sensors

MDPI

Article

Land Cover Classification of UAV Remote Sensing Based on Transformer–CNN Hybrid Architecture

Tingyu Lu [1,2], Luhe Wan [1,2,*], Shaoqun Qi [1,2] and Meixiang Gao [3]

[1] College of Geographical Sciences, Harbin Normal University, Harbin 150025, China; lutingyu@sina.com (T.L.); qsq-00@163.com (S.Q.)
[2] Heilongjiang Province Key Laboratory of Geographical Environment Monitoring and Spatial Information Service in Cold Regions, Harbin Normal University, Harbin 150025, China
[3] Department of Geography and Spatial Information Techniques, Ningbo University, Ningbo 315211, China; gmx102@163.com
* Correspondence: wanluhe@sina.com

Abstract: High-precision land cover maps of remote sensing images based on an intelligent extraction method are an important research field for many scholars. In recent years, deep learning represented by convolutional neural networks has been introduced into the field of land cover remote sensing mapping. In view of the problem that a convolution operation is good at extracting local features but has limitations in modeling long-distance dependence relationships, a semantic segmentation network, DE-UNet, with a dual encoder is proposed in this paper. The Swin Transformer and convolutional neural network are used to design the hybrid architecture. The Swin Transformer pays attention to multi-scale global features and learns local features through the convolutional neural network. Integrated features take into account both global and local context information. In the experiment, remote sensing images from UAVs were used to test three deep learning models including DE-UNet. DE-UNet achieved the highest classification accuracy, and the average overall accuracy was 0.28% and 4.81% higher than UNet and UNet++, respectively. It shows that the introduction of a Transformer enhances the model fitting ability.

Keywords: deep learning; transformer; CNN; semantic segmentation; UAV remote sensing

check for updates

Citation: Lu, T.; Wan, L.; Qi, S.; Gao, M. Land Cover Classification of UAV Remote Sensing Based on Transformer–CNN Hybrid Architecture. *Sensors* **2023**, *23*, 5288. https://doi.org/10.3390/s23115288

Academic Editor: David Lopez-Carr

Received: 6 April 2023
Revised: 29 May 2023
Accepted: 31 May 2023
Published: 2 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, remote sensing of unmanned aerial vehicles (UAVs), as a new way to acquire remote sensing data, has become a new technology to study earth characteristics and the properties of objects near the surface [1]. Different from traditional aircraft or satellite remote sensing platforms, UAVs can generate ultra-high-spatial-resolution digital images in a relatively low spatial range. In addition, compared with large remote sensing platforms, UAVs have faster response capability, shorter preparation time and lower operating costs [2]. UAV remote sensing is widely used in surveying and mapping, agriculture, environmental monitoring, resource surveys and military fields [3–6]. The rapid development of UAV remote sensing technology significantly reduces the cost of remote sensing data collection in a small range. A large number of UAV remote sensing public data sets that can be applied in computer vision tasks, such as target recognition, image classification and semantic segmentation, have been published one after another [7–9]. The convenience of UAV remote sensing provides opportunities for land cover mapping of different spatial ranges and different spatial resolutions. The use of lidar sensors can obtain vegetation height and structure information, which is useful in land cover classification [10,11].

At present, the deep learning algorithm is widely used in the field of remote sensing image classification. Zhu et al. used multi-spectral remote sensing image data from UAVs and artificial neural networks to classify vegetation on the riverbank and calculate the Simpson and Shannon–Wiener diversity index of the mountain vegetation community [12].

A Aeberli et al. combined a convolutional neural network with the local maximum filtering method to extract a single banana tree based on multi-temporal and multi-spectral remote sensing data from UAVs. The research results showed that the method had a high plant detection rate and was suitable for precision agriculture [13]. TD Camargo et al. used a lightweight deep residual convolutional neural network optimization algorithm and UAV remote sensing images to extract weeds in winter wheat planting areas, and the weed classification map drawn is rich in detail, which can provide data support for the precise spraying of pesticides [14]. L Hashemi-Beni et al. proposed a mapping method combining a full convolutional neural network (FCN) and regional growth (RG) when studying the extraction of flood areas from optical images [15]. To sum up, we can see that deep learning has good application prospects in remote sensing image classification tasks, which is due to deep learning's hierarchical representation of data and powerful feature extraction ability. The architecture of deep neural networks is constantly evolving, from the initial multi-layer perceptron to shallow neural networks, from deep neural networks to wider and deeper neural networks and from a single convolutional neural network to a complex neural network with convolutional and cyclic integration. In addition, regularization and optimization promote improvements and the design of deep learning in different aspects.

The advent of advanced structures and modules, such as residual connections, attention mechanisms and Transformers, further improves the accuracy of classification results and the robustness of models. The Transformer is a brand new neural network architecture originally used for natural language processing, characterized by its self-attentional mechanism and efficient computing power. With the continuous improvement in Transformer architecture, it has been used for image and video processing tasks in addition to natural language processing [16], such as Vis-Transformer (ViT) [17] and the end-to-end Vision-Language Cross-modal model (VL) [18] for computer vision. At present, there are two design paradigms for image classification models based on a Transformer. One is the combination of a convolutional neural network (CNN) and Transformer. The core idea is to capture local features using limited receptive fields of the CNN and then use the Transformer to learn long-distance dependency. Finally, the model preserves the global and local features to the greatest extent. The other method is a pure Transformer architecture, which applies a standard Transformer directly to the image classification task and then trains the image classification model in a supervised learning manner. Encouraged by the Vis-Transformer's performance, many researchers have applied the Transformer to remote sensing image classification tasks. Inspired by Vis-Transformer's performance, a number of researchers have applied Transformers to remote sensing image classification tasks, and many experiments have shown that the performance gains generated by a Transformer are obvious.

In order to solve the problem that continuous subsampling in deep convolutional neural network leads to a small receptive field, local features can only be extracted and long-distance dependence cannot be established, this paper introduces a visual Transformer model, integrates a convolutional neural network with a Transformer and designs a semantic segmentation network, DE-UNet, with a dual encoder. The feature of the network is that two branch networks focus on global features and local features. Finally, the two types of features are fused to provide richer semantic information for the model.

2. Study Areas and Dataset

2.1. Study Area

The research area is located in Acheng District, Harbin City, Heilongjiang Province, China (as shown in Figure 1). The UAV data collection range is about 1.72 square kilometers, and the altitude is between 368 m and 384 m. The main ground objects are corn, rice and water, and surrounding areas also contain shrubs and building land.

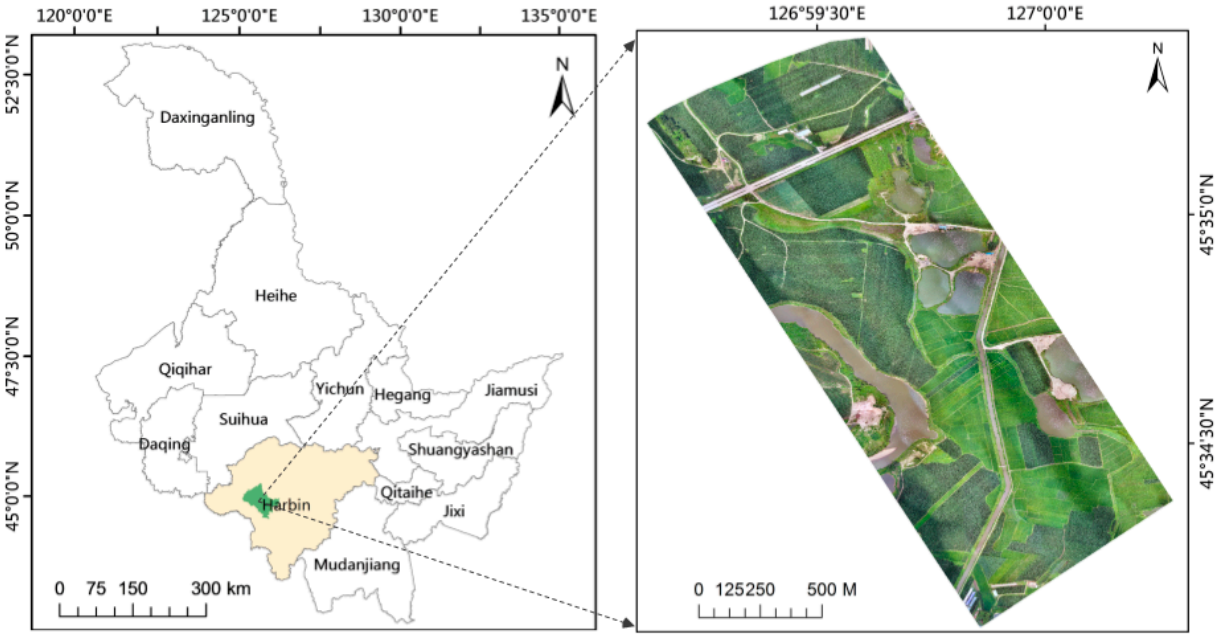


Figure 1. Location of the study area.

The multi-spectral images are taken by using a JR503 five-lens tilt photography gimlet mounted on a CW-10 composite wing UAV. The platform adopts four 40 mm tilt lenses and one 35 mm orthographic lens, with an image amplitude of 6000 × 4000 and up to 120 million pixels. It can output five channels of POS information at the same time to accurately record the exposure data of each camera. The spatial resolution is 0.04 m, the image is stored in JPEG format and the course and side overlaps are 75%. The shooting time was in the morning of 16 July 2022. The weather was clear and cloudless. The UAV flew 50 m high and took pictures with 60 m spacing and 90 m spacing. A snake-shaped route was selected, and the minimum circling radius was 120 m. The latitude of the plane is between 45°34′5.87″ N and 45°35′23.569″ N, and the longitude is between 126°59′7.569″ E–127°0′16.853″ E. The course setting and partial orthophoto data are shown in Figure 2.

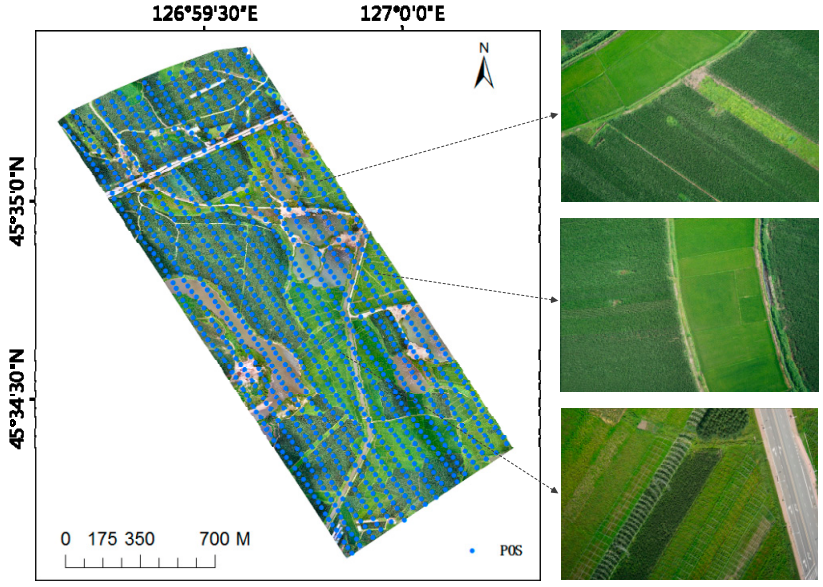


Figure 2. The flight path of the UAV, where each blue dot represents the location of the camera in space when the image was taken.

2.2. Construct Training Sets and Testing Sets

In the indoor operation phase, we used Agisoft Metashape and Pix4Dmapper to generate orthopoeia and digital surface model (DSM) from the drone data. The steps are as follows: prior to image processing, check the point position correspondence between images and POS (coordinates corresponding to the center point of each image) information, and delete the images taken automatically during takeoff and landing without corresponding POS points, as well as the photos taken at the end of the airline and during the adjustment of UAV flight height. Finally, 1800 ortho images were obtained, and a TIF image with spatial reference was output using the ortho Mosaic module of Pix4Dmapper. The elevation system was Dalian elevation system, and the projection coordinate system was WGS_1984_UTM_Zone_52N.

Corn, rice and water accounted for more than 90% of the whole research area. Therefore, this paper constructed training sets and test sets based on these three types of ground objects. The generated UAV images have a high spatial resolution. Combined with field mapping, high-precision digitization results of ground object labels can be produced with the image as the background. This part of work is completed under ArcGIS and ENVI platforms. The training set is used for model training. First, an area of about 72,000 square meters is selected from the research area as the training set (see Figure 3). The training data are a true-color image with a spatial resolution of 12,950 × 8686, and 2000 images of 224 × 224 size are randomly cut from the original image and label data. The training samples of this semantic segmentation task are constituted. The number of training samples of various ground objects is shown in Table 1. The size of the test set is 12,548 × 8060. In order to better test the generalization ability of the deep learning model, the test set and the training set do not overlap in space. Since the dimension of the input image of the model is 224 × 224 × 3, from the upper-left corner of the test set to the lower-right corner of the test set, according to the step of 112 pixels, the size of 224 × 224 images is clipped successively. After model prediction, the predicted images are spliced to restore the initial size of the test set. The number of samples in the test set is shown in Table 2.



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

Figure 3. (**a**). Original UAV image (**b**). Ground truth.

Table 1. Training samples.

No.	Class	Color	Sample Size
1	Corn		46,083,620
2	Rice		41,024,595
3	Water		6,814,415

Table 2. Test samples, where the resolution of the test set was 12,548 × 8060.

No.	Class	Color	Sample Size
1	Corn		5,874,169
2	Rice		64,240,779
3	Water		5,591,738

3. Methodology

In this paper, four classification algorithms are used to classify UAV image land cover, which are the traditional machine learning algorithm AdaBoost, two common semantic segmentation networks, UNet and UNet++, and our proposed DE-UNet.

3.1. Vision Transformer and Swin Transformer

Transformer network architecture is widely used in the field of natural language processing (NLP). Due to its excellent performance, Transformer network architecture has been introduced into the field of natural image and is currently a research hotspot in visual recognition tasks. Both Transformer and visual Transformer are based on the attention mechanism architecture, which is widely used in convolutional neural networks. The attention-enhanced convolutional network [19] connects the convolutional feature mapping with a set of feature mappings generated via self-attention to enhance the ability of the model to capture long-distance information. A large number of experiments show that the classification accuracy of the model is higher than that of the squeezing and excising network (SENet) when the number of parameters is similar, which verifies that the self-attention mechanism can enhance CNN. In the target detection task model DERT [20], only CNN is used to extract basic features, and Transformer is fully used in encoder and decoder parts, forming an encoder–decoder sequence prediction framework based on Transformer to predict all targets at one time. Vision Transformer (ViT) [16] completely gives up CNN and reuses Transformer architecture in natural language processing to solve the image problem. The input image is divided into several patches and then input into subsequent Transformer encoders in the form of a sequence. TNT (Transformer in Transformer) [21] model improves ViT by dividing patch in ViT into sub-patches and representing images with finer granularity. Research shows that TNT model has stronger learning ability and generalization ability. Swin Transformer [22] proposed hierarchical Transformer, which divides patch into windows and adopts local attention mechanism between windows. Meanwhile, in order to solve the interaction problem of information in windows between different layers, the window migration strategy is adopted to support cross-window connection between upper and lower layers. This layered design allows for great flexibility in modeling features at different scales.

Transformer adopts an encoder–decoder architecture and consists of three components: multi-head attention layer, feedforward neural network and layer standardization. Its structure is shown in Figure 4.

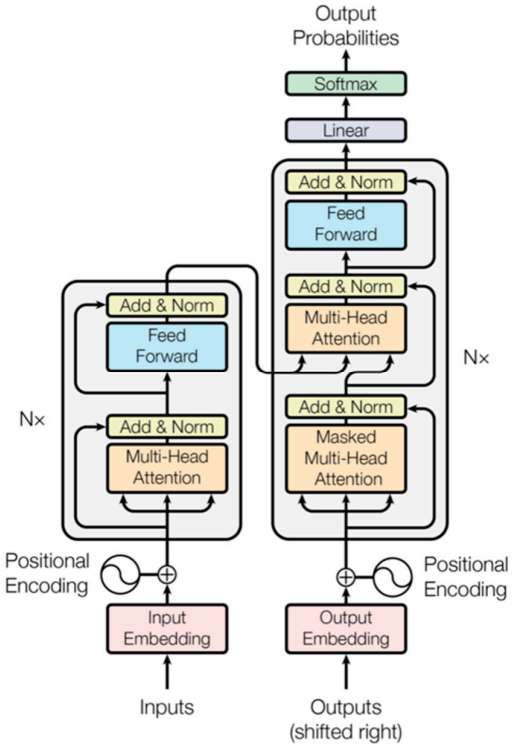


Figure 4. Structure of the original transformer [16].

In self-attention layer, three matrices, namely query vector Q (Query), key vector K (Key) and value vector V (Value), are converted from the average input vector. First, Q and K are dot products and usually the calculated result is divided by $\sqrt{d_k}$ to prevent the result from being too large, where d_k is the dimension of key vector K, and softmax is used to convert the result into probability distribution and then multiplied by matrix V to obtain the representation of weight summation, generating attention and vectors with greater probability to obtain extra attention. The calculation process is as follows:

$$S = Q \cdot K^T \tag{1}$$

$$S_n = \frac{S}{\sqrt{d_k}} \tag{2}$$

$$P = Softmax(S_n) \tag{3}$$

$$Z = V \cdot P \tag{4}$$

The complete calculation formula can be expressed as:

$$Attention(Q, K, V) = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{5}$$

In Transformer, the self-attention layer is improved by using multi-head self-attention mechanism. Single-head self-attention limits the model's ability to focus on multiple specific spatial locations. Multi-head attention uses different query matrices, key matrices and value matrices, which are randomly initialized and trained to project input vectors into different subspaces. Thus, the model can focus on multiple positions in different subspaces and generate further attention. The calculation process can be expressed by the following formula:

$$MultiHead(Q', K', V') = Concat(head_1, head_2 \ldots\ldots, head_h)W^O \tag{6}$$

$$head_i = Attention\left(QW_i^Q,\ KW_i^K,\ VW_i^V\right) \tag{7}$$

In the formula, the calculation process of $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, W_i^K and W_i^V is similar to that of W_i^V, where h is the amount of attention. Under the multi-head attention mechanism, each attention group maintains its own input and output weight matrix. The feedforward neural network is composed of two fully connected layers. The ReLU activation function is used after the first fully connected layer. There is a feedforward neural network in the encoder and decoder but they do not share parameters.

$$FN = max(0, X \cdot W_1 + b_1)W_2 + b_2 \tag{8}$$

There is a residual connection between the self-attention layer of the encoder and the feedforward neural network, after which another layer standardization operation is carried out. The calculation process is as follows:

$$output = LayerNorm(X + Attention(X)) \tag{9}$$

where X is the input of self-attention layer, and query vector Q, key vector K and value vector V are calculated from X.

The author of ViT believes that images can also be input into Transformer in the form of sequence–sequence just like word vectors in natural semantic processing, and the image patch sequence can be directly applied to Transformer to realize the classification of the whole image. The ViT model is composed of three parts: the image embedding layer, Transformer encoder and MLP. The embedding layer cuts the input image into patches of the same size. For a given image $H \times W \times C$, the number of images N is:

$$N = \frac{H}{p_1} \times \frac{W}{p_2} \tag{10}$$

where p_1 and p_2 are the height and width of the image, respectively. In practical application, $p_1 = p_2$ is usually set. The input of Transformer is a two-dimensional matrix. Therefore, it is necessary to first transform the three-dimensional image into two-dimensional input of (N, D). As shown in Figure 5, the embedding layer will transform the image into sequence data that can be processed by Transformer structure. Then, each patch is mapped to a one-dimensional vector, usually called Token, through linear mapping, and each Patch corresponds to a Token. Meanwhile, according to the position of patch in the input image, we add position information identified by the vector and finally input it into the Transformer encoder. Within the embedding layer, the process of mapping patch to one-dimensional vector is realized via the convolution layer. It can be seen that ViT belongs to the design paradigm combined with CNN and Transformer.



Figure 5. The embedding layer divides the image into equally sized parts.

After secondary feature extraction in Transformer, the final classification results can be obtained through multi-layer perceptron, which consists of a fully connected layer and activation function. Through the above analysis, we can understand that the image resolution processed using ViT is single and fixed. Even if global self-attention has global

modeling ability, it cannot obtain multi-dimensional features, and global self-attention significantly enhances the computational complexity. In order to extract multi-scale features and reduce the amount of computation, Swin Transformer model proposes hierarchical feature representation and window self-attention. The Swin Transformer architecture is shown in Figure 6.

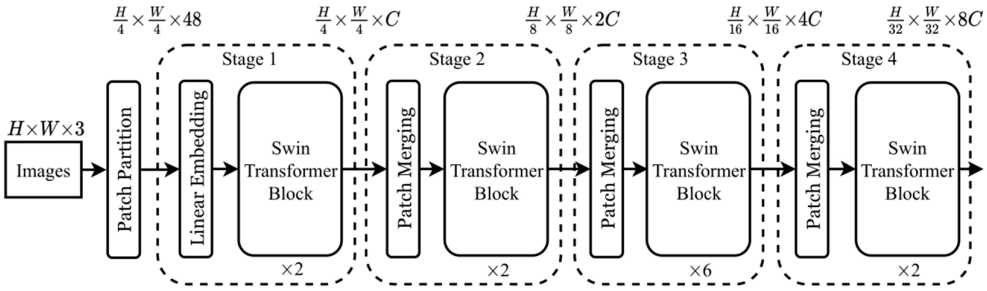


Figure 6. The structure of the Swin Transformer (image from [22]).

Swin Transformer is composed of 4 stages, each of which has similar functional units. In the first stage, images are first divided into patches of the same size and then input into Swin Transformer block. In the second stage to the fourth stage, the input patch is first merged, and four adjacent patches are merged into one patch. With the deepening of the network, the size of the feature map gradually decreases, and the perception range of each patch expands, so that multi-scale features can be extracted. This operation is similar to pooling in the convolutional neural network. Window self-attention is composed of two modules: standard window self-attention and moving window self-attention. The former limits the calculation of attention to a single window, thus reducing the amount of calculation, and the latter is set to solve the information interaction between windows.

3.2. Semantic Segmentation Network DE-UNet

Swin Transformer, with its excellent architecture design and efficient calculation, has become a popular backbone network. It has been widely used in computer vision tasks, especially in a series of visual downstream tasks, such as image segmentation, target detection and image classification. P Kyeong-Beom et al. [22] combined Swin Transformer with CNN to propose a semantic segmentation network, SwinE Net, for medical images. Xu X et al. [23] designed the spatial attention staggered cascade network framework SAIEC based on Swin Transformer, and they proved the effectiveness and feasibility of the model in remote sensing image target detection data set. Cui Zhang et al. [24] proposed SwinSUNet, a pure Transformer model with U-shaped structure, using Swin Transformer block. Experimental results show that SwinSUNet performs better than the traditional CNN method in the task of change detection. Inspired by these models, we combine Swin Transformer and UNet to propose a DE-UNet model with dual encoder, making Transformer's long-distance feature capture capability complementary to CNN's local feature extraction capability. Figure 7 shows the model architecture of DE-UNet.

On the far left of DE-UNet is the Swin-Transformer-based encoder, which consists of four stages, with two Swin Transformer blocks in the first, second and fourth stages. There are six Swin Transformer blocks in the third stage. Inside each Swin Transformer block are standard windows (W-Trans block) and moving windows (SW-Trans block). The middle part is UNet network encoder, which consists of two-dimensional convolution, activation function ReLU and maximum pooling. First, in Swin Transformer encoder, the input RGB image is divided into 4×4 patches, and each patch has a feature dimension of 48. After reshaping each patch into a vector, a linear embedding layer is applied to record the position information of each patch. After the first stage, we output the feature map $((H/4 \times W/4) \times 64)$, that is, the height and width are one quarter of the input image, and the number of features is 64. The features of the output are passed to the next second stage and are fused with the features extracted from the second convolution unit of the UNet

encoder and input to the next convolution unit. Similar operations also occur between the following stage and the convolutional unit. It needs to be noted that there is a Patch Merging module in the second to fourth stages, which downsamples input features, reduces resolution and adjusts the number of features, and its function is similar to the pooling operation in the convolutional neural network. As can be seen from Figure 7, the two encoders carried out feature fusion for a total of 4 times. With the increase in network depth, the image resolution is further reduced, and the number of features is constantly increased. After the first to fourth fusion, the output features are: $((H/4 \times W/4) \times 128)$, $((H/8 \times W/8) \times 256)$, $((H/16 \times W/16) \times 512)$ and $((H/32 \times W/32) \times 1024)$. On the right side of DE-UNet is the decoder of the model, which decodes the features extracted by the encoder. The image resolution is gradually restored to $H \times W$ through four upsampling operations (purple arrow in Figure 7). Finally, the activation function Softmax outputs the classification results. The design of dual encoder makes up for the shortcomings of convolutional neural network and makes full use of feature information of different levels, which helps to segment image more accurately. In the next part of the experiment setting, we apply the DE-UNet model to the semantic segmentation task based on the UAV remote sensing image with high spatial resolution and evaluate the performance of the model.

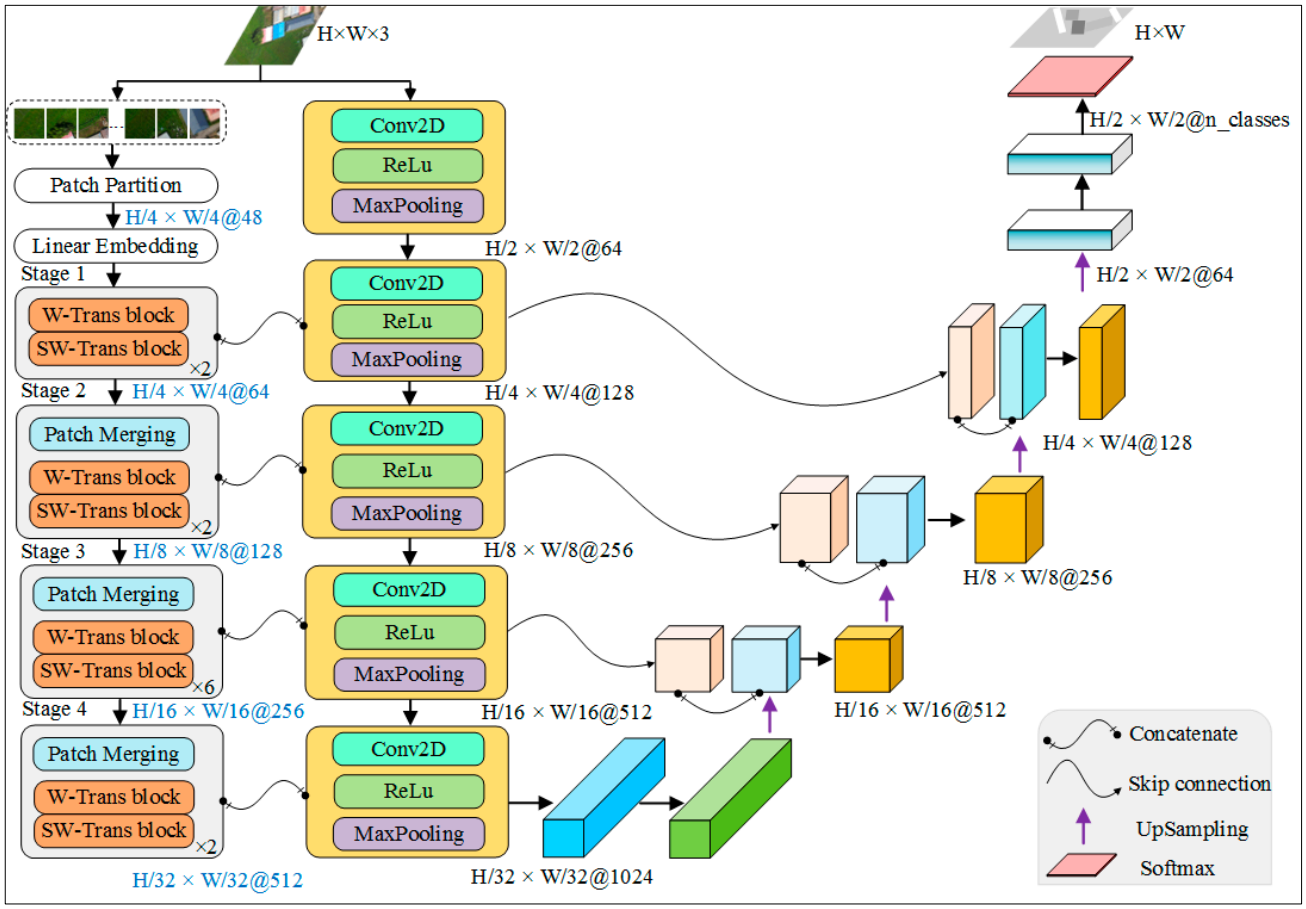


Figure 7. DE-UNet: a semantic segmentation network with dual encoder.

3.3. Baseline Classification Models

Two typical CNN semantic segmentation models and one machine learning model were developed for comparison.

3.3.1. AdaBoost

AdaBoost (Adaptive Boosting) is an ensemble learning algorithm proposed by Freund and Schapire [25], which controls deviation and variance and has an adaptive resampling

technique with enhanced prediction ability [26]. When a single linear classifier (weak classifier) cannot complete complex classification, multiple linear classifiers can be connected to form a strong classifier, which is the basic idea of AdaBoost. During operation, each sample of the training set is trained and assigned the same weight w, which constitutes vector D [27]. The weight distribution formula is shown as follows:

$$D_l = (w_{l1}, w_{l2}, \ldots, w_{lN}),\ w_{l1} = \frac{1}{N},\ i = 1, 2, \ldots, N \tag{11}$$

The weak classifier $G(x)$ is obtained by using training samples with weight distribution D_m for learning:

$$G_m(x) : \chi \to \{-1, 1\} \tag{12}$$

Then, in the iterative process, the sample weight is constantly adjusted to reduce the weight of samples with correct classification and increase the weight of samples with wrong classification, and the model is guided to learn again to continuously improve the prediction accuracy. AdaBoost assigns a weight value α to each weak classifier. These weights are obtained by calculating the error rate e of the weak classifier. The formula for calculating α is as follows:

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m} \tag{13}$$

The strong classifier is based on the weak classifier. The linear combination of the weak classifier and the final strong classifier can be expressed via the following formula:

$$f(x) = \sum_{m=1}^{M} \alpha_m G_m(x) \tag{14}$$

$$G(x) = sign(f(x)) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right) \tag{15}$$

AdaBoost has been proved to be an effective method for land cover classification [28–30]. In the experiment in this section, we use Python language and machine learning library scikit-learn to implement AdaBoost classifier. The base learning machine adopts decision tree. The maximum number of iterations was set to 400, and the learning rate was set to 0.8. The algorithm parameter used SAMME.R, which converges faster.

3.3.2. UNet

UNet is a deep neural network proposed by Ronneberger et al. [31] for medical image segmentation. It is considered to be a classical framework in semantic segmentation tasks, which is not only applicable to binary classification but also has excellent performance in multi-classification tasks, especially when there is imbalance between categories [32]. The core of UNet architecture is downsampling, upsampling and skip connection. The shallow network is used to extract the primary features of the image, while the deep network captures the advanced features, and skip connection realizes the splicing of the two levels of features. With its simple and efficient architecture design and good adaptability, it is widely used in many computer vision tasks, such as image classification and target detection. There are also a number of variations based on the UNet architecture. For example, Ozgun Cicek et al. extended UNet architecture to design 3D UNet, which can learn from sparsely annotated stereoscopic images and test the model's performance on complex 3D images without the use of pre-training network, achieving good segmentation results [33]. Wagner et al. used UNet to extract the forest types and tree species maps of 1600 square kilometers of tropical rain forest from the WorldView-3 satellite high-resolution image (0.3 m). The overall classification accuracy exceeded 95%, and the intersection ratio (IoU) reached 0.96 [34]. Shi et al. improved UNet and proposed CloudU-Net [35] for day–night satellite cloud image segmentation by introducing extended convolution and fully connected conditional random field (CRF). Experimental results show that the

segmentation effect of this network is better than UNet and FCN. Li et al. proposed MACU-Net using asymmetric convolution and multi-scale skip join, which achieved ideal classification accuracy in high-resolution remote sensing image segmentation [36].

3.3.3. UNet++

UNet++ [37] improved the UNet architecture by using more dense skip connections, so as to reduce the semantic loss of the feature map in the process of encoding and decoding and make full use of the multi-scale features of the image. The architecture of UNet++ is shown in Figure 8. The yellow and blue arrows represent the intensive connections added by the model. Meanwhile, in order to obtain more adequate training for the shallow network and solve the problems, such as the disappearance of training gradient and slow convergence rate, deep supervision (red line) is also added to the model.

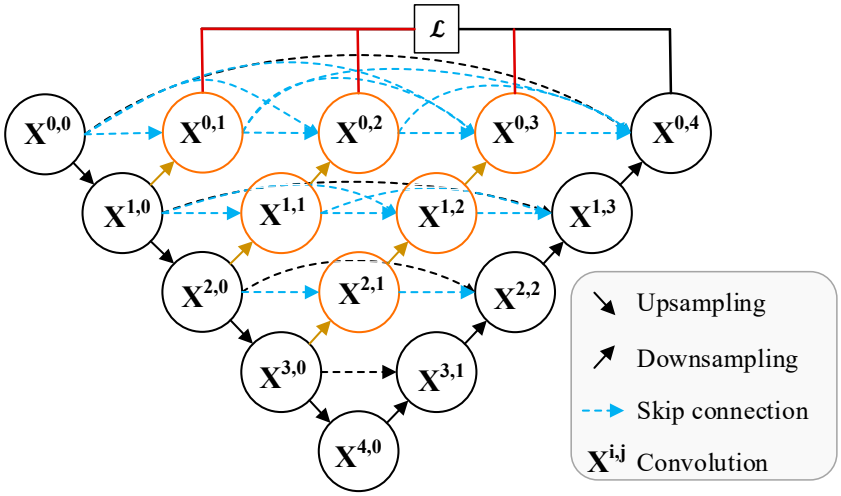


Figure 8. The architecture of UNet++.

Similar to UNet, there are also many variations of UNet++, and it is widely used in high-resolution remote sensing image classification and target detection tasks. Some experiments show that UNet++ with deep supervision mechanism has obvious performance gain compared with UNet [38–40].

4. Experimental Results and Discussion

4.1. Model Training

We used Python 3.7 and the deep learning framework Keras to construct the DE-UNet network. DE-UNet was compared with AdaBoost and another two deep learning models (UNet and UNet++). In order to fairly compare the impact of different methods on the classification results, the three deep learning models used the same hyperparameter settings during training.

The learning rate is 0.05, the batch size is set to 16, the weight attenuation is 0.001, the epoch is 150, Adam is used as the optimization function and the early stopping strategy is used in the entire training process. Training is stopped when the training accuracy does not increase for 30 consecutive rounds. The training process when the deep learning model achieves the highest classification accuracy is shown in Figure 9. As can be seen from Figure 9, the three deep learning models all converge to a higher accuracy but the convergence speed is slow. Among them, UNet has an early stop phenomenon, and no fitting occurred in all models.

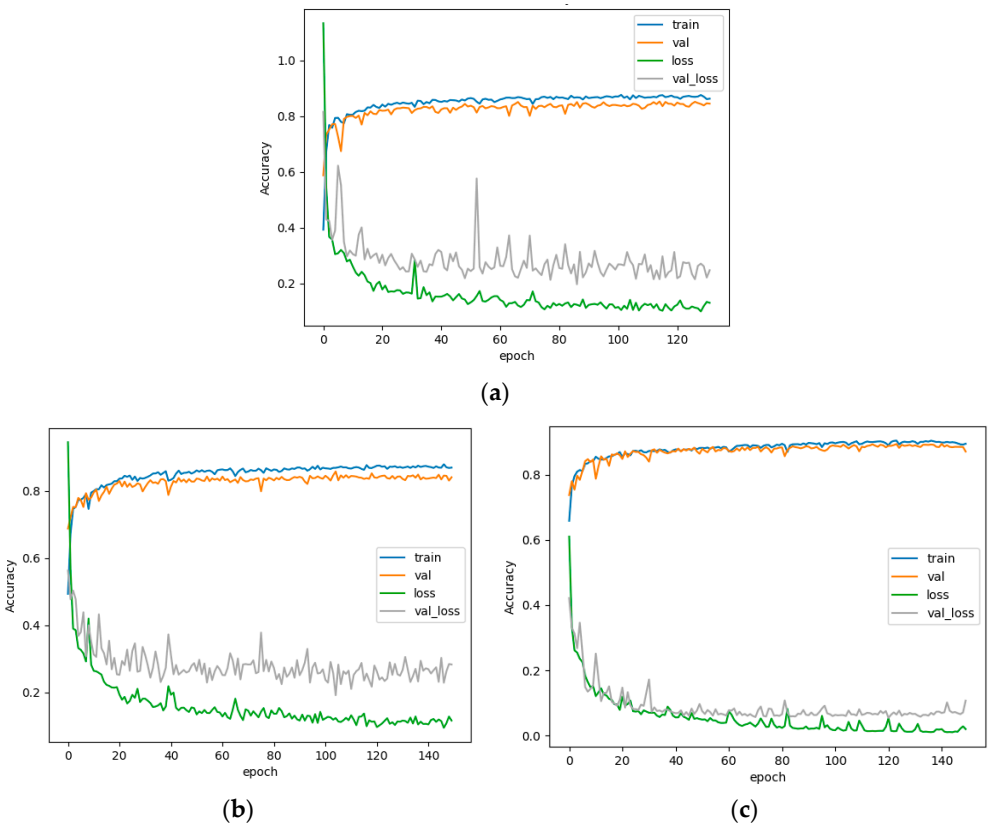


Figure 9. Loss and precision change curves of three deep learning algorithms during training. (**a**) UNet; (**b**) UNet++; (**c**) DE-UNet.

4.2. Classification Result and Evaluation

In order to better compare the performance of different classification algorithms, we trained each algorithm five times in the experiment and then calculated the average classification accuracy and standard deviation of various ground objects on the test set. Table 3 summarizes the average evaluation indexes obtained via different classification algorithms in five runs.

Table 3. Classification accuracy obtained by 4 different classification algorithms.

Class	AdaBoost Mean ± SD	UNet Mean ± SD	UNet++ Mean ± SD	DE-UNet Mean ± SD
Corn	48.69 ± 0.07%	94.32 ± 1.82%	96.02 ± 3.53%	96.07 ± 2.08%
Rice	85.83 ± 0.11%	96.92 ± 0.60%	98.36 ± 0.67%	98.46 ± 0.54%
Water	76.18 ± 0.24%	76.89 ± 3.10%	93.23 ± 6.56%	74.08 ± 1.17%
OA (%)	71.12 ± 0.05	94.23 ± 0.48	89.70 ± 1.80	94.51 ± 0.94
AA (%)	79.53 ± 0.03	90.20 ± 0.86	90.53 ± 1.86	88.74 ± 3.03
Kappa × 100	67.60 ± 0.02	89.08 ± 0.92	81.13 ± 3.60	89.74 ± 1.66

The statistical results in Table 3 show that the proposed DE-UNet model has the best performance, and the classification accuracy of the two crops and the overall classification accuracy are better than the other three algorithms. In five training sessions, the highest classification accuracy obtained using DE-UNet on corn and rice is 98.47% and 98.24%, respectively. The performance of UNet++ is slightly better than that of UNet. In terms of the classification accuracy of water bodies, UNet++ performs better but is not stable enough. The overall classification accuracy of the traditional machine learning algorithm, AdaBoost, is relatively low, especially in the recognition of corn and water bodies. The average classification accuracy of corn is only about 50% for five times. For different methods

achieving the highest overall classification accuracy, the corresponding classification map is shown in Figure 10.

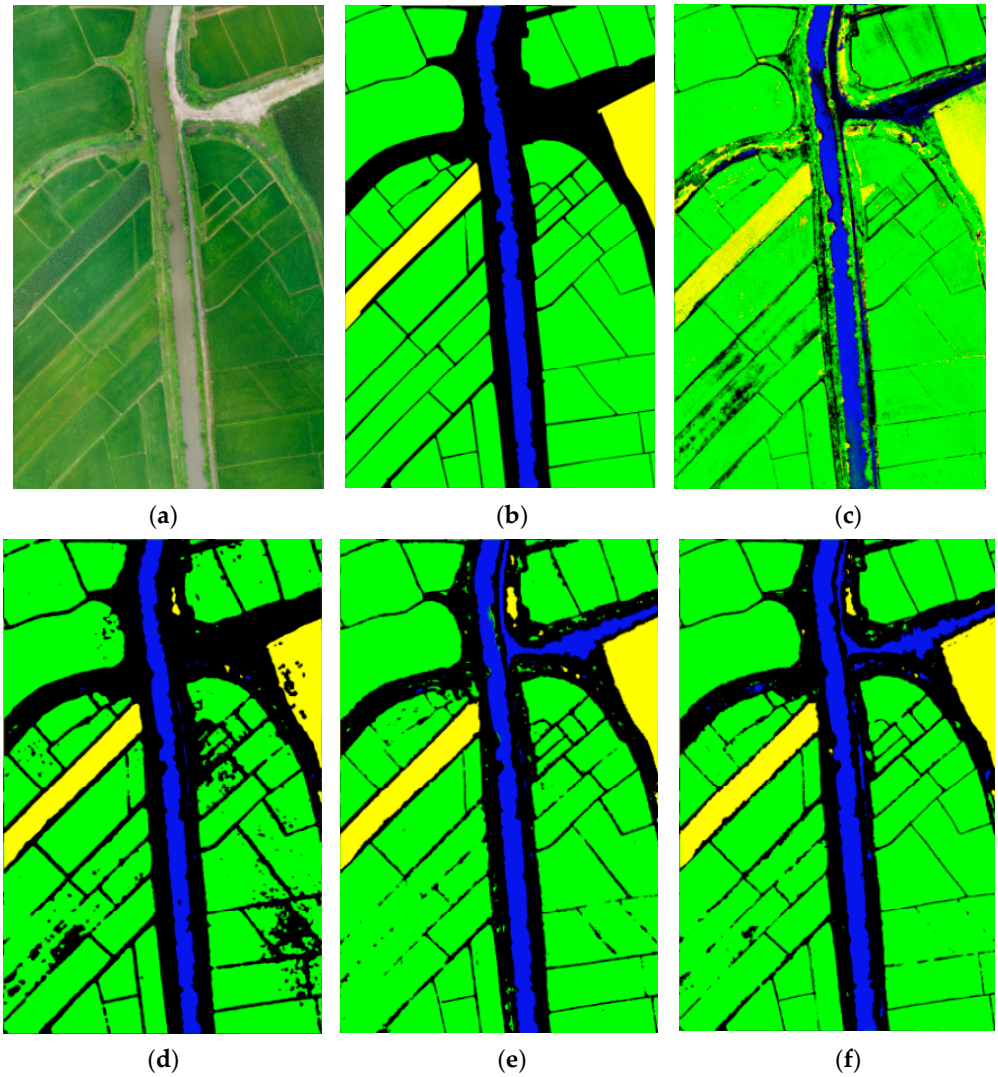


Figure 10. Classification results of 4 different classification algorithms (**a**). Original UAV image (**b**). Ground truth (**c**). AdaBoost (**d**). UNet (**e**). UNet++ (**f**). DE-UNet.

It can be seen from the classification map that there is a lot of noise in the classification results of AdaBoost, and the algorithm misclassification weeds on both sides of the river to rice and bushes on both sides of the road to corn. The classification effect of the three deep learning models on corn and rice is ideal, and the classification accuracy is over 90%. However, in the classification of water bodies, only UNet++ has a good performance. When marking ground objects, we take roads and bushes as the background, which poses a challenge to the model because there is a small amount of water on the road surface in this region. UNet++ and DE-UNet divided some waterlogged roads into water bodies, which affected the average classification accuracy and the overall classification accuracy.

By comparing the classification results of maize and rice, we found that the automatic extraction results of the two crops are highly consistent with the actual distribution, and the land boundaries are clear. However, there is obvious misclassification between water bodies and waterlogging roads and between water bodies and buildings, which is related to the fact that we did not label the roads and buildings in categories, so the model cannot effectively learn the features of these two types of ground objects.

After years of development, the control of a UAV aerial photography platform is becoming more and more simple, the data acquisition method is more flexible and efficient, the data quality is constantly improved and the cost of data acquisition is much lower than that of traditional aerial photography technology. Classification results based on UAV remote sensing images can be widely used in natural disaster assessments, planting area estimations and other applications.

4.3. Discussion

In this work, we proved that UAV images are suitable for land cover classification. This paper compared the traditional machine learning classification algorithm with the current popular advanced deep learning algorithm. The experimental results show that the traditional machine learning algorithm is not always good at dealing with the unbalanced classification problem. The deep learning classification method based on nonlinear input transformation (UNet, UNet++, DE-UNet) has better performance than the machine learning method (AdaBoost).

Among the three deep learning methods, DE-UNet with the dual encoder, proposed in this paper, is better than UNet and UNet++. As can be seen from Table 3 and Figure 10, the classification accuracy of both corn and rice is very high for the deep learning model, and the high misclassification rate between water body and background is the main issue affecting the overall classification accuracy of the model. In addition, as expected, the addition of a Swin Transformer significantly improves the classification accuracy for corn and rice. Although the deep learning method has obvious performance advantages in land cover classification using UAV remote sensing, there are still two limiting factors in the practical operation. First, the imbalance of samples resulted in low training accuracy and prediction accuracy of the model, which was specifically reflected in the obvious difference between the classification accuracy of the water body and the two crops. Second, there are many parameters in the deep learning model, so training a deep learning model requires a large computational cost. In the experiment, GPU (8G) acceleration is used in this study, and the training time of UNet, UNet++ and DE-UNet is about 5, 11 and 16 h, respectively, while the training time of AdaBoost is 2 h. How to balance the efficiency and precision of the model is one of the research directions in our follow-up work.

In this study, in addition to the proposed DE-UNet, we also chose UNet and UNet++, two commonly used deep learning supervised classification algorithms, which are widely used in the field of remote sensing image classification and also achieved satisfactory accuracy in this classification task. All these indicate that in the field of remote sensing image classification, the technology of intelligent extraction of image features using convolutional neural networks or vision Transformers is very important. The combination of the two is a widely used architecture design nowadays. It can even be predicted that a new computer vision application paradigm based on a vision Transformer will be formed in the future.

5. Conclusions

Deep learning has been widely used in the field of remote sensing. This paper studies the application potential of UAV remote sensing data in land cover maps and evaluates the latest deep learning supervised classification algorithm. We propose a semantic segmentation network that integrates a convolutional neural network and Swin Transformer, and we evaluate its performance on training sets and test sets. The experimental results show that compared with traditional machine learning methods, the three deep learning methods have consistent improvements in classification accuracy, the proposed DE-UNet method performs the best and the hybrid architecture of the Swin Tranformer fusion convolutional neural network has more powerful feature extraction ability. It was observed that, using the DE-UNet method, the average classification accuracy reached 96.07% for corn and 98.46% for rice. Meanwhile, the other two deep learning methods also achieved high classification accuracy.

It is worth mentioning that the experiment was conducted in July and the data were single-phase UAV images with high vegetation coverage. However, vegetation at different growth stages is different in spectral curves, spatial textures and other features, which challenges the generalization ability of the model. When the growth stage of vegetation changes, the model may not be able to give satisfactory results. A common method is to solve this problem through multi-temporal remote sensing, which is what we will pay attention to in future work.

Due to cloud cover, shadow generated by tall ground objects and random noise generated in the process of image acquisition, the quality of UAV images will be affected, which poses a great challenge to the classification algorithm, thus affecting the classification results. For oblique photogrammetry, in addition to obtaining visible light images, point cloud data of ground objects can also be generated. The data fusion of point cloud and visible images to generate new 3D scenes and classification based on a deep learning algorithm is the main direction of our future research.

Author Contributions: Investigation, writing—original, draft preparation, methodology, T.L.; conceptualization, formal analysis, review and editing, L.W.; investigation, data curation, S.Q.; supervision, funding acquisition, M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China, grant number 42271051.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: The study didn't involve humans.

Data Availability Statement: All data sets used and produced for the purposes of this study can be requested from the corresponding author.

Acknowledgments: We would like to thank the reviewers for their useful and constructive comments. We believe the manuscript is easier to read and follow thanks to their input.

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

References

1. Cao, H.; Gu, X.; Sun, Y.; Gao, H.; Tao, Z.; Shi, S. Comparing, validating and improving the performance of reflectance obtention method for UAV-Remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102391. [CrossRef]
2. Kelcey, J.; Lucieer, A. Sensor Correction of a 6-Band Multispectral Imaging Sensor for UAV Remote Sensing. *Remote Sens.* **2012**, *4*, 1462–1493. [CrossRef]
3. Yang, X.; Zan, M.; Maimaiti, M. Estimation of above ground biomass of Populus euphratica forest using UAV and satellite remote sensing. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 7.
4. Haala, N.; Kölle, M.; Cramer, M.; Laupheimer, D.; Mandlburger, G.; Glira, P. Hybrid Georeferencing, Enhancement and Classification of Ultra-High Resolution Uav LIDAR and Image Point Clouds for Monitoring Applications. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *5*, 2. [CrossRef]
5. Kuhn, J.; Casas-Mulet, R.; Pander, J.; Geist, J. Assessing Stream Thermal Heterogeneity and Cold-Water Patches from UAV-Based Imagery: A Matter of Classification Methods and Metrics. *Remote Sens.* **2021**, *13*, 1379. [CrossRef]
6. Bian, R.; Nian, Y.; Gou, X.; He, Z.; Tan, X. Analysis of Forest Canopy Height based on UAV LiDAR: A Case Study of Picea crassifolia in the East and Central of the Qilian Mountains. *Remote Sens. Technol. Appl.* **2021**, *36*, 10.
7. Calvin, H.; Zhe, X.; Salah, S. Feature Learning Based Approach for Weed Classification Using High Resolution Aerial Images from a Digital Camera Mounted on a UAV. *Remote Sens.* **2014**, *6*, 12037–12054.
8. Liang, Z.; Wang, J.; Xiao, G.; Zeng, L. FAANet: Feature-aligned attention network for real-time multiple object tracking in UAV videos. *Chin. Opt. Lett.* **2022**, *20*, 081101. [CrossRef]
9. Ye, L.; Vosselman, G.; Xia, G.S.; Yilmaz, A.; Yang, M.Y. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119.
10. Puliti, S.; Dash, J.P.; Watt, M.S.; Breidenbach, J.; Pearse, G.D. A comparison of UAV laser scanning, photogrammetry and airborne laser scanning for precision inventory of small-forest properties. *Forestry* **2019**, *93*, 150–162. [CrossRef]
11. Al-Najjar, H.A.H.; Kalantar, B.; Pradhan, B.; Saeidi, V.; Halin, A.A.; Ueda, N.; Mansor, S. Land cover classification from fused DSM and UAV images using convolutional neural networks. *Remote Sens.* **2019**, *11*, 1461. [CrossRef]

12. Zhu, H.; Huang, Y.; Li, Y.; Yu, F.; Tu, T.; Wang, W.; Zang, Y.; Li, J.; Luo, Y. Diversity of Plant Community in Flood Land of Henan Section of the Lower Yellow River based on Unmanned Aerial Vehicle Remote Sensing. *Wetl. Sci.* **2021**, *19*, 17–26.
13. Aeberli, A.; Johansen, K.; Robson, A.; Lamb, D.W.; Phinn, S. Detection of Banana Plants Using Multi-Temporal Multispectral UAV Imagery. *Remote Sens.* **2021**, *13*, 2123. [CrossRef]
14. de Camargo, T.; Schirrmann, M.; Landwehr, N.; Dammer, K.-H.; Pflanz, M. Optimized Deep Learning Model as a Basis for Fast UAV Mapping of Weed Species in Winter Wheat Crops. *Remote Sens.* **2021**, *13*, 1704. [CrossRef]
15. Hashemi-Beni, L.; Gebrehiwot, A. Flood Extent Mapping: An Integrated Method using Deep Learning and Region Growing Using UAV Optical Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *1*, 99. [CrossRef]
16. 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. [CrossRef]
17. Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; Fu, J. *Seeing Out of the Box: End-to-End Pre-Training for Vision-Language Representation Learning*; IEEE: Toulouse, France, 2021; pp. 12971–12980.
18. Chen, H.; Zhang, T.; Chen, R. Image classification method based on lightweight convolutional Transformer and its application in remote sensing image classification. *J. Electron. Inf. Technol.* **2022**, *44*, 1–9.
19. Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q.V. Attention Augmented Convolutional Networks. *arXiv* **2019**, arXiv:1904.09925. [CrossRef]
20. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
21. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. *arXiv* **2021**, arXiv:2103.00112. Available online: https://arxiv.org/abs/2103.00112 (accessed on 5 April 2023).
22. Kyeong-Beom, P.; Yeol, L.J. SwinE-Net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer. *J. Comput. Des. Eng.* **2022**, *2*, 616–632.
23. Xu, X.; Feng, Z.; Cao, C.; Li, M.; Wu, J.; Wu, Z.; Shang, Y.; Ye, S. An Improved Swin Transformer-Based Model for Remote Sensing Object Detection and Instance Segmentation. *Remote Sens.* **2021**, *13*, 4779. [CrossRef]
24. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5224713. [CrossRef]
25. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
26. Pham, B.T.; Tien Bui, D.; Prakash, I.; Dholakia, M.B. Hybrid Integration of Multilayer Perceptron Neural Networks and Machine Learning Ensembles for Landslide Susceptibility Assessment at Himalayan Area (India) Using GIS. *CATENA* **2017**, *149*, 52–63. [CrossRef]
27. Valdez, M.C.; Chang, K.-T.; Chen, C.-F.; Chiang, S.-H.; Santos, J.L. Modelling the Spatial Variability of Wildfire Susceptibility in Honduras Using Remote Sensing and Geographical Information Systems. *Geomat. Nat. Hazards Risk* **2017**, *8*, 876–892. [CrossRef]
28. Dou, P.; Chen, Y.; Yue, H. Remote-sensing imagery classification using multiple classification algorithm-based AdaBoost. *Int. J. Remote Sens.* **2018**, *39*, 619–639. [CrossRef]
29. Bigdeli, B.; Pahlavani, P.; Amirkolaee, H.A. An ensemble deep learning method as data fusion system for remote sensing multisensor classification. *Appl. Soft Comput.* **2021**, *110*, 107563. [CrossRef]
30. Zhu, S.; Deng, J.; Zhang, Y.; Yang, C.; Yan, Z.; Xie, Y. Study on distribution map of weeds in rice field based on UAV remote sensing. *J. South China Agric. Univ.* **2020**, *41*, 8.
31. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
32. Sarvamangala, D.; Kulkarni, R.V. Convolutional neural networks in medical image understanding: A survey. *Evol. Intell.* **2021**, *15*, 1–22. [CrossRef]
33. Cicek, O.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9901, pp. 424–432.
34. Wagner, F.H.; Sanchez, A.; Tarabalka, Y.; Lotte, R.G.; Ferreira, M.P.; Aidar, M.P.; Gloor, E.; Phillips, O.L.; Aragao, L.E. Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images. *Remote Sens. Ecol. Conserv.* **2019**, *5*, 360–375. [CrossRef]
35. Shi, C.; Zhou, Y.; Qiu, B.; Guo, D.; Li, M. CloudU-Net: A Deep Convolutional Neural Network Architecture for Daytime and Nighttime Cloud Images' Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1688–1692. [CrossRef]
36. Li, R.; Duan, C.; Zheng, S.; Zhang, C.; Atkinson, P.M. MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *99*, 1–5. [CrossRef]
37. Zhou, Z.; Siddiquee, M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support, Proceedings of the 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11045, pp. 3–11. [CrossRef]

38. Alexakis, E.B.; Armenakis, C. Evaluation of UNet and UNet++ Architectures in High Resolution Image Change Detection Applications. ISPRS—International Archives of the Photogrammetry. *Remote Sens. Spat. Inf. Sci.* **2020**, *43b3*, 1507–1514. [CrossRef]
39. Bao, Y.; Liu, W.; Gao, O.; Lin, Z.; Hu, Q. E-Unet++: A Semantic Segmentation Method for Remote Sensing Images. In Proceedings of the 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 18–20 June 2021; pp. 1858–1862. [CrossRef]
40. Raza, A.; Huo, H.; Fang, T. EUNet-CD: Efficient UNet++ for Change Detection of Very High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3510805. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.