agronomy

Article

Taoism-Net: A Fruit Tree Segmentation Model Based on Minimalism Design for UAV Camera

Yanheng Mai [1], Jiaqi Zheng [1], Zefeng Luo [1], Chaoran Yu [2], Jianqiang Lu [1,3], Caili Yu [4,*], Zuanhui Lin [1,*] and Zhongliang Liao [1]

[1] College of Electronic Engineering (College of Artificial Intelligence), South China Agricultural University, Guangzhou 510642, China; scau_myh@stu.scau.edu.cn (Y.M.); 2965499424@stu.scau.edu.cn (Z.L.); ljq@scau.edu.cn (J.L.)
[2] Vegetable Research Institute, Guangdong Academy of Agricultural Sciences Guangdong Key Laboratory for New Technology Research of Vegetables, Guangzhou 510000, China
[3] National Center for International Collaboration Research on Precision Agricultural Aviation Pesticides Spraying Technology (NPAAC), Guangzhou 510642, China
[4] Shanwei Institute of Technology, Center for Intelligent Perception and Internet of Things Research, Shanwei 516600, China
* Correspondence: scau_ai_team@163.com (C.Y.); lzh@scau.edu.cn (Z.L.); Tel.: +86-19120393823 (C.Y.); +86-13725492904 (Z.L.)

Abstract: The development of precision agriculture requires unmanned aerial vehicles (UAVs) to collect diverse data, such as RGB images, 3D point clouds, and hyperspectral images. Recently, convolutional networks have made remarkable progress in downstream visual tasks, while often disregarding the trade-off between accuracy and speed in UAV-based segmentation tasks. The study aims to provide further valuable insights using an efficient model named Taoism-Net. The findings include the following: (1) Prescription maps in agricultural UAVs requires pixel-level precise segmentation, with many focusing solely on accuracy at the expense of real-time processing capabilities, being incapable of satisfying the expectations of practical tasks. (2) Taoism-Net is a refreshingly segmented model, overcoming the challenges of complexity in deep learning, based on minimalist design, which is used to generate prescription maps through pixel level classification mapping of geodetic coordinates (the lychee tree aerial dataset in Guangdong is used for experiments). (3) Compared with mainstream lightweight models or mature segmentation algorithms, Taoism-Net achieves significant improvements, including an improvement of at least 4.8% in mIoU, and manifested a superior performance in the accuracy–latency curve. (4) "The greatest truths are concise" is a saying widely spread by ancient Taoism, indicating that the most fundamental approach is reflected through the utmost minimalism; moreover, Taoism-Net expects to a build bridge between academic research and industrial deployment, for example, UAVs in precision agriculture.

Keywords: precision agriculture; instance segmentation; UAV; deep learning; computer vision

Citation: Mai, Y.; Zheng, J.; Luo, Z.; Yu, C.; Lu, J.; Yu, C.; Lin, Z.; Liao, Z. Taoism-Net: A Fruit Tree Segmentation Model Based on Minimalism Design for UAV Camera. *Agronomy* **2024**, *14*, 1155. https://doi.org/10.3390/agronomy14061155

Academic Editor: Yanbo Huang

Received: 26 April 2024
Revised: 19 May 2024
Accepted: 21 May 2024
Published: 28 May 2024

1. Introduction

Fruit tree cultivation occupies a central role in the global agricultural production arena, standing as an indispensable component of the agricultural economy that furnishes a sturdy economic pillar for agricultural producers worldwide. Amidst the surge in global population and the escalating demand for healthy food products, compounded by the supply–demand imbalance engendered by the migration of agricultural labor to other sectors, the fruit cultivation industry confronts unparalleled challenges, necessitating technological innovations to adapt to the evolving market demands. Within this context, ensuring the robust growth of flowers and fruits, alongside the judicious implementation of thinning operations, imposes heightened requirements on the precision of canopy recognition technologies. The accurate assessment of tree canopy structures becomes vital for efficiently undertaking tasks such as pruning branches.

Over the course of recent decades, propelled by relentless technological advancement, UAV technology has undergone remarkable maturation within the agricultural sphere. Particularly noteworthy is the pervasive application of deep learning paradigms in the realm of computer vision, affording the feasible integration of UAVs endowed with such sophisticated models for agricultural pursuits. Dai Ming [1] et al. proposed a straightforward and efficient Transformer architecture, named FSRA, designed to tackle challenges such as positional drift and scale ambiguity inherent in UAV-based geolocation tasks. While this approach circumvents the need for additional supervision, the deployment of Transformers on edge devices remains fraught with challenges, chiefly due to constraints imposed by inference speed limitations. Onishi [2] attempted to use drone RGB images and deep learning techniques to classify individual trees, but ignored the challenge of different drone perspectives on segmentation tasks.

In R-CNN [3], the segmentation mask of the object is input into the network to extract features for classification. Faster R-CNN [4] and SPPNet [5] accelerate this process by pooling features from the global feature map. Early work used the mask of MCG [6] as input to extract features, and MaskR-CNN [7] is an effective framework belonging to this category. SharpMask [8] and LRR [9] fuse feature maps for fine detail segmentation. FCN [10], U-Net [11], and Noh [12] fuse information from lower layers through skip connections. Both TDM [13] and FPN [14] enhance the top-down path through horizontal connections. The BERT proposed by BaoHangbo [15] utilizes a pre training task called Masked Image Modeling (MIM) to recover original image labels, using a Transformer backbone network similar to ViT, which is widely used in classification task implementation.

Despite the commendable performance attained by convolutional neural networks and Transformers, the concomitant escalation in computational overhead and conundrums pertaining to stability have engendered certain impediments for UAV deployment. Consequently, antecedent iterations of UAV-deployed deep learning architectures have leaned towards the adoption of lightweight frameworks, albeit often at the expense of detection acuity. Conversely, certain deep learning models predicated upon cloud-based infrastructures for detection manifest latency phenomena. In response to these exigencies, we propose an efficacious and parsimonious model christened Taoism-Net. The technical roadmap and workflow of this article are shown in Figure 1.

The main contributions of this article are summarized as follows: (1) In practical verification, the feasibility of establishing high-performance neural networks without complex architectures such as residuals, high depth, and attention layers reflects a shift in design towards a simple and elegant paradigm. (2) Since we use a deep training strategy and concatenated activation function, the non-linear fitting ability used in the training and testing process of Taoism-Net improves its performance, and it is comparable to well-known deep neural networks, thus highlighting the potential of minimalism in deep learning.(3) The optimal speed–precision trade-off is achieved and the GPU delay is low, which indicates that the proposed Taoism-Net has advantages under the conditions of sufficient computing power and UAV edge computing. (4) We use Diceloss to solve sample imbalance problems and provide higher detection accuracy for single classification problems. In terms of visual neural network design, this article aims to build a practical bridge between academic research and industrial deployment, hoping to provide new insights for computer vision work in the actual agricultural field and promote more research on neural network architecture design.

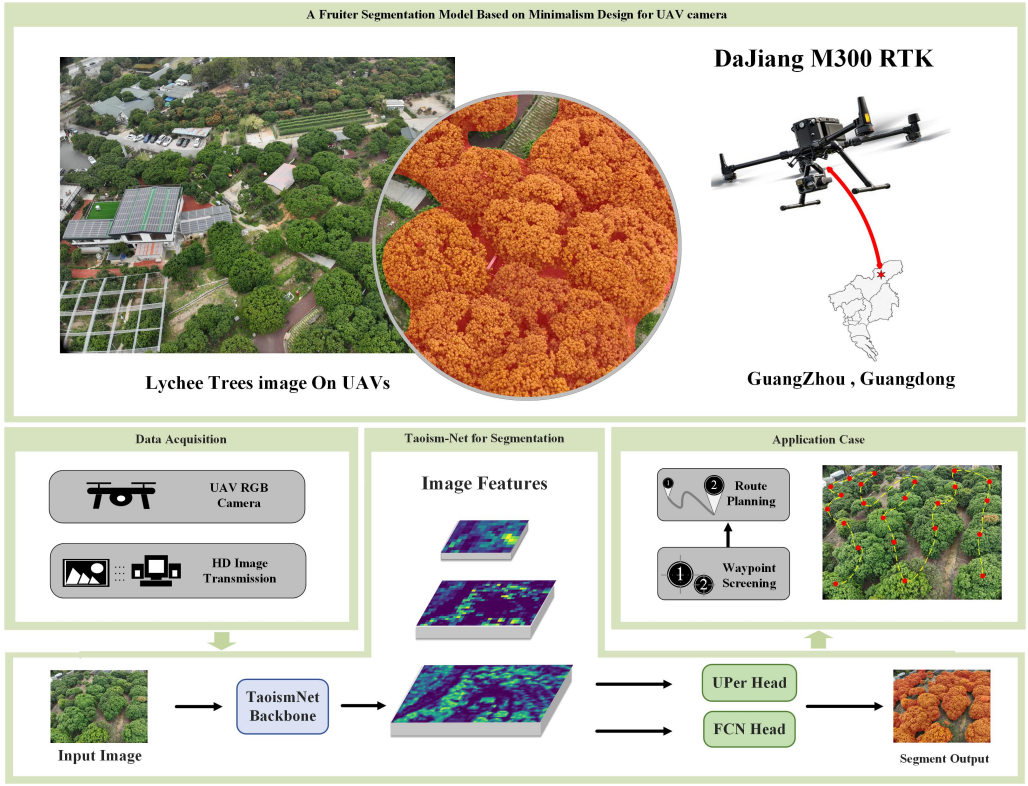


Figure 1. Technical roadmap and workflow. In the lychee orchard in Guangzhou, Guangdong Province, DJI M300 RTK took off to collect RGB images of lychee and other fruit trees from different angles. By downloading data such as relative altitude and focal length from DJI M300 RTK, it was determined that the real-world size corresponding to a single pixel in all images ranges from 1.24 m to 2.49 m. The trained model can effectively and accurately segment the fruit trees in the field of view, laying a visual algorithm foundation for mapping geodetic coordinates and other agricultural activities in the future. The figure shows the basic structure of the backbone network and the output segmentation heads, further serving various application scenarios, such as waypoint planning.

2. Materials and Methods

2.1. Materials of Dataset

The application of agricultural drones is witnessing rapid advancements; however, the progression of visual algorithms has veered towards a direction that contradicts practical applications. In recent years, state-of-the-art backbone networks have hinged on computationally intensive feature extraction to achieve enhanced accuracy. This paper introduces Taoism-Net, which benchmarks its model using images of lychee tree canopies captured by agricultural drones. The relative working heights, with reference to the takeoff point, range from 24.9 m to 50.0 m, averaging at 32.8 m. The dataset, sourced from the RGB channels of drone-mounted sensors (depicted in Figure 2), boasts a resolution of 5280 × 3956 pixels. The data were gathered at the Guangzhou Lychee Cultural Museum, spanning several acres and encompassing photographs of lychee canopies exhibiting diverse characteristics across different periods, with the collection period spanning from February to May, covering the stages from bud initiation to fruit maturity.In recognition of the variances in real-world scenarios and to assess the model's generalization capability, multiple lychee orchards were photographed. Considering image edge distortions, multi-angle captures were executed to accumulate a more comprehensive dataset. Relative to other prevalent segmentation models, our proposed method demonstrates structural superiority in mapping the representational space, ensuring effectiveness across datasets of similar canopy structures from different crops. Figure 2 presents exemplar images from the dataset, illustrating shooting angles varying approximately from 30° (Figure 2a) to around 60° (Figure 2b) and up to about 90° (Figure 2c). Furthermore, visual pixel points are mapped to geographical

coordinates, enabling the precise localization of lychee tree canopies. This foundational visual algorithmic approach paves the way for downstream agricultural tasks, furnishing them with accurate spatial information.

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

Figure 2. The returned image from UAVs RGB sensor at various views. (**a**) View at 30°; (**b**) view at 60°; (**c**) view at 90°.

2.2. Model of Concise Design

The fundamental design of neural networks has reached some consensus, with input images transformed from 3 channels to multiple channels and downsampled, and fully connected layers used for classification output. After each stage, the channels of features expand while the height and width decrease, and different networks utilize and stack different types of blocks to build deep models. Despite the success of existing deep networks, their works employ abundant complex layers to extract high-level features from the receptive field. For example, the convolutional layers of the classical AlexNet [16] use 96 convolutional kernels to operate on input images, with a kernel size of 11 × 11, while reducing the number of feature maps to avoid the risk of overfitting. The base version of VGG [17], VGG11, has 8 convolutional layers and 3 fully connected layers, with 5 pooling layers following each convolutional layer, making the architecture very complex. VGG19 has 16 convolutional layers, making the network even more complex.

Therefore, we propose Taoism-Net for the visual segmentation task of drones, emphasizing elegant and simple design while maintaining an excellent performance in latency. Taoism-Net achieves feature extraction by avoiding excessive depth, residual, and complex operations (such as self-attention), and the backbone network has no branches or additional blocks, as the impact of the residual method on performance improvement is small in practical testing. The working principle of its backbone network is shown in Figure 3. This paper proposes three types of Taoism-Net with different parameters, which can be used for the pre-training and knowledge distillation of the model. Table 1 shows the number of small convolution kernels inside the model, pooling layers, and the size of feature maps for each layer.

Table 1. Internal network structure of Taoism-Net of different sizes.

Sstages	Input Size	Layers	Taoism-Net-v6	Taoism-Net-v9	Taoism-Net-v13
Stem	2048 × 2048	Convolution Layer	[4 × 4, 512] stride 4		
Stage 1	512 × 512	Taoism-Net Block	[2 × 2, 1024] × 1 MaxPool 2 × 2	[2 × 2, 1024] × 2 MaxPool 2 × 2	[2 × 2, 1024] × 2 MaxPool 2 × 2
Stage2	128 × 128	Taoism-Net Block	[2 × 2, 2048] × 1 MaxPool 2 × 2		
Stage 3	32 × 32	Taoism-Net Block	[2 × 2, 4096] × 1 MaxPool 2 × 2	[2 × 2, 4096] × 3 MaxPool 2 × 2	[2 × 2, 4096] × 7 MaxPool 2 × 2

Table 1. *Cont.*

Sstages	Input Size	Layers	Taoism-Net-v6	Taoism-Net-v9	Taoism-Net-v13
Stage 4	8×8	Taoism-Net Block		$[2 \times 2, 4096] \times 1$	
Classifer	8×8	Fully Connected Layer		AvgPool 7×7 1×1, 1000	

The network structure parameters of different sizes for Taoism-Net. Its network mainly includes Convolution Layers, Taoism-Net Blocks, and Fully Connected Layers. Taoism-Net Block is used in stages 1–4, consisting of a combination of single or multiple 1×1 convolutional layers with 2×2 pooling layers. This article provides model sizes with different computational costs, and customs can also adjust the number of convolution kernels for the Taoism-Net Block according to the actual task needs.

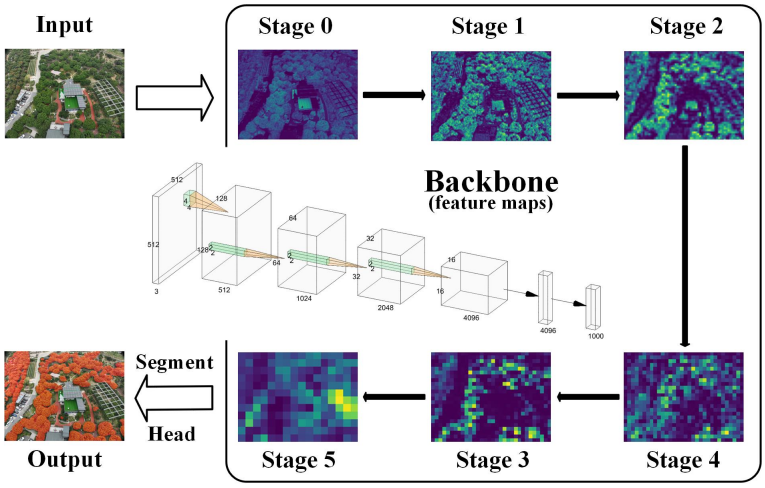


Figure 3. The working principle of backbone networks: taking six layers as an example, they extract local and global features layer by layer, fit nonlinear features through concurrent activation functions, and finally output segmentation results through UPerHead [18] and FCNHead [10].

2.3. Deep Training Strategy and Nonlinear Activation

To train the Taoism-Net, we embarked on a meticulous analysis of the intricacies surrounding its streamlined architectural design, culminating in the formulation of a sophisticated "deep training" strategy. This progressive strategy initiates with the integration of several layers, each imbued with nonlinear activation functions. As the iterative process of training unfolds, the model systematically prunes these nonlinear layers, affording a seamless amalgamation while upholding the imperative of swift inference.

Here, we illustrate the architecture of the network with 6 layers as an example. For the stem, we use a $4 \times 4 \times 3 \times C$ convolutional layer with a stride of 4, following popular settings in ResNet [19], Swin Transformer [18], and ConvNext [20], mapping images with 3 channels to features with C channels. In stages 1, 2, and 3, max-pooling layers with a stride of 2 are used to reduce size and feature maps, with the number of channels increasing by 2.

When it is necessary to focus on feature learning in the center, we usually use odd convolution kernels to reduce the dimensionality of the feature map. The comparison shown in Figure 4 illustrates the efforts made by previous scholars, who attempted to use two 3×3 convolution kernels instead of 5×5 or even larger 7×7 convolution kernels. In the experiment of the paper, it was found that for the visual task of extracting fruit tree canopy features based on drones, the neural network pays more attention to the texture features of the target. This can also be found in subsequent thermal map analysis, as shown in Figure 8. Therefore, by using two 2×2 convolutional kernels combined with activation functions, we can achieve sufficient detection accuracy and accelerate the training and inference speed of the model. In subsequent network layers, we do not increase the number

of channels, as it follows average pooling layers. The last layer is a fully connected layer for outputting classification results. The kernel size for each convolutional layer is 2×2, as our goal is to use minimal per-layer computational cost while preserving feature map information. Activation functions are applied after each 2×2 convolutional layer. To simplify the training process of the network, layer normalization is also included after each layer.

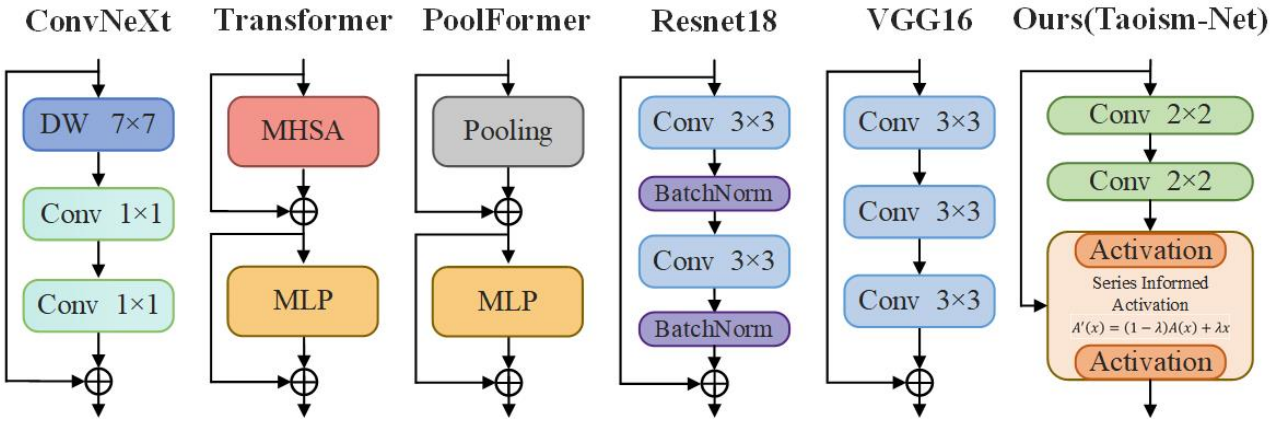


Figure 4. Comparison of different Transformer-based and convolution-based blocks.

As the iterative epochs of training, the intricacies of the activation functions gradually coalesce into a semblance of identity mappings. Upon the culmination of the training regime, the confluence of two convolutional operations into a singular entity emerges as a facile endeavor, thereby precipitating a discernible reduction in inference latency. This methodology, emblematic of its efficacy, finds resonance within the broader milieu of convolutional neural networks.

For the activation function $A(x)$ (which can be ordinary functions such as ReLU and Tanh), its linear transformation can be expressed as (1)

$$A'(x) = (1-\lambda)A(x) + \lambda x \tag{1}$$

where λ is a hyperparameter used to balance the nonlinearity of the modified activation function $A'(x)$.

The continuous superposition of non-linearity of activation functions is the core idea of deep networks. On the contrary, we instead stack simple activation functions concurrently; for example, the single activation function representing input x in a neural network is $A(x)$, which can be commonly used functions such as ReLU and Tanh. The concurrent superposition of $A(x)$ can be expressed as (2)

$$A_s(x) = \sum_{i=1}^{n} a_i A(x + b_i) \tag{2}$$

where n is the number of series informed activation functions, a_i is the scale and b_i is the bias of each activation, and these functions can be stacked simultaneously to greatly enhance the nonlinear characteristics of the activation functions. As shown in Figure 5, compared to traditional activation functions that are more difficult to calculate, the multi-layer stacked Series Informed Activation can exhibit faster training and prediction capabilities without losing non-linear fitting ability.

In order to further enrich the fitting ability of the backbone network, sequence-based functions learn global information by changing the input of adjacent regions, similar to BNET [21], where H, W, and C are their width, height, and number of channels, respectively. The activation function is represented as (3)

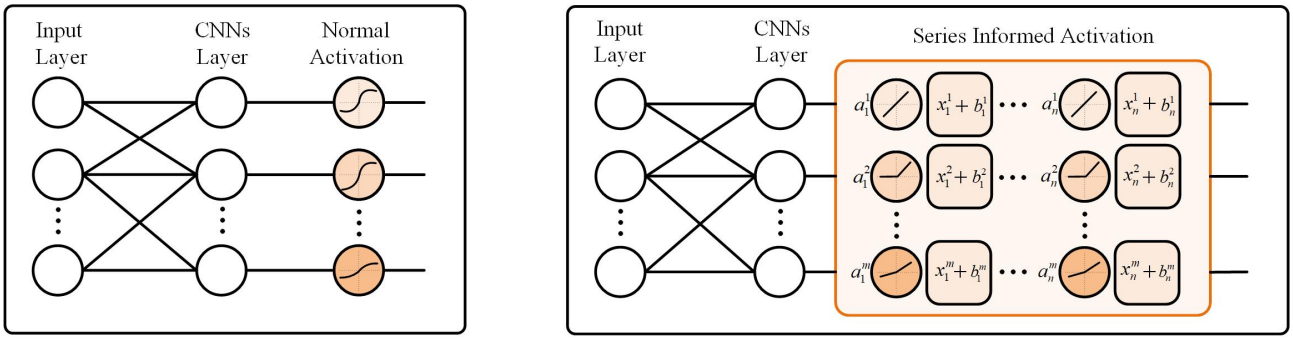
$$A_s(x_h, w, c) = \sum_{i,j \in \{-n,n\}} a_i, j, cA(x_{i+h,j+w,c} + b_c) \tag{3}$$

When $n = 0$, it is obvious that the series based activation function degenerates into a regular activation function, which can be seen as a general extension of existing activation functions. We use the ReLU variant as the basic activation function to construct the series, as its inference in GPU is efficient. The related computational cost can be expressed as Equations (4)–(6), and the computational cost of the above methods is much lower than most CNNs.

$$O(CONV) = H \times W \times C_{in} \times C_{out} \times k^2 \tag{4}$$

$$O(SA) = H \times W \times C_{in} \times n^2 \tag{5}$$

$$\frac{O(CONV)}{O(SA)} = \frac{H \times W \times C_{in} \times C_{out} \times k^2}{H \times W \times C_{in} \times n^2} = \frac{C_{out} \times k^2}{n^2} \tag{6}$$



(a) Activation Function of Traditional CNNs Backbone

(b) Series Informed Activation Function of Taoism-Net Backbone

Figure 5. Comparison with Series Informed Activation and Traditional CNNs Activation. Experiments have shown that a backbone using Series Informed Activation can achieve better nonlinear fitting ability at lower computational costs, which is crucial for the minimalist design of the model.

2.4. Dice Loss Function for Imbalance Smaple

In the dataset, there is only one category of samples, and the focal length of the drone lens and the altitude of flight cause a certain degree of sample imbalance. The commonly used cross entropy loss [22] is better at classification problems. At present, the popular Focal Loss [23] is more commonly used to solve sample imbalance in binary and even multi classification. However, the problem of sample imbalance is more common in drone image acquisition, such as the similarity of color features to other plants and the unclear edge features caused by texture features. Unlike cross entropy loss, Dice Loss is proposed in VNet [24], which does not consider the relationship between categories and only focuses on the overlap between predicted and actual results. Its calculation process can be formulated as Equations (7)–(9):

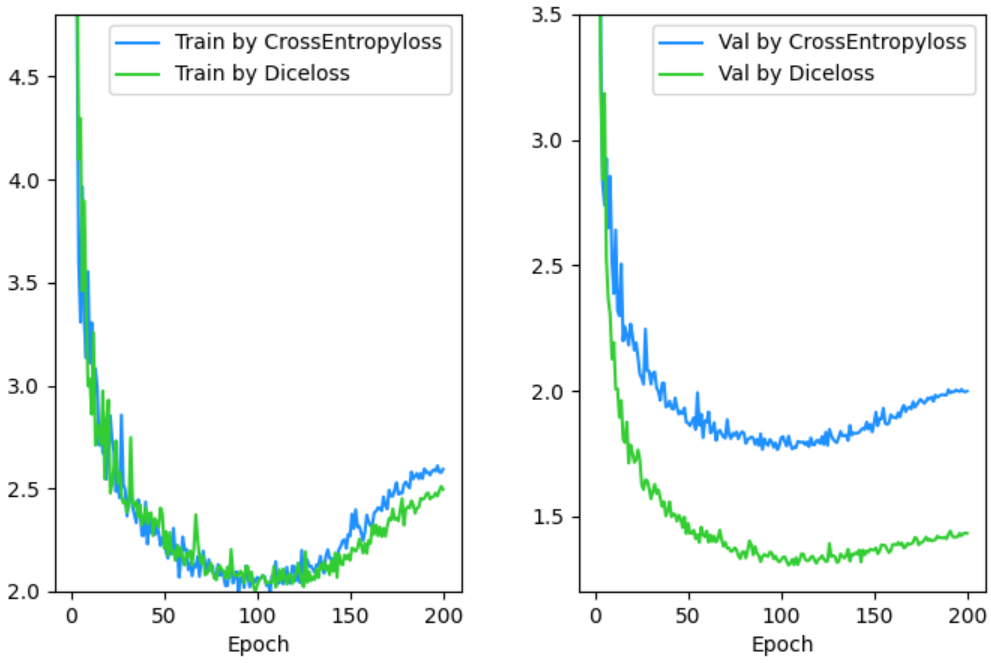
$$dice = \frac{2p1 * y1 + \epsilon}{p1 + y1 + \epsilon} \tag{7}$$

$$DL = 1 - \frac{2p1y1 + \epsilon}{p_1^2 + y_1^2 + \epsilon} \tag{8}$$

$$DSC = 1 - \frac{2(1 - p_1)p_1 \cdot y_1 + \epsilon}{(1 - p_1)p_1 + y_1 + \epsilon} \tag{9}$$

The Loss curve refers to the process of changing the loss function during model training. The loss function is a parameter used to evaluate the prediction deviation of a

model. The image shown in Figure 6b samples the segmentation validation loss using CrossEntropyLoss and DiceLoss during the training process, while the overall fusion validation loss comparison curve is shown in Figure 6a, with the green line representing the loss function using DiceLoss.



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

Figure 6. Comparison with DiceLoss and CrossEntropyLoss during Training and Validation. (**a**) Train loss; (**b**) Val loss.

3. Results

After multiple experiments, the training hyperparameters were finally determined, with an iteration count of 200 and a batch size of 16 per iteration. The initial learning rate was 0.05. The learning rate reduction method using cosine annealing has been contributed to adjust model's learning rate more precisely during the training process, improving convergence performance. When it comes to calculating the loss function, DiceLoss is used for solving the problem about objective imbalance. In this specific dataset, UAVs collect the images from various views and different distance, leading to a situation that some image is easy to segment but the other is difficult, and cause an unexpected shock during the training period. DiceLoss is derived from the dice coefficient and is a measurement function used to measure set similarity. It is usually used to calculate the similarity between two samples and, moreover, to successfully relieve the impact of an imbalanced sample. The code accompanying this article facilitates the conversion of model formats, as illustrated in Appendix A.

3.1. Comparison with Mainstream Segment Models

As shown in Table 2, whether compared to mainstream lightweight image segmentation models such as U-Net [25], PP-LiteSeg [26], or mature image segmentation algorithms such as DeepLabv3+ [27] and PSPNET [28], Taoism-Net has achieved significant improvement. For the dataset of this experiment, it achieved an improvement of at least 3.6% in mIoU value and won both in accuracy and speed. The method proposed in this paper benefits from the efficient backbone of Taoism, emphasizing the elegance and simplicity of design while maintaining excellent performance in computer vision tasks. Taoism-Net improves recognition accuracy by avoiding excessive depth, residuals, and complex operations such as self attention. Based on a series of simplified networks, it solves inherently complex

problems and is very suitable for environments with limited resources during inference operations on UAVs. As shown in Figure 7, Taoism-Net achieves a significant improvement in accuracy speed curve testing compared to mainstream segmentation models.

Table 2. Comparison with mainstream segmentation models.

Model	Param (/M)	FLOPs (/G)	Latency (/ms)	(fps) (/s)	mIoU (/%)	Accruancy (/%)
U-Net-vGG [25]	24.89	451.67	113.66	8.8	82.82	90.61
U-Net-ResNet50 [25]	43.93	184.10	125.95	7.9	84.71	91.72
PP-LiteSeg-STDC1 [26]	8.214	9.904	30	33.3	82.85	90.65
PP-LiteSeg-STDC2 [26]	12.25	25.06	32	31.25	84.32	91.52
HR-Net-W18 [29]	9.64	37.32	43.51	23.0	84.30	91.12
HR-Net-W32 [29]	29.54	90.93	80.87	8.21	83.71	91.14
HR-Net-W48 [29]	65.87	187.67	103.05	8.13	82.48	90.43
DeepLabv3-MoibleNet [27]	5.813	52.87	86.70	19.05	83.79	91.18
DeepLabv3-Xception [27]	54.71	166.84	109.61	8.15	82.90	90.66
PSPNET-MobileNet [28]	2.376	6.031	57.19	10.6	81.31	90.56
PSPNET-ResNet50 [28]	46.79	118.43	86.11	10.4	82.94	90.94
Taoism-Net-v6	71.44	233.63	**12.36**	**80.9**	85.74	93.52
Taoism-Net-v9	79.95	260.51	25.50	64.5	87.90	94.86
Taoism-Net-v13	88.45	287.40	47.85	44.9	**88.32**	**96.59**

The device used for the above benchmark test has an AMD Ryzen77840H w/Radeon 780M Graphics card (16 CPUs) with 16 GB of RAM, and an NVIDIA GeForce RT 4060 Laptop GPU graphics card with a VRAM of 7971 MB.

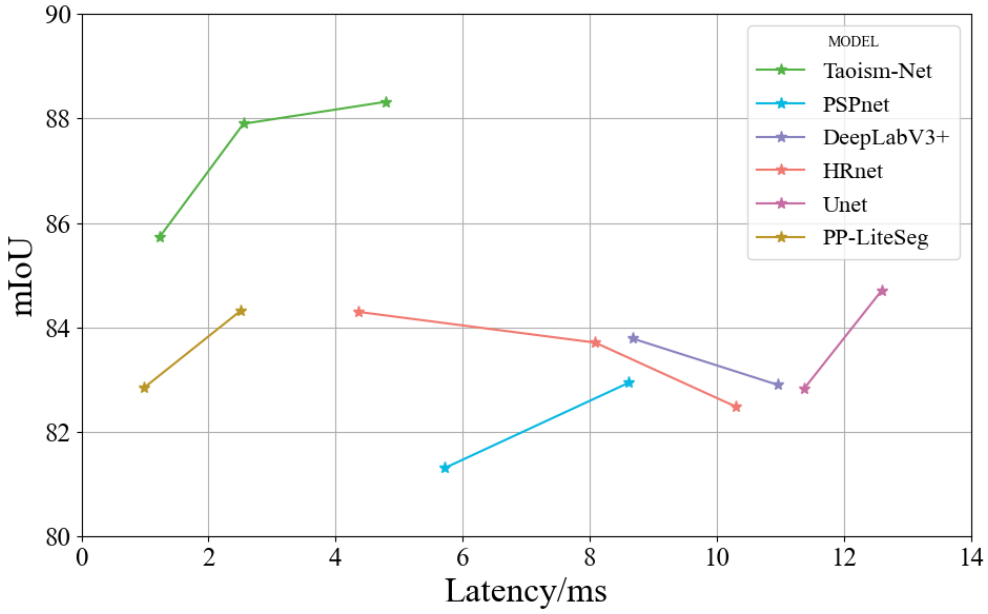


Figure 7. The device used for the above benchmark test has an AMD Ryzen77840H w/Radeon 780M Graphics card (16 CPUs) with 16 GB of RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU graphics card with a VRAM of 7971 MB.

3.2. Ablation Experiment

To verify the proposed backbone network and the Diceloss used in this study, ablation experiments were conducted on the dataset. A fair comparison was made using the encoder–decoder framework provided by mmsegmentation. Backbones were replaced and the same decoder was used. All other hyperparameter configurations were consistent. The ablation experiment results are shown in Table 3. From Table 3, it can be seen that the mIoU value of the benchmark model is 85.2%, and the mIoU value of Taoism-Net shows an increase of 1.4%; after adopting Diceloss, mIoU increased by 3.1%.

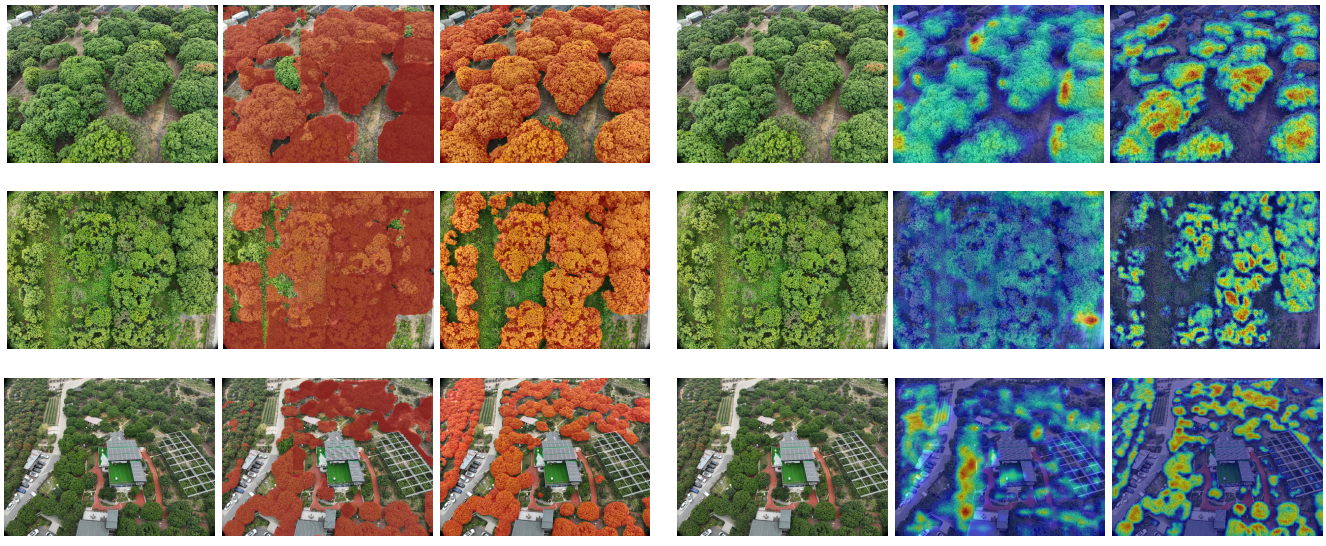
Table 3. Ablation Experiment.

Strategies	Param (/M)	FLOPs (/G)	Latency (/ms)	MIoU (/%)
Encoder–SwinTransformer [30]	59.02	237.08	86.32	85.20
Encoder–Taoism-Net [30]	88.45	287.40	40.27	86.49
Encoder–Taoism-Net and Diceloss [30]	88.45	287.40	37.85	88.32

The device used for the above benchmark test has an AMD Ryzen77840H w/Radeon 780M Graphics card (16 CPUs) with 16 GB of RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU graphics card with a VRAM of 7971 MB.

3.3. HeapMap Analysis Based on Grad-CAM++

In order to demonstrate the superiority of Taoism-Net in feature extraction, for the weight files that have already been trained, this paper uses the Grad CAM++ [31] method to draw a heatmap, which visualizes the attention of different models to different features. The darker the red color in the heatmap, the greater the contribution of the model's region to the final prediction result, and the higher the attention paid to this part of the image. As shown in Figure 8, Taoism-Net has a stronger focus and is concentrated at the center of the canopy, gradually decreasing from the center and edges, effectively distinguishing the target edges. In comparison to the baseline model, YOLOv8s-seg performs better in segmenting tree canopies when viewed at a 90-degree nadir angle but is also prone to missing detection objects to some extent. At a 30-degree perspective, its performance deteriorates particularly for distant, smaller objects. Conversely, Taoism-Net, leveraging composite 2×2 convolutional kernels in deeper layers, has demonstrated through experimental outcomes that it captures a richer context within its effective receptive field, thereby excelling in extracting abstract information from images. Consequently, this methodology not only achieves desirable segmentation results at a 90-degree viewpoint but also experiences reduced misdetection rates at 30 degrees. Furthermore, when models are pretrainied on a diverse array of crops to bolster generalization capabilities, our method consistently yields optimal outcomes across various crop types, underscoring its robustness and versatility in agricultural drone-based canopy segmentation tasks.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Segmentation Mask Segmentation Heapmap

Figure 8. Taoism-Net and YOLOv8s-Seg perform segmentation mask and heatmaps on images captured from different perspectives. The images captured by drones with viewing angles of 30°, 60°, and 90° are selected as example images. This article uses perspective images to make the three example images more representative. Among them, (**a**–**c**) show the comparison of segmentation

performance between the two models. From the figure, it can be seen that YOLOv8s Seg has problems of oversegmentation, undersegmentation, and repeated segmentation for different perspectives, flight heights, and complex road conditions. However, in the case of grass and trees alternating, it is easy to misjudge the grass as the fruit tree canopy. Taoism-Net can maintain stable and accurate segmentation results in the face of various angles, constantly changing flight altitudes, and complex road conditions. (**a**) Original image; (**b**) YOLOv8s-Seg; (**c**) Taoism-Net (Ours); (**d**) original image; (**e**) YOLOv8s-Seg; (**f**) Taoism-Net (Ours).

4. Discussion

The dataset for this paper was collected at the Guangzhou Lychee Cultural Museum, which spans several acres. During the acquisition process, the drone captured image data with a resolution of 5280 × 3956 through its RGB sensor channels, operating at a relative altitude between 24.9 m and 50.0 m. In the creation of the dataset, this paper fully considered the representativeness of the region, selecting a sufficiently large area that includes images of lychee canopies with different features from various stages. The paper also took into account the potential impact of different angles on the model, employing shooting methods from various angles. To prevent the influence of data from specific angles during the training process from being too significant, a method of random partitioning was used to construct the dataset. Considering that the model might be trained with datasets from other crops, when pre-training models based on a variety of crops, the model can maintain excellent generalization capabilities across various crops. The experimental results indicate that, without the need for complex convolution and residual calculations, deep and complex networks already possess sufficient nonlinearity, although adding residual methods can significantly simplify the training process and improve performance. When extracting deep features in the backbone network, using small convolution kernels such as 1 × 1 or 2 × 2 can also obtain good feature extraction while reducing computational overhead. After the Vision Transformer (ViT) [32] was proposed for computer vision in 2021, researchers became fascinated with it. Although such models greatly increase the model inference latency, the improvement in accuracy still leads to a rush to incorporate Transformer blocks and self-attention mechanisms into models, neglecting the original intention of M.Goccia [33] et al. in 1995 to compress computational load with convolutional networks. While acknowledging the significant contributions of large-parameter models to the field of computer vision, this paper focuses on using extremely simplified network designs, introducing a feature extraction strategy with multiple layers of small convolution kernels and employing concurrent activation functions to accelerate training and inference speeds. It is hoped that this will provide new insights for computer vision work in the practical agricultural field.

5. Conclusions

With the continuous loss of agricultural labor, the demand for unmanned agricultural operations is increasing. With the development of deep learning and neural network technology, it is possible to use instance segmentation models to generate agricultural prescription maps, and use UAVs to apply pesticides to fruit trees and other precision agricultural tasks.

In modern agricultural production, precise segmentation algorithms on drones are of great significance for improving production efficiency and quality. With the development of deep learning and neural network technology, a large number of segmentation models have achieved ideal results in the public dataset, but the computational cost in edge computing is still a challenging problem. This article proposes a neural network called Taoism-Net based on minimalist design, which designs a backbone with small convolutional kernels to achieve efficient feature extraction. Stacked deployment to hierarchical activation functions reduces computational overhead while increasing the network's non-linear expression ability. DiceLoss was used in training to alleviate the problem of imbalanced samples. By comparing it with different models, Taoism-Net shows its advantages in precision

and reasoning speed. This model can be deployed in batch on edge computing devices, supports TensorRT, Openvino and other model conversion schemes, and can generate real-time agricultural prescription maps during UAV operations to guide precision agricultural tasks, helping improve the efficiency and accuracy of UAV operations.

Taoism-Net adheres to the concept of "simplicity is the most important" from Taoism, breaking away from the mainstream of complex deep neural networks and adopting a minimalist network design, combining UAVs with visual neural networks. The work of this article builds a practical bridge between academic research and industrial deployment, hoping to provide new insights for computer vision work in the actual agricultural field and promote more research on neural network architecture design.

Author Contributions: Conceptualization, Y.M. and J.Z.; Methodology, Y.M.; Experiment conception and design, Y.M., J.Z. and Z.L. (Zefeng Luo); Experiment execution, Y.M., Z.L. (Zefeng Luo) and C.Y. (Chaoran Yu); Writing—original draft preparation, Y.M. and Z.L. (Zefeng Luo); Writing—review and editing, Y.M., J.Z., Z.L. (Zuanhui Lin) and Z.L. (Zhongliang Liao); Funding acquisition, J.L., C.Y. (Caili Yu), C.Y. (Chaoran Yu) and Z.L. (Zuanhui Lin). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by: the Key-Area Research and Development Program of Guangdong Province (Grant No.2023B0202090001); Key Research and Development Program of Guangzhou (2023B03J1392); The National Natural Science Foundation of China (42061046); The special projects in key fields of ordinary universities in Guangdong Province (2021ZDZX4111).

Data Availability Statement: The data used in this study are available from thecorresponding author upon reasonable request.

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

Appendix A

Our model weights are trained in Pytorch, and available Taoism-Net export formats are in the Table A1 below. You can predict or validate directly on exported models, and usage examples are shown for your model after export completes.

Table A1. The weights trained in this article can support model transformation based on the following parameters.

Format	Model	Metadata	Argument
Pytorch	Taoism.pt	✓	-
TorchScript	Taoism.torchScript	✓	imgsz, optimize
ONNX	Taoism.onnx	✓	imgsz, half, dynamic,simplify, opset
openVINO	Taoism.openvino_model	✓	imgsz, half, int8
TensorRT	Taoism.engine	✓	imgsz, half, dynamic,simplify, workspace
TF Lite	Taoism.tflite	✓	imgsz, half, int8

References

1. Dai, M.; Hu, J.; Zhuang, J.; Zheng, E. A transformer-based feature segmentation and region alignment method for uav-view geo-localization. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4376–4389. [CrossRef]
2. Onishi, M.; Ise, T. Explainable identification and mapping of trees using UAV RGB image and deep learning. *Sci. Rep.* **2021**, *11*, 903. [CrossRef] [PubMed]
3. 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.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
6. Arbeláez, P.; Pont-Tuset, J.; Barron, J.T.; Marques, F.; Malik, J. Multiscale combinatorial grouping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 328–335.

7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
8. Pinheiro, P.O.; Lin, T.Y.; Collobert, R.; Dollár, P. Learning to refine object segments. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 75–91.
9. Liu, G.; Lin, Z.; Yu, Y. Robust subspace segmentation by low-rank representation. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 663–670.
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
12. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
13. Zhu, H.; Li, X.; Zhang, P.; Li, G.; He, J.; Li, H.; Gai, K. Learning tree-based deep model for recommender systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1079–1088.
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. Bao, H.; Dong, L.; Piao, S.; Wei, F. Beit: Bert pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
16. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. [CrossRef]
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
20. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16133–16142.
21. Xu, Y.; Xie, L.; Xie, C.; Dai, W.; Mei, J.; Qiao, S.; Shen, W.; Xiong, H.; Yuille, A. Bnet: Batch normalization with enhanced linear transformation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9225–9232. [CrossRef] [PubMed]
22. Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 322–330.
23. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
24. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
25. Li, X.; Fang, Z.; Zhao, R.; Mo, H. MRI Image Segmentation of Brain Tumor Based on Res-UNet. In Proceedings of the 2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI), Orlando, FL, USA, 7–9 November 2023.
26. Peng, J.; Liu, Y.; Tang, S.; Hao, Y.; Chu, L.; Chen, G.; Wu, Z.; Chen, Z.; Yu, Z.; Du, Y.; et al. Pp-Liteseg: A Superior Real-Time Semantic Segmentation Model. *arXiv* **2022**, arXiv:2204.02681.
27. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
28. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
29. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5386–5395.
30. Contributors, M. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 20 May 2024).
31. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018.

32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
33. Goccia, M.; Bruzzo, M.; Scagliola, C.; Dellepiane, S. Recognition of Container Code Characters through Gray-Level Feature Extraction and Gradient-Based Classifier Optimization. In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh, Scotland, 3–6 August 2003; Volume 1, pp. 973–977.