AAFormer: A Multi-Modal Transformer Network for Aerial Agricultural Images

Yao Shen[1], Lei Wang[2], Yue Jin[1]
[1] China Pacific Insurance (Group) Co., Ltd.
{shenyao-001, jinyue-007}@cpic.com.cn
[2] East China Normal University
71205901087@stu.ecnu.edu.cn

Abstract

The semantic segmentation of agricultural aerial images is very important for the recognition and analysis of farmland anomaly patterns, such as drydown, endrow, nutrient deficiency, etc. Methods for general semantic segmentation such as Fully Convolutional Networks can extract rich semantic features, but are difficult to exploit the long-range information. Recently, vision Transformer architectures have made outstanding performances in image segmentation tasks, but transformer-based models have not been fully explored in the field of agriculture.Therefore, we propose a novel architecture called Agricultural Aerial Transformer (AAFormer) to solve the semantic segmentation of aerial farmland images. We adopt Mix Transformer (MiT) in the encoder stage to enhance the ability of field anomaly pattern recognition and leverage the Squeeze-and-Excitation (SE) module in the decoder stage to improve the effectiveness of key channels. The boundary maps of farmland are introduced into the decoder. Evaluated on the Agriculture-Vision validation set, the mIoU of our proposed model reaches 45.44%.

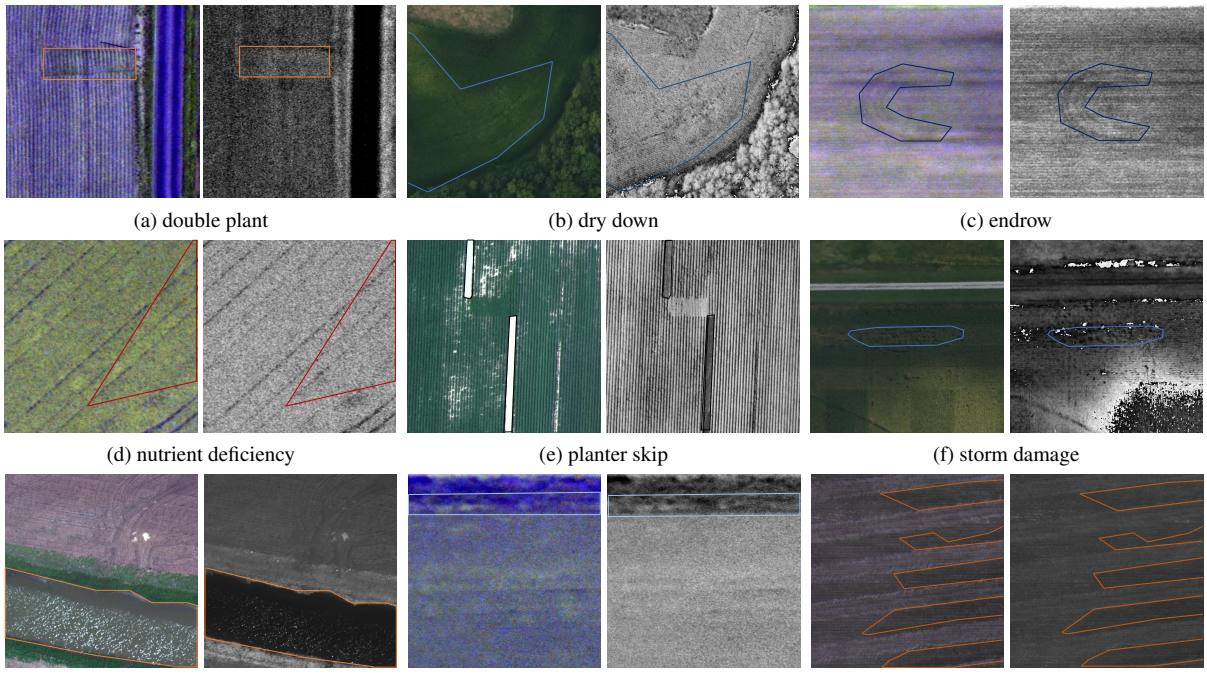
1. Introduction

Semantic segmentation of agricultural aerial images is one of the main research directions in the field of agricultural vision. An effective aerial farmland segmentation algorithm is very important for the detection of anomaly areas in the field, such as segmentation of drydown, endrow, nutrient deficiency, and so on. The recognition of anomaly mode is helpful to monitor the local state of farmland, assess the impact scope of natural disasters, and promote the efficiency of farmland survey in the field of agricultural insurance. Agricultural aerial image analysis also supports the formulation of national agricultural policies to enhance the yield of agricultural fields and regional economic development.

The Agriculture-Vision [6, 7] dataset contains agricultural aerial images with nine anomaly annotations: double plant, drydown, endrow, nutrient deficiency, planter skip, storm damage, water, waterway, and weed cluster. It is a large-scale dataset with high resolution images annotated by agronomic experts. The dataset contains 94986 high-quality aerial images in total. We visualize several samples in Figure 1. There are two main differences from common aerial image datasets. Firstly, the resolution of farmland aerial image is as high as 10 cm per pixel but the size and the shape of annotated anomaly area are irregular. As is shown in Figure 2, the total number of images and pixels marked in different anomaly pattern vary greatly. Secondly, farmland aerial images are multi-spectral. In addition to the common RGB channels, there are NIR (near-infrared) channels. Note that annotation labels are not mutually exclusive while multiple labels may belong to the same pixel. These differences make the semantic segmentation of agricultural aerial images more challenging than that in other domains.

For the semantic segmentation task of agricultural aerial images, the most common methods are Fully Convolutional Network (FCN) [15] and DeepLab series [5]. Although convolution neural network can extract rich feature information, it is difficult to process the information of image context, which is particularly significant for segmentation. However, the localization of receptive fields of the convolution layer limits the learning ability to a relatively small area.

In the field of natural language processing, Transformers achieves the most advanced performance in many tasks, due to the strong ability to capture context information. Inspired by this, researchers begin to apply Transformers to computer vision. It has been proved effective to encode the image into a series of embedded patches, and use self-attention module to learn context information from the patches. However, in the task of semantic segmentation in



(a) double plant (b) dry down (c) endrow

(d) nutrient deficiency (e) planter skip (f) storm damage

(g) water (h) waterway (i) weed cluster

Figure 1. The annotations of the agricultural aerial dataset are divided into nine categories: double plant, drydown, endrow, nutrient deficiency, plant skip, storm damage, water, waterway, and weed cluster. In each category, the left image represents RGB three-channel image and the right image represents near-infrared(NIR) single-channel image.

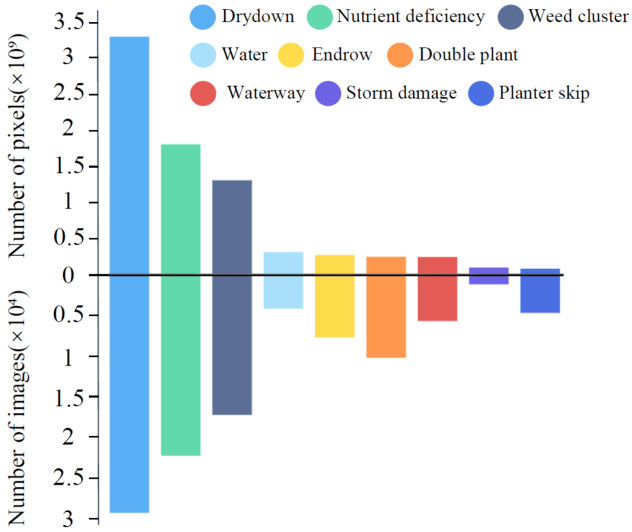


Figure 2. Imbalance of agricultural aerial dataset

agricultural field, the powerful modeling ability of Transformers has not been fully exploited.

In this paper, we propose a novel architecture AAFormer for semantic segmentation of agricultural aerial images. In order to make better use of Transformers to exploit more features from different channels of agricultural aerial images, we firstly combine RGB images and NIR (near-infrared) channel as the input of the model. In the encoding stage, we introduce Mix Transformer [19] which can effectively learn the long-range features from the embedded image patch and enhance the ability to recognize field anomaly patterns. In the decoding stage, we exploit features with different scales to generate the final segmentation results. At the same time, we adopt the SE [10] module which can filter out more noise and improve the accuracy of the model.

To sum up, our main contributions are as follows:

- We propose a novel segmentation architecture AAFormer to deal with the semantic segmentation of agricultural aerial images.
- We introduce a channel attention module in the decoder of AAFormer to improve the effect of key channels.
- We demonstrate the introduction of boundary into the decoder stage will refine the segmentation result.

2. Related Works

The semantic segmentation of aerial images is very important for the detection of anomaly patterns in farmland. Generally, the preliminary attempt of semantic segmentation task is based on Fully Convolutional Network (FCN) [15], SegNet [1] and Deeplab series [5] will use the multi-

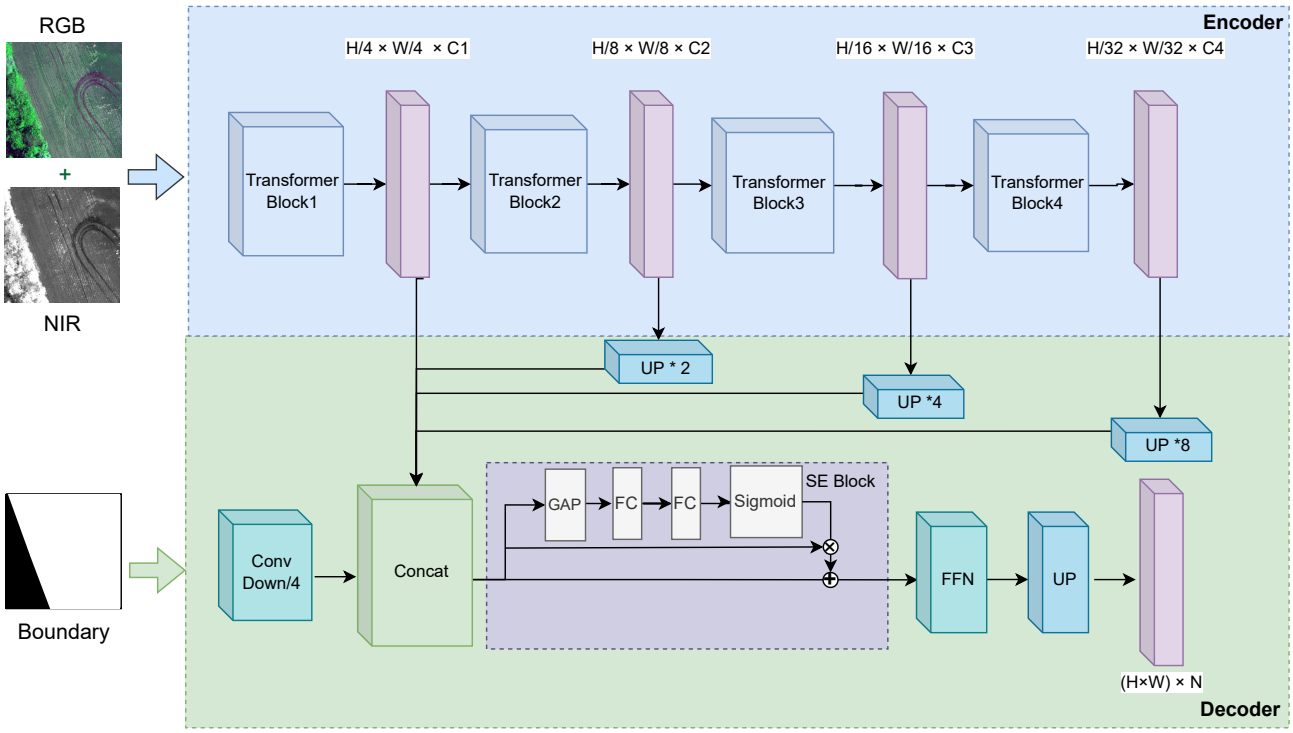


Figure 3. Overall architecture of network structure.

scale information in the image to improve the accuracy of the model.

In the research of agricultural aerial images, a semantic segmentation model [7] was constructed based on FPN to detect the field anomaly patterns of aerial images. Based on the self-constructing graph (SCG) [14], a multi-view self-constructing graph convolution network (MSCG-Net) [13] was proposed and an adaptive class weight loss was developed to solve the class imbalance problem. Considering the multi-modal nature of agricultural aerial images, switchable normalization block [21] was deployed to reduce feature divergence. In addition, Efficient Fused Pyramid Network (Fuse-PN) [11] was proposed for anomaly pattern segmentation in agricultural aerial images by enhancing features on different scales extracted from RGB and NIR.

Although these methods are meant for the semantic segmentation of agricultural aerial images, due to the limitations of the basic network, they do not perform well in learning global context and remote spatial dependence, which may improve the segmentation performance of the model to a certain extent.

The Transformer architecture was first used in machine translation tasks [18], and achieved the most advanced performance in NLP tasks [12]. Many researchers began to explore the application of Transformer architecture to computer vision tasks [9]. Through the unique structure, vision Transformer (ViT) [2] shows excellent ability in the field of video classification. Carion et al. [3] combined Transformer and CNN to detect objects. In SETR [22], a pure transformer is deployed to encode an image as a sequence of patches. And this encoder can be combined with a simple decoder to provide a powerful segmentation model. CoTr [20] bridges CNN and Transformer for 3D medical image segmentation.

At present, Transformer architecture has been applied to many fields of vision research, but the use of the Transformer has received little attention in the field of agricultural semantic segmentation. Therefore, we expect to leverage the powerful modeling ability of Transformer to solve the semantic segmentation of agricultural aerial images.

3. Method

Our proposed method is based on the encoder-decoder framework. The overall architecture of the model is shown in Figure 3.

3.1. Encoder

Agricultural aerial images contain three-channel RGB images and single-channel NIR near-infrared images. To make full use of the image information of RGB and NIR, we first splice them into four-channel images as the input of the model.

Table 1. Performance results of mIoUs and class IoUs on different models. Compared with other methods, our proposed method can further improve the mIoU of the model on the validation set.

Model	mIoU(%)	Background	Double plant	Drydown	Endrow	Nutrient deficiency	Planter skip	Water	Waterway	Weed cluster
DeepLabv3(os = 8) [4]	35.29	73.01	21.32	56.19	12.00	35.22	20.10	42.19	35.04	22.51
DeepLabv3+ (os = 8) [5]	37.95	72.76	21.94	56.80	16.88	34.18	18.80	61.98	35.25	22.98
DeepLabv3(os = 16) [4]	41.66	74.45	25.77	57.91	19.15	39.40	24.25	72.35	36.42	25.24
DeepLabv3+ (os = 16) [5]	42.27	74.32	25.62	57.96	21.65	38.42	29.22	73.19	**36.92**	23.16
Agriculture-Vision baseline [7]	43.40	74.31	28.45	57.43	21.74	38.86	33.55	**73.59**	34.37	28.33
AAFormer (ours) w/o boundary	45.31	**77.43**	**37.26**	59.88	24.13	42.41	**41.64**	69.29	26.63	29.13
AAFormer (ours)	**45.44**	76.85	37.06	**60.93**	**24.45**	**42.42**	41.35	69.17	26.89	**29.89**

To generate multi-scale feature maps, we leverage Mix Transformer (MiT) [19] as our encoder, as shown in Figure 4. An aerial image is divided into multiple smaller patches. Through four levels of the hierarchical Transformer blocks, the features of different levels are obtained from the patches. The main computational bottleneck of encoders is the self-attention layer. Efficient self-attention is introduced to reduce the computational complexity. The Mix-FFN module directly uses a 3×3 convolution operation to provide positional information for feed forward network (FFN) in view of the information leak by zero padding. In addition, the overlapped patch merging process is designed to ensure the continuity of local information around images or feature blocks.

Overall, this hierarchical structure is deployed to efficiently obtain a multi-level feature map to further improve the model performance.

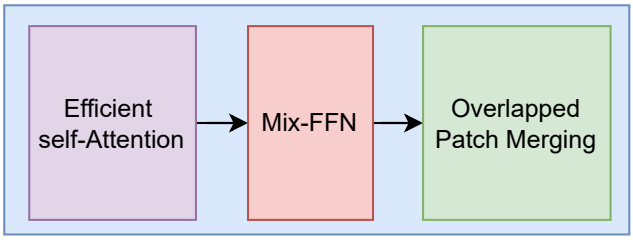


Figure 4. Transformer block

3.2. Decoder

In order to design an effective and efficient model, we adopt a light-weight decoder similar to SegFormer [19]. Benefiting from the powerful receptive field of vision Transformer encoder, we consider it is unnecessary to make too complex design on the decoder. The proposed decoder can be intuitively understood from figure 4. Firstly, feature maps of different scales are obtained from the 4 stages of the encoder. A fully-connected layer is utilized to uniformly adjust the characteristics of each stage to C channels. The width and height are aligned to (H / 4, w / 4) through the up sampling operation. Then, concatenate the feature maps in the channel direction. To raise the model's attention to important channels, we introduce a Squeeze-and-Excitation (SE) block [10] with residual connection. As is shown in Figure 3, the global compression feature is obtained through global average pooling (GAP). And the weight of each channel in the feature map is obtained by the bottleneck of two fully connected layers. We use the weighted features with residual connection to excite useful channels, so as to upgrade the effectiveness of the network. However, such a decoding structure has not made maximum use of the information from agricultural aerial photos. The Agriculture-Vision dataset also provides boundary maps, which identifies farmland and non-farmland areas. The existence of farmland edge is related to specific area types. Considering that the mask-like boundary map is at a higher semantic level, we introduce boundary feature into the decoder, instead of sending into the backbone as an additional channel of the image. This is an option in modeling practice, because the boundary annotation is not always available in other datasets.

3.3. Multi-label Classification Head

Different from ordinary semantic segmentation tasks, a region in farmland may belong to more than one type. In other words, there is a problem of label overlapping. Therefore, we adopt multi-label classification head to judge the existence of each type on every pixel. The loss function needs to adapt to this change. In this work, we propose a hybrid loss L in the combination of binary cross entropy loss L_{ce} and binary dice loss [17] L_{dice} with weights w. The loss functions can be written as

$$L_{ce} = \sum_{j=1}^{c}\sum_{i=1}^{n} y_{i,j}\log x_{i,j} + (1-y_{i,j})\log(1-x_{i,j}) \quad (1)$$

$$L_{dice} = \sum_{j=1}^{c} \frac{2\sum_{i=1}^{n} x_{i,j}y_{i,j}}{\sum_{i=1}^{n} x_{i,j}^2 + \sum_{i=1}^{n} y_{i,j}^2} \quad (2)$$

$$L = w_{ce}L_{ce} + w_{dice}L_{dice} \quad (3)$$

where c and n respectively indicates the number of categories and pixels. And x indicates a logit after sigmoid operation while y is the ground truth.

At inference, we utilize the argmax operation to select the most appropriate type of each pixel, for the convenience to compare with other methods.

4. Experiment

4.1. Implementation

The experiments are conducted on 2 Tesla V100 GPUs. We work on MiT-B3 [19] backbone which is pretrained on the ADE-20k [23] dataset. The code is implemented based on mmsegmentation [8] project. The batch size is set to 32. The experiments are trained for 80k iterations. We utilize AdamW [16] optimizer and set an initial learning rate of 0.00006. The warm-up strategy is used for 1.5k iterations.

4.2. Metric

In our experiment, there may be label overlapping in some annotations in the agricultural aerial image dataset. We set pixels with multiple labels, the prediction of any label will be regarded as the correct pixel classification of the label. On the contrary, the prediction without any basic truth label will be regarded as the wrong classification of all truth labels. The specific calculation method is shown in Formula (4), where n is the number of tag types of anomaly mode, TP_n is the true positive value, P_n and T_n are the prediction results and target of different categories respectively:

$$mIoU = \frac{1}{n}\sum_{i=1}^{9}\frac{TP_i}{(P_i + T_i - TP_i)} \tag{4}$$

4.3. Results

To demonstrate the superiority of our proposed method, we compare it with DeepLabv3 [4], DeepLabv3+ [5] and the baseline method of the dataset [7]. Table 1 shows the customized mIoU scores of various methods evaluated on the Agriculture-Vision validation set. It can be seen that our proposed method has achieved the highest mean score. Considering that the boundary graph is not introduced in the previous methods, we also provide the results without using boundary map. It is visible that our method has certain advantages. Figure 5 shows the visualization of our model on the Agriculture-Vision dataset. The first 2 rows represent RGB and NIR images respectively. The 3rd and 4th rows are the ground truth and prediction results on specific categories. The 5th row represents the dyeing effect of the prediction results on the RGB images with an opacity of 0.5.

Table 2. Discussions related to decoder design and boundaries.

Exp.	SE Block	Boundary	mIoU(%)
1	-	-	45.04
2	✓	-	45.31
3	✓	at input	45.32
4	✓	at decoder	**45.44**

4.4. Ablation Studies & Discussions

To verify the effectiveness of the SE block with residual connection in the decoder part, we make the comparison with an all-MLP decoder [19] (Exp.1) in Table 2. It can be seen that the adoption of SE block without boundaries (Exp.2) by the way proposed can improve 0.27% on the customized mIoU score due to the channel attention mechanism.

We also discuss the strategy of introducing boundaries from different processes of the model. It is shown that using boundary map (Exp.3) as the 5th channel of input has little impact on the results. Instead, our proposed approach (Exp.4) to introduce boundaries at decoder stage gets 0.12% better.

5. Conclusion & Future Works

In this paper, we propose a novel Transformer-based architecture, AAFormer, which realizes the semantic segmentation on agricultural aerial images. In the encoding stage, we introduce the MiT structure to learn the context information from the embedded image patches and enhance the ability to recognize the anomaly patterns of farmland. In the decoding stage, we exploit features with different scales to generate the final segmentation results. Meanwhile, the utilization of SE module and boundary maps is proved to be effective to refine the result of the model.

In the next step, we will aim at solving the problems of over-fitting and data imbalance. Various data augmentation and sampling-based approaches could be applied to this task. On the other hand, we look forward to samples of greater quantity and larger resolution to promote the development of the research.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. 2017. 2

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095*, 2(3):4, 2021. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-

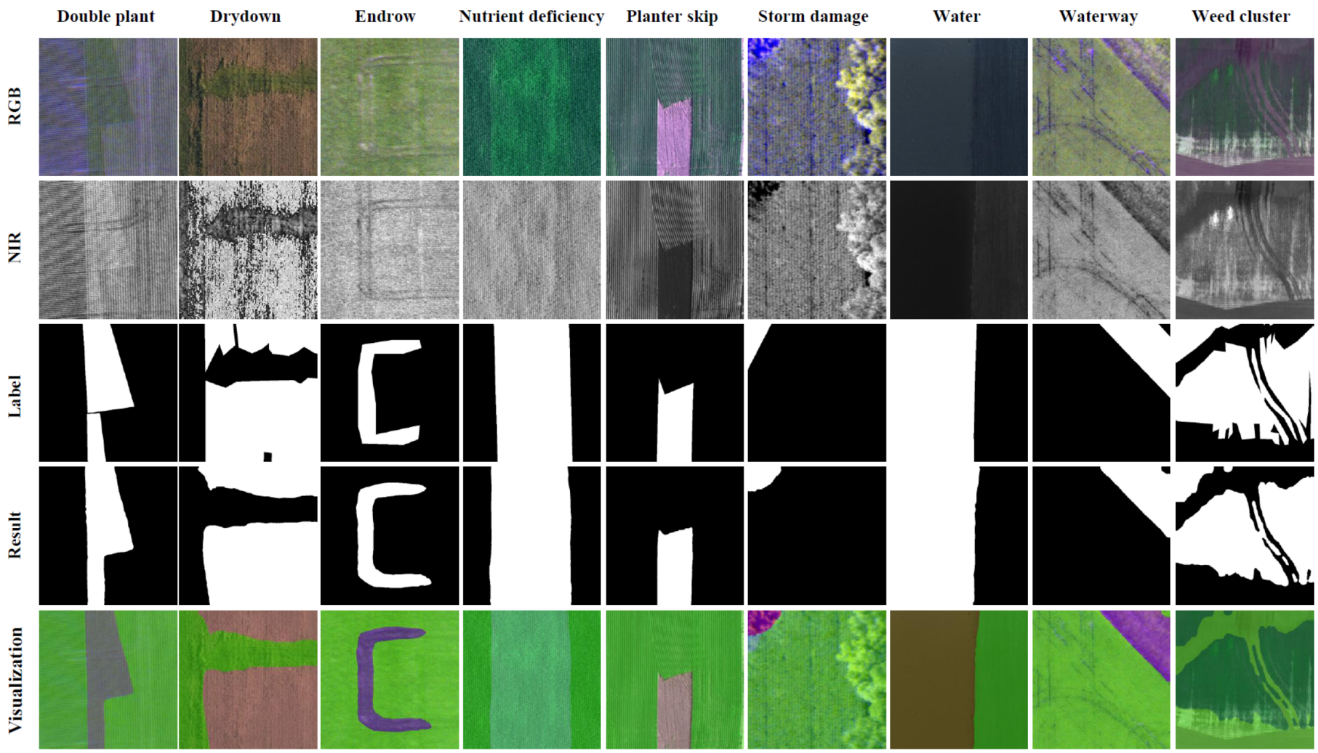


Figure 5. Semantic segmentation results of aerial images.

end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 4, 5

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 2, 4, 5

[6] Mang Tik Chiu, Xingqian Xu, Kai Wang, Jennifer Hobbs, Naira Hovakimyan, Thomas S. Huang, and Honghui Shi. The 1st agriculture-vision challenge: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[7] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G. Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, David Wilson, Adrian Tudor, Naira Hovakimyan, Thomas S. Huang, and Honghui Shi. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3, 5

[8] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation. 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2, 4

[11] Shubham Innani, Prasad Dutande, Bhakti Baheti, Sanjay Talbar, and Ujjwal Baid. Fuse-pn: A novel architecture for anomaly pattern segmentation in aerial agricultural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2021. 3

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 3

[13] Qinghui Liu, Michael Kampffmeyer, Robert Jenssen, and Arnt-Børre Salberg. Multi-view self-constructing graph convolutional networks with adaptive class weighting loss for semantic segmentation. *CoRR*, abs/2004.10327, 2020. 3

[14] Q. Liu, M. Kampffmeyer, R. Jenssen, and A. B. Salberg. Self-constructing graph convolutional networks for semantic labeling. 2020. 3

[15] Long, Jonathan, Shelhamer, Evan, Darrell, and Trevor. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2017. 1, 2

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[17] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 4

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[19] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 4, 5

[20] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021. 3

[21] Siwei Yang, Shaozuo Yu, Bingchen Zhao, and Yin Wang. Reducing the feature divergence of rgb and near-infrared images using switchable normalization. 2021. 3

[22] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 3

[23] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 5