RESEARCH ON SEMANTIC SEGMENTATION OF GREENHOUSE ROAD IMAGE
/
温室大棚道路图像的语义分割研究

YongZheng YANG, HongBo WANG*, ZhiCheng XIE, JunMao LI, ZiLu HUANG
College of Mechanical and Electronic Engineering, Inner Mongolia Agricultural University, Hohhot / China
Tel: +86 13739981395; E-mail: wanghb@imau.edu.cn
DOI: https://doi.org/10.35633/inmateh-71-65

Keywords: *road detection, SETR, DeepLabv3+, semantic segmentation*

ABSTRACT
To realize the automatic driving of agricultural machinery in the greenhouse, this paper uses image acquisition equipment to collect road images in the greenhouse and makes data sets, builds SETR (SEgmentation TRansformer) model based on Transformer framework and DeepLabv3+ model based on convolution neural network for semantic segmentation of road images in the greenhouse, and verifies the semantic segmentation ability of the two models to road images in the greenhouse. Several groups of training periods are set as observation points to observe the semantic segmentation effect of the two models on the greenhouse road image, and the test set which has not been trained by the model is used as the prediction object to verify the performance of the two models on the semantic segmentation of greenhouse road image. The SETR model reached 94.64% PA (Pixel Accuracy) on the greenhouse road data set, and 82.72% mIoU (Mean Intersection over Union), DeepLabv3+ model reached 90.80% PA and 72.35% mIoU on the greenhouse road data set. Both models have excellent performance in semantic segmentation of greenhouse road images, and the performance of SETR model is slightly better than that of DeepLabv3+ model. The semantic segmentation performance of the two models for greenhouse road images can meet the needs of actual deployment.

摘要
为实现农业机械在温室大棚中的自动驾驶，本文使用图像采集设备采集温室大棚中道路图像并制作了数据集，搭建基于 Transformer 框架的 SETR 模型与基于卷积神经网络的 DeepLabv3+模型对温室大棚中道路图像进行语义分割，验证两种模型对温室大棚中道路图像语义分割能力。设置了多组训练周期作为观测点，观测两种模型对温室大棚道路图像的语义分割效果，以未投入过模型训练的测试集作为预测对象，验证两种模型对温室大棚道路图像语义分割的性能。SETR 模型在温室大棚道路数据集上达到了 94.64%的 PA、82.72%的 mIoU,DeepLabv3+模型在温室大棚道路数据集上达到了 90.80%的 PA、72.35%的 mIoU，两种模型在对温室大棚道路图像的语义分割上均有优异表现，SETR 模型的性能略高于 DeepLabv3+模型，两种模型对温室大棚道路图像的语义分割性能均满足在实际部署的需求。

INTRODUCTION
With the development of artificial intelligence, technologies such as autopilot and robot automatic cruise are gradually applied in various industries *(Yoon et al., 2022; Daduna, 2020)*. The autopilot of intelligent mechanical equipment in the greenhouse, open-air farm and other environments is one of the important development directions of intelligent agriculture. As one of the basic tasks of self-driving, semantic segmentation is an important support for the panoramic analysis of roads and the pixel-level classification of road environment. Based on the analysis of the visual navigation extraction algorithm of agricultural robot based on dark primary colour, *Zhongkun Hou, (2020*), proposed a method of farmland image preprocessing and edge detection based on dark primary colour, which provides some reference for the analysis of visual navigation extraction algorithm of field navigation robot based on dark primary colour.

With the development of deep learning technology, more and more models are used in the field of image semantic segmentation, which refers to the pixel-level classification of images *(Xiongi et al., 2021)*.

Yong Zheng YANG, M.S. Stud.; HongBo Wang, Professor, Correspondent author; ZhiCheng XIE, M.S. Stud.; JunMao LI, M.S. Stud.; ZiLu HUANG, M.S. Stud.

For example, the full convolutional network FCN *(Long et al., 2015)* which first applied deep learning technology to semantic segmentation, FCN uses convolutional encoder-decoder architecture to upsample the image to generate image features, and classifies the image at the pixel level in an end-to-end way. Impressive results have been obtained on datasets with PASCAL VOC and Cityscapes. With the brilliance of FCN in the field of image semantic segmentation, there are some models such as Mask R-CNN *(He et al., 2017)*, DeepLab *(Chen et al., 2018)* and BiSeNet *(Yu et al., 2018)* with VGG, ResNet, DensenNet and other frameworks as the backbone of the network, which have excellent performance in image semantic segmentation. *Yu et al. (2021)* further improved and designed a new BiSeNe V2 model for real-time semantic segmentation with higher precision and higher efficiency for BiSeNet, and introduced a new guided aggregation layer to combine features, so that BiSeNet V2 has a better performance for semantic segmentation of data sets such as Cityscapes. *Ren et al. (2023)* improved the BiSeNet network architecture by sharing the head of the two-branch network, eliminated parameter redundancy and improved the extraction of shallow features of the image, and verified that the algorithm has a significant improvement in real-time performance and accuracy on multiple data sets. *Chen et al. (2018)* proposed a new model DeepLabv3+, by combining the advantages of pyramid pooling module and encoder-decoder structure, and verified that the encoder-decoder network structure is faster and stronger on PASCAL VOC and Cityscapes data sets. The HRNet designed by *Wang et al. (2020)* realizes the parallel connection of convolutional streams from high resolution to low resolution and the exchange of repeated information across resolutions, so that the image maintains a high-resolution representation in the whole process of coding, and the features have more accurate spatial information and richer semantics. *Fan et al. (2021)* designed a new network structure STDC on the basis of BiSeNet, and did a number of experiments on CamVid data set and Cityscapes data set to verify the effect of the model, and finally realized 76.8% mIoU with 97.0FPS. *Hu (2020)* introduced a new efficient and lightweight neural architecture RandLA-Net to realize the semantic segmentation of large-scale point cloud images. RandLA-Net can process 1 million points at a time, and the speed is greatly improved, which surpasses many advanced speech segmentation models on Semantic3D and SemanticKITTI data sets.

Transformer was first used in natural language processing. Unlike CNN, which relies on convolution and recursion to obtain information *(Vaswani et al., 2017)*. Transformer is completely based on multiple attention mechanism to obtain global context information, surpassing RNN *(Yin et al., 2017)*, Deep RNNs *(Shi et al., 2017)*, BiRNN *(Johansen et al., 2017)* and other models with excellent performance in natural language processing at that time. Because Transformer has excellent performance in natural language processing and its framework is different from the traditional convolution neural network, many visual attempts of Transformer framework began to emerge. VIT *(Dosovitskiy et al., 2020)* ushered in the era of the application of Transformer to computer vision, and achieved better results in image classification tasks using multiple small and medium-sized data sets for image recognition benchmark testing with fewer computing resources. The subsequent DETR *(Carion et al., 2020)* algorithm completes the application of Transformer in the field of target detection, and has a very excellent effect compared with the target detection algorithm based on CNN in the same period. Therefore, Transformer is widely used in the field of computer vision, in which SETR *(Zheng et al., 2021)* and Segformer *(Xie et al., 2021)* are the representatives of Transformer in semantic segmentation algorithms. SETR regards image semantic segmentation as a sequence-to-sequence prediction task, uses Transformer global context modelling, provides a powerful semantic segmentation model, and has achieved excellent results on data sets such as Cityscapes. Segformer unifies Transformer and lightweight MLP decoder, which can output multi-scale features and achieve local and global attention at the same time, and has a more powerful representation. Segformer verifies its strong semantic segmentation ability on multiple data sets. *Jin et al. (2021)* designed a new semantic segmentation network TrSeg based on Transformer, which can adaptively capture global context information in multi-scale. Several groups of experiments show that TrSeg has better performance than other multi-scale models to obtain information. *Gu et al. (2022)* proposed a semantic segmentation model HRViT, which integrates high-resolution multi-branch architecture and ViT. This model reduces the redundancy in the linear layer and increases the attention blocks with enhanced expressiveness. Finally, it is proved that the model has great potential in semantic segmentation on multiple semantic segmentation data sets. *Lu et al. (2021)* adopted a different way of building semantic segmentation than the traditional way of building semantic segmentation, put encoder-decoder into pre-training, only put classifiers into meta-learning, and designed a new classifier: Classifier Weight Transformer (CWT). CWT can dynamically adjust the weight of the classifier in an inductive way when the support set is put into training, so as to adapt to each query image. Although the model is simple, its performance is much better than other models through experiments.

With the development of deep learning, the model of image semantic segmentation becomes more efficient and accurate, which provides technical support for autopilot and panoramic analysis of roads. In this paper, the road environment image in the greenhouse is used to prepare the data set, and SETR model, which is the most mature and leading semantic segmentation model in Transformer framework, and DeepLabv3+ model, which also has excellent performance in convolution neural network, are selected to segment the self-built greenhouse road data set, and the feasibility of semantic segmentation of greenhouse road is verified. It provides a theoretical basis for automatic driving of intelligent agricultural machinery in the greenhouse.

MATERIALS AND METHODS

Network framework of SETR

The design of encoder in SETR *(Zheng et al., 2021)* takes Transformer as the main body and is composed of layer Normalization, multi-head attention layer (Multi-head self-attention MSA) and multi-layer perceptron (Multilayer Perceptron MLP) blocks. Because Transformer accepts the input of one-dimensional sequence embedded by features (where L is the length of the sequence and C is the size of the hidden channel), the image of H × W needs to be segmented into grids containing multiple slices p of 16 × W hand 16, and each slice p is mapped into a sequence with dimension C through the linear projection (linear mapping) function. In order to encode the spatial information of the slice, the literature learns a specific embedded pi for each location I to get the final sequence as input $E = \{e_1 + p_1, e_2 + p_2, ..., e_L + p_L\}$.

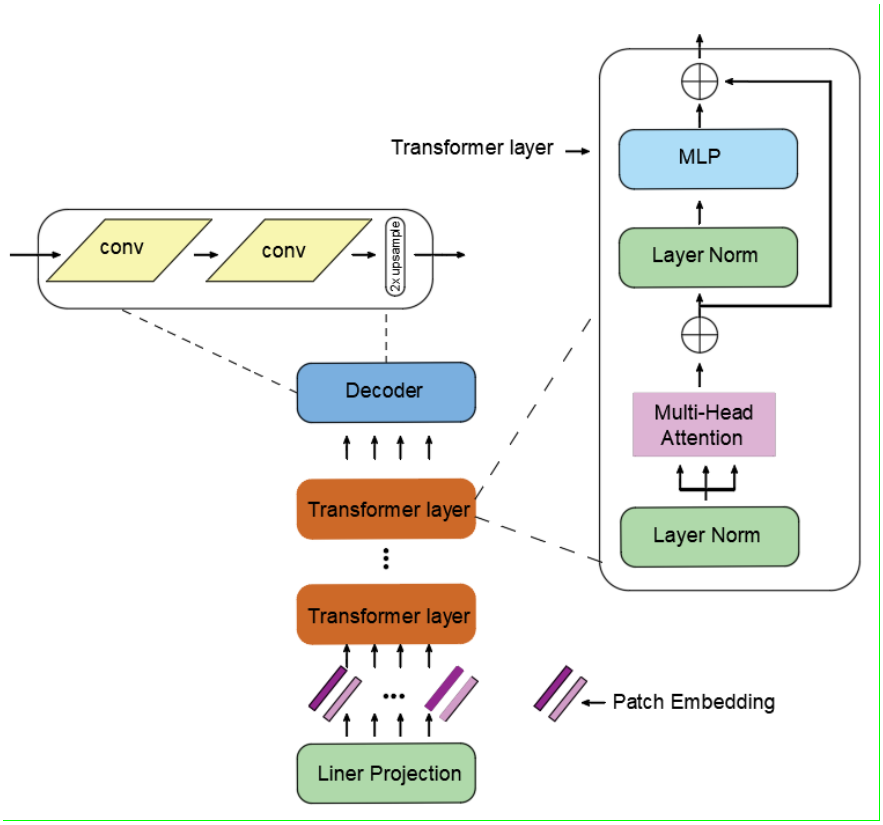


Fig. 1 - The structure of SETR Model

In reference *Zheng et al., (2021)*, three modules are designed for Decoder: Naive upsampling (Naive), Progressive UPsampling (PUP) and Multi-Level feature Aggregation (MLA). Among the three kinds of Decoder, the architecture of Naive is the simplest. Naive uses a two-layer network to map features to the number of target categories, and then directly uses bilinear upsampling to get the original image resolution. In order to avoid excessive noise, PUP returns to the original image resolution by gradually double up sampling. MLA uses a multi-level feature fusion strategy similar to the pyramid feature fusion strategy, and in order to enhance the interaction between different layers of features, it uses the method of down-to-down fusion (element-wise addition), and finally 4x bilinear up-sampling back to the original image resolution. The SETR model takes up more computing resources. Due to the limitations of the experimental hardware, this paper chooses Naive, which has the simplest architecture, as the Decoder of the SETR model. The structure of the SETR model is shown in Figure 1.

Network framework of DEepLabV3+

As a continuation of DeepLabv3 semantic segmentation model, Deeplabv3+ *(Chen et al., 2018)* has better ability to recover object edge information and takes up less computing resources. Deeplabv3+ uses Deeplabv3 as encoder, so DeepLabv3+ can use hole convolution (Atrous convolution) to extract deep convolution neural network to calculate features at any resolution, and obtain more informative features by changing the size of encoder output stride. Deeplabv3+ proposes a simple and efficient decoder. Firstly, the output features of the encoder are 4 times bilinear up-sampled and connected with the low-level features of the backbone network, and the training difficulty is reduced and the features are refined through multiple convolution layers. Finally, a simple 4-fold bilinear up-sampling is performed to return to the original image resolution. On the backbone network, Deeplabv3+ adopts Xception *(Chollet, 2018)*, which is faster and more effective in image classification, and replaces all the maximum pool layer (max pooling) in Xception with deep separable convolution layer (depthwise separable convolution). The structure of the Deeplabv3+ network is shown in Figure 2.

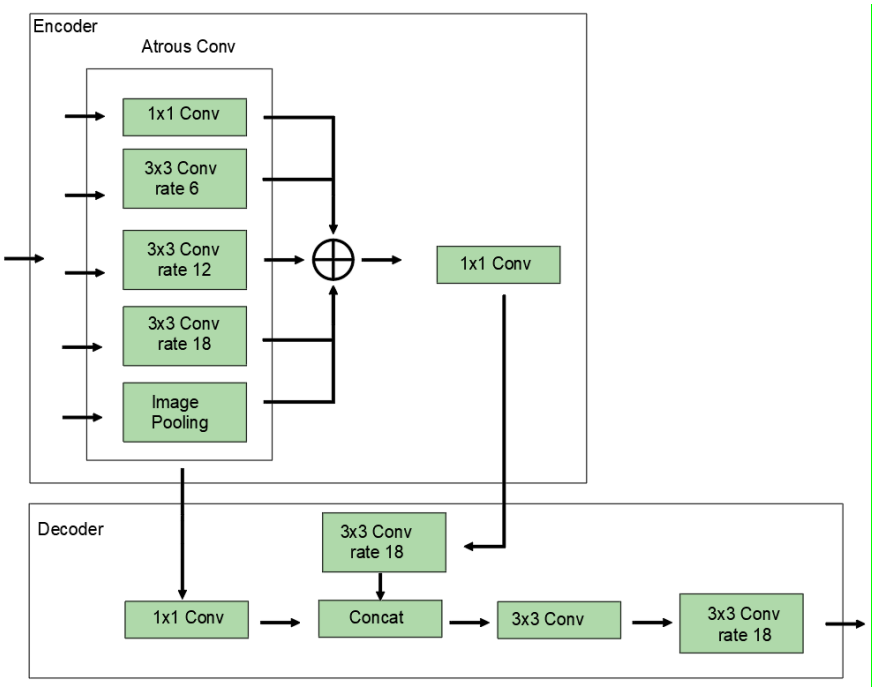


Fig. 2 - The structure of Deeplabv3+ Network

Image acquisition of datasets

The image of establishing the data set is collected in a tomato greenhouse in city of Hohhot. The captured image is shown in Figure 3.

Fig. 3 - Road images collected in greenhouse

Fig. 4 - Image acquisition equipment

The image is divided into a greenhouse corridor road and an inter-crop road. The tracked intelligent car used in the image acquisition equipment is equipped with an Astra Pro Plus camera, and the image acquisition equipment is shown in Figure 4. Control the image acquisition equipment to move in the greenhouse, simulate the moving road condition of intelligent agricultural machinery in the greenhouse, and record the road and around the road when moving in real-time. The images are collected repeatedly in multiple greenhouse corridors and roads among crops to enrich the data set. One image is extracted from the real-time recorded video every 12 frames, and a total of 1400 field road images are collected by manual screening, of which 1000 are used as a training set, 200 as verification set and 200 as a test set.

Image calibration of datasets

Using LabelMe tool to calibrate the collected data sets, the calibrated categories are: intercrop road-road1, greenhouse corridor-road2, greenhouse crops-crop, tomato-tomato on crops, plastic film-background1 in greenhouse, earthen wall in greenhouse-background2, brick wall in greenhouse-background3, which are divided into 7 categories, and the contours of all kinds of targets in the image are depicted. The calibrated image is generated into a label image, in which both the original image and the label image are RGB images, and the sizes of the three channels of each category are: road1 [128,0,0], road2 [128,128,128], crop [0,128,0], tomato [227,139,6], background1 [0,0,128], background2 [128,128,0], background3 [128,0,128]. The original image and its label image are shown in Figure 5. The calibrated training set and verification set are used in the model training, and the uncalibrated test set is not put into the model training to test the prediction ability of the model to the road in the greenhouse.

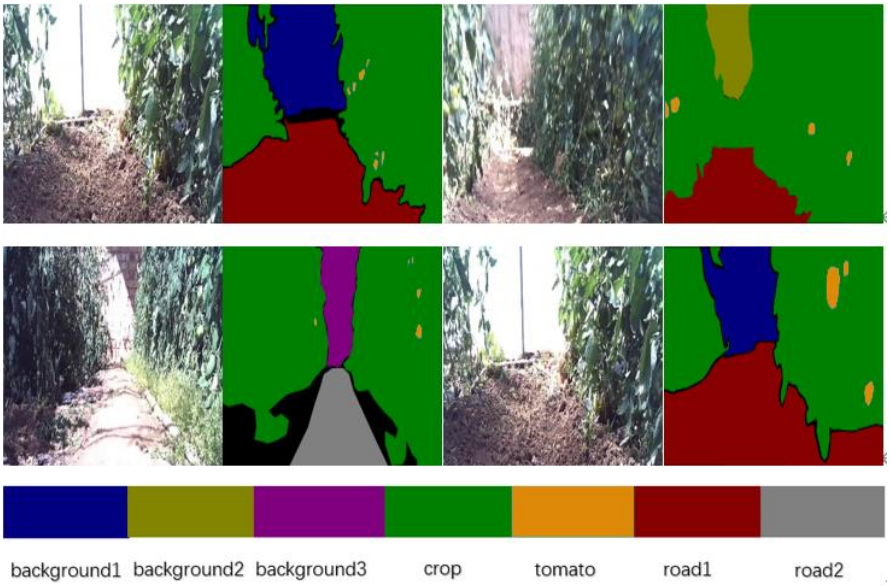


Fig. 5 - Road image and its label image

RESULTS

Experimental environment and evaluation indicators

In this paper, PyTorch deep learning framework is used to build SETR and Deeplabv3+ models, and the development language is Python3.9. Five NVIDIA Tesla V100 GPU 32GB were used to train the SETR and Deeplabv3+ models. In the training strategy, random gradient descent (SGD) is used as the optimizer, the initial learning rate is set to 0.001, the training cycle (epoch) is set to 500, and the learning rate decreases by 10% every 50 training cycles, and five images per batch (batch) are input into the model for training. The evaluation index selects pixel accuracy (Pixel Accuracy, PA) and average intersection ratio (Mean Intersection over Union, MIoU). The effect of the model is evaluated by the score of the two models on the two evaluation indexes. The evaluation index formula is as follows:

$$PA = \frac{\sum_{i=0}^{N} p_{ii}}{\sum_{i=0}^{N}\sum_{j=0}^{N} p_{ij}} \quad (1)$$

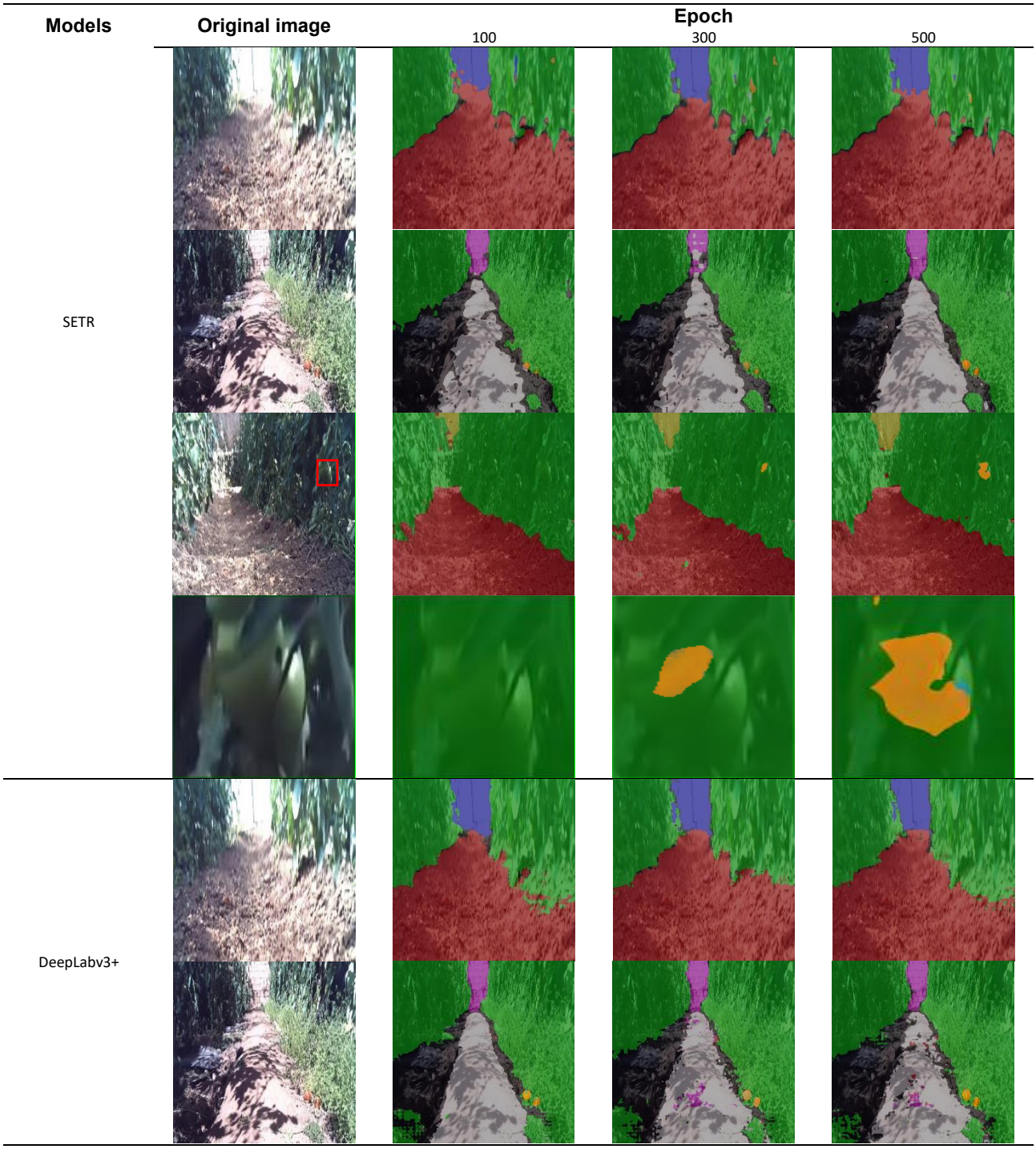
$$mIoU = \frac{1}{N+1}\sum_{i=0}^{N}\frac{p_{ii}}{\sum_{j=0}^{N}p_{ij}+\sum_{j=0}^{N}p_{ji}-p_{ii}} \quad (2)$$

In the formula N is the number of categories, p_{ii} is the real case, p_{jj} is the true negative case, p_{ji} is the false negative case, p_{ij} is the false positive case.

Analysis of the training effect of two models

The prepared training set and verification set are put into the built SETR model and DeepLabv3+ model for training, respectively, and the training effect of the model is observed at 100,300,500 epochs.

The semantic segmentation effect of the model under the three training cycles is shown in Figure 6. The training effect of the model is evaluated according to PA and mIoU evaluation indicators, as shown in Table 1.



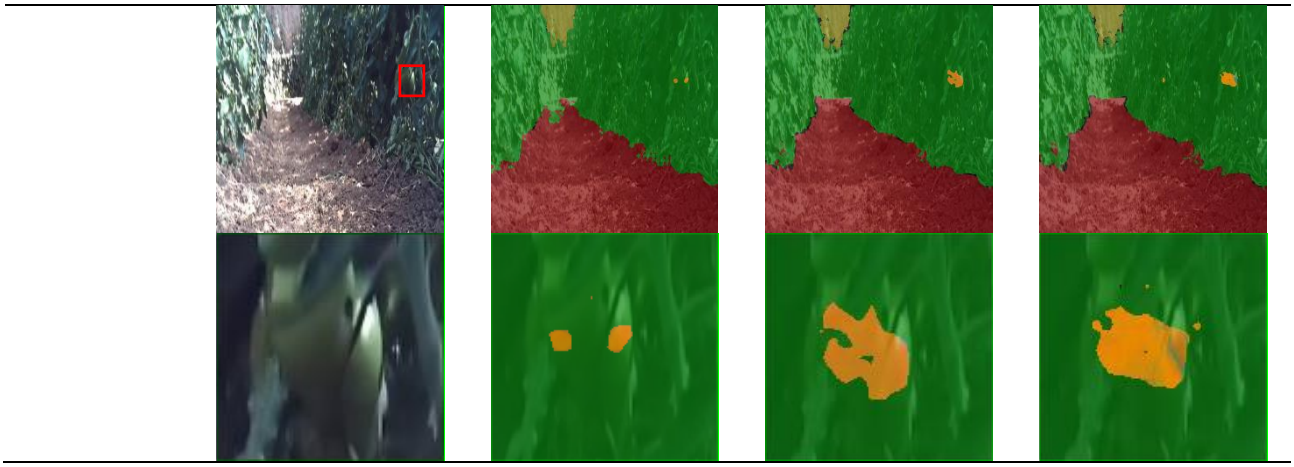


Fig. 6 - The Semantic Segmentation effect of each training epoch of the two models

According to the analysis of Figure 6, both SETR and DeepLabv3+ models can segment the semantics of each target in the greenhouse road image after 100 times of training, and the target contour is basically consistent with the actual contour, but there are many phenomena of segmentation errors, especially the edge between the targets is not well segmented, and the small target tomato in the greenhouse road image is not well segmented. After 300, 500 times of training, the semantic segmentation ability of SETR model to the field road image is gradually enhanced, and the segmentation error phenomenon is gradually reduced. After 500 times of training, the model can segment the edge of the target very well, and can complete the segmentation of small target tomatoes. After 300, 500 times of training, the semantic segmentation ability of DeepLabv3+ model for the greenhouse road image does not change much, especially for greenhouse corridor-road2, the semantic segmentation effect is poor, there are still big segmentation errors, the edge segmentation between multiple targets is not clear enough, and there is interference between targets, so the semantic segmentation of small target tomato can be completed. Although the two models can complete the semantic segmentation of tomatoes in the field road after 500 times of training, the segmented tomato contours are not completely consistent with the actual contours and some tomatoes cannot be recognized. It should be due to the small number of tomato targets in the image and too little semantic information of the tomato in the greenhouse road data set.

Table 1

Evaluation index scores of each training epoch of the two models

Models	Epoch					
	100		300		500	
	PA / %	mIoU / %	PA / %	mIoU / %	PA / %	mIoU / %
SETR	93.78	81.34	94.95	86.98	96.46	90.45
DeepLabv3+	92.71	80.69	93.37	82.55	93.34	81.55

The analysis of Table 1 shows that after 100, 300, 500 times of training, the PA and mIoU values of the SETR model are gradually increasing. Combined with the analysis of Figure 7, it shows that the semantic segmentation ability of the SETR model to the greenhouse road image is enhanced, and the PA and mIoU of the SETR model reach 96.46% and 90.45% respectively after 500 times of training. After 100, 300, 500 times of training in the DeepLabv3+ model, the values of PA and mIoU of the model did not change significantly, the PA of the model was about 93.3%, and the mIoU was about 81.5%, indicating that the DeepLabv3+ model did not improve its ability of semantic segmentation of field roads with the increase of training times, which was consistent with the analysis of Figure 7. In the case of 500 times of training of the two models, the two evaluation indexes of the SETR model were much higher than those of the DeepLabv3+ model, with PA 3.12% higher and mIoU higher by 8.9%.

To sum up, both SETR model and DeepLabv3+ model can complete the task of semantic segmentation of greenhouse road images after different times of training. SETR model has significantly improved the ability of semantic segmentation of greenhouse road images after different times of training, while DeepLabv3+ model has not significantly improved the ability of semantic segmentation of greenhouse road images after many trainings. And after 500 times of training, the semantic segmentation ability of SETR model is obviously better than that of DeepLabv3+ model.

Analysis of prediction results of two models

In order to test the generalization ability of the two models and whether there is a fitting phenomenon in the two models, the untrained test set is taken as the prediction object, and the trained SETR model and Deeplabv3 model are used to predict the test set. The semantic segmentation effect of the two models is shown in Figure 7, and the prediction ability of the model is evaluated by three evaluation indexes of PA, mIoU, as shown in Table 2.

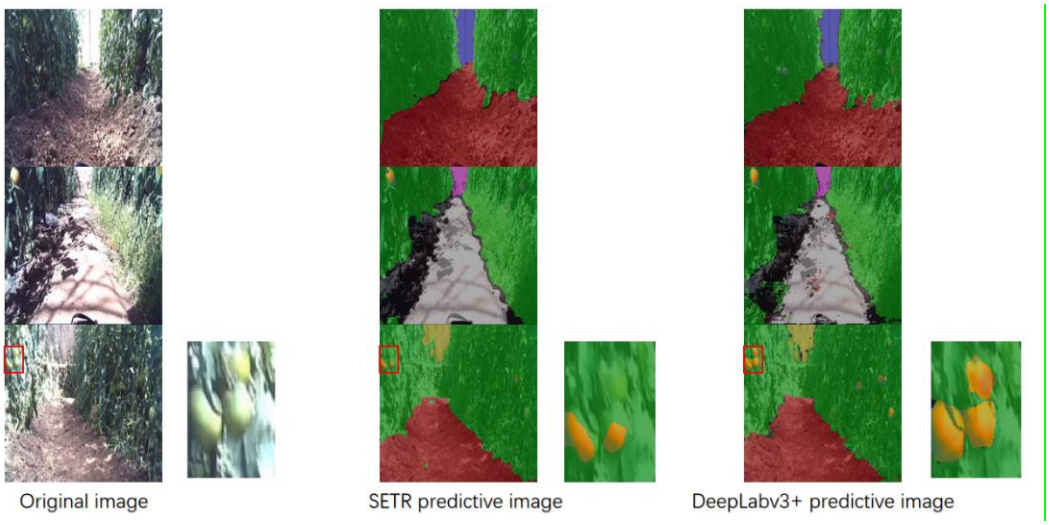


Fig. 7 - Semantic Segmentation effect of two models on Test set

Table 2

Evaluation Index score of semantic Segmentation of Test set by two models

Models	Evaluation index	
	PA/%	mIoU/%
SETR	94.64	82.72
DeepLabv3+	90.80	72.35

According to the analysis of Figure 7, SETR model and DeepLabv3+ model predict the test set, and realize the semantic segmentation of each target in the greenhouse road image. The target contours predicted by SETR model and DeepLabv3+ model is basically consistent with the actual contours of the target, but there are also some errors. Generally speaking, the effect of SETR model on image semantic segmentation is higher than that of DeepLabv3+ model. The image predicted by SETR model has fewer errors, the edges between targets are clearer, and the predicted target contours are more consistent with the actual contours of targets. However, the semantic segmentation effect of SETR model on small target tomatoes is slightly worse than that of DeepLabv3+ model.

It can be seen from Table 2 that SETR model and DeepLabv3+ model have high scores on the performance of semantic segmentation of greenhouse roads. Among them, the PA of SETR model reaches 94.64%, the PA of SETR model reaches 82.72%, the PA of DeepLabv3 + model reaches 90.80%, and the score of 72.35% SETR model is higher than that of DeepLabv3+ model in all indicators of semantic segmentation of greenhouse road data sets. Compared with Table 2 and Table 1, both SETR model and DeepLabv3 model have some overfitting in the semantic segmentation of greenhouse road data sets. Compared with the SETR model, the PA decreased by 1.82% and the MIoU decreased by 7.73%. Compared with the DeepLabv3+ model, the PA decreased by 2.54% and the SETR model decreased by 9.2%. The overfitting problem of the DeepLabv3 + model is slightly more serious than the DeepLabv3 + model. There is a certain overfitting problem in the two models, which should be caused by the insufficient number of images in the data set, and the images are subjective by manual calibration, and there is a certain deviation between the calibrated target contour and the real contour, thus affecting the semantic segmentation ability of the model to the image. According to the score of semantic segmentation of each excellent model in each open data set, the performance of semantic segmentation of greenhouse road data set by SETR model and DeepLabv3+ model can meet the requirements of actual deployment.

CONCLUSIONS

In order to realize the automatic driving of agricultural machinery in the greenhouse, it is necessary to analyse the road environment of the greenhouse. Semantic segmentation can achieve pixel-level classification of images and panoramic analysis of the environment, which is an important support for autopilot. Semantic segmentation of greenhouse road image can provide agricultural machinery with road information, crop information and other environmental information, which is an important support for agricultural machinery to perceive the changing greenhouse environment.

In this paper, according to the demand of automatic driving of agricultural machinery in the greenhouse, a tracked intelligent car with camera is used as image acquisition equipment and several images of greenhouse road are collected. After the manual calibration of the greenhouse image, a greenhouse road data set is established. Two models of SETR based on Transformer framework and DeepLabv3+ based on convolution neural network are built, and several groups of semantic segmentation experiments are carried out on the self-built greenhouse road data set. The two models have accomplished the task of semantic segmentation of greenhouse images. SETR model achieves PA94.64%, mIoU82.72%, DeepLabv3+ model achieves PA90.80%, mIoU72.35%, SETR model has higher semantic segmentation performance than DeepLabv3+ model in greenhouse road image. In terms of the semantic segmentation performance of the greenhouse road image, the two models can meet the deployment requirements of automatic driving of agricultural machinery in the greenhouse. This paper realizes the semantic segmentation of the road data set of self-built greenhouse through two semantic segmentation models of DeepLabv3+ and SETR, accomplishes one of the important tasks of automatic driving of agricultural machinery in greenhouse, and provides an important theoretical support for agricultural machinery to realize automatic driving of greenhouse.

REFERENCES

[1] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. (2020). End-to-end object detection with transformers. *In Computer Vision – ECCV 2020* (Springer International Publishing), Vol. 12346, pp. 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[2] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2018). Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 40 issue 4, pp. 834-848. https://doi.org/10.1109/tpami.2017.2699184

[3] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *In Proceedings of the European conference on computer vision (ECCV)*, pp. 801-818. https://doi.org/10.1007/978-3-030-01234-2_49

[4] Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2017.195

[5] Daduna, J R. (2020). Automated and autonomous driving in freight transport-opportunities and limitations. *In Computational Logistics: 11th International Conference, ICCL 2020,* vol. 12433, pp. 457-475. https://doi.org/10.1007/978-3-030-59747-4_30

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929

[7] Fan, M. Y., Lai, S. Q., Huang, J. S, Wei X. M., Chai Z. H., Luo J. F., Wei X. L. (2021). Rethinking BiSeNet for real-time semantic segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9716-9725. https://doi.org/10.1109/cvpr46437.2021.00959

[8] Gu, J., Kwon, H., Wang, D., Ye, W., Li, M., Chen, Y. H., Pan, D. Z. (2022). Multi-scale high-resolution vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 12094-12103. https://doi.org/10.1109/cvpr52688.2022.01178

[9] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. *In Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969. https://doi.org/10.1109/iccv.2017.322

[10] Hu, Q., Yang, B., Xie, L., Rosa, S., Markham, A. (2020). RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 11105-11114. https://doi.org/10.1109/cvpr42600.2020.01112

[11] Jin, Y, Han, D, Ko, H. (2021). TrSeg: transformer for semantic segmentation. *Pattern Recognition Letters*, vol. 148, issue 4, pp. 29-35. https://doi.org/10.1016/j.patrec.2021.04.024

[12] Johansen, A. R., Sønderby, C. K., Sønderby, S. K., & Winther, O. (2017). Deep recurrent conditional random field network for protein secondary prediction. *In Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics,* pp. 73-78.

[13] https://doi.org/10.1145/3107411.3107489

[14] Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440. https://doi.org/10.1109/cvpr.2015.7298965

[15] Lu, Z., He, S., Zhu, X., Zhang, L., Song, Y. Z., Xiang, T. (2021). Simpler is better: Few-shot semantic segmentation with classifier weight transformer. *In Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 8741-8750. https://doi.org/10.1109/iccv48922.2021.00862

[16] Ren, F. L., Yang, L., Zhou H., B., He X., Xu, W. X. (2023). Real-time semantic segmentation based on improved BiSeNet (基于改进BiSeNet的实时图像语义分割). *Optics and Precision Engineering*, vol. 31, issue 8, pp. 1217-1227. https://doi.org/10.37188/OPE.20233108.1217

[17] Shi, H, Xu, M, Li, R. (2017). Deep learning for household load forecasting – a novel pooling deep RNN. *IEEE Transactions on Smart Grid,* vol. 9, issue. 5, pp. 5271-5280. https://doi.org/10.1109/tsg.2017.2686012

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

[19] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, issue 10, pp. 3349-3364. https://doi.org/10.1109/tpami.2020.2983686

[20] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, pp. 12077-12090. https://doi.org/10.48550/arXiv.2105.15203

[21] Xiong, W., Tong, L., Jin, J. Y., Wang, C. S., Wang, J., Zeng, C. Y. (2021). Research on semantic segmentation algorithm based on convolutional neural network (基于卷积神经网络的语义分割算法研究). *Application Research of Computers*, vol. 38, issue 4, pp. 1261-1264. https://doi.org/10.19734/j.issn.1001-3695.2019.12.0705

[22] Yin, W., Kann, K., Yu, M., Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*. https://doi.org/10.48550/arXiv.1702.01923

[23] Yoon, J. Y., Jeong, J., Sung, W. (2022). Design and implementation of HD mapping, vehicle control, and V2I communication for robo-taxi services. *Sensors*, vol. 22, issue 18, pp. 7049. https://doi.org/10.3390/s22187049

[24] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N. (2021). BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, vol. 129, pp. 3051-3068. https://doi.org/10.1007/s11263-021-01515-2

[25] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N. (2018). BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV),* pp. 325-341. https://doi.org/10.1007/978-3-030-01261-8_20

[26] Zheng, S. X., Lu, J. C., Zhao, H. S., Zhu, X. T., Luo, Z. K., Wang, Y. B., Fu, Y. W., Feng, J. F., Xiang, T., Philip, H. S. Torr, Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 6881-6890. https://doi.org/10.1109/cvpr46437.2021.00681

[27] Zhongkun, H. (2020). Analysis of visual navigation extraction algorithm of farm robot based on dark primary colour. *INMATEH-Agricultural Engineering*, *62*(3). https://doi.org/10.35633/inmateh-62-23