**Image Segmentation using Deeplabv3+**

**Review 1:**

Aiming at the problem that the deeplabv3+ model is not accurate in segmentation of the image target edge, the image feature fitting is slow, and the attention information cannot be effectively used. This paper proposes an improved Deeplabv3+ network that can be effectively applied to the semantic segmentation task of SAR images. It is proposed to add a feature cross attention module (FCA) to the model. The cross-attention network is composed of two branches and a feature cross attention module. Among them, the shallow branch is used to extract low-level spatial information, and the deep branch is used to extract high-level context features to make important feature extraction more refined. This paper designs and realizes the connection between Feature Cross Attention module and Deeplabv3+ coding module, input the output features of the Deeplabv3+ encoding module into the feature cross attention module for convolution operation to realize the recalibration of the original features. The decoding module of Deeplabv3+ obtains spatial features and channel features from two branches respectively, and then merges the obtained features to obtain more important features.

**Reference:**

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; IEEE: Piscataway, NJ, USA, 2015; pp. 3431–3440.

**Review 2 :**

Semantic segmentation of SAR images is a basic and important problem in remote sensing image interpretation. Its purpose is to assign a category label to each pixel in SAR images. Due to their all-weather and rich feature information, SAR images have received special attention from scholars at home and abroad in recent years which is of great significance to promote the development of SAR image processing technology. With the rapid development of Deep Learning, Deep Convolutional Neural Networks (DCNN) are vital for feature extraction and characterization. Fully Convolutional Networks (FCN) based on the emergence of the end-to-end classical semantic segmentation model, have achieved great success. However, at the same time, due to the fixed network structure, FCN also reveals many disadvantages. Without considering the global context information, the sampling of the feature map will be restored to the image size of the original image, resulting in inaccurate pixel positioning. Ronner berge proposed a U-Net network for biomedical image segmentation. the experimental results of this article are listed, including FPPM model parameter determination experiments, ablation experiments, and results on synthetic and SAR images. Compared to the original Deeplabv3+ network, the segmentation accuracy has been improved by 3.64% and mIoU improved by 2.1%, shortening the training time by 19 ms.by 3.64% and mIoU improved by 2.1%, shortening the training time by 19 ms.

**Reference:**

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; Springer: Cham, Switzerland, 2015; pp. 234–241.

**Review 3:**

This paper tried to address the issue of semantic segmentation of dark images using a proposed framework with three Deeplab v3 variations. Though the approach is technically sound, the experiments seem to be weak with such a small dataset. Please see the comments below:

It may not be necessary to give such a detailed explanation of what semantic segmentation is (lines 144- 175), since it should be commonly known by the readers. The reviewer appreciates the detailed critical analysis of existing works in Table It seems like the proposed parallel DCNN shown in Figure 2 is a standard model ensembling process with merging the labels at the end. The images were collected in sequence with a large overlap between each other. The random selection of training and testing sets may cause the issue of overfitting the network which increases its performance since one image in the testing set may have a large overlap with the image in the training set. The experiments showed a performance improvement on the image dataset.

**Reference:**

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890

**Review 4:**

This section provides an overview of some of the most prominent deep learning architectures used by the computer vision community, including convolutional neural networks(CNN). It is worth mentioning that in some cases it can be trained on applications/datasets (assuming a sufficient quantity of labeled training data), but in many cases there are not enough labeled data available to train a model from scratch and one can use transfer learning to tackle this problem. In image segmentation case, many people use a model trained on imagenet the higher-level layers learn features from increasingly wider receptive fields.

**Reference:**

R. Szeliski, Computer vision: algorithms and applications. Springer Science & Business Media, 2010.

**Review 5:**

This work is considered a milestone in image segmentation, demonstrating that deep networks can be trained for semantic segmentation in an end-to-end manner on variable size images. However, despite its popularity and effectiveness, the conventional FCN model has some limitations it is not fast enough for real-time inference, it does not take into account the global context information in an efficient way, and it is not easily transferable to 3D images. Several efforts have attempted to overcome some of the limitations of the FCN. A CNN+CRF model. The coarse score map of CNN is sampled via interpolated interpolation,and fed to a fully connected CRF to refine the segmentation result. FCNs have been applied to a variety of segmentation problems, such as brain tumor segmentation instance aware semantic segmentation skin lesion segmentation and iris segmentation.

**Reference:**

H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890

**Review 6:**

Another popular family of deep models for image segmentation is based on the convolutional encoder-decoder architecture. There are several models initially developed for medical/biomedical image segmentation, which are inspired by FCNs and encoder-decoder models. U-Net [49], and V-Net [50] are two well-known such architectures, which are now also being used outside the medical domain. Ronneberger et al. proposed the U-Net for segmenting biological microscopy images. Their network and training strategy relies on the use of data augmentation to learn from the very few annotated images effectively. The U-Net architecture comprises two parts, a contracting path to capture context, and a symmetric expanding path that enables precise localization and pixel of the input image.

**Reference:**

G. Song, H. Myeong, and K. Mu Lee, “Seednet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1760–1768.

**Review 7:**

3D image datasets are popular in robotic, medical image analysis, 3D scene analysis, and construction applications. Three dimensional images are usually provided via meshes or other volumetric representations, such as point clouds. Here, we mention some of the popular 3D datasets. Stanford 2D-3D: This dataset provides a variety of mutually registered modalities from 2D, 2.5D and 3D domains, with instance-level semantic and geometric annotations, and is collected in 6 indoor areas. It contains over 70,000 RGB images, along with the corresponding depths, surface normals, semantic annotations, global XYZ images as well as camera information. This dataset contains a variety of common urban road objects, collected in the central business district of Sydney, Australia. There are 631 individual scans of objects across classes of vehicles pedestrians, signs and trees.

**Reference:**

Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, “Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3029–3037.

**Review 8**:

To improve the segmentation performance of grapevine leaf black rot spots, a deep learning network based on the DeepLab v3+ was constructed. It is the third version of DeepLab, with high segmentation effectiveness and speed. In the improved DeepLab v3+ network constructed in this paper, the residual part in the backbone network ResNet101 incorporates a plugand-play attention mechanism module. This improves the performance of various CNNs without increasing the complexity of the model. Moreover, a feature fusion branch based on a feature pyramid network (FPN) was added to the DeepLabv3+ encoder, which performs feature fusion on high-resolution and low-resolution feature maps. Finally, in the improved DeepLab v3+, one 4-fold up-sampling is replaced with two 2-fold up-sampling. Furthermore, the continuity of pixels in the obtained images is stronger and the network segmentation effect is improved. The, two fully-connected layers capture the non-linear cross-channel interaction. Finally,a Sigmoid function generates the channel weights with a value between 0 and 1. This weight is added to the feature channel as a weight to generate the next level of input data.

**Reference:**

Appeltans, S., Pieters, J. G., and Mouazen, A. M. (2021). Detection of leek white tip disease under field conditions using hyperspectral proximal sensing and supervised machine learning. Comput. Electron. Agric. 190:106453. doi: 10.1016/j.compag.2021.106453

**Review 9:**

The training dataset with annotation information was fed into the improved DeepLab v3+ network for training. The network was trained for 120 epochs, which required around 8.3h. During the training process, the training model was saved once every 1 epoch, and a total of 120 completed models were saved. The convergence of the model can be  reflected by the loss values generated during the training process. It shows the changes in the loss values of the training data and validation data in the training set during the training process. The training loss and validation loss gradually converged to stability during the training process. This also demonstrates that the use of deep learning methods can reduce subjective errors caused by manual segmentation. The red markers indicate that the leaf edges were misidentified as spots and segmented by the network model due to shadows. Experiments with the Plant Village dataset demonstrated that the improved DeepLab v3+,which incorporates an attention mechanisms property.

**Reference:**

Zhao, Z.; Chen, K.; Yamane, S. CBAM-Unet++:easier to find the target with the attention module “CBAM”. In Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Kyoto, Japan, 12-15 October 2021

**Review 10:**

In this paper, the experiments were carried out on some of the improvements proposed to the paper and the experimental results proved its effectiveness. First, to verify the effectiveness of the CA attention mechanism module, the Deeplabv3+ model, equipped with the CA module, was compared with the original model through experiments. It was verified through the analysis of experimerntal results. Then, in order to prove the applicability of the focal loss function, a comparative experiment was carried out with the Deeplabv3+ model based on the cross-entropy loss function. The experimental results show that the focal loss function can significantly improve the segmentation effect compared with the cross-entropy loss function in data set segmentation with uneven samples. Secondly, the influence of the improved ASPP module proposed in this paper on the performance of the Deeplabv3+ model is compared and based on the above improvements, the effectiveness of the FPPM proposed in this paper on SAR image semantic segmentation is further verified. In the Table 3, we have listed all the changes to the Deeplabv3+ network.

**Reference:**

Linsley, D.; Dan, S.; Eberhardt, S.; Serre, T. Learning what and where to attend. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018