THE EYE IN THE SKY

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Team Members:

Dishank Bansal - dishank.bansal@iitkgp.ac.in
Sayan Sinha - sayan.sinha@iitkgp.ac.in
Siddhant Haldar - siddhanthaldar14@iitkgp.ac.in

CONTENTS

INTRODUCTION

In present times, as multi-spectral imagery is gaining pace, the availability of satellite imaging data from remote sensing satellites has resulted in a considerable effort being devoted to the classification of this data with of aim of producing high-quality thematic maps and establishing accurate inventories of spatial classes. This events demands the design and implementation of a satellite image classification pipeline for a given dataset of satellite images. We were provided with 14 high-resolution multispectral satellite images having four bands - Blue, Green, Red and Near-Infrared - along with the ground truth segmentation maps depicting 8 labelled classes - Roads, Buildings, Trees, Grass, Bare Soil, Water, Railways and Swimming pools. Hence, this model can be approached as a per-pixel classification problem which follows the same principles as that of semantic segmentation. The domain of semantic segmentation aims at producing a pixel-wise probability distribution depiction the most probable class at each pixel. Here, the primary aim is to partition the given image into regions or segments such that pixels belonging to a region are more similar to each other than pixels belonging to different regions.

Satellite image segmentation is gaining pace in present times with the technique finding applications in several domains. Satellite imaging is widely used for segmentation of the earth's surface for the detection of prominent waterbodies, vegetation and the development of maps utilizing the segmentation of roads, buildings, railway lines and the like. It is also used for monitoring the spatial displacement of human population and for evaluating the available surface for implanting solar panels on roofs. Satellite image classification also finds an important application in the assessment of building damages following a natural disaster, allowing the formulation of an adequate response in targeted areas and is also used for quantifying the total forest area and the rate of deforestation in a particular area.

Looking back in history, a common technique used for classification of multi-spectral images is to exploit the spectral signature of each pixel. The spectral signature of a pixel provides information about the reflectance and absorptance of surfaces depicted in the image. Such properties depend on the physical as well as chemical properties (such as colour, texture, etc) of the surface and hence, every object in this world has an unique spectral footprint. Researchers have been working on developing models that focus on learning the spectral footprint of the classes which are to be classified. Some common models which are used in this regard are Support Vector Machines (SVMs) and neural networks among others.

A common approach while performing semantic segmentation is to also utilize the spatial information i.e., the information provided by the neighboring pixels, to predicted the probability of the presence of a certain class at each pixel of the image. An effective image processing algorithm in this regard is the Watershed Algorithm. However, recent works have established the superiority of convolutional neural networks, with CNN-based models achieving superhuman accuracy when it comes to certain tasks [4]. Although CNNs have proved to be extremely effective for tasks related to image segmentation, one problem with this approach is that a large amount of data to train such networks. Thus, in this report, we propose a deep learning based semantic segmentation architecture which aims at achieving effective classification of the given images in spite of being limited by the availability of comparatively less training da

Classification Approach

2.1 Motivation

Traditionally, techniques for satellite image segmentation have been focusing on learning the spectral signals of classes that are to be segmented. However, such an approach has its limitations as even though it utilizes the information obtained from the physical and chemical properties of the reflecting surfaces, it misses out on the spatial context provided by the image.

Modern state-of-the-art image segmentation techniques address this problem by utilized deep learning techniques such as the use of convolutional neural networks to avail the spatial information provided by the incoming data. However, to train such a network, a large amount of data is usually required. Hence, we aim to develop a model that combines both the spectral and spatial information available from the incoming data and that can also be trained using as less as 14 images. We develop an ensemble classifier comprising a couple of primary classifiers, the results from which are combined using a meta-classifier which produces the final segmentation map. Such an approach is expected to work better than using individual classifiers since our model now utilizes both, the spatial and the contextual, information and makes a more informed decision to generate a more accurate pixel to pixel mapping for the segmentation model.

2.2 Methodology

2.2.1 Pre-Processing

The given dataset comprised 16 bit 4-channel multispectral images comprising Blue, Green, Red and Near-Infrared bands. Owning to 16 bit memory, values in images reaches to 4000 whereas for viewing image it has to be in 0 to 255 . Hence, we normalize the given images using a couple of techniques. We tried Min-Max normalization and Mean+/-2xStandard Deviation normalization. We found out that later one provide better contrast, making scene clearly distinct. Fig. 1 shows the visualization images after normlization.

Min = Channel.min
Max = Channel.max
Channel = (Channel – Min)/(Max-Min)

Min = Mean – 2 x Std. Deviation
Max = Mean + 2 x Std. Deviation
Channel = (Channel – Min)/(Max-Min)

Figure 1 : Comparison of the proposed image normalization methods

We also incorporated the Normalized Difference Water Index (NDWI) and Normalized Difference Vegetation Index (NDVI) in our input data. NDVI is commonly used for vegetation index whereas NDWI is the most appropriate for water body segmentation. Hypothising that using NDWI and NDVI index will imporve network ability to segment water and grasslands, we also added these along with 4 bands. Hence, our input data comprised 6 channels - Blue, Gree, Red, Near-Infrared, NDWI and NDVI.

$$NDVI = \frac{NIR - RED}{NIR + REd} \qquad NDWI = \frac{GREEN - NIR}{GREEN + NIR}$$

2.2.2 Model

A lot of past work in the field of multi-spectral image classification has been focusing on the use of spectral information of each pixel for developing segmented maps of the input image. With the rise of deep learning, researchers have been adopting convolution neural networks (CNNs) for building semantic segmentation models which take into account the spatial information provided by the neighbourhood of a pixel. However, such models require a large amount of data to train. Hence, the main challenge in this task was to develop a models that utilizes both the spectral and spatial information of the image and can also be trained on less training data. Hence, we have developed a pipeline that assists the model to learn the necessary information irrespective of having limited training data.

Architecture:

We develop a deep learning model which takes the 6-channel image as input and provides the segmented map as the output. Our pipeline comprises an ensemble of a couple of networks. One of the models is a 4-layer deep neural network which performs binary segmentation corresponding to each class, the results of which are merged to produce the final pixel-wise probability distribution. The second is a modified U-Net [2] utilizing dense blocks [1], which we call *Dense U-Net,* which produces the pixel wise probability map for corresponding 8 classes taking into account the spatial information at each pixel. The outputs from these 2 networks are merged using a meta-classifier to produce the final segmented map. Using an ensemble of the two models helps us reduce the variance in the output predictions, thus, reducing the tendency of the model to overfit. The architecture proposed by us is demonstrated in Fig. 2.

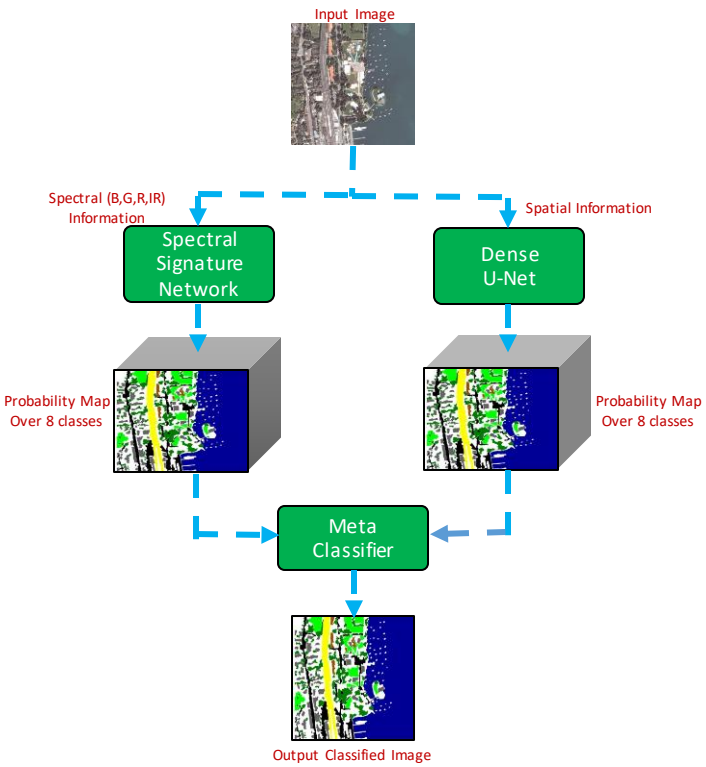

Figure 2 : The proposed architecture comprising the Spectral Signature Network, Dense U-Net and the Meta-Classifier Network

Spectral Signature Network:

The 4-layer deep neural network which performs binary segmentation corresponding to each class. Hence, we have 8 binary classifiers for each of the 8 labelled classes which is used to build the 8 channel segmentation map produce by the primary 4-layer neural network. As each surface has an unique spectral signature defined by their physical and chemical properties. This model aims to learn the spectral signature corresponding to each segmented class. The proposed spectral signature network is shown in Fig. 3.

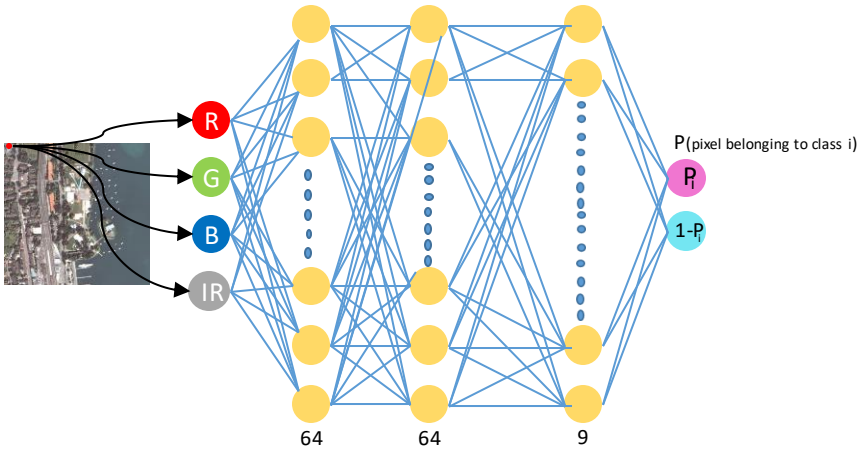

Figure 3 : The 4-layer binary classification network, termed as the Spectral Signature Network

Dense U-net:

The second branch is a modified U-Net model utilizing dense blocks, which we call *Dense U-Net.* The architecture is inspired from a new CNN architecture, Densely Connected Convolutional Networks (DenseNets), which has shown excellent results on image classification tasks. The idea of DenseNets [5] is based on the observation that if each layer is directly connected to every other layer in a feed-forward fashion then the network will be more accurate and easier to train. U-Net is built from a downsampling path, an upsampling path and skip connections. Fig 4 shows an outline of the architecture. The advantages of using skip connections are:

- They help the upsampling path recover spatially detailed information from the downsampling path, by reusing features maps.
- They aid the flow of gradients to the encoder part during training, thus minimizing the risk of vanishing gradient as model depth increases.

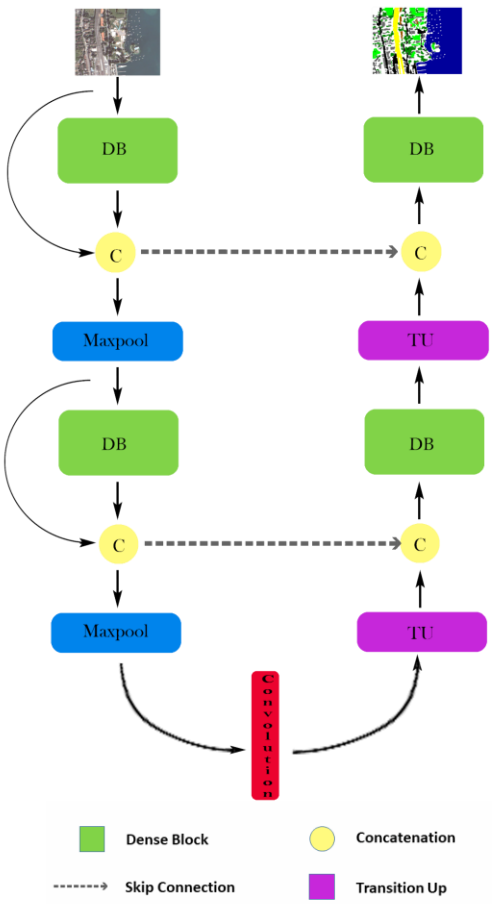

Figure 4 : The proposed Dense U-Net Model

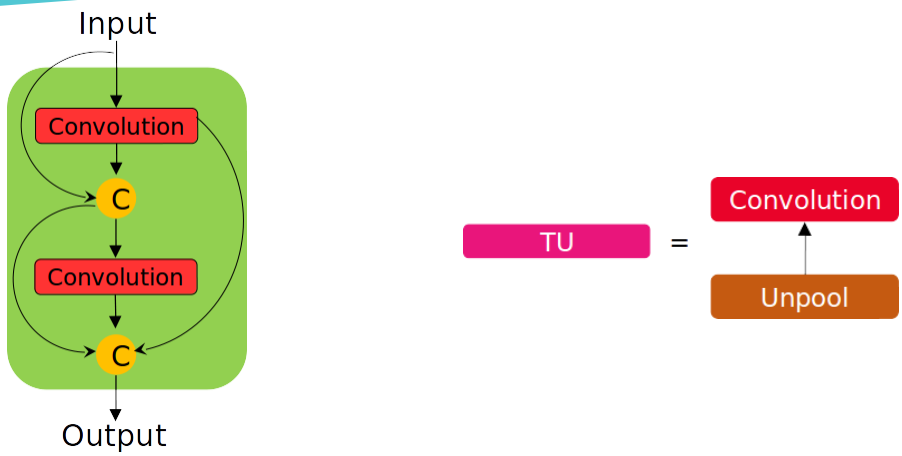

Figure 5 : Components of a dense block(DB) [left] and the Transition Unpool (TU) block [right]

The goal of our model is to exploit the feature reuse by extending the more sophisticated DenseNet architecture, while avoiding the feature explosion at the upsampling path of the network. Another modification in our network is that we have used unpooling layers followed by convolution layer instead of transposed convolution blocks in our upsampling layers. The unpooling layer upsamples a coarsely-resolved feature map from the preceding decoder block to a finer resolution by copying the values at the pooling indices to blocks that are double the size. Such an unpooling layer ensures that spatial information remains preserved, in contrast to using interpolation or the use of transposed convolutions for upsampling. The components of a dense block and Transition up have been shown in Fig 5.

Meta Classifier:

In order to integrate the spectral signature learned by the first neural network based model and the spatial information learned by the Dense U-Net model, we use the outputs obtained from the two networks as inputs for a meta-classifier. In the meta-classifier, the use of convolutions over all channels would result in mixing up of the information from all channels which is expected to harm the model performance. Instead we train the meta-classifier to learn weights for each pair of channels. Hence, our final model takes a class wise weighted sum of the probability distribution obtained from the two preceding classifiers to produce the final class wise probability distribution.

2.2.3 POST-PROCESSING

The proposed model provides us with a 9 channel pixel-wise probability distribution. Obtaining the segmentation map by considering the channelwise maximum score for each pixel results in omission of the under-represented classes like yellow and brown from the segmented map. Therefore, we devised a novel post processing technique in order to address this issue.

We observe that due to the lack of sufficient data for the yellow and brown classes, the proposed model is unable to learn these under-represented classes. Therefore, we utilize the pixel-wise probability distribution corresponding to each class obtained from the 8 binary classifiers in the spectral signature model. We obtain suitable probability thresholds for each class to achieve a high precision with a decent value of recall for each class. After obtaining the class wise probability thresholds, we marks the pixels which have just one channel satisfying the probability threshold corresponding to a certain class. These pixels are marked to represent the class whose threshold is satisfied at that pixel.

The remaining pixels which do not satisfy this criteria are subjected to being marked by the class corresponding the the channel wise maximum score at that pixel. Since we obtain suitable thresholds for all classes, this method ensures that the under-represented classes get adequate representation, atleast at the pixel at which they have a relatively high confidence score.

2.3 Implementaion

The dataset provided for this competition comprised 14 images having variable sizes. Hence, we have focused on developing an architecture that is independent of the size of the input images. Our model is composed of two primary networks - a fully connected network that maps individual pixel values and a *Dense U-Net* model which is fully convolutional network. Hence, we observe that neither of the two models is dependent on the size of the input image.

For training the network we used categorical cross entropy loss. Cross-entropy loss, or log loss, measures the performance of a classification model whose output is a probability value between 0 and 1.

$$Loss = -\sum_{i} p_i log(q_i)$$

p is the target probability distribution of classes and q is the output of the model. This loss ensured that q approaches to p.

For updating the weights of the model, we used Adam Optimizer with a learning rate of 0.0002 and the exponential decay rate for the first moment estimates was set to 0.5. The network was trained using 4-fold cross validation technique.

For training the purposes, the provided images had been divided into overlapping patches of size equal to the dimension of the smallest image in the training dataset, i.e., 622x782 pixels. In order to increase the training data, we have used data augmentation where each image was subjected to rotations at intervals of 45 degrees thus obtaining 8 images from each 622x782 dimensional image. Since satellite images depict a top view image of the world, they are not required to conform to any sense of spatial consistency. Hence, this allows us to rotate the images thus enabling the network to get exposed to multiple views of the training data.

A problem that people commonly come across in semantic segmentation is the global boundary effect. In semantic segmentation, as we move towards the boundary of the image, the accuracy of the predictions. In order to tackle this problem, we used reflective padding on the input in order to increase its size. The output obtained on passing this image through the network is then cropped in order to remove the padded layers. This ensures that the outer pixels of the image with poor prediction accuracy is removed, thus, producing better output predictions.

During test time, in order to reduce the variance of the predictions, we applied a test time augmentation technique. Each test image is passed through the network twice with 90 degress change in orientation. The outputs are obtained corresponding to the two images, The output corresponding to the rotated images is again made upright by applying a 90 degree anticlockwise rotation and the outputs corresponding to the two images were averaged with geometric mean to obtain the final segmented map. The proposed test time augmentation technique is shown in Fig. 6.

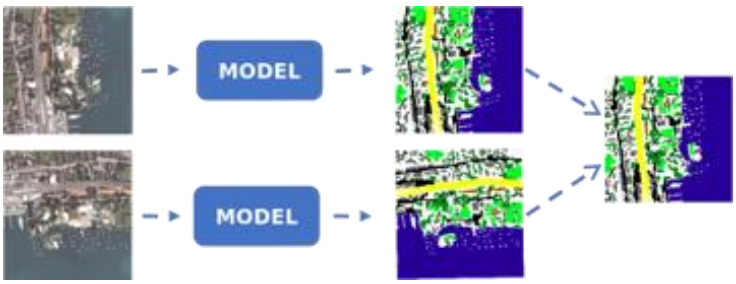

Figure 6: Test-time Augmentation : Upright image[top] and rotated image[bottom] combined to produce the output segmentation map

3 CLASSIFICATION RESULTS

The provided dataset contains a white class which represents the unlabelled pixels. These are pixels which represent objects in the image which could not be annotated and hence, **we have not considered the white pixels for training and validation** of our model. Below(Fig. 7) are the classification results of a couple of validation images.

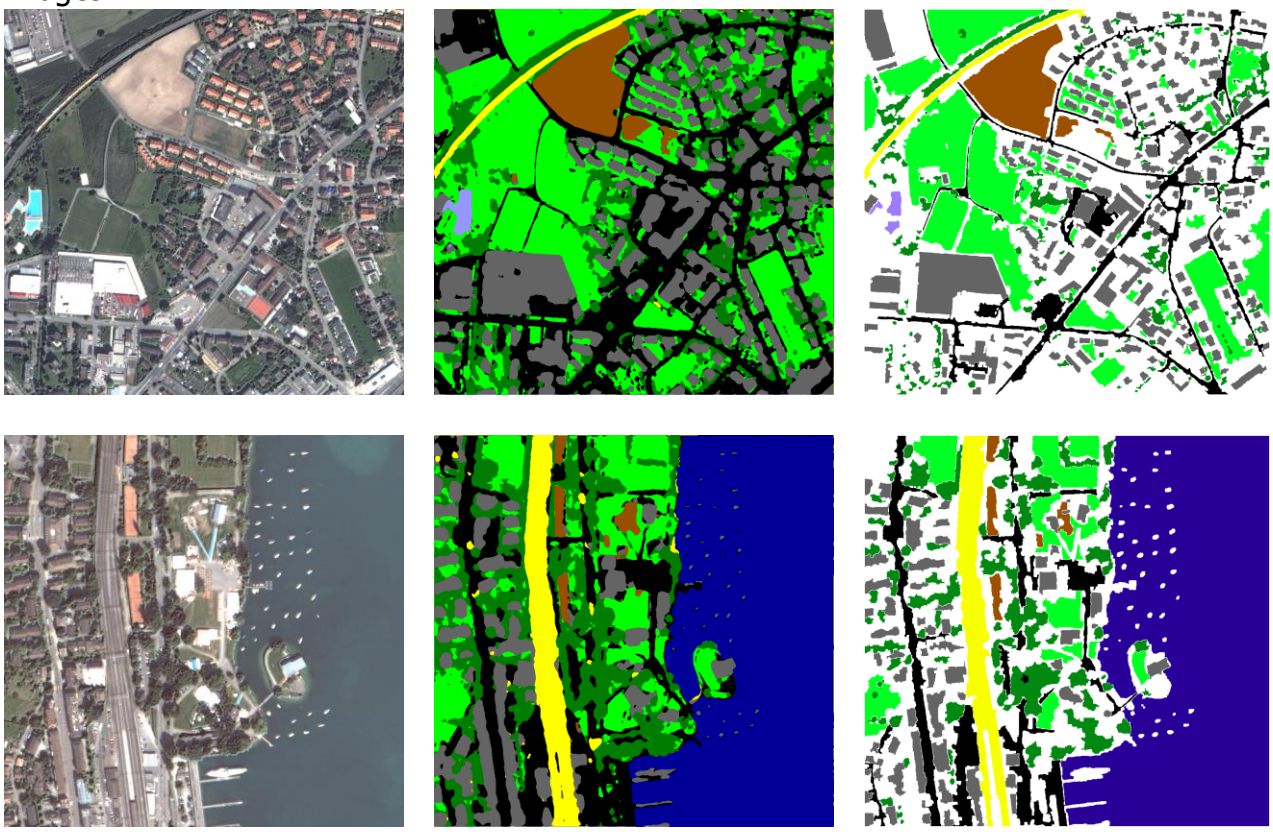

Figure 7: From left to right : Original image(considering only BGR channels)[left], predicted segmentation map[middle] and ground truth segmentation map[right]

Evaluation Metrics

The confusion matrix, kappa coefficient and the overall accuracy reported here has also been calculated by ignoring the pixels that have been labelled white in the ground truth images.

Confusion Matrix

The following confusion matrix is calculated for the validation images.

Ground Truth \ Predicted	Roads	Buildings	Trees	Grass	Bare Soil	Water	Railroad	Swim. Pools
Roads	547223	161611	9587	2801	1979	721	1587	1052
Buildings	102369	1099843	4071	889	630	728	2046	462
Trees	8624	2700	356841	17691	1	44	66	0
Grass	1190	36	2963	68137	3363	0	0	1
Bare Soil	5633	2598	625	1278	15124	0	2	0
Water	19928	8091	657	133	6154	384414	582	0
Railroad	6817	45898	0	0	1262	0	862	0
Swim. Pools	0	0	0	0	0	0	0	4405

	Roads	Buildings	Trees	Grass	Bare Soil	Water	Railways	Swim. Pools
Precision	0.7910	0.8327	0.9522	0.7493	0.5304	0.9961	0.1675	0.7441
Recall	0.7532	0.9082	0.9245	0.9002	0.5987	0.9154	0.0157	1.0

Kappa Cofficient

Cohen's kappa coefficient (κ) is a statistic which measures inter-rater agreement for qualitative (categorical) items.

Validation – 0.8787

Training – 0.9576

Overall Accuracy

The overall classification accuracy measures the percentage of pixel in the predicted segmentation map which represent the same class as the corresponding pixel in the ground truth segmentation map.

Validation – 85.30%

Training – 94.67%

Conclusion 5

Satellite image segmentation, an emerging field as it is, is finding widespread application in several domains. In this report, we propose a novel approach for satellite image segmentation which combines the spectral signature at each pixel with the spatial information due to the neighbours around each pixel to produce the final segmentation map. Using an ensemble of classifiers combined using a meta-classifier helps reduce the variance in the output predictions. We also make use of a test time augmentation technique to further reduce the variance in the model predictions. Here, we obtained an accuracy of **85.3%** on the validation images and **94.67**% accuracy on training images.

In the near future, we plan to verify the efficiency of this model in a deterministic fashion by applying Layerwise Relevance Propagation(LRP) on this model and related models in order to validate the superior performance of our proposed model in comparison to the existing models for satellite image segmentation.

References 6

1. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017, July). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on* (pp. 1175-1183). IEEE.
2. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.
3. Roy, A. G., Conjeti, S., Karri, S. P. K., Sheet, D., Katouzian, A., Wachinger, C., & Navab, N. (2017). ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. Biomedical optics express, 8(8), 3627-3642.
4. 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).
5. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017, July). Densely connected convolutional networks. In CVPR (Vol. 1, No. 2, p. 3).