remote sensing

MDPI

Article

CTFuseNet: A Multi-Scale CNN-Transformer Feature Fused Network for Crop Type Segmentation on UAV Remote Sensing Imagery

Jianjian Xiang , Jia Liu *, Du Chen , Qi Xiong and Chongjiu Deng

School of Computer Science, China University of Geosciences, Wuhan 430074, China
* Correspondence: liujia@cug.edu.cn

Abstract: Timely and accurate acquisition of crop type information is significant for irrigation scheduling, yield estimation, harvesting arrangement, etc. The unmanned aerial vehicle (UAV) has emerged as an effective way to obtain high resolution remote sensing images for crop type mapping. Convolutional neural network (CNN)-based methods have been widely used to predict crop types according to UAV remote sensing imagery, which has excellent local feature extraction capabilities. However, its receptive field limits the capture of global contextual information. To solve this issue, this study introduced the self-attention-based transformer that obtained long-term feature dependencies of remote sensing imagery as supplementary to local details for accurate crop-type segmentation in UAV remote sensing imagery and proposed an end-to-end CNN–transformer feature-fused network (CTFuseNet). The proposed CTFuseNet first provided a parallel structure of CNN and transformer branches in the encoder to extract both local and global semantic features from the imagery. A new feature-fusion module was designed to flexibly aggregate the multi-scale global and local features from the two branches. Finally, the FPNHead of feature pyramid network served as the decoder for the improved adaptation to the multi-scale fused features and output the crop-type segmentation results. Our comprehensive experiments indicated that the proposed CTFuseNet achieved a higher crop-type-segmentation accuracy, with a mean intersection over union of 85.33% and a pixel accuracy of 92.46% on the benchmark remote sensing dataset and outperformed the state-of-the-art networks, including U-Net, PSPNet, DeepLabV3+, DANet, OCRNet, SETR, and SegFormer. Therefore, the proposed CTFuseNet was beneficial for crop-type segmentation, revealing the advantage of fusing the features found by the CNN and the transformer. Further work is needed to promote accuracy and efficiency of this approach, as well as to assess the model transferability.

Keywords: precision agriculture; UAV remote sensing; semantic segmentation; deep learning; CNN; transformer; feature fusion

Citation: Xiang, J.; Liu, J.; Chen, D.; Xiong, Q.; Deng, C. CTFuseNet: A Multi-Scale CNN-Transformer Feature Fused Network for Crop Type Segmentation on UAV Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 1151. https://doi.org/10.3390/rs15041151

Academic Editor: Adel Hafiane

Received: 29 December 2022
Revised: 16 February 2023
Accepted: 18 February 2023
Published: 20 February 2023

1. Introduction

Global food security is a worldwide issue and presents serious challenges. Precision agriculture is crucial to ensuring food security, as well as the social and economic development [1]. Accurate and timely crop-type-segmentation information facilitates the precision agriculture activities, e.g., irrigation scheduling, yield estimation, land use assessment, as well as governmental agricultural policy decisions [2].

As compared to the manual surveys that consume a large amount of human and material resources, remote sensing is capable of acquiring information in a rapid, large-scale, non-destructive way and has been involved in various fields, e.g., crop biomass estimation, crop yield estimation, crop area estimation, plant protection, and agricultural disaster prediction [3]. In addition to spaceborne and airborne platforms, unmanned aerial vehicles (UAVs) have recently emerged as effective tools for agriculture monitoring. UAVs are able to fly at low altitudes and obtain images with very high spatial resolution and are not affected by clouds [4]. Due to these outstanding advantages, UAVs have been applied

in agricultural applications, such as weed mapping [5], insect detection [6], plant disease detection [7], crop yield prediction [8], etc.

Machine learning (ML) algorithms, such as maximum likelihood [9], support vector machines [10], and random forests [11], have long been applied in a variety of agriculture applications to construct linear or non-linear patterns and correlations from samples. A typical ML-based crop-type-segmentation pipeline includes the study area selection, data acquisition, data preprocessing, feature selection or transformation, and the application of ML algorithms. However, the performance of ML-based methods relies heavily on handcrafted feature extraction techniques and expertise. Deep-learning (DL) models have demonstrated superior performance in accuracy, efficiency, and generalization, over traditional machine-learning algorithms in complex tasks, due to their powerful automatic feature extraction and non-linear expression capabilities. In the past years, DL has achieved success in various fields, including natural language processing, computer vision, etc., as well as in the remote sensing community.

Early attempts demonstrated that crop-type determination using DL-based image classification algorithms could achieve high accuracy [12,13]. However, DL-based image classification could only determine the type of the input images and often required classification with a sliding window on large-scale images to generate crop maps, which limits its efficiency. Meanwhile, semantic segmentation predicts the class of each pixel for the images, which is more accurate and efficient for crop type mapping. Since the fully convolutional network (FCN) [14] first expanded the use of end-to-end convolutional neural networks (CNNs) for semantic segmentation, the accuracy of image semantic segmentation has been continuously improved, and many advanced CNN-based semantic segmentation networks, e.g., SegNet [15], U-Net [16] and DeepLab [17] have been proposed. Research on crop-type segmentation on UAV remote sensing images is emerging [18,19] and becoming mainstream. These networks are typically using an encoder–decoder architecture, which uses CNNs for feature extraction in the encoder and pixel-level class segmentation in the decoder. In the feature extraction of the encoder, the contextual information of an image is crucial [20,21]. However, due to the limitation of the receptive field of CNNs [22], only limited contextual information can be processed, and the restriction of long-range scene perception of the full image may lead to the wrong categorization of the crop field when processed by segmentation networks.

In recent years, an attention mechanism that allocates limited attention resources to quickly filter valuable information has been widely applied in the field of computer vision, including such methods as CBAM [23], SENet [24], and GSoP [25] for CNNs. Recently, transformers with an architecture based on an self-attention mechanism proposed by Google have achieved state-of-the-art results in natural language processing [26]. Dosovitskiy et al. proposed a vision transformer (ViT) based on the concept of transformers, in which they had replaced the CNN architecture with a self-attentive mechanism and attained excellent results in computer vision tasks [27]. A transformer was able to extract the global features of images, and transformer-based semantic segmentation algorithms such as SETR [21], SegFormer [28], and Swin Transformer [29] have achieved state-of-the-art results and been applied in remote sensing applications, such as land-cover mapping [30].

However, when applying a transformer-based networks to crop-type segmentation with UAV remote sensing images directly, satisfactory results can not always be obtained mainly due to the following reasons. Spatial information and local features, e.g., edge details, color, texture and shape of crops, in remote sensing images are crucial factors affecting the semantic segmentation results [31]. However, a transformer divides the high-resolution UAV imagery into patches, which may lead to an incomplete structure of crop-field edges [32]. In addition, transformers compress the image into one-dimensional tokens and send them to the transformer encoder as a sequence, which then leads to the loss of spatial information and local information. This is not conducive to recovering the detailed information in the decoder stage of semantic segmentation networks, resulting in a decrease in the accuracy [33].

Recently, there have been a few attempts to use CNNs and transformers simultaneously for crop type mapping. Li et al. [34] combined the ability of a CNN to express spatial and spectral correlations of remote sensing images with that of a transformer to obtain temporal correlations, and they installed a CNN and a transformer on the same processing flow. Wang et al. [32] designed a coupled CNN–transformer network using different feature-fusion modules for different resolutions and additional loss functions during the model training process. In the wider range of areas for remote sensing image interpretation, more research towards combining CNN and transformer architectures to improve performance have been conducted, such as for object detection [35], image pan-sharpening [36], SAR image classification [37], and fine-grained ship classification [38]. In these studies, combining a CNN and a transformer to take advantage of both methods' benefits is challenging. Typically, the features extracted by the CNN have been fed into the transformer, or vice versa. Meanwhile, the separate global and local feature extraction processes provide the flexibility to switch backbone networks and use their pre-trained weights. However, fully aggregating and decoding the global and local features is difficult. In computer vision, there have been efforts in aggregating and decoding local and global features to improve performance. For the CNN-based semantic segmentation, U-Net [16] used concatenation and deconvolution to aggregate features at different scales while PSPNet [39] and DeepLabV3+ [40] employed pyramid-structured decoders to aggregate and decode features at various scales. For the transformer-based semantic segmentation, SegFormer [28] integrated a lightweight decoder composed of various multi-layer perceptrons (MLP) and a Swin transformer [29] utilized UperHead from UperNet [41] to aggregate the self-attention generated from windows of different sizes. The aforementioned works that aggregated local features from CNNs and global features from transformers are discussed in Section 5.3.

Therefore, in this study, we proposed an end-to-end CNN–transformer feature-fused network (CTFuseNet) to achieve accurate crop-type segmentation in order to solve the main issues, as previously mentioned: (1) CNN-based crop segmentation methods lack long-range contextual information, which may lead to the wrong categorization of the crop field; (2) transformer-based methods slice the high-resolution UAV imagery into patches, which may lead to the loss of spatial information; and (3) fully extracting, aggregating and decoding the multi-scale global and local features with a transformer and a CNN has yet to be fully examined. Comprehensive experiments with advanced segmentation networks including U-Net, PSPNet, DeepLabV3+, DANet, OCRNet, SETR and SegFormer, as well as the performance analysis of the proposed module, were conducted. The experimental results showed the effectiveness of the proposed method. Specifically, the main contributions of this paper are listed, as follows:

- We proposed CTFuseNet, a global–local feature-fused network based on transformer and CNN architectures for accurate crop-type segmentation. It provides a parallel structure of CNN and transformer branches in the encoder to extract both local and global semantic features of remote sensing imagery and outputs multi-scale features for aggregation.
- We design a new lightweight feature-fusion module to flexibly aggregate the multi-scale local and global features output from CNN and transformer branches in the proposed CTFuseNet.
- The FPNHead from feature pyramid network servers as the decoder in the proposed CTFuseNet, instead of maintaining the original All-MLP decoder in SegFormer. It allows decoding features at different scales, thus adapting to the fused features and further improving the accuracy.

The rest of this paper is organized as follows. In Section 2, the proposed CTFuseNet is presented, including the parallel structure of CNN and transformer branches in the encoder, the feature-fusion module, and the decoder for the multi-scale global–local features. The study area, the dataset, the experimental settings, and the methods for performance comparison are provided in Section 3. The experimental results and the comprehensive

analysis are presented in Section 4. The discussion about the proposed method is given in Section 5. Finally, conclusions are drawn in Section 6.

2. Methodology

The proposed end-to-end CNN–transformer feature-fused network (CTFuseNet) for crop-type semantic segmentation had three major components: a parallel structure of a CNN and a transformer to extract multi-scale local and global features in the encoder, a newly designed multi-scale feature-fusion module, and the FPNHead from the feature pyramid network that served as the decoder and output the crop-type segmentation results. The overall architecture of the proposed network is shown in Figure 1. The following subsections describe the proposed method in detail.

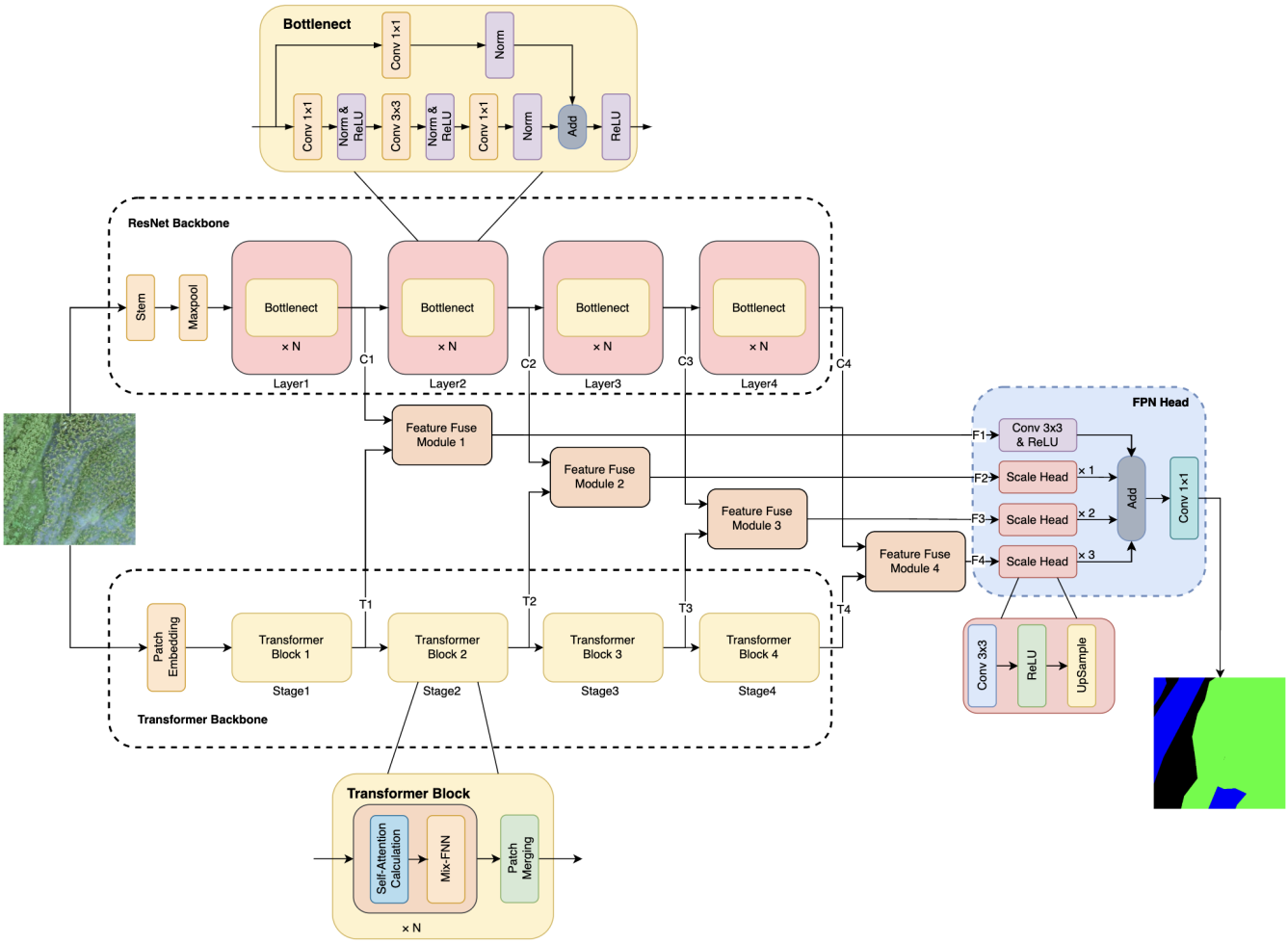

Figure 1. The framework of the proposed CTFuseNet for accurate crop-type segmentation.

2.1. Parallel Structure of CNN and Transformer for Local and Global Feature Extraction

As shown in the upper part of Figure 1, the proposed CTFuseNet consisted of a parallel structure of CNN and transformer branches. The CNN branch was designed to extract the local details of the remote sensing images. In this paper, ResNet [42], a CNN-based image classification network, was used as the feature extraction backbone in the CNN branch. It should be noted that the CNN branch enabled the integration of state-of-the-art CNNs to enhance performance. The ResNet was composed of several residual blocks called bottlenect, which contain a residual structure that utilized shortcut connections to transfer features from shallow layers to deep layers and resolved the degradation issue in deep CNN networks. According to the combinations of the bottleneck number in Layers

1–4, the ResNet could be divided into ResNet18, ResNet34, ResNet50, ResNet101 and ResNet152, for example, the number of bottlenecks in Layers 1–4 of ResNet50 was 3, 4, 6, and 3, respectively.

In the CNN branch of the proposed network in this paper, a remote sensing image of $H \times W \times 3$ (H and W represent the resolution of the input image, and the number of channels of RGB images was 3) was first processed into a $\frac{H}{2} \times \frac{W}{2} \times 64$ feature map by Stem (composed of conv and activation functions) and Maxpool, and the features were extracted at different levels through four layers. Each layer downsampled the input feature map and input it into several bottleneck structures, for example, the input of Layer 2 was $\frac{H}{4} \times \frac{W}{4} \times C$ and the output was $\frac{H}{8} \times \frac{W}{8} \times 2C$. Finally, four feature maps are obtained, and their resolutions and channel numbers were $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, $\frac{H}{32} \times \frac{W}{32}$, and C, $2C$, $4C$, $8C$, respectively. These four feature maps with different resolutions and a different number of channels supported the network in order to model local features at different scales.

As shown in the bottom part of Figure 1, the transformer branch in the CTFuseNet was designed to extract the global contextual information of remote sensing images. In this paper, SegFormer [28], a state-of-the-art transformer-based segmentation network, was adopted in the transformer branch. The encoder of SegFormer used a hierarchical transformer encoder without position encoding, and its encoder structure is shown in Figure 1. In the transformer branch with SegFormer, a remote sensing image of $H \times W \times 3$ was first divided into a series of 4×4 patches as an input feature sequence, which are called tokens, and then the transformer encoder was used to extract the global features of input tokens. Finally, four feature maps of different scales and with resolutions, such as $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, were obtained.

The transformer feature-extraction stage consisted of a patch-embedding module for dividing the input images into patches, and a transformer encoder consisting of four transformer blocks, which is shown in Figure 1. Each transformer block extracted the deep global features of the input features and reduced the resolution to $\frac{1}{2}$ of the original. Each transformer block contained several self-attention calculation modules and Mix-FNN modules, and a patch-merging module, which were detailed, as follows.

The formula of self-attention calculation module in the original ViT [27] proposed by Dosovitskiy et al. is shown as Equation (1). On this basis, the self-attention calculation module in SegFormer performed a reshape operation on K to reduce the computational complexity. Its formula is shown as Equations (2) and (3).

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_{head}}}\right)V \tag{1}$$

$$\widehat{K} = \text{Reshape}\left(\frac{N}{R}, C \cdot R\right)(K) \tag{2}$$

$$K_{new} = \text{Linear}(C \cdot R, C)\left(\widehat{K}\right) \tag{3}$$

where K is the input sequence to be reshaped, N is the length of the input feature sequence, C is the channel number of the feature sequence, and R is the ratio of the Reshape operation. In addition, $\text{Reshape}\left(\frac{N}{R}, C \cdot R\right)(K)$ refers to reshaping K to the shape of $\frac{N}{R} \times (C \cdot R)$, and $\text{Linear}(C \cdot R, C)\left(\widehat{K}\right)$ refers to linear transformation of K.

The mix-FFN module was used to replace the positional encoding in ViT to obtain the spatial information of the input image. The formula is as follows:

$$x_{out} = \text{MLP}(\text{GELU}(\text{Conv}(\text{MLP}(x_{in})))) + x_{in} \tag{4}$$

where x_{in} is the input features from self-attention calculation module, and GELU is the activation function. The Conv is a convolutional operation with a kernel size of 3×3.

The patch-merging module restored the feature sequence in the form of a one-dimensional token to a feature map with the shape of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, where i denotes the i-th layer.

2.2. Multi-Scale Feature-Fusion Module

The multi-scale feature-fusion module was designed to aggregate the local features from the CNN branch and the global features from the transformer branch. It efficiently and flexibly fused the features with different resolutions and channel numbers, and its structure is shown in Figure 2.

In the feature-fusion module, the features extracted by CNN and transformer (named C and T) were first passed through a convolutional network with a kernel size of 1×1, respectively, to adjust their dimensions. The concatenation was then performed to merge C and T. In this stage, the features merged by the concatenation were separated into a third dimension. The merged features were fed into a 1×1 convolution, followed by a normalization operation and an activation function. Finally, the features from the CNN and transformer branches were fully fused through a layer of 1×1 convolution. In addition, a residual operation was used to accelerate the convergence of the model.

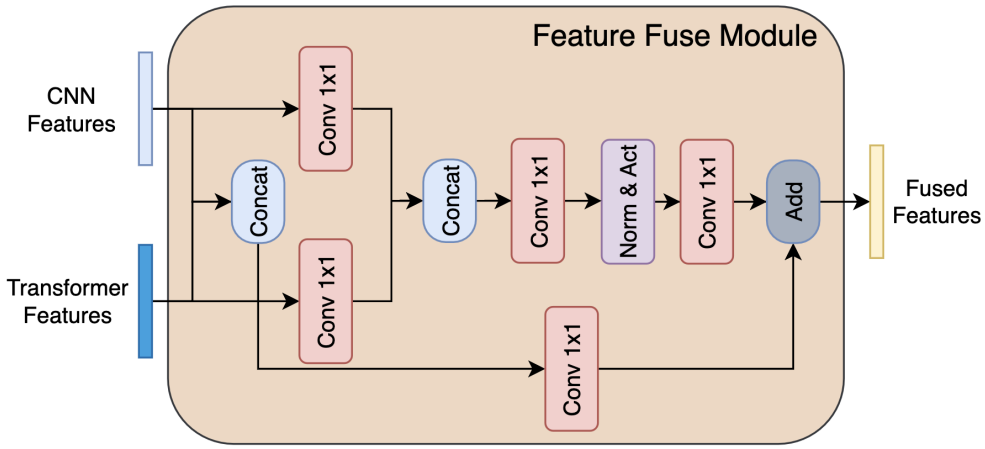

Figure 2. The structure of the proposed feature-fusion module.

2.3. FPNHead for Multi-Scale Feature Decoding

In the proposed CTFuseNet in Figure 1, four features extracted through CNN with the resolution of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$ (C1, C2, C3, and C4, respectively), and four features extracted from transformer with the same resolution (T1, T2, T3, and T4, respectively) were obtained. The features with the same resolution were fused separately, and finally four fusion features (F1, F2, F3, and F4, respectively) were obtained.

SegFormer originally designed a lightweight All-MLP decoder, i.e., MLPHead that is capable of extracting the global feature and contextual information for a transformer. However, the fused features in the CTFuseNet contained both information extracted from the transformer and the CNN. Meanwhile, the feature pyramid network [43] was first designed to parse the different levels of contextual information from the CNN with different receptive fields for object detection.Considering its effectiveness on decoding features of low-resolution, semantically strong features and high-resolution, semantically weak features, it had been adapted to a semantic segmentation network [44].

Therefore, in this paper, the FPNHead from the feature pyramid network served as the decoder for the local features and global fusion features in the proposed CTFuseNet. Its structure is shown on the right of Figure 1. For features F1, F2, and F3 with a spatial resolution less than $\frac{H}{4} \times \frac{W}{4}$, several scale heads were performed to restore their spatial resolution to $\frac{H}{4} \times \frac{W}{4}$. Subsequently, four feature maps with the same resolution were integrated in order to finally create the prediction map, which was generated through a 1×1 convolution.

2.4. Objective Function

The proposed CTFuseNet used the cross-entropy loss L_{CE} to optimize its performance, as shown in Equation (5), where N is the number of pixels in the image, C is the number of classes, $y_{i,c}$ is the ground-truth label for the i^{th} pixel and c^{th} class, and $p_{i,c}$ is the predicted probability for the i^{th} pixel and c^{th} class. The objective function was used to minimize the cross-entropy loss L_{CE}.

$$L_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{5}$$

3. Dataset and Experimental Settings

3.1. Study Area and Dataset

The benchmark dataset in this paper was the barley remote sensing dataset (Barley dataset) derived from the UAV remote sensing images of a barley cultivation site in Xingren City, Guizhou Province, provided by "2019 County Agricultural brain AI challenge" on the Alibaba Cloud Tianchi Platform [45]. Xingren City is located between 104°54′–105°34′ E and 25°18′–25°47′ N, in Qianxinan Buyei and Miao Autonomous Prefectures, Guizhou Province (as shown in Figure 3). The climate of Xingren City is suitable for the growth of corn, flue-cured tobacco, barley, and other crops.

The dataset contained 4 labeled RGB images, with sizes of $44,343 \times 3360$, $18,576 \times 68,363$, $44,647 \times 32,881$ and $55,128 \times 49,447$ pixels. The labels were given in the form of single channel images, and the value of each pixel corresponded to the category of crops in the RGB images. Specifically, the pixel value of "flue-cured tobacco" was 1, "corn" was 2, "barley" was 3, "building" was 4, and all other positions were regarded as "other" with pixel value of 0 (as shown in Figure 4). Table 1 shows the dataset information in detail.

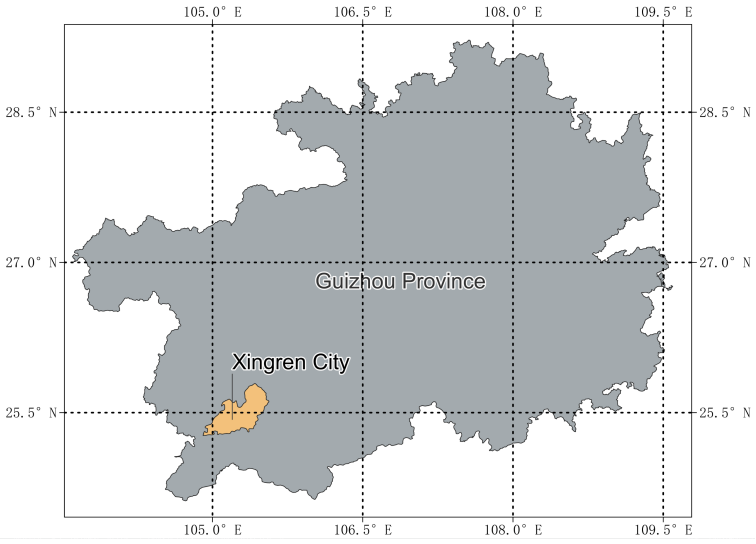

Figure 3. The study area in this paper.

Table 1. Detail of the benchmark Barley dataset.

Resolutions	Categories	Pixel Value
$44,343 \times 3360$ $18,576 \times 68,363$ $44,647 \times 32,881$ $55,128 \times 49,447$	flue-cured tobacco	1
	corn	2
	barley	3
	building	4
	background	0

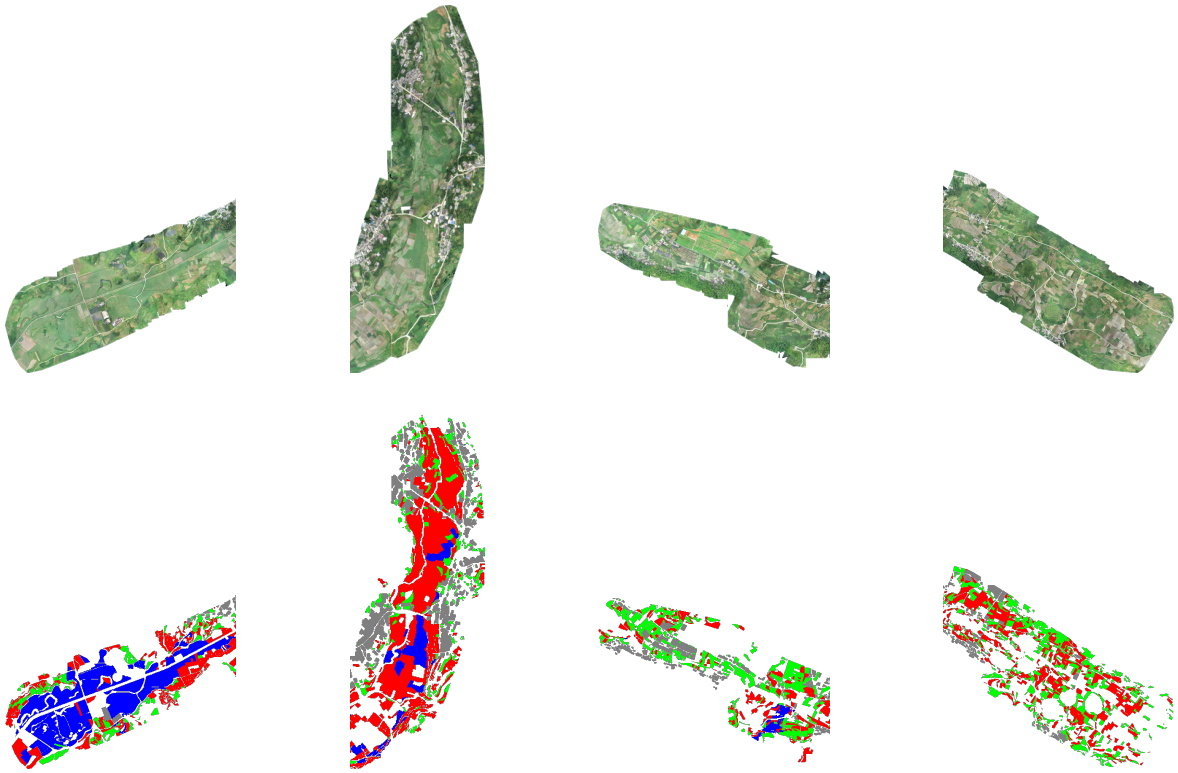

Figure 4. A panoramic view of the dataset. The top row is the RGB UAV remote sensing image, and the bottom row is the corresponding label.

3.2. Data Preprocessing

To facilitate the training and testing of the Barley dataset, we cropped the original 4 large-scale images with a sliding window without overlapping, and the crop size was 512×512. Subsequently, we sliced out the fully transparent images and obtained $13,037$ labeled 512×512 UAV remote sensing images. Finally, we divided the training and testing data by the ratio of $8:2$, that is, $10,429$ for training and 2608 for testing. Figure 5 shows examples of the cropped remote sensing images and their corresponding labels. To further enhance the diversity of training data, we performed data augmentation during model training, including random cropping, scaling, and flipping, as well as photometric distortions, such as changes in brightness, contrast, hue, saturation, and random light noise.

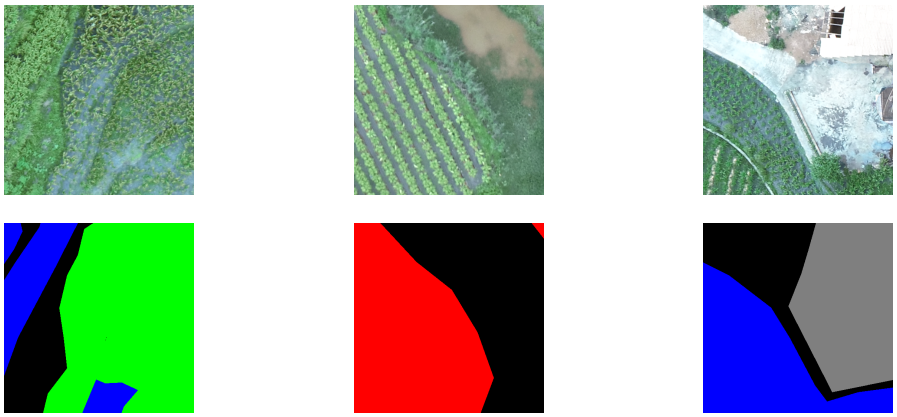

Figure 5. Cropped remote sensing images. The top row is the RGB image, and the bottom row is the corresponding label, in which black represents the background, red represents flue-cured tobacco, green represents corn, blue represents barley, and gray represents building.

3.3. Experimental Setup

The hardware platform included an Intel Xeon E5-2680 v4 CPU (https://www.intel.sg/content/www/xa/en/products/sku/91754/intel-xeon-processor-e52680-v4-35m-cache-2-40-ghz/specifications.html (accessed on 17 February 2023)), 64G of RAM, and an NVIDIA RTX3090 (https://www.nvidia.com/en-us/geforce/graphics-cards/30-series/rtx-3090-3090ti (accessed on 17 February 2023)) graphics card with 24G of video memory. The software platform was Ubuntu 20.04 with CUDA11.1, Python 3.7 and PyTorch 1.8.1 installed.

For the network training, we use the Adam optimizer with a batch size of 8 and an initial learning rate of 0.0001. The learning rate decayed by *gamma* every 10 epochs, and the value of the *gamma* was set to 0.5. At the same time, in order to avoid over-fitting, the dropout ratio was set to 0.1.

3.4. Evaluation Criteria

Mean intersection over union (mIoU) and pixel accuracy (PA), which are commonly used evaluation metrics for semantic segmentation [46,47], were adopted to evaluate the precision of the crop-type segmentation in our experiment.

Accordingly, $mIoU$ calculated the ratio of the intersection and the concatenation of the actual values and the predicted values, so it was able to capture both the precision and the recall in a single score, and its formula is shown below.

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}} \tag{6}$$

where k represents the number of crop-type segmentation categories, P_{ii} indicates the number of pixels correctly predicted, and P_{ij} represents the number of pixels whose actual value is i but predicted to be j.

In addition, PA is the ratio of the number of correctly predicted pixels to the total number of pixels, and its formula is shown below.

$$PA = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} P_{ij}} \tag{7}$$

3.5. Methods for Performance Comparison

In this paper, we compared the proposed CTFuseNet with seven advanced semantic segmentation methods listed below. For DANet, DeepLabV3+, and PSPNet, ResNet50 was utilized as the backbone. For SegFormer, to balance the computational complexity and precision, SegFormer-B3 was used as the backbone. For our proposed CTFuseNet, ResNet50 and SegFormer-B3 were used as the two-branched feature-extraction backbones.

- U-Net [16] uses a U-shaped structured network to propagate contextual information from low-resolution layers to high-resolution layers via upsampling.
- PSPNet [39] exploits the capability of global contextual information by different region-based contextual aggregation through a pyramid pooling module.
- DeepLabV3+ [40] is the latest version of the DeepLab series networks, which uses atrous convolution to expand the receptive field and depth-wise separable convolution to improve feature-extraction efficiency.
- DANet [48] models semantic information using attention mechanism in both spatial and channel dimensions.
- OCRNet [49] explicitly transforms the pixel classification problem into an object–region classification problem using the object–contextual-representations approach.
- SETR [21] is a transformer-based network that replaces CNN completely with self-attention modules.
- SegFormer [28] is a simple and efficient, yet powerful, semantic segmentation network that is based on a transformer.

4. Results and Analysis

4.1. Crop-Type Segmentation Performance

The segmentation accuracy of the proposed CTFuseNet and those of all the models used for comparison are shown in Table 2. Examples of the segmentation results are shown in Figure 6. Taking mIoU as the main indicator, the U-Net, based on the CNN architecture of an encoder–decoder, achieved a score of 78.73%. The DANet, DeepLabV3+, PSPNe, and OCRNet, which introduced attention mechanism and integrated multi-scale features, improved the segmentation accuracy of U-Net by 6.2%, 5.8%, 6.5%, and 6.1%, respectively, reaching 83.64%, 83.27%, 83.82%, and 83.56%, respectively. This showed that the attention mechanism and the multi-scale feature fusion significantly improved the crop-type segmentation performance.

After introducing the self-attention-based transformer into the crop-type segmentation, Table 2 shows the results of SETR and SegFormer. The encoder of SETR, a transformer-based network, completely relied on a self-attention mechanism. SETR improved the accuracy by 6%, as compared to U-Net, but was slightly lower than DANet, DeepLabV3+, and the other CNN-based models. SegFormer outperformed all CNN-based models and SETR in terms of crop-type segmentation accuracy. It used a transformer exclusively in the feature-extraction stage and fused features from different scales in the decoding stage, and the accuracy achieved was 84.47% for mIoU and 91.92% for PA.

Combining both advantages of a CNN and a transformer, the proposed CTFuseNet achieved the highest segmentation of 85.33% for mIoU and 92.46% for PA. Moreover, as shown in Figure 6, the CNN-based networks resulted in biases, especially at the edge of large ranges of crops in complex scenes (column 2 in Figure 6) while the transformer-based networks were not accurate for the segmentation of small ranges of crops (column 6 in Figure 6), and erroneous segmentation occurred in areas where crops were sparsely grown (column 7 in Figure 6). The proposed CTFuseNet obtained the best segmentation results.

Table 2. Performance comparison of crop-type segmentation networks. The best results are in bold.

Method	IoU(%) per Category					PA(%)	mIoU(%)
	Background	Flue-Cured Tobacco	Corn	Barley	Building		
U-Net	84.25	93.14	73.85	73.54	68.89	89.17	78.73
DANet	87.44	94.67	76.67	76.68	80.75	91.51	83.64
DeepLabV3+	87.36	94.45	76.39	78.13	80.01	91.36	83.27
PSPNet	87.86	94.37	77.43	78.77	80.65	91.69	83.82
OCRNet	87.78	93.98	76.71	77.44	81.89	91.48	83.56
SETR	87.07	94.09	76.46	77.31	79.96	91.14	82.98
SegFormer-B3	88.25	**95.39**	77.08	78.98	82.67	91.92	84.47
CTFuseNet (SegFormer-B3 + ResNet50)	**88.89**	95.16	**78.61**	**80.84**	**83.13**	**92.46**	**85.33**

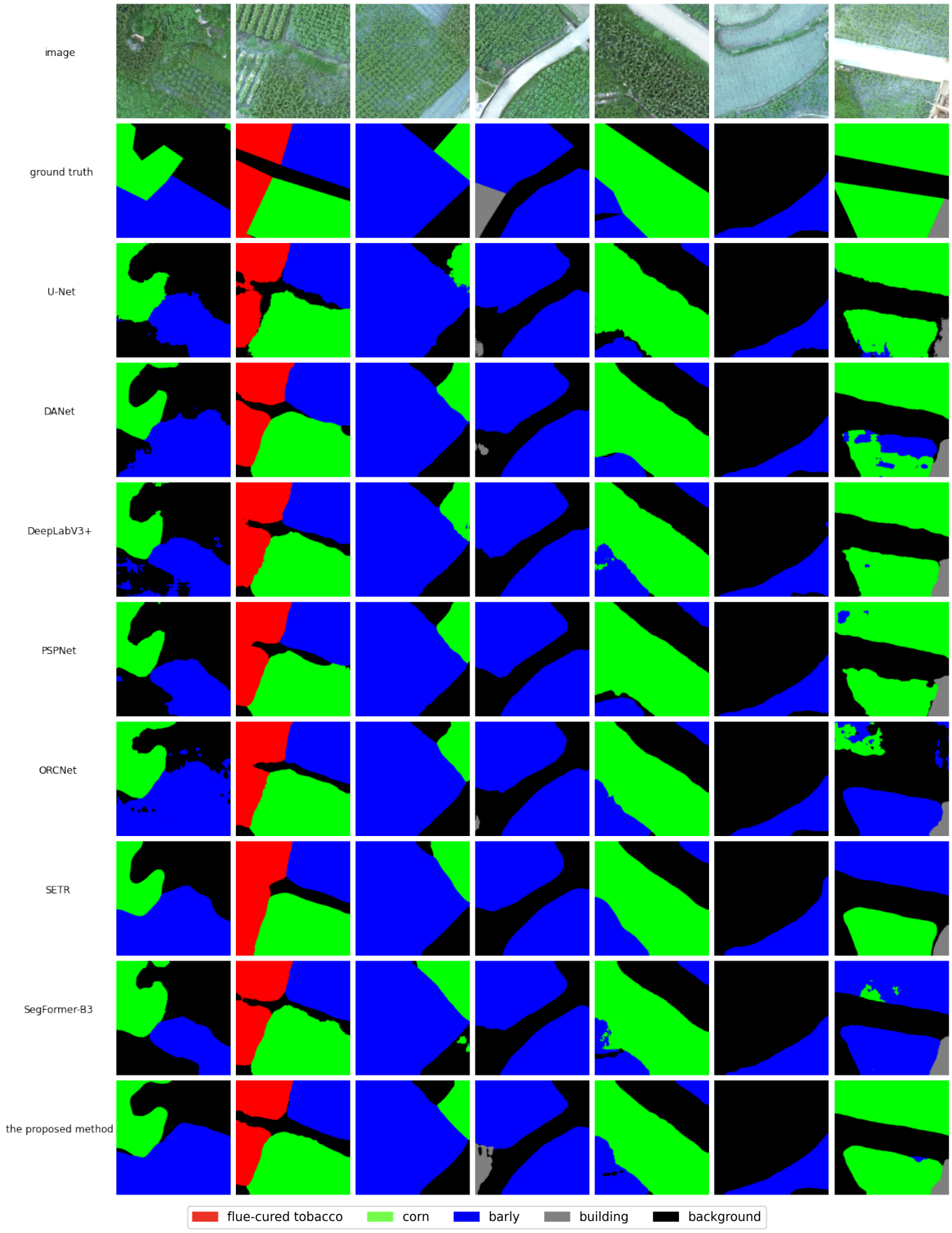

Figure 6. Examples and comparison of crop-type segmentation results.

4.2. Performance of Feature-Fusion Module

In addition to the crop-type segmentation performance of the proposed CTFuseNet, we analyzed the impact of our proposed feature-fusion module (CTFuse module). In particular, the dimensions of the features extracted by the CNN and the transformer were different, so we used Conv2d to adjust their dimensions before adding the features.

The segmentation accuracy is listed in Table 3. As compared to the mIoU of 84.47% and the PA of 91.92% for the SegFormer-B3 without the feature fusion, simply adding the respective features of a CNN and a transformer achieved an mIoU of 84.98% and a PA of 92.25%. It showed that the CNN features, indeed, improved the segmentation accuracy. By comparing the feature-fusion strategies, we found that simply adding them did not make full use of the CNN features. By using the proposed CTFuse Module, we obtained an mIoU of 85.33%, which was 1.02% higher than the original SegFormer-B3 and 0.41% higher than the adding strategy.

In addition, from the perspective of the model inference speed in terms of Img/s, i.e., the speed of the inference pictures per second, the CTFuse module was a lightweight module and had only a slight impact on the inference speed.

Table 3. Performance analysis of the feature fusion strategies. The best results are in bold.

Method	Fuse Strategy	Decoder	PA(%)	mIoU(%)	Img/s
SegFormer-B3	-	MLPHead	91.92	84.47	**18.0**
SegFormer-B3 + ResNet50	Conv2d & Add	FPNHead	92.25	84.98	14.3
CTFuseNet (SegFormer-B3 + ResNet50)	CTFuse Module	FPNHead	**92.46**	**85.33**	14.0

4.3. Performance of Decoders

In the semantic segmentation network based on the encoder–decoder structure, the role of decoder was to decode the features extracted in the encoder and restore them to the original spatial resolution of the image in order to output the prediction map. Decoders were one of the key factors affecting the segmentation accuracy. The original SegFormer designed a lightweight All-MLP decoder (MLPHead), which considered the characteristics of its hierarchical transformer encoder. However, in our proposed dual-branch CTFuseNet, both a transformer and a CNN were used to extract features. Thus, the FPNHead from the feature pyramid network served as the decoder for the fusion features in this work, and the results are shown in Table 4.

After fusing the CNN features, the mIoU increased from 84.47% to 85.2%, as compared to the original SegFormer with MLPHead for feature decoding, and the PA improved from 91.92% to 92.41%. The mIoU and PA continued to increase to 85.33% and 92.46%, respectively, with the FPNHead for multi-scale features. Due to the pyramid decoding structure, FPNHead considered both global and local features and was suitable for decoding the fused features from the CNN and the transformer. In addition, from the perspective of efficiency, the inference speed of the network with FPNHead was similar to that with MLPHead.

Table 4. Performance analysis of the decoders. The best results are in bold.

Method	Decoder	PA(%)	mIoU(%)	Img/s
SegFormer-B3	MLPHead	91.92	84.47	**18.0**
SegFormer-B3 + ResNet50	MLPHead	92.41	85.2	14.0
CCTFuseNet (SegFormer-B3 + ResNet50)	FPNHead	**92.46**	**85.33**	14.0

5. Discussion

5.1. Classification vs. Segmentation for Crop Mapping

Currently, the mainstream crop-mapping methods for remote sensing images include image classification and semantic segmentation. Since image classification can only predict the category information of the input image, a common practice for crop mapping based on image classification has been cropping the image into patches using a fixed-size sliding window, inputting them into an image classification network, and then mapping the category information based on the location of the patches in the whole image [50], as shown in Figure 7. In addition, since only a few pixels are contained in a single patch, many studies have attempted to improve the classification accuracy using adding additional spectral and temporal information from hyperspectral images [51], multispectral images [52], synthetic aperture radar (SAR) images [53,54], time series data [34,50], etc.

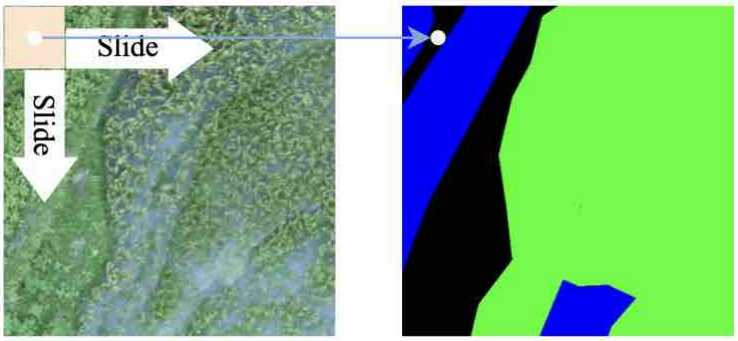

Figure 7. An illustration of the image classification-based method for crop mapping.

Conversely, semantic segmentation was able to process the entire image and output the classification information of each pixel in the entire image directly, which was more efficient for crop mapping. Semantic segmentation networks usually use the encoder–decoder structure, where the encoder is used to extract features and reduce the spatial dimension, and the decoder is used to parse features and recover the spatial dimension. Since the input of the segmentation network was the entire image containing all pixels, it enabled the semantic segmentation network to handle information over long distances and achieve high accuracy and efficiency.

Therefore, in this paper, we exploited the semantic-segmentation-based method for accurate and efficient crop mapping on large-scale RGB UAV images. In addition to enhancing the feature extraction and fusion by the proposed dual-branch CTFuseNet and CTFuse module, specifically for the importance of the decoder in the semantic segmentation networks, we adapted the decoder to the fused features. Our proposed method showed an accuracy of 85.33% in mIoU and 92.46% in PA, outperforming the state-of-the-art networks in comparison.

5.2. Local–Global Fusion Features of CTFuseNet

The results in Section 4 demonstrated the effectiveness of our proposed CTFuseNet by taking advantage of the extraction ability of the CNN in local features and that of the transformer in global features. Furthermore, we visualized heat maps by using gradient-weighted class activation mapping (Grad-CAM) [55] to better illustrate the features from the CNN and the transformer branches, as well as the fused results. Grad-CAM was able to generate heat maps for the gradients of the target classes flowing into the final convolutional layer and to highlight focal regions, and it was also applicable for the transformer-based networks [56].

In particular, the results of the single-branch CNN-based DeepLabV3+ (with ResNet50 as backbone), the single-branch transformer-based SegFormer (with SegFormer-B3 as backbone), and the proposed CTFuseNet (ResNet50 + SegFormer-B3) for different crops

were evaluated. For DeepLabV3+, we used the final bottleneck in its backbone as the target layer to compute Grad-CAM. For SegFormer, the LayerNorm layer in its last transformer encoder was used. For the proposed CTFuseNet, the final convolutional layer in the CTFuse module was used. By overlaying the heat map onto the remote sensing image, we observed the regions the model prioritized, as shown in Figures 8 and 9.

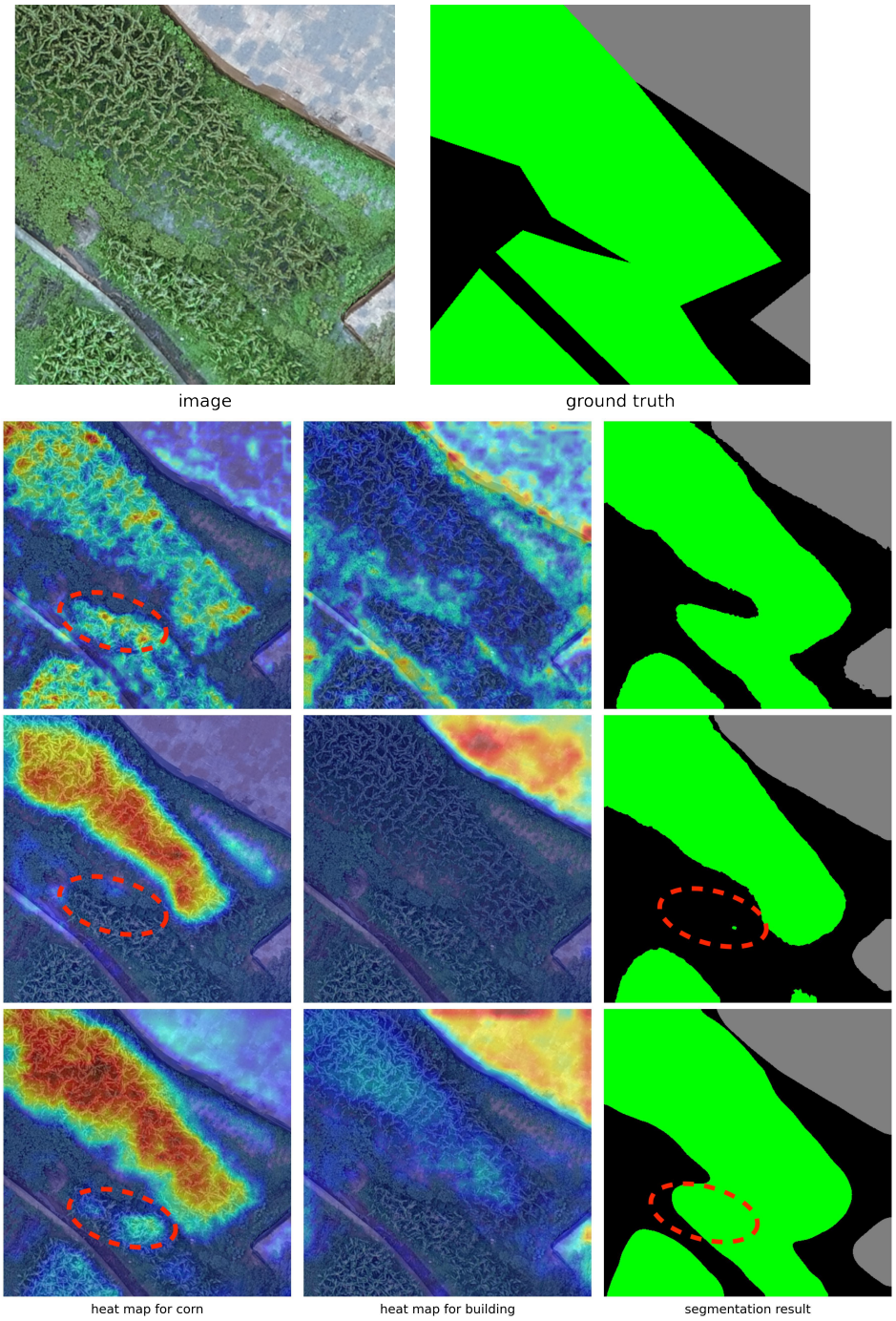

Figure 8. An example of heat map of corn and building. The rows from top to bottom are results from DeepLabV3+, SegFormer and the proposed CTFuseNet, respectively.

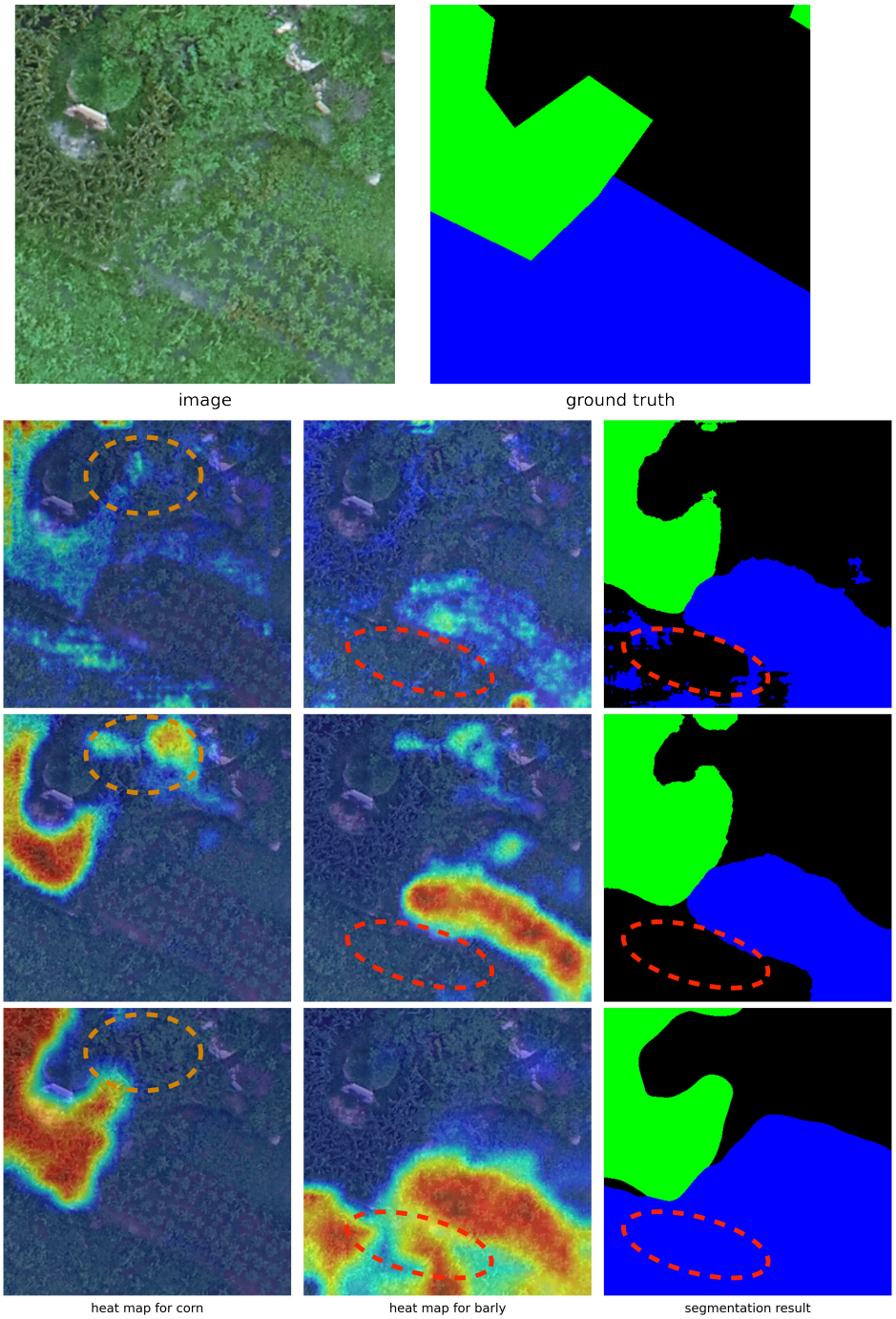

Figure 9. An example of heat map of corn and barley. The rows from top to bottom are results from DeepLabV3+, SegFormer and the proposed CTFuseNet, respectively.

The top and the second rows in Figure 8 show that the CNN-based DeepLabV3+ prioritized local, small-scale feature information, and the transformer-based SegFormer was more suited for extracting feature over long distances at large scales. However, for the crop-type segmentation task, the remote sensing images not only captured information about the crops but also about the underlying soil, so the local information, such as texture and color, were also important. Therefore, from the segmentation results, the transformer-based SegFormer was not accurate in segmenting the corn. In contrast, for our proposed dual-branch CTFuseNet, the local information obtained from the CNN branch compensated for the transformer branch, thus achieving better segmentation results.

Another example of a heat map for core and barley is presented in Figure 9. In the modeling of the barley features in the lower left corner of Figure 9, DeepLabV3+ identified only a small fraction of objects while SegFormer could not identify any. Due to the proposed CTFuseNet's ability to model both local features and long-range global features, it surpassed the single-branch DeepLabV3+ and SegFormer's ability to identify barley.

With the addition of the results mentioned in Section 4, our model segmented the barley type with an accuracy of 80.84% in IoU, outperforming DeepLabV3+'s 78.13% and SegFormer-B3's 78.98%. Similarly, it segmented the corn type with an IoU of 78.61%, as compared to DeepLabV3+'s 76.39% and SegFormer-B3's 77.08%, which improved by 2.9% and 2.0%, respectively.The above heat maps illustrate the features that contributed to the crop-type segmentation improvement of the proposed fusion network.

5.3. Fusion Strategies of CNN and transformer

Given the huge advantages transformers have over CNN for global feature extraction and the modeling of long-range semantic information, many researchers have started to integrate the merits of transformers in the field of vision, such as through ViT [27] for image classification, DETR [57] for object detection, SETR [21], and SegFormer [28] for semantic segmentation, and they have achieved the state-of-the-art performance. However, transformer-based networks often require cutting images into small patches, which causes loss of spatial information, while CNN-based networks are excellent for extractin local information. Therefore, a series of networks to combine CNNs and transformers have emerged, such as TransUNet [58], CoAtNet [59], CvT [60], Conformer [61], CMT [33], etc.

These networks usually combine CNN and Transformer in two ways. One uses a single-branch junction structure, where the features extracted by the CNN are then fed into the Transformer, or the features extracted by the Transformer are then fed into the CNN, with the goal of extracting both local features and global features in a single processing flow. The other is to use a dual-branch structure to extract the features using a CNN and a Transformer in different branches and then to fuse the results, which is exemplified in the feature-coupling unit (FCU) in Conformer [61].

We adopted the latter approach in this paper, i.e., we proposed a two-branch network that used a CNN and a transformer for feature extraction in two branches and then fused the features using the CTFuse module. This approach provided the ability to use existing advanced networks and their pre-trained weights to achieve satisfactory results with only minor fine-tuning, which mitigated having to train the model from the very beginning. This was beneficial to the field of remote sensing since labeled data is generally limited. In addition, the feature-extraction backbone networks on both branches could be readily replaced with lightweight networks, enabling easy switching during tasks that require high computational speeds.

5.4. Statistical Significance Analysis

In this study, we employed the McNemar test [62] to evaluate the statistical significance of the differences in accuracy between our proposed CTFuseNet method and the methods adopted for comparison [63,64]. The McNemar test compared the classification results of two methods on the same set of samples and recorded the classification results for each method in a 2×2 contingency table, as shown in Table 5. Where f_{11} is the proportion of samples classified correctly by both classification methods, f_{12} is the proportion classified correctly by classification method 1 while incorrectly by method 2, and f_{21} is the proportion classified incorrectly by method 1 while correctly by method 2, and f_{22} is the proportion that classified incorrectly by both methods.

Table 5. Contingency table for McNemar test between two classification results.

		Classification 2		
		Correct	**Incorrect**	$\sum$
Classification 1	Correct	f_{11}	f_{12}	$f_{11} + f_{12}$
	Incorrect	f_{21}	f_{22}	$f_{21} + f_{22}$
	$\sum$	$f_{11} + f_{21}$	$f_{12} + f_{22}$	1

For the comparison of different algorithms, the McNemar test statistic was then calculated by Formula (8). As it follows a Chi-squared distribution with one degree of freedom, the p-value was obtained from the chi-squared distribution in table [65] or estimated by statistical software [66]. To evaluate the statistical significance of the test, the p-value was compared to a predetermined significance level α, which was typically set at 0.05. If the p-value was lower than α, the null hypothesis that the two classification methods had equivalent performance was rejected. However, if the p-value was greater than α, there was not enough evidence to reject the null hypothesis.

$$\chi^2 = \frac{(f_{12} - f_{21})^2}{f_{12} + f_{21}} \tag{8}$$

We adopted the starting point U-Net and the transformer-based SegFormer with the closest performance as baselines and employed the McNemar test on the test set to evaluate the statistical significance of the proposed CTFuseNet against the baselines. As our model was a multi-class semantic segmentation model, we assessed its performance by transforming a pixel-by-pixel classification task, which resulted in a total of $512 \times 512 \times 2608$ samples.

The comparison between CTFuseNet and U-Net yielded a McNemar test statistic of 11.9949 and a p-value of 0.00053, where the p-value was much lower than the significance level 0.05, indicating a statistically significant improvement in performance by CTFuseNet, as compared to U-Net. Similarly, the comparison between CTFuseNet and SegFormer resulted in a McNemar test statistic of 22.7454 and a p-value of 1.85×10^{-6}, further highlighting the superior performance of CTFuseNet over SegFormer.

The statistical analysis demonstrated that the proposed CTFuseNet outperformed the benchmarks with statistical significance, providing evidence for the effectiveness of our approach in crop-type segmentation.

5.5. Model Transferability

Changing environmental and image scene in the diverse geographic, temporal, and sensor conditions may cause data distribution shifts [67,68]. Therefore, the deep learning models trained on a certain set of labeled images are generally restricted to a specific dataset [69] and may not yield satisfactory performance for unseen data when applied to different image acquisition conditions [70]. The model transferability across domains, e.g., geo-locations, time, and sensors, have been a significant challenge and major concern in the remote sensing community [71].

The proposed CTFusedNet in this paper integrated a parallel structure of CNN and transformer branches (i.e., ResNet and SegFormer, respectively) that maintained their independence. Current research has shown the generalization capability of ResNet [72,73] and Vision transformer [73] by mathematical analysis and experimental results. In addition, data augmentation is also an effective way to diversify the training dataset and improve the generalization of the DL-based method [69], and therefore, we performed data augmentation, e.g., random cropping, scaling, flipping, and photometric distortions.

However, the benchmark dataset had a limited spatio-temporal configuration, so the model generalization ability was evaluated by the testing dataset. For data from different geo-locations, time, or sensors, transfer learning is a general practice considered to foster model generalization and improve performance [70,74]. Moreover, efforts have been divide into the following aspects, i.e., weakly supervised learning, optimal model and

feature selection, semi-supervised domain adaptation, deep metric based methods for few-shot learning, and meta-learning with deep metric embedding [67]. The comprehensive assessment and promotion of model transferability across domains is a critical direction in our future work.

6. Conclusions

In this paper, considering that the existing CNN-based crop segmentation methods lack long-range contextual information, which limits its accuracy, we introduced a self-attention based transformer and provided a flexible, parallel method for coupling CNN and transformer architectures to obtain long-term feature dependencies as supplementary while preserving local details for accurate crop-type segmentation on UAV remote sensing imagery. Therefore, we proposed an end-to-end CNN–transformer feature-fused network (CTFuseNet) to fully aggregate and decode the multi-scale global and local features based on the transformer and CNN architectures.

The CTFuseNet provided a parallel global–local feature extraction structure based on the transformer and CNN architectures in the encoder for accurate crop-type segmentation. A new feature–fusion module, CTFuse, was designed to flexibly fuse the extracted multi-scale features from the CNN and the transformer. In addition, FPNHead from the feature pyramid network served as the decoder to adapt to the fused multi-scale features and further improve the accuracy. The experimental results showed that for crop-type segmentation, according to the benchmark dataset, we achieved the highest accuracy with an mIoU of 85.33% and a PA of 92.46%, which was an increase of 2.4% and 1.0%, respectively, as compared to the single-branch networks, i.e., the CNN-based ResNet50 and the transformer-based SegFormer-B3. It also outperformed the typical networks, including U-Net, PSPNet, DeepLabV3+, DANet, OCRNet, SETR, and SegFormer. It should be noted that, due to the high flexibility of the CTFuse module we developed, the proposed CTFuseNet provided a framework for combining more advanced transformer or CNN branches and to further enhance the crop-type segmentation capabilities.

Despite the performance improvements achieved by the proposed CTFuseNet, there are still challenges, especially for real-time crop-type segmentation tasks onboard UAVs. In the future, in addition to designing feature-fusion modules and decoders with better performance, we will prioritize lightweight crop-type segmentation networks through model compression technologies, such as model pruning and knowledge distillation, to reduce the model size and computation complexity, as well as to promote the inference efficiency. Furthermore, we will thoroughly assess and improve the model robustness and transferability on diverse datasets.

Author Contributions: Conceptualization, J.X. and J.L.; literature investigation and analysis, J.X., J.L. and D.C.; writing—original draft preparation, J.X. and J.L.; writing—review and editing, J.X., J.L., D.C., Q.X. and C.D.; visualization, J.X. and J.L.; supervision, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant No. 41901376, and in part by the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing KLIGIP-2022-B08.

Data Availability Statement: The Barley dataset can be obtained on https://tianchi.aliyun.com/dataset/74952 (accessed on 28 December 2022).

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

References

1. FAO. The Future of Food and Agriculture–Trends and Challenges. *Annu. Rep.* **2017**, *296*, 1–180.
2. Yi, Z.; Jia, L.; Chen, Q. Crop Classification Using Multi-Temporal Sentinel-2 Data in the Shiyang River Basin of China. *Remote Sens.* **2020**, *12*, 4052. [CrossRef]
3. Mulla, D.J. Twenty Five Years of Remote Sensing in Precision Agriculture: Key Advances and Remaining Knowledge Gaps. *Biosyst. Eng.* **2013**, *114*, 358–371. [CrossRef]

4. Liu, J.; Xiang, J.; Jin, Y.; Liu, R.; Yan, J.; Wang, L. Boost Precision Agriculture with Unmanned Aerial Vehicle Remote Sensing and Edge Intelligence: A Survey. *Remote Sens.* **2021**, *13*, 4387. [CrossRef]
5. Valente, J.; Doldersum, M.; Roers, C.; Kooistra, L. Detecting Rumex Obtusifolius Weed Plants In Grasslands from UAV RGB Imagery Using Deep Learning. *ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2019**, *IV-2/W5*, 179–185. [CrossRef]
6. Furuya, D.E.G.; Ma, L.; Pinheiro, M.M.F.; Gomes, F.D.G.; Goncalvez, W.N.; Marcato Junior, J.; Rodrigues, D.d.C.; Blassioli-Moraes, M.C.; Michereff, M.F.F.; Borges, M.; et al. Prediction of Insect-Herbivory-Damage and Insect-Type Attack in Maize Plants Using Hyperspectral Data. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102608. [CrossRef]
7. Abdulridha, J.; Batuman, O.; Ampatzidis, Y. UAV-Based Remote Sensing Technique to Detect Citrus Canker Disease Utilizing Hyperspectral Imaging and Machine Learning. *Remote Sens.* **2019**, *11*, 1373. [CrossRef]
8. Apolo-Apolo, O.E.; Martínez-Guanter, J.; Egea, G.; Raja, P.; Pérez-Ruiz, M. Deep Learning Techniques for Estimation of the Yield and Size of Citrus Fruits Using a UAV. *Eur. J. Agron.* **2020**, *115*, 126030. [CrossRef]
9. Feng, S.; Zhao, J.; Liu, T.; Zhang, H.; Zhang, Z.; Guo, X. Crop Type Identification and Mapping Using Machine Learning Algorithms and Sentinel-2 Time Series Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3295–3306. [CrossRef]
10. Useya, J.; Chen, S. Comparative Performance Evaluation of Pixel-Level and Decision-Level Data Fusion of Landsat 8 OLI, Landsat 7 ETM+ and Sentinel-2 MSI for Crop Ensemble Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4441–4451. [CrossRef]
11. Hariharan, S.; Mandal, D.; Tirodkar, S.; Kumar, V.; Bhattacharya, A.; Lopez-Sanchez, J.M. A Novel Phenology Based Feature Subset Selection Technique Using Random Forest for Multitemporal PolSAR Crop Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4244–4258. [CrossRef]
12. Zhao, J.; Zhong, Y.; Hu, X.; Wei, L.; Zhang, L. A Robust Spectral-Spatial Approach to Identifying Heterogeneous Crops Using Remote Sensing Imagery with High Spectral and Spatial Resolutions. *Remote Sens. Environ.* **2020**, *239*, 111605. [CrossRef]
13. Lei, L.; Wang, X.; Zhong, Y.; Zhao, H.; Hu, X.; Luo, C. DOCC: Deep One-Class Crop Classification via Positive and Unlabeled Learning for Multi-Modal Satellite Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102598. [CrossRef]
14. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
15. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv* **2016**, arXiv:1511.00561.
16. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
17. 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. *arXiv* **2017**, arXiv:1606.00915.
18. Yang, M.D.; Tseng, H.H.; Hsu, Y.C.; Tseng, W.C. Real-Time Crop Classification Using Edge Computing and Deep Learning. In Proceedings of the 2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 10–13 January 2020; pp. 1–4. [CrossRef]
19. Osco, L.P.; Nogueira, K.; Marques Ramos, A.P.; Faita Pinheiro, M.M.; Furuya, D.E.G.; Gonçalves, W.N.; de Castro Jorge, L.A.; Marcato Junior, J.; dos Santos, J.A. Semantic Segmentation of Citrus-Orchard Using Deep Neural Networks and Multispectral UAV-based Imagery. *Precis. Agric.* **2021**, *22*, 1171–1188. [CrossRef]
20. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
21. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. *arXiv* **2021**, arXiv:2012.15840.
22. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems, Proceedings of the Thirtieth Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016*; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29.
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19. [CrossRef]
24. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* **2019**, arXiv:1709.01507.
25. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global Second-Order Pooling Convolutional Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3024–3033.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
28. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems, Proceedings of the Conference on Neural Information Processing Systems, Virtual, 6–14 December 2021*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 12077–12090.

29. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
30. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
31. Xu, Z.; Zhang, W.; Zhang, T.; Li, J. HRCNet: High-Resolution Context Extraction Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 71. [CrossRef]
32. Wang, H.; Chen, X.; Zhang, T.; Xu, Z.; Li, J. CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1956. [CrossRef]
33. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. CMT: Convolutional Neural Networks Meet Vision Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12165–12175. [CrossRef]
34. Li, Z.; Chen, G.; Zhang, T. A CNN-Transformer Hybrid Approach for Crop Classification Using Multitemporal Multisensor Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 847–858. [CrossRef]
35. Li, Q.; Chen, Y.; Zeng, Y. Transformer with Transfer CNN for Remote-Sensing-Image Object Detection. *Remote Sens.* **2022**, *14*, 984. [CrossRef]
36. Li, S.; Guo, Q.; Li, A. Pan-Sharpening Based on CNN plus Pyramid Transformer by Using No-Reference Loss. *Remote Sens.* **2022**, *14*, 624. [CrossRef]
37. Liu, X.; Wu, Y.; Liang, W.; Cao, Y.; Li, M. High Resolution SAR Image Classification Using Global-Local Network Structure Based on Vision Transformer and CNN. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4505405. [CrossRef]
38. Huang, L.; Wang, F.; Zhang, Y.; Xu, Q. Fine-Grained Ship Classification by Combining CNN and Swin Transformer. *Remote Sens.* **2022**, *14*. [CrossRef]
39. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
40. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
41. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified Perceptual Parsing for Scene Understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
43. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
44. Kirillov, A.; Girshick, R.; He, K.; Dollar, P. Panoptic Feature Pyramid Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 6399–6408.
45. Tianchi. Barley Remote Sensing Dataset. 2020. Available online: https://tianchi.aliyun.com/dataset/74952 (accessed on 28 December 2022).
46. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
47. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene Parsing through ADE20K Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5122–5130. [CrossRef]
48. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. *arXiv* **2019**, arXiv:1809.02983.
49. Yuan, Y.; Chen, X.; Wang, J. Object-Contextual Representations for Semantic Segmentation. In *Lecture Notes in Computer Science, Proceedings of the 16th European Conference Computer Vision (ECCV 2020), Glasgow, UK, 23–28 August 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 173–190. [CrossRef]
50. Li, J.; Shen, Y.; Yang, C. An Adversarial Generative Network for Crop Classification from Remote Sensing Timeseries Images. *Remote Sens.* **2021**, *13*, 65. [CrossRef]
51. Zhong, Y.; Hu, X.; Luo, C.; Wang, X.; Zhao, J.; Zhang, L. WHU-Hi: UAV-borne Hyperspectral with High Spatial Resolution (H2) Benchmark Datasets and Classifier for Precise Crop Identification Based on Deep Convolutional Neural Network with CRF. *Remote Sens. Environ.* **2020**, *250*, 112012. [CrossRef]
52. Gogineni, R.; Chaturvedi, A.; B S, D.S. A Variational Pan-Sharpening Algorithm to Enhance the Spectral and Spatial Details. *Int. J. Image Data Fusion* **2021**, *12*, 242–264. [CrossRef]
53. Qu, Y.; Zhao, W.; Yuan, Z.; Chen, J. Crop Mapping from Sentinel-1 Polarimetric Time-Series with a Deep Neural Network. *Remote Sens.* **2020**, *12*, 2493. [CrossRef]
54. Shakya, A.; Biswas, M.; Pal, M. Fusion and Classification of Multi-Temporal SAR and Optical Imagery Using Convolutional Neural Network. *Int. J. Image Data Fusion* **2022**, *13*, 113–135. [CrossRef]

55. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
56. Gildenblat , J. PyTorch Library for CAM Methods, 2021. Available online: https://github.com/jacobgil/pytorch-grad-cam (accessed on 29 September 2022).
57. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Lecture Notes in Computer Science, Proceedings of the 16th European Conference on Computer Vision (ECCV 2020), Glasgow, UK, 23–28 August 2020;* Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229. [CrossRef]
58. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:cs/2102.04306. https://doi.org/10.48550/arXiv.2102.04306.
59. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. CoAtNet: Marrying Convolution and Attention for All Data Sizes. In Advances in *Neural Information Processing Systems, Proceedings of the Conference on Neural Information Processing Systems, Online Event, 6–14 December 2021*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 3965–3977.
60. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. CvT: Introducing Convolutions to Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 22–31.
61. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local Features Coupling Global Representations for Visual Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 367–376.
62. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef]
63. Foody, G.M. Thematic Map Comparison. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 627–633. . [CrossRef]
64. Crisóstomo de Castro Filho, H.; Abílio de Carvalho Júnior, O.; Ferreira de Carvalho, O.L.; Pozzobon de Bem, P.; dos Santos de Moura, R.; Olino de Albuquerque, A.; Rosa Silva, C.; Guimarães Ferreira, P.H.; Fontes Guimarães, R.; Trancoso Gomes, R.A. Rice Crop Detection Using LSTM, Bi-LSTM, and Machine Learning Models from Sentinel-1 Time Series. *Remote Sens.* **2020**, *12*, 2655. [CrossRef]
65. Greenwood, P.E.; Nikulin, M.S. *A Guide to Chi-Squared Testing*; John Wiley & Sons: Hoboken, NJ, USA, 1996; Volume 280.
66. Seabold, S.; Perktold, J. statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010.
67. Lunga, D.; Arndt, J.; Gerrand, J.; Stewart, R. ReSFlow: A Remote Sensing Imagery Data-Flow for Improved Model Generalization. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 10468–10483. [CrossRef]
68. Tong, X.Y.; Xia, G.S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote. Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
69. Zhang, E.; Liu, L.; Huang, L.; Ng, K.S. An automated, generalized, deep-learning-based method for delineating the calving fronts of Greenland glaciers from multi-sensor remote sensing imagery. *Remote. Sens. Environ.* **2021**, *254*, 112265. [CrossRef]
70. Qin, R.; Liu, T. A Review of Landcover Classification with Very-High Resolution Remotely Sensed Optical Images-Analysis Unit, Model Scalability and Transferability. *Remote. Sens.* **2022**, *14*. [CrossRef]
71. Xiong, Y.; Guo, S.; Chen, J.; Deng, X.; Sun, L.; Zheng, X.; Xu, W. Improved SRGAN for Remote Sensing Image Super-Resolution Across Locations and Sensors. *Remote. Sens.* **2020**, *12*. [CrossRef]
72. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019; pp. 3141–3149. [CrossRef]
73. He, F.; Liu, T.; Tao, D. Why ResNet Works? Residuals Generalize. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 5349–5362. [CrossRef] [PubMed]
74. Zhu, Q.; Zhang, Y.; Wang, L.; Zhong, Y.; Guan, Q.; Lu, X.; Zhang, L.; Li, D. A Global Context-aware and Batch-independent Network for road extraction from VHR satellite imagery. *ISPRS J. Photogramm. Remote. Sens.* **2021**, *175*, 353–365. [CrossRef]