EXTRACTING WATER BODIES IN RGB IMAGES USING DEEPLABV3+ ALGORITHM

Akula Harika [a], Ramesh Sivanpillai [b, *], Sajith Variyar V. V [a], Sowmya V [a]

[a] Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India – 641112. (ORCID: 0000-0002-5301-9229) - akulaharika12@gmail.com, (ORCID: 0000-0003-3944-8155) - vv_sajithvariyar@cb.amrita.edu, (ORCID: 0000-0003-3745-6944) - v_sowmya@cb.amrita.edu
[b] Wyoming GIS Center, University of Wyoming, Laramie, WY 82072, USA (ORCID: 0000-0003-3547-9464) – sivan@uwyo.edu

KEY WORDS: Sentinel 2A/B, Multi-scale Features, Encoder-Decoder, Atrous Spatial Pyramid Pooling, ASPP, Water quality, NIR.

ABSTRACT:

Deep Learning algorithms are increasingly used for mapping waterbodies in remotely sensed images. DeepLabV3+ is an image segmentation method that includes ASPP and encoder-decoder to retrieve pyramid spatial features at different scales and structural information respectively. Previous studies have shown that DeepLabV3+ can accurately map waterbodies in false colour infrared images. However, ability of DeepLabV3+ for extracting waterbodies in RGB images is unknown. This study tested DeepLabv3+ algorithm to extract waterbodies in the RGB bands. Sentinel 2A/B images (n = 2841) and their corresponding annotations were downloaded from Kaggle (host of public datasets) and subset images (n = 10405) of 100 x 100 pixels were cropped. From these subset images, 8941 were used for training and validation and 1464 were used for testing the trained model. Dice and Jaccard/Intersection over Union (IoU) were used for evaluating the output generated by the model. The network was trained for 50 epochs with 32 iterations in each epoch. The model trained at the end of 30th epoch was selected as final based on minimum information loss (0.0743). The average Dice and Jaccard/IoU scores for the output images were 0.8412 and 0.7169 respectively. The high scores obtained in this study indicate that DeepLabV3+ can be used for identifying waterbodies in RGB or true-colour images.

1. INTRODUCTION

Remotely sensed data are used for monitoring changes in water bodies and other earth surface features. Previous studies have reported the importance of the spectral information collected in the infrared regions for classifying pixels corresponding to water class in satellite and aerial images. Normalized Difference Water Index (NDWI) (McFeeters, 1996) and Modified NDWI (MNDWI) (Xu, 2006) are commonly used spectral indices for distinguishing pixels corresponding to water bodies. These indices measure the difference in the spectral information collected in green and near-IR (NDWI) or mid-IR (MNDWI) bands respectively. However, not all sensors collect spectral data in the infrared regions. Hence, it is not possible to compute NDWI and MNDWI values from the images they acquire.

Distinguishing pixels corresponding to water bodies in RGB images is relatively difficult with statistical/pixel clustering techniques because of the spectral overlap between earth surface features. Variations in water quality (turbidity, presence of floating biological materials) can increase this overlap, making it difficult to classify water bodies from other surface features in RGB images.

Newer methods that rely on pattern recognition has shown to overcome many of the limitations associated with traditional image classification methods for distinguishing earth surface features in remotely sensed images. Mainly, Deep Learning (DL) methods has shown improvements in various applications under airborne and space borne platforms. Mohan et al. (2018) used advanced DL methods on aerial images for vehicle detection using Alexnet and VGG-16, epiphyte segmentation (Shashank et al., 2020) and effect of annotation and loss function (Aswin et al., 2021). Sunil et al. (2021) used Faster R-CNN network to identify oil pads in high resolution aerial images. Previous Studies have reported that DL methods were able to better distinguish surface features in remotely sensed images and achieve accuracy. DL methods assign class label to image pixels to understand higher-level semantics. Several DL methods have been used for identifying and classifying water body in remotely sensed images.

Zhang et al. (2007) used layered feed-forward Neural Network classifier to classify pixels corresponding to water bodies in Landsat Thematic Mapper (TM) images. Isikdogan et al. (2017) proposed a Deepwater map technique to identify water pixels in Landsat images with different land cover classes and clouds/cloud shadows etc. Li et al. (2019) used FCN model to extract water bodies in VHR images collected by the GaoFen-2 satellite. Dong et al. (2019) introduced SNS-CNN architecture which is the modified Unet to segment water bodies in optical remote sensing images downloaded from Google Earth™. Multi-scale feature extraction is a critical and important task in multi-spectral image segmentation. From the limitations of the above-mentioned methods which uses only normal convolution suffers from multi-scale feature extraction. Previous and recent studies demonstrate the importance and need for multi-scale feature extraction for classifying earth surface features in multispectral images. Atrous convolution-based models capture multi-scale information in cascade or parallel context by adopting multiple rates. DeepLabV3+ is one of those models in the DeepLab series which is a widely successful DL algorithm which fulfilled the need of multi-scale feature extraction.

* Corresponding author

1.1 Overview of DeepLab Neural Network

DeepLab is a state-of-the-art segmentation model introduced by Google. DeepLabV1 is an advancement over the earlier standard models called Fully Convolutional Network (FCN) (Long et al., 2015) or several Deep Convolutional Neural Network (DCNN). One of the limitations of DCNN was the reduced spatial resolution in the output feature maps. To address this limitation, DeepLabV1 removed the down-sampling operator from the last few pooling layers of DCNN and replaced it with atrous convolution layer to increase the sampling rate. After acquiring the input images, DeepLabV1 passes them through DCNN followed by a couple of atrous convolution layers resulting in a coarse-grained feature map. Bilinear interpolation techniques are used for recovering the images in original spatial resolution. DeepLabV2 uses Atrous Spatial Pyramid Pooling (ASPP) and applies atrous convolution with different sampling rates to the feature map generated by DCNN. It enables to account for information captured at different scales and improves accuracy. DeepLabV3 uses an improved ASPP that includes batch normalization and image-level features. One of the main challenges in image segmentation is to capture sharper object boundaries. DeepLab3+ addresses this challenge by introducing a decoder in it which was not present in the earlier versions of DeepLab models.

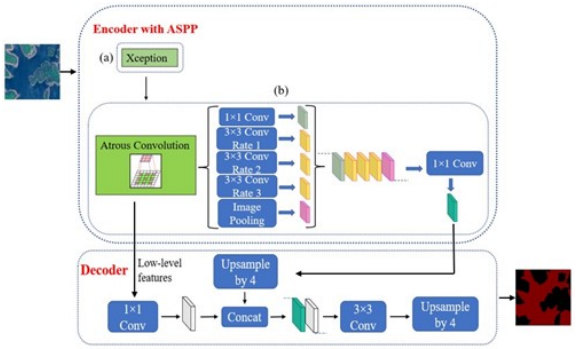

Figure 1. Overview of DeepLabv3+ Architecture – Consisting of Encoder-Decoder with Atrous Spatial Pyramid Pooling. (a) Xception is the Encoder for generating features. (b) Atrous spatial pyramid pooling in which atrous convolution is employed at multiple rates on the features generated by the encoder. Decoder is to upsample the relevant features generated by encoder.

DeepLabV3+ consists of an encoder-decoder (Ronneberger, 2015) architecture with ASPP (He et al., 2015) in between the encoder and decoder modules (Figure 1). Encoder captures texture information including edges followed by pooling operations to reduce the spatial dimension of feature maps. Decoder recovers the detailed information of the feature maps and the corresponding spatial dimension by up sampling the features. This encoder-decoder architecture proved to be useful for various applications. DeepLabV3+ architecture that combines the encoder-decoder and ASPP was introduced by Chen et al. (2018).

Xception is used as the encoder network (Figure 1) in DeepLabV3+ network (Chen et al., 2018). Encoder network gradually reduces the size of the feature maps and captures high-level sematic information. In this network, output stride is defined as the ratio of input image resolution to the final output resolution. ASPP (Figure 1) encodes multi-scale information through several rates using the atrous convolution followed by pooling those multi-scale features. First atrous convolution is applied over the input feature map with rate *r* that corresponds to the stride with which the filter must move. The value of *r* assigned is directly related to the dimension of the input image. For smaller images (e.g., 100 x 100 pixels), low *r* values (2 and 4) are assigned. Higher r values (6, 12, 24 etc.) will be assigned for larger images (256 x 256, 512 x 512). Finally, decoder (Figure 1) reads the features generated by the encoder after applying atrous convolution at multiple rates. Features are up sampled and concatenated with the low-level features generated in the encoder part.

This study tested whether DeepLabV3+ can correctly identify pixels corresponding to water bodies using only the RGB (true color) bands acquired by the sensors onboard Sentinel 2A/B satellites.

2. MATERIALS AND METHODS

2.1 True colour satellite and mosaic images

Escobar published the RGB bands of Sentinel-2 images of water bodies in Kaggle (Escobar, n.d) under "Satellite images of water bodies". This is a publicly available dataset consisting of 2841 images in different dimensions ranging between 69 x 5 and 6683 x 5640 pixels. This dataset consisting of RGB and their corresponding mask images was downloaded as a single compressed file (247 MB). The masks were generated with Normalized Difference Water Index (NDWI) derived from bands 8 and 3 of the Sentinel-2 A/B satellite. Pixels corresponding to water are highlighted in white colour while the rest of the features are represented in black (background). Water bodies in these images represented clear and turbid water, and in different proportions. Majority of the images were in good quality (contrast) whereas 5% of the images were in poor quality (contrast or haze).

2.2 DeepLabV3+ network

The current version of DeepLabV3+ network was downloaded from GitHub (Tomar, 2021) as a single zip file. The zip file consisted of separate sub-folders for input images, model architecture, training the model, predicting output and evaluation measures as Phython-3.9 script files. These files were initially downloaded to a Windows 10 laptop (Intel core i5, 8th Generation, 1.60 GHz, 64 bits processor with 8 GB RAM). Script files were edited in Notepad++ and were uploaded to a LINUX kernel-based server with Ubuntu 10.04.6 LTS which was remotely accessed through the Windows 10 laptop.

2.3 Training and Test images

Sentinel-2 A/B images and their corresponding masks were cropped (pixel dimension of 100 x 100) resulting in 10405 images. Approximately 15% of the subset images were set aside as test images (n = 1464). The rest of the subset images (n = 8941) were split in 8:2 ratio for training and validating the DeepLabV3+ network.

2.4 Training the DeepLabV3+ network

DeepLabV3+ network was set to train for 50 epochs with 32 iterations within each epoch. During the training process, weights and bias values are passed and adjusted using an optimizer. Weight is a learnable parameter that transforms input

image/data. Bias is a constant that helps the network that can fit best model for the given input images. The network has several filters/kernels that performs the convolution, and outputs a feature map from each of the convolution layers along with the weights. These weights are updated during the back-propagation process. The total number of learnable/trainable parameters from DeepLabV3+ network is 41044130. Based on the loss at the end of each epoch, the learnable parameters were adjusted. The training loss indicates if the model can fit the training data i.e., whether the model has enough information to process the required information from the input images. The validation loss indicates how well the model is able to predict the validation images.

2.5 Model evaluation

Test images (n = 1464) were used for evaluating the performance of the trained DeepLabV3+ model. Jaccard and Dice scores were computed using the True Positive (TP), False Positive (FP), and False Negative (FN) values.

Jaccard index or Intersection over Union (IoU) measures the similarity between the model predicted output and NDWI generated mask images, and is computed using the following equation:

$$\text{Jaccard score} = (TP) / [(TP) + (FP) + (FN)] \quad (1)$$

Dice index or F1 score is twice the intersection of mask and predicted images over the sum of the pixels. It varies from Jaccard index which only counts the TP values once in the numerator and denominator.

$$\text{Dice score} = (2 \times TP) / [(2 \times TP) + (FP) + (FN)] \quad (2)$$

Both scores are used as a similarity measure between the model predicted and mask images, and range between 0 (no similarity) and 1 (high similarity). In this study, the evaluation metrics for the water class were computed and reported.

3. RESULTS AND DISCUSSION

3.1 Training and Validation loss

Training loss was high until the fifth epoch which indicated that the DeepLabV3+ network learned few information from the input (training) images (Figure 2). After the sixth epoch, the network's learning improved but there were fluctuations until 15th epoch. This could be due to learning difficulties encountered by the model during the early epochs.

The validation loss steadily declined until the 15th epoch. This could be a result of model overfitting the images used for validation. After the 15th epoch, the trained model was able to generate output (predicted) mask that better matched with the corresponding ground truth mask. The validation loss remained a constant at the end of the 25th epoch and until the 35th epoch. Minor fluctuations were noticed past 35th epoch. This could be due to the repetition of the validation samples after the 35th epoch because of learning patterns beyond the target (water) class. Learning patterns beyond the information included in the mask could result in higher validation loss values.

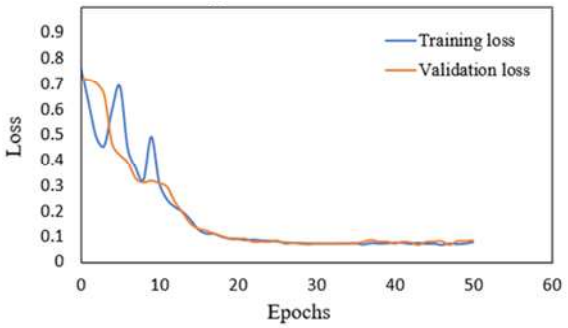

Figure 2. Plot depicting training (blue line) and validation (orange line) losses

The first minimum validation loss was reported at 25th epoch and remained stagnant until the 35th epoch. The trained model at the end of the 30th epoch was selected to evaluate the test images. The validation loss at the end of the 30th epoch was 0.0743. The time taken for training and validating the DeepLabV3+ was 13 hours. The weights saved in the checkpoints were used for evaluating the 1464 test images.

3.2 Evaluation of model performance

Trained DeepLabV3+ model was evaluated on 1464 test images. The statistical summary of the Jaccard and Dice scores are listed in Table 1.

Statistical summary of evaluation metrics	Jaccard	Dice
Average	0.7169	0.8412
Minimum	0	0
Maximum	1	1
Standard deviation	0.115	0.093

Table 1. Statistical summary of the evaluation metrics obtained by comparing the DeepLabV3+ model predicted images to the NDWI derived mask images (n = 1464).

The average Jaccard and Dice scores for the test images were 0.7169 and 0.8412 respectively. The Jaccard/IoU is above average while the Dice score indicates very good agreement between the predicted and actual mask images. These scores indicate the trained model was able to predict most of the target pixels. However, the range of these metric scores indicate that there was a wide range of variation in the model's ability to correctly predict the target class in the test images. The minimum score (0) indicates that the model failed to predict all target pixels in some images. The maximum score (1) indicates a perfect match between the model predicted output and NDWI derived mask images.

Figure 3 highlights the results from select (n = 7) RGB, corresponding mask, and model predicted output images. Their Jaccard and Dice scores are included in the caption.

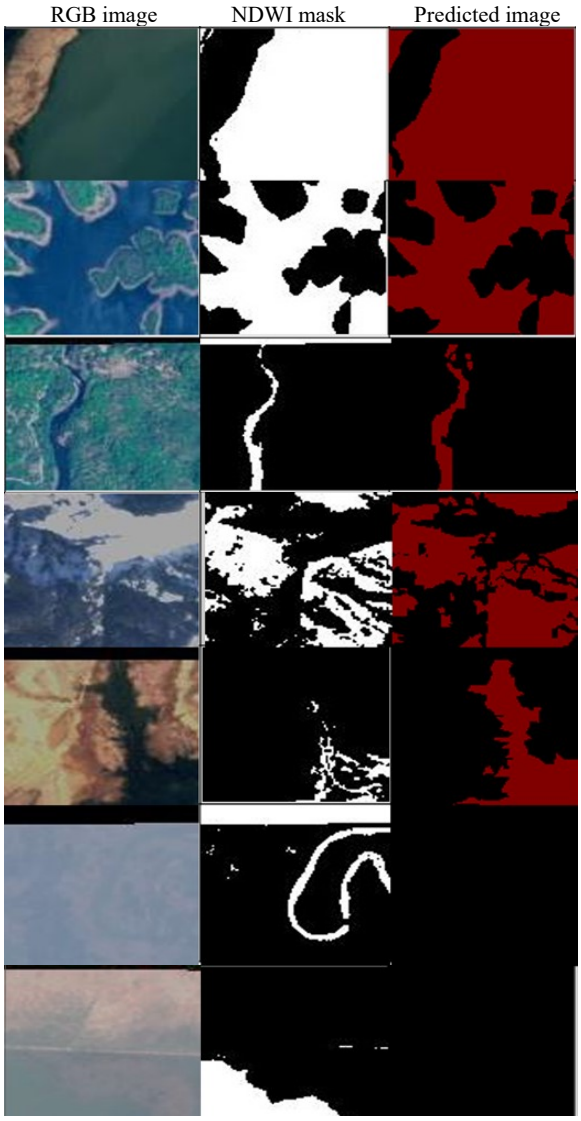

Figure 3. Select input RGB (left), corresponding NDWI derived mask (middle), and the trained DeepLabV3+ model predicted (right) masks. The Jaccard scores for the 7 sample images (top to bottom) were 0.8436, 0.9141, 0.7526, 0.4169, 0.3658, 0 and 0 respectively. The corresponding Dice scores were 0.8924, 0.9732, 0.8025, 0.5472, 0.4198, 0 and 0 respectively.

From the sample images presented in Figure 3, it is evident that the DepLabV3+ model was able to correctly predict water (target) class under certain conditions. Irrespective of the number of target pixels (occupancy) in an image, the model correctly identified most of the target pixels in rows 1-3. These images were of good quality, and the water bodies were also relatively clear in them.

In some images, DeepLabV3+ model identified both clear and turbid water pixels, more than the target pixels in the mask images (Figure 3, rows 4 and 5). Previous studies have shown that NDWI is less effective to identify turbid water. Hence NDWI generated masks could have excluded some or all of pixels corresponding to turbid water. This mismatch between the predicted and mask images would have resulted in lower Jaccard and Dice scores.

When the overall image quality was poor due to haze or contrast, the DeepLabV3+ model was unable to identify any of the target pixels (Jaccard score = Dice score = 0). Since the mask images were generated with the NIR band, the quality of RGB bands did not influence the identification of water bodies. Since DeepLabV3+ was not trained with the NIR band, it was unable to correctly identify the water bodies in the poor-quality images.

4. CONCLUSION AND FUTURE WORK

Based on the results obtained in this study, DeepLabV3+ can be used for identifying pixels corresponding to water bodies in the RGB bands of the Sentinel-2A/B images. The predicted images were comparable to their corresponding mask images when the image and water quality were higher.

Future work must focus on analysing the conditions that resulted in poor prediction, and suitable modifications have to be made to the training or validation steps. These modifications will improve the model's ability to predict water bodies in poor quality images.

ACKNOWLEDGEMENTS

Authors thank Prof. K. P. Soman, Head, Center for Computational Engineering and Networking (CEN) at Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, and UW Wyoming GIS centre for their valuable support.

REFERENCES

Aswin, S., Sajithvariyar, V., Sivanpillai, R., Sowmya, V., Brown, G. K., Shashank, A., Soman, K., 2021. Effect of annotation and loss function on epiphyte identification using conditional generative adversarial network. *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, IEEE, 1–6. DOI: 10.1109/ICAECT49130.2021.9392478.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV),* 801–818. DOI: 10.1007/978-3-030-01234-2_49.

Dong, S., Pang, L., Zhuang, Y., Liu, W., Yang, Z., Long, T., 2019. Optical remote sensing water-land segmentation representation based on proposed sns-cnn network. *IGARSS 2019- 2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 3895–3898. DOI: 10.1109/IGARSS.2019.8898367.

Escobar, F., (n.d). "Satellite Images of Water Bodies", https://www.kaggle.com/datasets/franciscoescobar/satellite images-of-water-bodies, CC BY-NC-SA 4.0, (6 June 2022).

He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, 37(9), 1904–1916. DOI: 10.1109/TPAMI.2015.2389824.

Isikdogan, F., Bovik, A. C., Passalacqua, P., 2017. Surface water mapping by deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(11), 4909– 4918. DOI: 10.1109/JSTARS.2017.2735443.

Li, L., Yan, Z., Shen, Q., Cheng, G., Gao, L., Zhang, B., 2019. Water body extraction from very high spatial resolution remote sensing data based on fully convolutional networks. *Remote Sensing*, 11(10), 1162. DOI: 10.3390/rs11101162.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431– 3440. DOI: 10.1109/CVPR.2015.7298965.

McFeeters, S. K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7), 1425–1432.

Mohan, V. S., Sowmya, V., Soman, K., 2018. Deep neural networks as feature extractors for classification of vehicles in aerial imagery. *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, 105–110. DOI: 10.1109/SPIN.2018.8474153.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer- Assisted Intervention*, Springer, 234–241. DOI: 10.1007/978-3-319-24574-4_28.

Shashank, A., Sajithvariyar, V., Sowmya, V., Soman, K., Sivanpillai, R., Brown, G., 2020. Identifying epiphytes in drones' photos with a conditional generative adversarial network (C-GAN). *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 44, 99–104. DOI: 10.5194/isprs-archives-XLIV-M-2-2020-99-2020.

Sunil, A., Sajithvariyar, V. V., Sowmya, V., Sivanpillai, R., Soman, K. P., 2021. Identifying oil pads in high spatial resolution aerial images using faster r-cnn, *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences.*, XLIV-M-3-2021, 155–161, DOI: 10.5194/isprs-archives-XLIV-M-3-2021-155-2021.

Tomar, N., 2021. "Human-Image-Segmentation-with-DeepLabV3Plus-in-TensorFlow", https://github.com/nikhilroxtomar/Human-Image-Segmentation-with-DeepLabV3Plus-in-TensorFlow (6 June 2022).

Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14), 3025–3033. DOI: 10.1080/01431160600589179.

Zhang, Y., Gao, J., Wang, J., 2007. Detailed mapping of a salt farm from Landsat TM imagery using neural network and maximum likelihood classifiers: a comparison. *International Journal of Remote Sensing*, 28(10), 2077–2089. DOI: 10.1080/01431160500406870.