*mathematics*

MDPI

*Article*

# A Lightweight YOLOv5-Based Model with Feature Fusion and Dilation Convolution for Image Segmentation

**Linwei Chen [1] and Jingjing Yang [2,*]**

[1] College of Information Engineering, China Jiliang University, Hangzhou 310018, China; linwc10@163.com
[2] School of Information Science and Engineering, Hebei North University, Zhangjiakou 075000, China
* Correspondence: landryyang@hebeinu.edu.cn

**Abstract:** Image segmentation has played an essential role in computer vision. The target detection model represented by YOLOv5 is widely used in image segmentation. However, YOLOv5 has performance bottlenecks such as object scale variation, object occlusion, computational volume, and speed when processing complex images. To solve these problems, an enhanced algorithm based on YOLOv5 is proposed. MobileViT is used as the backbone network of the YOLOv5 algorithm, and feature fusion and dilated convolution are added to the model. This method is validated on the COCO and PASCAL-VOC datasets. Experimental results show that it significantly reduces the processing time and achieves high segmentation quality with an accuracy of 95.32% on COCO and 96.02% on PASCAL-VOC. The improved model is 116 M, 52 M, and 76 M, smaller than U-Net, SegNet, and Mask R-CNN, respectively. This paper provides a new idea and method with which to solve the problems in the field of image segmentation, and the method has strong practicality and generalization value.

**Keywords:** YOLOv5; deep learning; feature fusion; dilated convolution; MobileViT; image segmentation

**MSC:** 68T05; 68T07; 94A08



**Citation:** Chen, L.; Yang, J. A Lightweight YOLOv5-Based Model with Feature Fusion and Dilation Convolution for Image Segmentation. *Mathematics* **2023**, *11*, 3538. https://doi.org/10.3390/math11163538

Academic Editors: Lunke Fei, Yongbing Zhang and Jie Wen

Received: 15 July 2023
Revised: 12 August 2023
Accepted: 15 August 2023
Published: 16 August 2023



## 1. Introduction

Image segmentation is one of the important research directions in the field of computer vision, which aims to classify each pixel in an image into different objects or regions for the deep understanding of the image [1]. Recently, image segmentation techniques have undergone a transformation from traditional methods based on rules and thresholds to those based on machine learning and deep learning [2].

Early image segmentation methods mainly relied on manually designed features and manually judged thresholds for segmentation. These methods were simple and easy to understand but lacked the flexibility to adapt to complex scenes and changing data. With the development of machine learning and deep learning, data-driven image segmentation methods based on data have gradually become mainstream [3–8]. Among them, the convolutional neural network is a widely used deep learning model with strong feature extraction and nonlinear modeling capabilities, which can automatically learn complex feature representations [2,9,10]. FCN is the first convolutional network segmentation method proposed [11], which obtains feature maps of different scales by performing convolution operations on the input image and restores the feature maps to the original scale through upsampling. Subsequently, a series of improved models were proposed, such as U-Net, SegNet, and DeepLab [12–14], which further improved segmentation performance by introducing methods such as skip connections, atrous convolution, and spatial pyramid pooling [14,15]. However, these algorithms still have certain limitations. First, they usually require large computational resources and high time costs, so it is difficult to achieve real-time segmentation; second, the segmentation accuracy is affected when dealing with problems with complex textures, pose changes, etc. Therefore, improving

segmentation efficiency and accuracy is a current research hotspot and challenge in image segmentation [9,16,17].

In recent years, target detection techniques have also been gradually applied to image segmentation tasks. YOLOv5, as the most representative object detection model, has good performance in terms of speed and accuracy and can be applied to different scenes [18]. Meanwhile, techniques such as feature fusion, dilated convolution, and MobileViT are also widely used in image segmentation, which all help to improve the feature extraction capability and segmentation performance of the model.

This paper proposes a novel image segmentation of a universal target algorithm based on YOLOv5. This proposed method combines feature fusion, dilation convolution, and MobileViT. Among them, feature fusion is a method by which to fuse feature maps of different scales, which can improve the detection ability of the model for small and distant targets [19,20]; dilation convolution is a method by which to capture a more extensive range of information by increasing the field of perception size in the spatial domain, which can effectively solve the problem of tiny structures existing inside the segmented objects [21,22]; MobileViT is a lightweight model that can achieve high-performance image segmentation through model compression and acceleration [23–25].

In summary, current research in image segmentation focuses on using deep learning techniques to improve segmentation efficiency and accuracy. This study's main contributions can be summarized as follows: first, an image segmentation framework is proposed; the segmentation accuracy and efficiency of the model are further improved by the introduction of feature fusion, dilated convolution, and MobileViT techniques; second, this improved network is validated on several datasets and has strong logical reasoning ability and independent learning ability, which can promote the development of artificial intelligence technology [26]; third, this approach has good segmentation ability and generalization performance, which helps to solve problems in the field of computer vision. In addition, the method can be considered to be applied to other related fields, such as target tracking and scene understanding.

## 2. Related Work

Ultralytics proposed YOLOv5 in May 2020 [27]. The YOLOv5 series include YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. YOLOv5s is the best network with the smallest depth and feature map width compared with the other three models. It consists of four main components: the input module, the backbone network, the neck network, and the prediction network [28].

Input module: The data are loaded at the input terminal, and the YOLOv5 network preprocesses the input image at this stage. First, the input image is resized to the specified size. Data enhancement, random cropping, and random scheduling are also used in this module. The above data enhancement methods enrich the dataset and enhance the robustness of the model [29,30].

Backbone network: The backbone network of YOLOv5 aims to extract generic features of the target and mainly consists of SPP, CBL, CSPDarknet53, and Focus [30]. The idea of SPP is to eliminate the requirement of fixed input image size and generate fixed-length images, while the method can effectively solve problems such as target deformation. The SPP layer is designed to allow the input image to be of any size, but the output to the fully linked layer is a vector of fixed dimensions. In addition, SPP only performs one convolution calculation on the original image; the feature map of the whole image can be obtained, and then the candidate frames are mapped and the characteristics of the same dimension are obtained and then classified, so it saves a lot of computation time. CBL consists of convolution, batch normalization, and Leaky ReLU activation function. Compared with YOLOv4, YOLOv5 uses two CSP structures: CSP1_X and CSP2_X, one for the backbone network and the other for the neck network [28,31]. The key role of Focus is to slice the image before it enters the backbone. The output space is quadrupled by Focus operation, and the original three channels become twelve channels in order to

obtain a double-downsampled feature map without information loss after the convolution operation. The Focus slicing operation is shown in Figure 1.

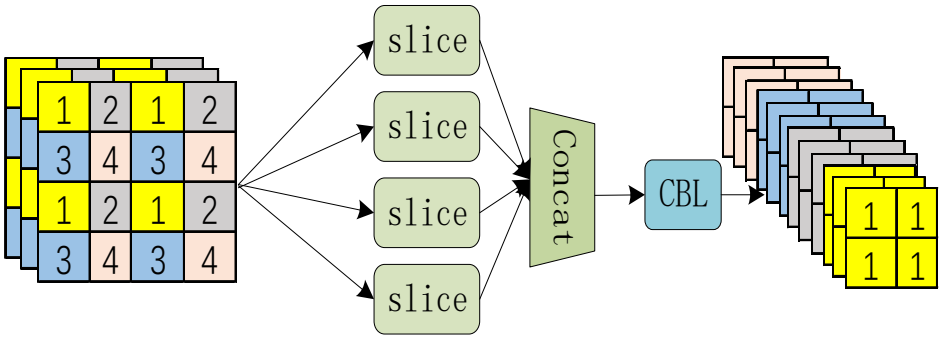


**Figure 1.** Focus slicing operation.

Neck network: The neck is located between the backbone and the prediction network and uses the structure connected by Feature Pyramid Networks, which aims to further enhance the diversity of features. In addition, the neck structure of YOLOv5 adopts CSP2 [25,27].

Prediction network: Prediction is the output side, completing the object detection results' output.

## 3. The Proposed Segmentation Approach

In order to segment the generic targets more effectively and accurately, the original YOLOv5 network structure is improved in this paper as follows.

In the original YOLOv5 model, feature fusion occurs after the convolutional layers, between the last convolutional layer and the prediction layer. Here, in this study, feature fusion is added to the backbone network of the YOLOv5 model. The reasons are as follows: (1) feature fusion can improve the perceptual capability of the model, enabling it to better process information of different scales and semantics in images; this helps to improve the detection accuracy and robustness of the model; (2) feature fusion can reduce the dimensionality of the feature map, thus reducing the computational complexity and speeding up the training and inference of the model; (3) feature fusion can promote the exchange and sharing of information between different levels in the feature graph, enhancing the expressiveness and generalization ability of the model.

Dilated convolution is added to the YOLOv5 model. This is because dilated convolution can increase the perceptual field, i.e., the response of each element in the output tensor to a larger region in the input tensor, which improves the perception of the model. In addition, the dilated convolution can also control the spacing between elements in the convolution kernel by adjusting the dilation factor, thus adapting to different scales of targets and scenes.

Using MobileViT in the YOLOv5 model can provide users a good experience and improve the model operation's efficiency, and lightweight networks have proven their necessity in practical applications [32,33].

The proposed approach for segmentation in this paper is shown in Figure 2.

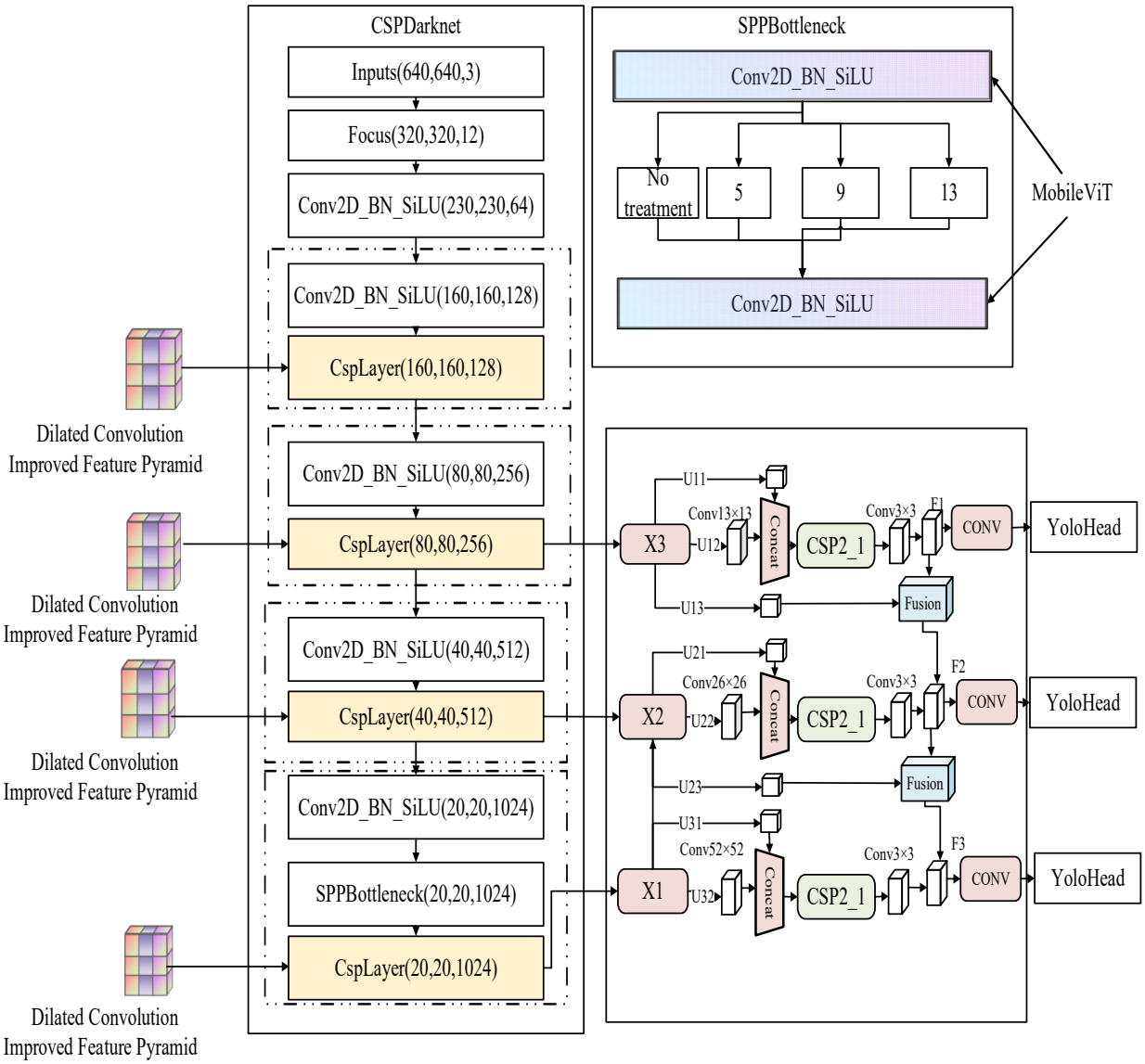


**Figure 2.** Proposed segmentation approach.

### *3.1. Improved YOLOv5s Model Based on Feature Fusion*

This paper proposes the multi-level YOLOv5 (M-YOLOv5s) algorithm based on multi-level feature fusion. The feature pyramid structure based on multi-level fusion is a multi-level feature fusion structure added in the middle of the feature pyramid structure so that the network can better use the location information of shallow features and fully integrate the semantic information with the location information. The structure includes a top-down convolutional neural network and a multi-level feature fusion network structure, where X1~X3 denote the lower, lower middle, and middle effective feature layers of the network, respectively; F1~F3 denote the features after multi-level fusion; and U1~U3 are the forward propagation processes of the network.

The feature pyramid network based on multi-level feature fusion performs the convolution operation with the same size of the feature layers extracted from the bottom up to obtain shallow features. It compresses the number of channels in depth to obtain richer location information to enhance the network's ability to locate the target. Through feature overlay, the deep semantic information interacts with the shallow localization information, and the number of channels is adjusted by the feature integration and convolution operation, which enhances the location localization ability and improves the robustness of the network. The quality of multi-scale feature fusion is enhanced by passing back the

multi-level fused features and fusing them with the effective feature layer again. Finally, the feature image input to the detection side of the network is made more detailed by refining the multi-level fused features. The feature pyramid network structure based on multi-layer fusion is more comprehensive for feature information acquisition and fully uses the location information of shallow features. Compared with the feature pyramid network structure, the feature pyramid network with multi-level feature fusion has higher feature information density with the same feature layer depth, and the multi-scale fusion of the feature pyramid structure makes the location information of the shallow features and the semantic information of the deep features more fully integrated. The structure is shown in Figure 3.

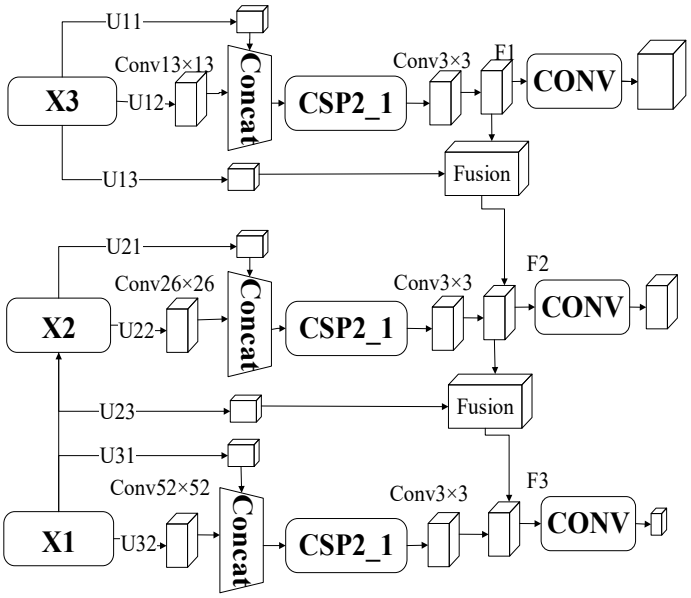


**Figure 3.** Feature pyramid network structure based on multi-level feature fusion.

For X1, convolution size of 52 × 52 is selected to obtain the shallow features, and the feature integration is performed with the convolution size of 3 × 3 and the channels number of 128.

For X2, the X2 features of U21 are superimposed with the shallow features obtained via the convolution of U22 with a convolution size of 26 × 26. Feature integration and channel compression with a convolution size of 3 × 3 and a channel number of 256 are performed to obtain F2, and the F2 features are feature fused with the U23 features. The F2 detection branch performs the same work as F1.

For X3, first, U11 retains the original lower valid feature layer information, and U12 extracts features with a convolution size of 1313 and compresses the number of channels to obtain shallow features. The features obtained from U11 and U12 are superimposed. Then they integrate the features with a convolution size of 33 and a channel number of 1024 to obtain F1. The output of F1 is divided into detection and fusion branches. After performing feature integration, the features have richer location information and retain the original deep semantic information. Finally, the detection branch of F1 features performs feature refinement of F1 features with 1 × 1 input to the detection part of the network, and the F1 fusion branch performs feature fusion with U13 features to prepare for multi-size feature fusion.

### *3.2. Addition of Dilated Convolution Module*

In this section, we improve the structure for the feature fusion part based on the feature fusion improved YOLOv5s network. M-YOLOv5s adds the multi-level feature fusion structure proposed in this section to the feature pyramid network and introduces

the extended convolution module, which constitutes this section's model, namely, Dilated convolution-M-YOLOv5s (DM-YOLOv5s).

Dilated convolution adds a dilation rate parameter compared to the original normal convolution operation. Unlike traditional convolution, the dilated convolution introduces a dilation factor in the convolution kernel, which makes a fixed number of holes between the elements in the convolution kernel. Figure 4a,b shows the normal and dilated convolution with an expansion rate of 4. The size of the convolution kernel after the expansion rate enlargement is calculated by the following formula:

$$Kernelsize = N + (S - 1) \times (N - 1), \tag{1}$$

where $S$ is the expansion rate and $N$ is the original convolution kernel size. Figure 4a shows the convolution kernel for normal convolution with size 33; for Figure 4b, the convolution kernel becomes 99, but the computational parameters are not increased, and the perceptual field becomes larger as the convolution kernel size becomes larger. The addition of the extended convolution module alleviates the problem of imbalance in the interaction between the deep feature layer and the shallow feature information. Taking the lower effective feature layer as an example, the dilated convolution module is added to the original structure compared to the multi-level feature fusion structure.

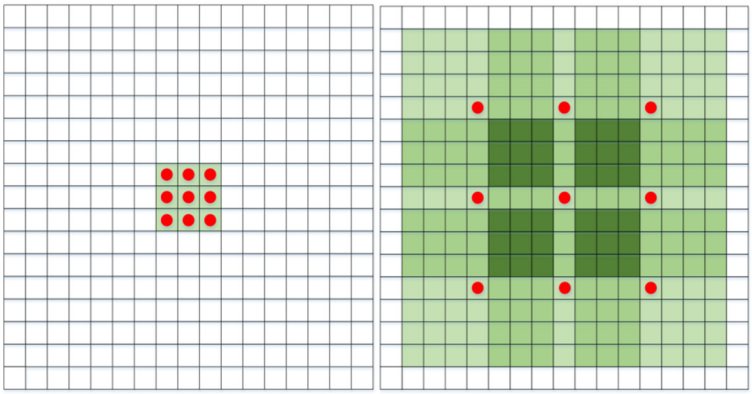


**Figure 4.** Schematic diagram of dilated convolution.

Using dilated convolution instead of regular convolution for feature extraction, the setting of the extension rate for the extended convolution needs to be calculated in a certain way to prevent the input and output feature map sizes from being unequal. The specific formula for the input and output feature map sizes of the dilated convolution is as follows:

$$Z_2 = \frac{Z_1 + 2P - d(N-1) - 1}{S} + 1, \tag{2}$$

where P is the zero-filling row, d is the expansion rate, N is the original convolutional kernel size, and S is the step size. $Z_1$ and $Z_2$ are the input and output feature map sizes, respectively.

In this paper, the expansion rate of the extended convolution is set to 2, S is set to 1, and N is set to 3. To ensure that the input and output sizes are the same, their filling zero rows are 2. The calculation formula for the increased sensory field by the extended convolution set in the target detection network is as follows:

$$W_2 = W_1 + (K - 1) \times S, \tag{3}$$

where $W_1$ is the perceptual field of the previous layer, $K$ is the convolutional kernel size after adding the expansion rate, and S is the step size. The initial network perceptual field is 1, and the perceptual field ratios of the multi-level feature fusion feature pyramid network

using normal convolution and extended convolution are 7:11, 9:13, and 11:15, respectively. Adding the extended convolution module to the multi-level feature fusion feature pyramid network expands the overall network perceptual field, and the increase in the perceptual field makes the network more sensitive to image information.

Dilation controls the spacing between samples within the convolution kernel by inserting a certain number of zeros within the kernel. This allows for the expansion of the effective receptive field of the convolution kernel while maintaining the original receptive field. This operation enables dilated convolution to capture a broader range of contextual information, thereby enhancing its ability to comprehend object structures and semantics within the receptive field. In the context of image segmentation tasks, feature pyramid networks are designed to obtain multi-scale feature representations to address objects of varying sizes. Dilated convolution modules can be incorporated within the feature pyramid at different levels, offering expanded contextual information for each level. By employing dilated convolution modules, feature pyramid networks can effectively capture object boundaries, textures, and semantic information across different scales. Consequently, this enhances the accuracy and robustness of image segmentation tasks by enabling the network to better perceive intricate details and contextual variations.

### *3.3. DM-YOLOv5s Model Based on MobileViT Lightweight*

MobileViT is a variant of the Vision Transformer (ViT) architecture tailored for mobile device scenarios. ViT, an attention-based image classification network, has achieved remarkable performance in image classification tasks. MobileViT achieves efficient execution on resource-constrained mobile devices by lightweighting and optimizing the ViT structure. Compared with the ordinary ViT, combining convolution and Transformer can obtain robust and high-performance ViT. To guarantee accuracy, it also has better real-time performance. A lightweight network model-based MobileViT is proposed. In this paper, we use MobileViT, a lightweight-based network, as the backbone network of the model. The convolutional neural network is good at extracting local feature information, and Vision Transformer based on a self-attentive mechanism is good at extracting global feature information. MobileViT network treats Vision Transformer as convolution, combining the advantages of convolutional neural network and Vision Transformer to build a lightweight and general network model. The network structure of MobileViT is shown in Table 1.

**Table 1.** Network structure of MobileViT.

| Input | Operator | #Out | L | s |
|---|---|---|---|---|
| 2562 × 3 | conv2d | 16 | - | 2 |
| 1282 × 16 | MV2 | 32 | - | 1 |
| 1282 × 32 | MV2 | 64 | - | 2 |
| 642 × 64 | MV2 | 64 | - | 1 |
| 642 × 64 | MV2 | 64 | - | 1 |
| 642 × 64 | MV2 | 96 | - | /2 |
| 322 × 96 | MVIT | 96 | 2 | 1 |
| 322 × 96 | MV2 | 128 | - | 2 |
| 162 × 128 | MVIT | 128 | 4 | 1 |
| 162 × 128 | MV2 | 160 | - | 2 |
| 82 × 160 | MVIT | 160 | 3 | 1 |
| 82 × 160 | Conv2d | 640 | - | 1 |
| 82 × 640 | Avgpool 8 × 8 | - | - | - |
| 12 × 640 | FC | - | - | - |
| 12 × k | Conv2d\ | K< | - | - |

Input indicates the input size of each module in the network; Operator indicates the module experienced by each feature layer; #out indicates the number of channels output after each feature layer; L indicates the number of Transformer modules in the MVIT

module; s indicates the step length of each operation; MVIT indicates MobileViT module; and MV2 indicates MobileNetV2 module.

The shallow features contain more location information, and the deep features contain more semantic information. MobileViT network is an image classification network, and the task in this paper is a target detection task. The features of the tenth layer of the MobileViT network have been downsampled by 32 times, and a large amount of location information will be lost if the features continue to be extracted and the subsequent network involves the classification task. Therefore, the tenth layer of the MobileViT network is discarded, and the remaining network is used as the backbone feature extraction network of the model [32].

The flow of the proposed MobileViT-DM-YOLOv5s (MDM-YOLOv5s) algorithm in this paper is shown in Algorithm 1.

**Algorithm 1:** The Flow of MobileViT-DM-YOLOv5s Algorithm

```
Begin
  // Define inputs and outputs
Input: image
Output: segmentation_result
  // Defining M-YOLOv5s Algorithm for Multi-Level Feature Fusion
  // Define feature pyramid structure
  X1, X2, X3 = feature_pyramid(image)
  // Define multi-layer feature fusion network structure
  F1 = X1
  F2 = fuse_features(X2, F1)
  F3 = fuse_features(X3, F2)
  // Defining top-down convolutional neural networks
  U1 = upsample(F3)
  U2 = fuse_features(U1, F2)
  U3 = fuse_features(U2, F1)
  // Define DM-YOLOv5s model
  M-YOLOv5s = YOLOv5s(feature=F3)
  DM-YOLOv5s = DilatedConvolution(M-YOLOv5s)
  // Defining the MobileViT Network
  MobileViT = MobileViT(feature=U3)
  // Performing image segmentation tasks
  segmentation_result = MobileViT(DM-YOLOv5s(image))
End
```

## 4. Experiment Section

### *4.1. Experimental Setup*

#### 4.1.1. Experimental Platform

The experiments' platforms are AMDR7-5800H3 and NVIDIA GeForce RTX2060 (6 GB). The experimental development environment is Python, and the PyTorch framework builds the detection model. The experiments use the stochastic gradient descent method to optimize the model. Epochs are set to 200 to ensure that the model can fully learn the features in the dataset. The input image size for YOLOv5 model training: the image size should be specified for YOLOv5 model training and set to 640 $\times$ 640. Batch size refers to the number of samples used in each training and is set to 32. The learning rate is set to 0.01, the regularization factor is set to 0.0005, and the number of iterations is set to 500.

#### 4.1.2. Dataset

Choosing a suitable dataset is crucial for image segmentation experiments. First, a good dataset should be representative and diverse, able to cover a variety of situations such as different scenes, different lighting conditions, and different object morphologies, thus ensuring the generalization performance of the algorithm. Second, the quality of the annotation of the dataset is also very important; the annotation should be accurate,

detailed, and consistent, and different types of annotations may be required for different tasks. Finally, the dataset size should also be large enough to train deep learning models and perform adequate validation and comparison. Therefore, choosing the right dataset can improve the reliability and generalization of experimental results.

In this paper, we choose COCO and PASCAL-VOC datasets to conduct ablation experiments and comparison experiments on the models. Choosing two datasets for experiments can help evaluate image segmentation algorithms' performance and generalization ability on different datasets. Comparing the two datasets can reveal their differences.

COCO: The COCO dataset contains over 80 common object classes, such as people, animals, vehicles, and furniture. The main features of the COCO dataset are diversity and complexity, containing a large number of images and multiple object instances covering a wide range of scales, poses, and occlusions.

PASCAL-VOC: The PASCAL-VOC dataset mainly contains 20 common object classes, such as people, cars, airplanes, and animals. Each image is accurately annotated with object bounding boxes; for some images, there is also a pixel-level semantic segmentation annotation of the objects.

#### 4.1.3. Mean Average Precision

In this paper, Average Precision (*AP*) and mean Average Precision (*mAP*) are used for evaluation and comparison [34]. The mean average precision (*mAP*) is used to measure the accuracy of target detection and represents the performance of the algorithm in terms of detection accuracy over the entire dataset, as follows:

$$mAP = \frac{AP}{m} \tag{4}$$

where $m$ is the total number of object categories in the dataset.

### *4.2. Ablation Experiments*

#### 4.2.1. Ablation Experiment of Feature Fusion

For the feature extraction part of YOLOv5s, the feature pyramid structure with multi-level feature fusion is established to enhance the interaction between the shallow localization information and the deep semantic information, making the multi-scale feature fusion more adequate. To investigate the effect of the feature pyramid structure of multi-level feature fusion on multi-scale feature fusion, the effectiveness of the improved network structure is illustrated in terms of function loss and mean-average accuracy [35].

(1) Loss comparison

As shown in Figure 5, YOLOv5s represents the loss of the YOLOv5s algorithm using the original feature pyramid, and M-YOLOv5s represents the loss of the YOLOv5s algorithm using the multi-level feature fusion feature pyramid structure.

In the same rounds, the loss of the M-YOLOv5s algorithm using the multi-level feature fusion feature pyramid structure is 0.19% lower than that of the YOLOv5s algorithm, and the overall curve shows that the loss of the M-YOLOv5s algorithm is not more obviously perturbed than that of the YOLOv5s algorithm at around 50 rounds and the loss of the network decreases more rapidly in the early training period, and the overall trend is flatter. The lower loss function curve during training means that the algorithm can predict the segmentation results of the images more accurately, thus demonstrating the generalization performance and effectiveness of the M-YOLOv5s algorithm on the test set.

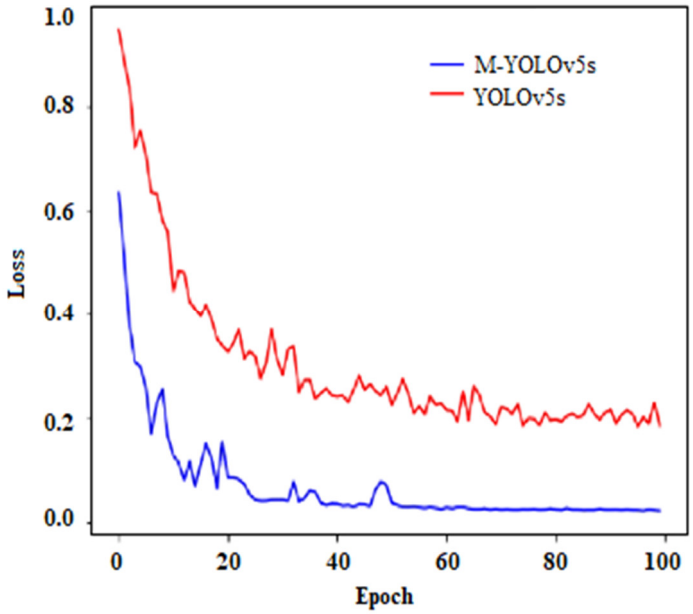


**Figure 5.** Comparison of feature pyramid network loss for multi-level feature fusion.

(2) Accuracy comparison

As shown in Table 2, the M-YOLOv5s algorithm fully combines shallow and deep features to improve the quality of multi-scale feature fusion. In terms of detection accuracy, the M-YOLOv5s algorithm has improved by 1.58%, which illustrates that feature fusion between multi-level features promotes the information flow between shallow and deep feature layers, and also illustrates the effectiveness of multi-level feature fusion feature pyramid structure for improving network performance. M-YOLOv5s algorithm can accurately divide and segment different objects or regions in the image.

**Table 2.** Comparison of image segmentation accuracy between M-YOLOv5s and YOLOv5s.

| Backbone Networks | Feature Pyramid Network | mPA (%) |
|---|---|---|
| YOLOv5s | FPN | 92.12 |
| M-YOLOv5s | M- FPN | 93.70 |

4.2.2. Ablation Experiment of Dilated Convolutional

After using the feature pyramid structure with multi-level feature fusion, the network accuracy is improved to some extent, but the feature pyramid structure with multi-level feature fusion improves the accuracy at the expense of feature map size.

In general, the loss of feature map size can lead to the imbalance between the semantic information of deep features and the location information of shallow features, and the interaction of their feature information is limited, resulting in the missing detection of some objects. Therefore, the dilated convolution module is added to the shallow feature layer of the multi-level feature fusion feature pyramid structure to prevent the loss of image detail information due to the deepening of the network so that the deep feature layer can expand the perceptual field without losing the image size and verify the effect on loss and accuracy under the balanced feature information interaction by adding the dilated convolution module.

(1) Loss comparison

As shown in Figure 6, M-YOLOv5s in the figure indicates the loss of the algorithm without adding the dilated convolution module, and DM-YOLOv5s indicates the loss of the algorithm with the dilated convolution module added in the shallow features.

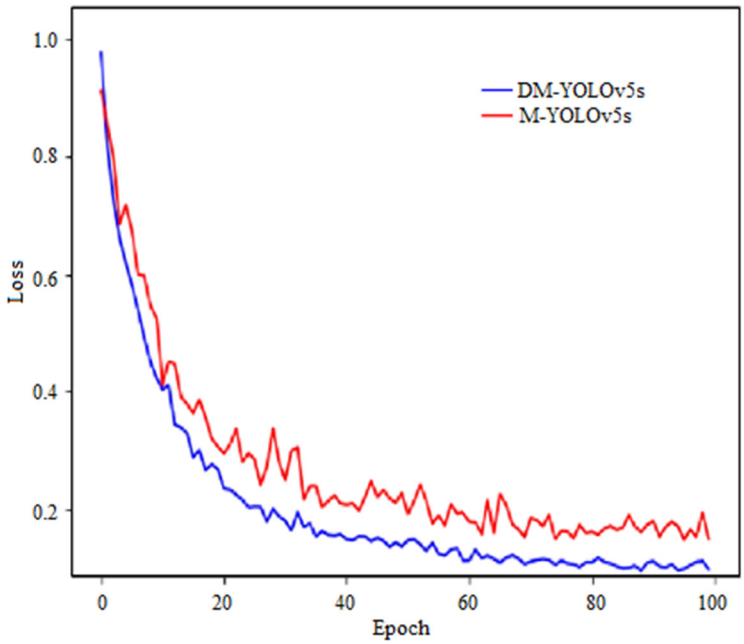


**Figure 6.** Comparison of dilated convolution loss.

The DM-YOLOv5s algorithm loss is reduced by 0.69% under the same rounds, and the two curves have matched the overall curve, but the perturbation of DM-YOLOv5s algorithm loss is flatter around round 50, and the flat loss curve means that the loss function decreases more slowly when training the model, but the performance of the model is more stable on both training and test data without overfitting or underfitting problems. Such a situation usually indicates that the model is close to the optimal state and can perform well on new data. The perturbation at 50 rounds is due to the freeze training method chosen in this paper, which opens the original weights after the first 50 rounds and trains them together, so there is a certain loss of perturbation, but the experimental results show that the loss of perturbation after adding the original weights can be mitigated by optimizing the structure to achieve a better training effect.

(2) Accuracy comparison

As shown in Table 3, the accuracy is improved by 2.21% when dilated convolution is used, and the improvement effect of the network illustrates that its network performance has reached the optimum for M-YOLOv5s. After adding dilated convolution, the information interaction between shallow and deep features has tended to be balanced and stable, the feature flow between scales is sufficient, and the overall network performance is close to saturation.

**Table 3.** Comparison of image segmentation accuracy between M-YOLOv5s and DM-YOLOv5s.

| Image Segmentation Model | Dilated Convolution Module | mPA (%) |
|---|---|---|
| M-YOLOv5s | - | 93.70 |
| DM-YOLOv5s | V | 95.91 |

4.2.3. Ablation Experiment of Lightweight Model

To evaluate the improved part of this paper more comprehensively, ablation experiments were conducted on the MDM-YOLOv5s network, and the results are shown in Table 4.

**Table 4.** Comparison of comprehensive performance of DM-YOLOv5s and MDM-YOLOv5s image segmentation.

| Image Segmentation Model | MobileviT | mAP (%) | Size/MB | Time/s |
|---|---|---|---|---|
| DM-YOLOv5s | - | 95.91 | 317 | 0.045 |
| MDM-YOLOv5s | V | 95.32 | 302 | 0.035 |

In this case, the comprehensive performance depends not only on the accuracy of the model but is also influenced by the model size. Both metrics need to be considered considering the trade-off between computational resources and storage capacity required in practical application scenarios. Their accuracies are relatively close to each other for the two models, MDM-YOLOv5s and DM-YOLOv5s. If these models need to be used in resource-constrained environments, model size may need to be considered as an important metric, in which case MDM-YOLOv5s will be more advantageous.

Beyond model size, processing speed is a pivotal performance metric. We conducted comparative analyses through statistical data to assess the time required by the DM-YOLOv5s model and the MDM-YOLOv5s model for processing a single image. The DM-YOLOv5s model typically requires 0.04 to 0.06 s per image. Nevertheless, the average processing time of the MDM-YOLOv5s model has been reduced to 0.035 s. This signifies that while sustaining high performance, our enhancement approach further amplifies image processing speed, rendering it more suitable for real-time application scenarios.

### *4.3. Comparison Experiment*

This section compares the MDM-YOLOv5s network with other image segmentation algorithms, and from Table 5, we can learn that the improved algorithms in this paper have higher accuracy than other algorithms.

**Table 5.** Comparison of accuracy of different image segmentation algorithms.

| Image Segmentation Model | COCO Precision (%) | PASCAL-VOC Precision (%) | mAP (%) | Model Size/MB |
|---|---|---|---|---|
| U-Net | 89.78 | 85.76 | 87.77 | 418 |
| SegNet | 90.82 | 89.76 | 90.29 | 354 |
| Mask R-CNN | 91.85 | 92.49 | 92.17 | 378 |
| MDM-YOLOv5s | 95.32 | 96.02 | 95.67 | 302 |

From the experimental results in the above table, it can be seen that the MDM-YOLOv5s algorithm is more accurate than U-Net, SegNet, and Mask R-CNN for image segmentation, respectively, and the accuracy curves of each algorithm model are shown in Figure 7.

The experimental results in the above table show that the MDM-YOLOv5s model is the smallest, reducing 116 M, 52 M, and 76 M compared to U-Net, SegNet, and Mask R-CNN, respectively. It means that the MDM-YOLOv5s model requires less storage space and computational resources to accomplish the same task. This is valuable for practical applications because it allows the model to run more efficiently on embedded devices or in constrained environments. MDM-YOLOv5s employs techniques such as multi-scale fusion, dilation convolution, and model lightweight to effectively compress the model structure and the number of parameters while ensuring accuracy. In contrast, U-Net, SegNet, and Mask R-CNN usually require more parameters and computational resources to achieve higher accuracy and robustness. Therefore, the small model of the MDM-YOLOv5s model is a great advantage and is especially suitable for scenarios with limitations on computational resources and storage space.

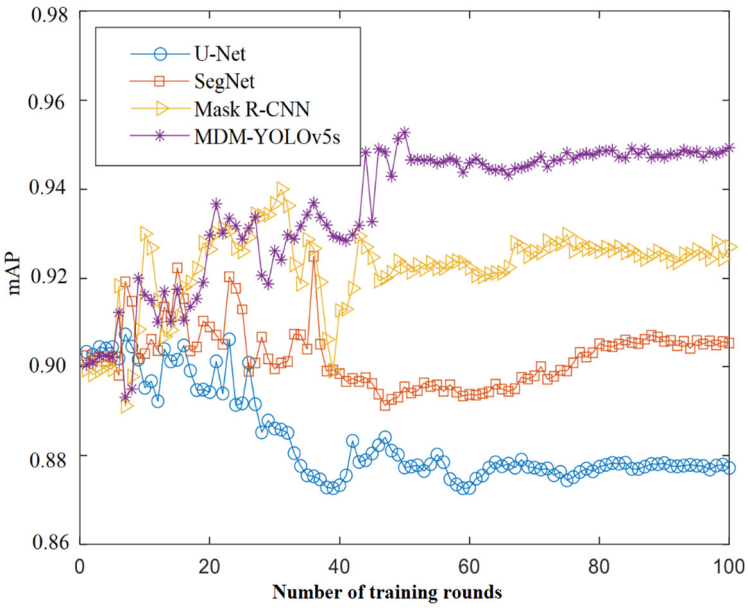


**Figure 7.** Accuracy comparison curves of different models.

The delay of the target segmentation model refers to the time required to go from the input image to the output target bounding box. Comparing the latencies of target segmentation models can help us evaluate the usability and applicability of different models in real-world scenarios, especially in real-time applications. By comparing the latencies between models, we can understand which parts of the model are more time-consuming and thus optimize for those parts to improve the efficiency and speed of the model. Comparing the latencies of models can also help us evaluate the required hardware device resources, including CPU, GPU, memory, etc., to ensure that the model can run well on the selected hardware. Therefore, this paper compares the latency of MDM-YOLOv5s, U-Net, SegNet, and Mask R-CNN as shown in Figure 8.

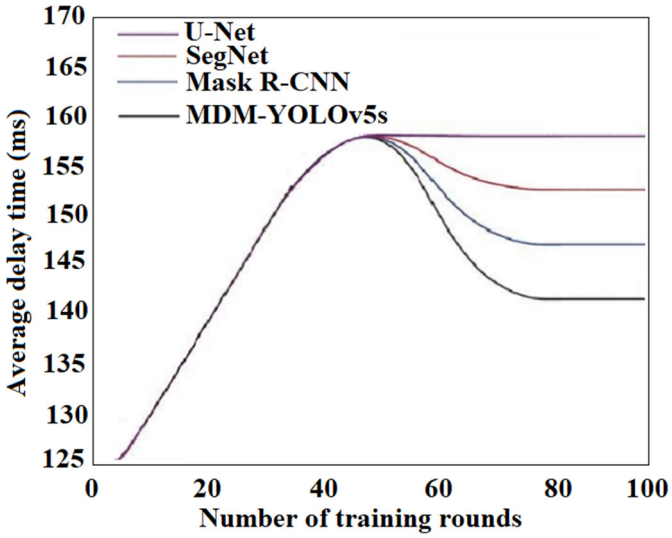


**Figure 8.** Comparison curves of latency of different models.

The MDM-YOLOv5s model is compared with U-Net, SegNet, and Mask R-CNN, and it is found that the latency of MDM-YOLOv5s is lower than the other three algorithms in image segmentation. This means that in practical applications, the MDM-YOLOv5s model has higher real-time performance and responsiveness and can detect and segment the target of the input image faster.

As a result, the MDM-YOLOv5s model can complete the inference task in a shorter time, speed up the image segmentation and improve the processing efficiency. The MDM-YOLOv5s model occupies less storage space, reducing the storage cost and making the operation and deployment more convenient. At the same time, the MDM-YOLOv5s model can be applied to some resource-constrained scenarios, such as mobile devices and embedded systems, extending the application scope of image segmentation technology. In addition, the MDM-YOLOv5s model has a simpler structure and fewer parameters, so it requires low conditional requirements, and the model training may be more stable and reliable. The model segmentation effect graph is shown in Figure 9.

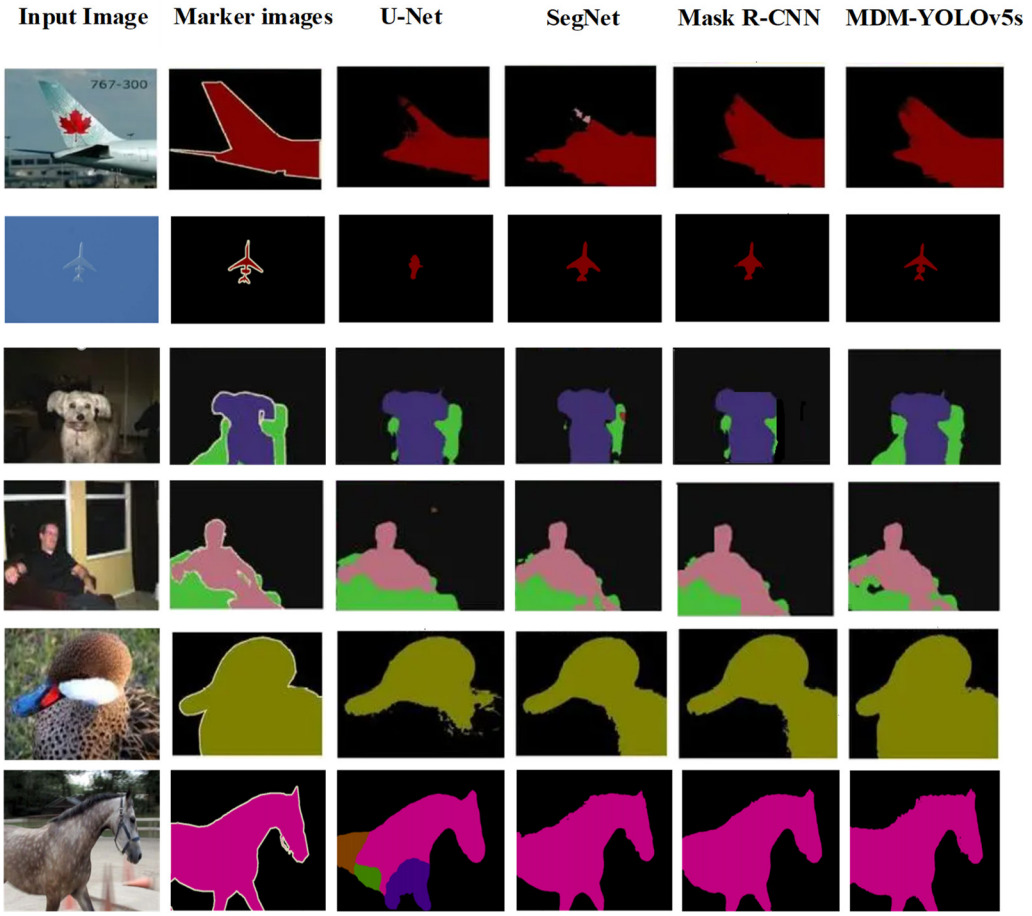


**Figure 9.** Graph of segmentation effects of different models.

Comparing the MDM-YOLOv5s model with U-Net, SegNet, and Mask R-CNN, it is found that the accuracy of MDM-YOLOv5s is higher than the other three algorithms in image segmentation. The high accuracy of the MDM-YOLOv5s model can be attributed to the following aspects:

1. Feature fusion technique: The MDM-YOLOv5s model uses the feature fusion technique to fuse features at different levels, thus improving the model's understanding of images and segmentation accuracy;
2. Dilated convolution technique: The MDM-YOLOv5s model adopts the dilated convolution technique, which can effectively expand the perceptual field and improve the model's ability to capture image details, thus improving the segmentation accuracy;
3. MobileViT technology: The MDM-YOLOv5s model also adopts MobileViT technology, which can effectively reduce the model parameters and computation volume, thus improving the model operation speed and efficiency;
4. YOLOv5s structure: The MDM-YOLOv5s model is improved based on the YOLOv5s structure, and YOLOv5s itself is an efficient target detection algorithm with a simple

structure, small computation, and fast speed, and these advantages also provide the basis for the high accuracy of the MDM-YOLOv5s model.

Compared with the excellent image segmentation model TransUNet introduced in recent years [36], the comparison mainly includes the mean average precision (mAP), model size, and processing time of the single image. Experimental results show that the MDM-YOLOv5 algorithm outperforms the TransUNet algorithm in accuracy by 1.34 percentage points. Compared to TransUNet, the model size of MDM-YOLOv5 has significantly decreased. The processing time for MDM-YOLOv5 is reduced by 0.135 s compared with TransUNet, indicating that MDM-YOLOv5 is more efficient in terms of processing speed.

In summary, the MDM-YOLOv5s model uses a variety of advanced techniques and combines the advantages of YOLOv5s, thus achieving high accuracy in image segmentation tasks. Compared with U-Net, SegNet, Mask R-CNN, and TransUNet, the MDM-YOLOv5s model is optimized in feature fusion, dilation convolution, MobileViT technique, and structure, and thus has higher accuracy in image segmentation.

## 5. Conclusions

In this research, we explored the combination of advanced techniques such as feature fusion, dilated convolution, and MobileViT with YOLOv5s and applied them to the image segmentation task. This approach can significantly shorten processing time through experimental validation while maintaining high segmentation quality and has strong practical and generalization value. The contributions of this research are as follows: first, a novel image segmentation framework is proposed to take advantage of YOLOv5s' target detection, making full use of its fast feature; second, the segmentation accuracy and efficiency of the model are further improved by the introduction of feature fusion, dilated convolution and MobileViT techniques. Overall, this research provides a new idea and method with which to solve the complex problems in image segmentation.

The future outlook is that the proposed algorithm can continue to be improved and optimized, and more technical means can be added to improve the segmentation efficiency and accuracy further. In addition, the method can be applied to other related fields, such as target tracking and scene understanding. In summary, this study provides useful references and inspiration for future research in the field of image segmentation.

**Author Contributions:** Conceptualization, L.C. and J.Y.; methodology, L.C.; software, L.C.; validation, L.C. and J.Y.; formal analysis, L.C.; investigation, L.C.; resources, L.C. and J.Y.; data curation, L.C. and J.Y.; writing—original draft preparation, L.C.; writing—review and editing, J.Y. and L.C.; supervision, J.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Natural Science Foundation of Hebei, grant number F2021405001.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: https://cocodataset.org/, (accessed on 27 March 2022).

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

## References

1. Kurban, T.; Civicioglu, P.; Kurban, R.; Besdok, E. Comparison of evolutionary and swarm based computational techniques for multi-level color image thresholding. *Appl. Soft Comput.* **2014**, *23*, 128–143. [CrossRef]
2. Liu, X.; Deng, Z.; Yang, Y. Recent progress in semantic image segmentation. *Artif. Intell. Rev.* **2019**, *52*, 1089–1106. [CrossRef]
3. Pal, N.R.; Pal, S.K. A review on image segmentation techniques. *Pattern Recognit.* **1993**, *26*, 1277–1294. [CrossRef]
4. Patra, S.; Gautam, R.; Singla, A. A novel context sensitive multi-level thresholding for image segmentation. *Appl. Soft Comput.* **2014**, *23*, 122–127. [CrossRef]
5. Dutta, P.K. Image segmentation based approach for the purpose of developing satellite image spatial information extraction for forestation and river bed analysis. *Int. J. Image Graph.* **2019**, *19*, 1950002. [CrossRef]
6. Wen, J.; Fang, X.Z.; Cui, J.R.; Fei, L.K.; Yan, K.; Chen, Y.; Xu, Y. Robust sparse linear discriminant analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 390–403. [CrossRef]

7. Bao, X.; Jia, H.; Lang, C. A novel hybrid harris hawks optimization for color image multi-level thresholding segmentation. *IEEE Access* **2019**, *7*, 76529–76546. [CrossRef]
8. Khan, A.; Irtaza, A.; Javed, A.; Nazir, T.; Malik, H.; Malik, K.; Khan, M. Defocus blur detection using novel local directional mean patterns (LDMP) and segmentation via KNN matting. *Front. Comput. Sci.* **2022**, *16*, 104–116. [CrossRef]
9. Nanda, N.; Kakkar, P.; Nagpal, S. Computer-aided segmentation of liver lesions in CT scans using cascaded convolutional neural networks and genetically optimised classifier. *Arab. J. Sci. Eng.* **2019**, *44*, 4049–4062. [CrossRef]
10. Thyreau, B.; Taki, Y. Learning a cortical parcellation of the brain robust to the MRI segmentation with convolutional neural networks. *Med. Image Anal.* **2020**, *14*, 101639. [CrossRef]
11. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651. [CrossRef]
12. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]
13. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
14. Han, L.; Chen, Y.H.; Li, J.M.; Zhong, B.; Sun, M. Liver segmentation with 2.5 D perpendicular UNets. *Comput. Electr. Eng.* **2021**, *91*, 107118. [CrossRef]
15. Huynh, C.; Tran, A.T.; Luu, K.; Hoai, M. Progressive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16750–16759. [CrossRef]
16. Fan, H.; Sun, Y.; Zhang, X.J.; Zhang, C.C.; Li, X.J.; Wang, Y. Magnetic-resonance image segmentation based on improved variable weight multi-resolution Markov random field in undecimated complex wavelet domain. *Chin. Phys. B* **2021**, *30*, 748–761. [CrossRef]
17. Kotte, S.; Kumar, P.R.; Injeti, S.K. An efficient approach for optimal multi-level thresholding selection for gray scale images based on improved differential search algorithm. *Ain Shams Eng. J.* **2018**, *9*, 1043–1067. [CrossRef]
18. Huang, J.Y.; Cui, H.; Ma, J.; Hao, Y. Research on an aerial object detection algorithm based on improved YOLOv5. In Proceedings of the 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA), Changchun, China, 20–22 May 2022; pp. 396–400. [CrossRef]
19. Zhou, Q.; Wang, R.; Hu, H.M.; Tan, Q.; Zhang, W.J. Referring image segmentation with attention guided cross modal fusion for semantic oriented languages. *Front. Comput. Sci.* **2022**, *16*, 175–177. [CrossRef]
20. Li, Z.L.; Zhang, Q.J.; Long, T.; Zhao, B.J. A parallel pipeline connected-component labeling method for on-orbit space target monitoring. *Syst. Eng. Electron.* **2022**, *33*, 1095–1107. [CrossRef]
21. Xia, H.; Sun, W.; Song, S.; Mou, X. Md-net: Multi-scale dilated convolution network for CT images segmentation. *Neural Process Lett.* **2020**, *51*, 2915–2927. [CrossRef]
22. Wu, Y.; Lin, L. Automatic lung segmentation in CT images using dilated convolution based weighted fully convolutional network. *J. Phys. Confer. Ser.* **2022**, *1646*, 012032. [CrossRef]
23. Dong, X.; Yan, S.; Duan, C. A lightweight vehicles detection network model based on YOLOv5. *Eng. Appl. Artif. Intell.* **2022**, *113*, 104914. [CrossRef]
24. Liu, H.; Sun, F.; Gu, J.; Deng, L.J. Sf-yolov5: A lightweight small object detection algorithm based on improved feature fusion mode. *Sensors* **2022**, *22*, 5817. [CrossRef]
25. Zhou, L.; Wei, S.Y.; Cui, Z.M.; Fang, J.Q.; Yang, X.T.; Ding, W. Lira-YOLO: A lightweight model for ship detection in radar images. *Syst. Eng. Electron. Technol.* **2020**, *31*, 950–956. [CrossRef]
26. Wen, J.; Liu, C.L.; Deng, S.J.; Liu, Y.C.; Fei, L.K.; Yan, K.; Xu, Y. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–13. [CrossRef]
27. Yang, G.H.; Feng, W.; Jin, J.T.; Lei, Q.J.; Li, X.H.; Gui, G.C.; Wang, W.J. Face mask recognition system with YOLOV5 based on image recognition. In Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020; pp. 1398–1404. [CrossRef]
28. Wang, Z.; Jin, L.; Wang, S.; Xu, H. Apple stem/calyx real-time recognition using YOLO-v5 algorithm for fruit automatic loading system. *Postharvest Biol. Technol.* **2022**, *185*, 111808. [CrossRef]
29. Lei, F.; Tang, F.F.; Li, S.H. Underwater target detection algorithm based on improved YOLOv5. *J. Mar. Sci. Eng.* **2022**, *10*, 310. [CrossRef]
30. Mathew, M.P.; Mahesh, T.Y. Leaf-based disease detection in bell pepper plant using YOLOv5. *Signal Image Video Process* **2022**, *16*, 841–847. [CrossRef]
31. Dewi, C.; Chen, R.C.; Jiang, X.; Yu, H. Deep convolutional neural network for enhancing traffic sign recognition developed on yolov4. *Multimed Tools Appl.* **2022**, *81*, 37821–37845. [CrossRef]
32. Zhou, L.; Gao, R.; Wang, J. A self-supervised, few-shot semantic segmentation study based on mobileViT model structure. In Proceedings of the 2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT), Jilin, China, 28–30 April 2023; pp. 917–921. [CrossRef]
33. Aiadi, O.; Khaldi, B. A fast lightweight network for the discrimination of COVID-19 and pulmonary diseases. *Biomed. Signal Process Control* **2022**, *78*, 103925. [CrossRef]

34. Csurka, G.; Larlus, D.; Perronnin, F.; Meylan, F. What is a good evaluation measure for semantic segmentation? In Proceedings of the British Machine Vision Conference, Meylan, France, 16–19 January 2013; pp. 1–11. [CrossRef]
35. Zhang, Y.; David, P.; Foroosh, H.; Gong, B. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 1823–1841. [CrossRef]
36. Chen, J.N.; Lu, Y.Y.; Yu, Q.H.; Luo, X.D.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y.Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.