electronics

MDPI

Article

TransAttention U-Net for Semantic Segmentation of Poppy

Zifei Luo [1,2], Wenzhu Yang [1,2,*], Ruru Gou [1,2] and Yunfeng Yuan [1,2]

[1] School of Cyber Security and Computer, Hebei University, Baoding 071002, China
[2] Hebei Machine Vision Engineering Research Center, Hebei University, Baoding 071002, China
* Correspondence: wenzhuyang@hbu.edu.cn; Tel.: +86-15720127565

Abstract: This work represents a new attempt to use drone aerial photography to detect illegal cultivation of opium poppy. The key of this task is the precise segmentation of the poppy plant from the captured image. To achieve segmentation mask close to real data, it is necessary to extract target areas according to different morphological characteristics of poppy plant and reduce complex environmental interference. Based on RGB images, poppy plants, weeds, and background regions are separated individually. Firstly, the pixel features of poppy plant are enhanced using a hybrid strategy approach to augment the too-small samples. Secondly, the U-Shape network incorporating the self-attention mechanism is improved to segment the enhanced dataset. In this process, the multi-head self-attention module is enhanced by using relative position encoding to deal with the special morphological characteristics between poppy stem and fruit. The results indicated that the proposed method can segmented out the poppy plant precisely.

Keywords: semantic segmentation; U-Shape network; multi-head self-attention

1. Introduction

Image semantic segmentation is an important direction in the field of machine vision and is a fundamental task for labeling and classifying each pixel. Segmentation of an image is the process of dividing the image into several regions with different characteristics based on the similarity, processing the image from the pixel level.

The recognition and segmentation of green crops is one of the significant studies in agricultural machinery vision systems [1], where the target regions are extracted according to different morphological features of crops, which are widely used in scenarios such as yield estimation [2], precision agriculture [3], and variety monitoring [4]. Meanwhile, advances in the technology of electronic and avionics systems for UAVs mainly include cost reduction and miniaturization of equipment, bringing efficiency gains for green crop segmentation [5–7].

Poppy is an annual herbaceous plant. The stem is 30–80 cm tall and the flower buds are ovoid, long-stalked and pendulous when not in bloom. The juice extracted from the capsule is processed into opium, morphine, and heroin. As a result, the opium poppy has become an important source of drugs in the world. The cultivation of opium poppy should be strictly regulated.

In agricultural production, traditional methods of yield estimation of plants often utilize manual methods that are severely labor intensive. Unlike common crops, opium poppies are often grown on a small scale and require more precise segmentation. The main identifying features of poppy are the plant morphology and the characteristics of the stamens and fruits. Poppies have large leaves with irregularly wavy edges, and the base of the leaf encloses the stem in a clasping pattern. However, the petal color of the poppy varies from pink to red due to the variety, so the flower color is not a good feature. By now, there is no existing method for segmentation of poppy images.

Citation: Luo, Z.; Yang, W.; Gou, R.; Yuan, Y. TransAttention U-Net for Semantic Segmentation of Poppy. *Electronics* **2023**, *12*, 487. https://doi.org/10.3390/electronics12030487

Academic Editor: Cheng-Chi Lee

Received: 17 December 2022
Revised: 11 January 2023
Accepted: 16 January 2023
Published: 17 January 2023

The main goal of our work is to perform accurate segmentation of poppy images to distinguish poppies from other crops. In order to achieve accurate segmentation of poppy, we need to solve the following difficulties:

- Poppy is not a common crop, which is difficult to collect and the number of samples is small to support the training of a high- accuracy network.
- RGB images can mostly retain the morphological features of poppy and other crops, but there will be interferences in the images. The similar tree crowns or other plant interferences can easily lead to the problem of segmentation errors.
- The change in shooting height brings complex scale features.
- The plant morphology of opium poppy is special, and there is often a long distance between the fruit and stem of the same poppy plant, which can lead to loss of edge information. Due to the lack of ability of traditional methods to process the rich semantic information of images, many works have been done on agricultural yield estimation using deep learning techniques. As an important tool for image processing, convolutional neural networks can make full use of the semantic information of images. However, limitations still exist due to convolutional operations in modeling long-distance relationships. Therefore, these architectures usually yield weaker performance, especially for poppy images with large differences in structural textures, shape features. In this regard, we use CNN as the backbone to integrate transformer technology, which has excellent attention mechanism. The combination with CNN effectively avoids the high computing power required for a network built by transformer only.

In this paper, the enhanced images are trained using an improved deep network model to effectively segment poppies and other species of plants. The size of poppy cultivation is estimated by using the segmentation results obtained from the deep network. Our key ideas are as follows:

- We use ExG (Excess Green index), CIVE (Color Index of Vegetation Extraction) and the combined vegetation indices COM (Combination index) to process the original image instead of the original RGB image as the network input for the distinctive features of poppy (leaf texture, fruit shape) [8]. This is useful for the task of semantic segmentation of poppy with few samples and inter-class interference.
- A U-shaped network—TAU-Net is improved using Transformer for the semantic segmentation task of poppy images captured by UAVs. The backbone network incorporates both CNN networks and self-attentive mechanisms. Poppy features vary widely at different scales. Unlike the network built by the self-attentive module only or the convolution module only, TAU-Net takes advantage of the transformer to perceive the whole image instead of the original U-net which has a restricted field of perception. The improved network improves the robustness of the network to scale changes without imposing high computational effort.
- Poppy images collected by UAVs have high resolution and pixels with high spatial structure. There is a remote dependency between the fruit and the rootstock of the plant of poppy. In this paper, the number of tokens is huge when using transformer in encoding stage. The relative position encoding method learns the relationship between tokens to maintain more accurate position information.

2. Related Works

2.1. Plant Image Enhancement

One of the drawbacks of using supervised methods in agricultural image segmentation tasks is the lack of sufficient labeled datasets, which often affects the training process of the network [9]. The robustness and generality of deep learning models are influenced by the diversity and amount of training data. For example, in the segmentation of plant pests and diseases, cases have different onset conditions and some lesions have sparse image samples. It is difficult to have enough data to support training in practical projects [10]. In small sample problems, traditional methods usually perform geometric or color transformations on existing data, however, they do not substantially increase the dataset. There are other

ideas to solve this problem. One of the approaches is to use multiple channels of an image [11]. Multiple channels segmented from the image (raw RGB data, vegetation indices, HSV color channels, and Canny edge detector) are used as input data for the CNN, which enhances the generalization of the model with limited training data.

2.2. Artificial Neural Networks for Vegetation Image Segmentation

Before the rise of neural network models, there are many traditional methods designed for solving the problems of agricultural image classification and semantic segmentation. Representative traditional algorithms include threshold-based methods, clustering-based methods, wavelet transforms, support vector machine, or Hough transform. In vegetation image segmentation [12], traditional methods follow the setting of equality between vegetation and other objects in one image, which are not fully effective at different stages of plant growth. On the other hand, they have limitations due to the influence of changing light conditions on the obtained vegetation segmentation results. There have been many excellent methods to form segmentation of vegetation images using neural networks.

Fully convolutional network (FCN) is a pioneering work in deep learning for semantic segmentation, establishing a framework for a generic network model for semantic segmentation of images (i.e., pixel-level classification of targets) and providing key ideas for the development of encoder-decoder networks [13]. SegNet is one of the classic models of encoder-decoder network [14]. It follows the FCN architecture and the semantic segmentation network is obtained by VGG16 [15]. Although SegNet is fast in convergence, it does not fully consider the pixel-to-pixel relationship. In the field of agriculture, SegNet is more advantageous in large volume target extraction for high spatial resolution remote sensing images. When extracting sunflower planting areas, SegNet achieved the best accuracy of 89.8% with image fusion performed. Although SegNet is fast in convergence, it does not fully consider the pixel-to-pixel relationship. DeepLab improves FCN by employing atrous convolution [16]. DeepLab network is excellent at extracting more dense features. At the same time, DeepLab largely expands the receptive field and obtains a multi-scale global background, but is still limited by the local area. To segment the lychee trunk [17], the Xception feature extraction model used by DeepLabV3+ was improved at different layers [18]. The MIoU obtained by the network is 76%. DeepLabV3+ network is excellent at extracting more dense features. At the same time, DeepLabV3+ largely expands the receptive field and obtains a multi-scale global background, but is still limited by the local area. Secondly, the huge number of parameters bring tremendous computational burden.

The U-Net architecture based on FCN, demonstrated excellent segmentation performance [19]. But usually, the sizes of pixel blocks are much smaller than the whole image. The algorithm can only extract some local feature information, which leads to the limitation of the classification performance. In weed image segmentation [20], the simplified U-Net obtained the IoU of 89.45% on the validation set.

2.3. Attentive Mechanisms

The essence of the attention mechanism is the specific selection of input data, which focuses on the most critical information in a large amount of data by highlighting the key inputs on the output and suppressing the non-important information. Self-attention is a variation of the attention mechanism that relies less on external information and is better at capturing the internal relevance of data or features [21].

For vision, attention mechanisms can be divided into channel attention, spatial attention, temporal attention, and branching attention according to dimensions [22]. The residual attention network [23], which consists of multiple attention modules stacked, is able to quickly collect global information of images and combine the global information with the original feature maps, but suffers from the high computational load. The SENet network adaptively adjusts the feature map channel weights by establishing interdependencies between the feature map channels through compression-excitation methods [24].

The attention mechanism is fused into the depth model used for semantic segmentation to take advantage of it and make the network more sensitive to the focus regions.

Transformer originated in the field of natural language processing and has been a great success in the field of vision as well [21]. ViT is an encoder-decoder network with only attention modules stacked up, which implements image serialization by decomposing patches [25]. On this basis, DeiT has higher efficiency [26]. It has been shown that the self-attentive layer can achieve high precision semantic segmentation instead of convolutional operation, but it brings a huge computational cost.

3. Methodology

Semantic segmentation of UAV images is a computationally intensive task, especially in segmenting poppy images with complex features and inter-class interferences. Besides, the sample size of poppy dataset is small. Therefore, the U-Net network with simple structure, small number of parameters and good performance in previous small sample agricultural image segmentation tasks is preferred for this task. The U-Net network has both a systolic path that captures contextual information and a symmetric extended path that allows precise localization. These paths allow the network to propagate contextual information to higher resolutions and fuse feature information at different scales in poppy images.

However, U-Net has limitations: patch redundancy and difficulty in handling the existence of remote dependencies between the poppy's fruit and plant rootstocks.

Some of the poppy images are shown in Figure 1. These images have differences in scale, angle, and light intensity.

Figure 1. Poppy images have differences in (**a**) shooting, (**b**) light intensity and (**c**) different scales.

We improve an end-to-end neural network incorporating self-attentive mechanism for achieving accurate segmentation. The segmentation system can be divided into two steps: firstly, for the challenge of too small samples, we divide and rotate the poppy images, then, calculate different vegetation indices to separate channels. The processed images are used as additional representations to support CNN training. Secondly, we design an encoder-decoder semantic segmentation network TAU-Net incorporating attention mechanisms to semantically annotate the input data. The following summary provides a detailed description of the above steps.

3.1. Input Presentation

To reduce the impact on the original image distribution, the RGB images are often used directly as input to the neural network. The optimizer is allowed to decide how to adjust the parameters to train the data. The limited number of samples result in this approach being insufficient to train a segmentation network with good accuracy. To solve this problem, we borrowed the vegetation index approach, which has an excellent track record in agriculture, to derive additional representations from the original RGB images. In

poppy images, the features that significantly differ between poppy plants and other plants contain: (1) the special jagged texture features of the leaves; (2) the round texture features and color features of the fruits. We selected four indices that are sensitive to poppy features: ExG, CIVE, GB, COM for image enhancement.

ExG, CIVE can effectively enhance the feature differences between plants and other categories (soil, eaves). I_R, I_G and I_B respectively represent the normalized pixel values of the corresponding bands.

$$I_{ExG} = 2 * I_G - I_B - I_R \tag{1}$$

$$I_{GIVE} = 0.441 * I_R - 0.811 * I_G + 0.385 * I_B + 18.78745 \tag{2}$$

GB (Green-Blue index) serves to extract the plant texture features well and enhance the feature differences between poppies and other plants. Meanwhile, we took advantage of the strengths of each factor and combined them as COM to further expand the dataset.

$$I_{GB} = I_G - I_B \tag{3}$$

$$I_{COM} = I_R + I_B \tag{4}$$

As shown in the Figure 2, the transformed image is connected to the channelized input RGB image and the normalized input image is built. These additional representations not only help to learn the weight parameters, leading to better generalization properties of the network, but also obtain better performance in separating poppy vegetation.

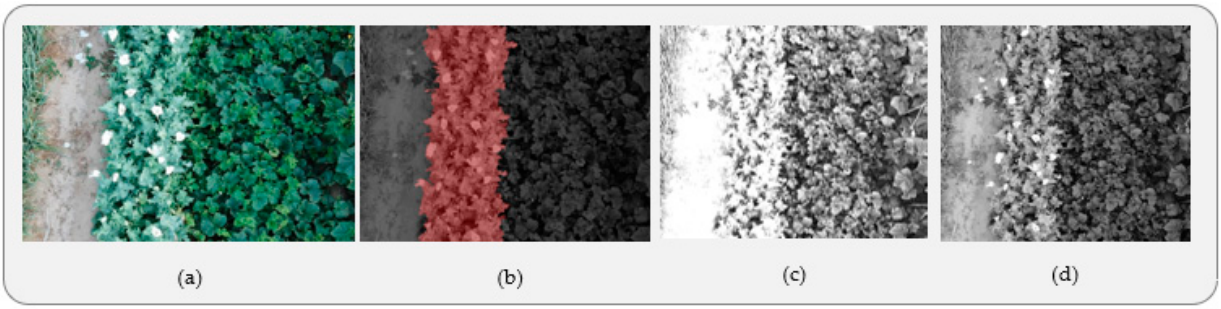

Figure 2. (**a**) The original poppy image, (**b**) ground truth and (**c**,**d**) the augmented image.

3.2. Network Architecture

We use the classical segmentation network U-Net as the backbone. U-Net network consists of three components: (1) Down-sampling stage that abstracts features level by level. (2) Up-sampling stage that reconstructs the features. (3) A final convolutional layer to achieve classification. We propose an end-to-end network with U-Net as the backbone: TAU-Net (Transformer Attention U-Net) network. TAU-Net takes advantage of the transformer to sense the whole image instead of the original U-net which has a restricted sense field. It is very suitable for poppy segmentation tasks with large feature differences at different scales. The network model diagram is shown in Figure 3.

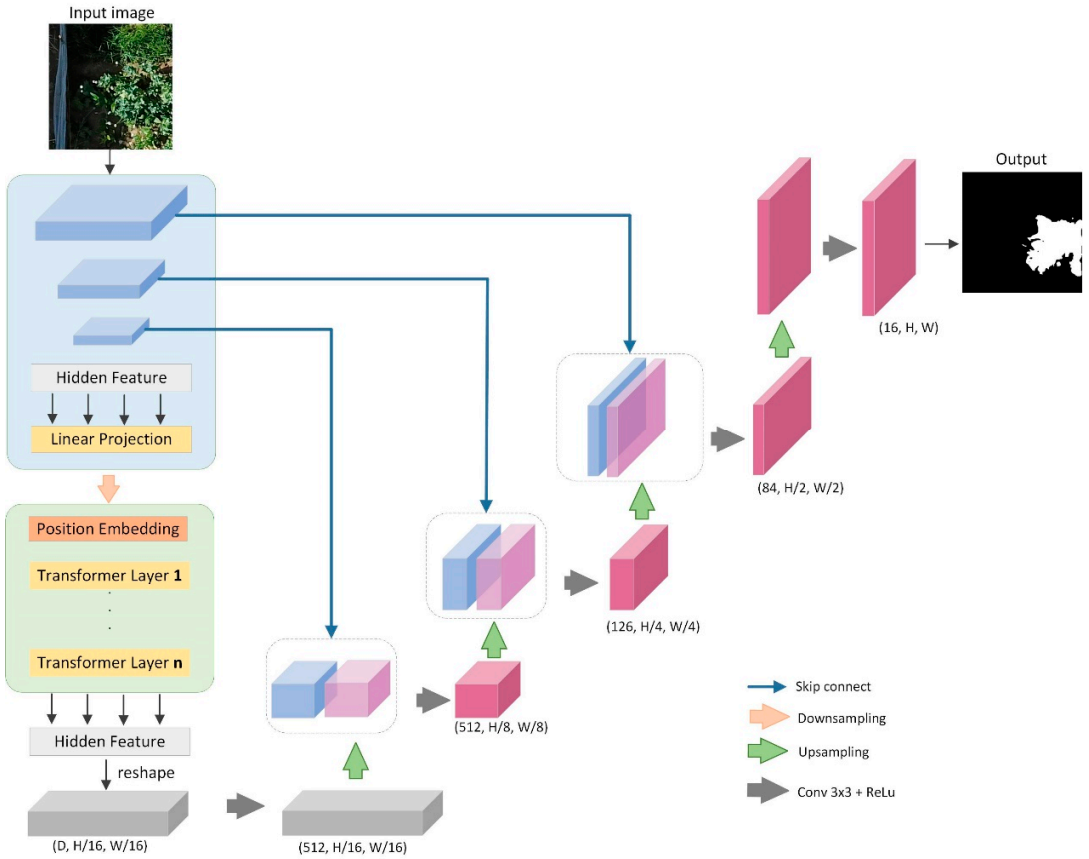

Figure 3. The TAU-Net network model.

Input: We use the normalized dataset after the first stage of enhancement as the input to the TAU-Net network. For alignment, the size of the input image is adjusted to $x \in \mathbb{R}^{H \times W \times C}$ where the spatial resolution is expressed as $H \times W$ and the number of channels is C. The adjusted image is the input of first convolutional layer.

CNN-based encoder: We use different convolutional modules in different stages of the encoder-decoder network. The down-sampling stage uses ResNet to encode the image into a high-level feature representation.

Transformer-based encoder: There are various interference factors in natural environment, such as allium fistulosum (similar fruit morphology features) and white plastic (similar flower color features). The self-attention module can suppress irrelevant and interfering features, effectively extract global information of feature maps in each down-sampling stage. The first step in using the transformer as an encoder is tokenization of the input. We reshape the input x into a flat two-dimensional sequence We reshape the input x into a flat two-dimensional sequence $\left\{ x_p^i \in \mathbb{R}^{P^2.C} \middle| i = 1, \ldots, N \right\}$. The size of the patch is $P \times P$, the length of the input sequence is $N = \frac{HW}{P^2}$, and the dimension of each patch is $d_x = P^2C$. The second step is patch embedding. The patch x_P is mapped to an embedding space of dimension D in the linear projection section. The third part is the transformer layer. Transformer layer consists of the L-layer multi-head attention module and the multilayer perceptron. The sequence inputs are later normalized and input to the Multi-Head Attention (MHA) module and the Multi-Layer Perceptron (MLP) module, and the residual connections reduce information loss. The MHA module obtains a receptive field containing the whole image by learning the relationship between each pixel. x denotes the different embeddings of the feature map. The module has three inputs: query matrix Q, key matrix K, and value matrix V. In this module, Q, K, V matrices have equal size correspond to three embedding matrices WQ, WK, WV. One row of attention matrix A

represents the correlation of a single element of Q with all elements of K and implements attention computation in multiple headers.

In the task of semantic segmentation of poppy, the number of tokens is large. At the same time, pixels are highly spatially structured, and the special botanical morphology of poppy leads to a remote dependency between the fruit and the plant rootstock. Relative position encoding (RPE) is usually computed by means of a look-up table with learnable parameters that interact with the query and key in the self-attention module. Such a scheme allows the module to encode the relative distance between input tokens combining relative position encoding of different embedding features, capturing very long dependencies between tokens. The improved encoding module is able to handle longer sequences, maintaining the translation invariance required for semantic segmentation, and is able to further improve the representational power capability of the model.

The core of transformer is the self-attention mechanism. Self-attention mechanism maps a query Q and a set of key values to an equal sequence length output. The output sequence is $z = (z_1, \ldots, z_n)$ where $z_i \in \mathbb{R}^{d_z}$ is represented as a linearly transformed weighted sum of input parameters.

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_i W^V) \tag{5}$$

Each weighting coefficient α_{ij} is calculated using Softmax function.

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})} \tag{6}$$

In the above equation e_{ij} is calculated using scaled dot product attention:

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}} \tag{7}$$

where the projections $W^Q, W^K, W^V \in \mathbb{R}^{d_x \times d_z}$ are parameter matrices and are unique per layer.

Self-attention can be expressed after adding relative position encoding as:

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_i W^V + p_{ij}^V) \tag{8}$$

$$e_{ij} = \frac{(x_i W^Q + p_{ij}^Q)(x_j W^K + p_{ij}^K)^T}{\sqrt{d_z}} \tag{9}$$

We use $p_{ij}^V, p_{ij}^Q, p_{ij}^K \in \mathbb{R}^{d_z}$ where $d_z = d_x$ to encode the relative positions added to the query vector Q, the key vector K and the value vector V, respectively. They are used to represent the relative position relationship between two tokens. The representation of the tokens position pipe is added during the computation of self-attention.

Relative position encoding of the bias mode has been added to the MHA module. The improved MHA is shown in the Figure 4.

The output at layer l is expressed as Formulas (10) and (11).

$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1} \tag{10}$$

$$z_\ell = MLP(LN(z'_\ell)) + z'_\ell \tag{11}$$

Decoder: The decoder is used to decode the hidden features. The final mask is obtained after several decoding modules. The decoding stage contains deconvolution, 3×3 convolution, ReLU. Meanwhile, the skip connection achieves multi-scale high-resolution feature information from the down-sampling stage. This design improves the learning rate

while preserving the edge information and compensates for the loss of low-level details due to transformer.

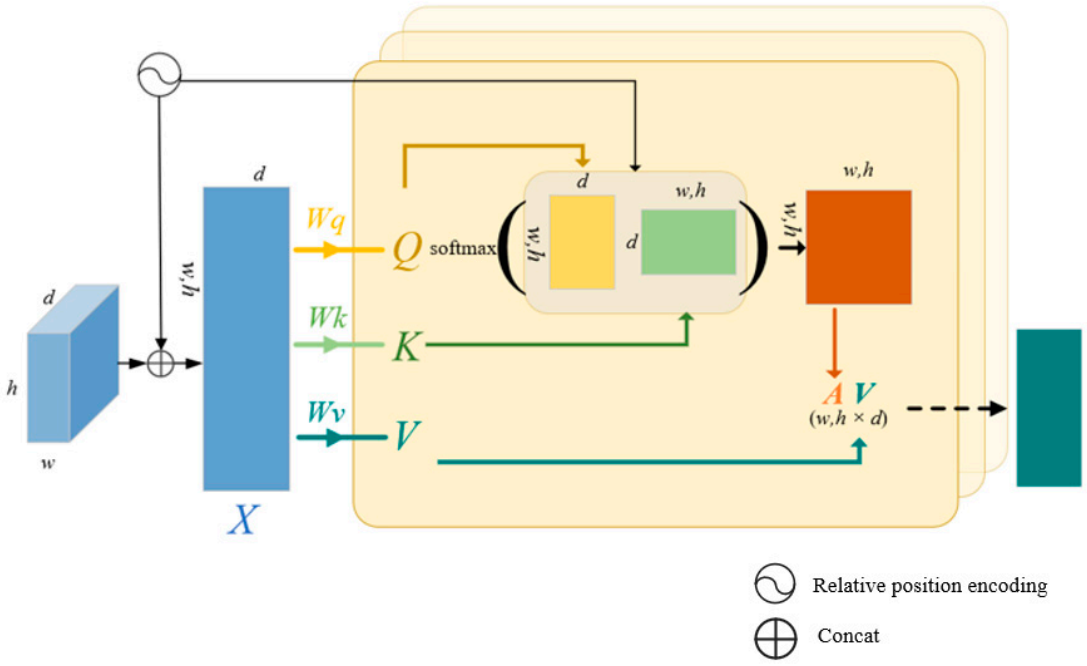

Figure 4. The MHA model with relative position encoding.

4. Experimental Evaluation and Discussion

4.1. Experimental Conditions and Dataset

The UAV images were collected during the flowering and fruiting stages of the poppy. These images contain 54 near-ground images with 3000 × 4000 resolution. To verify the experimental effect, poppies were photographed from different scales. The images were taken with different levels of shading and crop intermingling. We use labelme software to label the images into two categories, poppy or background, and divide the training set and test set according to the ratio of 4:1. The image preprocessing details are given in the above section.

4.2. Evaluation Metrics

In this study, two metrics were used to compare the experimental effects of poppy segmentation: the Dice Score, the training time per image. The dice loss comes from dice coefficient, a metric function used to evaluate the similarity of two samples, mostly used in semantic segmentation tasks. The formula is as follows.

$$dice = \frac{2 \times TP}{2 \times TP + FP + FN}$$

4.3. Comparison Experiment

For the design of the hybrid encoder, we chose ResNet-50 as the encoder part of the CNN in. The backbone network was pre-trained on ImageNet. Due to the Transformer requirement, we adjusted the input resolution to 224 × 224 and reduced to full resolution using up-sampling in the decoder.

We set the batch size to 2, 4, 8, 16 and epoch to 100, 150 respectively. Because of the small sample size, the learning rate is set to 0.01. The weight decay to 1×10^{-4}. The best seg-mentation result is obtained when the batch size is 4, epoch is 100 and learning rate is 0.01. All experiments were performed using a single NVIDIA RTXA4000 GPU. We conducted the main experiments on the Poppy dataset, comparing the improved TAU-Net with the advanced segmentation networks U-Net and DeepLabV3+ on the same dataset.

The loss function used is the loss function of Dice Loss combined with BCE. DiceBCE Loss has the following advantages.

- When a poppy image appears the poppy sample only occupies a very small area, this situation of extreme imbalance between the foreground and background. If a 512 × 512 image has only a 10 × 10 split sample, BCE cannot solve this extremely uneven situation, but Dice Loss is not affected by the foreground size.
- When a poppy image includes a large poppy sample and a small poppy sample at the same time, this situation belongs to the unbalanced content of the split. If a 512 × 512 image has a 10 × 10 and a 200 × 200 segmentation sample, Dice Loss will tend to learn the large block and ignore the small sample, but BCE will still learn the small samples.

Combining Table 1 and Figure 5 shows that U-Net and DeepLabV3+ can all predict the soil-poppy boundary better.

Table 1. The segmentation results of TAU-Net on the poppy dataset. The segmentation evaluation values represent the average of 5 experiments. Besides, Table 1 also contains the results of U-Net and DeepLabV3+ comparing experiments on the same dataset.

Module	Dice Score	Run Time (s)
U-Net	0.74	**1.4**
DeepLabV3+	0.66	2.03
TAU-Net	**0.77**	1.87

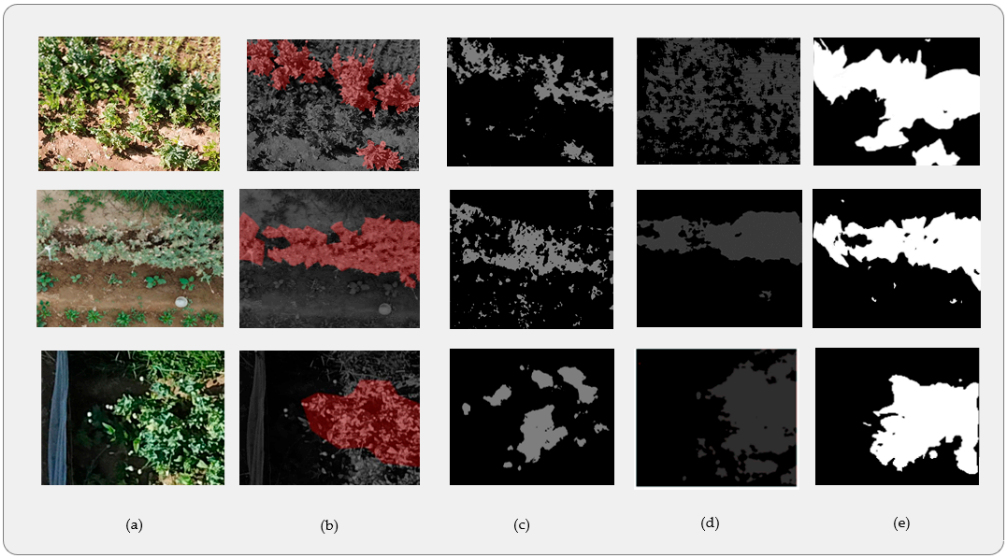

Figure 5. (**a**) Poppy image, (**b**) ground truth and segmentation masks of: (**c**) U-Net, (**d**) DeepLabV3+, (**e**) TAU-Net (with PRE).

In the U-Net segmentation results, false positives are more frequent. The main problem is that small patches of weeds are incorrectly segmented as poppies, but the correct prediction rate is higher for other plants in large scale.

The accuracy of DeepLabV3+ decomposition results was better than that of U-Net, especially for the densely planted poppy areas with vertical shots. However, when the shooting angle changes, the prediction accuracy of DeepLabV3+ becomes very low and the number of parameters is large, and the training speed is slower than that of U-Net.

The TAU-Net can not only predict the boundary between soil and poppy well, but also has fewer false positives, which indicates that TAU-Net has an advantage over other methods in suppressing noise. Meanwhile, the introduction of Transformer did not impose an excessive computational burden on the network.

4.4. Ablations Study

In this paper, we conducted ablation experiments on TAU-Net to observe the effect of different settings on the experimental results. The first experiment we conduct is to analyze the effect of excluding the addition of the relative position encoding MHA module. The second experiment is to observe the effect of changing the number of layers of the transformer. The first row of each table shows the segmentation performance of each default TAU-Net. The remaining rows of each model were trained with the same hyperparameters and the same settings. The impact of the relative position encoding MHA module as shown in Table 2.

Table 2. Ablation performance of the MHA module for relative position encoding.

Method	Dice Score	Run Time (s)
No relative position code added	0.74	**1.84**
Add relative position code	**0.77**	1.87

From the Figure 6, the improved MHA module effectively improves the fruit segmentation at the edges.

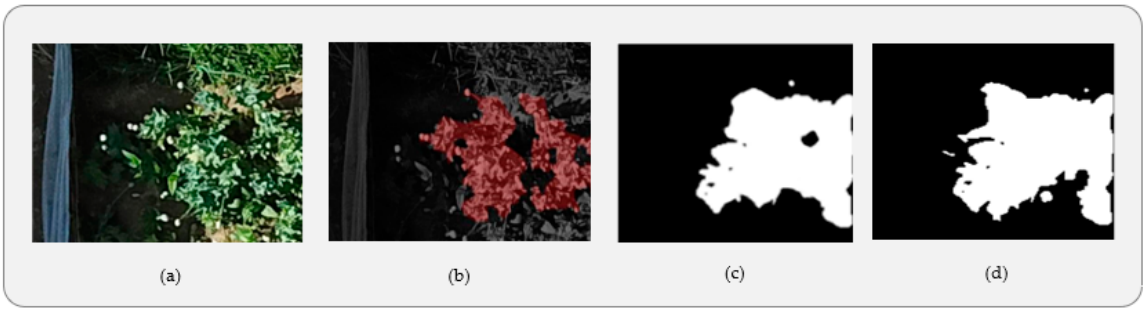

Figure 6. (**a**) Poppy image, (**b**) ground truth and segmentation results of (**c**) TAU-Net without RPE, (**d**) TAU-Net with the RPE modified MHA module.

5. Conclusions

In this paper, we improve a U-shaped network TAU-Net using Transformer for the semantic segmentation task of poppy images captured by UAV. The backbone network incorporates both CNN networks and self-attentive mechanisms. Poppy features vary widely at different scales. Unlike the network built by the self-attentive module only or the convolutional module only, TAU-Net takes advantage of the transformer to remove irrelevant or noisy regions from the features and better extracts the poppy information, and also unlocks the limitations of the original U-Net sensory field. The improved network improves robustness to scale changes without high computational effort. Poppy images collect by UAV have high resolution, pixels have high spatial structure, and there is a remote dependency between the poppy's fruit and plant rootstock. The relative position method learns the relationship between tokens, retains more accurate position information, and achieves the optimization of segmentation boundary information.

However, the method proposed in this paper still has some limitations:

- There are false positives in the experimental results, which reduce the experimental accuracy.
- The proposed method is weak in handling intercropping.
- Image acquisition in natural environment is affected by lighting conditions. The feature performance of poppies varies greatly under different light intensities, which have a large impact on the learning ability of the network.
- The robustness of the network to changes in photo angles also needs to be enhanced.

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

Funding: This work was supported by the Natural Science Foundation of Hebei Province under Grant (F2022201003) and the Post-graduate's Innovation Fund Project of Hebei University under Grant (HBU2022ss037).

Data Availability Statement: No public dataset available.

Acknowledgments: We thank the High-Performance Computing Center of Hebei University for the equipment and other help offered.

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

References

1. Mavridou, E.; Vrochidou, E.; Papakostas, G.A.; Pachidis, T.; Kaburlasos, V.G. Machine vision systems in precision agriculture for crop farming. *J. Imaging* **2019**, *5*, 89. [CrossRef] [PubMed]
2. Narin, O.G.; Abdikan, S. Monitoring of phenological stage and yield estimation of sunflower plant using Sentinel-2 satellite images. *Geocarto Int.* **2022**, *37*, 1378–1392. [CrossRef]
3. Aslan, M.F.; Durdu, A.; Sabanci, K.; Ropelewska, E.; Gültekin, S.S. A comprehensive survey of the recent studies with UAV for precision agriculture in open fields and greenhouses. *Appl. Sci.* **2022**, *12*, 1047. [CrossRef]
4. Rehman, A.; Saba, T.; Kashif, M.; Fati, S.M.; Bahaj, S.A.; Chaudhry, H. A revisit of internet of things technologies for monitoring and control strategies in smart agriculture. *Agronomy* **2022**, *12*, 127. [CrossRef]
5. Hassan, M.A.; Javed, A.R.; Hassan, T.; Band, S.S.; Sitharthan, R.; Rizwan, M. Reinforcing Communication on the Internet of Aerial Vehicles. *IEEE Trans. Green Commun. Netw.* **2022**, *6*, 1288–1297. [CrossRef]
6. Hassan, M.A.; Ali, S.; Imad, M.; Bibi, S. New Advancements in Cybersecurity: A Comprehensive Survey. In *Big Data Analytics and Computational Intelligence for Cybersecurity*; Springer: Cham, Switzerland, 2022; pp. 3–17.
7. Lateef, S.; Rizwan, M.; Hassan, M.A. Security Threats in Flying Ad Hoc Network (FANET). In *Computational Intelligence for Unmanned Aerial Vehicles Communication Networks*; Springer: Cham, Switzerland, 2022; pp. 73–96.
8. Kitzler, F.; Wagentristl, H.; Neugschwandtner, R.W.; Gronauer, A.; Motsch, V. Influence of Selected Modeling Parameters on Plant Segmentation Quality Using Decision Tree Classifiers. *Agriculture* **2022**, *12*, 1408. [CrossRef]
9. Kamilaris, A.; Prenafeta-Boldú, F.X. A review of the use of convolutional neural networks in agriculture. *J. Agric. Sci.* **2018**, *156*, 312–322. [CrossRef]
10. Yuan, Y.; Chen, L.; Wu, H.; Li, L. Advanced agricultural disease image recognition technologies: A review. *Inf. Process. Agric.* **2021**, *9*, 48–59. [CrossRef]
11. Milioto, A.; Lottes, P.; Stachniss, C. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 2229–2235.
12. Mardanisamani, S.; Eramian, M. Segmentation of vegetation and microplots in aerial agriculture images: A survey. *Plant Phenome J.* **2022**, *5*, 20042. [CrossRef]
13. 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.
14. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional Encoder-Decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
15. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
16. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
17. Peng, H.; Xue, C.; Shao, Y.; Chen, K.; Xiong, J.; Xie, Z.; Zhang, L. Semantic segmentation of litchi branches using DeepLabV3+ model. *IEEE Access* **2020**, *8*, 164546–164555. [CrossRef]
18. 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.
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
20. Zou, K.; Chen, X.; Wang, Y.; Zhang, C.; Zhang, F. A modified U-Net with a specific data argumentation method for semantic segmentation of weed images in the field. *Comput. Electron. Agric.* **2021**, *187*, 106242. [CrossRef]

21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; LGB: Los Angle, CA, USA, 2017; Volume 30.
22. Mou, L.; Zhao, Y.; Chen, L.; Cheng, J.; Gu, Z.; Hao, H.; Qi, H.; Zheng, Y.; Frangi, A.; Liu, J. CS-Net: Channel and spatial attention network for curvilinear structure segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2019; pp. 721–730.
23. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
26. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–14 August 2021; pp. 10347–10357.