered environments. The CGAM dynamically adjusts weights for feature maps by evaluating the relevance of various locations in the input images, thereby enhancing the model's understand of image features and improving detection performance [37]. In the CGAM, the input image is segmented into groups, with each potentially encapsulating distinct semantic information

and neighboring pixels. The input sequence first undergoes linear mappings to generate queries (Q), keys (K), and values (V). The CGAM applies grouped attention and computes attention weights within every set using Q, K, and V to produce the group's attention output. This process can be cascaded by concatenating or weight summing the outputs from multiple groups. The cascaded outputs are further linearly transformed to yield the final output of the CGAM. This progressive focusing process enhances features' refinement across various levels, thereby boosting the model's capability to perceive and accurately depict features, as illustrated in Figure 4. Moreover, we proposed a context-guided feature fusion (CGFF) module as an extension of the CCFF fusion block, replacing traditional downsampling with context-guided downsampling to preserve the contextual information of targets. The original RT-DETR uses a 3×3 convolution operation with a stride of 2, followed by batch normalization (BN) and a sigmoid linear unit activation function. The CGFF block effectively harmonizes the incorporation of global contextual information with local details. The context-guided downsampling block is another essential innovation in our approach and preserves critical contextual information, enabling the model to maintain a high level of detection accuracy even in the presence of scale variations and dense object scenes. The block comprises four main components, as shown in Figure 5: one local feature extractor, one surrounding feature extractor, one joint feature extractor, and one global feature extractor. Among these, the local feature extractor employs 3×3 convolution layer to capture local features from neighboring pixels. The surrounding context feature extractor uses dilated convolution with a 3×3 kernel and a dilation rate of 2 to enlarge the receptive field and capture contextual features. The joint feature extractor concatenates outputs from the previous stages, applying a pair of BN and parametric rectified linear unit (PReLU) operations. At the end of this module, there is a global feature extractor comprising a global average pooling layer with two fully connected layers, focusing on both spatial and channel aspects.

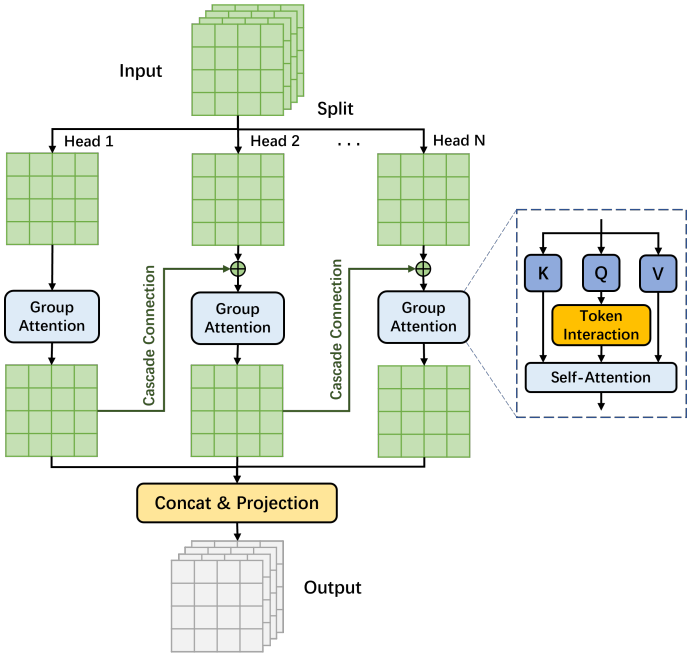

Figure 4. Diagram of the cascaded group attention module.

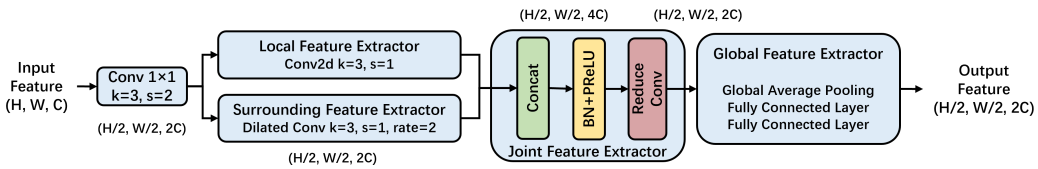

Figure 5. Diagram of the context-guided downsampling block.

The following formulas express the overall steps of the neck part of the proposed MCG-RTDETR:

$$Q = K = V = Flatten(P_5), \tag{1}$$

$$F_5 = Reshape(CGAM(Q, K, V)), \tag{2}$$

$$Output = CGFF(P_2, P_3, P_4, F_5)), \tag{3}$$

where $CGAM$ represents the cascaded group attention module, and *Reshape* denotes the operation of restoring flatten features to the identical shape as P_5.

3.4. Predict Head

To tackle the difficulties presented by diversity scales and dense or complex scenes, we optimized the detection head by adjusting output layers suitable for small objects and tweaking relevant parameters, as depicted in Figure 6. Compared to the basic detection head, the P2 detection head adds a head for small target detecting. Combining the advantages of both the original head and P2 head of the RT-DETR method, we devised our detection head to better integrate high-resolution and low-level feature maps, resulting in improved sensitivity to targets, as shown in the Figure 6c. We effectively captured the multi-scale attributes of the objects through the combination of these feature maps with different scales from the previous extractor and utilizing semantic information.

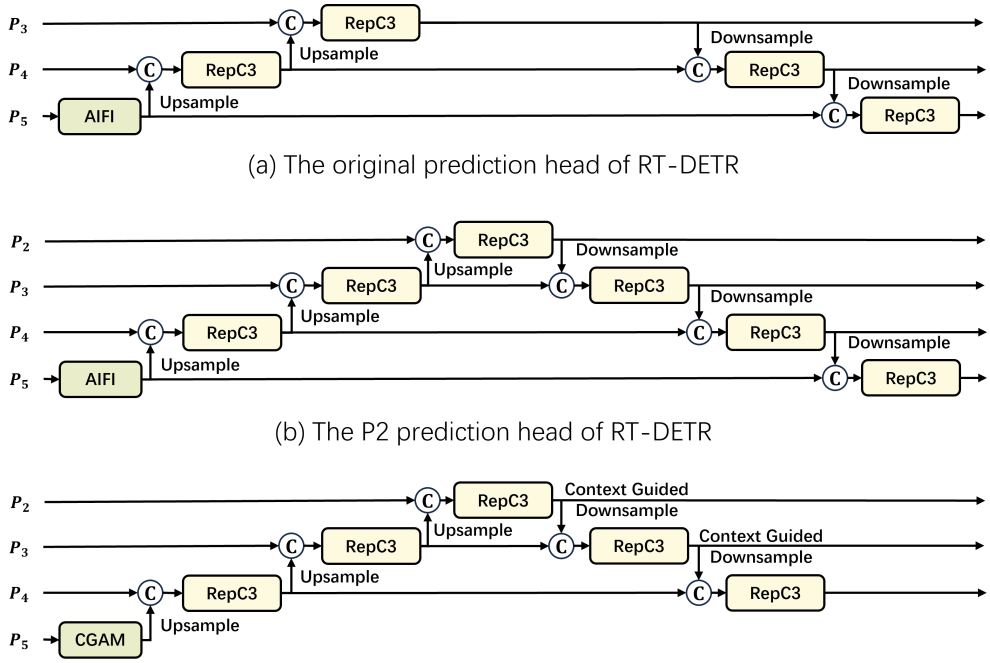

Figure 6. The diagram of prediction heads.

Our goal in modifying the prediction head is to enhance the precision and resilience of detecting targets. This approach has demonstrated its effectiveness in boosting detection performance, enabling the model to perceive objects spanning various scales, categories, and shapes.

3.5. IoU-Aware Query Selection

In detection task, objects or targets are localized by predicting bounding boxes. The intersection over union (IoU) measures the ratio of the intersection area to the union area between the ground truth and the predicted bounding boxes [8], and it can be calculated as follows:

$$IoU = \left| \frac{B \cap B^{gt}}{B \cup B^{gt}} \right|, \tag{4}$$

where B and B^{gt} are the predicted and the ground truth bounding boxes, respectively.

The IoU-aware query selection mechanism involves training the model to assign low classification scores to features with low IoU values and high classification scores to features with high IoU values [10]. Consequently, these predicted bounding boxes that correspond to the encoder features, as determined by the model based on classification scores, exhibit both high IoU and classification scores. The incorporation of the IoU score into the classification branch's objective function ensures that both the classification and localization of positive samples are consistent. We redefine the detector's optimization function as follows:

$$L(\hat{y}, y) = L_{bbox}(\hat{b}, b) + L_{cls}(\hat{c}, \hat{b}, y, b) = L_{bbox}(\hat{b}, b) + L_{cls}(\hat{c}, c, IoU), \tag{5}$$

where $\hat{y}$ and y represent the predicted and ground truth, respectively. Specifically, $\hat{y} = \{\hat{c}, \hat{b}\}$, and $y = \{c, b\}$, where c and b denote the categories and bounding boxes.

4. Experiments and Results

4.1. Dataset and Implementation Details

The VisDrone2019 dataset is an authoritative resource in the international drone vision community, and it features diverse multi-scene and multi-task shooting captured by various drones across 14 cities in China, environments (urban and rural), sparse or dense scenes, weather conditions (cloudy and sunny), and lighting conditions (day and night). The dataset contains 10 classes, including pedestrians, people, bicycles, cars, vans, trucks, tricycles, awning tricycles, buses, and motors [14]. For our experiments, we partitioned the images into three sets: a training set with 6471 samples, a validation set with 548 samples, and a test set with 1610 samples, following the original division protocol within the VisDrone 2019 challenge [14]. We considered multiple IoU thresholds by using the COCO metrics [38] to evaluate at diverse levels of positioning accuracy. Illustrative samples from the VisDrone2019 dataset are depicted in Figure 7.

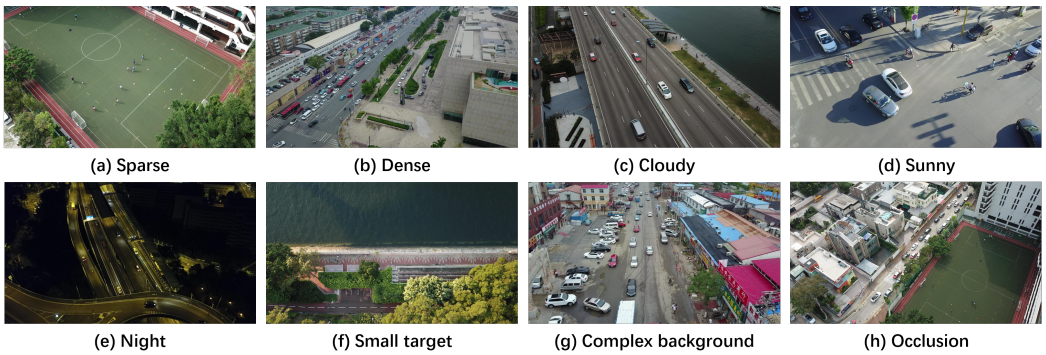

Figure 7. Illustrative samples of the VisDrone2019 dataset.

Our proposed MCG-RTDETR was implemented on the Pytorch 2.1.0 platform. During both the training and inference phases, the size of the inputs was fixed as 640 × 640. The training epoch was set to 300 using the Adam with decoupled weight decay (AdamW) optimizer, with a weight decay of 0.0001 and momentum of 0.9. The batch size remained constant at four, and the initial learning rate was 0.0001. Each experimental trial was performed using an NVIDIA RTX 4080 graphics processing unit (GPU). In addition, the state-of-the-art methods were trained and validated under the same settings as each original paper using MMDetection [39].

4.2. Evaluation Metrics

The precision, recall, and mean average precision (mAP) can gauge the performance of detectors. Quantitatively assessing the effectiveness of algorithms involves using the frames per second (FPS) rate to evaluate its speed. Precision quantifies the proportion of correctly identified positive among all samples predicted as positive. Meanwhile, recall measures the ratio of actual number of positive samples in the predict sample to the number of samples predicted. Here are the definitions of precision and recall:

$$Precision = \frac{TP}{TP + FP}, \tag{6}$$

$$Recall = \frac{TP}{TP + FN}, \tag{7}$$

where TP represents the detector correctly detecting annotated objects, FP represents the detector incorrectly predicting background regions as annotated object, and FN represents the detector incorrectly predicting annotated as background regions.

The average precision quantifies the model's accuracy in correctly detecting targets by averaging across various confidence thresholds. The mAP offer a holistic evaluation of the precision–recall trade-off by averaging the AP scores across all classes. The definition of AP is given as follows:

$$AP = \int_0^1 P(R)dR, \tag{8}$$

Here is the definition of the mAP:

$$mAP = \frac{1}{N}\sum_{n=1}^{N} AP_n, \tag{9}$$

where N denotes the categories number, and AP_n denotes the AP of class n.

The FPS is given by

$$FPS = \frac{s}{T}, \tag{10}$$

where s is the count of samples, and T is the required processing time.

4.3. Ablation Experiments

The ablation experiments evaluated the MCG-RTDETR framework on the VisDrone2019 dataset, with RT-DETR-r18 serving as the baseline network. Several modifications were implemented to enhance the model's effectiveness. Backbone improvements: The original BasicBlock in the backbone part was replaced with the DualConv (BasicBlock with dual convolution filter), and deformable convolution was used to enhance feature representation. Neck modification: The cascaded group attention module (CGAM) was utilized in place of the AIFI module in the latest stage of the backbone. And the neck part was further modified with the introduction of the the context-guided downsampling(CGD) replacing the traditional convolution downsampling operation. Redesigned prediction head: The prediction head was redesigned as P3 to bolster the target perception across diverse scales, complex scenes, shapes, and categories by comparing with the original head and P2.

Tables 1 and 2 present detailed performance metrics, showcasing improvements across different model configurations. Notably, the outcomes highlight that every modification led to ameliorated evaluation metrics. The baseline RT-DETR-r18 achieved 25.0% AP, while the proposed MCG-RTDETR configuration (RT-DETR-DualConv-DeConv-CGD-P3-CGAM) reached 29.7% AP, demonstrating systematic improvements through each enhancement stage. According to the COCO metric evaluation criteria, the proposed MCG-RTDETR showed marked improvements across the AP_{50} (58.2%), AP_{75} (26.3%), and AP for small objects (26.2%), medium objects (51.0%), and large objects (73.5%). The baseline RT-DETR-r18 achieved 34.4% AR_{100}, while ours reached 39.2% AR_{100}, indicating enhanced recall capabilities. Our method improved the AR across AR_1 (3.7%), AR_{10} (21.7%), and AR for

small (36.3%), medium (59.4%), and large objects (80.4%). For instance, the overall AP improved by 4.7% and the AR_{100} by 4.8%. Compared to the baseline, the AP_s increased by 4.5% and the AR_s increased by 4.8% in the MCG-RTDETR approach. These changes particularly enhanced the capabilities to identify small and complex targets in UAV images. A 6.8% increase in the AP_l and a 7.8% increase in the AR_l highlight detection improvements in large objects. By enhancing the backbone, feature extraction modules, fusion method, and prediction heads, the MCG-RTDETR approach significantly boosted the AP and AR metrics on the VisDrone2019 dataset, proving its effectiveness in real scenarios. In addition, Tables 1 and 2 list the giga floating point operations per second (GFLOPs) and the number of model parameters (M) of different methods, which help to better understand the contribution of each module. We found that DualConv played a significant role in reducing the computational complexity.

Table 1. Performance comparison on VisDrone2019 dataset according to average precision (%).

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	$GFLOPs$
RT-DETR-r18	25.0	52.4	20.2	21.7	46.1	66.7	57.0
RT-DETR-DualConv	24.8	52.4	19.7	21.8	44.8	69.5	47.3
RT-DETR-DualConv-CGD	24.6	51.5	19.8	21.2	46.0	70.6	52.1
RT-DETR-DualConv-P2	27.7	55.1	23.9	24.5	47.9	64.5	68.6
RT-DETR-DualConv-CGD-P2	28.1	55.7	24.4	24.9	48.7	69.5	75.1
RT-DETR-DualConv-CGD-P3	28.9	57.1	25.3	25.7	49.3	69.0	90.8
RT-DETR-DualConv-DeConv-CGD-P2	28.0	55.5	24.4	24.7	48.0	66.2	73.9
RT-DETR-DualConv-DeConv-CGD-P2-CGAM	28.8	57.1	25.0	25.3	50.4	68.3	74.0
RT-DETR-DualConv-DeConv-CGD-P3	29.1	57.5	25.4	25.8	49.8	70.3	89.6
RT-DETR-DualConv-DeConv-CGD-P3-CGAM	29.7	58.2	26.3	26.2	51.0	73.5	89.7

Table 2. Performance comparison on VisDrone2019 dataset according to average recall (%).

Methods	AR_1	AR_{10}	AR_{100}	AR_s	AR_m	AR_l	$Params$
RT-DETR-r18	3.5	19.4	34.4	31.5	54.5	72.6	19.88
RT-DETR-DualConv	3.4	19.2	34.4	31.6	53.5	77.8	15.88
RT-DETR-DualConv-CGD	3.5	19.3	33.9	31.0	54.4	78.3	18.33
RT-DETR-DualConv-P2	3.7	20.7	37.9	35.2	56.3	72.6	14.60
RT-DETR-DualConv-CGD-P2	3.6	20.9	38.2	35.5	56.9	77.4	17.20
RT-DETR-DualConv-CGD-P3	3.7	21.4	38.8	36.0	57.7	75.7	19.55
RT-DETR-DualConv-DeConv-CGD-P2	3.6	20.8	38.2	35.7	55.8	73.5	20.46
RT-DETR-DualConv-DeConv-CGD-P2-CGAM	3.6	21.1	39.1	36.3	58.8	73.9	20.29
RT-DETR-DualConv-DeConv-CGD-P3	3.7	21.5	38.6	35.8	58.0	77.0	22.81
RT-DETR-DualConv-DeConv-CGD-P3-CGAM	3.7	21.7	39.2	36.3	59.4	80.4	22.64

4.4. Comparisons of Performance

The capability of our MCG-RTDETR was compared with several state-of-the-art object detectors on the VisDrone2019 dataset. Table 3 lists the qualitative results of the MCG-RTDETR with Faster R-CNN [5], RetinaNet [6], Cascade R-CNN [15], GFL [40], ATSS-dyhead [41,42], TOOD [43], RTMDET-tiny [44], YOLOX-tiny [45], YOLOv5 [19], YOLOv8 [20], RT-DETR-r18 [10], and RT-DETR-r50 [10]. As shown in Table 3, the MCG-RTDETR achieved the highest AP (29.7%) among all comparison models, significantly outperforming the baseline RT-DETR by 4.7% and the next best model, TOOD (26.3%), by 3.4%. This demonstrates the effectiveness of our enhancements in improving the overall detection accuracy. The MCG-RTDETR showed substantial improvements in the AP_{50} (58.2%) and AP_{75} (26.3%) compared to the baseline RT-DETR-r18, with increases of 5.8% and 6.1%, respectively. These gains indicate better precision at higher IoU thresholds, reflecting improved localization accuracy. The substantial increase in AP_{50} highlights the model's robust efficiency in correct target detection with a 50% IoU threshold.

Table 3. Comparison with state-of-the-art models on the VisDrone2019 dataset.

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	$GFLOPs$	$Params(M)$	$FPS(s)$
Faster R-CNN [5]	24.3	39.6	25.9	15.4	36.4	45.0	208	41.39	38.2
RetinaNet [6]	17.3	29.1	17.9	8.1	29.4	35.2	210	36.52	36.1
Cascade R-CNN [15]	25.1	39.8	26.7	15.7	37.6	46.3	236	69.29	13.7
GFL [40]	24.7	39.8	25.6	15.0	37.1	47.4	206	32.28	36.7
ATSS-dyhead [41,42]	26.3	41.5	27.7	16.2	40.1	55.7	110	38.91	24.7
TOOD [43]	26.3	41.9	27.5	16.8	38.5	49.0	199	32.04	34.7
RTMDET-tiny [44]	19.9	33.2	20.2	10.0	31.7	42.9	8.03	4.88	90.5
YOLOX-tiny [45]	18.9	34.5	18.3	11.9	27.7	29.6	7.58	5.04	235.7
YOLOv5 [19]	19.1	32.9	19.0	9.9	30.4	38.6	7.1	2.50	235.5
YOLOv8 [20]	19.7	33.7	19.6	10.3	31.1	38.7	8.1	3.01	223.0
RT-DETR-r18 [10]	25.0	52.4	20.2	21.7	46.1	66.7	57.0	19.88	125.3
RT-DETR-r50 [10]	25.6	54.6	19.7	22.6	45.5	67.0	129.6	41.97	75.3
MCG-RTDETR	29.7	58.2	26.3	26.2	51.0	73.5	89.7	22.64	84.1

The MCG-RTDETR excelled across all object scales, with significant improvements: an AP_s of 26.2%, which is 4.5 percentage points higher than RT-DETR-r18; an AP_m of 51.0%, showing an improvement of 4.9 percentage points; and an AP_l of 73.5%, marking an increase of 6.8 percentage points. These improvements demonstrate the MCG-RTDETR's enhanced capability to detect objects of varying sizes, especially large objects, where it showed the most significant gain. Moreover, MCG-RTDETR balanced the performance and efficiency well. The GFLOPs of ours came out to 89.7 M, which is higher than some lightweight models but lower than more computationally intensive models like the Cascade R-CNN (236 GFLOPs). This indicates that a moderate computational overhead for the performance was achieved. And the parameters count came out to 22.64 M, which is a moderate model size that supports efficient training and deployment. The FPS came out to 84.1, providing a good trade-off between accuracy and speed, rendering it well-suited in real-time scenarios. It offers a competitive inference speed compared to other high-performance models, ensuring that it can be used effectively in practical scenarios. The RT-DETR-r50 and RT-DETR-r18 yielded lower AP scores (25.6% and 25.0%, respectively) compared to the MCG-RTDETR, demonstrating that the proposed enhancements significantly boost detection performance. The improvements in the AP, AP_{50}, and AP_{75} for the MCG-RTDETR indicate that our modifications to the backbone, feature extraction modules, and prediction heads contributed to superior detection capabilities.

Our study on the VisDrone2019 dataset includes both quantitative metrics and qualitative assessments, enriching our analysis with visual samples. Figure 8 illustrates exemplary results visualizing the detection capability of ours, MCG-RTDETR, alongside these state-of-the-art detectors. These comparisons highlight MCG-RTDETR's effectiveness in challenging scenarios such as scenes with occlusion and complex environmental factors, vertical shooting angle during daylight, cloudy with intricate background, very small targets, low-light and night scenes, and dynamic objects like vehicles at night (labeled (a–f) in Figure 8). To accommodate the detail and clarity required, these images are presented across multiple pages. The first column is the name of each method, including Ground Truth, Faster R-CNN [5], RetinaNet [6], Cascade R-CNN [15], GFL [40], ATSS-dyhead [41,42], TOOD [43], RTMDET-tiny [44], YOLOX-tiny [45], YOLOv5 [19], YOLOv8 [20], RT-DETR-r18 [10], RT-DETR-r50 [10], and our proposed MCG-RTDETR.

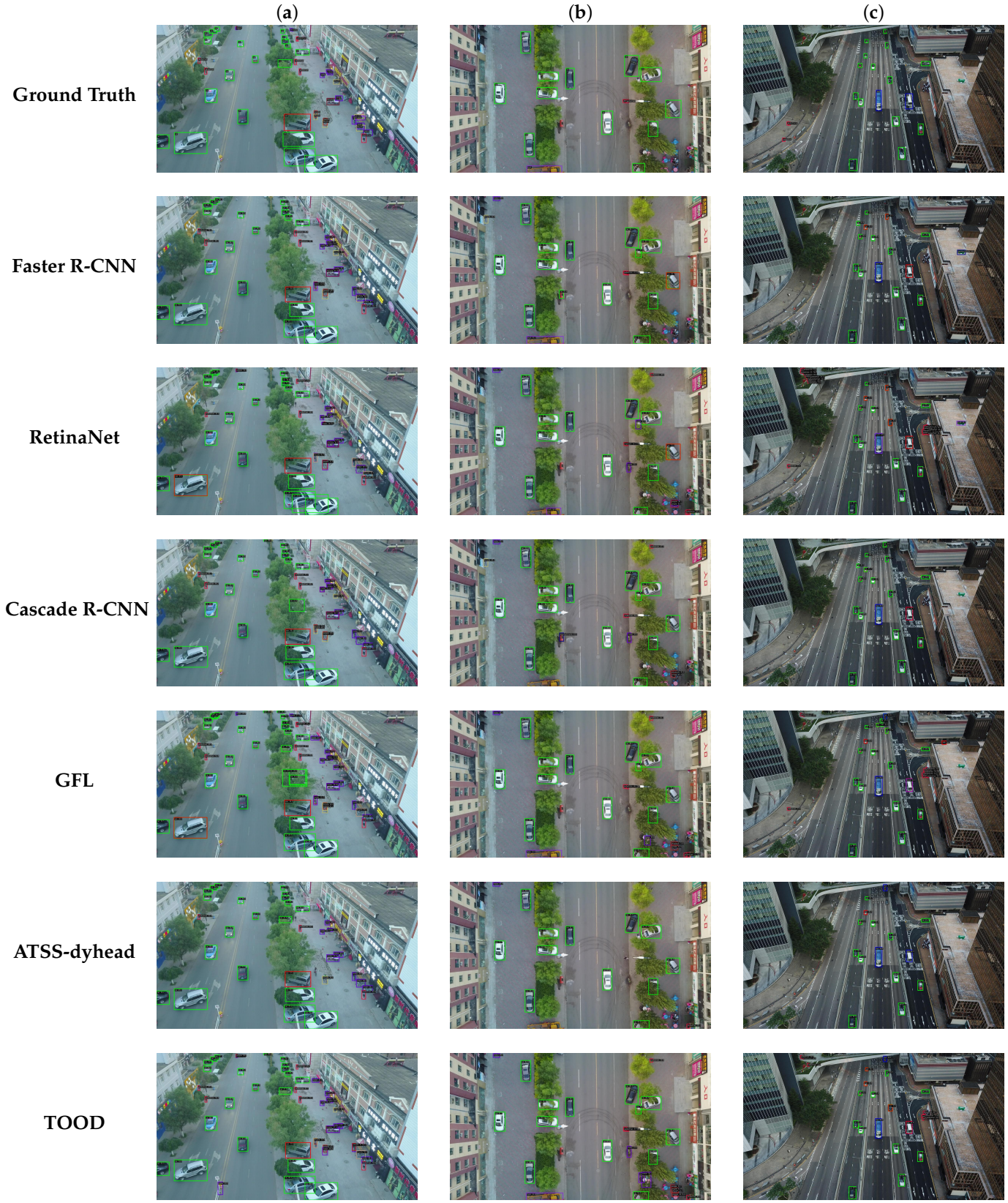

Figure 8. *Cont.*

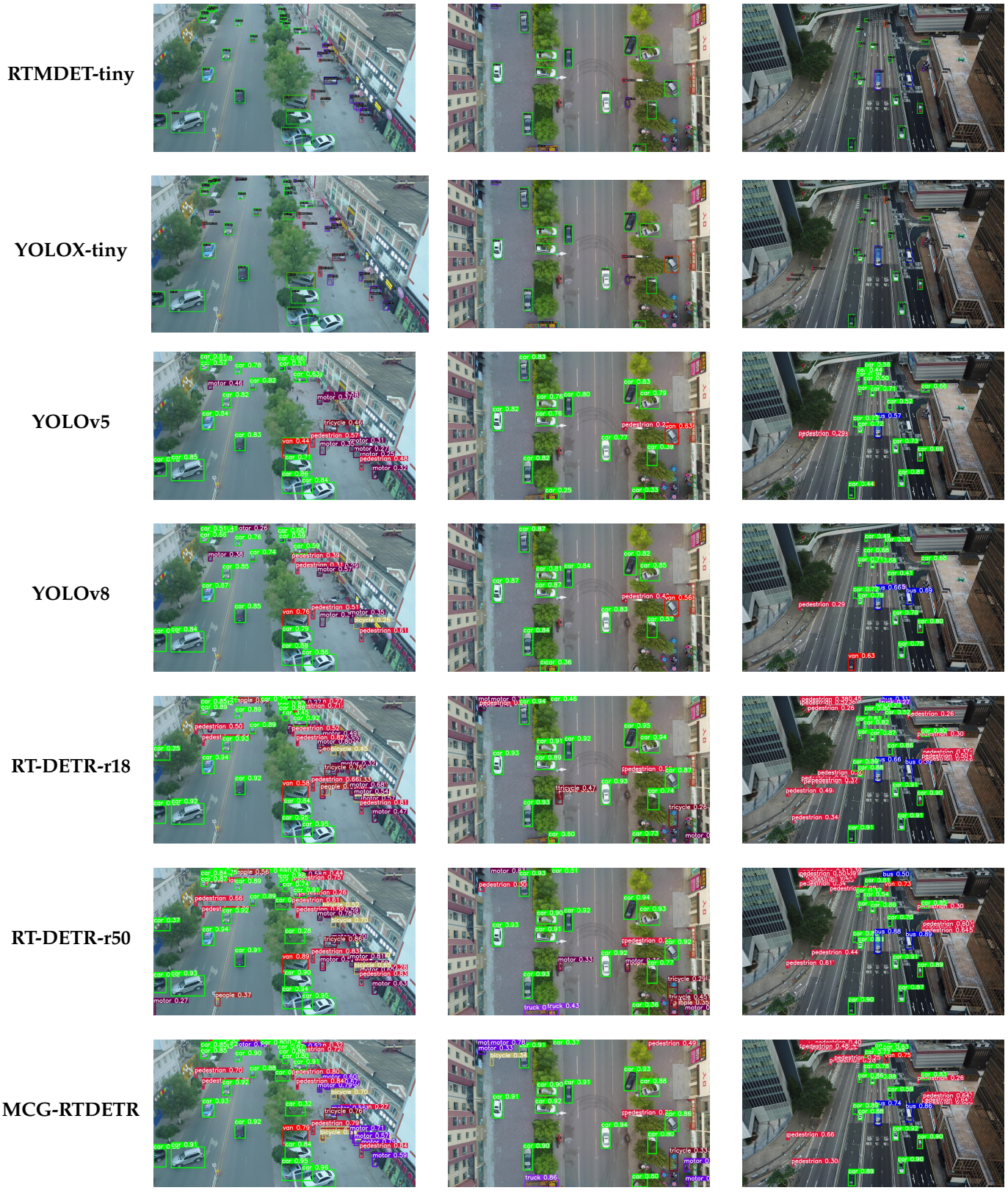

Figure 8. *Cont.*

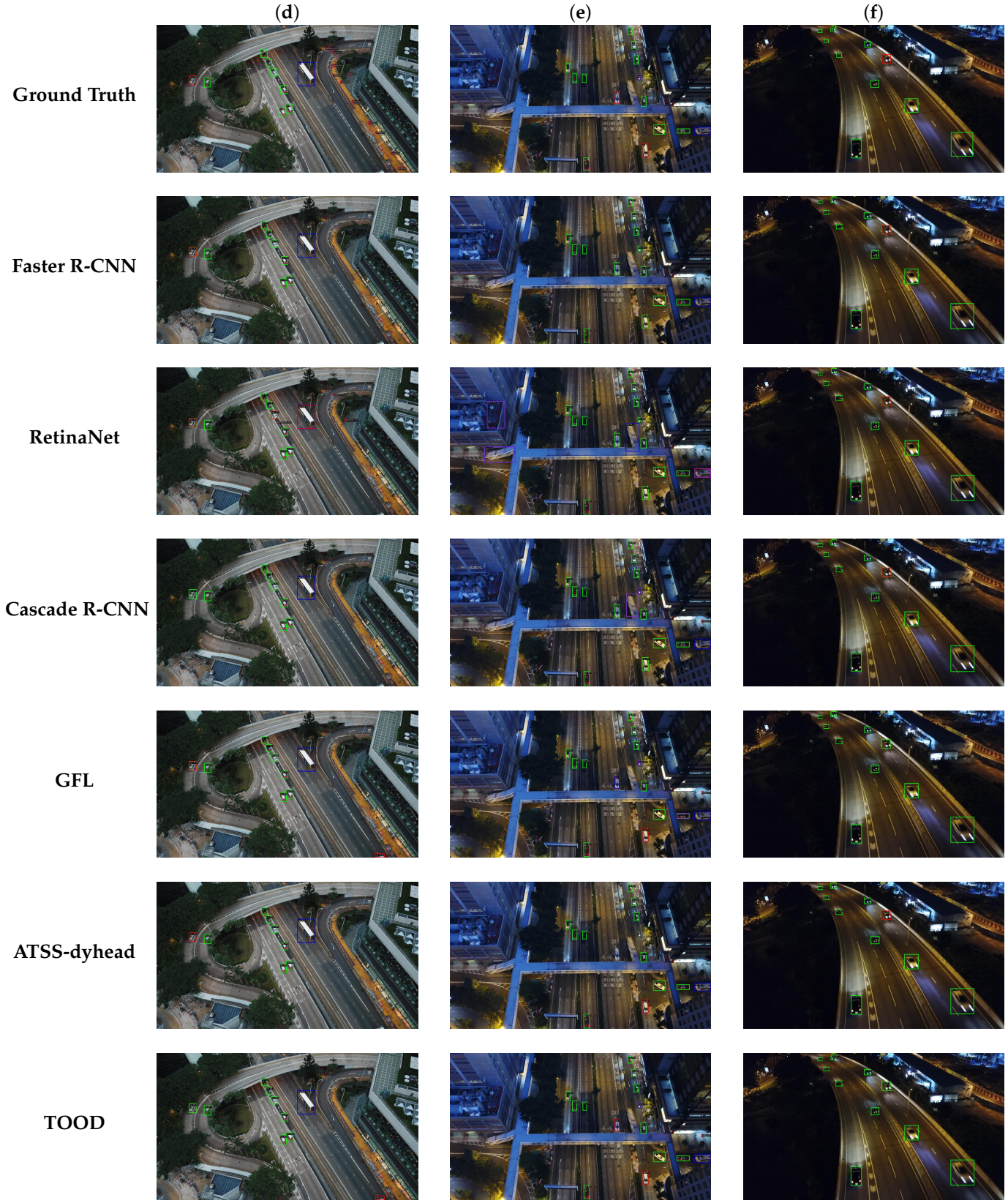

Figure 8. *Cont.*

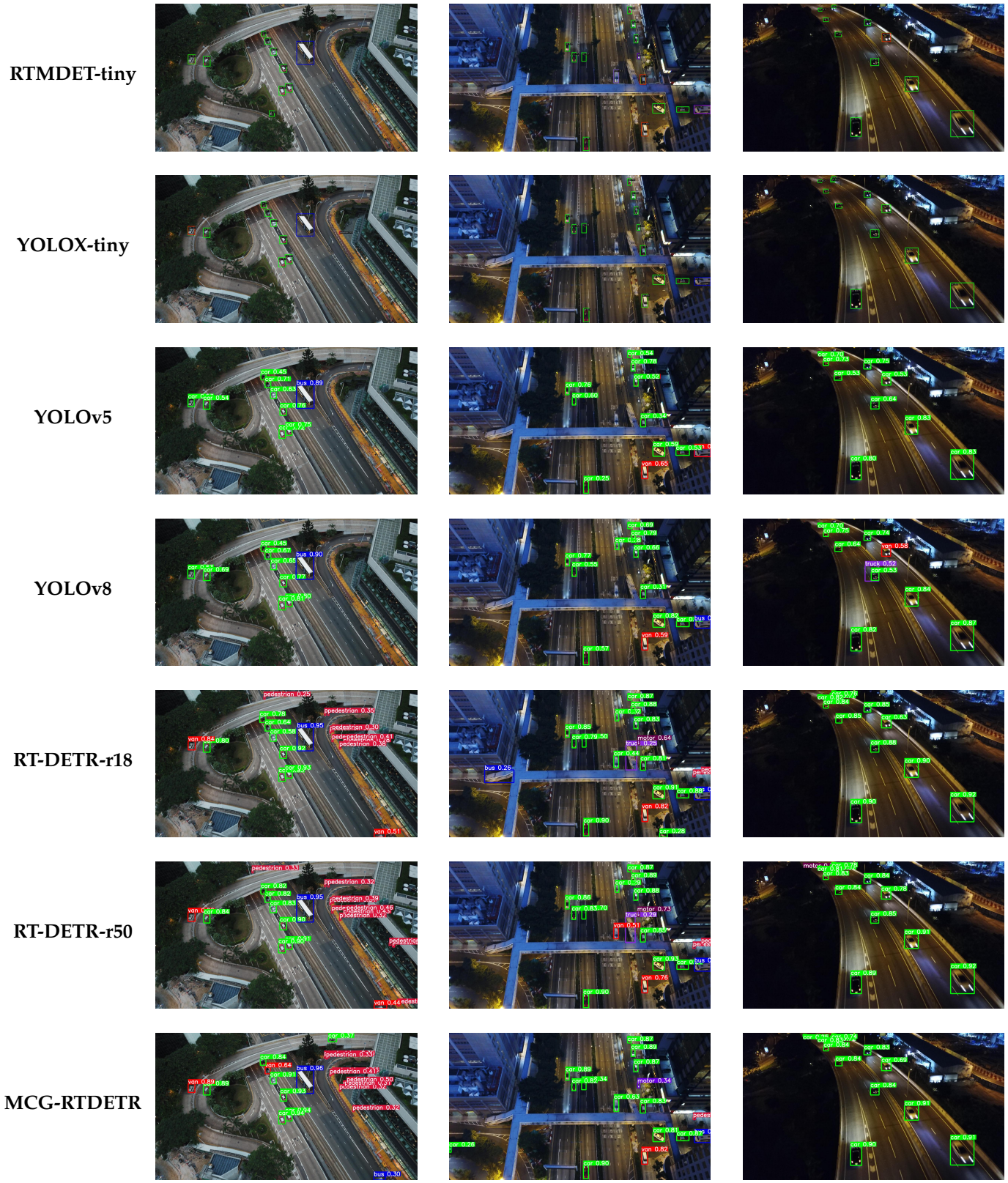

Figure 8. Visible object detection results of the proposed MCG-RTDETR and some state-of-the-art detectors on complex detection scenes of VisDrone2019 dataset. (**a**) depicts scenes with occlusion and complex environmental factors, (**b**) depicts vertical shooting angle during daylight, (**c**) depicts cloudy with intricate background. (**d**) depicts very small targets, (**e**) depicts low-light and night scene, (**f**) depicts dynamic objects like vehicles at night.

Across these diverse conditions, the MCG-RTDETR exhibited robust performance. Unlike baseline YOLO-based models and the original RT-DETR, which prioritize speed but may falter in accuracy under adverse conditions, the MCG-RTDETR excelled in both detection accuracy and efficiency. Its ability to accurately recognize and localize targets of different sizes, regardless of environmental challenges, underscores its practical applicability for real-time deployment. The experimental results underscore the MCG-RTDETR's superiority, validating its effectiveness through rigorous quantitative metrics and qualitative evaluations. This approach contributes significantly to advancing object detection technology, providing reliable solutions for complex real-world scenarios.

5. Discussion

The experimental findings in Section 4 validate that our MCG-RTDETR model substantially improves both the efficiency and precision of identifying objects in UAV imagery. Ablation experiments confirm the effectiveness of these enhanced modules, which include the dual convolution module, deformable convolution module, cascaded guided attention module (CGAM), and the context-guided feature fusion (CGFF) structure with context guided downsampling.

Integrating the dual convolution and deformable convolution modules into the backbone for feature extraction reduces computational costs and the parameter quantity while increasing detection accuracy. This improvement is due to the capability of grouped and deformable convolutions to extract features by prioritizing global information regions and attenuating unrelated background details. Moreover, a CGAM strengthens feature representation in feature maps through a group attention mechanism. Within the neck, a CGFF structure boosts cross-scale feature fusion and representation abilities effectively. The context-guided downsampling operation captures local details and global dependencies, allowing the structure to concentrate more on targets while minimizing interference from the background. Furthermore, we adjusted the detection head, which reduced the computational complexity and parameter count. Ablation experiments show that integrating the CGFF with P3 outperformed the P2 prediction head, highlighting the complementary nature of these approaches.

Comprehensive analysis of the MCG-RTDETR reveals improved object detection capabilities even under challenging scenarios with various weather conditions and complex backgrounds. Our proposed method, MCG-RTDETR, stands out across both the quantitative metrics and qualitative evaluations with high accuracy and recall values. Comparisons with other famous object detectors, including the Faster R-CNN, RetinaNet, Cascade R-CNN, GFL, ATSS-dyhead, TOOD, RTMDET-tiny, YOLOX-tiny, YOLOv5, YOLOv8, and original RT-DETR with a r18 and r50 backbone network, show that MCG-RTDETR excels in both precision and speed efficiency. Our model's resilience in addressing challenges like object occlusion, low visibility, and dynamic targets has been demonstrated through visual analysis across different environmental conditions. In future applications, the proposed algorithm is anticipated to offer practicality and efficiency in detecting small targets in complex environments for UAV-based tasks. This enhancement has potential benefits for military reconnaissance, ecological protection, natural disaster monitoring, and rescue operations.

6. Conclusions

In this study, we present the MCG-RTDETR method, which is RT-DETR-based and augmented with multi-convolution (dual convolution and deformable convolution modules), a cascaded group attention module, a context-guided feature fusion structure with context-guided downsampling operation, and a more flexible prediction head for precise object detection in UAV images. Following an extensive analysis of current state-of-the-art algorithms, our method was shown to effectively learn joint features from local details and surrounding context, enhancing global feature connection and feature fusion. The dual convolution and deformable convolution resulted in a reduction in computational costs and memory usage compared to the original RT-DETR method, meeting requirements for real-time detection. In comparisons with benchmarks like Faster R-CNN, RetinaNet,

Cascade R-CNN, and some YOLO series (YOLOX, YOLOv5, YOLOv8), the MCG-RTDETR consistently delivered competitive results in both quantitative metrics and qualitative assessments on the VisDrone2019 dataset. Therefore, our approach serves as a valuable theoretical reference for addressing similar difficulties of object detection in UAV imagery. Experimental outcomes demonstrate our proposed method's scalability and robust performance, with significant potential for practical applications such as smart city surveillance and autonomous driving.

Author Contributions: Conceptualization, C.Y. and Y.S.; Methodology, C.Y.; Software, C.Y.; Validation, C.Y.; Formal analysis, C.Y.; Investigation, C.Y. and Y.S.; Resources, Y.S.; Writing—original draft preparation, C.Y.; Writing—review and editing, C.Y. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00251595) and by the MSIT, Korea, under the ITRC support program (IITP-2024-RS-2023-00258639) supervised by the IITP.

Data Availability Statement: Data are contained within the article.

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

References

1. Wu, X.; Li, W.; Hong, D.; Tao, R.; Du, Q. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *Geosci. Remote Sens.* **2022**, *10*, 91–124. [CrossRef]
2. Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; Gould, S. Image retrieval on real-life images with pre-trained vision-and-language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 2125–2134.
3. Reis, D.; Kupec, J.; Hong, J.; Daoudi, A. Real-time flying object detection with YOLOv8. *arXiv* **2023**, arXiv:2305.09972.
4. Ye, B.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Joint feature learning and relation modeling for tracking: A one-stream framework. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 341–357.
5. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
6. 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.
7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
8. Terven, J.; Cordova-Esparza, D. A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *arXiv* **2023**, arXiv:2304.00501.
9. 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; pp. 213–229.
10. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. *arXiv* **2023**, arXiv:2304.08069.
11. Zhong, J.; Chen, J.; Mian, A. DualConv: Dual convolutional kernels for lightweight deep neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 9528–9535. [CrossRef] [PubMed]
12. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
13. Wu, T.; Tang, S.; Zhang, R.; Zhang, Y. CGNet: A light-weight context guided network for semantic segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 1169–1179. [CrossRef] [PubMed]
14. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Republic of Korea, 27–28 October 2019; pp. 213–226.
15. 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, FL, USA, 18–22 June 2018; pp. 6154–6162.
16. 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 European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
17. 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, 26–30 June 2016; pp. 770–778.

18. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
19. Jocher, G. YOLOv5 by Ultralytics (Version 7.0). 2020. Available online: https://zenodo.org/records/7347926 (accessed on 18 December 2023).
20. Solawetz, J. What is YOLOv8? The Ultimate Guide. 2023. Available online: https://blog.roboflow.com/whats-new-in-yolov8/ (accessed on 18 December 2023).
21. 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.
22. 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, 18–22 June 2018; pp. 8759–8768.
23. Yu, C.S.; Shin, Y. SAR ship detection based on improved YOLOv5 and BiFPN. *ICT Express* **2023**, *10*, 28–33. [CrossRef]
24. 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, 13–19 June 2020; pp. 10781–10790.
25. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
26. Lin, Q.; Ding, Y.; Xu, H.; Lin, W.; Li, J.; Xie, X. Ecascade-RCNN: Enhanced cascade RCNN for multi-scale object detection in UAV images. In Proceedings of the International Conference on Automation, Robotics and Applications, Prague, Czech Republic, 4–6 February 2021; pp. 268–272.
27. Chen, C.; Gong, W.; Chen, Y.; Li, W. Object detection in remote Sensing images based on a scene-contextual feature pyramid network. *Remote Sens.* **2019**, *11*, 339. [CrossRef]
28. Zhang, X.; Izquierdo, E.; Chandramouli, K. Dense and small object detection in UAV vision based on cascade network. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Seoul, Republic of Korea, 27–28 October 2019; pp. 118–126.
29. Liu, Y.; Ding, Z.; Cao, Y.; Chang, M. Multi-scale feature fusion UAV image object detection method based on dilated convolution and attention mechanism. In Proceedings of the International Conference on Information Technology: IoT and Smart City, Xi'an, China, 25–27 December 2020; pp. 125–132.
30. Liang, X.; Zhang, J.; Zhuo, L.; Li, Y.; Tian, Q. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 1758–1770. [CrossRef]
31. Liu, M.; Wang, X.; Zhou, A.; Fu, X.; Ma, Y.; Piao, C. UAV-YOLO: Small object detection on unmanned aerial vehicle perspective. *Sensors* **2020**, *20*, 2238. [CrossRef] [PubMed]
32. Zhang, Z.; Liu, Y.; Liu, T.; Lin, Z.; Wang, S. DAGN: A real-time UAV remote sensing image vehicle detection framework. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1884–1888. [CrossRef]
33. Liu, Z.; Gao, G.; Sun, L.; Fang, Z. HRDNet: High-resolution detection network for small objects. In Proceedings of the IEEE International Conference on Multimedia and Expo, Shenzhen, China, 5–9 July 2021; pp. 1–6.
34. Zhang, P.; Zhong, Y.; Li, X. SlimYOLOv3: Narrower, faster and better for real-time UAV applications. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 37–45.
35. Wang, F.; Wang, H.; Qin, Z.; Tang, J. UAV target detection algorithm based on improved YOLOv8. *IEEE Access* **2023**, *11*, 116534–116544. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L., Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
37. 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, 18–22 June 2023; pp. 14420–14430.
38. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
39. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXix:1906.07155.
40. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
41. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
42. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7373–7382.
43. Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing aided hyper inference and fine-tuning for small object detection. In Proceedings of the IEEE International Conference on Image Processing, Bordeaux, France, 16–19 October 2022; pp. 966–970.

44. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An empirical study of designing real-time object detectors. *arXiv* **2022**, arXiv:2212.07784.
45. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.