



## Article

# YOLO-DCTI: Small Object Detection in Remote Sensing Base on Contextual Transformer Enhancement

Lingtong Min<sup>1</sup>, Ziman Fan<sup>1</sup> , Qinyi Lv<sup>1,\*</sup>, Mohamed Reda<sup>2</sup> , Linghao Shen<sup>3</sup> and Binglu Wang<sup>3</sup>

<sup>1</sup> School of Electronic Information, Northwestern Polytechnical University, Xi'an 710072, China; minlingtong@nwpu.edu.cn (L.M.); fanziman@mail.nwpu.edu.cn (Z.F.)

<sup>2</sup> Department of Avionics, Military Technical College, Cairo 4393010, Egypt; mohamedredaismail@mtc.edu.eg

<sup>3</sup> School of Automation, Northwestern Polytechnical University, Xi'an 710072, China; shenly@mail.nwpu.edu.cn (L.S.); binglu\_wang@mail.nwpu.edu.cn (B.W.)

\* Correspondence: lvqinyi@nwpu.edu.cn

**Abstract:** Object detection for remote sensing is a fundamental task in image processing of remote sensing; as one of the core components, small or tiny object detection plays an important role. Despite the considerable advancements achieved in small object detection with the integration of CNN and transformer networks, there remains untapped potential for enhancing the extraction and utilization of information associated with small objects. Particularly within transformer structures, this potential arises from the disregard of the complex and the intertwined interplay between spatial context information and channel information during the global modeling of pixel-level information within small objects. As a result, valuable information is prone to being obfuscated and annihilated. To mitigate this limitation, we propose an innovative framework, YOLO-DCTI, that capitalizes on the Contextual Transformer (CoT) framework for the detection of small or tiny objects. Specifically, within CoT, we seamlessly incorporate global residuals and local fusion mechanisms throughout the entire input-to-output pipeline. This integration facilitates a profound investigation into the network's intrinsic representations at deeper levels and fosters the fusion of spatial contextual attributes with channel characteristics. Moreover, we propose an improved decoupled contextual transformer detection head structure, denoted as DCTI, to effectively resolve the feature conflicts that ensue from the concurrent classification and regression tasks. The experimental results on the Dota, VISDrone, and NWPU VHR-10 datasets show that, on the powerful real-time detection network YOLOv7, the speed and accuracy of tiny targets are better balanced.

**Keywords:** small object detection; remote sensing images; transformer; YOLOv7



**Citation:** Min, L.; Fan, Z.; Lv, Q.; Reda, M.; Shen, L.; Wang, B. YOLO-DCTI: Small Object Detection in Remote Sensing Base on Contextual Transformer Enhancement. *Remote Sens.* **2023**, *15*, 3970. <https://doi.org/10.3390/rs15163970>

Academic Editors: Zhitong Xiong, Qiang Li and Muhammad Shahzad

Received: 14 July 2023

Revised: 5 August 2023

Accepted: 8 August 2023

Published: 10 August 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Remote sensing object detection is a prominent and consequential application within the realm of remote sensing image processing [1]. It aims to accurately identify and locate specific target instances within an image. Within this domain, remote sensing small object detection holds particular importance as it focuses on detecting objects in remote sensing images that occupy a very small area or consist of only a few pixels. Detecting small objects is considerably more challenging than detecting larger objects, resulting in lower accuracy rates [2]. In recent years, small object detection based on convolutional neural networks (CNNs) has rapidly developed with the rapid growth of deep learning [3]. Small object detection often faces challenges such as limited information on small objects, scarcity of positive samples, and imbalanced classification. To tackle this challenge, researchers and experts have put forth diverse deep neural network methodologies, encompassing CNNs, GANs, RNNs, and transformers, to tackle the detection of small objects, encompassing remote sensing small objects. To improve the detection of small objects, Liu W et al., proposed the YOLOV5-Tassel network, which introduced the SimAM module in front of each detection head to extract the features of interest [4]. Li J. et al., suggested using

GAN models to generate high-resolution images of small objects, narrowing the gap between small and large objects, and improving the detection capability of tiny objects [5]. Xu W et al. integrated contextual information into the Swin Transformer and designed an advanced framework called the foreground-enhanced attention Swin Transformer (FEA-Swin) [6]. Although the accuracy of detecting small objects has improved, the speed has been somewhat compromised. Zhu X. et al., proposed the YOLOv5-THP model, which is based on YOLOv5 and adds a transformer model with an attention mechanism to the detection head [7]. While this enhances the network's performance in detecting small objects, it also brings a significant computing burden.

In the field of remote sensing, detecting small objects remains challenging due to large image scales, complex and varied backgrounds, and unique shooting perspectives. Cheng et al. propose a model training regularization method that enhances the performance of detection of small or tiny objects in remote sensing by exploiting and incorporating global contextual cues and image-level contextual information [8]. Liu J. et al., added a dilated convolution module to the FPN and designed a relationship connection attention module to automatically select and refine features, combining global and local attention to achieve the detection task of small objects in remote sensing [9]. Cheng et al., proposed an end-to-end cross-scale feature fusion (CSFF) framework based on the feature pyramid network (FPN), which inserted squeeze-and-excitation (SE) modules at the top layer to achieve better detection of tiny objects in optical remote sensing images [10]. Dong et al., proposed a CNN method based on balanced multi-scale fusion (BMF-CNN), which fused high- and low-level semantic information to improve the detection performance of tiny objects in remote sensing [11]. Liang X. et al., proposed a single-shot detector (FS-SSD) based on feature fusion and scaling to better adapt to the detection of tiny or small objects in remote sensing. FS-SSD added a scaling branch in the deconvolution module and used two feature pyramids generated by the deconvolution module and feature fusion module together for prediction, improving the accuracy of object detection [12]. Xu et al., designed a transformer-guided multi-interaction network (TransMIN) using local-global feature interaction (LGFI) and cross-view feature interaction (CVFI) modules to enhance the performance of small object detection in remote sensing. However, this improvement unavoidably introduces a computational burden [13]. Li et al., proposed a transformer that aggregates multi-scale global spatial positions to enhance small object detection performance but it also comes with a computational burden [14]. To reduce the computational cost of the transformer, Xu et al., improved the lightweight Swin transformer and designed a Local Perception Swin transformer (LPSW) backbone network to enhance small-scale detection accuracy [15]. Gong et al., designed an SPH-YOLOv5 model based on Swin Transformer Prediction Heads (SPHs) to balance the accuracy and speed of small object detection in remote sensing [16]. Although many experts and scholars are studying the balance between detection accuracy and inference speed, achieving an elegant balance remains a challenging problem [17–21].

Considerable advancements have been achieved in the utilization of transformers [6,7,13–16] for small object detection within the remote sensing domain. The exceptional performance of the Contextual Transformer (CoT) [22] in harnessing spatial contextual information, thereby offering a fresh outlook on transformer design, merits significant attention. In the domain of remote sensing, small target pixels are characterized by a scarcity of spatial information but a profusion of channel-based data. Consequently, the amalgamation and modeling of spatial and channel information assume paramount importance. Furthermore, transformers impose notable demands on computational resources and network capacity, presenting a challenge in striking an optimal balance between detection accuracy and processing speed for small object detection in the remote sensing discipline. Meanwhile, Bar M et al. demonstrated that the background is critical for human recognition of objects [18]. Empirical research in computer vision has also shown that both traditional methods [19] and deep learning-based methods [12] can enhance algorithm performance by properly modeling spatial context. Moreover, He K. et al., have proven that residual

structures are advantageous for improving network performance [17,20]. Finally, we note that the classification and regression tasks of object detection focus on the salient features and boundary features of the target, respectively [23]. Therefore, a decoupled detection head incorporating residual structure as well as channel and spatial context knowledge should have a positive impact on the detection of small or tiny objects.

We propose a new detection framework, YOLO-DCTI, for detecting small or tiny objects in remote sensing images. By introducing a global residual structure and a local fusion structure into the contextual transformer (CoT), and designing an improved decoupled contextual transformer detection head structure (DCTI) based on CoT, we have achieved improved detection performance for small or tiny objects on the powerful single-stage benchmark network YOLOv7. The main contributions of this paper can be summarized as follows:

1. We have developed the CoT-I module, an extension of the original CoT framework, which integrates global residual structures and local fusion modules. This integration facilitates the extraction of spatial context background features and the fusion of channel features, thereby enabling the network to learn deeper-level characteristics. In comparison to the conventional CoT approach, the inclusion of global residual structures empowers the network to capture more profound features, while the incorporation of local fusion structures seamlessly combines background context features with channel features.
2. We introduce an efficient decoupled detection head structure DCTI, leveraging the CoT-I framework, to mitigate the limited exploration and utilization of salient region features and boundary-adjointing features arising from the interdependence of classification and regression tasks within most object detection heads. This decoupled design allows the classification task to emphasize salient region features, while the regression task focuses on boundary-surrounding features. Concurrently, CoT-I effectively exploits and harnesses the feature relationships between spatial context and channels, facilitating the detection of small objects in remote sensing and yielding substantial improvements in detection accuracy.
3. Despite the escalation in model parameters and the consequential inference latency resulting from the adoption of our proposed DCTI structure, the integration of global residual connections and local fusion strategies yields a notable enhancement in inference accuracy without incurring any detrimental impact on the inference speed. Comparative evaluation against the baseline YOLO v7 model showcases a substantial improvement in the inference accuracy specifically for diminutive targets, mAP@0.5:0.95 surging from 61.8 to 65.2. Additionally, our model achieves a reduction of 0.3ms in the inference speed per image with dimensions of  $640 \times 640$ .

## 2. Related Work

### 2.1. Transformer Framework for Object Detection

The transformer structure, based on self-attention, first appeared in NLP tasks. Compared to modern convolutional neural networks (CNN) [24], the Vision Transformer has made impressive progress in the field of computer vision. After Dosovitskiy A et al. successfully introduced transformers into computer vision [25], many scholars turned to transformers [26–28]. In object detection, DETR [29] and Pix2seq [30] are the earliest transformer detectors that define two different object detection paradigms. However, transformers have many parameters, require high computing power and hardware, and are not easily applicable. To apply transformers on mobile devices, Mehta S et al. proposed a lightweight MobileViT series [31–33], which achieved a good balance between accuracy and real-time performance, and has been widely used in risk detection [34], medicine [35], and other fields. A major advantage of transformers is that they can use the attention mechanism to model the global dependence of input data, obtain longer-term global information, and ignore the connection between local contexts. To address this problem, Li Y. et al., proposed a lightweight CoT [22] self-attention module to capture contextual background

information on 2D feature maps. It can extract information between local contexts while capturing global dependencies for more adequate information exchange. In this paper, we use CoT to exploit the global characteristics of spatial context and channels. Based on the original structure, we added the global residual and local fusion structures to further tap and utilize the characteristics of space and channels.

## 2.2. YOLO Framework for Object Detection

In 2015, YOLO [36] introduced a one-stage object detection method that combined candidate frame extraction, CNN feature learning, and NMS optimization to simplify the network structure. The detection speed was nearly 10 times faster than R-CNN, making real-time object detection possible with the computing power available at that time. However, it was not suitable for detecting small objects. YOLOv2 [37] added optimization strategies such as batch normalization and a dimensional clustering anchor box based on v1 to improve the accuracy of object regression and positioning. YOLOv3 [38] added the residual structure and FPN structure based on v2 to further improve the detection performance of small objects. The network framework structure after YOLOv3 can be roughly divided into three parts, backbone, neck, and head. Subsequent versions have optimized internal details to varying degrees. For example, YOLOv4 [39], based on v3, further optimized the backbone network and activation function, and used Mosaic data enhancement to improve the robustness and reliability of the network. YOLOv5 [40] added the focus structure based on v4 and accelerated the training speed by slicing. YOLOv6 [41] introduced RepVGG in the backbone, proposed a more efficient EfficientRep block, and simplified the design of the decoupling detection head to improve the detection efficiency. YOLOv7 [42] adopted the E-ELAN structure in the neck part, which reduces the inference speed, and used the auxiliary head training method. At present, YOLOv7 is one of the more advanced object detection networks due to its real-time characteristics. It is widely used in fields with high time requirements such as industrial equipment inspection [43], sea rescue [44], and aquaculture [45]. Therefore, we use YOLOv7, one of the powerful benchmarks, as the benchmark model.

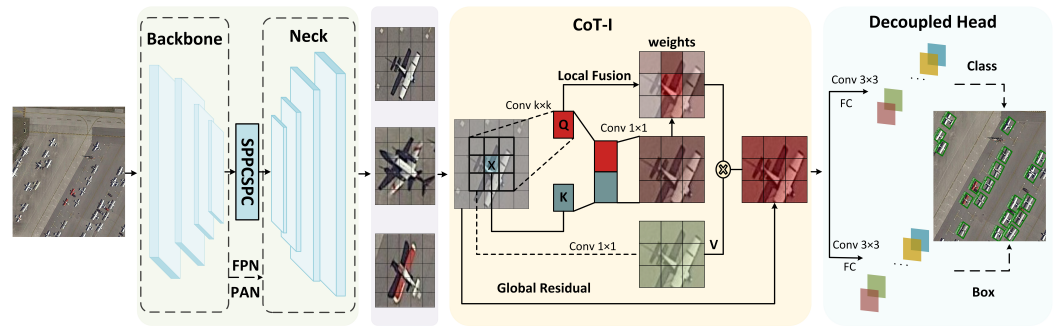
## 2.3. Detection Head Framework for Object Detection

In the object detection task, there are two tasks: classification and regression, which respectively output the classification and bounding box position of the object. Song G. et al., pointed out that the focus of the classification and regression tasks is different [23]. Specifically, classification pays more attention to the texture content of the object, while regression pays more attention to the edge information of the object. Wu Y et al. suggested that it may be better to divide classification and regression tasks into FC-head and Conv-head [46]. In the single-stage model, YOLOX [47] adopts the decoupling head structure that separates the classification and regression branches and adds two additional  $3 \times 3$  convolutional layers. This improves detection accuracy at the cost of inference speed. Building upon this approach, YOLOv6 takes into account the balance between the representation ability of related operators and the hardware computing overhead and adopts the Hybrid Channels strategy to redesign a more efficient decoupling head structure that reduces the cost while maintaining accuracy. They also mitigate the additional latency overhead of  $3 \times 3$  convolutions in the decoupled detection head. Feng C. et al., use feature extractors to learn task interaction features from multiple convolutional layers to enhance the interaction between classification and localization [48]. They also pointed out that the interaction characteristics of different tasks may vary due to the classification and localization goals. To resolve the feature conflict introduced between the two tasks, they designed a layer attention mechanism that focuses on different types of features such as different layers and receptive fields. This mechanism helps to resolve a certain degree of feature conflict between the two tasks.



### 3. Proposed Method

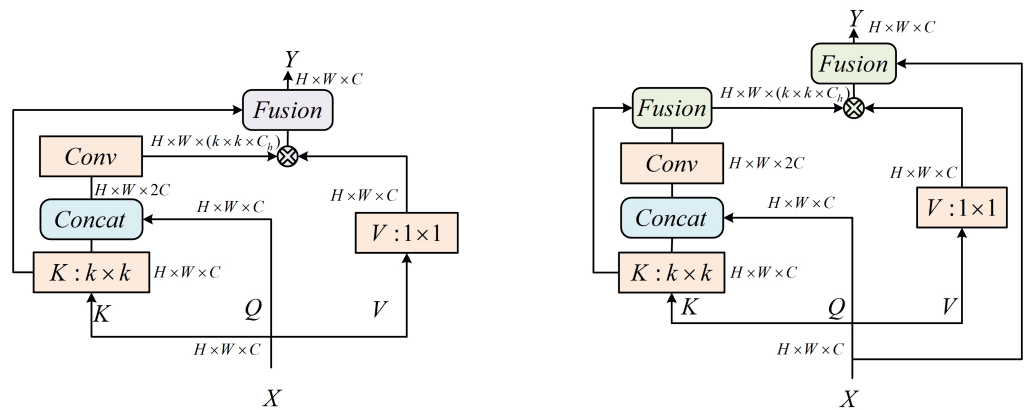
We present an improved decoupled contextual transformer (YOLO-DCTI) for the detection of tiny or small objects in the domain of remote sensing. Our proposed framework is built upon the foundation of YOLOv7. The comprehensive architecture of our framework is depicted in Figure 1. Our contributions begin with the feature  $X \in \mathbb{R}^{H \times W \times C}$  obtained after the backbone, FPN, and PAN stages. In this section, we first give a brief overview of the widely adopted Contextual Transformer (CoT) framework in object detection. Subsequently, we introduce an enhanced variant named CoT-I, which incorporates a global residual structure and a local fusion structure into the CoT module. The global residual mechanism integrates input information with self-attention features, while the local fusion mechanism combines spatial contextual information with channel-based information. Ultimately, we integrate the CoT-I module into a decoupled detection head named DCTI, enabling the establishment of global interdependencies between the classification and regression tasks through the utilization of self-attention mechanisms. This integration facilitates the comprehensive exploration and exploitation of a wider spectrum of channel features and spatial contextual features.



**Figure 1.** The overall framework of our YOLO-DCTI; DCTI consists of CoT-I and Decoupled-Head. The FPN features are input to the CoT-I module for comprehensive modeling of global contextual information and spatial relationships. Subsequently, a dual-branch architecture is employed to effectively extract and discriminate both category-specific and localization-specific information.

#### 3.1. Contextual Transformer

In this section, we present the formulation of the Contextual Transformer (CoT) framework, illustrated in Figure 2 (left). The input  $X$ , obtained from the Feature Pyramid Network (FPN), undergoes three transformation matrices:  $W_q$ ,  $W_k$ ,  $W_v$ . These matrices yield  $K = XW_k$ ,  $Q = XW_q$ ,  $V = XW_v$ . Specifically,  $W_q$  is an identity matrix,  $W_k$  represents spatial convolution using a  $k \times k$  kernel, and  $W_v$  signifies spatial convolution with a  $1 \times 1$  kernel. The output  $Y$  is mathematically expressed as:



**Figure 2.** The detailed structures of (left) CoT and (right) our CoT-I block.

$$Y = K + f_c(K, Q)W_\delta W_\theta \times V \quad (1)$$

In the above equation, the function  $f_c()$  denotes concatenation along the channel dimension  $C$ , while  $W_\delta$  and  $W_\theta$  correspond to  $1 \times 1$  convolutions in the spatial domain. The symbol  $\times$  represents matrix multiplication. For brevity, we omit the transformation of the channel dimension  $C$  and the batch normalization (BN) operation during the  $1 \times 1$  convolutions.

It is evident that the learned key matrix  $K$  captures significant information from neighboring pixels within the spatial domain, incorporating essential static spatial context information. Subsequently,  $Q$  and  $K$  are concatenated along the channel dimension, followed by the application of two  $1 \times 1$  convolutions:  $W_\delta$  with an activation function and  $W_\theta$  without an activation function. Matrix multiplication is then performed with  $V$ , resulting in a matrix  $T$  enriched with dynamic contextual information, which can be used as follows:

$$T = f_c(K, Q)W_\delta W_\theta \quad (2)$$

This resultant matrix  $T$  is subsequently fused with the static contextual information  $K$  to derive the final output  $Y$ . Notably,  $Y$  incorporates both dynamic contextual information and static contextual information. CoT demonstrates exceptional performance in leveraging contextual information; however, it partially overlooks the joint contribution of channel information and contextual information.

### 3.2. Improved Contextual Transformer (CoT-I)

First of all, the ResNet [20] and Densenet [17] models prove the effectiveness of the residual structure in the model, which helps optimize the deeper layer of the architecture. We add a residual structure from the input to the output, called the global residual structure.

Secondly, we note that the intermediate matrix  $T$  captures relatively persistent global dependencies along the channel dimension, whereas the static contextual information  $K$  contains abundant spatial contextual information. The interrelation between them is more closely intertwined compared to the dynamic contextual information. Hence, we integrate the intermediate matrix  $T$  with the static contextual information  $K$  to achieve a more comprehensive representation of spatial and channel features. This fusion can be mathematically expressed as follows:

$$Y = X + [f_c(K, Q)W_\delta W + K] \times V \quad (3)$$

where  $+$  represents through attention mechanism to complete the fusion of static  $K$  and intermediate vector  $T$ .

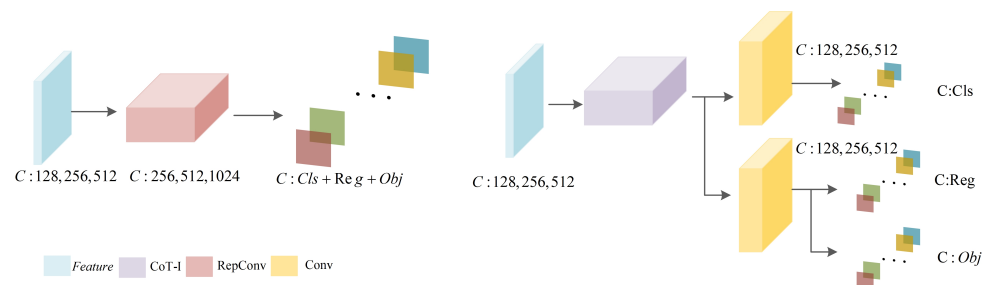
The global residual mechanism plays a pivotal role in facilitating the fusion of input information and self-attention features. By incorporating a global residual structure, the method adeptly amalgamates and consolidates information from diverse hierarchical levels of input, thereby engendering a more comprehensive and all-encompassing feature representation. This ensemble effectively captures long-range dependencies and augments the overall discriminative capacity of the model.

The local fusion mechanism concentrates on integrating spatial context information with channel-based information. Through the integration of these two distinct information sources, the method proficiently models the intricate interplay between neighboring pixels and harnesses the wealth of channel-based information inherent in remote sensing data. This fusion framework empowers the model to capture intricate details and contextual cues with greater precision, thereby engendering superior performance in the detection of small objects.

### 3.3. Improved Decoupled Contextual Transformer Detection Head (DCTI)

Our DCTI structure starts with features  $X \in \mathbb{R}^{H \times W \times C}$  obtained after the backbone, FPN, and PAN stages.  $X$  has three different dimensions of features; ( $H$ ,  $W$ , and  $C$ ) are (20, 20, 128), (40, 40, 256), and (80, 80, 512), respectively. For the purpose of enhancing the

lucidity of the presentation, we solely exemplify the variations in channel  $C$ , as depicted in Figure 3. The coupled detection head depicted (left) leverages  $1 \times 1$  convolutions and the RepConv module to enhance feature information along the channel dimension. Despite its simplicity, this approach is remarkably effective. However, in the context of object detection tasks, it becomes evident that the classification task primarily emphasizes salient feature information, while the regression task is more concerned with capturing boundary feature information related to the targets. Employing a shared Repconv module for both tasks inevitably introduces conflicts between them.



**Figure 3.** Comparison between (left) YOLOv7 head and (right) ours.

Moreover, we observe the exceptional capability of CoT-I in capturing global information for effective modeling. In comparison to  $1 \times 1$  convolutions, CoT-I exhibits superior feature exploration in both the classification and regression tasks.

Consequently, we propose the DCTI, building upon the CoT-I framework, as illustrated (right). Initially, CoT-I assimilates the feature information derived from the feature pyramid and computes an intermediate variable  $T$ , which encompasses abundant spatial information, employing Equation (1). Subsequently, a local fusion strategy is employed to merge the spatial and channel information, yielding the dynamic feature  $K1$  through the utilization of Equation (3). Finally, employing a global residual strategy, the input from the feature pyramid is combined with the dynamic feature to produce the output  $Y$ . The decoupled detection head further individually models  $Y$  to obtain category information and bounding box information.

#### 4. Results

To evaluate the performance of the YOLOv7-DCTI algorithm for remote sensing small object detection, training, and testing were conducted on the Dota-small dataset [49]. Furthermore, to assess the algorithm's overall performance, this experiment included training and testing on the VISDrone dataset [50] and NWPU VHR-10 datasets [51]. A comparison was made between seven different networks, namely Faster RCNN, SSD, YOLO v5s, YOLO v5l, YOLO v5m, YOLO v7-tiny, and YOLOv7, using the aforementioned three datasets. To ensure fairness among the YOLO series, a batch size of 16 was utilized during the training process and pre-trained weights were not employed. The data augmentation strategy [52,53] remained consistent with other training conditions. During testing, the NMS [54] threshold was uniformly set to 0.65 and the batch size was uniformly set to 32.

##### 4.1. Datasets

###### 4.1.1. Dota-Small

In recent years, several remote sensing datasets have been developed. This paper focuses on the extraction of small or tiny objects from the DoTAv1.0 dataset, which consists of 2000 aerial images of 2000 cities and over 190,000 fully labeled objects, each of which comprises eight positional parameters ( $x1, y1, x2, y2, x3, y3, x4, y4$ ). In this study, we have selected a dataset that includes five categories of small objects, namely small vehicles, large vehicles, planes, storage tanks, and ships. However, due to the large image size in the DoTAv1.0 dataset, direct training is not feasible. Therefore, we have cropped the images to a size of  $1024 \times 1024$ , resulting in a total of 8624 images. These images were subsequently divided into three sets according to the train:val:test ratio, with a split of 8:2:2.

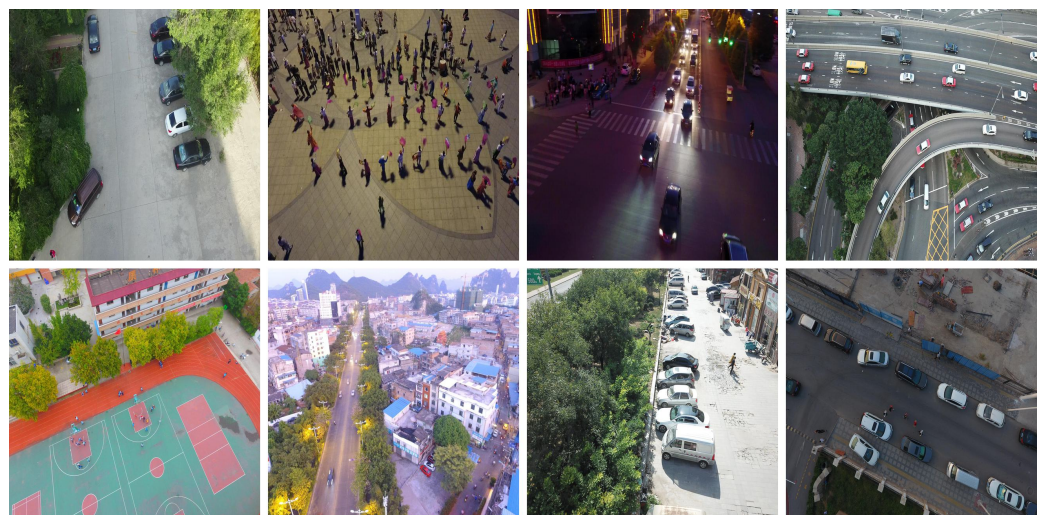
Among these, 5176 images were used for training, while 1724 images were allocated to validation and testing. The five types of objects included in the dataset are small vehicle, large vehicle, plane, storage tank, and ship, as illustrated in Figure 4. We have set three sets of anchors with the following dimensions: (10,10, 13,22, 24,12), (23,24, 28,37, 45,25), and (49,50, 91,88, 186,188).



**Figure 4.** Some selected unprocessed images of Dota-small.

#### 4.1.2. VisDrone

VisDrone is a widely recognized and highly demanding aerial photography dataset that is extensively used in UAV (Unmanned Aerial Vehicle) applications. It features a meticulous manual annotation process that has accurately labeled and classified 342,391 objects into 10 distinct categories. However, the official evaluation portal for the test-challenge set is unavailable, so we have utilized the test-dev set for evaluating our proposed method. Figures 5–7 showcases a selection of unprocessed images extracted from the VisDrone dataset. In our experiments, we have employed three sets of anchor dimensions: (3,4, 4,8, 8,7), (7,14, 14,9, 13,20), and (25,13, 27,27, 51,40).

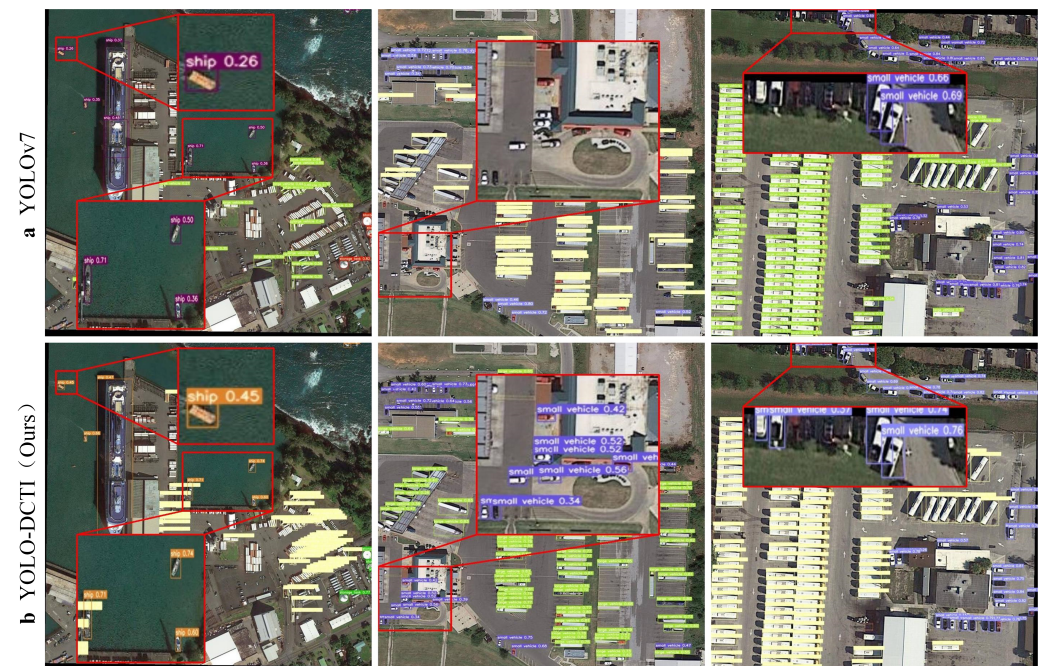


**Figure 5.** Some selected unprocessed images of VISDrone.





**Figure 6.** Some selected unprocessed images of NWPU VHR-10.



**Figure 7.** Recognition results in crowded environments of Dota-small dataset. (a) Recognition results of the YOLOv7 network. (b) Recognition results of the YOLO-DCTI network.

#### 4.1.3. NWPU VHR-10

To assess the generalization capability of our proposed method, we conducted experiments on the NWPU VHR-10 dataset. This dataset is specifically designed for geospatial object detection and comprises ten different object categories, namely airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. In our study, we randomly divided the dataset into three sets: 60% of the images were allocated to the training set, 20% to the validation set, and the remaining 20% to the testing set. Figure 8 showcases a selection of unprocessed images from the NWPU VHR-10 dataset. For our experiments, we employed three sets of anchor dimensions: (24,23, 30,30, 32,46), (47,32, 52,51, 74,61), and (88,96, 205,104, 150,194).





**Figure 8.** Some examples of detection results on the Dota-small dataset using YOLO-DCTI.

#### 4.2. Experimental Environment and Settings

The experiment was conducted using a 64-bit Windows 10 operating system. The GPU utilized was NVIDIA GeForce RTX3090 and the deep learning framework employed was Torch v1.10.0. To evaluate the performance of object detection methods, this paper adopts the common indicators of object detection. This paper employs several metrics, including accuracy, recall, average precision, mean average precision, and average inference time per image (ms). The accuracy rate measures the proportion of correctly predicted samples out of the total tested samples. The recall rate indicates the proportion of positive samples that are accurately predicted. AP is calculated as the area under the precision–recall curve. mAP represents the average of AP values across all categories. Specifically, mAP@0.5:0.95 refers to the average mAP value computed at ten IoU thresholds (0.50, 0.55, ..., 0.95). On the other hand, mAP@0.5 denotes the mAP value computed at an IOU threshold greater than 0.5.

#### 4.3. Experimental Results and Analysis

Experiments were conducted on three publicly available datasets, namely VISDrone, NWPU VHR-10, and Dota-small, to assess the efficacy of the proposed method. The experimental evaluation was carried out in four distinct stages: (1) Validation of the method's feasibility on the Dota-small dataset, including a comparative analysis against other object detection techniques to showcase its effectiveness. (2) Examination of the method's generalization capability using the VISDrone dataset. (3) Further validation of the method's performance on the NWPU VHR-10 dataset. (4) Evaluation of the performance of our model, encompassing inference speed, model parameters, and detection accuracy, and conducting a comparative analysis with existing models. (5) Execution of ablation experiments to scrutinize the effectiveness of each step within the proposed method and assess the optimal parameter configuration.

##### 4.3.1. Experiments on Dota-Small Dataset

To validate the proposed model, we trained it for 200 epochs on the Dota-small dataset. To ensure fairness, we used a batch size of 24 during the training process and did not use pre-trained weights. We kept the data augmentation strategy [52,53] and other training conditions consistent. During testing, the NMS [54] threshold was uniformly

set to 0.65 and the batch size was uniformly set to 32. In our experimental analysis, we conducted a comparative evaluation of the YOLOv7-DCTI algorithm with mainstream object detection algorithms using the Dota-small dataset generated specifically for this study. The outcomes of these experiments are documented in Table 1, encompassing five distinct categories: small vehicles, large vehicles, planes, storage tanks, and ships. The AP@0.5 values provided in the table indicate the average recognition accuracy achieved by each algorithm for individual categories. Additionally, the columns denoted as mAP@0.5 and mAP@0.5:0.95 represent the average recognition accuracy across all categories.

The Dota-small dataset contains predominantly small or tiny objects with limited information and is characterized by complex and variable image backgrounds. Distinguishing these objects from the background presents significant challenges, as some objects may be partially occluded, further complicating detection. In comparison to other mainstream object detection algorithms, the enhanced network proposed in this study demonstrates a notable accuracy advantage in detecting small or tiny objects.

Our proposed method achieves the highest accuracy rate of 65.2% for small or tiny objects in the Dota-small dataset, surpassing YOLOv7 by 3.4%. In comparison to these four object detection algorithms, Faster R-CNN, SSD, YOLOv5l, and YOLOv7-tiny, the mean average precision (mAP) at the intersection over union (IoU) threshold range of 0.5 to 0.95 showed improvements of 19.3%, 45.2%, 4.5%, and 12.1%, respectively.

Although the detection speed of YOLOv7 and YOLOv5 is similar to that of our proposed method, their mAP scores are lower. In scenarios where the differentiation among the YOLO series detection heads is minimal, the proposed method achieves higher mAP at the IoU range of 0.5 to 0.95, while maintaining similar detection speeds. This demonstrates that the proposed method effectively compensates for the differences in detection heads and offers greater advantages.

The introduction of a Contextual Transformer (CoT) in the decoupled head unavoidably leads to a slight sacrifice in reasoning speed. However, the global residual structure and local fusion structure do not experience any decrease in speed. As a result, the detection speed of the proposed structure remains largely unaffected even with increasing complexity, achieving a favorable balance between inference speed and detection accuracy.

**Table 1.** Comparison of detection accuracy of different object detection algorithms on Dota-small dataset.

| Model             | Small Vehicle | Large Vehicle | Plane | Storage Tank | Ship | mAP@0.5 | mAP@0.5:0.95 |
|-------------------|---------------|---------------|-------|--------------|------|---------|--------------|
| SSD [55]          | 26.9          | 47.4          | 79.9  | 35.2         | 28.9 | 43.7    | 20.0         |
| Faster R-CNN [56] | 67.8          | 78.8          | 93.4  | 74.6         | 58.6 | 74.6    | 45.9         |
| YOLOv5s [40]      | 82.9          | 83.4          | 89.0  | 77.5         | 86.4 | 83.8    | 51.9         |
| YOLOv5l [40]      | 89.0          | 91.5          | 95.5  | 84.5         | 90.4 | 90.2    | 60.7         |
| YOLOv5m [40]      | 88.9          | 91.2          | 92.6  | 86.3         | 90.5 | 89.9    | 60.8         |
| YOLOv7-tiny [42]  | 84.5          | 87.8          | 92.8  | 77.3         | 86.2 | 85.7    | 53.1         |
| YOLOv7 [42]       | 89.9          | 91.9          | 95.6  | 84.9         | 91.1 | 90.7    | 61.8         |
| Ours              | 90.0          | 92.6          | 96.8  | 85.7         | 91.5 | 91.4    | 65.2         |

As illustrated in Figure 7, YOLO-DCTI demonstrates superior performance in accurately detecting objects with unclear features or small sizes, even in complex backgrounds. It exhibits no omissions or false detections, unlike YOLOv7, which is prone to such errors. The proposed YOLO-DCTI algorithm in this paper excels at identifying small or tiny objects in challenging scenarios, yielding relatively high prediction probabilities. In contrast, YOLOv7 may struggle to accurately recognize small or tiny objects, resulting in lower recognition probabilities compared to YOLO-DCTI.

Figure 8 depicts the detection outcomes obtained through the utilization of the YOLO-DCTI methodology. On the whole, the performance is laudable; nonetheless, certain instances of missed detections persist in densely populated target scenes exhibiting analogous configurations. This phenomenon can be ascribed to the intricacy involved in discerning like attributes within such highly congested environments. As a result, this

specific issue accentuates the imperativeness of conducting more extensive and meticulous investigations in future research endeavors.

#### 4.3.2. Experiments on VisDrone Dataset

Table 2 presents the experimental results, demonstrating the strong performance of the proposed method on the VisDrone dataset. YOLO-DCTI achieves a noteworthy improvement of 0.2% in mAP@0.5:0.95 compared to the original method while maintaining a comparable detection speed. Notably, YOLOv5x achieves a mAP@0.5 of 48.1% on this dataset, while YOLOv7 achieves a mAP@0.5 of 49.2%. However, it is important to note that both networks employ coupled detection heads, which are unable to effectively address the inherent discrepancy between classification and regression tasks, resulting in slightly lower detection accuracy compared to our proposed method. As shown in Figure 9, the effectiveness of our method is reflected in its effective detection of small and dense objects (e.g., people and cars).

**Table 2.** Experimental results on VISDrone dataset.

| Model       | P1   | P2   | B1    | C    | V    | T1   | T2   | A    | B2   | M    | mAP@0.5 | mAP@0.5:0.95 |
|-------------|------|------|-------|------|------|------|------|------|------|------|---------|--------------|
| SSD         | 0.05 | 0.02 | 0.009 | 0.36 | 0.08 | 0.05 | 0.0  | 0.0  | 0.19 | 0.02 | 0.081   | 0.042        |
| Faster RCNN | 37.5 | 19.4 | 13.3  | 71.9 | 42.5 | 42.8 | 19.8 | 18.1 | 58.4 | 34.4 | 35.8    | 20.2         |
| YOLOv5s     | 34.9 | 24.8 | 11.6  | 75.0 | 48.5 | 46.4 | 22.5 | 20.6 | 64.8 | 35.9 | 38.5    | 20.0         |
| YOLOv5l     | 45.5 | 33.5 | 19.6  | 80.3 | 55.0 | 55.6 | 33.1 | 28.1 | 70.5 | 46.7 | 46.8    | 25.3         |
| YOLOv5x     | 44.9 | 33.7 | 20.3  | 80.7 | 56.4 | 58.3 | 33.4 | 32.6 | 72.4 | 48.0 | 48.1    | 26.2         |
| YOLOv7-tiny | 33.6 | 23.9 | 10.8  | 73.1 | 45.4 | 41.2 | 21.6 | 19.5 | 59.3 | 35.7 | 36.4    | 18.3         |
| YOLOv7      | 46.5 | 34.3 | 22.2  | 81.2 | 58.3 | 59.3 | 34.6 | 31.9 | 73.3 | 50.0 | 49.2    | 27.2         |
| Ours        | 48.7 | 36.2 | 22.6  | 82.1 | 58.2 | 60.0 | 34.5 | 31.4 | 72.9 | 51.2 | 49.8    | 27.4         |

P1: pedestrian, P2: people, B1: bicycle, C: car, V: van, T1: truck, T2: tricycle, A: awning–tricycle, B2: bus, M: motor.



**Figure 9.** Some examples of detection results on the VisDrone dataset using YOLO-DCTI.

#### 4.3.3. Experiments on NWPU VHR-10 Dataset

Based on the results presented in Table 3, the proposed method exhibits robust performance on the NWPU VHR-10 dataset. Notably, our model achieves impressive AP@0.5 scores for the ten object categories: 99.6%, 93.0%, 96.8%, 99.5%, 90.9%, 95.0%, 99.2%, 91.5%, 98.9%, and 90.3%. Moreover, our model achieves a mAP@0.5 score of 95.5%. Figure 10 presents the detection results of the YOLO-DCTI model on the NWPU VHR-10 dataset, showcasing its efficacy in detecting targets of diverse scales.

To visually illustrate the effectiveness of the method, Figure 11 shows the detection results and grad-cam map of YOLO-DCTI on the NWPU VHR-10 dataset. These numbers provide convincing evidence of our model's ability to accurately identify object locations

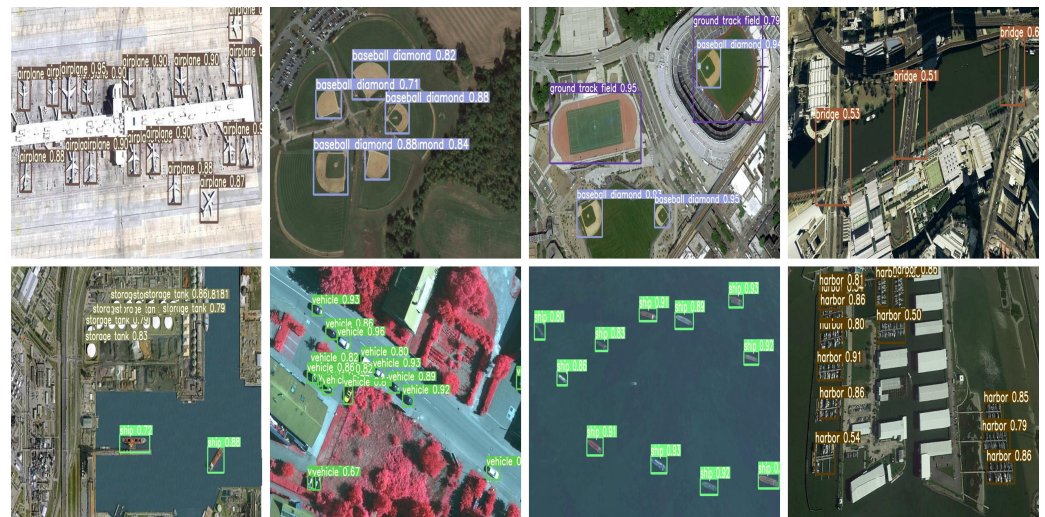


and assign appropriate attention to them. Overall, the experimental results highlight the robust performance of the proposed method in object detection tasks, thus indicating its potential in various real-world applications.

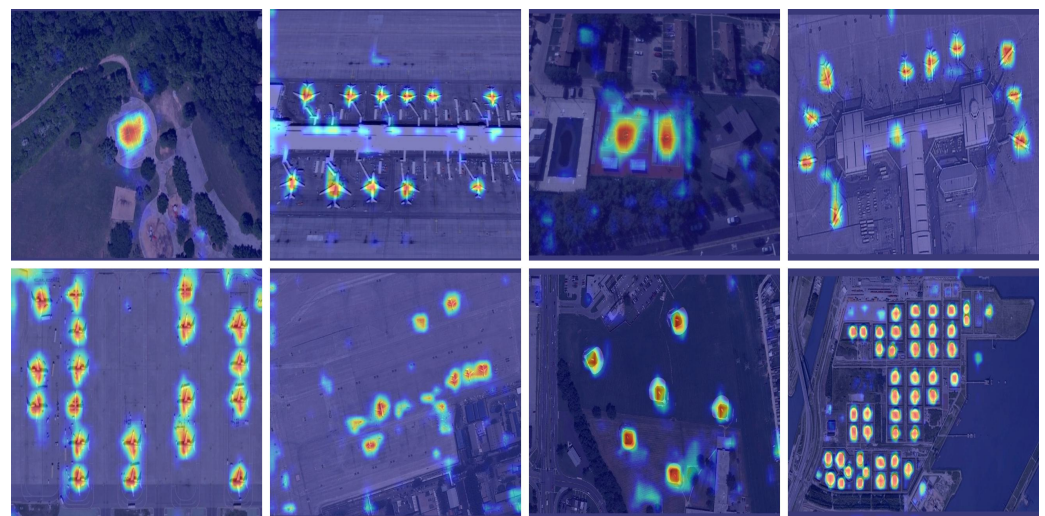
**Table 3.** Experimental results on NWPU VHR-10 dataset.

| Method       | A    | SH   | ST   | BD   | TC   | BC   | GTF  | H    | B    | V    | AP@0.5 |
|--------------|------|------|------|------|------|------|------|------|------|------|--------|
| SSD          | 90.4 | 60.9 | 79.8 | 89.9 | 82.6 | 80.6 | 98.3 | 73.4 | 76.7 | 52.1 | 78.4   |
| Faster R-CNN | 94.6 | 82.3 | 65.3 | 95.5 | 81.9 | 89.7 | 92.4 | 72.4 | 57.5 | 77.8 | 80.9   |
| YOLOv5s      | 99.2 | 84.1 | 97.5 | 99.4 | 89.5 | 79.1 | 97.6 | 74.0 | 82.1 | 80.7 | 88.3   |
| YOLOv5l      | 99.5 | 93.3 | 99.3 | 98.9 | 87.9 | 77.0 | 99.2 | 90.3 | 78.7 | 91.4 | 91.6   |
| YOLOv5x      | 99.3 | 90.7 | 99.6 | 98.5 | 89.5 | 86.3 | 99.3 | 83.7 | 83.4 | 91.2 | 92.2   |
| YOLOv7-tiny  | 98.8 | 89.0 | 98.9 | 99.2 | 85.8 | 68.8 | 98.3 | 83.6 | 72.0 | 77.5 | 87.2   |
| YOLOv7       | 99.5 | 91.6 | 99.4 | 98.9 | 90.1 | 96.8 | 99.4 | 86.7 | 94.1 | 90.4 | 94.7   |
| Ours         | 99.6 | 93.0 | 96.8 | 99.5 | 90.9 | 95.0 | 99.2 | 91.5 | 98.9 | 90.3 | 95.5   |

A: airplane, SH: ship, ST: storage tank, BD: baseball diamond, TC: tennis court, BC: basketball court, GTF: ground track field, H: harbor, B: bridge, V: vehicle, mAP: mAP@0.5:0.95.



**Figure 10.** Some examples of detection results on the NWPU VHR-10 dataset using YOLO-DCTI.



**Figure 11.** A selection of class activation maps exported using grad-cam on NWPU VHR-10 datasets.

#### 4.3.4. Comparison Experiment on Inference Speed and Model Parameters

To elucidate the equilibrium achieved by our YOLO-DCTI model concerning inference speed and detection performance, we conducted a comprehensive comparative analysis of

various models using the Dota-small dataset, as presented in Table 4. This analysis incorporates essential metrics such as mAP@0.5:0.95, inference speed, and model parameters. The evaluation employed test images with a resolution of  $640 \times 640$  and the inference speed was quantified in milliseconds (ms). Our model showcased a total of 37.67 M parameters, closely akin to the 39.46 M parameters of Faster RCNN. However, our model gained a notable advantage due to the absence of candidate box generation operations, leading to commendable inference speed. While our inference speed aligns with YOLOv5l and YOLOv7, the distinctive structural design of DCTI fortifies its capacity to effectively capture features of small targets. Although our approach may not outperform others in terms of model parameters and inference speed, it successfully achieves a favorable equilibrium between inference speed and detection accuracy.

**Table 4.** Inference speed and parameter comparison.

| Model        | mAP@0.5:0.95 | Inference (ms) | Parameter (M) |
|--------------|--------------|----------------|---------------|
| SSD          | 20.0         | 3.4            | 13.00         |
| Faster R-CNN | 45.9         | 26.5           | 39.46         |
| YOLOv5s      | 51.9         | 0.6            | 1.89          |
| YOLOv5l      | 60.7         | 4.4            | 47.08         |
| YOLOv5x      | 60.8         | 7.8            | 87.03         |
| YOLOv7-tiny  | 53.1         | 1.0            | 5.77          |
| YOLOv7       | 61.8         | 4.4            | 34.81         |
| Ours         | 65.2         | 4.7            | 37.67         |

#### 4.3.5. Ablation Experiment

The Dota-small dataset was used to conduct an ablation experiment aiming to investigate the impact of different structures on the final detection results. The obtained test results are presented in Table 5.

**Table 5.** Comparison of detection performance of different categories.

| Model           | Small Vehicle | Large Vehicle | Plane | Storage Tank | Ship | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 |
|-----------------|---------------|---------------|-------|--------------|------|-----------|--------|---------|--------------|
| Baseline        | 59.8          | 67.6          | 68.8  | 52.6         | 61.2 | 89.4      | 85.5   | 90.7    | 61.8         |
| Decoupled       | 60.6          | 71.7          | 73.2  | 54.2         | 63.7 | 90.8      | 85.5   | 91.1    | 64.7         |
| CoT             | 60.1          | 71.5          | 73.2  | 55.2         | 63.7 | 91.2      | 85.6   | 91.3    | 64.7         |
| Global-Residual | 60.8          | 71.6          | 73.7  | 55.0         | 63.6 | 89.7      | 87.4   | 91.3    | 64.9         |
| Local-Fusion    | 61.2          | 72.1          | 73.7  | 54.9         | 64.1 | 90.0      | 86.9   | 91.4    | 65.2         |

After incorporating CoT in the decoupled head, the mAP@0.5:0.95 value increased by 0.4%. CoT helps in identifying small or tiny objects by exploiting spatial context and global channel information. The addition of global residual structure and local fusion to CoT led to an improvement in mAP@0.5:0.95 by 1.0%. CoT-I further fuses spatial context and channel features, enabling the network to learn more about small object information, thereby enhancing detection performance. After incorporating the CoT-I structure in YOLOV7, mAP@0.5:0.95 increased by 1.5%, providing further evidence that CoT-I can enhance detection accuracy.

We conducted a comparative analysis of the model's performance regarding speed and parameters, as presented in Table 6. The results demonstrate that the incorporation of decoupled heads and CoT introduced a speed latency of 0.1 ms and 0.2 ms, respectively, in comparison to the baseline, along with an augmentation of 0.4 M and 2.86 M parameters. However, it is noteworthy that the inclusion of Global-Residual and Local-Fusion did not impose any discernible burden on the inference speed and parameter requirements.



**Table 6.** Inference speed and parameter comparison.

| Model           | Inference (ms) | Parameter (M) |
|-----------------|----------------|---------------|
| Baseline        | 4.4            | 34.81         |
| Decoupled       | 4.5            | 35.11         |
| CoT             | 4.7            | 37.67         |
| Global-Residual | 4.7            | 37.67         |
| Local-Fusion    | 4.7            | 37.67         |

We analyzed the kernel sizes utilized in CoT-I and present our findings in Table 7. In Equation (1), the kernel size, denoted as  $W_k$ , is specifically referred to as  $k = 3, 5$ , and  $7$ . Our analysis reveals that the model achieves the highest inference speed when employing a kernel size of  $k = 3$ . On the other hand, adopting a kernel size of  $k = 5$  results in improved detection accuracy, albeit with a certain trade-off in terms of inference speed. Notably, when utilizing a kernel size of  $k = 7$ , both the model's detection accuracy and inference speed significantly decrease. These observations suggest that expanding the perception range does not necessarily lead to performance enhancement.

**Table 7.** Comparison of improved contextual transformer parameters.

| Model   | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 | Inference (ms) | Parameter (M) |
|---------|-----------|--------|---------|--------------|----------------|---------------|
| $k = 3$ | 91.2      | 85.6   | 91.3    | 64.7         | 4.7            | 37.67         |
| $k = 5$ | 90.4      | 86.6   | 91.4    | 65.2         | 5.8            | 41.62         |
| $k = 7$ | 89.7      | 87.0   | 91.3    | 64.7         | 6.8            | 47.55         |

## 5. Conclusions

This research proposes the framework YOLO-DCTI for remote sensing small or tiny object detection based on YOLOv7 and an improved Context Transformer (CoT-I), which improves detection accuracy by mining and utilizing upper and lower spatial features and channel features. Specifically, we designed an efficient decoupled detection head structure DCTI by introducing CoT and embedding it into YOLOv7 to obtain long-term dependent features on channel and context spatial features. Furthermore, we introduce an innovative CoT variant, CoT-I, which incorporates a global residual structure and a local fusion structure. The global residual structure plays a critical role in merging and integrating information from various input levels, thereby yielding a more comprehensive and holistic feature representation. Similarly, the local fusion structure assumes a vital role in modeling the intricate interactions among neighboring pixels while leveraging the abundant channel-based information prevalent in remote sensing data. Extensive experiments demonstrate that our method can improve detection accuracy by mining and utilizing more features. Although the improvement in detection accuracy is accompanied by a slight loss of detection speed, the balance between detection accuracy and speed is crucial for remote sensing object detection, provided that the speed meets the application requirements. Despite the advancements in detection accuracy demonstrated by our model, occasional instances of missed detections persist in densely populated scenarios featuring small targets with similar features. In light of this, we aim to undertake more comprehensive research in our future endeavors to delve deeper into this matter.

**Author Contributions:** Conceptualization, L.M. and Z.F.; methodology, L.M. and Z.F.; software, L.M. and Z.F.; validation, L.M., Z.F. and Q.L.; formal analysis, L.M. and Z.F.; investigation, L.M. and Z.F.; resources, L.M. and Z.F.; data curation, L.M. and Z.F.; writing—original draft preparation, L.M. and Z.F.; writing—review and editing, B.W. and L.S.; visualization, L.M. and Z.F.; supervision, L.M. and Z.F.; project administration, L.M. and Z.F.; funding acquisition, L.M. and M.R.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China under Grant 62206221, the Fundamental Research Funds for the Central Universities, the Postdoctoral Science Foundation of China under Grant 2022M710393, and the Fourth Special Grant of China Postdoctoral Science Foundation (in front of the station) 2022TQ0035.

**Data Availability Statement:** Not applicable.

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

## References

1. Wang, B.; Zhao, Y.; Li, X. Multiple instance graph learning for weakly supervised remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5613112. [\[CrossRef\]](#)
2. Tong, K.; Wu, Y. Deep learning-based detection from the perspective of tiny objects: A survey. *Image Vis. Comput.* **2022**, *123*, 104471. [\[CrossRef\]](#)
3. Wu, Y.; Zhang, K.; Wang, J.; Wang, Y.; Wang, Q.; Li, Q. CDD-Net: A context-driven detection network for multiclass object detection. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 8004905. [\[CrossRef\]](#)
4. Liu, W.; Quijano, K.; Crawford, M.M. YOLOv5-Tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 is based on transfer learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8085–8094. [\[CrossRef\]](#)
5. Li, J.; Liang, X.; Wei, Y.; Xu, T.; Feng, J.; Yan, S. Perceptual generative adversarial networks for small object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1222–1230.
6. Xu, W.; Zhang, C.; Wang, Q.; Dai, P. FEA-swin: Foreground enhancement attention swin transformer network for accurate UAV-based dense object detection. *Sensors* **2022**, *22*, 6993. [\[CrossRef\]](#)
7. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
8. Cheng, G.; Lang, C.; Wu, M.; Xie, X.; Yao, X.; Han, J. Feature enhancement network for object detection in optical remote sensing images. *J. Remote Sens.* **2021**, *2021*, 9805389. [\[CrossRef\]](#)
9. Liu, J.; Yang, D.; Hu, F. Multiscale object detection in remote sensing images combined with multi-receptive-field features and relation-connected attention. *Remote Sens.* **2022**, *14*, 427. [\[CrossRef\]](#)
10. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L. Cross-scale feature fusion for object detection in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 431–435. [\[CrossRef\]](#)
11. Dong, Z.; Lin, B. BMF-CNN: An object detection method based on multi-scale feature fusion in VHR remote sensing images. *Remote Sens. Lett.* **2020**, *11*, 215–224. [\[CrossRef\]](#)
12. 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. [\[CrossRef\]](#)
13. Xu, G.; Song, T.; Sun, X.; Gao, C. TransMIN: Transformer-Guided Multi-Interaction Network for Remote Sensing Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *20*, 6000505. [\[CrossRef\]](#)
14. Li, Q.; Chen, Y.; Zeng, Y. Transformer with transfer CNN for remote-sensing-image object detection. *Remote Sens.* **2022**, *14*, 984. [\[CrossRef\]](#)
15. Xu, X.; Feng, Z.; Cao, C.; Li, M.; Wu, J.; Wu, Z.; Shang, Y.; Ye, S. An improved swin transformer-based model for remote sensing object detection and instance segmentation. *Remote Sens.* **2021**, *13*, 4779. [\[CrossRef\]](#)
16. Gong, H.; Mu, T.; Li, Q.; Dai, H.; Li, C.; He, Z.; Wang, W.; Han, F.; Tuniyazi, A.; Li, H.; et al. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. *Remote Sens.* **2022**, *14*, 2861. [\[CrossRef\]](#)
17. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
18. Bar, M. Visual objects in context. *Nat. Rev. Neurosci.* **2004**, *5*, 617–629. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Carbonetto, P.; De Freitas, N.; Barnard, K. A statistical model for general contextual object recognition. In Proceedings of 8th European Conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 350–362.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Part IV 14, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.
21. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7310–7311.
22. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1489–1500. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Song, G.; Liu, Y.; Wang, X. Revisiting the sibling head in object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11563–11572.

24. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A survey of visual transformers. *arXiv* **2021**, arxiv:2111.06091. [[CrossRef](#)]
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arxiv:2010.11929.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
27. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
28. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.
29. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Part I 16, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
30. Chen, T.; Saxena, S.; Li, L.; Fleet, D.J.; Hinton, G. Pix2seq: A language modeling framework for object detection. *arXiv* **2021**, arxiv:2109.10852.
31. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arxiv:2110.02178.
32. Mehta, S.; Rastegari, M. Separable self-attention for mobile vision transformers. *arXiv* **2022**, arXiv:2206.02680.
33. Wadekar, S.N.; Chaurasia, A. Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv* **2022**, arxiv:2209.15159.
34. Tong, H.; Peng, T.; Jiang, X. A Lightweight Risk Advertising Image Detection Method Based on Mobile-ViT. In Proceedings of the 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dalian, China, 11–12 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1249–1253.
35. Marefat, A.; Joloudari, J.H.; Rastgarpour, M. *A Transformer-Based Algorithm for Automatically Diagnosing Malaria Parasite in Thin Blood Smear Images Using MobileViT*; Technical Report; EasyChair: Manchester, UK, 2022.
36. 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.
37. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
38. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arxiv:1804.02767.
39. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arxiv:2004.10934.
40. Yolov5. Available online: <https://github.com/ultralytics/yolov5> (accessed on 15 March 2023).
41. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arxiv:2209.02976.
42. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
43. Hussain, M.; Al-Aqrabi, H.; Munawar, M.; Hill, R.; Alsoufi, T. Domain Feature Mapping with YOLOv7 for Automated Edge-Based Pallet Racking Inspections. *Sensors* **2022**, *22*, 6927. [[CrossRef](#)] [[PubMed](#)]
44. Zhao, H.; Zhang, H.; Zhao, Y. Yolov7-sea: Object detection of maritime uav images based on improved yolov7. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 233–238.
45. Jiang, K.; Xie, T.; Yan, R.; Wen, X.; Li, D.; Jiang, H.; Jiang, N.; Feng, L.; Duan, X.; Wang, J. An Attention Mechanism-Improved YOLOv7 Object Detection Algorithm for Hemp Duck Count Estimation. *Agriculture* **2022**, *12*, 1659. [[CrossRef](#)]
46. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10186–10195.
47. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arxiv:2107.08430.
48. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE Computer Society: Piscataway, NJ, USA, 2021; pp. 3490–3499.
49. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
50. 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 Workshops, Seoul, Korea, 27–28 October 2019.

51. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [[CrossRef](#)]
52. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arxiv:1710.09412.
53. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2918–2928.
54. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; IEEE: Piscataway, NJ, USA, 2006; Volume 3, pp. 850–855.
55. 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, Part I 14, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
56. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1497. [[CrossRef](#)]

**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.