information

Article

A Novel Dynamic Contextual Feature Fusion Model for Small Object Detection in Satellite Remote-Sensing Images

Hongbo Yang [1,2] and Shi Qiu [1,*]

[1] Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China; yanghongbo@opt.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: qiushi@opt.ac.cn

Abstract: Ground objects in satellite images pose unique challenges due to their low resolution, small pixel size, lack of texture features, and dense distribution. Detecting small objects in satellite remote-sensing images is a difficult task. We propose a new detector focusing on contextual information and multi-scale feature fusion. Inspired by the notion that surrounding context information can aid in identifying small objects, we propose a lightweight context convolution block based on dilated convolutions and integrate it into the convolutional neural network (CNN). We integrate dynamic convolution blocks during the feature fusion step to enhance the high-level feature upsampling. An attention mechanism is employed to focus on the salient features of objects. We have conducted a series of experiments to validate the effectiveness of our proposed model. Notably, the proposed model achieved a 3.5% mean average precision (mAP) improvement on the satellite object detection dataset. Another feature of our approach is lightweight design. We employ group convolution to reduce the computational cost in the proposed contextual convolution module. Compared to the baseline model, our method reduces the number of parameters by 30%, computational cost by 34%, and an FPS rate close to the baseline model. We also validate the detection results through a series of visualizations.

Keywords: small object detection; satellite remote-sensing image processing; computer vision; deep learning

Citation: Yang, H.; Qiu, S. A Novel Dynamic Contextual Feature Fusion Model for Small Object Detection in Satellite Remote-Sensing Images. *Information* **2024**, *15*, 230. https://doi.org/10.3390/info15040230

Academic Editor: Marco Leo

Received: 26 March 2024
Revised: 15 April 2024
Accepted: 16 April 2024
Published: 18 April 2024

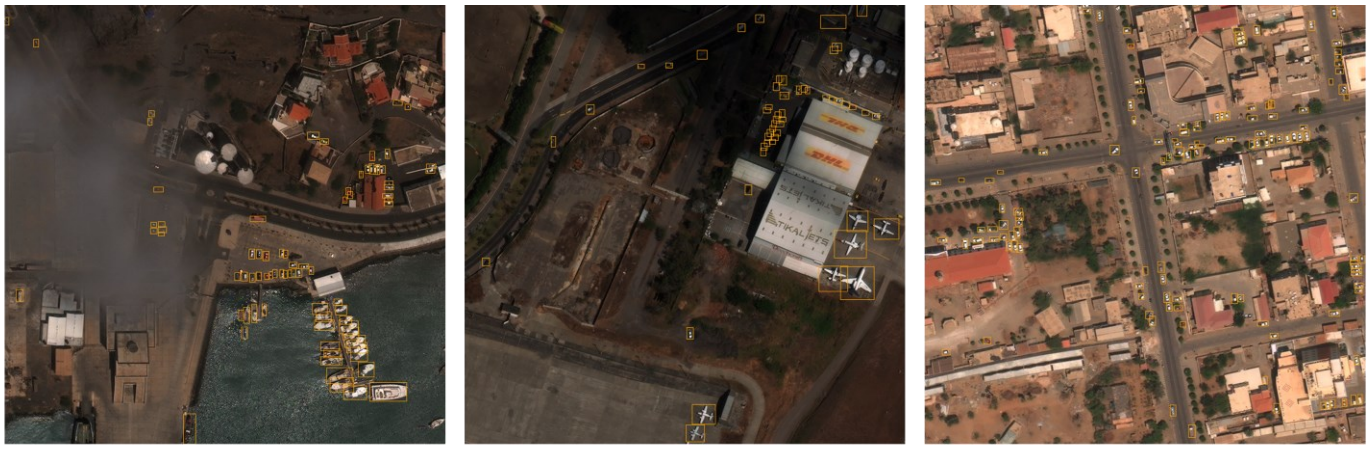
1. Introduction

With the rapid development of optical satellite remote-sensing technology, quantitative analysis and processing of satellite images are pivotal in the field of remote-sensing image interpretation. Remote-sensing image object detection, which involves identifying and locating objects of interest within images, is applied in various domains, including urban planning, disaster assessment, precision agriculture, and environmental observation. In recent years, high-resolution optical satellites have been able to capture sub-meter-level images and even videos of the Earth's surface. The total amount of data has expanded rapidly as a result. Convolutional neural networks [1] (CNNs), which can automatically extract image features and demonstrate strong generalization performance, have become a mainstream research focus.

Satellite remote-sensing images exhibit distinct characteristics, including small and densely distributed objects, the overhead perspective, complex scenes, and potential cloud and shadow interference. Research on object detection models commonly aims to improve the detection performance of natural images. The precision index improvement on the COCO Benchmark [2] serves as an effective reference for enhancing these models. However, the improvements may not work when applied to satellite images. It is specifically manifested in the following situation: previous research on remote-sensing object detection has achieved remarkable precision when dealing with common large-scale ground objects, such

as buildings and ports; however, when it comes to small and dense objects like vehicles, airplanes, and ships, the detection precision tends to be suboptimal.

As shown in Figure 1, satellite remote-sensing images exhibit distinct differences from natural images. When viewed from a vertical overhead perspective, the objects of interest, such as vehicles, airplanes, and ships, possess orientation characteristics—they can face any direction. These ground objects have small pixel sizes, and most of them belong to small objects, making their detection more difficult than common objects. These small objects lack prominent appearance and texture features, and some of them appear like mosaics. Furthermore, the dense distribution of objects brings additional challenges to object detection.



(a) ships (b) airplanes (c) vehicles

Figure 1. Examples of satellite images. We show the distribution of three categories of ground objects of interest, ships, aircraft, and vehicles in satellite images.

There are two mainstream object detection frameworks: two-stage and single-stage approaches. Due to the balance between precision and efficiency, the single-stage 'You Only Look Once' (YOLO) series [3–6] object detectors have attracted the most attention in the field of real-time object detection. YOLO directly predicts the category, location, and confidence of objects in the input images, removing redundant predictions by non-maximum suppression (NMS). However, YOLO is not effective at detecting small objects.

Context and multi-scale representation are the two pillars for small object detection. Therefore, we focus on these two key points to improve YOLO detection.

If the detector only focuses on local features within the bounding box, distinguishing ground objects from the background becomes difficult [7]. However, considering features of context around the objects makes correct detection easier, as shown in Figure 2. The car object within the orange bounding box is mosaic-like, with an abstract appearance and difficult to identify. As the range expands to the red box, it can be initially inferred that there is a car on the road within the orange box. Within the blue box, the car object can be easily detected. Therefore, contextual features play a crucial role in detecting small ground objects in satellite images.

The low-level feature maps represent local details in CNNs, while high-level feature maps convey more semantic information [8]. Multi-scale feature representation combines the local information from low-level feature maps and rich semantic information from high-level feature maps. A series of feature pyramids [9–13] are proposed to improve multi-scale detection. These studies focus on the cross-layer interaction of features and the effective flow of multi-scale information. Nevertheless, feature representation inside the layer is rarely considered. Without region proposals, the YOLO algorithm generates feature maps in which the features of the background constitute the majority but the feature representation of small objects is weak. Consequently, the features of small objects tend

to disappear as the layers go deeper. The feature representation of objects is the basis of multi-scale representation; in other words, it is necessary to pay attention to optimizing the intra-layer feature representation. The attention-guided feature fusion can help the model focus on features belonging to small objects and learn them effectively. Additionally, the representation of multi-scale contextual features can also be learned better.

(**a**) Object in bounding box (**b**) Object with little context (**c**) Object with lots of context

Figure 2. Context-assisted object detection. As the range expands, more context is involved, making detection easier.

In this paper, we improve the small object detector by focusing on contextual information and multi-scale feature fusion. We integrate a context extraction module into the backbone network to learn the reliance of objects on contextual features. We also improve the upsampler of the feature pyramid to enhance multi-scale feature interaction. Different from previous research, we realize that the optimization of intra-layer feature representation can not only help improve the learning of small object features, but also improve the detection combined with multi-scale context. Consequently, we employ the attention mechanism to guide multi-scale feature fusion. Based on the above improvements, the proposed framework is named 'Dynamic Contextual Feature Fusion'. In general, we propose a lightweight small object detection model for satellite images, integrating the proposed Dynamic Contextual Feature Fusion in the YOLO framework (DCFF-YOLO), which improves the detection precision while reducing the parameter quantity. The contributions of our work are summarized as follows:

(1) We proposed the lightweight Context Convolution Block to extract both local features and contextual features, using channel-wise group convolution to reduce the parameters and computations. In our work, the Context Convolution Block is designed to replace the C2f block in the YOLOv8 backbone network;
(2) We employ the DySample [9] module, a dynamic upsampler, to improve the upsample path of the multi-scale feature fusion;
(3) To help the model focus on the features of objects and learn the intra-layer feature representation of objects, we build the attention-guided feature fusion by integrating the Convolutional Block Attention Module (CBAM) [10] attention in the feature pyramid;
(4) We have applied the proposed DCFF-YOLO to satellite image datasets, and the evaluation shows that our research takes advantage of the average precision index and efficiency.

The structure of this paper is arranged as follows: Section 2 introduces the progress of related work; Section 3 describes the proposed method; Section 4 shows the experimental results; Sections 5 and 6 are the discussion and conclusion of the work.

2. Related Works

2.1. Object Detection in Satellite Remote-Sensing Images

The object detection frameworks can be categorized into two-stage and single-stage approaches. The two-stage object detector first extracts image features and generates region proposals using the CNN backbone network. Subsequently, the two-stage detector performs fine classification and regression within these region proposals. R-CNN [11]

innovatively extracts image features by CNN and generates region proposals. Fast R-CNN [12] employs ROI (Region of Interest) Pooling to obtain a fixed-length feature vector. Faster R-CNN [13] proposed the Region Proposal Network to replace Selective Search. FPN [14] introduced a feature pyramid network for multi-scale feature fusion. Mask R-CNN [15] proposes ROI Align to preserve the spatial position information of feature points in the maps. On the other hand, the single-stage object detector directly predicts the location and category of the objects without generating explicit region proposals. Due to the need for real-time performance, the single-stage 'You Only Look Once' (YOLO) series object detection models have attracted the most attention in the field of applied research. Cross Stage Partial Darknet [5] with 53 convolutional layers (CSPDarknet53) was designed and modified as the backbone network of YOLOv5. YOLOX [16] proposed the decoupled head to bring anchor-free detection to YOLO. The C2f (CSP bottleneck including 2 convolutional layers with shortcut) block replaces the C3 convolution block in YOLOv8's backbone. As shown in Figure 3, the typical YOLO framework comprises 3 main structures: backbone, neck, and head. The CSPDarknet53 is modified as the backbone network and extracts features from the input images. Three feature branches P3, P4, and P5 of the CSPDarknet53 are designed as inputs for the FPN structure to fuse multi-scale features in the neck. Then three detection heads corresponding to three branches decode features at different scales and predict bounding boxes of objects.

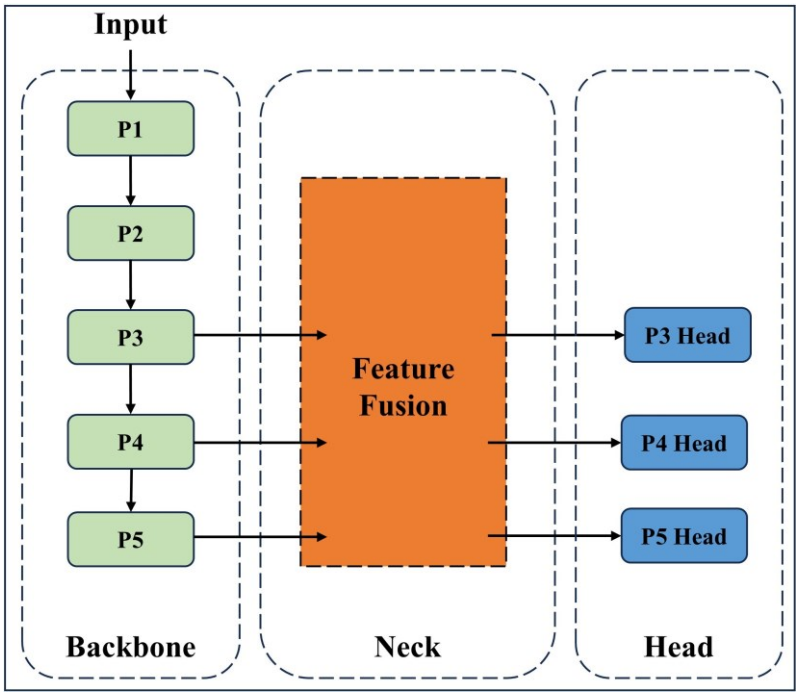


Figure 3. Overview of the YOLO framework. YOLO consists of three main structures: backbone to extract features, neck for feature fusion, and head to predict the boxes. P1~P5 are repeated convolution modules. As the layers go deeper, the corresponding receptive fields expand.

Object detection in remote-sensing images can further help city traffic observation and statistics of disaster damage tasks. Some studies learn the rotation-invariant features to improve object detection in remote-sensing images [17–19]. LSK-Net [20] extracts long-distance semantic features by large convolution kernel in remote sensing images. FFCA [21] proposes an efficient detector for small object detection in remote-sensing images based on YOLOv5. Zhang et al. [22] introduce few-shot learning into the remote-sensing image object detection task to alleviate the need for annotations. Due to the scarcity of large-scale publicly available satellite datasets, there is a relative lack of research on small object detection in satellite remote-sensing images. YOLT [23] uses sliding window slice input to

process large-scale remote-sensing images. YOLOS [24] adds a tiny object detection head corresponding to the P2 feature branch based on YOLOX.

2.2. Small Object Detection, Context Extraction, and Multi-Scale Feature Fusion

Small objects are typically defined as objects smaller than 32 × 32 pixels in the image. Due to the lack of features, small object detection is a challenging task [25,26]. In satellite remote-sensing images, a notable characteristic is distance consistency, which means all ground objects lie on a plane with approximately the same distance from the imaging sensor. Consequently, small objects constitute the absolute majority and are densely distributed in these images, significantly amplifying the difficulty of object detection in the context of satellite remote sensing.

Due to the subsampling and pooling process, deep CNN networks generate hierarchical feature maps corresponding to various spatial resolutions. In shallow-layer feature maps, small objects that are less than 32 × 32 pixels can be visible, but semantic features are lacking in these layers, which is critical for identifying the objects. Small objects usually exhibit lower resolutions and possess insufficient features. Consequently, the features of small objects tend to disappear in the deep-layer feature maps through downsampling.

Context extraction and multi-scale feature representation are two key points of small object detection. Context features can be divided into local contextual features and global contextual features (or semantic contextual features). Local contextual features reflect the association between objects and surrounding background pixels. On the other hand, global contextual features refer to the dependence of an object on the surrounding scene. Leveraging contextual features to improve small object detection is a widely adopted practice. For instance, FD-SSD [27] employs multi-branch dilated convolution to extract context features and improve the general detector. AC-FPN [28] develops an attention-guided context feature pyramid, introduces the Contextual Feature Extraction Module (CEM) into the FPN structure to extract contextual features, and the Attention guidance Module (AM) to discard redundant contextual features. However, AC-FPN [28] just extracts contextual in the top-level feature map of the FPN, and the multi-scale representation of contextual features is not considered. Large-kernel convolution with a larger receptive field is also used to extract the context reliance. LSK-Net [20] employs a set of convolution kernels of different sizes to extract features and achieves good performance in aerial imagery but suboptimal in satellite imagery, potentially reflecting suboptimization of small objects. LSKA [29] proposes a lightweight large-kernel convolution based on depth convolution and dilated convolution. The highlight of this study is the introduction of the attention mechanism into the kernel to optimize the intra-layer feature representation.

Different categories of objects rely on context to varying degrees. To qualitatively describe the dependence of the detection for vehicles, airplanes, and ships, we employ the dependency coordinate system shown in Figure 4, considering both their individual texture features and context features.

In the feature fusion step, the low-level feature maps and high-level feature maps are fused to combine local and semantic features, and further handle multi-scale detection. FPN [14] proposed the Feature Pyramid Network structure to fuse multi-scale feature maps from top to bottom. PANet [30] introduces two-way feature fusion, incorporating a bottom-up fusion way. BIFPN [31] and HSFPN [32] introduce distinctive feature fusion blocks for weighted feature fusion. ASFF [33] and AFPN [34] modify the pipeline of adjacent-layer fusion to achieve long-distance cross-layer fusion. Nonetheless, the above improvements focus on cross-layer feature interaction and rarely consider the intra-layer feature representation. The semantic features of small objects in remote-sensing images are weak and can easily vanish from deep layers after downsampling, which potentially leads to incorrect feature fusion. We have experimented with these fusion improvements and obtained unexpected performance degradation. The experiment results are described in the experiments and results section.

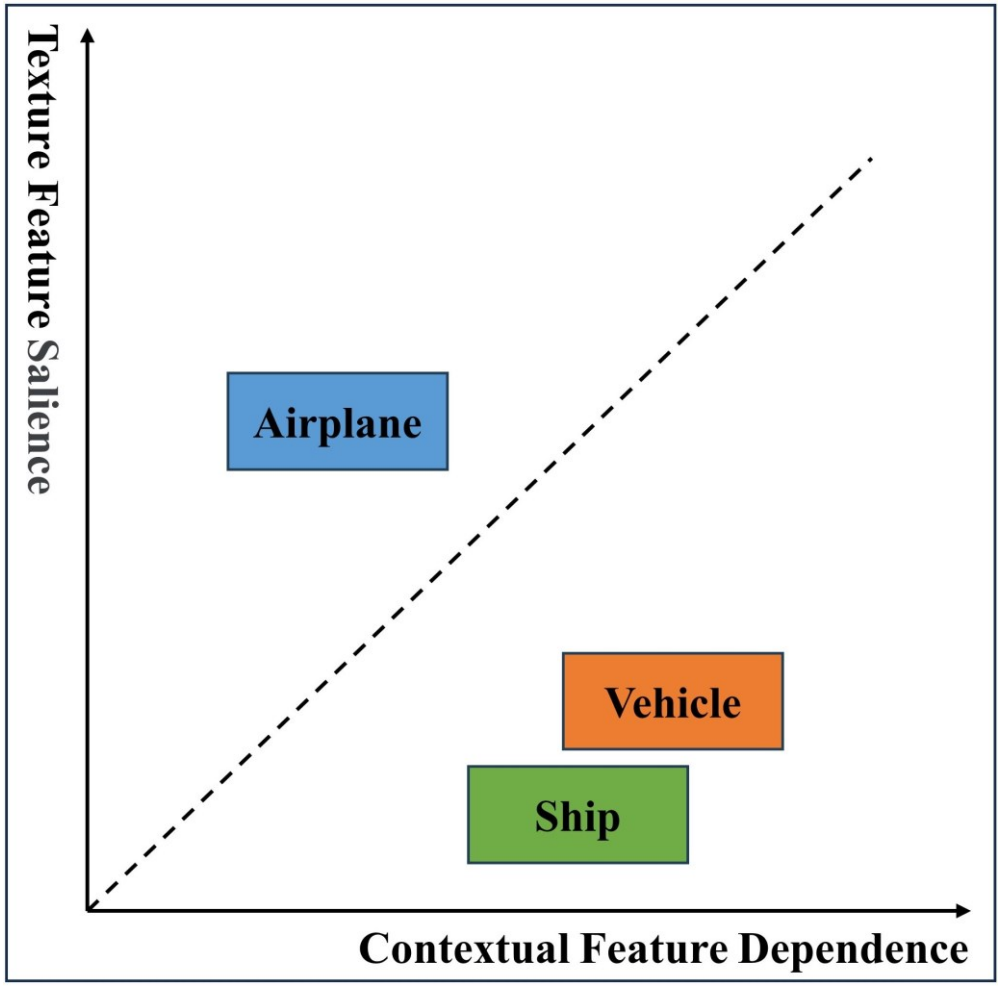


Figure 4. The dependency coordinate system that qualitatively describes the dependence of the detection for different classes. Airplanes have more prominent texture features, while the detection of ships and vehicles relies more on contextual features. Inshore ships are difficult to distinguish from the surrounding context, resulting in lower detection accuracy.

Low-level features are also incorporated into feature fusion to enhance small object detection. YOLOv8 provides a modified model that introduces the P2 feature layer into the fusion neck and adds a corresponding detection head. As a result, the model parameters and the amount of computation increase significantly. However, some categories of objects in remote-sensing images exhibit mosaic-like patterns and lack distinct texture features, so introducing the P2 layer features may not work.

3. Method

3.1. Overall Architecture

The overview of the proposed DCFF-YOLO framework is shown in Figure 5. We develop a novel lightweight dynamic contextual feature fusion model for small object detection tasks in satellite remote-sensing images. The framework consists of three main component networks that extract complementary features for small object detection: (1) the contextual backbone network containing the lightweight Context Convolutional Blocks, (2) the attention-guided feature fusion neck structure, and (3) the detection head.

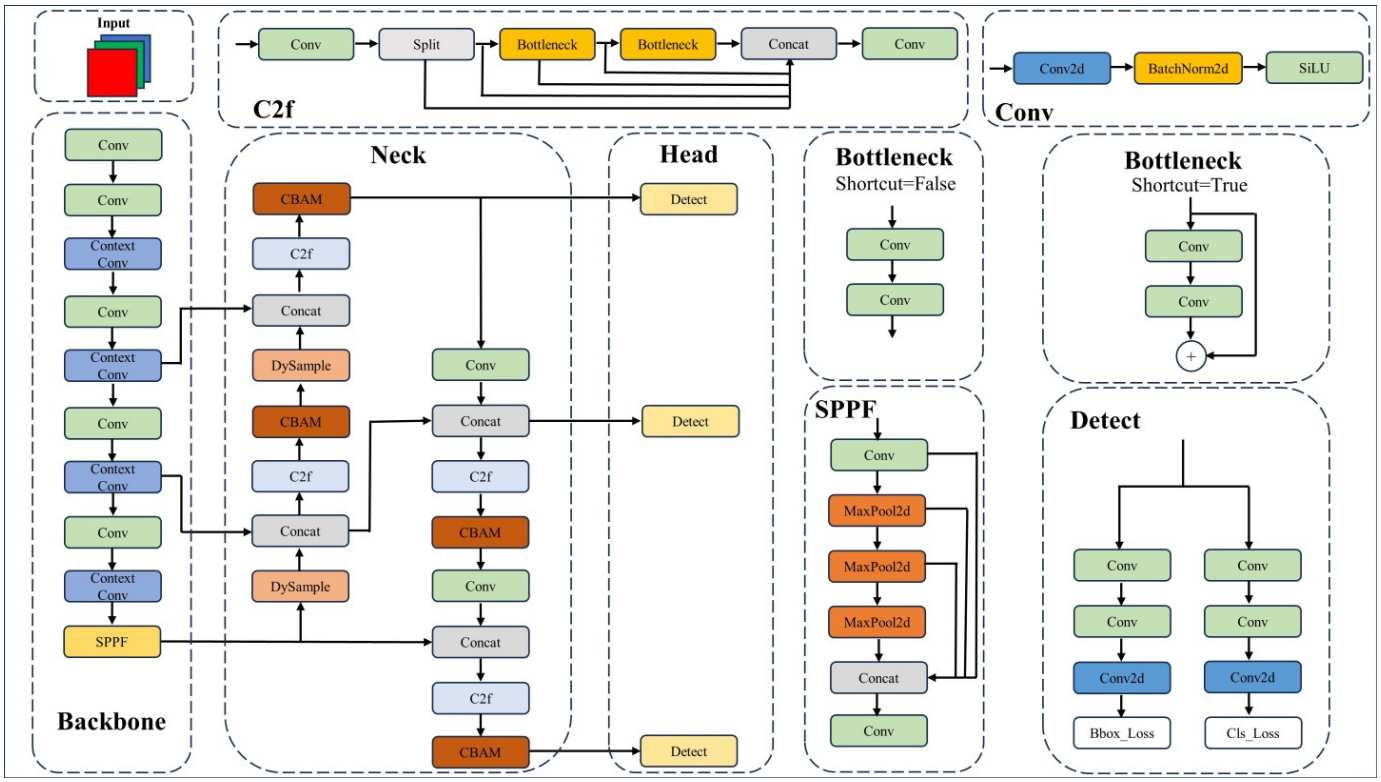


Figure 5. The overview of the proposed DCFF-YOLO model architecture. The framework is mainly comprised of the Contextual Convolution backbone, the attention-guided feature fusion neck, and the decoupled YOLO head.

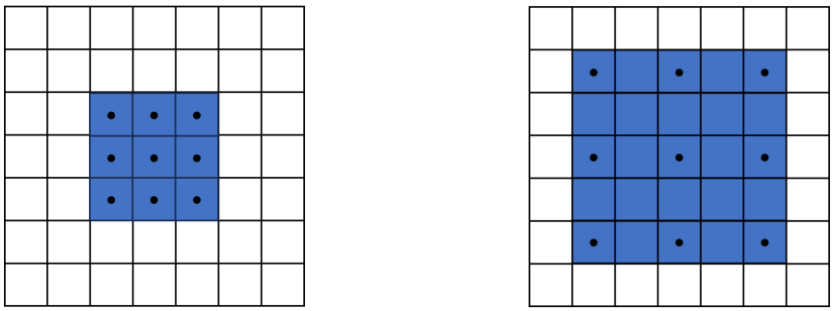
First, we propose the lightweight Context Convolution Block, which not only extracts local and contextual features but also utilizes channel-wise group convolution to reduce complexity. The Context Convolution Block is designed to replace the C2f block in the backbone network. In this paper, we place the Context Convolution Block after the P3, P4, and P5 layers, respectively. The integration of both local and surrounding context features yields more accurate detection maps. Secondly, we employ the DySample [9] block to replace the simple upsampling in the feature fusion neck structure. DySample, which is an ultra-lightweight and effective dynamic upsampler, can adaptively adjust the weight of the convolution kernel according to the high-level feature map input, and further enhance the dense detection. Thirdly, we introduce CBAM [10] attention in the neck structure to guide the detection head focus on potential object features. CBAM [10] is a lightweight attention mechanism that combines spatial domain and channel domain attention. Essentially, the attention mechanism decomposes features, further adjusting weights to improve the representation of key features and discard redundant features.

3.2. Context Convolution Block

Dilated convolution expands the receptive field by altering the spacing of the convolution kernel grid. In Figure 6a, a standard convolution employs a 3×3 convolution kernel, resulting in a receptive field size of 3. Alternatively, Figure 6b demonstrates that dilated convolution samples points on a broader pixel grid with a dilation rate of 2. The dilated convolution enlarges the receptive field to a size of 5, thereby extracting contextual information.

Although dilated convolution expands the receptive field, the number of convolution sampling points remains unchanged. In other words, dilated convolution involves sparse sampling. Therefore, the dilation rate within a certain range may be effective, or it can lead to discontinuity in sampling and irrelevance of long-distance sampling points. In satellite images, a small range of surrounding context is sufficient to improve the detection of small objects. However, an overly large receptive field may incorporate features from

other objects, especially in scenes with densely distributed objects. Based on the above considerations, a small and judiciously applicable expansion rate of 2 is used in this paper.



(**a**) Standard convolution (**b**) Dilated convolution

Figure 6. Standard and dilated convolution. Black dots represent the sampling points.

The structure of the Context Convolution Block is shown in Figure 7. Channel-wise grouped convolution is employed to reduce the computation. The block contains two branches: the local branch $f_{local}(\cdot)$ uses standard convolution to extract local features and the context branch $f_{context}(\cdot)$ uses dilated convolution to extract surrounding context features with a dilation rate of 2. The reliance of detection on local and contextual features is subject to variation in feature maps with receptive fields of different scales. Therefore, the features from both branches are concatenated to form the joint feature, which serves as the input for the Multi-Layer Perceptron (MLP). In this step, learnable weights are allocated to both local and contextual features.

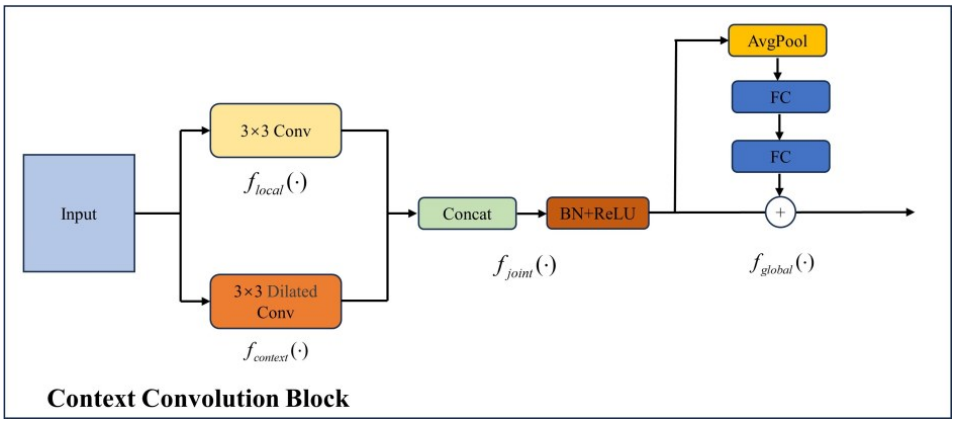


Figure 7. The structure of Context Convolution Block, which consists of local extractor $f_{\text{local}}(\cdot)$, surrounding context extractor $f_{\text{context}}(\cdot)$, joint feature extractor $f_{\text{joint}}(\cdot)$, and global extractor $f_{\text{global}}(\cdot)$.

By concatenating and regularizing the outputs of both branches, the joint feature $f_{\text{joint}}(\cdot)$ can be obtained.

$$\begin{aligned} f_{joint}(F) &= \text{ReLU}(\text{BatchNorm}(f_{local}(F), f_{context}(F))) \\ &= \text{ReLU}(\text{BatchNorm}(F_{local}, F_{context})) \end{aligned} \tag{1}$$

Equation (1) represents the calculation of the joint features, where F_{local} and $F_{context}$ represent the features obtained by the local feature extractor $f_{local}(\cdot)$ and the contextual feature extractor $f_{context}(\cdot)$, respectively, and ReLU denotes the activation function.

Finally, $f_{\text{global}}(\cdot)$ assigns learnable weights to local and contextual features by employing a pooling layer and two fully connected layers to refine the joint feature of both local features and the surrounding context features. The weighted feature is computed as:

$$\begin{aligned} f_{weighted}(F) &= \mathrm{MLP}^2(\mathrm{AvgPool}(\mathrm{ReLU}(\mathrm{BatchNorm}(f_{local}(F), f_{context}(F))))) \\ &= W_1 W_0(\mathrm{ReLU}(F_{local}^{Avg}, F_{context}^{Avg})) \end{aligned} \tag{2}$$

where W_0 and W_1 denote the weights of the two fully connected layers, and F_{local}^{Avg} and $F_{context}^{Avg}$ denote the output of F_{local} and $F_{context}$, respectively, after average pooling.

The global feature is computed as:

$$f_{global}(F) = f_{joint}(F) + f_{weighted}(F) \tag{3}$$

Instead of using the 1×1 pointwise convolution following depthwise convolution, we introduce an additional branch involving pooling and MLP layers to adjust the weight of local features and context features. In our experimental validation, 1×1 convolution can transmit information between channels, thereby breaking the independence of local features and surrounding context features and resulting in suboptimal detections.

3.3. Attention-Guided Feature Fusion

The FPN [14] and PAN [30] are combined as the feature fusion neck structure commonly known as PA-FPN in YOLO. An overview of the YOLO framework with the PA-FPN structure is shown in Figure 8.

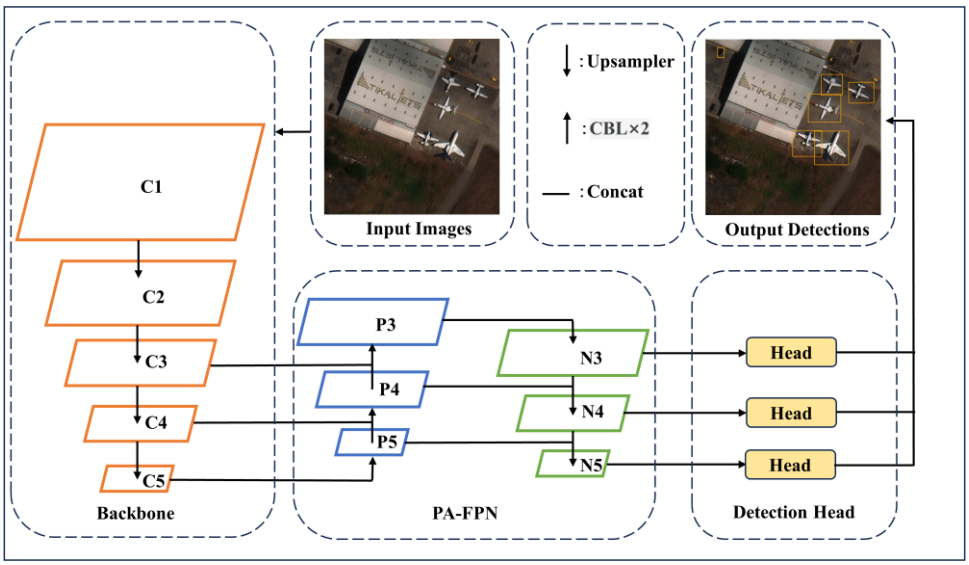


Figure 8. Overview of the YOLO framework with PA-FPN structure. The upsampler and downsampler are asymmetrical in the PA-FPN structure.

A series of structural diagrams, similar to Figure 8, conceals the asymmetry between the low-level feature downsampling and high-level feature upsampling processes. Specifically, the downsampler is learnable and relatively complex, whereas the upsampler is parameterless and very simple.

When focusing on the PA-FPN network, we observe that the low-level feature map undergoes size reduction through two repeated layers: Convolution+Batch Normalization+Activation Function. The layer is named Convolution Block Layer, abbreviated as CBL, which is a learnable layer. Upsampling is about modeling geometric information of images or feature maps. Notably, the PA-FPN network employs a straightforward nearest upsampling function to enlarge the scale of feature maps. This upsampling function is a simple linear mapping function.

Given an image, a binary function can be used to represent the mapping of pixel coordinates to pixel values:

$$v = f(x, y) \tag{4}$$

where v denotes the pixel value, (x,y) denotes the position of pixel, $f(\cdot)$ denotes the pixel value distribution.

When the image is upsampled by the ratio of s, the enlarged image is mapped to:

$$v = f(\left[\frac{x'}{s}\right], \left[\frac{y'}{s}\right]) \tag{5}$$

where (x',y') denotes the position of pixel in the enlarged image, and $[\cdot]$ denotes the rounding operator.

During the nearest upsampling process, the position and texture information of the feature points is partially lost. In the feature fusion step, it may lead to the misalignment of high-level semantic features and low-level local features, impeding the feature representation of small objects.

Dynamic upsampler is employed in the feature map upsampling process to adjust the upsampling parameters for the input feature map. This helps improve feature representation while maintaining an acceptable computational cost. DySample [9] is an ultra-lightweight and effective dynamic upsampler commonly used for image amplification. We introduce the DySample block to replace the naive upsampling function in the feature fusion step. The structure of DySample is shown in Figure 9.

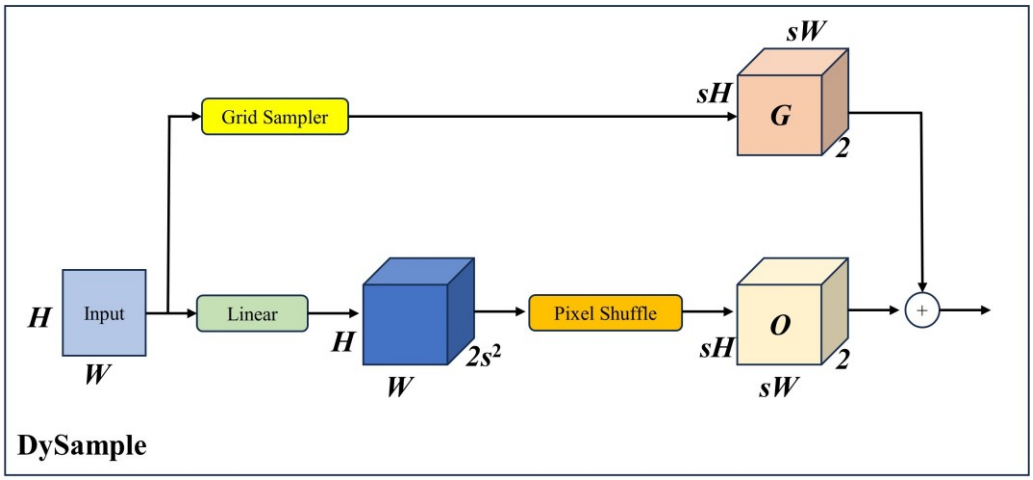


Figure 9. The structure of the DySample block. In the figure, s denotes the upsampling ratio, ***G*** denotes the grid sampling point coordinates, and ***O*** denotes the point position offsets generated by the dynamic sampling point generator.

Based on point sampling theory, DySample achieves dynamic convolution by adjusting the position offset of the sampling point, which is different from the kernel-based upsampler. At a complexity close to linear interpolation, it is designed to enhance the feature representation.

CBAM [10] (Convolutional Block Attention Module) is a lightweight and general attention module that combines channel and spatial attention mechanisms. Figure 10 shows the structure of CBAM attention. We integrate CBAM attention into the fusion neck for adaptive feature refinement.

The contextual convolution backbone extracts both local and contextual features, and transfers features into the fusion neck hierarchically. A sequence of layers in the depth of the network corresponds to spatial receptive fields of different scales, which means that detection at various levels relies on distinct dependencies related to contextual features and local features. The CBAM [10] attention assists in learning the importance weights of these contextual and local features, thereby helping the model to focus on the salient features

of objects. In this paper, CBAM is placed at the fusion neck, following the C2f module, as shown in Figure 5.

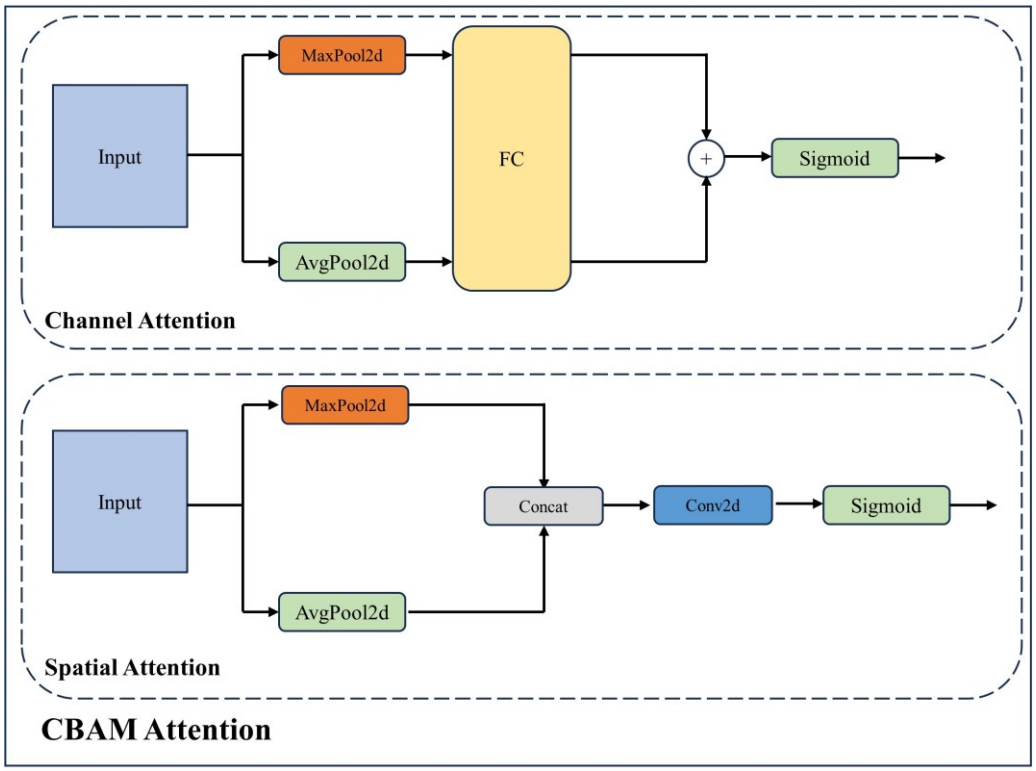


Figure 10. The structure of CBAM [10] attention. CBAM attention is composed of channel attention and spatial attention. The channel attention is globally applied, while the spatial attention focuses locally.

For channel attention, the input is dimensionally reduced through pooling in the spatial dimension, and then the channel attention is learned by the fully connected layer.

$$\begin{aligned} Attention_{channel}(F) &= \text{sigmoid}(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \\ &= \text{sigmoid}(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))) \end{aligned} \tag{6}$$

Equation (1) represents the calculation procedure of channel attention, where F denotes the input features, AvgPool and MaxPool denote the average and maximum pooling layers, and MLP denotes the fully connected layer, whose weights are denoted as W_1 and W_0.

When it comes to spatial attention, the input features are dimensionally reduced through average pooling and maximum pooling in parallel, and then are concatenated as the input of a convolution layer. The spatial attention is computed as:

$$\begin{aligned} Attention_{spatial}(F) &= \text{sigmoid}(f[\text{AvgPool}(F), \text{MaxPool}(F)]) \\ &= \text{sigmoid}(f[F^s_{avg}, F^s_{\max}]) \end{aligned} \tag{7}$$

where $f(\cdot)$ denotes a convolution operation and $sigmoid(\cdot)$ denotes the activation function.

The two attention modules, channel and spatial, focus on different aspects: 'what' and 'where,' respectively. In this paper, the spatial-channel sequential process is designed to highlight features of small objects while suppressing redundant features.

4. Experiments and Results

4.1. Metrics

Precision, recall, and mean average precision (mAP) are usually used as the accuracy indexes. Precision is computed as:

$$P = \frac{TP}{TP + FP} \tag{8}$$

where P denotes the precision index, TP denotes the number of true-positive samples, and FP denotes the number of false-positive samples. True-positive refers to objects correctly detected by the model, while false-positive denotes samples that are incorrectly recognized as objects by the model but actually belong to the background in the ground truth.

Recall is computed as:

$$R = \frac{TP}{TP + FN} \tag{9}$$

where R denotes the recall index and FN denotes the number of false-negative samples. False-negative refers to the missed objects.

We can obtain the precision-recall curve of the detection results. The average precision (AP) is computed as:

$$AP = \int_0^1 P(R)dR \tag{10}$$

The computation of mAP (mean average precision) of multiple category detection is given as:

$$mAP = \frac{1}{N}\sum_{k=1}^{N} AP_k \tag{11}$$

where N denotes the number of categories.

In YOLO, the running time of the object detection neural network mainly consists of three parts: image preprocessing time t_{pre}, inference time $t_{inference}$, and post-processing time t_{post}. The preprocessing refers to the algorithm program decoding images into tensors and assigning grids to input images. Inference time refers to the inference time of the neural network. Post-processing mainly takes time in the NMS process.

$$t_d = t_{pre} + t_{inference} + t_{post} \tag{12}$$

Frame Per Second (FPS) is commonly used to reflect the average operating efficiency of the object detection neural network in processing the images of the current dataset. It is defined as the ratio of the number of images to the total running time or as the reciprocal of the mean running time for each image:

$$FPS = \frac{N}{\sum_{i=1}^{N} t_d^i} = \frac{1}{Avg(t_d)} \tag{13}$$

where N denotes the number of images, t_d^i refers to the running time corresponding to the i^{th} image, and $Avg(t_d)$ refers to the mean running time of each image.

4.2. Implementation Details and Dataset

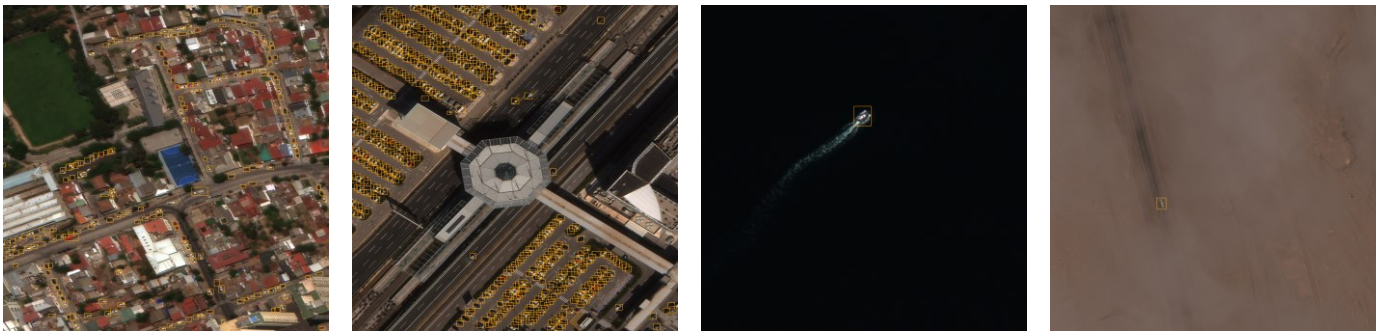
The experiments are implemented on the platform with 4 NVIDIA 3090(24 G) GPUs and Intel® Xeon® Platinum 8260 CPU @ 2.40 GHz, 96 cores, 192 threads. The system version is Ubuntu 18.04 LTS. We conduct the experiment in a virtual environment with PyTorch 2.2.1 and CUDA 12.1. The input image size is 800 × 800 pixels and the batch size of the training data is 16. We employed the stochastic gradient descent (SGD) function as the optimizer to train the model. The hyperparameters of SGD are set to 0.937 of the momentum, 0.01 for the initial learning rate, and 1×10^{-4} for the final learning rate. Every training process lasts 300 epochs. The intersection over union (IoU) threshold for NMS during validating is set to 0.5. Considering the substantial disparity between natural images and satellite images, we refrain from utilizing pre-trained weight initialization parameters, so as to minimize any unwarranted interference at the beginning of the model training process.

The xView dataset is a large-scale public satellite object detection benchmark, including 846 large-scale satellite images, covering an impressive area of over 2000 square kilometers on the Earth's surface. The original xView dataset contains not only small objects, but also larger objects such as windmills, bridges, and airports. We processed the format of annotations and removed the annotations of larger-sized objects. After processing the dataset, a total of 9296 images are obtained, containing 445,473 objects. We have reconstructed the labels of the objects into three categories: vehicle, ship, and airplane. Table 1 shows the number of objects in these categories and the mean pixel size of the bounding boxes.

Table 1. Statistics of the processed xView dataset.

Class	Objects Quantity	Mean Size
Vehicle	442,636	12.4
Ship	1606	14.1
Airplane	1231	27.9

The mean sizes of objects in all selected categories are very small, making detection difficult. Furthermore, the xView dataset covers a wide range of land surface and contains diverse scenes, including some extremely challenging areas. Figure 11 displays several challenging image examples, including complex backgrounds, extremely dense object distributions, heavy reliance on contextual features, and occlusion caused by clouds.



(**a**) Complex background (**b**) Dense distribution (**c**) Reliance on context (**d**) Occlusion

Figure 11. Examples of challenging images.

With reference to the characteristics of the dataset images, we discuss the difficulties in object detection in satellite images. A small bounding box positioning deviation causes a significant decrease in the Intersection over Union (IoU) rate of small objects, but has less impact on large objects. As a result, small objects have strict requirements for detector bounding box positioning. Difficulties also arise due to some challenging scenes: the appearance of some objects is not obvious, while the background features of the scene are complex, as shown in Figure 11a; some objects are very densely distributed and the bounding boxes overlap each other, making it difficult for the detector to distinguish the objects, as shown in Figure 11b; some objects rely heavily on contextual features, such as the ship relying on the wake in Figure 11c; cloud coverage obscures object features, as shown in Figure 11d.

4.3. Quantitative Results

For a fair comparison, we turn off the online data augmentations during training. We deploy our proposed method based on the YOLOv8-s model and set the YOLOv8-s model as the baseline. We transplanted the state-of-the-art feature fusion methods and models based on YOLOv8, trained them on the xView dataset, and compared the performance of these models and improvements with the model we proposed. We compared the baseline

model, multiple SOTA models and their variants, and the proposed model. The results are shown in Table 2.

Table 2. Performance comparison of different models for object detection on the xView dataset. The best results are in bold.

Model	$mAP_{50:95}$/%	mAP_{50}/%
Baseline	33.8	75.5
YOLOX [16]	20.4	52.7
Mask-RCNN [15]	22.2	59.6
Swin-Transformer [35]	26.2	64.3
YOLOv8-s-P2	30.6	70.4
LSK-Net [20]	25.3	65.4
LSKA [29]	36.2	75.8
AFPN [34]	23.1	57.8
ASF-YOLO [36]	30.7	72.5
BIFPN [31]	27.9	71.2
HS-FPN [32]	33.1	72.5
DCFF-YOLO (ours)	**37.3** (+3.5)	**77.6** (+2.1)

The compared models consist of three groups: backbone networks, context extraction networks, and multi-scale fusion improvements. The comparison results will be discussed in groups. (1) Backbone group: YOLOX [16], Mask RCNN [15], Swin-Transformer [35], and YOLOv8-s-P2 are in the compared backbone group. YOLOX, Mask RCNN, and Swin-Transformer are not good at small object detection. Thanks to the exquisite module design, the baseline model, YOLOv8-s, has achieved a high average precision on the xView dataset, significantly surpassing the previous backbone models. YOLOv8-s-P2 denotes an additional tiny-objection detect head with a P2 feature branch in the YOLOv8 neck structure. Nevertheless, the performance of YOLOv8-s-P2 is lower than the baseline YOLOv8-s model, revealing that low-level features perhaps mislead detection. (2) Context extraction group: Two large kernel convolutional networks, LSK-Net [20] and LSKA [29], are in the group. The performance of LSK-Net on the dataset is poor. Meanwhile, the lightweight designed LSKA has gained performance improvements compared to the baseline. (3) Multi-scale fusion group: AFPN [34], ASF-YOLO [36], BIFPN [31], and HS-FPN [32] refer to the fashion or state-of-the-art feature fusion models, which are believed to improve the fusion step and enhance the detection performance. Unexpectedly, these models exhibit an average precision index degradation on the xView dataset. Notably, the AFPN model costs the largest computation and gains the most performance degradation, which may be caused by unguided cross-layer feature concatenation. Our proposed model has achieved the best average precision index, improving the 3.5 $mAP_{50:95}$ index compared to the baseline model.

With reference to the performance comparison, we discuss the improvements for small object detection in satellite images. First, context is beneficial for small object detection, which can be verified by the improvement of LSKA and our proposed DCFF-YOLO. Second, unguided cross-layer feature interaction leads to performance degradation, as verified by the multi-scale group. Optimizing the intra-layer representation of small objects is the basis for cross-layer interaction. Third, LSKA introduces an attention mechanism to the backbone. The comparison of LSK-Net and LSKA reveals the benefits of the attention-guided intra-layer representation.

We conduct an ablation study to verify the contribution of each proposed module, as shown in Table 3.

The Context Convolution block, DySample [9], and CBAM [10] attention module have effectively improved the model performance by solving small object detection of satellite remote-sensing images. Comparing the Context Convolution block and DySample [9], the former improves both the $mAP_{50:95}$ and mAP_{50}, while the latter mainly improves the mAP_{50} index. It reveals that contextual information helps detect more objects and achieve more accurate localization, while dynamic convolution tends to help find more objects by

improving high-level feature upsampling. When combining all three modules, it is observed that the mAP and mAP_{50} indexes of the model are boosted remarkably. It underscores the valuable contribution of the CBAM [10] attention in guiding the feature fusion.

Table 3. The result of the ablation study. The best results are in bold.

Method			Metrics	
Context Conv	**DySample [9]**	**CBAM [10]**	**$mAP_{50:95}$/%**	**mAP_{50}/%**
			33.8	75.5
✓			35.8	77.0
	✓		35.3	75.6
✓	✓		35.8	75.5
✓	✓	✓	**37.3**	**77.5**

For the goal of real-time object detection, we propose a lightweight convolution module and integrate the lightweight upsampling and attention module into the network to build the proposed lightweight model based on YOLO for small object detection. We conducted a comprehensive comparison, analyzing the parameters, computational cost, and the frame per second (FPS) index of the proposed model among our proposed model, the baseline model, and comparison methods. The comparison is shown in Table 4.

Table 4. The scale and efficiency comparison of models. The best results are in bold. An upward arrow indicates that the higher the indicator, the better, while a downward arrow indicates the opposite.

Model	↓Parameters	↓FLOPS	↑FPS
Baseline	11.1 M	28.4 G	168.1
AFPN [34]	8.9 M	38.5 G	57.4
BIFPN [31]	**7.3 M**	25.2 G	138.8
Context Conv	7.5 M	19.5 G	**168.4**
DCFF-YOLO (ours)	7.8 M	**18.8 G**	157.8

We have proposed a lighter and more efficient object detection model that outperforms the baseline model on satellite image datasets. In Table 4, 'Context Conv' refers to the model where we incorporate the proposed Context Convolution block into the YOLOv8-s architecture. The indexes highlight the contribution of the Context Convolution block for light-weighting. Building upon this foundation, our proposed DCFF-YOLO detector only introduces just a few additional parameters, resulting in further improvements to average precision.

4.4. Visualization

4.4.1. Training Process

We record the training process as shown in Figure 12. We can find that the proposed method has advantages in both accuracy and convergence.

4.4.2. Visualization of the Feature Response

To demonstrate the effectiveness of our improvements, we visualize the model's feature maps in the form of heatmaps in Figure 13. Compared to the baseline model, the feature response of DCFF has the following improvements: (1) DCFF detects more ground -truth objects. (2) DCFF handles dense detection better. (3) The feature responses of DCFF at the object boundary and in surrounding contexts are stronger, showcasing the effect of the contextual convolution block and the attention module.

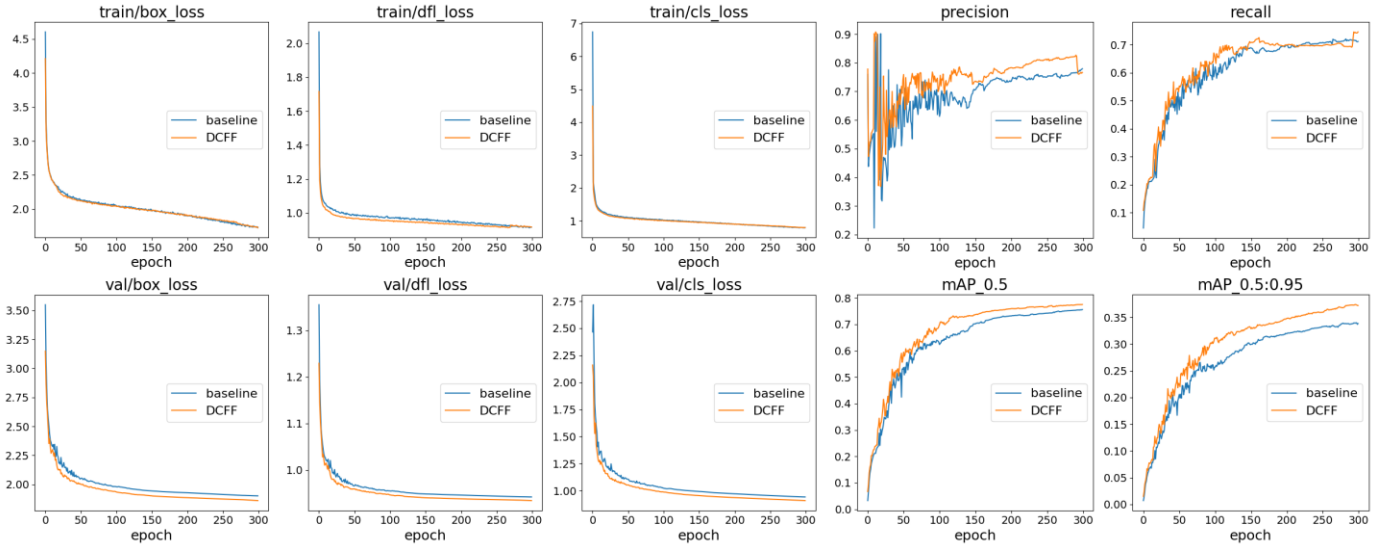
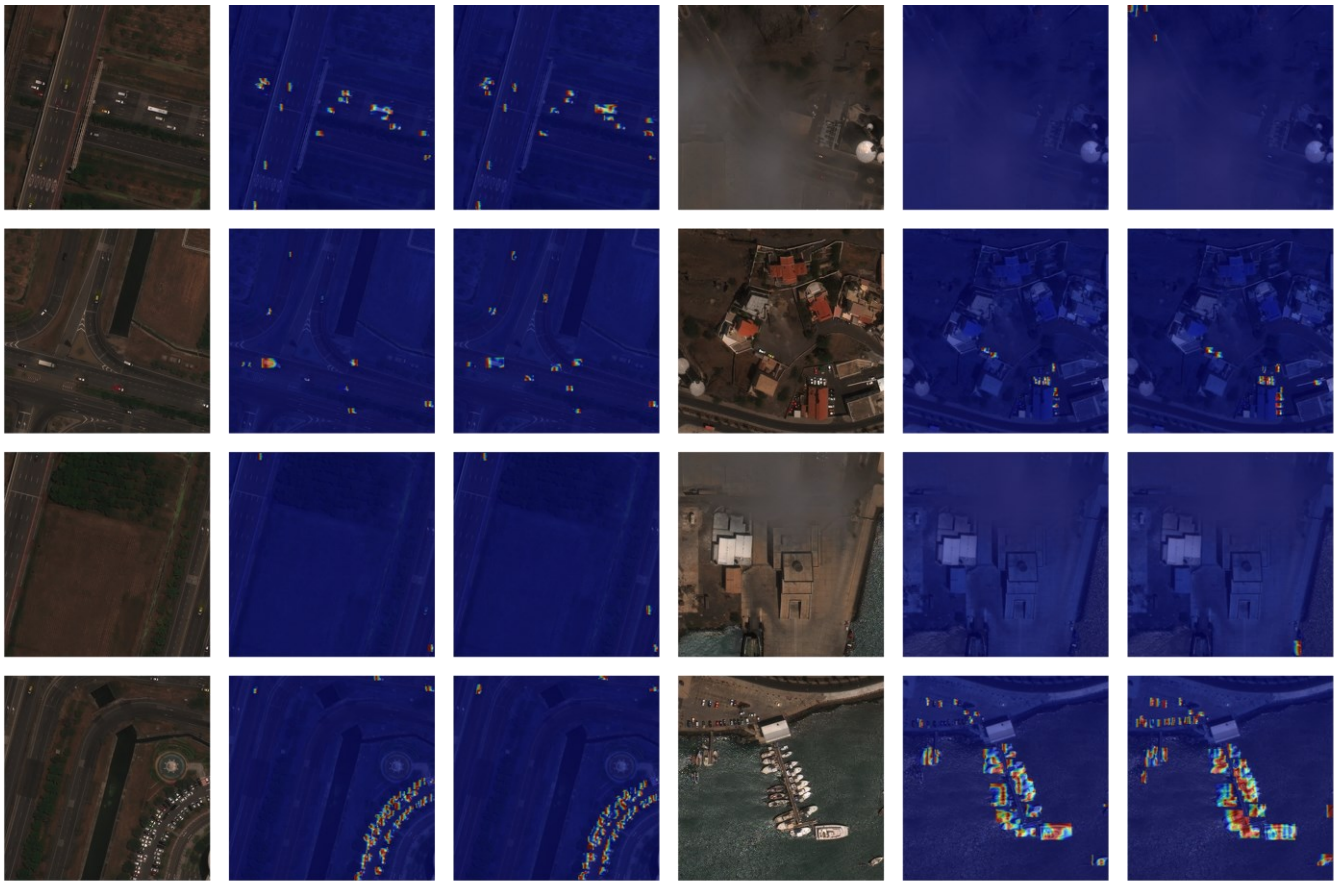


Figure 12. The comparison of the training process. Our proposed model (DCFF) exhibits a lower loss value and stays ahead of the baseline in the curve of the average precision index on the validation set.



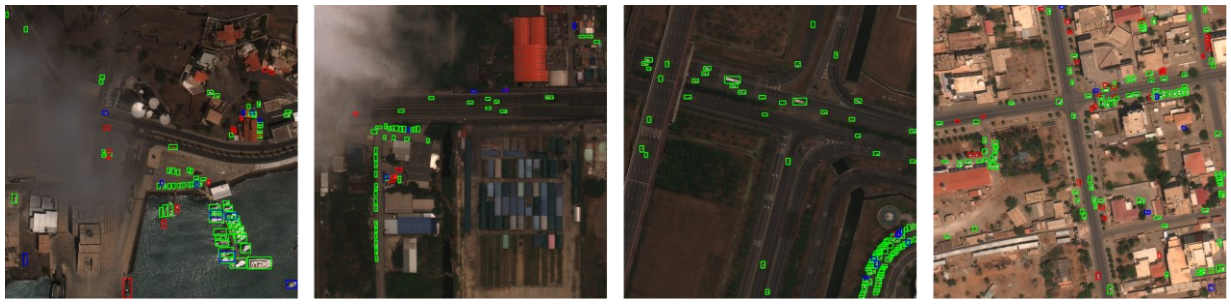
(a) source images (b) baseline (c) DCFF (d) source images (e) baseline (f) DCFF

Figure 13. The heatmap of the feature responses. The redder the response, the more significant the feature representation; conversely, the bluer it is, the closer it is to the background.

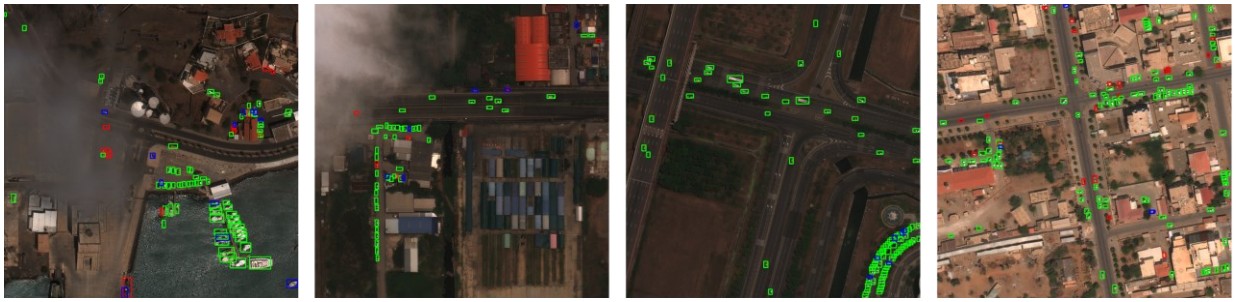
4.4.3. Visualization of the Detection

We have also conducted a comparative visualization experiment, displaying TP, FP, and FN objects in Figure 14. TPs (green boxes) represent correct detections, FPs (blue boxes)

correspond to wrong detections, which denote detections of the background as objects, and FNs (red boxes) denote missed detections.



(**a**) Baseline



(**b**) DCFF (proposed)

Figure 14. Visualization of detections. Green boxes indicate correct detections, red boxes indicate missed detections, and blue boxes indicate redundant detections.

When we focus on the red boxes (missed detections), our proposed DCFF demonstrates a slight advantage in dense detection compared to the baseline model. The advantage is mainly reflected in densely distributed scenes, for example, in the middle left of the 2nd (from left to right) image and bottom right of the 3rd image in Figure 14b.

When it comes to the blue boxes (wrong detections), DCFF recognizes objects from the background more accurately. For example, in the 1st image of Figure 14a the baseline mistakenly detects the warehouse on the shore as an object, while DCFF avoids this wrong detection in Figure 14b. Furthermore, there are some missing annotations in the dataset, which means that some blue boxes are actually correct detections.

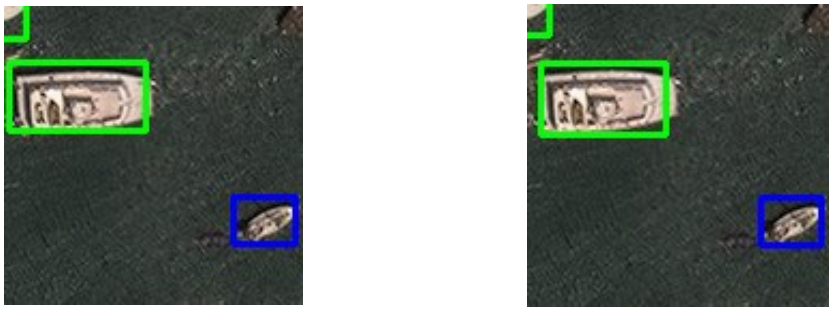
5. Discussion

In satellite images, small objects may lack discernible texture features. Our research demonstrates that enhancing small object detection involves emphasizing contextual features. This approach aligns with the method of combining the surrounding context to identify objects for humankind. However, introducing low-level features without proper arrangement can escalate computational complexity and potentially result in reduced average precision. Through a series of experiments and analyses, we have found the effectiveness of adjusting intra-layer feature representation. In addition, the attention mechanism is crucial in multi-scale feature fusion and can facilitate other potential improvements.

At present, there are only a limited number of large-scale public satellite image object detection datasets. The average precision index of object detection is notably affected by the quality of the dataset. An example is taken from Figure 14 to show a missing annotation in Figure 15. Obviously, both detectors correctly detect the ship object.

Our improvements rely heavily on accurate annotations, but it is hard to accurately label all satellite image objects. This limitation manifests in the improvement of precision indexes. In typical object detection scenarios with normal-style images, improvements resulting from model enhancements are often more pronounced in the mAP_{50} indicator than in the $mAP_{50:95}$. However, our research demonstrates a different trend: the improvement in

$mAP_{50:95}$ surpasses that of mAP_{50}. This discrepancy may arise from the model's challenge in identifying objects that were previously missed by the baseline model.



(**a**) Baseline (**b**) DCFF (ours)

Figure 15. Missing annotation of the dataset are shown in the blue box. Both of the baseline and DCFF correctly detect the ship object, but these detections are marked as wrong detections. The green boxes indicate the other correct detections.

6. Conclusions

In this paper, we have developed a lightweight and high-performance object detection model tailored for satellite remote-sensing images. The objects of interest are mostly small objects, which poses great challenges to detection. We improve the small object detector by extracting local and contextual features, optimizing the intra-layer feature representation, and introducing an attention mechanism to the multi-scale feature fusion. We propose the Context Convolution block to extract local and contextual features to enhance the detection. The Context Convolution block both extracts more features and reduces computational cost. In the feature fusion step, we employ the DySample [9] upsampler and CBAM [10] attention to improve fusing features. As observed in comparative experiments, unguided cross-layer concatenation of features will lead to degradation of detection performance. Due to the contribution of these modules, we have achieved a 3.5% $mAP_{50:95}$ improvement compared to the baseline model, shedding light on the mechanism of small object detection in satellite images. Additionally, our model boasts a lightweight design. The proposed Context Convolution block reduces the computational cost, and the two introduced modules also follow a lightweight style. Consequently, we achieve the average precision improvement with only 70% of the parameters and 66% of the FLOPS, and the FPS remains comparable to the baseline model.

Author Contributions: Methodology and writing, H.Y.; supervision, S.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the China Postdoctoral Science Foundation (No. 2023M740760), Light of West China (No. XAB2022YN10), and Shaanxi Key Research and Development Plan (2024SF-YBXM-678).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The xView dataset is accessed on 22 February 2018, and available at the following link: http://xviewdataset.org/. Please contact the authors for data requests.

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

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25, Proceedings of the 26th Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012*; Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012.

2. 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.
3. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
4. Redmon, J.; Farhadi, A. Yolov3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
5. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2022**, arXiv:2209.02976.
6. 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.
7. Wu, T.; Tang, S.; Zhang, R.; Cao, J.; Zhang, Y. Cgnet: A Light-Weight Context Guided Network for Semantic Segmentation. *IEEE Trans. Image Process.* **2020**, *30*, 1169–1179. [CrossRef] [PubMed]
8. Sengupta, A.; Ye, Y.; Wang, R.; Liu, C.; Roy, K. Going Deeper in Spiking Neural Networks: VGG and Residual Architectures. *Front. Neurosci.* **2019**, *13*, 95. [CrossRef] [PubMed]
9. Liu, W.; Lu, H.; Fu, H.; Cao, Z. Learning to Upsample by Learning to Sample. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 6027–6037.
10. 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.
11. 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.
12. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
13. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28, Proceedings of the Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; MIT Press: Cambridge, MA, USA, 2015.
14. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
15. 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.
16. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding Yolo Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
17. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
18. Liu, Z.; Hu, J.; Weng, L.; Yang, Y. Rotated Region Based CNN for Ship Detection. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 900–904.
19. Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; Xu, C. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11207–11216.
20. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; Li, X. Large Selective Kernel Network for Remote Sensing Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 16794–16805.
21. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for Small Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611215. [CrossRef]
22. Zhang, F.; Shi, Y.; Xiong, Z.; Zhu, X.X. Few-Shot Object Detection in Remote Sensing: Lifting the Curse of Incompletely Annotated Novel Objects. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5603514. [CrossRef]
23. Li, W.; Li, W.; Yang, F.; Wang, P. Multi-Scale Object Detection in Satellite Imagery Based on YOLT. In Proceedings of the IGARSS 2019, IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 162–165.
24. He, Q.; Sun, X.; Yan, Z.; Li, B.; Fu, K. Multi-Object Tracking in Satellite Videos with Graph-Based Multitask Modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5619513. [CrossRef]
25. Wang, X.; Wang, A.; Yi, J.; Song, Y.; Chehri, A. Small Object Detection Based on Deep Learning for Remote Sensing: A Comprehensive Review. *Remote Sens.* **2023**, *15*, 3265. [CrossRef]
26. Yavariabdi, A.; Kusetogullari, H.; Celik, T.; Cicek, H. FastUAV-Net: A Multi-UAV Detection Algorithm for Embedded Platforms. *Electronics* **2021**, *10*, 724. [CrossRef]
27. Lim, J.-S.; Astrid, M.; Yoon, H.-J.; Lee, S.-I. Small Object Detection Using Context and Attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021; pp. 181–186.
28. Cao, J.; Chen, Q.; Guo, J.; Shi, R. Attention-Guided Context Feature Pyramid Network for Object Detection. *arXiv* **2020**, arXiv:2005.11475.

29. Lau, K.W.; Po, L.-M.; Rehman, Y.A.U. Large Separable Kernel Attention: Rethinking the Large Kernel Attention Design in CNN. *Expert Syst. Appl.* **2024**, *236*, 121352. [CrossRef]
30. Kim, S.-W.; Kook, H.-K.; Sun, J.-Y.; Kang, M.-C.; Ko, S.-J. Parallel Feature Pyramid Network for Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 234–250.
31. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
32. Chen, Y.; Zhang, C.; Chen, B.; Huang, Y.; Sun, Y.; Wang, C.; Fu, X.; Dai, Y.; Qin, F.; Peng, Y.; et al. Accurate Leukocyte Detection Based on Deformable-DETR and Multi-Level Feature Fusion for Aiding Diagnosis of Blood Diseases. *Comput. Biol. Med.* **2024**, *170*, 107917. [CrossRef] [PubMed]
33. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:191109516.
34. Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; Liang, R. AFPN: Asymptotic Feature Pyramid Network for Object Detection. In Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, HI, USA, 1–4 October 2023; pp. 2184–2189.
35. 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, QC, Canada, 10–17 October 2021; pp. 10012–10022.
36. Kang, M.; Ting, C.-M.; Ting, F.F.; Phan, R.C.-W. ASF-YOLO: A Novel YOLO Model with Attentional Scale Sequence Fusion for Cell Instance Segmentation. *arXiv* **2023**, arXiv:231206458.