*remote sensing*

MDPI

*Article*

# A Convolution with Transformer Attention Module Integrating Local and Global Features for Object Detection in Remote Sensing Based on YOLOv8n

**Kaiqi Lang [1,2], Jie Cui [1,2], Mingyu Yang [1], Hanyu Wang [1], Zilong Wang [1,2] and Honghai Shen [1,*]**

[1] Key Laboratory of Airborne Optical Imaging and Measurement, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; langkaiqi20@mails.ucas.ac.cn (K.L.); cuijie22@mails.ucas.ac.cn (J.C.); yangmingyu@ciomp.ac.cn (M.Y.); wanghanyu@ciomp.ac.cn (H.W.); wangzilong21@mails.ucas.ac.cn (Z.W.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: shenhh@ciomp.ac.cn

**Abstract:** Object detection in remote sensing scenarios plays an indispensable and significant role in civilian, commercial, and military areas, leveraging the power of convolutional neural networks (CNNs). Remote sensing images, captured by crafts and satellites, exhibit unique characteristics including complicated backgrounds, limited features, distinct density, and varied scales. The contextual and comprehensive information in an image can make a detector precisely localize and classify targets, which is extremely valuable for object detection in remote sensing scenarios. However, CNNs, restricted by the essence of the convolution operation, possess local receptive fields and scarce contextual information, even in large models. To address this limitation and improve detection performance by extracting global contextual information, we propose a novel plug-and-play attention module, named Convolution with Transformer Attention Module (CTAM). CTAM is composed of a convolutional bottleneck block and a simplified Transformer layer, which can facilitate the integration of local features and position information with long-range dependency. YOLOv8n, a superior and faster variant of the YOLO series, is selected as the baseline. To demonstrate the effectiveness and efficiency of CTAM, we incorporated CTAM into YOLOv8n and conducted extensive experiments on the DIOR dataset. YOLOv8n-CTAM achieves an impressive 54.2 mAP@50-95, surpassing YOLOv8n (51.4) by a large margin. Notably, it outperforms the baseline by 2.7 mAP@70 and 4.4 mAP@90, showcasing its superiority with stricter IoU thresholds. Furthermore, the experiments conducted on the TGRS-HRRSD dataset validate the excellent generalization ability of CTAM.

**Keywords:** object detection; remote sensing; Transformer; CNNs; contextual information



**Citation:** Lang, K.; Cui, J.; Yang, M.; Wang, H.; Wang, Z.; Shen, H. A Convolution with Transformer Attention Module Integrating Local and Global Features for Object Detection in Remote Sensing Based on YOLOv8n. *Remote Sens.* **2024**, *16*, 906. https://doi.org/10.3390/rs16050906

Academic Editors: Silvia Liberata Ullo and Li Zhang

Received: 10 January 2024
Revised: 1 March 2024
Accepted: 1 March 2024
Published: 4 March 2024



## 1. Introduction

With the advancement of remote sensing technologies, images captured by various crafts and satellites have an enormous quantity and high spatial resolution. These images contain significant information crucial for a wide range of applications, such as land planning, forest protection, traffic monitoring, disaster detection, and personnel rescue. Object detection plays a fundamental yet important role in remote sensing image processing. It can extract valuable information from images by localizing and classifying regions of interest. However, traditional object detection algorithms such as Histogram of Oriented Gradients (HOG) [1] and Scale-Invariant Feature Transform (SIFT) [2] rely on handcrafted features tailored to specific scenes, resulting in inferior efficiency, accuracy, and generalization.

In recent years, CNNs have rapidly revolutionized various fields in computer vision (CV), such as image classification, object detection, instance segmentation, and pose estimation. Object detection, as one of the primary tasks, is an indispensable component in industry detection, security surveillance, and autonomous driving. Since the success of

AlexNet [3], scholars and researchers have increasingly focused on the CNN-based detectors. These detectors have gradually surpassed traditional methods and taken a dominant position in object detection.

CNN-based detectors can be broadly categorized into two-stage and one-stage detectors. Two-stage algorithms treat the detection process as region proposal generation and proposal-wise prediction. The R-CNN series are the representative detectors in two-stage algorithms, including R-CNN [4], Fast R-CNN [5], Faster R-CNN [6], Mask R-CNN [7], and Cascade R-CNN [8], which have salient performance and wide applications in object detection. On the other hand, one-stage detectors directly produce dense predictions through CNNs, making them more efficient and suitable for various real-time applications. SSD [9], the YOLO series, and RetinaNet [10] are the well-known detectors in one-stage algorithms. The YOLO series is a big family of detectors including YOLOv1-v8 [11–18] and other variants, with the goal of achieving better and faster performance. RetinaNet employs ResNet [19] as the backbone and designs focal loss to alleviate the imbalance of hard and easy samples. In addition, to eliminate the limitations of anchor boxes, several anchor-free detectors have been proposed, such as FCOS [20], CenterNet [21], and RepPoints [22]. FCOS leverages the center of grids and down-sample strides to replace anchor boxes and incorporates the center-ness prediction to suppress low-quality boxes. CenterNet detects each object as top-left, bottom-right, and central keypoints and designs cascade corner pooling and center pooling to enrich features. RepPoints utilizes a set of keypoints to represent objects and automatically learns important information.

The models mentioned above are primarily designed for nature images and exhibit excellent performance on datasets like MS COCO [23]. However, remote sensing images, owing to the top-down view and long capturing distance, possess distinctive characteristics compared with nature images, including complicated backgrounds, limited features, distinct density, and varied scales, as illustrated in Figure 1. The contextual and comprehensive information in an image is beneficial for recognizing hard targets, which is extremely valuable in remote sensing scenarios. Nevertheless, with the limitation of local computation in the convolution operation, the detectors based on CNNs have local receptive fields and lack interaction among the data of distant positions in an image. To address this problem, researchers in the field of remote sensing have made substantial efforts and innovations. In [24], CBAM is employed to connect the backbone of YOLOv3 with an auxiliary network. RSADet [25], a two-stage CNN framework, introduces a scale attention module to fuse spatial and channel information. SRAF-Net [26] combines deformable convolution and context attention by designing a context-based deformable module. GRS-Det [27] utilizes Gaussian-Mask to enhance the perception of ships with contextual information. While these models manage to address the problem of lacking contextual information to some extent, there remains a scarcity of long-distance interactions in these CNN-based detectors. Moreover, unsupervised remote sensing image analysis methods such as Spatial-Spectral Masked Auto-encoder [28] and Nearest Neighbor-Based Contrastive Learning [29] show promising prospects in object detection.

Neck Attention Block (NAB) [30] is an effective block and achieves salient performance in small object detection. However, due to channel attention used in NAB, the improved model lacks sufficient global information for extracting and fusing important features. In this paper, we introduce the Convolution with Transformer Attention Module (CTAM) to enhance contextual and comprehensive information for CNN-based detectors, aiming at improving localization capacity and detection accuracy in remote sensing scenarios. CTAM is a novel plug-and-play attention module composed of two key components: a convolutional bottleneck block and a simplified Transformer layer. The convolutional bottleneck block is a bottleneck structure utilized to extract local features and retain position biases, and the simplified Transformer layer is designed to capture long-range dependency and provide global contextual information.

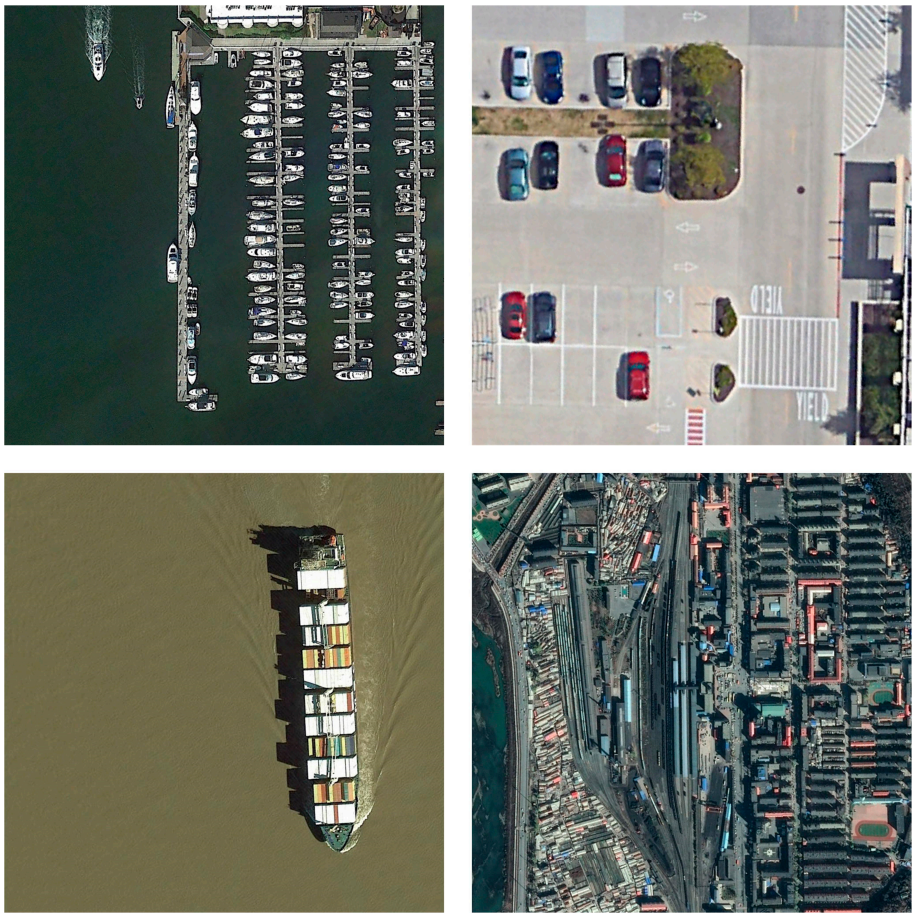

**Figure 1.** Various targets in remote sensing images.

To demonstrate the effectiveness and efficiency of CTAM, we selected YOLOv8n as the baseline, which achieves better performance while maintaining fast detection speed. We improved YOLOv8n with CTAM and conducted extensive experiments on the DIOR dataset [31]. YOLOv8n-CTAM surpasses the baseline by 2.8 mAP@50-95, with only a slight increase in detection time (0.2 ms). Notably, YOLOv8n-CTAM exhibits higher superiority with stricter IoU thresholds, such as mAP@70 and mAP@90, indicating CTAM makes the model focus on the central regions of targets and enhances localization capacity by integrating local features with global information. Compared with state-of-the-art detectors, it achieves cutting-edge performance while maintaining extremely fast speed. The results obtained on the TGRS-HRRSD dataset [32] further demonstrate the excellent generalization ability of CTAM.

The main contributions of this paper are as follows:

(1) We construct CTAM, a novel plug-in-play attention module, which effectively addresses the limitations of both CNNs and Transformer. CTAM facilitates the integration of local features and global contextual information and significantly enhances YOLOv8n's localization capacity.
(2) In contrast to the original Transformer applied in CV, we design a simplified Transformer structure by eliminating universal yet unnecessary operations for remote sensing scenarios, resulting in superior performance.
(3) We conducted extensive experiments on the DIOR and TGRS-HRRSD datasets, explicitly demonstrating the positive impact of CTAM. It improves localization capacity and exhibits noteworthy effectiveness, efficiency, and generalization ability.

The remainder of this paper is organized as follows: In Section 2, we provide an overview of related works concerning Transformer and the YOLO series. Section 3 offers

a detailed description of CTAM and the improved model. Section 4 presents the specific datasets used in our study and the experiments and analysis of CTAM. Finally, the conclusion is drawn in Section 5.

## 2. Related Works

### *2.1. The YOLO Series*

The primary goal of the YOLO series is to make object detection better, faster, and more scalable. YOLOv1, as the pioneering model, treats object detection as a regression problem. It segments an image into multiple grids, and each grid is responsible for predicting bounding boxes and class probabilities. With the improvements including a multi-scale training method, anchor boxes, and a new backbone, YOLOv2 achieves a tradeoff between accuracy and speed. YOLOv3 designs a stronger backbone called Darknet-53, implements multi-scale prediction, and incorporates various data augmentation techniques. YOLOv4 further optimizes the backbone and establishes CSPDarknet-53 based on CSPNet. Additionally, it introduces many methods to enhance the detection performance without increasing the inference cost. To accommodate various detection tasks and ease deployment, YOLOv5 designs five models with different computational costs. In order to mitigate the overfitting of the model, it proposes Mosaic, a novel data augmentation method, feeding a mixed image composed of several images into the model during training. YOLOX [33] switches YOLOv3 to an anchor-free manner with a leading label assignment strategy and constructs a decoupled head to suppress the conflict between regression and classification. For diverse platforms and applications, YOLOv6, an anchor-free detector like FCOS, constructs a fundamental CSPStackRep block for the backbone and adopts Task Alignment Learning [34] as the label assignment method. To enhance the capacity for learning and converging, YOLOv7 proposes Extended-ELAN to regulate the gradient paths. As the latest version in the YOLO series, YOLOv8 creates an anchor-free detector based on 'C2F' with more gradient paths. It constructs a decoupled head with Distribution Focal Loss [35]. It supports various CV tasks, including classification, detection, segmentation, pose estimation, and tracking.

Due to its exceptional accuracy, flexibility, and efficiency, the YOLO series has extensive and significant applications in remote sensing scenarios. For instance, aiming at identifying high-density targets in UAV aerial images, MS-YOLOv7 [36] combines the original model with the Swin Transformer unit and proposes a new pyramidal pooling module. To address the dense occlusion in low-resolution images from the TinyPerson dataset, TOD-YOLOv7 [37] appends a tiny object detection layer and designs a recursive gated convolution module. UAV-YOLOv8 [38] utilizes Wise-IoU as the box regression loss to enhance localization ability in the UAV scenario. For UAV object detection, [39] modifies YOLOv8 with Bi-PAN-FPN and improves 'C2F' with GhostblockV2. In [40], YOLOv7-TS devises a Feature Map Extraction Module to reduce information loss.

In the YOLOv8 variants, YOLOv8n stands out for its fast speed, salient performance, and flexible deployment, so we select it as the baseline for real-time object detection in diverse remote sensing scenarios. With the goal of enhancing global contextual information for CNN-based detectors, we propose CTAM to integrate long-range dependency with local features, and YOLOv8-CTAM achieves superior performance on the DIOR and HRRSD datasets among various detectors.

### *2.2. Transformer*

Transformer [41] was initially conceived to model long-range dependency and introduce parallel computation for natural language processing (NLP). The remarkable breakthrough achieved by Transformer in NLP has motivated many scholars and researchers to explore its potential applications in CV.

Transformer-based backbones have made enormous progress in fundamental CV tasks, such as image classification, object detection, and semantic segmentation. Vision Transformer (ViT) [42], a pioneering vanilla Transformer model, achieves competitive

performance in image classification compared with state-of-the-art CNNs. ViT converts from 2D images to sequence data through a process of flattening and mapping image patches, which are fed into standard Transformer encoders with position embeddings. It finally employs a multi-layer perception (MLP) head for category classification. In contrast, DeiT [43] introduces the token-based knowledge distillation method, aiming at reducing the reliance on large amounts of data while achieving better performance on ImageNet-1K. Swin Transformer [44] adopts a hierarchical structure similar to the CNN-based models. To restrict the computational complexity posed by high-resolution images, it computes self-attention in non-overlapping local windows instead of global dependency. Additionally, it allows cross-window communication via shifted windows. CSWin Transformer [45] designs horizontal and vertical stripes to calculate self-attention in parallel. In comparison to Swin Transformer, it expands the local receptive field while constraining the computational complexity. BiFormer [46] proposes a novel sparse self-attention called Bi-Level Routing Attention, which filters irrelevant key–value pairs and applies self-attention to the remaining pairs to alleviate the heavy computational burden and high memory usage of Transformer.

Treating object detection as a set prediction problem, DETR [47] utilizes a CNN-based backbone to extract features and adopts a Transformer encoder–decoder and a feed-forward network to obtain detection results. It avoids the heavy computational cost produced by Transformer-based backbones while capturing global contextual information. Building on the success of DETR, various DETR variants have been proposed, such as Deformable DETR [48], Conditional DETR [49], and Lite DETR [50].

In contrast to natural scenes, long-range dependency captured by Transformer is more significant for object detection in remote sensing. TPH-YOLOv5 [51] designs a Transformer prediction head instead of the original head and incorporates an additional scale for tiny objects. Moreover, it adopts CBAM to identify dense objects. As a result, TPH-YOLOv5 achieves superior performance on the VisDrone dataset. Lu et al. [52] select CSWin Transformer as the backbone and construct a hybrid patch embedding module and a slicing-based inference method for UAV image object detection. Based on Swin Transformer, Xu et al. [53] designed a local-perception backbone to improve small object detection.

In this paper, we assert that the original Transformer retains substantial potential in remote sensing scenarios. Consequently, we devise a simplified Transformer and incorporate it into CTAM to integrate global contextual information with local features. The feasibility of the simplified Transformer is demonstrated in Section 4.5.

## 3. Materials and Methods

### *3.1. CTAM*

With the limitations of the top-down view, long capturing distance, and complex interference, remote sensing images exhibit some characteristics that differ from nature images, including complicated backgrounds, limited features, distinct density, and varied scales. Global contextual and comprehensive information can help detectors recognize targets, which is exceptionally valuable for object detection in remote sensing scenarios. Nevertheless, typical CNN-based detectors, restricted by the nature of convolution operation, severely lack global interaction. To address this pivotal problem, we construct a novel plug-and-play module named CTAM, aiming at integrating global contextual information with local features. It is composed of two primary components: a simplified Transformer layer, responsible for capturing long-range dependency, and a convolutional bottleneck block, responsible for extracting local features and providing inductive biases for the other component.

As depicted in Figure 2, the simplified Transformer layer contains two reshape operations and an easier Transformer variant. Although many models in CV, such as ViT, apply a standard Transformer to various tasks, we assume that it could not be the optimal form for object detection in remote sensing. The experiments described in Section 4 indicate that, at least for CTAM, the original Transformer is not a suitable form. For this paper, the

simplified Transformer removes LayerNorm [54], Dropout, and GeLU [55] and utilizes single-head attention to compute self-attention. It can be broadly divided into two parts: multi-head attention (MHA) and multi-layer perception (MLP). To process 2D feature maps, we flatten the map $X \propto R^{H \times W \times C}$ into $X_t \propto R^{(H*W) \times C}$ to serve as the input for the simplified Transformer, where (H, W, C) represents the resolution of the feature map. The matrices of Query, Key, and Value are computed as

$$Q,\ K,\ V = \text{Split}(\ \text{Linear}(\ X_t)\ ) \tag{1}$$

'Linear' refers to a fully connected layer, and 'Split' is an operation that segments a matrix into chunks along the channel dimension. Q, K, and V maintain the same sizes as $X_t$. Then, the computation of self-attention is as follows:

$$\text{Attention} = \text{SoftMax}(QK^T/C^{0.5})\ V \tag{2}$$

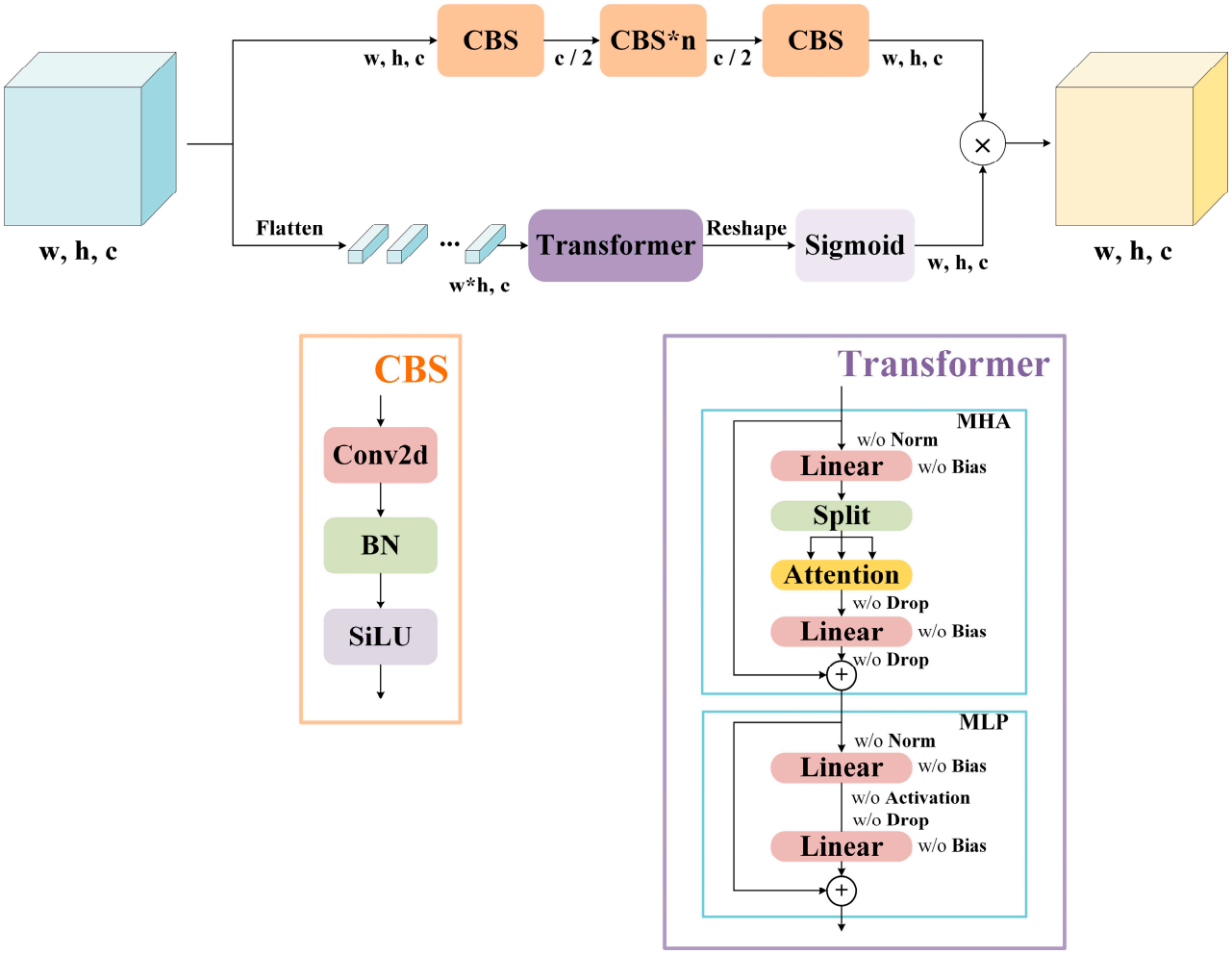


**Figure 2.** The structure of CTAM. Width, height, and channel are denoted as w, h, and c, respectively. 'w/o' is an abbreviation for 'with or without'. 'Norm', 'Drop', and 'Activation' represent LayerNorm, Dropout, and GeLU, respectively. 'Bias' is the bias in the linear layer.

The final output of MHA with a residual connection, denoted as $X_{sha}$ , can be expressed as

$$X_{sha} = \text{Linear}(\text{Attention}) + X_t \tag{3}$$

MLP is composed of two fully connected layers without GeLU in the first layer. The entire process can be defined as

$$X_{mlp} = \text{Linear}(\ \text{Linear}(\ X_{sha}\ )\ ) + X_{sha} \tag{4}$$

where $X_{mlp}$ represents the output of MLP.

In summary, we express the output of the simplified Transformer layer $X_{tran}$ as

$$X_{tran} = Transformer(X) \tag{5}$$

Regarding the convolutional bottleneck block, it contains a stack of 'CBSs' (Conv-BN-SiLU), composed of one convolutional layer, Batch Normalization [56], and SiLU [57], as shown in Figure 2. The first and final 'CBSs' are used for channel compression and expansion. Those in between are employed for feature extraction and fusion. To balance the performance and computational cost of CTAM, we introduce the hyperparameter n to control the quantity of the middle 'CBS'. The result of the convolutional bottleneck block $X_{conv}$ can be computed as

$$X_{conv} = CBS_1 \ldots CBS_{n+2}(\mathrm{X}) \tag{6}$$

Inspired by NAB using traditional attention mechanisms to regulate the feature map after convolutional layers instead of the original map, we employ element-wise multiplication to fuse the features generated by the convolutional bottleneck block and the simplified Transformer layer. In this manner, we can acquire the features of each grid, guided by both global contextual and local information. CTAM is a complementary integration that introduces local concentration for Transformer and long-range dependency for the CNN. With the improvement of CTAM, YOLOv8n exhibits stronger localization capacity and better performance, as detailed in Section 4. Ultimately, the whole CTAM can be formally expressed as

$$X_{ctam} = X_{conv} \otimes X_{tran} \tag{7}$$

where $X_{ctam}$ and '$\otimes$' refer to the output of CTAM and element-wise multiplication, respectively.

### *3.2. YOLOv8n-CTAM*

YOLOv8, as one of the most cutting-edge models in the YOLO series, further boosts performance and flexibility across various tasks and applications. Like other common detectors, YOLOv8 can be divided into three components, backbone, path aggregation structure [58], and detection head. Figure 3 explicitly depicts the architecture of YOLOv8n-CTAM.

A preprocessed remote sensing image with the resolution of (800, 800, 3) serves as the input image and is fed into the backbone. Note that YOLOv8 proposes 'C2F' as the basic unit instead of 'C3' in YOLOv5, featuring more gradient paths. Through a sequence of 'Stage', we can obtain three feature maps with 8×, 16×, and 32× downsampling rates, respectively. Subsequently, these maps are sent into the path aggregation structure, composed of top-down and bottom-up paths. This structure aims to enhance localization information for the coarse maps and contextual information for the fine-grained maps. Finally, the detection head utilizes the augmented feature maps to predict the category and bounding box for each grid. To mitigate the conflict between classification and regression, YOLOv8 designs a decoupled head and adopts the general distribution to model bounding box representation.

YOLOv8 develops some variants with different widths and depths for various applications. YOLOv8n acquires the fastest detection speed and the smallest memory usage by decreasing its width and depth. Therefore, we select YOLOv8n as the baseline to satisfy the requirement of real-time detection. CTAM is inserted between the path aggregation structure and the detection head to integrate global contextual information with local features for object detection in remote sensing scenarios. The visualization results in Section 4 adequately demonstrate the effectiveness of CTAM.

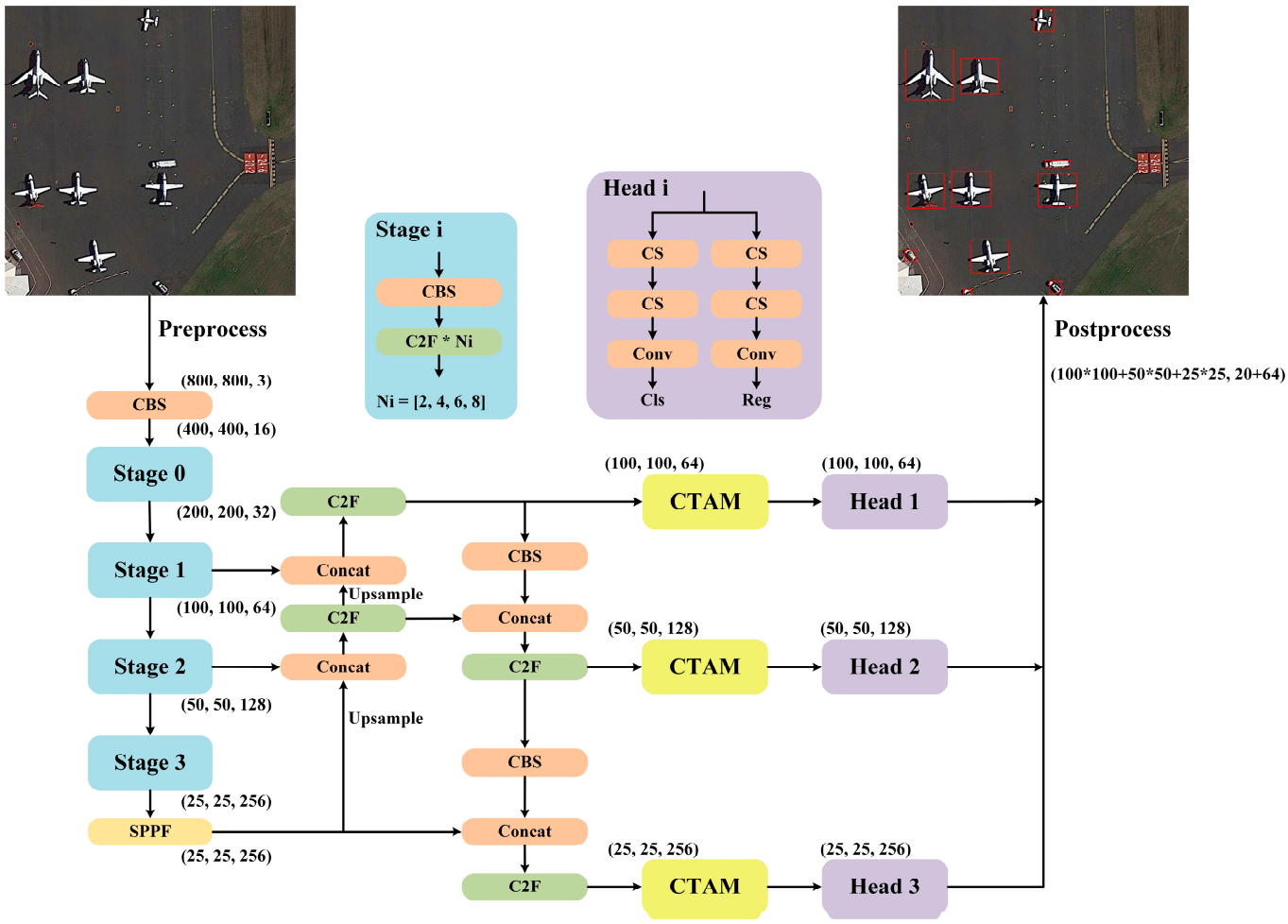


**Figure 3.** The architecture of YOLOv8n-CTAM. Here, '(800, 800, 3)' represents the resolution of input images. 'Concat' is the operation that concatenates tensors along the channel dimension. 'SPPF' is the improved version of 'SPP' in YOLOv5. 'CS' denotes a convolutional layer with SiLU.

## 4. Results

### *4.1. Experimental Environment and Settings*

All experiments were carried out on a Linux operating system (Ubuntu 20.04) with an Intel(R) Core (TM) i9-10940X CPU and two Nvidia RTX-3090 GPUs for distributed training. The deep learning framework was Pytorch 1.13 based on Python 3.9.16, CUDA 11.7, and Torchvision 0.14.1.

Hyperparameter settings play a significant role in the training process and greatly impact the final detection accuracy. To ensure a fair comparison, each model in this paper adopted the same hyperparameters outlined in Table 1. 'Image size' is the resolution of input images, restricting the sizes of targets and computational cost. 'Epoch' represents the number of iterations that a detector is trained on a dataset. Appropriate epochs make a model achieve excellent performance while saving computational resources. 'Learning rate', 'Momentum', and 'Weight decay' regulate the convergence rate and training stability. 'Mosaic' is a valuable measure for alleviating data overfitting. In addition, 'n (#CBS)', utilized to control local feature extraction, is introduced into the convolutional bottleneck block of CTAM. According to the experiments in Section 4.5, YOLOv8n is improved with CTAM (n = 2) to integrate long-range dependency with local features.

**Table 1.** Hyperparameter settings.

| Hyperparameter | Details |
|---|---|
| Image size | (800, 800) |
| Epoch | 100 |
| Batch size | 16 |
| Learning rate | 0.01~0.0001 |
| Momentum | 0.937 |
| Weight decay | 0.0005 |
| Mosaic | True (close at the last 10 epochs) |
| n (#CBS) | 2 |

### *4.2. Evaluation Metrics*

To evaluate the effectiveness and efficiency of CTAM, we adopt common metrics in object detection, including precision, recall, average precision (AP), mean average precision (mAP), model parameters, FLOPs, and detection time. Precision denotes the proportion of true positive samples among the total positive samples, and recall measures the proportion of true positive samples among the total true samples. The AP value for each category is obtained by calculating the area under the precision–recall curve, and mAP denotes the mean of AP values across all categories. The AP and mAP can be expressed as

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

$$mAP = \frac{1}{nc}\sum_1^{nc} AP_i \tag{9}$$

where 'P(R)' denotes the precision–recall curve and 'nc' represents the number of categories.

To evaluate the performance of the detector more comprehensively and accurately, we utilized different IoU thresholds to acquire corresponding mAP values. A higher threshold signifies a more rigorous criterion for the overlaps between bounding boxes and ground truth boxes. Specifically, mAP@50 represents an mAP value computed with an IoU threshold of 0.5. mAP@50-95 is the average of the mAP values under the IoU thresholds between 0.5 and 0.95, with a step of 0.05. To explicitly verify the localization capacity of CTAM, mAP@50, mAP@70, mAP@90, and mAP@50-95 were adopted as the evaluation criteria in the next experiments.

### *4.3. Datasets*

DIOR is a large-scale, diverse, and publicly available remote sensing dataset containing 23,463 images and 192,472 instances. It is divided into three subsets: a training set (5862 images), a validation set (5863 images), and a test set (11,738 images). It has 20 categories: airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, harbor, golf course, ground track field, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill. Each image in DIOR is standardized to the resolution of (800, 800, 3). The sizes of bounding boxes range from 2 to 764 pixels, posing a considerable challenge for object detection on the DIOR dataset. Each object in DIOR is annotated with a horizontal bounding box. In comparison to VEDAI, HRSC2016, and COWC, DIOR has more images and instances, which is beneficial for the robustness and generalization of detectors. For this paper, we conducted extensive experiments on the DIOR dataset to demonstrate the efficiency and effectiveness of CTAM.

Aiming at validating the generalization ability of CTAM, experiments were conducted on the TGRS-HRRSD dataset, another large-scale remote sensing dataset. This dataset possesses 21,761 images categorized into 13 classes, and the mean scale per class ranges from 42 to 277 pixels. Furthermore, it elaborately balances the number of each category. The comprehensive results are detailed in the next section.

### *4.4. Experiments on the DIOR Dataset*

To testify to the efficiency and effectiveness of CTAM, we initially trained YOLOv8n on the training and validation sets of the DIOR dataset with 100 epochs and evaluated its performance on the test set. For a fair comparison, we improved YOLOv8n with CTAM (n = 2) using the same settings and strategies. The experimental results for all categories are documented in Table 2. YOLOv8n-CTAM achieves 84.6 precision, 68.5 recall, and 54.2 mAP@50-95. It outperforms the baseline by a large margin, indicating the effectiveness of CTAM. 'Time' represents the total time, including preprocessing, inference, and post-processing time on an NVIDIA RTX 3090 with a batch size of 16. Due to the calculation of global contextual information occurring in the feature maps with 8×, 16×, and 32× strides, YOLOv8n-CTAM has a slight growth in 'FLOPs', 'Param', and 'Time'. It remains an extremely lightweight detector meeting the real-time requirement. Furthermore, with an increasing IoU threshold, YOLOv8n-CTAM achieves progressively better performance, surpassing the baseline by 1.8 mAP@50, 2.7 mAP@70, and 4.4 mAP@90. These results provide substantial evidence that CTAM can enhance localization capacity and detection performance by introducing global contextual information.

**Table 2.** The experimental results of YOLOv8n and YOLOv8n-CTAM on the DIOR dataset for all categories.

| Model | Precision | Recall | mAP@50 | mAP@70 | mAP@90 | mAP@50-95 | FLOPs | Param | Time |
|---|---|---|---|---|---|---|---|---|---|
| YOLOv8n | 82.7 | 67.2 | 74.7 | 62.5 | 21.9 | 51.4 | 8.1 G | 3.0 M | 2.1 ms |
| YOLOv8n-CTAM | 84.6 | 68.5 | 76.5 | 65.2 | 26.3 | 54.2 | 11.2 G | 4.7 M | 2.3 ms |

Table 3 presents the performance of the baseline and YOLOv8n-CTAM across each category. In almost all classes, YOLOv8-CTAM displays higher accuracy, recall, and mAP@50-95 compared with the baseline, especially in golf field detection, where it exceeds the baseline by 5.0 precision, 4.4 recall, and 11.4 mAP@50-95. Apparently, we can confirm that CTAM is beneficial for multi-scale target detection in remote sensing scenarios. Furthermore, we analyze the training processes of both detectors, as shown in Figure 4. YOLOv8n-CTAM exhibits a faster convergence rate in both regression and classification loss. Notably, the loss curves of both models show a rapid decline in the last 10 epochs, indicating that closing Mosaic in the last 10 epochs can lead to an enhancement in the final performance.

**Table 3.** Comparison of YOLOv8n and YOLOv8n-CTAM on the DIOR dataset across each category.

| Category | YOLOv8n | | | YOLOv8n-CTAM | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | AP@50-95 | Precision | Recall | AP@50-95 |
| Airplane | 95.3 | 64.1 | 54.4 | 96.8 | 65.5 | 57.1 |
| Airport | 78.2 | 79.2 | 54.1 | 79.7 | 82.1 | 60.1 |
| Baseball field | 95.3 | 69.2 | 68.3 | 95.7 | 71.4 | 69.2 |
| Basketball court | 91.7 | 85.1 | 75.3 | 93.4 | 85.4 | 76.8 |
| Bridge | 72.3 | 35.8 | 24.6 | 73.4 | 36.9 | 27.0 |
| Chimney | 94.9 | 72.4 | 64.2 | 95.3 | 73.0 | 65.6 |
| Dam | 63.4 | 63.5 | 35.4 | 70.5 | 65.4 | 40.4 |
| Expressway service area | 78.8 | 81.0 | 57.7 | 85.0 | 82.5 | 64.7 |
| Expressway toll station | 85.3 | 59.1 | 52.1 | 88.8 | 61.0 | 53.0 |
| Harbor | 74.0 | 58.3 | 45.1 | 76.9 | 59.6 | 47.6 |
| Golf field | 72.2 | 75.9 | 53.8 | 77.2 | 80.3 | 64.2 |
| Ground track field | 70.6 | 80.3 | 59.9 | 70.9 | 81.2 | 62.4 |
| Overpass | 82.0 | 52.3 | 40.8 | 84.0 | 54.1 | 43.1 |
| Ship | 92.2 | 85.4 | 55.8 | 93.5 | 85.4 | 56.7 |
| Stadium | 83.6 | 61.2 | 58.9 | 84.4 | 64.5 | 61.4 |
| Storage tank | 95.4 | 58.5 | 47.1 | 96.2 | 52.8 | 46.7 |
| Tennis court | 95.5 | 86.0 | 76.9 | 95.7 | 85.3 | 77.6 |

**Table 3.** *Cont.*

| Category | YOLOv8n | | | YOLOv8n-CTAM | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **AP@50-95** | **Precision** | **Recall** | **AP@50-95** |
| Train station | 59.4 | 59.1 | 30.8 | 59.2 | 64.6 | 35.1 |
| Vehicle | 87.3 | 37.2 | 28.9 | 88.5 | 36.9 | 29.4 |
| Windmill | 86.9 | 80.3 | 44.0 | 87.7 | 81.9 | 46.6 |

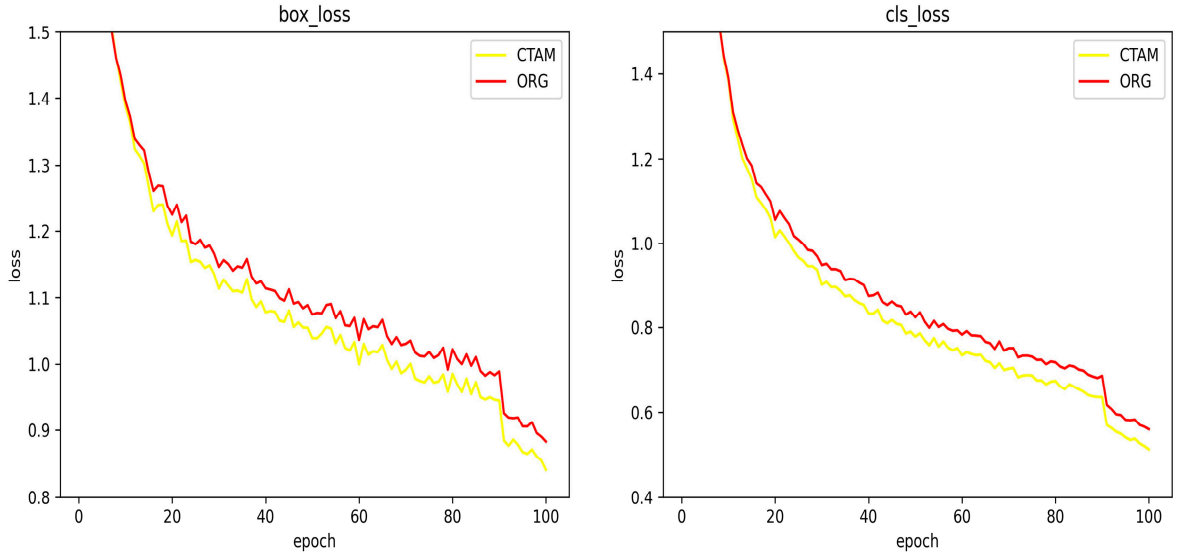


**Figure 4.** The loss curves of the baseline and YOLOv8n-CTAM. 'ORG' and 'CTAM' denote the original YOLOv8n and YOLOv8n-CTAM, respectively.

In comparison with the state-of-the-art detectors, YOLOv8n-CTAM achieves the most cutting-edge performance, as listed in Table 4. Specifically, the improved detector outperforms the well-known detectors in natural scenes by a large margin. In the field of remote sensing, it also surpasses SCRDet++ with ResNet-101 and CANet by 1.4 and 2.2 mAP@50, respectively. Above all, YOLOv8n-CTAM is a lightweight detector with an impressive 435 frames per second (FPS) on a single NVIDIA RTX 3090. YOLOv8n-CTAM demonstrates considerable potential for various applications and deployments in diverse remote sensing scenarios.

**Table 4.** The experimental results of different detectors on the DIOR dataset. '--' denotes difficult-to-obtain data.

| Detector | Year | mAP@50 | Parameters |
|---|---|---|---|
| Mask R-CNN | 2017 | 63.5 | 44.4 M |
| SSD | 2016 | 58.6 | 34.3 M |
| YOLOv3 | 2018 | 57.1 | 61.9 M |
| FCOS | 2019 | 69.4 | 32.3 M |
| CenterNet | 2019 | 69.4 | 32.3 M |
| DETR | 2020 | 68.6 | 54.7 M |
| SCRDet++ (ResNet-101) [59] | 2022 | 75.1 | -- |
| CANet [60] | 2020 | 74.3 | 64.6 M |
| YOLOv5n | 2020 | 72.0 | 1.9 M |
| YOLOv8n | 2023 | 74.7 | 3.0 M |
| YOLOv8n-CTAM | 2023 | 76.5 | 4.7 M |

Some detection results obtained by YOLOv8n-CTAM are displayed in Figure 5. YOLOv8n-CTAM successfully overcomes the challenges posed by remote sensing images, including complicated backgrounds, limited features, distinct density, and varied

scales. It achieves salient performance across various scenes and multi-scale categories. Although it may miss some targets or yield incorrect results in extremely hard situations, the acceptable detection accuracy with the remarkably low computational burden renders YOLOv8n-CTAM flexible and robust for deployment on real-time hardware platforms. In conclusion, the integration of global and local information within CTAM can compensate for inherent drawbacks in CNN and Transformer, leading to excellent localization capacity and detection accuracy in remote sensing images.

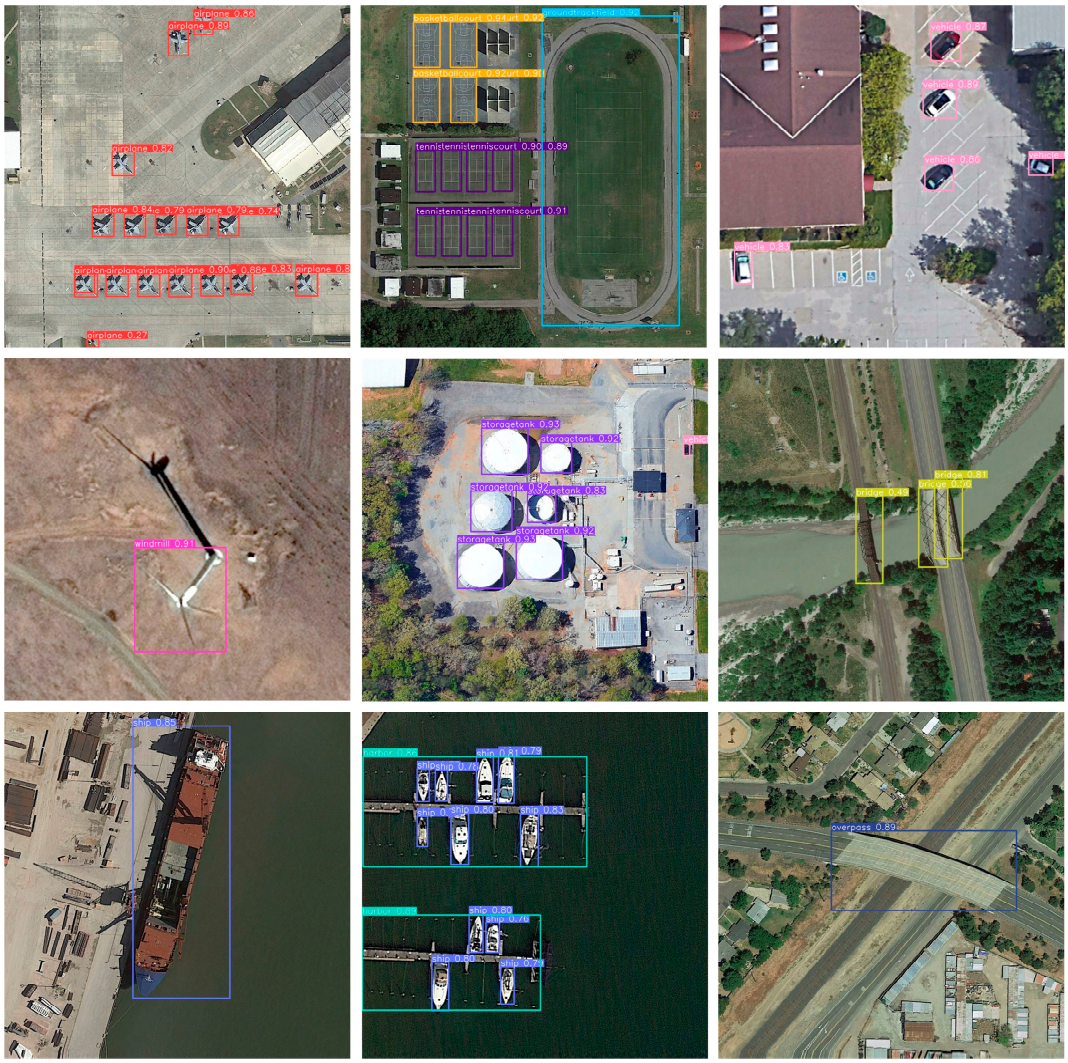

**Figure 5.** Detection examples of YOLOv8n-CTAM on the DIOR dataset.

### *4.5. Ablation Study*

#### 4.5.1. The Simplified Transformer in CTAM

Transformer is a powerful structure that can acquire long-range dependency by calculating scaled dot-product attention among all positions. To address the critical limitation of CNNs lacking global contextual information, we incorporate Transformer into CTAM to integrate local features with contextual and comprehensive information. Although the standard Transformer derived from NLP has been widely applied in CV, we assume that sequences and images have essential differences, and the standard Transformer can be optimized to achieve better performance for object detection in remote sensing scenarios. In this paper, we delve into a detailed analysis of the Transformer structure and construct a simplified Transformer layer in CTAM to accommodate the task of remote sensing object detection.

Extensive experiments for the optimal Transformer encoder structure were conducted on the DIOR dataset, as documented in Table 5. At first, 'Initial' has the worst mAP@50-95 among all Transformer variants. The absence of LayerNorm in 'A' results in an improvement of 0.5 mAP@50-95 compared with 'Initial', indicating that LayerNorm, widely applied in NLP, may hinder detection performance in remote sensing scenarios. Subsequently, the variants with different numbers of heads, specifically two and four, exhibit identical performance to the variant with '#Heads' = 1. Hence, we removed this hyperparameter and viewed it as a constant. Similarly, 'Dropout' has a negative impact on the detection accuracy, so it was set to 0. Finally, we eliminated the activation function 'GeLU' from MLP and constructed the simplified Transformer layer for CTAM. Moreover, the investigation of the influence of biases in MHA and MLP illustrates that the simplified Transformer encoder with both biases achieves the most salient performance on the DIOR dataset.

**Table 5.** Ablation study for the Transformer structures. 'Initial' and 'Simplified' denote the initial and final Transformer structures, respectively. 'A'–'G' are the Transformer variants with different hyperparameters. The number of heads in self-attention is represented as '# Heads'.

| Transformer | LayerNorm | GeLU | MHA bias | MLP bias | # Heads | Dropout | mAP@50-95 |
|---|---|---|---|---|---|---|---|
| Initial | √ | √ | √ | √ | 1 | 0 | 53.3 |
| A | | √ | √ | √ | 1 | 0 | 53.8 |
| B | | √ | √ | √ | 2 | 0 | 53.8 |
| C | | √ | √ | √ | 4 | 0 | 53.8 |
| D | | √ | √ | √ | 1 | 0.1 | 53.6 |
| E | | | | | 1 | 0 | 53.6 |
| F | | | | √ | 1 | 0 | 53.4 |
| G | | | √ | | 1 | 0 | 53.4 |
| Simplified | | | √ | √ | 1 | 0 | 54.2 |

#### 4.5.2. The Number of 'CBSs'

In the convolutional bottleneck block of CTAM, the number of 'CBSs' serves as a hyperparameter introduced to regulate the extraction of local features and restrict computational complexity, as depicted in Figure 2. We varied the value of 'n' within the range [0, 1, 2, 3], and the corresponding experimental results are listed in Table 6. YOLOv8n with CTAM (n = 2) achieves the best precision, recall, mAP@50, and mAP@50-95, illustrating that it adequately extracts and fuses local features while increasing negligible computational burden. Consequently, CTAM (n = 2) is considered as the default module for YOLOv8n due to its optimal performance.

**Table 6.** Ablation study for the number of 'CBSs' within the convolutional bottleneck block.

| # CBSs | Precision | Recall | mAP@50 | mAP@50-95 | Parameters |
|---|---|---|---|---|---|
| Baseline | 82.7 | 67.2 | 74.7 | 51.4 | 3.0 M |
| n = 0 | 83.8 | 68.0 | 75.7 | 53.0 | 4.3 M |
| n = 1 | 84.1 | 67.7 | 75.7 | 53.4 | 4.5 M |
| n = 2 | 84.6 | 68.5 | 76.5 | 54.2 | 4.7 M |
| n = 3 | 83.9 | 68.7 | 76.0 | 53.6 | 4.9 M |

#### 4.5.3. Comparison with Traditional Attention Modules

Traditional attention modules in CV such as SE [61], CBAM [62], and ECA [63] have widespread applications in CV. To enhance features and suppress noises, these modules utilize the information extracted from the feature map to recalibrate themselves. However, like NAB, we claim this way is inflexible and harmful for feature extraction. In contrast, CTAM utilizes the global information generated by the simplified Transformer layer to integrate with the local features of the convolutional bottleneck block. For a fair comparison, we replaced CTAM with SE, CBAM, and ECA in the same positions, and the corresponding ex-

perimental results are displayed in Table 7. Compared with the original model, SE, CBAM, and ECA are nearly useless in performance, but CTAM brings an obvious improvement.

**Table 7.** Ablation study for traditional attention modules.

| Model | Precision | Recall | mAP@50 | mAP@50-95 |
|---|---|---|---|---|
| Baseline | 82.7 | 67.2 | 74.7 | 51.4 |
| YOLOv8n-SE | 83.7 | 67.1 | 74.9 | 51.4 |
| YOLOv8n-CBAM | 83.1 | 66.9 | 74.7 | 51.2 |
| YOLOv8n-ECA | 82.4 | 66.9 | 74.2 | 51.0 |
| YOLOv8n-CTAM | 84.6 | 68.5 | 76.5 | 54.2 |
| YOLOv8n-CTAM (Local) | 83.0 | 67.9 | 75.3 | 53.1 |
| YOLOv8n-CTAM (Global) | 82.4 | 66.4 | 74.0 | 51.0 |

Moreover, YOLOv8n-CTAM (Local) and YOLOv8n-CTAM (Global) represent YOLOv8n-CTAM only with the convolutional bottleneck block and the simplified Transformer layer, respectively. Their results in Table 7 demonstrate that the integration of local features and global attention is indispensable and significant.

### *4.6. Visualization*

To further comprehend the influence of employing CTAM between the path aggregation structure and the detection head in YOLOv8n, we visualize the feature maps before and after the employment of CTAM, as depicted in Figure 6. The raw image contains two small-size vehicles and two medium-size airplanes. The detection results generated by YOLOv8n-CTAM exhibit highly accurate bounding boxes and reliable probabilities, demonstrating the effectiveness of CTAM. YOLOv8n is structured with three branches for multi-scale prediction, where the feature maps with 8×, 16×, and 32× downsampling rates are responsible for small-size, medium-size, and large-size targets, respectively. In the small-scale branch, the feature map behind CTAM exhibits higher and more centralized attention towards the two vehicles, compared with the map before CTAM. Meanwhile, in the medium-scale branch, the former feature map focuses on multiple parts surrounding the two airplanes, while the attention of the latter map converges on the centers of the airplanes. Since the responses mainly concentrate on the first two feature maps for these targets, the visualization and discussion of the large-scale branch are omitted.

This visualization provides valuable insights into how CTAM influences the feature extraction and fusion in YOLOv8n. The detailed comparison illustrates that CTAM enables YOLOv8n to focus on the central regions of targets and generate extremely accurate bounding boxes by integrating local features with global contextual information. This visualization corresponds with the conclusion that CTAM can significantly improve the localization capacity according to the mAP values with different IoU thresholds.

### *4.7. Experiments on the TGRS-HRRSD Dataset*

To validate the generalization ability of CTAM, we also conducted experiments on the TGRS-HRRSD dataset, a multi-scale remote sensing dataset containing 55,740 instances and 13 categories. For a fair comparison, we adopted the consistent hyperparameters and strategies used in the DIOR dataset and trained detectors on the train–validate set. As listed in Table 8, YOLOv8n-CTAM outperforms the baseline by 0.9 mAP@50 and 2.1 mAP@50-95. Compared to typical detectors, YOLOv8n-CTAM achieves a superior performance on the TGRS-HRRSD dataset while maintaining a rapid detection time. Compared with lightweight models, YOLOv8n-CTAM exceeds YOLOv4-Tiny and YOLOv4-Tiny-NAB by a large margin. Hence, these experiments indicate that CTAM is not limited to a specific dataset and exhibits excellent generation ability in various remote sensing scenarios.

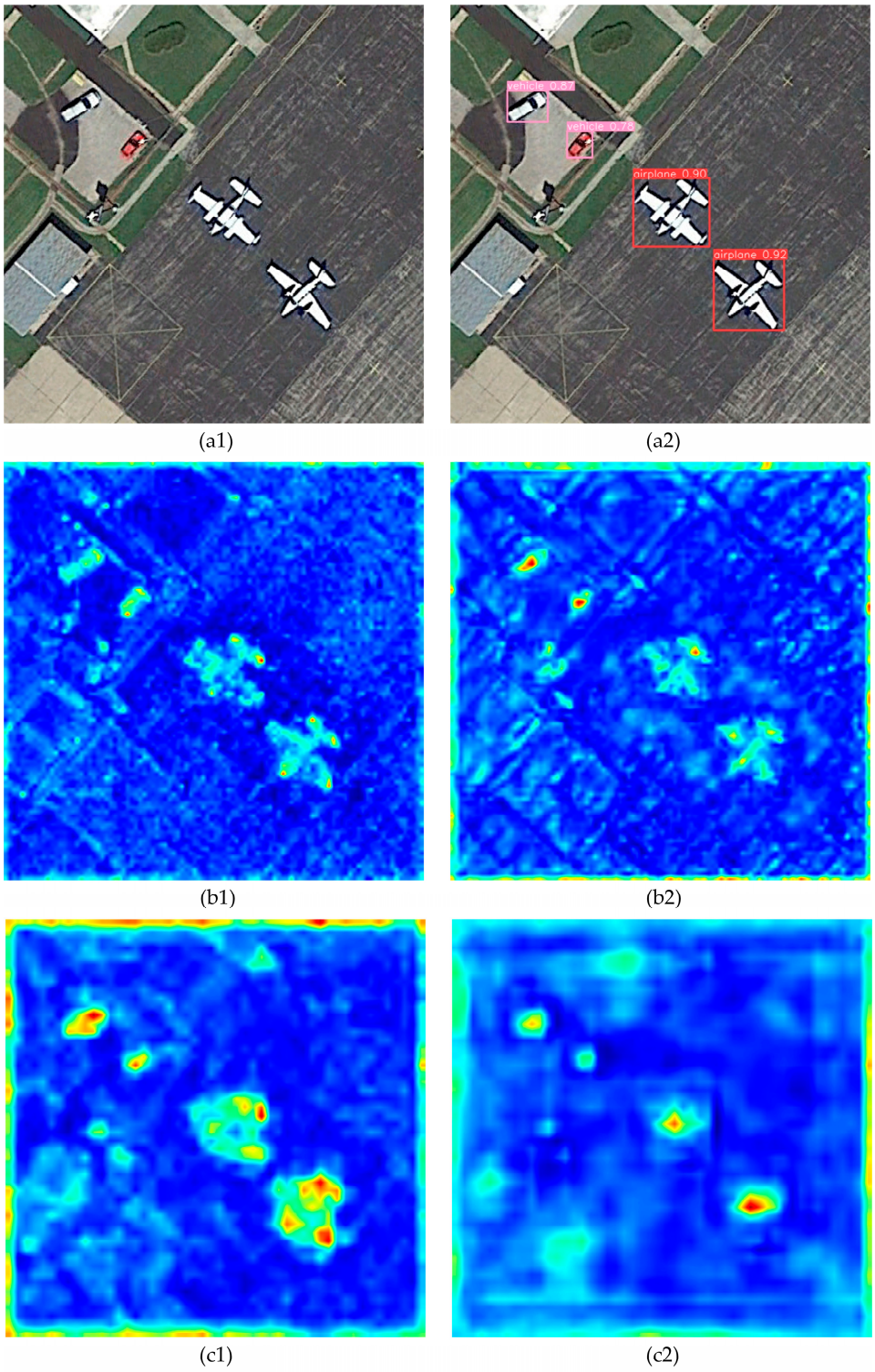


**Figure 6.** The visualization of YOLOv8n-CTAM. (**a1**,**a2**) display the raw image and detection results. (**b1**,**b2**) represent the feature maps before and after CTAM in the small-scale branch. Similarly, (**c1**,**c2**) denote the feature maps in the medium-scale branch. The visualization in the large-scale branch is omitted because both feature maps have scarce responses for the targets.

**Table 8.** The experimental results of different detectors on the TGRS-HRRSD dataset.

| Detector | Year | mAP@50 | Parameters |
|---|---|---|---|
| SSD | 2016 | 73.0 | 34.3 M |
| YOLOv3 | 2018 | 91.1 | 61.9 M |
| FCOS | 2019 | 86.9 | 32.3 M |
| YOLOv4-Tiny | 2020 | 83.4 | 5.9 M |
| YOLOv4-Tiny-NAB | 2022 | 87.2 | 7.1 M |
| YOLO-RS | 2023 | 87.7 | -- |
| AALFF | 2023 | 88.9 | 21.7 M |
| YOLOv8n | 2023 | 91.9 | 3.0 M |
| YOLOv8n-CTAM | 2023 | 92.8 | 4.7 M |

## 5. Conclusions

Remote sensing images have complicated backgrounds, limited features, distinct densities, and varied scales, rendering global contextual information extremely significant and valuable for object detection. However, CNN-based detectors with the limitation of local receptive fields have difficulty in capturing long-range dependency, resulting in inferior performance. To eliminate this inherent deficiency, we make the following contributions in this paper:

(1) We construct a novel plug-in-play attention module called CTAM, composed of a convolutional bottleneck block and a simplified Transformer layer. It can integrate local features with global contextual information through the interaction between the two components.
(2) We design a simplified Transformer in CTAM that is unlike the standard Transformer encoder widely applied in CV, and we demonstrate its validity in various remote sensing scenarios.
(3) For real-time object detection in remote sensing, we adopt YOLOv8n as the baseline and introduce CTAM to build YOLOv8n-CTAM. Extensive experiments demonstrate that YOLOv8n-CTAM achieves cutting-edge performance and generalization ability while maintaining an extremely rapid inference speed.
(4) The visualization of CTAM explicitly explains why CTAM can enhance localization capacity and improve detection accuracy by incorporating global information into local features.

Despite the remarkable performance and efficiency displayed by CTAM, it also brings unacceptable computational complexity and memory usage due to the defect of self-attention. In the future, we will optimize the computation of self-attention in CTAM and further explore the feasibility and flexibility of designing a backbone based on CTAM for object detection in remote sensing.

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

**Funding:** This research was funded by the National Key R&D Program of China, Grant No. 2022YFB3902300.

**Data Availability Statement:** All data used during the study have been uploaded at https://github.com/youjiaowuya/CTAM (accessed on 30 October 2023).

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

## References


1. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
2. Lowe, D.G. Object Recognition From Local Scale-Invariant Features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Corfu, Greece, 20–27 September 1999; pp. 1150–1157.
3. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
4. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection And Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
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. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
8. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving Into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single Shot Multibox Detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
10. 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.
11. 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.
12. 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.
13. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
14. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. Scaled-Yolov4: Scaling Cross Stage Partial Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13029–13038.
15. Yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 10 September 2023).
16. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
17. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: TRAINABLE BAG-OF-FREEBIES SETS NEW STATE-OF-THE-ART FOR REAL-TIME OBJECT DETECTORS. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
18. Yolov8. Available online: https://github.com/ultralytics/ultralytics (accessed on 10 September 2023).
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning For Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
20. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
21. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint Triplets For Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
22. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. Reppoints: Point Set Representation For Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9657–9666.
23. 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 Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
24. Qu, Z.; Zhu, F.; Qi, C. Remote sensing image target detection: Improvement of the YOLOv3 model with auxiliary networks. *Remote Sens.* **2021**, *13*, 3908. [CrossRef]
25. Yu, D.; Ji, S. A new spatial-oriented object detection framework for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
26. Sun, X.; Liu, Y.; Yan, Z.; Wang, P.; Diao, W.; Fu, K. SRAF-Net: Shape robust anchor-free network for garbage dumps in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6154–6168. [CrossRef]
27. Zhang, X.; Wang, G.; Zhu, P.; Zhang, T.; Li, C.; Jiao, L. GRS-Det: An anchor-free rotation ship detector based on Gaussian-mask in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3518–3531. [CrossRef]
28. Lin, J.; Gao, F.; Shi, X.; Dong, J.; Du, Q. SS-MAE: Spatial–Spectral Masked Autoencoder for Multisource Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [CrossRef]
29. Wang, M.; Gao, F.; Dong, J.; Li, H.-C.; Du, Q. Nearest Neighbor-Based Contrastive Learning for Hyperspectral and LiDAR Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]

30. Lang, K.; Yang, M.; Wang, H.; Wang, H.; Wang, Z.; Zhang, J.; Shen, H. Improved One-Stage Detectors with Neck Attention Block for Object Detection in Remote Sensing. *Remote Sens.* **2022**, *14*, 5805. [CrossRef]
31. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
32. Zhang, Y.; Yuan, Y.; Feng, Y.; Lu, X. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5535–5548. [CrossRef]
33. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
34. 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; pp. 3490–3499.
35. 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.
36. Zhao, L.; Zhu, M. MS-YOLOv7: YOLOv7 Based on Multi-Scale for Object Detection on UAV Aerial Photography. *Drones* **2023**, *7*, 188. [CrossRef]
37. Tang, F.; Yang, F.; Tian, X. Long-Distance Person Detection Based on YOLOv7. *Electronics* **2023**, *12*, 1502. [CrossRef]
38. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef]
39. Li, Y.; Fan, Q.; Huang, H.; Han, Z.; Gu, Q. A Modified YOLOv8 Detection Network for UAV Aerial Image Recognition. *Drones* **2023**, *7*, 304. [CrossRef]
40. Zhao, S.; Yuan, Y.; Wu, X.; Wang, Y.; Zhang, F. YOLOv7-TS: A Traffic Sign Detection Model Based on Sub-Pixel Convolution and Feature Fusion. *Sensors* **2024**, *24*, 989. [CrossRef]
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
43. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers & Distillation Through Attention. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10347–10357.
44. 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, Virtual, 18–24 July 2021; pp. 10012–10022.
45. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. Cswin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12124–12134.
46. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. BiFormer: Vision Transformer with Bi-Level Routing Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333.
47. 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.
48. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
49. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional Detr For Fast Training Convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3651–3660.
50. Li, F.; Zeng, A.; Liu, S.; Zhang, H.; Li, H.; Zhang, L.; Ni, L.M. Lite DETR: An Interleaved Multi-Scale Encoder For Efficient Detr. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18558–18567.
51. 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.
52. Lu, W.; Lan, C.; Niu, C.; Liu, W.; Lyu, L.; Shi, Q.; Wang, S. A CNN-Transformer Hybrid Model Based on CSWin Transformer for UAV Image Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1211–1231. [CrossRef]
53. 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]
54. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
55. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
56. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating Deep Network Training By Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
57. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **2018**, *107*, 3–11. [CrossRef]

58. 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.
59. Yang, X.; Yan, J.; Liao, W.; Yang, X.; Tang, J.; He, T. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2384–2399. [CrossRef]
60. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-layer attention network for small object detection in remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 2148–2161. [CrossRef]
61. Hu, J.; Shen, L.; Sun, G. Squeeze-And-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
62. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
63. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention For Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.