DO-UNET, DO-LINKNET: UNET, D-LINKNET WITH DO-CONV FOR THE DETECTION OF SETTLEMENTS WITHOUT ELECTRICITY CHALLENGE

Ruoxian Feng, Mengjiao Wang, Xuanming Zhang, Jun Zhang, Licheng Jiao, Xu Liu, Fang Liu

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education,
School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi Province 710071, China

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9553097

ABSTRACT

In this paper, two semantic segmentation models, DO-UNet and DO-LinkNet, are presented for the detection of human settlements, and a threshold-based model is proposed to detect areas with electricity. In DO-UNet and DO-LinkNet, the conventional convolutional layer is replaced with depthwise over-parameterized convolutional layer. Also, an extra pooling operation is carried out in the last layer since the size of the input images is different from that of the labels. Depthwise over-parameterized convolutional layer enhances the convolutional layer with an additional depthwise convolution. Pooling operation can accelerate training speed, increase the receptive field in feature extraction, and reduce the requirement of network complexity. In the detection of settlements without electricity challenge track, our best F1-score on the validation set and the test set are 0.8820 and 0.8798, respectively.

Index Terms— Semantic Segmentation, DO-UNet, DO-LinkNet, DO-Conv

1. INTRODUCTION

Extracting the information we need from satellite images has been a hot research topic recently and has a wide range of applications. In the Track DSE of 2021 IEEE GRSS Data Fusion Contest [1], a binary classification task is defined: each pixel is marked as either human settlements without electricity or other class. In this paper, we treat this task as a binary semantic segmentation task to generate pixel-level markers for regions.

In recent years, Convolutional Neural Network has shown advantages in the field of image semantic segmentation. Fully Convolutional Network [2], a landmark work, extends end-to-end convolutional networks to semantic segmentation, and its fully connected layer structure is still used in the most advanced segmentation models.

UNet [3] modifies and expands the network architecture of Fully Convolutional Network, and proposes the overlap-tile strategy, which effectively improves the training performance on small datasets. It has a good performance in both medical image processing and natural image understanding tasks. Later, SegNet, DeepLab series, PSPNet, as well as some lightweight networks such as LinkNet also have a good performance.

In Track DSE, the given dataset is composed of 98 tiles of 800×800 pixels. Each tile includes 98 channels from four different satellites images. Thus, choosing appropriate data as the input of the model is of great importance. Since the size of the labels is different from that of the input images, how to make the network extract the required information needs to be fully considered. Besides, the information on satellite images is often more complex, and the area we need to take into account is only a small part of the image, in which case, it is important to retain detailed spatial information. Based on the above problems, we propose two semantic segmentation networks, DO-UNet and DO-LinkNet, which are proved to having a good performance in this contest.

Firstly, for Sentinel-2 and Landsat 8 multispectral dataset, multi-channel fusion is carried out on different bands and visualized with The Environment for Visualizing Images (ENVI). Then, the visualized images of diffrent combination of channels are tested on UNet. After several comparative experiments, the Sentinel-2 RGB channels are selected as the input data for all subsequent models, because it has a much better effect than other datasets. After the detection of settlements, detection model of areas with electricity based on the idea of threshold is carried out on The Suomi Visible Infrared Imaging Radiometer Suite (VIIRS) night time dataset. In this way, the areas with electricity are detected. Based on this, the output results of DO-UNet and DO-LinkNet are modified. Finally, we obtained 0.8820 and 0.8798 F1-score on the validation set and the test set, respectively.

2. METHOD

2.1. Data Pre-Processing

The task of Track DSE is to detect human settlements that do not have access to electricity using the given data. We detect human settlements and areas that do not have access to electricity respectively. In order to detect human settlements, ENVI, a remote sensing image processing software, is used to

Thanks to IPIU for cluster resource support

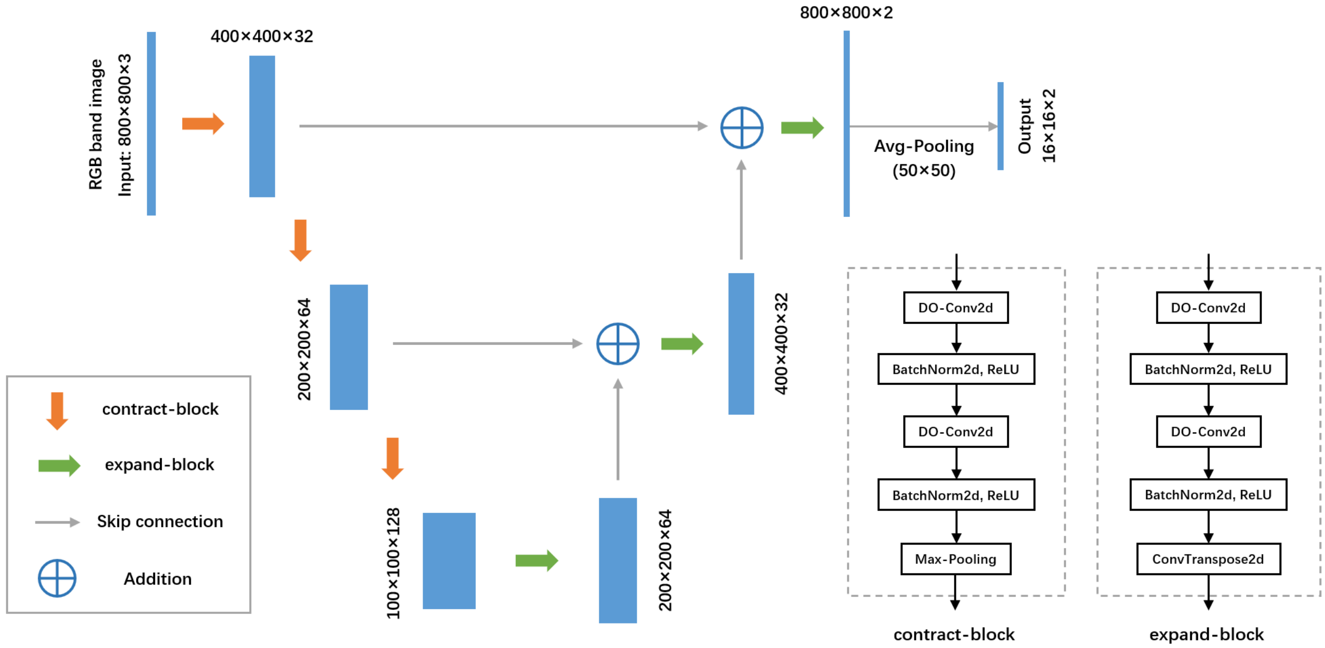

Fig. 1. DO-UNet architecture. The input to DO-UNet is the Sentinel-2 RGB channels image. DO-UNet can be divided into two parts: encoder and decoder. Contract-block and expand-block can expand the input by twice its original size and shrink it by half, respectively. In the contract-block and expand-block, the convolutional layers are replaced with DO-Conv layers, and the Batch Normalization (BN) and ReLU are added to accelerate the convergence and improve the performance of the network. In the last layer, Avg-Pooling of size 50×50 is used, and a 16×16 binary classification map is output.

fuse and visualize images of selected channels. After visualizing, RGB images with values ranging from 0 to 255 can be got. By analyzing RGB images generated by the combination of different channels and verifying their effect of detecting human settlements, RGB channels of Sentinel-2 are chosen as the input data for DO-UNet and DO-LinkNet. After selecting data without cloud, 120 training images are obtained. In addition, comparative experiments show that adding images of other channels as the input data will lead to the deterioration of model performance.

2.2. DO-Conv

Convolutional layers play a key role in the construction of convolutional neural networks. DO-Conv [4] is a new convolutional layer which combines the conventional convolution and the depthwise convolution. This combination leads to over-parameterization because of the increase in learnable parameters. The advantages of over-parameterization have been demonstrated by Jinming Cao et al.

2.3. Network Architecture

In Track DSE, the size of the input images is 800×800, while the size of the corresponding labels is 16×16. Therefore, a pooling layer with a stride of 50 is added to the last layer of the network to obtain the output images the same size as the labels. Compared with the graph cutting method, this method can accelerate the training speed of the network and reduce the requirement of network complexity.

The architecture of DO-UNet is shown in Fig.1. DO-UNet is mainly divided into two parts: encoder and decoder. The main component of the encoder are contract-blocks, and the decoder is mainly composed of expand-blocks. The main structure of the contract-block is the interconnection of DO-Conv layers, BN layers and Max-Pooling layers. BN can accelerate the convergence of the network, so as to improve the performance of the network. In addition, after the DO-Conv layer, ReLU is added. The only difference between expand-block and contract-block is that the Max-Pooling layer is changed to ConvTranspose2d to increase the size of the feature map to complete the decoding operation. Finally, an Avg-Pooling layer with a stride of 50 is added to the last layer of the network to obtain the output images the same size as the labels.

As shown in Fig.2, DO-LinkNet consists of three main parts: encoder, center part and decoder. The decoder is the same as that of LinkNet. Since the input resolution is 800×800 and the output resolution is 16×16, it is very important to enlarge the receptive field in the process of feature extraction. Therefore, dilated convolution is used in center part, with dilation rate of 1, 2, 4, 8, which is the same as the setting of D-LinkNet [5]. In order to improve the network per-

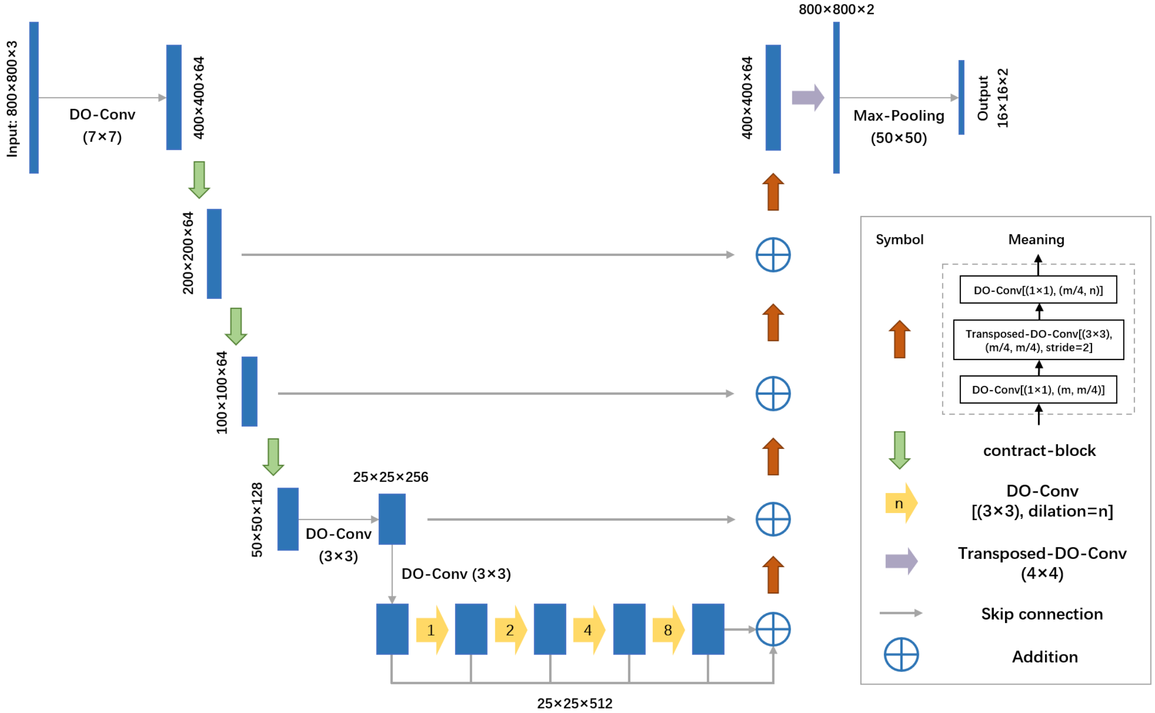

Fig. 2. DO-LinkNet architecture. DO-LinkNet is divided into three parts: encoder, center part and decoder. DO-LinkNet uses contract-block as the encoder and uses the same decoder as LinkNet. The center part can expand the receptive field, thus retaining detailed spatial information. Like DO-UNet, DO-LinkNet also replaces all convolutional layers with DO-Conv layers. Finally, Max-Pooling is carried out to obtain the 16×16 binary classification map.

formance, the conventional convolutional layers are replaced with DO-Conv layers. The result of binary classification map is obtained after 50×50 max pooling.

2.4. Detection of Areas with Electricity

The VIIRS night time dataset is able to detect areas with electricity. The 9 images corresponding to each area are taken as the input data of the detection model of areas with electricity, so there are 9 input channels. The threshold of the detection model of areas with electricity is determined by the F1-score of the detection of electricity on the VIIRS training dataset. If more than two of the nine results indicate that an area has access to electricity, the area is determined to have access to electricity.

3. EXPERIMENTS

In Track DSE, we use PyTorch as the deep learning framework. All models are trained on 4 NVIDIA GeForce RTX 2080Ti GPUs.

3.1. Dataset

The contest dataset is composed of 98 tiles of 800×800 pixels, distributed respectively across the training, validation and test sets as follows: 60, 19, and 19 tiles. Each tile includes 98 channels from four different satellites images. All the images have been resampled to a Ground Sampling Distance (GSD) of 10 m. Thus each tile corresponds to a 64km^2 area.

3.2. Implementation Details

In the process of training, we added Spatial and Channel Squeeze and Channel Excitation Block (scSE) [6] to UNet for comparative experiment. This module can enhance meaningful feature and suppress useless feature.

In order to prevent overfitting and enhance the robustness of the model, we selected the following three data augmentation methods through comparative experiments: flip, perspective and Contrast Limited Adaptive Histogram Equalization (CLAHE). During data loading, each kind of data augmentation occurs with a probability of 0.5 for each image input.

Test Time Augmentation (TTA) was applied to the test set images, including horizontal flipping, vertical flipping, and diagonal flipping. Each prediction is then averaged to generate the binary output.

On the validation set and test set, we conducted voting fusion of the results of multiple models with good performance, and modified the fusion results by using the detection model of areas with electricity to get the final predicted results. The models involved in the fusion on the test set came from the best 7 models on the validation set.

3.3. Results

The results of some single models are shown in Table 1. The results of multi-model fusion before and after modification on the test set are shown in Table 2.

Model	F1-score on validation set
UNet	0.8071
UNet①	0.8234
D-LinkNet①②	0.8418
UNet①②	0.8424
UNet①②③	0.8594
DO-LinkNet①②	0.8619
DO-UNet①②	0.8641
DO-UNet①②④	0.8687
DO-UNet①②⑤	0.8699

Table 1. F1-score on the validation set for single model with different strategies. Where, ① represents TTA, ② represents flip data augmentation, ③ represents scSE, ④ represents CLAHE data augmentation, and ⑤ represents perspective data augmentation.

Fusion results on test set	Results after modification
0.8640	0.8826
0.8591	0.8787
0.8534	0.8798

Table 2. DO-UNet and DO-LinkNet model fusion results before and after modification on the test set.

3.4. Analysis

As can be seen from Table 1, flip data augmentation, scSE, TTA, perspective data augmentation, and CLAHE data augmentation can improve F1-score by 0.0190, 0.0170, 0.0163, 0.0058 and 0.0046, respectively. DO-Conv can improve the results of UNet by 0.0217 and D-LinkNet by 0.0201.

It can be seen in Table 2 that the detection model of areas with electricity can improve F1-score by up to 0.0264.

4. CONCLUSION

In this paper, we present semantic segmentation networks DO-UNet and DO-LinkNet to detect human settlements, and a threshold-based model to identify areas with electricity. DO-Conv is used to replace the conventional convolutional layer, contributing to performance enhancements without increasing the computational complexity, while dilated convolution is used to expand receptive field and retain spatial information. A pooling layer is used to get the output images the same size as the labels, which accelerates the training speed of the network and reduces the requirement of network complexity. Experiments on the validation set and the test set show that lightweight networks DO-UNet and DO-LinkNet are very suitable for small datasets.

5. ACKNOWLEDGEMENT

The authors would like to thank the IEEE GRSS Image Analysis and Data Fusion Technical Committee, Hewlett Packard Enterprise, SolarAid, and Data Science Experts for organizing the Data Fusion Contest.

6. REFERENCES

[1] Naoto Yokoya, Pedram Ghamisi, Ronny Hansch, Colin Prieur, Hana Malha, Jocelyn Chanussot, Caleb Robinson, Kolya Malkin, and Nebojsa Jojic, "2021 data fusion contest: Geospatial artificial intelligence for social good [technical committees]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 287–C3, 2021.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 79, pp. 3431–3440, 2015.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.

[4] Jinming Cao, Yangyan Li, Mingchao Sun, Ying Chen, Dani Lischinski, Daniel Cohen-Or, Baoquan Chen, and Changhe Tu, "Do-conv: Depthwise over-parameterized convolutional layer," *Computer Vision and Pattern Recognition*, June 2020.

[5] Lichen Zhou, Chuang Zhang, and Ming Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 182–186, 2018.

[6] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," *Springer, Cham*, 2018.