remote sensing

MDPI

Article

Real-Time UAV Patrol Technology in Orchard Based on the Swin-T YOLOX Lightweight Model

Yubin Lan [1,2,3,4], Shaoming Lin [1,2,3,4], Hewen Du [1,2,3,4], Yaqi Guo [1,2,3,4] and Xiaoling Deng [1,2,3,4,*]

[1] College of Electronic Engineering (College of Artificial Intelligence), South China Agricultural University, Guangzhou 510642, China
[2] Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou 510642, China
[3] National Center for International Collaboration Research on Precision Agricultural Aviation Pesticide Spraying Technology, Guangzhou 510642, China
[4] Guangdong Engineering Technology Research Center of Smart Agriculture, Guangzhou 510642, China
* Correspondence: dengxl@scau.edu.cn

Abstract: Using unmanned aerial vehicle (UAV) real-time remote sensing to monitor diseased plants or abnormal areas of orchards from a low altitude perspective can greatly improve the efficiency and response speed of the patrol in smart orchards. The purpose of this paper is to realize the intelligence of the UAV terminal and make the UAV patrol orchard in real-time. The existing lightweight object detection algorithms are usually difficult to consider both detection accuracy and processing speed. In this study, a new lightweight model named Swin-T YOLOX, which consists of the advanced detection network YOLOX and the strong backbone Swin Transformer, was proposed. Model layer pruning technology was adopted to prune the multi-layer stacked structure of the Swin Transformer. A variety of data enhancement strategies were conducted to expand the dataset in the model training stage. The lightweight Swin-T YOLOX model was deployed to the embedded platform Jetson Xavier NX to evaluate its detection capability and real-time performance of the UAV patrol mission in the orchard. The research results show that, with the help of TensorRT optimization, the proposed lightweight Swin-T YOLOX network achieved 94.0% accuracy and achieved a detection speed of 40 fps on the embedded platform (Jetson Xavier NX) for patrol orchard missions. Compared to the original YOLOX network, the model accuracy has increased by 1.9%. Compared to the original Swin-T YOLOX, the size of the proposed lightweight Swin-T YOLOX has been reduced to two-thirds, while the model accuracy has slightly increased by 0.7%. At the same time, the detection speed of the model has reached 40 fps, which can be applied to the real-time UAV patrol in the orchard.

Keywords: model lightweight; real-time UAV patrol orchard; layer pruning; deployment model

Citation: Lan, Y.; Lin, S.; Du, H.; Guo, Y.; Deng, X. Real-Time UAV Patrol Technology in Orchard Based on the Swin-T YOLOX Lightweight Model. *Remote Sens.* **2022**, *14*, 5806. https://doi.org/10.3390/rs14225806

Academic Editor: Eben Broadbent

Received: 10 October 2022
Accepted: 14 November 2022
Published: 17 November 2022

1. Introduction

The fruit industry accounts for a large proportion of the Chinese agricultural economy. Traditional orchards are mainly human-managed, requiring lots of human resources to take care of them and wasting considerable agricultural resources due to the uneven application of pesticides and fertilizers or failure detection diseases and pests in time by the manual patrol. Therefore, the smart orchard will be the development direction in the future; it can achieve efficient management and the sustainable development of orchards through accurate sensor technology and continuously advanced intelligent technology [1].

To explore the smart orchard, some studies investigated orchard systems to realize remote temperature and humidity detection, water drip irrigation, intelligent fertilization, insect and pest monitoring, supplement and intelligent illumination and control, among others, which can effectively help the growers to plant fruit trees with the main management problems, thus improving the intelligent management level of the smart orchard [2].

Currently, the remote sensing technology of the crop mainly includes satellite, UAV and ground remote sensing technology, which has been widely used in precision agriculture.

Compared to satellite remote sensing and ground observation, UAV remote sensing has become an important new technology because it can collect the crop growth conditions of real-time data in a large area, visually monitor the growth of crops and conduct precise management immediately [3]. UAV patrol has the advantages of high efficiency, time saving and energy saving. Some UAV remote sensing systems or equipment have been developed in recent years. For example, Zhang Kexin et al. designed a quadrotor with the function of orchard autonomous patrol detection for the collection and transmission of orchard information [4]. Gao Xu et al. used the quadrotor UAV as the carrier and it was equipped with a video acquisition device to conduct aerial patrol, searching and collecting video image data of citrus park, and completing the orchard inspection task [5]. Nikolaos Stefas et al. proposed an autonomous aerial system that can safely navigate within the orchard row [6].

Deep learning has made great progress in machine vision tasks, such as image classification and object detection [7]. Combing deep learning and UAV remote sensing in real-time for an orchard can understand the overall situation of the orchard faster and more efficiently. Shi et al. [8] combined the deep learning training model and UAV remote sensing data to identify flos lonicerae. Deng et al. [9] used UAV technology and optimized ResNet101 network for pine wood nematode disease trees to achieve high accuracy and large-scale detection. Jiawei et al. [10] used the YOLACT instance segmentation algorithm to segment the litchi canopy in low-altitude remote sensing, and the AP reached 96.25% on the test set, which laid the foundation for the precise management of litchi plants. However, the application of the deep learning real-time detection algorithm to UAV patrol orchard task is seldom reported.

The deep learning-based object detection algorithms are generally divided into two categories: The two-stage method and one-stage method. In 2014, R.Gerishick et al. proposed a two-stage detection algorithm RCNN [11], which was the first to apply a convolutional neural network (CNN) to the target detection task, using the selective search algorithm [12] to generate candidate regions. Subsequently, two-stage detection algorithms, such as Fast RCNN [13] and Faster RCNN [14], appeared. However, due to the large number of prediction boxes, these two-stage methods are computation-intensive and slow in detection speed, which are not suitable for real-time detection tasks and embedded platforms. In 2015, J. Redmon et al. proposed a single-stage detection algorithm, YOLO (You Only Look Once) [15], regarding object detection as a regression problem and directly obtaining category and bounding box information. Then, YOLOv2 [16], YOLOv3 [17], SSD [18], YOLOX [19] and other one-stage detection networks were proposed, which greatly improve the detection speed and meet the requirements of real-time detection. In particular, the YOLOX algorithm proposed by Zheng Ge et al. further improves the detection accuracy while retaining the advantages of YOLO series algorithms, and it is one of the most powerful target detection algorithms at present.

However, the one-stage CNN model still has typical problems, such as the lack of the ability and low accuracy to extract remote features from global information. Inspired by the use of self-attention in the Transformer [20], many computer vision tasks propose to use the self-attention mechanism to effectively overcome the limitations of CNN in order to mine remote correlation dependencies in the text. The self-attention mechanism can obtain the relationship between global elements more quickly, focus on different areas of the image and integrate information of the whole image. Vision Transformers (ViT) [21] is a representative and state-of-the-art (SOTA) work in the field of image recognition. It only uses a self-attention mechanism, which makes the image recognition rate much higher than the CNN-based model.

However, ViT is still not suitable for the mission of real-time UAV patrol orchard, because ViT generates a single low-resolution patch feature map, and its computational complexity is quadratic to the input image size due to the calculation of global self-attention. Moreover, ViT focuses too much on the overall semantic information and ignores the local structural features. The Swin Transformer [22] proposed by Liu solves these problems by

introducing some CNN features and constructing hierarchical feature maps by merging image blocks. In addition, the Swin Transformer has linear computational complexity related to the input image size. Therefore, the Swin Transformer is regarded as suitable for the feature extraction network for the lightweight object detection algorithm in this study.

In order to satisfy the demand of the high real-time performance of UAV patrol orchard task, we must reduce the number of parameters and computing operations. The model pruning techniques are popular because of their simplicity in practice and promising compression rate and have achieved great success in the field of convolution neural networks (CNNs) for many vision tasks [23–25]. However, many works using model pruning techniques have also been proposed to compress the Transformer [26–28]. MAO et al. proposed Block-wise Structured Sparsity Learning (BSSL) to analyze the Transformer model property. Then, based on the characters derived from BSSL, we apply Structured Hoyer Square (SHS) to derive the final pruned models, named TPrune [29]. The model can achieve 1.16×–1.92× speedup on mobile devices.

In this study, the Swin-T YOLOX lightweight network was proposed and can meet the need of high accuracy and high real-time performance of UAV patrol orchard task. Based on the active learning strategy, the dataset from low-altitude images of citrus groves collected by UAVs was first annotated and constructed. A variety of advanced data enhancement methods are used to expand the dataset to enhance the robustness and generalization ability of the model. The model layer pruning technology was used to reduce the parameter quantity of the Swin Transformer in the backbone network as much as possible without losing the model accuracy. Finally, the lightweight Swin-T YOLOX model was deployed on the embedded GPU module (NVIDIA Jetson Xavier NX) for practical application. The overall flow chart of the study is shown in Figure 1.

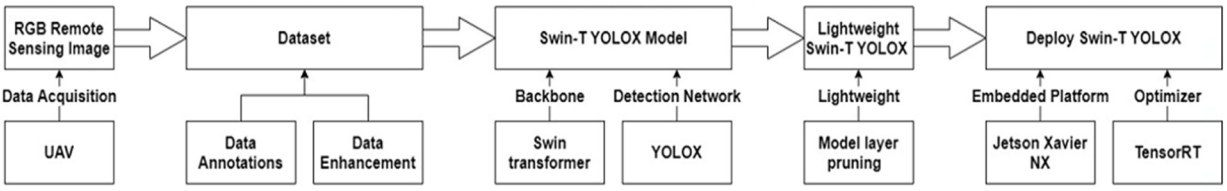

Figure 1. The overall flow chart of the study.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing

2.1.1. UAV Remote Sensing Image Acquisition

The test of UAV patrol orchard was conducted in a citrus orchard in Boluo County (23°29′56.74″N–114°28′4.11″E), Huizhou City, Guangdong Province, China. There were 324 citrus trees in the orchard. The DJI PHANTOM 4 UAV, as shown in Figure 2, equipped with a 1/2.3″ complementary metal-oxide semiconductor (CMOS) with 12.4 million effective pixels, was used to collect RGB images of the above orchard. The flying altitude was set to 20 m, and the pictures were taken at 4000 × 3000 pixels. An example of the UAV image is shown in Figure 3. Each image taken by UAV was cut into multiple 640 × 640 pixel images. The task of the orchard patrol is to detect the anomaly in orchard, including two types of abnormal targets, one is yellowing plants suspected of disease in the citrus canopy, and the other is uneven planting or felled areas in the orchard (as shown in Figure 3). The original dataset consists of 1007 cropped and filtered images. The open source software LabelImg was used to label images, which generates an XML file containing labels and location information. The labeled dataset was randomly allocated to train set, valid set and test set according to 8:1:1, corresponding to 806, 101 and 100 images. In addition, the program was executed to convert the PASCALL-VOC format to the COCO 2017 format.

Figure 2. The DJI PHANTOM 4.

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

Figure 3. The red squares in figure (**a**,**b**) are (**a**) Uneven planting or felled areas, (**b**) yellow suspected plant; (**c**) normal citrus tree.

2.1.2. UAV Remote Sensing Image Data Enhancement

Data augmentation increases the size and diversity of the training dataset, helps to solve overfitting problems, and improves the robustness and generalization ability of the model. Standard data enhancement techniques create new samples by applying simple geometric transformations (such as rotation, scaling, cropping, shifting, and flipping) and color space application transformations (such as adjusting the brightness, contrast, or color saturation of an image), or use a combination of them. More sophisticated techniques based on random erasing and image-mixing have been introduced to generate more challenging samples for the model, such as the Cutout [30] and Mixup [31] techniques. In Cutout, a random fixed-size region of the image is intentionally replaced with black pixels or random noise. Mixup combines two images with their class labels through linear interpolation to create a new training instance.

In this study, the dataset used a hybrid data enhancement method, that is, randomly select one of three enhancement techniques (standard, Cutout and Mixup) for each batch in the training stage. The standard data enhancement used vertical and horizontal flipping and randomly adjusted the brightness and color of the image. For the Cutout technique, the number of holes was set to 15 and the cutout region size was set to 10 × 10 pixels. The proportion of the enhancement training set in the Mixup technology was set to 0.5. After data enhancement, the dataset consists of 3725 images, including 3222 images in the train set, 402 images in the valid set, and 101 images in the test set. In order to verify the recognition accuracy of the model on the actual orchard scene, data enhancement is not used in the test set.

2.2. Proposed Swin-T YOLOX Lightweight Network

2.2.1. Swin-T YOLOX Algorithm

YOLOX was proposed by Ge Z et al. in 2021, surpassing all YOLOv3 to YOLOv5 at that time. Compared to YOLOv3, YOLOX improved 3%AP on COCO dataset. Compared to YOLOv5, YOLOX-L has a 1.8%AP improvement on COCO dataset. YOLOX -tiny and

YOLOX -Nano have a 10% and 1.8%AP improvement compared to YOLOV4-Tiny and NanoDet, respectively. It also provides version deployments supporting ONNX, TensorRT, NCNN, and Openvino. In YOLOX, the anchor free detector replaces the anchor mechanism and the darkNet53 of YOLOv3 is used as the baseline of YOLOX. The overall model of YOLOX is mainly composed of three key parts: Backbone, neck and YOLO Head.

Backbone was used for feature extraction of the YOLOX, which uses the CNN-based CSPDarknet backbone network to capture local feature information through a convolution kernel, often ignoring the relationship with global feature information. Unlike CNN, the Transformer has a strong ability to focus on global information modeling. The Swin Transformer model is an improved version based on the Transformer recently proposed by Microsoft. It not only has the ability of the Transformer to focus on global information modeling, but also uses the method of moving windows to realize the cross window connection, so that the model can focus on the relevant information of other adjacent windows. Cross window feature interaction extends the acceptance domain to a certain extent, which brings higher efficiency. The overall structure of the Swin transformer is shown in Figure 4.

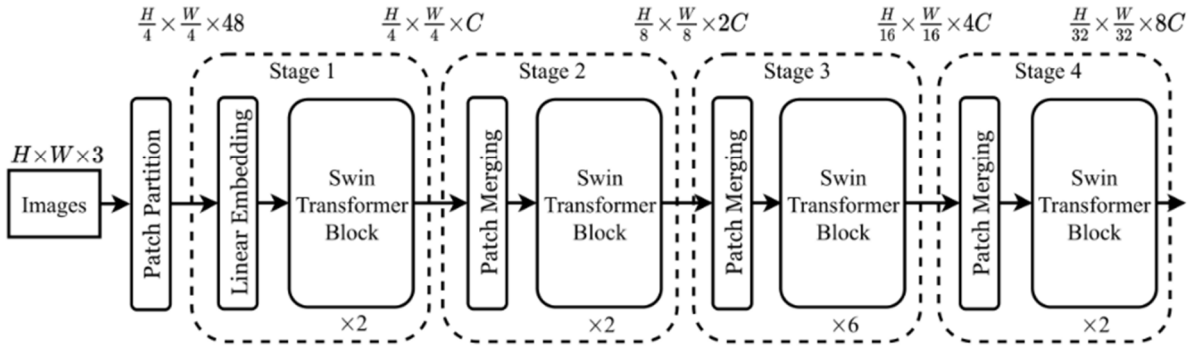

Figure 4. The overall structure of the Swin transformer.

Compared to ViT and CNN, the Swin Transformer can extract better remote sensing image features. Therefore, in this study, the Swin Transformer was used to replace Darknet53 as the backbone network of YOLOX. This model was named Swin-T YOLOX and was constructed as follows: Three feature layers were obtained from the Swin Transformer and were fused in the neck part, respectively. YOLO Head is divided into classifiers and regressors to judge the feature points and to determine whether there is a corresponding object, the overall structure of the Swin-T YOLOX is shown in Figure 5.

2.2.2. Lightweight of Swin-T YOLOX Model

Transformer is a deep architecture with millions of parameters, hundreds of attention heads, and multiple layers. In general, models with large architectures tend to produce better results. However, the enormous computational complexity and the huge memory requirement associated with these models make them impractical for deployment and prone to overfitting. Therefore, to meet the needs of orchard UAV real-time patrol, Swin-T YOLOX needs to be lightweight.

The Swin transformer block is the core part of the Swin transformer algorithm. The detailed structure is shown in Figure 6. The block is composed of window multi-head self-attention (*W-MSA*), shifted windows multi-head self-attention (*SW-MSA*) and multilayer perceptron (*MLP*). A layernorm (*LN*) layer is inserted in the middle to make the training more stable and a residual connection is used after each module. This part can be expressed as Equation (1).

$$\begin{aligned} \hat{z}^{l} &= W-MSA\left(LN(\hat{z}^{l-1})\right)+\hat{z}^{l-1}, \\ z^{l} &= MLP\left(LN(\hat{z}^{l})\right)+\hat{z}^{l}, \\ \hat{z}^{l+1} &= SW-MSA\left(LN(\hat{z}^{l})\right)+\hat{z}^{l}, \\ z^{l+1} &= MLP\left(LN(\hat{z}^{l+1})\right)+\hat{z}^{l+1}, \end{aligned} \tag{1}$$

where $\hat{z}^{l}$ and z^{l} represent the outputs of (*S*)*W-MSA* and multilayer perception (*MLP*) on block *l*. Each input in the formula is normalized by the layer norm (*LN*).

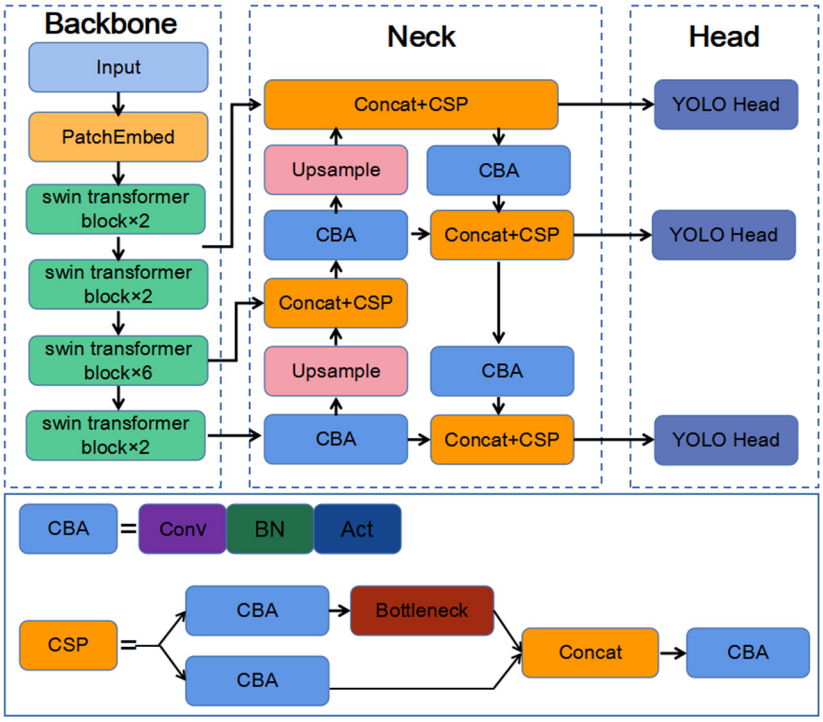

Figure 5. The overall structure of Swin-T YOLOX.

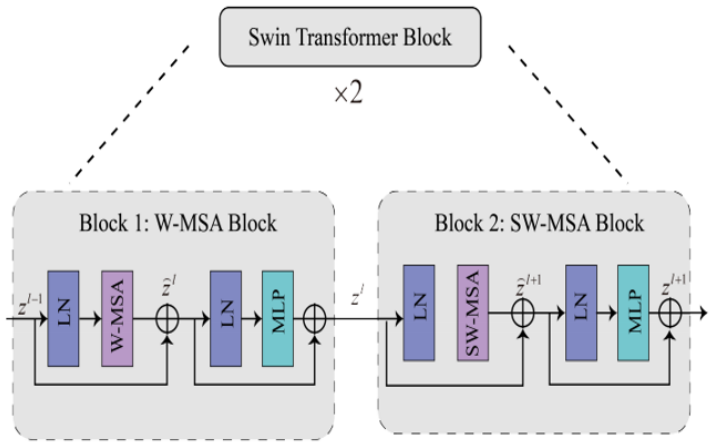

Figure 6. Swin Transformer Block structure.

Different versions of Swin Transformer differ in the number of Swin Transformer block layers, the size of hidden dimensions, the number of attention heads used by the *W-MSA* and *SW-MSA* layers, and the size of the *MLP* classifier. As shown in Table 1, the "Swin-T" model has 12 Swin Transformer block layers, with hidden size 768, and uses 24 attention

heads in the attention layer. Other versions use deeper network layers, attention heads, hidden dimensions, and a larger number of model parameters.

Table 1. Parameter statistics of tiny, small, base and large versions of the Swin Transformer.

Model	Number of Layers	Hidden Dimensions	Number of Attention Heads	Model Parameters
Swin-T	12	768	24	29 M
Swin-S	24	768	24	50 M
Swin-B	24	1024	32	88 M
Swin-L	24	1536	48	197 M

According to the overall structure of the Swin Transformer, the feature maps corresponding to the four stages are down sampled by 4 times, 8 times and 16 times. Different multiples can obtain the extraction effect of feature maps of different scales, which makes the backbone network not only applicable for image classification, but also for object detection and image segmentation. From the hierarchical feature map constructed by the Swin Transformer (Figure 7), it can also be seen that, with the increase in down-sampling times, the granularity of its feature extraction is constantly increasing. From the shallow layer to deep layer, its receptive field is also gradually expanding. However, in this study, the target detected by UAV remote sensing only accounts for a small part of the image, as shown in Figure 3. The expansion of the receptive field cannot improve the accuracy of target recognition, but the deep Swin Transformer block layers adds many parameters to the model.

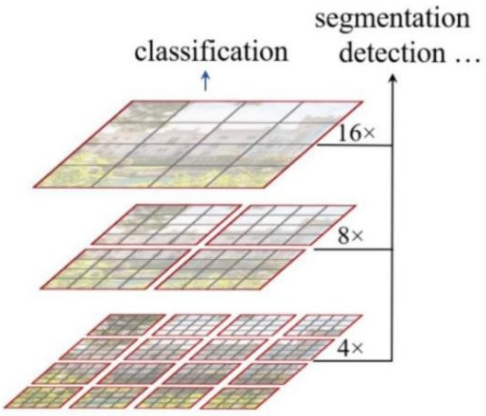

Figure 7. The hierarchical feature map constructed by the Swin Transformer.

Model compression techniques are aimed at producing a lighter version of the model without reducing the original accuracy. Knowledge distillation and model pruning are commonly used as compression approaches [32,33]. In order to remove the redundancy of multiple layers and multiple attention heads in the Swin Transformer, this study proposes a model compression method based on pruning the Swin Transformer block layers, that is, extracting smaller models with different depths from full-size models. The aim is to explore the tradeoff between model performance and model depth to determine the extreme compression architecture that provides the best accuracy.

Based on the Swin-T version, the model was first trained with the maximum number of layers (i.e., 12 layers), and the role of each layer was analyzed from the perspective of quantitative analysis. In order to better understand the network behavior and the area of attention of each layer's attention head, the output representation of each layer of the backbone network Swin Transformer was extracted and the attention map of each layer was visualized, as shown in Figure 8. Then, the same parameters were set to train the Swin-T YOLOX model, twelve groups of experiments were conducted, each group pruned

one layer of the Swin Transformer without repetition, and the model accuracy and number of parameters after pruning were counted.

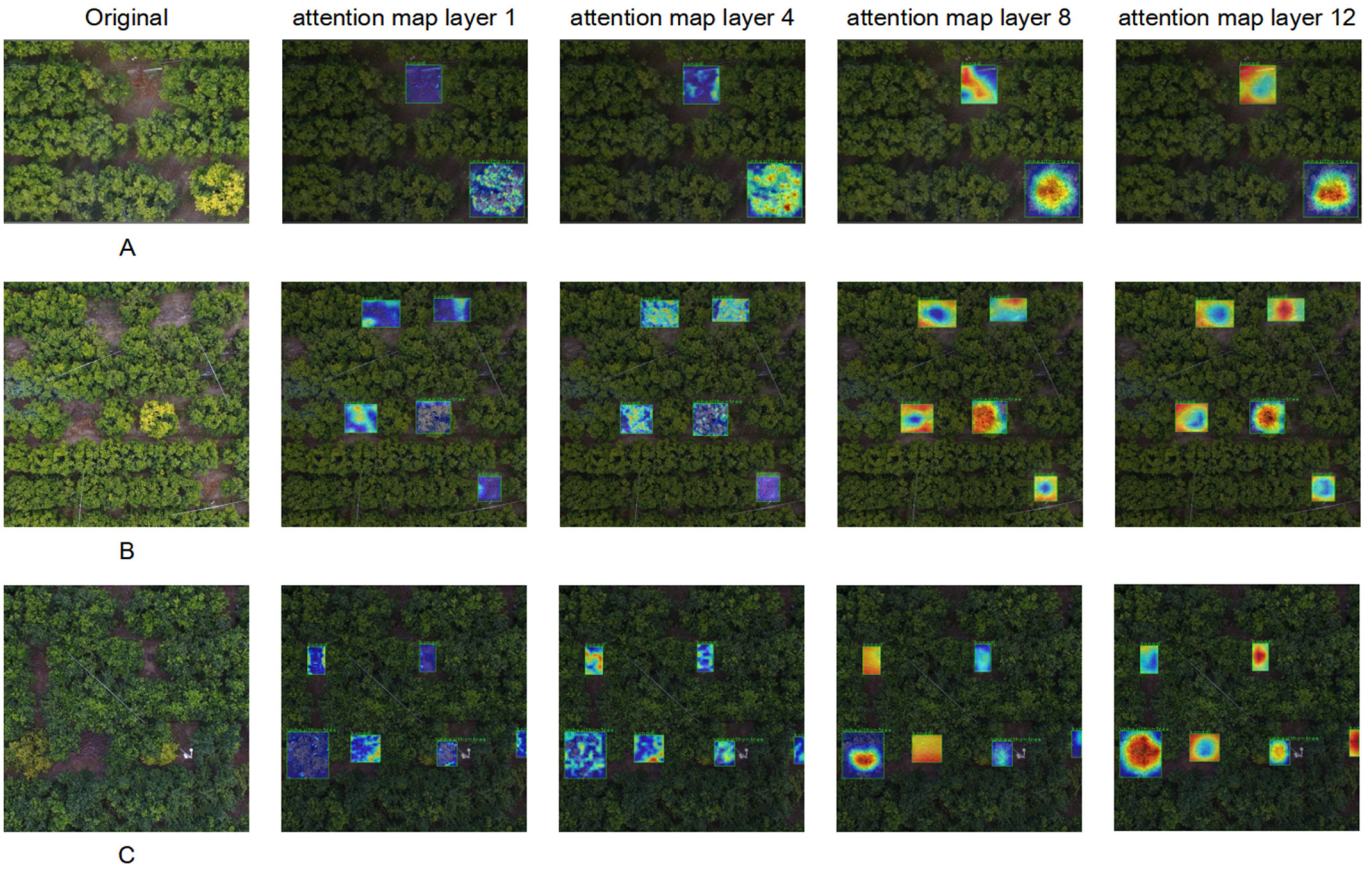

Figure 8. (**A**–**C**) Three groups of pictures show Attention maps for the 1st, 4th, 8th and 12th layers of the Swin Transformer. The three images in the first column are sample images.

Figure 8 shows the output examples of four different Swin Transformer layers (layers 1, 4, 8, and 12). As can be seen, the network gradually learns to focus on the regions with the most representative of the two taxonomic groups. For category 2—citrus yellowing suspected disease plants—the network only pays attention to the area with obvious yellow leaves in the first layer, while the heat concentration of all the yellow leaves in the fourth layer increases significantly. The distribution of heat concentration is clear in the eighth and twelfth layers, even matching the area of the objective tree crown. For category 1—uneven planting or felled areas—the change in attention heat map also has a similar response. The network pays attention to the dead leaves on the ground in the first layer, but also focus on some unrelated areas such as the edge near the tree crown in image C. These problems are gradually improved in subsequent layers. In the fourth and eighth layers, compared to the first layer, it is obvious that the heat concentration increases in the areas of bare and dead leaves. This means that attention will increase as the Swin Transformer layer deepens. In addition, the attention map provided by layer 8 has more visual similarities to the attention map provided by the last layer, with a slight increase in attention compared to the last layer.

Figure 9 shows the model accuracy and the number of model parameters obtained by pruning different layers of the Swin Transformer network. For the bar chart of accuracy, from layer 3 to layer 8, the pruning of each layer has an impact on the model accuracy close to 2%. However, pruning the shallow and deep layer of the Swin Transformer will not seriously reduce the model accuracy because the granularity of the feature extraction in shallow layers is too small, and the attention mechanism does not focus on the target object, so pruning the shallow layers will not seriously affect the model accuracy. However, the

granularity of the feature extraction in deep layers is too large, and the features extracted in the previous layers are enough to recognize the target object. Therefore, pruning the deep layer will not reduce the model accuracy. It is worth noting that pruning layer 10 can even improve the accuracy of the model. Moreover, from the analysis of model parameters, as the number of pruning layers becomes deeper, the number of model parameters decreases faster. This means that pruning the deep layers of the model is a reasonable choice from the perspective of model identification performance. Through a layer-by-layer comparison, the scheme of pruning the last three layers of the Swin Transformer network was adopted to realize the lightweight model.

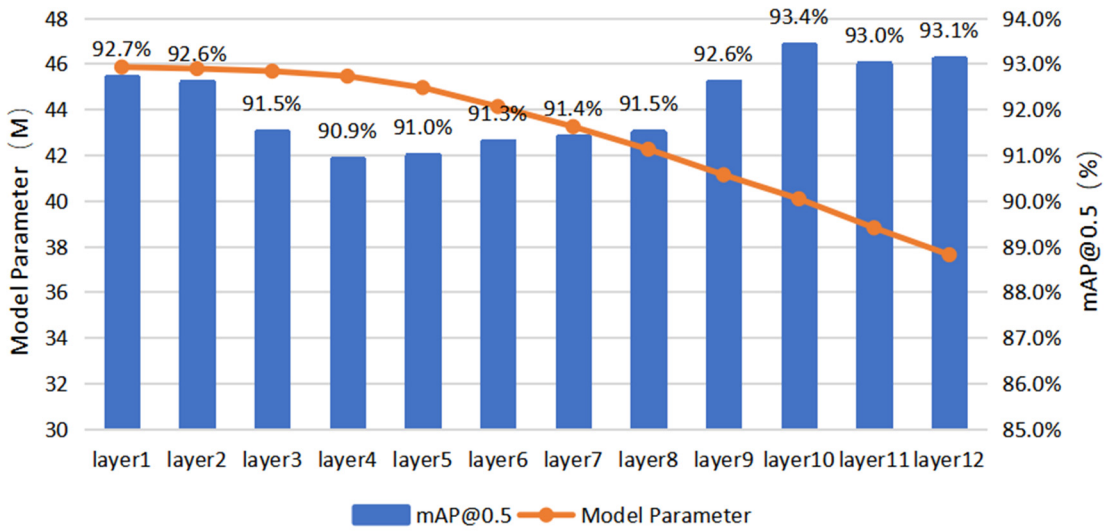

Figure 9. Pruning the Swin-T YOLOX model identification accuracy, model parameters and their changes at each layer of the Swin Transformer.

2.2.3. Deployment Based on Swin-T YOLOX Lightweight Model

The Jetson Xavier NX platform, which is based on the ARM architecture, was used for edge computing. The platform is approximately 70 × 45 mm in size and feature computing resources, such as 6-core Carmel ARM cpus and 384 NVIDIA CUDA® Cores, providing up to 21 TOPS computing power, its low power consumption, high performance, large memory bandwidth and other characteristics make it very suitable for airborne image data processing. TensorRT built into the Jetson edge computing platform is a high-performance deep learning Inference optimizer, which can provide low-latency and high-throughput deployment detection and is widely used in embedded platforms or autonomous driving platforms. TensorRT mainly optimizes the trained model for the acceleration of the detection phase.

In this work, the trained Swin-T YOLOX lightweight model was converted into the ONNX format, and the ONNX parser in TensorRT was used to parse the model and build TensorRT engine with the TRT format. As lower data accuracy leads to lower memory footprint and latency, even a smaller model size, the model calculation accuracy was set to a 16-bit floating-point when the detection engine was built (it is usually set to a 32-bit floating-point when training). Experimental results show that the model of the Swin-T YOLOX lightweight model optimized by TensorRT can effectively reduce the floating point arithmetic and improve the detection speed of the algorithm while ensuring the accuracy.

3. Experiments and Result

3.1. Model Evaluation Metrics

In this study, the size of the model, the speed of reasoning and the detection accuracy were all adopted as the lightweight model evaluation metrics. Detection speed Frame Per Second (FPS) is used to evaluate the detection frame rate of the model, that is, the

number of images that can be detected in a second. The recognition accuracy is calculated by Equation (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where *TP* means true positive, *FP* means false positive, *TN* means true negative and *FN* means false negative. If the IoU between the detection box and the true box is greater than the threshold (it was set to 0.5 in our experiments), the detection box is marked as *TP*. Otherwise, it is marked as *FP*, and if there is no detection box matching in the true box, it is marked as *FN*. The performance of the model can usually be evaluated by precision (*Pr*) and recall (*Re*), which are calculated by Equations (3) and (4).

$$Pr = \frac{TP}{FP + TP} \tag{3}$$

$$Re = \frac{TP}{FN + TP} \tag{4}$$

In order to combine the two metrics, average precision (*AP*) is introduced to measure the detection accuracy, as defined in Equation (5).

$$AP = \int_0^1 Pr(Re) dRe \tag{5}$$

The value of *AP* is equal to the area under the precision-recall curve, and the higher the *AP* value, the higher the accuracy of the network. In the task of multi-class targets detection, the detection accuracy of the model is evaluated by calculating the average value of all types of *AP* (*mAP*), which is defined in Equation (6).

$$mAP = \frac{1}{C} \sum_{c=1}^{C} AP(c) \tag{6}$$

where *C* is the number of target categories. In this study, *mAP* was adopted to evaluate the detection accuracy of the model.

3.2. Performance of UAV Patrol Orchard Anomaly Detection

In order to test the comparative performance, the same software and hardware platforms were used for the experiment in this study. The hardware platform was equipped with Intel(R) Core(TM) I7-10700 CPU@2.90 GHz (32G RAM) and NVIDIA GeForce RTX 3090 graphics processor (24G RAM). The software environment was CUDA 11.1, CUDNN 8.0, and Python 3.8. The model was optimized by the SGD (stochastic gradient descent) method. Training epochs was set to 300 and batch size was set to 16. The initial learning rate was set to 0.001 and SGD momentum was set to 0.9. Each model configuration was trained 5 times and the mAP@0.5 and the mAP@ [0.5:0.95] of all configurations were recorded.

The original YOLOX models, such as YOLOX-M and YOLOX-Large DarkNet53 as the backbone network, were compared with Swin-T YOLOX. In terms of model depth, YOLOX-M (25.3 M) has a similar number of parameters as Swin-T and YOLOX-L (54.3 M) has a larger number of parameters. The performance comparison of the UAV patrol orchard anomaly detection is shown in Table 2.

The results in Table 2 show that the performance of Swin-T YOLOX is significantly better than DarkNet53 in the task of anomaly detection of the UAV patrol. Compared to the YOLOX-L model, Swin-T YOLOX has fewer parameters, but higher mAP@0.5 and mAP@ [0.5:0.95]. Compared to YOLOX-M, Swin-T YOLOX has a significant advantage in detection accuracy. The reason is that the Swin Transformer can capture the relationship between global feature information better than DarkNet53, showing that the Swin Transformer is more suitable for the real-time UAV patrol orchard task in this study.

Table 2. Performance comparison of the YOLOX models of different backbone networks in the UAV patrol orchard anomaly detection.

Model	Model Parameters	mAP@0.5	mAP@ [0.5:0.95]
DarkNet53 + YOLOX-M	25.3 M	92.1%	66.3%
DarkNet53 + YOLOX-L	54.3 M	93.0%	69.9%
Swin-T YOLOX(proposed)	29.0 M	93.3%	73.6%

3.3. Performance Comparison of Model Pruning Schemes

Table 3 shows the overall performance of the Swin-T YOLOX model after multi-layer pruning of the Swin Transformer network. It can be seen from the table that the recognition performance of the model does not decrease or even improve when the 10th, 11th and 12th layers are pruned, but the number of parameters and calculation amounts are only about two-thirds of that of the original Swin-T YOLOX model. However, starting from the 9th layer, the more the shallows of the Swin Transformer are pruned, the recognition accuracy not only gradually decreases, but the number of parameters and calculation amount also decrease slightly. The experimental results show that the best solution can be obtained by pruning the last three layers of the Swin Transformer network, namely the 10th, 11th and 12th layers.

Table 3. Accuracy, model parameters and calculation of the Swin-T YOLOX model for multi-layer pruning of the Swin Transformer network.

The Number of Pruning Layers	mAP@0.5	Model Parameters (M)	Calculation (G)
None	93.3%	45.89	209.70
12 layer	93.3%	38.81	182.80
11~12 layer	94.1%	31.72	155.90
10~12 layer	94.0%	29.95	146.35
9~12 layer	92.4%	28.18	136.79

3.4. Comparison of Model Deployment Schemes

In order to realize real-time UAV patrol orchard, the model transplantation to the embedded platform deployment is essential. In this paper, the Swin-T YOLOX model and the pruned Swin-T YOLOX model were deployed on different hardware platforms, and 50 sample images in the test set are randomly selected to form the test library. Table 4 is the performance comparison of the average detection speed and average recognition accuracy of the test set. The deployment on Jetson included two schemes: One is to run the detection program directly on Jetson without using the TensorRT optimizer; the other is to deploy the detection program using the TensorRT optimizer.

Table 4. Overall performance comparison of the Swin-T YOLOX model deployed using different deployment platforms.

Platform	Model	Speed (fps)	Accuracy
GPU server	Swin-T YOLOX	22	93.3%
GPU server	Lightweigh Swin-T YOLOX	48	94.0%
Jetson	Swin-T YOLOX	10	93.4%
Jetson	Lightweigh Swin-T YOLOX	20	93.9%
Jetson + TensorRT	Swin-T YOLOX	24	93.2%
Jetson + TensorRT	Lightweigh Swin-T YOLOX	40	94.0%

As shown in Table 4, the detection speed of the lightweigh Swin-T YOLOX model is greatly improved compared to the original model. In the GPU server environment, the detection speed reaches 48 fps, and the accuracy reaches 94%. After layer pruning and TensorRT optimization, the Swin-T YOLOX reaches 40fps on the Jetson Xavier NX,

and the recognition rate is still high. The accuracy does not decrease while pruning, but slightly increases. The experimental results show that the proposed model, Swin-T YOLOX, after pruning and optimization, can effectively compress the model size and floating point arithmetic, thus guaranteeing the accuracy while improving the detection speed of the algorithm. The visualization of the orchard UAV patrol is shown in Figure 10.

Figure 10. The effect of the UAV patrol orchard. orange box indicates uneven planting or felled areas, blue box indicates yellow suspected disease plants.

4. Discussion

4.1. Data Process

The UAV remote sensing panoramic image in this study is too large, which is not conducive to real-time analysis, so the image processing was carried out by cutting the image. One of the tasks of the real-time UAV patrol orchard is to detect uneven planting or cutting areas where there is bare land among the trees planted in rows. However, pictures taken from the perspective of UAV tend to block such target areas in the edge area. As shown in Figure 11, there were no citrus trees in the original three locations in the red box, but most of the area was obscured by the adjacent tree crowns. In order to improve the recognition accuracy, it is necessary to cut out the edge area of the picture, that is, the area where the UAV cannot observe the ground vertically and keep the area in the blue box, as shown in Figure 11. Therefore, it is necessary to increase the overlap rate of view coverage when the UAV plans the path so as to ensure that the UAV can patrol the whole orchard. While this processing can improve the recognition accuracy, it increases the flight distance and time of the UAV.

4.2. Method of Model Lightweight

In order to meet the requirements of embedded terminal deployment on UAVs, most people will choose to train small models and optimize them through some tools. However, this paper chooses the method of layer pruning for the large model to compress the size of the model, because the performance of the model after training a large model then compressing it into a small model is better than that of directly training a small model. Moreover, the large model has the advantages of lower sensitivity and easier compression, which is the experimental conclusion proposed by Li et al. [34]. The experimental results show that the large model has high accuracy and robustness after compression. Through the compression of the model, the speed of the pre-trained large model is faster than that of the small model in the inference stage.

Figure 11. Vertical view of a citrus orchard from a drone. The blue square shows the area where the UAV can vertically observe the ground. The red box shows the problem of canopy occlusion.

4.3. Application and Future Work Directions

The main contributions of this paper are as follows:

(1) A lightweight network Swin-T YOLOX is proposed to be used for real-time UAV patrol orchard task. In this algorithm, the Swin Transformer is used to replace Darknet53 as the backbone network of YOLOX, which can significantly improve the model accuracy;
(2) Layer pruning technology is used to reduce the number of parameters and calculation amount of the Swin-T YOLOX;
(3) Deploy the Swin-T YOLOX model to the Jetson Xavier NX edge computing platform using the TensorRt optimizer, and test the detection performance of the algorithm on the embedded platform.

However, there is still an obvious shortage in this paper. In the UAV patrol orchard anomaly detection task in this study, only yellow plants can be detected from RGB remote sensing images, it is hard to confirm whether they are diseased plants. Sickness, nitrogen deficiency, potassium, magnesium and other nutrients, root damage or improper fertilization will lead to the yellowing of canopy leaves. It is difficult to judge the specific situation of citrus trees only from the UAV RGB images, ground diagnosis is required. However, the real-time UAV patrol orchard mode can quickly locate suspicious plants and abnormal areas in the whole orchard, accurately record the location information of abnormal plants or areas and greatly improve the patrol efficiency.

In the future, this research will continue to be conducted in the following two directions. First, in addition to the vertical angle adopted in this paper, a variety of shooting angles can be explored in the future, and the detailed diagnosis of abnormal areas can be explored through a variety of sensing methods. Second, since the multi-spectrum is widely used to analyze the nutritional status and insect pests of citrus trees, the Swin-T YOLOX lightweight model will be explored to the field of a UAV multispectral real-time remote sensing agricultural situation in the future.

5. Conclusions

This paper introduces a lightweight object detection model, Swin-T YOLOX, for UAV patrol orchard anomaly detection by replacing the original YOLOX backbone network with the Swin Transformer and applying layer pruning to cut the last three layers of the Swin Transformer network to form a lightweight model. After replacing the backbone, the accuracy of the Swin-T YOLOX achieved an accuracy improvement of 1.9% compared to the YOLOX model. After pruning, the number of parameters is two-thirds of that of

the original Swin-T YOLOX, and the recognition performance is improved by 0.7%. The lightweight Swin-T YOLOX model can be deployed on the Jetson Xavier NX by TensorRT. Working as a mobile terminal, the detection speed reaches 40 fps and the accuracy rate is maintained at 94.0%, which well completes the goal of the lightweight target detection algorithm that takes into account both detection accuracy and processing speed. The model can be applied to the mission of the real-time UAV patrol orchard.

Author Contributions: S.L. conceptualized the experiment, selected the algorithms, collected and analyzed the data, and wrote the manuscript. H.D. and Y.G. trained the algorithms, collected and analyzed data, and wrote the manuscript. X.D. and Y.L. supervised the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was supported by the Key-Areas of Artificial Intelligence in General Colleges and Universities of Guangdong Province (Grant No. 2019KZDZX1012), the Laboratory of Lingnan Modern Agriculture Project (Grant No. NT2021009), the Key-Area Research and Development Program of Guangzhou (Grant No. 202103000090), the Key-Area Research and Development Program of Guangdong Province (Grant No. 2019B020214003), the National Natural Science Foundation of China (Grant No. 61675003), the National Natural Science Foundation of China (Grant No. 61906074), and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2019A1515011276).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available because the dataset involves the research work of other scholars in our team.

Acknowledgments: S.L. thank H.D. for supporting and helping, X.D. and Y.L. for supervision, and Y.G. for revising the article.

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

References

1. Wang, H.; Zhou, W.; Li, J. Current Status, Problems and Development Trend of the Wisdom Agriculture Research in China. *J. Anhui Agric. Sci.* **2016**, *44*, 279–282.
2. Wang, L.; Li, K.; Shan, H. Design of Small-scale Intelligent Orchard System. *Agric. Eng.* **2021**, *11*, 55–61.
3. Fan, Y.; Zhang, Z.; Chen, G.; Li, B. Research on Monitoring and Analysis System of Corn Growth in Precision Agriculture Based on Internet of Things. *J. Agric. Mech. Res.* **2018**, *40*, 223–227.
4. Zhang, K.; Zhang, S.; Lian, X. Design of cruise inspection system for four-rotor autonomous aircraft in orchard. *J. Chin. Agric. Mech.* **2017**, *38*, 81–85.
5. Gao, X.; Xie, J.; Hu, D. Application of Quadrotor UAV in the Inspection System of Citrus Orchard. *Process Autom. Instrum.* **2016**, *36*, 26–30.
6. Nikolaos, S.; Haluk, B.; Volkan, I. Vision-based monitoring of orchards with UAVs. *Comput. Electron. Agric.* **2019**, *163*, 104814.
7. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
8. Shi, T.; Zhang, X.; Guo, L. Research on remote sensing recognition of wild planted Lonicera japonica based on deep convolutional neural network. *China J. Chin. Mater. Med.* **2020**, *45*, 5658–5662.
9. Deng, X.; Tong, Z.; Lan, Y. Detection and location of dead trees with pine wilt disease based on deep learning and UAV remote sensing. *AgriEngineering* **2020**, *2*, 294–307. [CrossRef]
10. Mo, J.; Lan, Y.; Yang, D. Deep learning-based instance segmentation method of litchi canopy from UAV-acquired images. *Remote Sens.* **2021**, *13*, 3919. [CrossRef]
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 (CVPR), Columbus, OH, USA, 20–23 June 2014.
12. Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis. (IJCV)* **2013**, *104*, 154–171. [CrossRef]
13. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), Boston, MA, USA, 7–10 June 2015.
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7–12 December 2015.
15. 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 (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
16. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

17. Farhadi, A.; Redmon, J. Yolov3: An incremental improvement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
18. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 10–16 October 2016.
19. Zheng, G.; Songtao, L.; Feng, W.; Zeming, L.; Jian, S. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
20. 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 Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 3–7 May 2021.
22. 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 International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
23. Wang, H.; Hu, X.; Zhang, Q. Structured pruning for efficient convolutional neural networks via incremental regularization. *IEEE J. Sel. Top. Signal Process.* **2019**, *14*, 775–788. [CrossRef]
24. Liu, Z.; Li, J.; Shen, Z.; Huang, G.; Yan, S.; Zhang, C. Learning efficient convolutional networks through network slimming. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
25. Li, H.; Kadav, A.; Durdanovic, I.; Samet, H.; Graf, H.P. Pruning filters for efficient convnets. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
26. Qi, P.; Sha, E.H.M.; Zhuge, Q. Accelerating framework of transformer by hardware design and model compression co-optimization. In Proceedings of the IEEE/ACM International Conference On Computer Aided Design (ICCAD), Wuxi, China, 22–23 December 2021.
27. Yu, S.; Chen, T.; Shen, J. Unified visual transformer compression. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 25–29 April 2022.
28. Hou, Z.; Kung, S.Y. Multi-dimensional model compression of vision transformer. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Taiyuan, China, 27–28 August 2022.
29. Mao, J.; Yang, H.; Li, A.; Li, H.; Chen, Y. TPrune: Efficient transformer pruning for mobile devices. *ACM Transact. Cyber-Phys. Syst.* **2021**, *5*, 1–22. [CrossRef]
30. DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* **2017**, arXiv:1708.04552.
31. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond empirical risk minimization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
32. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015.
33. Han, S.; Mao, H.; Dally, W.J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In Proceedings of the International Conference on Learning Representations (ICLR), SAN Juan, PR, USA, 2–4 May 2016.
34. Li, Z.; Wallace, E.; Shen, S. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In Proceedings of the International Conference on International Conference on Machine Learning (ICML), Online, 13–18 July 2020.