mathematics

MDPI

Article

A Cross-Modal Feature Fusion Model Based on ConvNeXt for RGB-D Semantic Segmentation

Xiaojiang Tang [1], Baoxia Li [1], Junwei Guo [2], Wenzhuo Chen [1], Dan Zhang [2] and Feng Huang [2,*]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[2] College of Science, China Agricultural University, Beijing 100083, China
* Correspondence: huangfeng@cau.edu.cn

Abstract: Semantic segmentation, as the pixel level classification with dividing an image into multiple blocks based on the similarities and differences of categories (i.e., assigning each pixel in the image to a class label), is an important task in computer vision. Combining RGB and Depth information can improve the performance of semantic segmentation. However, there is still a problem of the way to deeply integrate RGB and Depth. In this paper, we propose a cross-modal feature fusion RGB-D semantic segmentation model based on ConvNeXt, which uses ConvNeXt as the skeleton network and embeds a cross-modal feature fusion module (CMFFM). The CMFFM designs feature channel-wise and spectral-wise fusion, which can realize the deeply feature fusion of RGB and Depth. The in-depth multi-modal feature fusion in multiple stages improves the performance of the model. Experiments are performed on the public dataset of SUN-RGBD, showing the best segmentation by our proposed model ConvNeXt-CMFFM with the highest mIoU score of 53.5% among the nine comparative models. The outstanding performance of ConvNeXt-CMFFM is also achieved on our self-built dataset of RICE-RGBD with the highest mIoU score and pixel accuracy among the three comparative datasets. The ablation experiment on our rice dataset shows that compared with ConvNeXt (without CMFFM), the mIoU score of ConvNext-CMFFM is increased from 71.5% to 74.8% and its pixel accuracy is increased from 86.2% to 88.3%, indicating the effectiveness of the added feature fusion module in improving segmentation performance. This study shows the feasibility of the practical application of the proposed model in agriculture.

Keywords: RGB-D semantic segmentation; feature fusion; multi-modality

MSC: 68T07

Citation: Tang, X.; Li, B.; Guo, J.; Chen, W.; Zhang, D.; Huang, F. A Cross-Modal Feature Fusion Model Based on ConvNeXt for RGB-D Semantic Segmentation. *Mathematics* **2023**, *11*, 1828. https://doi.org/10.3390/math11081828

Academic Editors: Xinchao Zhao, Xingquan Zuo, Yinan Guo and Kunpeng Kang

Received: 28 February 2023
Revised: 5 April 2023
Accepted: 11 April 2023
Published: 12 April 2023

1. Introduction

Semantic segmentation is an important task in computer vision and its purpose is to divide the input image into multiple regions with coherent semantic meaning to complete pixel-dense scene understanding for many real-world applications, such as autonomous driving [1], robot navigation [2] and so on. In recent years, with the rapid development of deep learning [3–7], pixel-based semantic segmentation of RGB images has received more and more attention and has achieved remarkable progress in segmentation accuracy [6,7]. However, due to the characteristics of RGB images, current deep semantic segmentation models cannot always extract correct features in some specific cases. For example, when two objects have similar colors or textures, it is hard to differentiate between them through pure RGB image. In order to solve these problems, some researchers use additional information to assist in semantic segmentation.

In recent years, with the rapid development of RGB-D sensors, in addition to RGB information, Depth information can also be acquired. Depth data can show the structure and geometric information of objects in the scene and can be used as supplementary data for the simultaneous RGB data so as to obtain richer features such as color, texture and

shape, and improve the accuracy of semantic segmentation. Many works have proved that spatial information is very helpful for improving the accuracy of semantic segmentation, and affirmed the effectiveness of learning from complementary patterns. With the rapid development of convolutional neural networks (CNN or ConvNet), researchers have proposed various CNN-based methods to use Depth information for RGB-D segmentation. In segmentation tasks, two mainstream designs have been widely used, namely single-stream design and two-stream design. In a single-stream design, Depth information is simply spliced directly with RGB at the input end to form a 4-channel (RGB-D) input or a 6-channel (RGB-HHA where HHA is encoded from Depth referring to dispersion, height above ground and normal angle) input, and then a single CNN module is used for further processing. However, RGB and Depth are fundamentally different. RGB values capture photometric appearance attributes in projected image space, while Depth represents geometric attributes. Although adjacent pixels are close to each other on the image plane, they are not necessarily geometrically coherent in a 3D space. Therefore, simply stitching RGB and Depth cannot fully explore the correlation between RGB and Depth images. In a dual stream design, the architecture uses parallel encoders, and RGB and Depth are processed using separate streams. However, most studies have focused on designing frameworks for processing RGB and Depth images, ignoring the complementarity of features between RGB and Depth, resulting in excessive reliance on individual learning streams, resulting in the increase in computational costs. Of course, this also leads to insufficient feature fusion between RGB and Depth images, resulting in low segmentation accuracy. In this article, considering early feature fusion and supplementation of RGB and Depth images, we propose a cross-modal feature fusion RGB-D semantic segmentation model based on ConvNeXt [8]. By adding cross-modal feature fusion modules after different levels, more sufficient complementarity and fusion of RGB and Depth features were achieved and the accuracy of RGB-D semantic segmentation was improved. The main contributions of this article are as follows.

(1) A cross-modal RGB feature and deep feature fusion module is proposed. Through cross-modal information interaction, the generalization ability of the model is improved, and the inference ability of the model is also improved through the cross-attention mechanism.

(2) An RGB-D semantic segmentation model based on ConvNext's parallel dual two-branch structure is constructed, which can maintain the strong feature extraction capabilities of the RGB and Depth branches by cross-modal feature fusion and effectively integrate and fuse RGB and Depth features. This model exhibits better segmentation performance for large datasets.

2. Related Work

In earlier studies, researchers manually customized the fusion features of RGB-D. In recent years, RGB-D semantic segmentation methods based on deep learning have dominated the mainstream due to the great advantage of deep learning in the ability to extract features [9–22]. ACNet [9] proposed a channel attention module to fuse RGB features and Depth features. The authors of references [10,11] used three channels of horizontal parallax, ground height, angle and gravity to HHA encode Depth images, and this method of processing Depth information has been widely used in later algorithms. FuseNet [12] introduced a fuse layer to embed Depth features into RGB features. The authors of references [13,14] proposed an efficient feature fusion module for objects containing different levels of information by adopting multimodal feature fusion and multi-level feature refinement to capture RGB-D features. LSD-GF [15] introduced a gated fusion layer to adjust the RGB and Depth contributions on each pixel. Depth-aware convolution and pooling were achieved by integrating geometric information into RGB features [16]. CFN [17] utilized Depth information to segment an image into layers representing similar visual features. SCN [18] utilized Depth data to flexibly select useful contextual information for image regions where different objects coexist. J. McCormac et al. [19] superimposed RGB and Depth features into four channels to improve semantic segmentation.

The two-stream structures became the mainstream framework for future RGB-D semantic segmentation due to the high efficiency and variability. Qi et al. [20] introduced a 3D graph neural network to model accurate context through geometric cues provided by the Depth data. Zhang et al. [21] proposed a novel Task-Recursive Learning (TRL) framework to jointly and recurrently conduct three representative tasks containing depth estimation, surface normal prediction and semantic segmentation. Zhou et al. [22] proposed a pattern-structure diffusion (PSD) framework to mine and propagate task-specific and task-across pattern structures in the task-level space for joint depth estimation, segmentation and surface normal prediction. Since RGB-D contains the information of two different modes, the fusion of RGB and Depth information becomes an effective method to improve the accuracy of semantic segmentation [23–28]. Fan et al. [23] constructed an encoder network with two ConvNext-T backplates for each of RGB and Depth, and a decoder network composed of multi-scale supervision and multi-granularity segmentation branches to achieve scene segmentation at different scales. Yang et al. [24] proposed a new framework, MGCNet, that guides the fusion of patterns through differential exploration to reduce collaborative conflicts. In the decoder, a gating feature was proposed to avoid the exclusion of inter-layer information and capture contextual information adequately. Bai et al. [28] proposed a two-branch network called the differential convolution attention network (DCANet), which composed of a pixel differential convolution attention and a set differential convolution attention, and was used to fuse local and global information of two-mode data. Wu et al. [29] proposed a new framework for integrating Depth information into RGB CNN to guide feature extraction on RGB images. Some researchers focus on 3D scene completion, using deep learning methods and RGB-D data to achieve semantic segmentation and complete of three-dimensional scenes [30–32]. These semantic segmentation networks open up new ways to accurately extract Depth information. However, the difficult problem of ways to fully integrate RGB-D information persist. We can conclude that the key challenge of RGB-D semantic segmentation is how to identify the difference between RGB features and Depth features and integrate them effectively and use them to achieve higher segmentation accuracy.

3. Method

3.1. Framework Overview

In this paper, a cross-modal feature fusion RGB-D semantic segmentation model based on ConvNeXt is proposed. The framework of the model is shown in Figure 1. We employ two parallel RGB branches and a Depth branch to extract features from RGB and Depth images. At the same time, the cross-modal feature fusion module is used to supplement the features of RGB and Depth branch, that is, the features of different modalities are supplemented by this module, and then the supplemented features are fused to achieve cross-modal feature fusion.

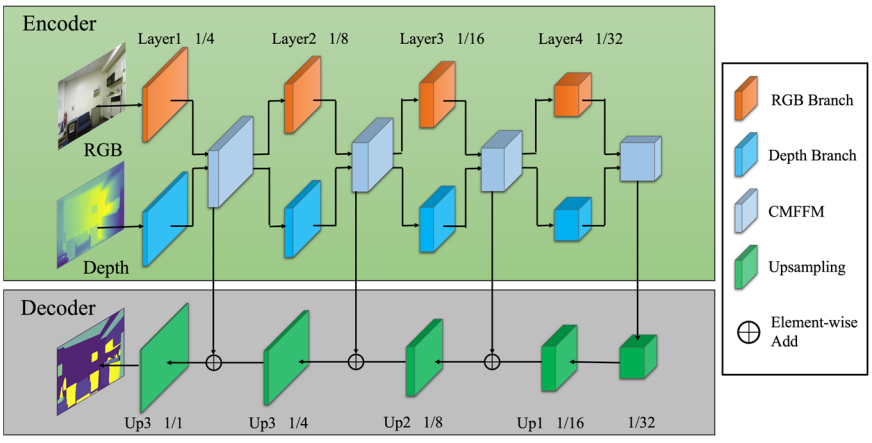

Figure 1. Framework Overview.

The encoder is used to extract RGB and Depth features at different levels, and then the decoder is used to convert the feature maps at different levels into the final semantic map. In order to improve the special utilization of different levels, we introduce multi-level feature supplement in the decoder and use the fused features in different levels in the encoder as the supplement to improve the robustness of the model. The model encoder consists of 4 stages. Later, cross-modal feature fusion module (CMFFM), which is presented in Section 3.3, is embedded in each stage, the RGB features and Depth features are sent to the next layer through the CMFFM, and the fusion features of RGB and Depth are sent to the decoder for feature supplementation. In Figure 1, the functions of downsampling and upsampling are to resize the image size. For example, the "1/4" in encoder and decoder part means the spatial size is reduced and enlarged to 1/4 and 4 times of the original size, respectively.

3.2. ConvNeXt

Since the proposal of VIT [33], it has rapidly replaced the convolutional network as the state-of-the-art image classification model. Using Transformer as the backbone network and introducing convolutional neural networks (ConvNet) enables Transformer to be applied in a variety of visual tasks, such as object detection, semantic segmentation, etc. ConvNeXt builds a network entirely composed of standard ConvNet modules based on the design of VIT and ResNet, which is superior to Transformer in accuracy and scalability while maintaining the simplicity and efficiency of standard ConvNet. The network structure of ConvNeXt consists of four layers shown in Figure 2. Layers 1, 2, and 4 contain three basic blocks, and Layer 3 contains 27 basic blocks. Each basic block contains three convolutional layers, and the Gaussian Error Linear Unit (GELU) [34] activation function and the simpler Layer Normalization (LN) [35] are used in each basic block. RGB image and Depth image are input through their respective branches in the model (Figure 1). After passing through the 1st downsampling layer, the RGB data and Depth data have the same data shape with 192 channels through the convolution operation (in Figure 2), and then the 192-channelled data are sent to Layer 1 for processing, and after the fusion module (i.e., CMFFM) of Layer 1, the data is sent to Layer 2 (for further downsampling) and the decoder modular (for upsampling with feature supplement) at the same time, then to Layers 3 and 4. Each Layer is connected by the downsampling layer.

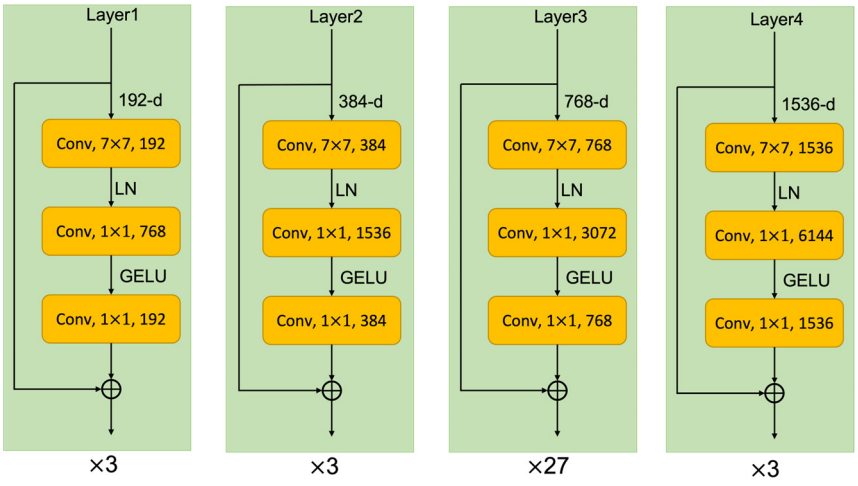

Figure 2. ConvNeXt structure.

3.3. Cross-Modal Feature Fusion Module (CMFFM)

We propose a cross-modal feature fusion module (CMFFM) to fuse RGB and Depth features, as shown in Figure 3. In order to improve the extraction and fusion capabilities of the multimodal features of the model, CMFFM processes the input data from RGB and depth features. For the input RGB feature, $F_{RGB} \in \mathbb{R}^{H \times W \times C}$ and the Depth feature

$F_{Depth} \in \mathbb{R}^{H\times W\times C}$, global max pooling and global average pooling are adopted along the channel-wise phase to retain more information, and finally four result vectors are obtained. Connecting vectors $Y_{RGB} \in \mathbb{R}^{2C}$ and $Y_{Depth} \in \mathbb{R}^{2C}$ are obtained by the RGB and the Depth feature, respectively. After MLP operations on Y_{RGB} and Y_{Depth}, respectively, the weights $W^C_{RGB} \in \mathbb{R}^{2C}$ and $W^C_{Depth} \in \mathbb{R}^{2C}$ are obtained through the Sigmoid function to split W^C_{RGB} into $W^{C1}_{RGB} \in \mathbb{R}^C$ and $W^{C2}_{RGB} \in \mathbb{R}^C$ and split W^C_{Depth} into $W^{C1}_{Depth} \in \mathbb{R}^C$ and $W^{C2}_{Depth} \in \mathbb{R}^C$.

$$\begin{array}{c} W^{C1}_{RGB}, W^{C2}_{RGB} = \mathrm{F}_{split}(Sigmoid(MLP(Y_{RGB}))) \\ W^{C1}_{Depth}, W^{C2}_{Depth} = \mathrm{F}_{split}\Big(Sigmoid\Big(MLP\Big(Y_{Depth}\Big)\Big)\Big) \end{array}, \tag{1}$$

where F_{split} means to split a vector into two vectors. The weights $F^C_{RGB} \in \mathbb{R}^{H\times W\times C}$ and $F^C_{Depth} \in \mathbb{R}^{H\times W\times C}$ of the RGB modality and the Depth modality in the channel-wise phase are calculated by the following formulas:

$$\begin{array}{c} F^C_{RGB} = \lambda_{C1} W^{C1}_{RGB} * F_{RGB} + \lambda_{C2} W^{C2}_{RGB} * F_{Depth} \\ F^C_{Depth} = \lambda_{C1} W^{C1}_{Depth} * F_{Depth} + \lambda_{C2} W^{C2}_{Depth} * F_{RGB} \end{array}, \tag{2}$$

where $*$ represents multiplication, λ_{C1} and λ_{C2} are hyperparameters, which are both set to 0.5 in this paper. In the spatial-wise phase, F_{RGB} and F_{Depth} are connected, and after two convolution layers with a convolution kernel of 1×1 and a RELU function, the Sigmoid function is used to obtain the feature map $Z \in \mathbb{R}^{H\times W\times 2}$, and then split it into two weight maps W^S_{RGB} and W^S_{Depth}. The calculation formulas for the weights $F^S_{RGB} \in \mathbb{R}^{H\times W\times C}$ and $F^S_{Depth} \in \mathbb{R}^{H\times W\times C}$ in the spatial-wise phase are calculated as follows:

$$Z = Conv_{1\times 1}\Big(RELU\Big(Conv_{1\times 1}\Big(F_{RGB}\Big|\Big|F_{Depth}\Big)\Big)\Big), \tag{3}$$

$$W^S_{RGB}, W^S_{Depth} = \mathrm{F}_{split}(Sigmoid(Z)), \tag{4}$$

$$\begin{array}{c} F^S_{RGB} = W^S_{RGB} * F_{RGB} \\ F^S_{Depth} = W^S_{Depth} * F_{Depth} \end{array}, \tag{5}$$

where $|.|$ represents the connection operation, $Conv_{1\times 1}$ is the convolution operation with the convolution kernel of 1×1. The final RGB feature output $F^{out}_{RGB} \in \mathbb{R}^{H\times W\times C}$, Depth feature output $F^{out}_{Depth} \in \mathbb{R}^{H\times W\times C}$ and fusion feature output $F_{Fusion} \in \mathbb{R}^{H\times W\times C}$ are calculated as follows:

$$\begin{array}{c} F^{out}_{RGB} = F_{RGB} + \lambda_C F^C_{RGB} + \lambda_S F^S_{RGB} \\ F^{out}_{Depth} = F_{Depth} + \lambda_C F^C_{Depth} + \lambda_S F^S_{DEpth} \end{array}, \tag{6}$$

$$F_{Fusion} = Conv_{1\times 1}\Big(RELU\Big(Conv_{1\times 1}\Big(F^{out}_{RGB}\Big|\Big|F^{out}_{Depth}\Big)\Big)\Big) + Conv_{1\times 1}\Big(F^{out}_{RGB}\Big|\Big|F^{out}_{Depth}\Big). \tag{7}$$

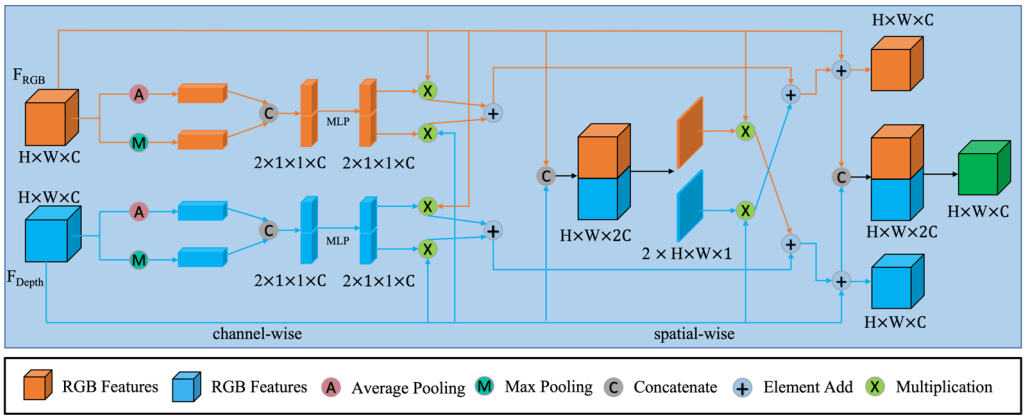

Figure 3. CMFFM structure.

After CMFFM, F_{RGB}^{out} and F_{Depth}^{out} are used as the input of the next layer again, and F_{Fusion} is sent to the decoder as a supplementary feature of the features of different levels in the decoder.

4. Results

4.1. Experimental Parameters and Evaluation Indexes

Because the deep neural network training process has many iterations and a large number of matrix operations and requires a large amount of computing resources, the high-performance graphics processing unit (GPUs) is indispensable. In this experiment, NVIDIA GeForce RTX 3090 is used with graphics memory of 24 GB. The CPU model is Intel(R) Core (TM) i9-10900K (3.70 GHz) with the memory size of 128 GB. The operating system is Ubuntu20.04, and the model was implemented using PyTorch deep learning framework, while CUDA Toolkit 11.1 and CUDNN V8.0.4 are used for computation acceleration. Anaconda3.6 and Python are used as the development environment and programming language for the model. In our model, the weights for all layers of the network are initialized to a commonly normal distribution with the mean of 0, the variance of 0.01 and the deviation of 0. The two parameters λ_{C1} and λ_{C2} in Equation (2) are both initialized to 0.5. The two parameters λ_C and λ_S of Equation (6) are both initialized to 0.5. In model training, two public datasets (NYUDv2 and SUN-RGBD) and a self-built Rice-RGB-D dataset are used. For public datasets, the size of the images, which are input to the model, is 480 × 640, and for the rice dataset, the size is 512 × 160. Adam is used as the optimizer with a learning rate of 2×10^{-5} and weight decay of 5×10^{-4}. We adopt a poly learning rate schedule with factor $(1 - iter/iter_{max})^{0.9}$ and use cross-entropy as the loss function, which is defined as

$$Loss = -\frac{1}{n}\sum_{x}\left[y\ln y' + (1-y)\ln(1-y')\right], \tag{8}$$

where y and y' denote the expected and actual output. We use the batch size of eight and epoch of 200 to train on all datasets. Experimental operation parameters are shown in Table 1.

Table 1. Running parameter table.

Parameter		Values
Operating system		Ubuntu20.04
CPU		Intel(R) Core (TM) i9-10900K (3.70 GHz)
GPU		GeForce RTX 3090
Development environment		Anaconda3.6
Framework		Pytorch1.8
Input size	NYUDv2	480 × 640
	SUN-RGBD	480 × 640
	RICE-RGBD	512 × 160
Learning Rate		2×10^{-5}
Batch size		8
Epoch		200

To evaluate the performance of the different methods, we use prevailing Pixel Accuracy (Pixel Acc.) and mean Intersection over Union (mIoU) as evaluation indicators, which are defined as [36]

$$Pixel\ Acc. = \sum_i \frac{n_{ii}}{s}, \tag{9}$$

$$mIoU = \frac{1}{n_c} \sum_i \frac{n_{ii}}{(s_i - n_{ii} + \sum_j n_{ji})}, \tag{10}$$

where n_{ji} is the number of pixels with ground-truth class j predicted as class i (when j equals to $i, n_{ji} = n_{ii}$), n_c is the total number of classes, s_i is the number of pixels with ground truth class i, and s is the total number of all pixels.

4.2. Public Datasets

We conduct experiments on two public benchmark datasets: NYUDv2 [37], SUN-RGBD [38]. The NYUDv2 dataset contains 1449 RGB-D images of 40 classes, of which 795 are used for training and the remaining 654 are used for testing. The SUN-RGBD dataset has 37 categories and contains 10,335 RGB-D images (5285 for training and 5050 for testing). In the experiment, the data processing includes adjusting resolution, data enhancement, data labeling and data normalization. We adjust all RGB images, Depth images and ground real images to a high spatial resolution of 480 × 640. During training, we use data enhancement to improve data diversity including random scaling, cropping and flipping to the inputs of RGB image and Depth image, respectively. For RGB images, we further enhance them by applying random hue, brightness, and saturation adjustments. The RGB images and Depth images are normalized to 0–1.

During the experiment, we use the same experimental conditions to repeat the experiment for five times to ensure the validity of the evaluation results. The experimental results are the average of five replicates. Through the experiments on NYUDv2 and SUN-RGBD datasets, it can be observed from Tables 2 and 3 that our proposed model achieves good results. For example, the application of our model on the NYUDv2 dataset shows the mIoU of 51.9% (the third out of the nine comparable models) and the pixel accuracy of 76.8% (the fourth out of the nine models). The application of our model on the SUN-RGBD dataset shows the mIoU (53.5%) is the best among the nine comparative models and the pixel accuracy (82.5%) is the third out of the nine models.

Table 2. Comparison with the state-of-the-art models on the NYUDv2 dataset (the top ones are marked in bold).

Method	mIoU (%)	Pixel Acc. (%)
3DGNN [20]	43.1	-
Kong et al. [39]	44.5	72.1
RAFNet [40]	47.5	73.8
ACNet [9]	48.3	-
CANet [41]	51.2	76.6
NANet [36]	51.4	77.1
DCANet [28]	53.3	78.2
MGCNet [24]	**54.5**	**78.7**
ConvNeXt-CMFFM	51.9	76.8

Table 3. Comparison with the state-of-the-art models on the SUN-RGBD dataset (the top ones are marked in bold).

Method	mIoU (%)	Pixel Acc. (%)
3DGNN [20]	45.9	-
Kong et al. [39]	45.1	80.3
RAFNet [40]	47.2	81.3
ACNet [9]	48.1	-
CANet [41]	48.1	81.6
NANet [36]	48.8	82.3
DCANet [28]	49.6	82.6
MGCNet [24]	51.5	**86.5**
ConvNeXt-CMFFM	**53.5**	82.5

Figure 4 shows the segmentation results of our model on NYUDv2 dataset (three indoor scenes including wall, chairs, sofa, clothes and so on in different colors), from which it can be observed that our model can well segment various objects in the scenes with the good segment effect close to the truth.

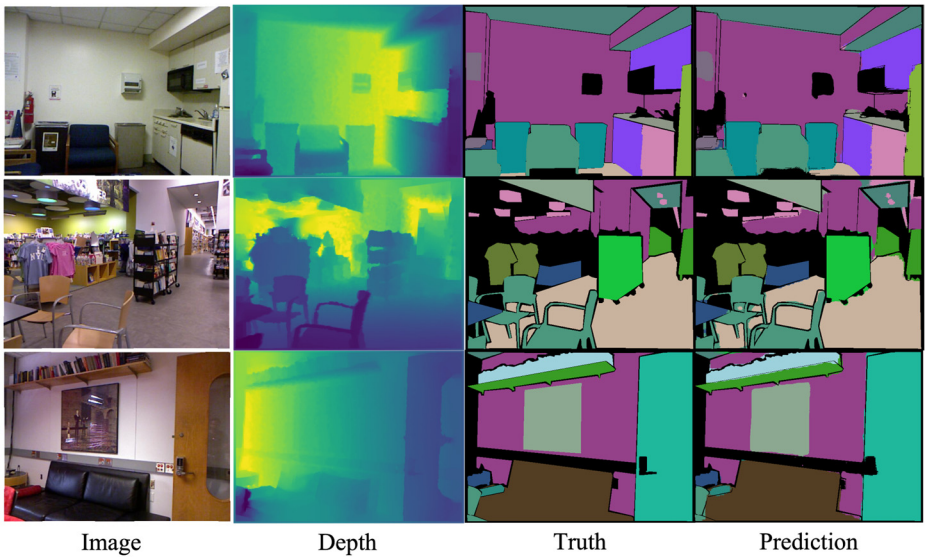

Figure 4. NYUDv2 segmentation results.

4.3. Our Rice Dataset

At the same time, we conduct practical application tests on the self-built RICE-RGBD image dataset, which contains two types of data, namely RGB-D data of single rice plant and whole rice cluster, as shown in Figure 5. There are 10,000 RICE-RGBD images in total in the dataset, with 6000 for training and 4000 for testing. Similar to the public dataset, the

data processing includes adjusting resolution, data enhancement, data labeling and data normalization. We adjust all RGB images, Depth images and ground real images to a high spatial resolution of 512 × 160. During training, we us data enhancement to improve data diversity including random scaling, cropping and flipping to the inputs of RGB image and Depth image, respectively. For RGB images, we further enhance them by applying random hue, brightness and saturation adjustments. Both the RGB images and Depth images are normalized to 0–1. The ground real images of our rice dataset are labeled. As can be seen from the Figure 5, ConvNeXt-CMFFM is able to well segment the spike and straw of rice. This experiment shows that our model can be well used in practical agricultural application.

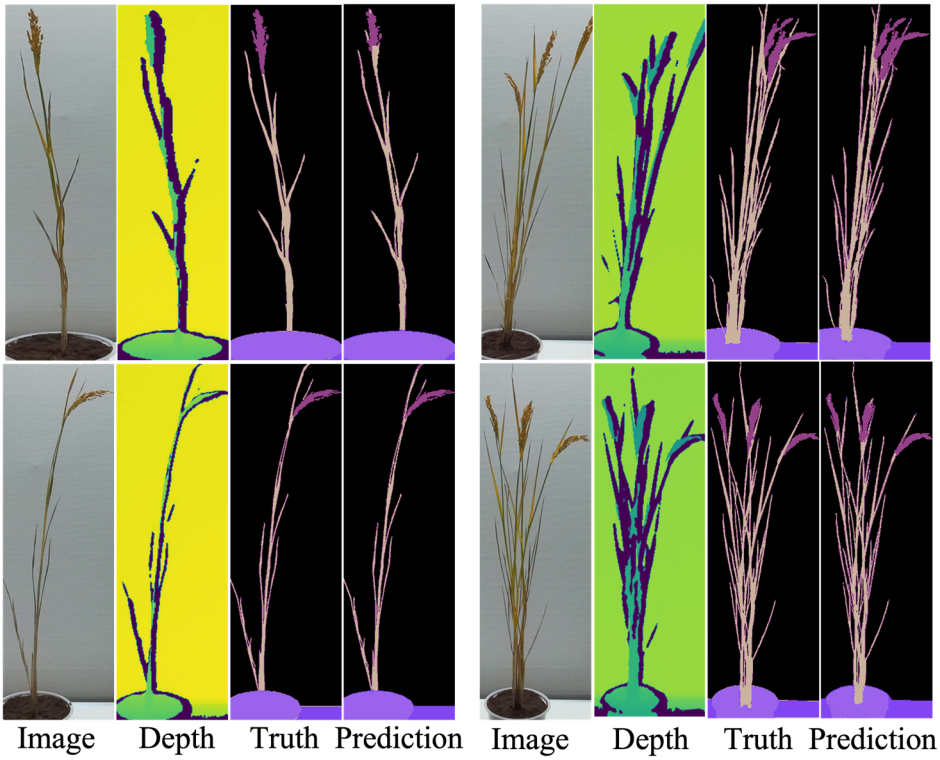

Figure 5. RICE-RGBD segmentation results.

4.4. Ablation Study

To further illustrate the validity of our proposed CMFFM, we compared the mIoU and Pixel Acc. metrics of ConvNext with and without the CMFFM on three datasets and the results are shown in Table 4. It shows that ConvNext-CMFFM performs better than ConvNeXt on the three datasets. Specifically, for our rice dataset, compared with ConvNeXt (without CMFFM), the mIoU score of ConvNext-CMFFM is increased from 71.5% to 74.8% and its pixel accuracy is increased from 86.2% to 88.3%. This indicates that CMFFM fusion of data features of different modes can help the model extract more important data features, improve the generalization ability and inference ability of the model, and further promote the segmentation performance of the model.

Table 4. Comparison results.

Method	Dataset	mIoU (%)	Pixel Acc. (%)
ConvNeXt-CMFFM	NYUDv2	51.9	76.8
	SUN-RGB	53.5	82.5
	RICE-RGBD	74.8	88.3
ConvNeXt	NYUDv2	50.2	76.1
	SUN-RGB	50.9	79.9
	RICE-RGBD	71.5	86.2

4.5. Discussion

In our model, the CMFFM proposed by us integrates RGB and Depth image features through channel-wise and spatial-wise phases. Meanwhile, RGB feature and Depth feature complement each other and participate in the calculation of the next layer of their respective branches. The ConvNeXt further enhances CNN feature mining capabilities. It can be seen from the experimental results that the segmentation results of our method on the public datasets of NYUDv2 and SUN-RGBD are competitive. Through the comparison of ablation studies, it can be determined that the proposed CMFFM improves the accuracy of model segmentation, which indicates that our module can realize the deep fusion of RGB features and Depth features. Furthermore, our proposed CMFFM can also be embedded into other backbone networks to improve the segmentation accuracy of models.

In addition, it can be seen that on the dataset of NYUDv2, the segmentation accuracy of our model is lower than that of DCANet and MGCNet, but on SUN-RGBD dataset, the segmentation accuracy of our model is the best with the highest mIoU. It shows that our model has better performance on the dataset with more training samples (for example, 10,335 samples of SUN-RGBD dataset) and slightly poor performance on the dataset with fewer data samples (for example, 1449 samples of NYUDv2). This indicates that although CMFFM proposed by us can integrate the features of RGB and Depth relatively well, it still requires a large number of training samples to learn more critical features, and there is still great room for improvement in the feature fusion of RGB and Depth.

At present, the parameters of our model are 309×10^6 and FLOPs are 408×10^9, which can meet the general desktop requirements. The processing speed of the model for a group of data (including one RGB image and one Depth image) is about 0.03 s, that is, the processing speed is about 33 frames per second, which can meet the requirements of common real-time line tasks. However, on embedded devices, the resource requirement of our model makes it difficult to apply it in embedded devices. Reducing the resource overhead of our model will be the future improvement goal.

Thus, the limitations of this model can be summarized as follows. First, the proposed method still needs to be improved for small sample data sets. Second, the two-stream structure makes the model cost a lot of resources, which limits its ability to be transplanted to portable embedded platforms. Finally, our method has a high requirement on the quality of Depth images. When collecting data in actual agricultural scenarios, due to the impact of equipment and environment, the collected depth information may be missing. At this point, this model is not sufficient to fuse the features of RGB and depth.

In future work, we will further explore the correlation between RGB and Depth to achieve a deeper cross-modal feature fusion to employ fewer training samples, learn more key features and reduce the dependence of the model on data. Meanwhile, we are considering to design a lightweight framework to reduce the model's demand for resources and improve and expand its real-time application performance on embedded system. In addition, in actual agricultural scenarios, we will also consider using RGB image and the complementary Depth information to perform the semantic segmentation of RGB-D.

5. Conclusions

In this paper, we propose a cross-modal feature fusion RGB-D semantic segmentation model based on ConvNeXt to better utilize multi-stage RGB-D features for semantic segmentation. In particular, a novel Cross-modal Feature Fusion Module (CMFFM), embedded in multiple stages of the model, is able to capture both spatial- and channel-wise features in RGB-D features at various stages. Extensive experimental results confirm the effectiveness of our method on the public NYUDv2 and SUN-RGBD datasets. The results on our self-built rice dataset also confirm the practical agricultural application of our method.

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

Funding: This research was funded by the National Natural Science Foundation of China (No. 11675261 and 12075315) and the horizontal projects of China Agricultural University (No. 202105511011054 and 202005511011203).

Data Availability Statement: This study uses the NYUDv2 dataset from https://cs.nyu.edu/~silberman/projects/indoor_scene_seg_sup.html (accessed on accessed on 13 March 2022), the SUN-RGBD dataset from https://rgbd.cs.princeton.edu/ (accessed on accessed on 15 March 2022) and RICE-RGBD dataset from the corresponding author.

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

References

1. Sun, L.; Yang, K.; Hu, X.; Hu, W.; Wang, K. Real-Time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images. *IEEE Robot. Autom. Lett.* **2020**, *5*, 5558–5565. [CrossRef]
2. Seichter, D.; Köhler, M.; Lewandowski, B.; Wengefeld, T.; Gross, H.M. Efficient RGB-D semantic segmentation for indoor scene analysis. In Proceedings of the IEEE International Conference on Robotics and Automation, Hongkong, China, 21–23 April 2021; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2021; Volume 2021, pp. 13525–13531.
3. Mohammed, M.S.; Abduljabar, A.M.; Faisal, M.M.; Mahmmod, B.M.; Abdulhussain, S.H.; Khan, W.; Liatsis, P.; Hussain, A. Low-cost autonomous car level 2: Design and implementation for conventional vehicles. *Results Eng.* **2023**, *17*, 100969. [CrossRef]
4. Faisal, M.M.; Mohammed, M.S.; Abduljabar, A.M.; Abdulhussain, S.H.; Mahmmod, B.M.; Khan, W.; Hussain, A. Object de-tection and distance measurement using AI. In Proceedings of the International Conference on Developments in Esystems Engineering, DeSE, Sharjah, United Arab Emirates, 7–10 December 2021; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2021; Volume 2021, pp. 559–565.
5. Duarte, J.; Martínez-Flórez, G.; Gallardo, D.I.; Venegas, O.; Gómez, H.W. A bimodal extension of the epsilon-skew-normal model. *Mathematics* **2023**, *11*, 507. [CrossRef]
6. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
7. 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] [PubMed]
8. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–23 June 2022; pp. 11976–11986.
9. Hu, X.; Yang, K.; Fei, L.; Wang, K. ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1440–1444.
10. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 345–360.
11. Gupta, S.; Arbelaez, P.; Malik, J. Perceptual organization and recognition of indoor scenes from RGB-D images. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 564–571.
12. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In Proceedings of the Computer Vision—ACCV 2016, Taipei, Taiwan, 20–24 November 2016; Lai, S.H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 213–228.
13. Lee, S.; Park, S.J.; Hong, K.S. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2017; Volume 2017, pp. 4990–4999.
14. Chen, X.; Lin, K.Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-directional cross-modality feature propagation with seperation-and-aggregation gate for RGB-D semantic segmentation. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 561–577.

15. Cheng, Y.; Cai, R.; Li, Z.; Zhao, X.; Huang, K. Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 1475–1483.
16. Wang, W.; Neumann, U. Depth-aware CNN for RGB-D segmentation. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 144–161.
17. Lin, D.; Chen, G.; Cohen-Or, D.; Heng, P.A.; Huang, H. Cascaded feature network for semantic segmentation of RGB-D images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2017; Volume 2017, pp. 1320–1328.
18. Lin, D.; Zhang, R.; Ji, Y.; Li, P.; Huang, H. SCN: Switchable context network for semantic segmentation of RGB-D images. *IEEE Trans. Cybern.* **2020**, *50*, 1120–1131. [CrossRef] [PubMed]
19. McCormac, J.; Handa, A.; Davison, A.; Leutenegger, S. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In Proceedings of the IEEE International Conference on Robotics and Automation, Singapore, Singapore, 29 May–3 June 2017; pp. 4628–4635.
20. Qi, X.; Liao, R.; Jia, J.; Fidler, S.; Urtasun, R. 3D graph neural networks for RGBD semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2017; Volume 2017, pp. 5209–5218.
21. Zhang, Z.; Cui, Z.; Xu, C.; Jie, Z.; Li, X.; Yang, J. Joint task-recursive learning for RGB-D scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2608–2623. [CrossRef] [PubMed]
22. Zhou, L.; Cui, Z.; Xu, C.; Zhang, Z.; Wang, C.; Zhang, T.; Yang, J. Pattern-structure diffusion for multi-task learning. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; IEEE Computer Society: Washington DC, USA, 2020; pp. 4513–4522.
23. Fan, J.; Zheng, P.; Lee, C.K.M. A multi-granularity scene segmentation network for human-robot collaboration environment perception. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Kyoto, Japan, 23–27 October 2022; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2022; Volume 2022, pp. 2105–2110.
24. Yang, E.; Zhou, W.; Qian, X.; Yu, L. MGCNet: Multilevel gated collaborative network for RGB-D semantic segmentation of indoor scene. *IEEE Signal Process. Lett.* **2022**, *29*, 2567–2571. [CrossRef]
25. Hua, Z.; Qi, L.; Du, D.; Jiang, W.; Sun, Y. Dual attention based multi-scale feature fusion network for indoor RGBD semantic segmentation. In Proceedings of the International Conference on Pattern Recognition, Montreal, QC, Canada, 21–25 August 2022; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2022; Volume 2022, pp. 3639–3644.
26. Wu, P.; Guo, R.; Tong, X.; Su, S.; Zuo, Z.; Sun, B.; Wei, J. Link-RGBD: Cross-guided feature fusion network for RGBD semantic segmentation. *IEEE Sensors J.* **2022**, *22*, 24161–24175. [CrossRef]
27. Chen, J.; Zhan, Y.; Xu, Y.; Pan, X. FAFNet: Fully aligned fusion network for RGBD semantic segmentation based on hierarchical semantic flows. *IET Image Process.* **2023**, *17*, 32–41. [CrossRef]
28. Bai, L.; Yang, J.; Tian, C.; Sun, Y.; Mao, M.; Xu, Y.; Xu, W. DCANet: Differential convolution attention network for RGB-D semantic segmentation. *arXiv* **2022**, arXiv:2210.06747. [CrossRef]
29. Wu, Z.; Allibert, G.; Stolz, C.; Ma, C.; Demonceaux, C. Depth-adapted CNNs for RGB-D semantic segmentation. *arXiv* **2022**, arXiv:2206.03939. [CrossRef]
30. Cai, Y.; Chen, X.; Zhang, C.; Lin, K.Y.; Wang, X.; Li, H. Semantic scene completion via integrating instances and scene in-the-loop. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; IEEE Computer Society: Washington DC, USA, 2021; pp. 324–333.
31. Price, A.; Huang, K.; Berenson, D. Fusing RGBD tracking and segmentation tree sampling for multi-hypothesis volumetric segmentation. In Proceedings of the IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May–5 June 2021; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2021; Volume 2021, pp. 9572–9578.
32. Li, S.; Zou, C.; Li, Y.; Zhao, X.; Gao, Y. Attention-based multi-modal fusion network for semantic scene completion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11402–11409.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
34. Hendrycks, D.; Gimpel, K. Gaussian error linear units (GELUs). *arXiv* **2016**, arXiv:1606.08415. [CrossRef]
35. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450. [CrossRef]
36. Zhang, G.; Xue, J.H.; Xie, P.; Yang, S.; Wang, G. Non-local aggregation for RGB-D semantic segmentation. *IEEE Signal Process. Lett.* **2021**, *28*, 658–662. [CrossRef]
37. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from RGBD images. In Proceedings of the Computer Vision—ECCV 2012, Florence, Italy, 7–13 October 2012; pp. 746–760.
38. Song, S.; Lichtenberg, S.P.; Xiao, J. SUN RGB-D: A RGB-D scene understanding benchmark suite. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2012; IEEE Computer Society: Washington DC, USA, 2015; pp. 567–576.
39. Kong, S.; Fowlkes, C. Recurrent scene parsing with perspective understanding in the loop. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE Computer Society: Washington DC, USA, 2018; pp. 956–965.

40. Yan, X.; Hou, S.; Karim, A.; Jia, W. RAFNet: RGB-D attention feature fusion network for indoor semantic segmentation. *Displays* **2021**, *70*, 102082. [CrossRef]
41. Zhou, H.; Qi, L.; Huang, H.; Yang, X.; Wan, Z.; Wen, X. CANet: Co-attention network for RGB-D semantic segmentation. *Pattern Recognit.* **2022**, *124*, 108468. [CrossRef]