Segmentation on Unbalanced Aerial Imagery

THESIS SUBMITTED TO THE UNIVERSITY OF CALCUTTA
IN THE FULFILMENT OF THE REQUIREMENT
FOR THE DEGREE OF
BACHELOR OF TECHNOLOGY
IN THE COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

Submitted by

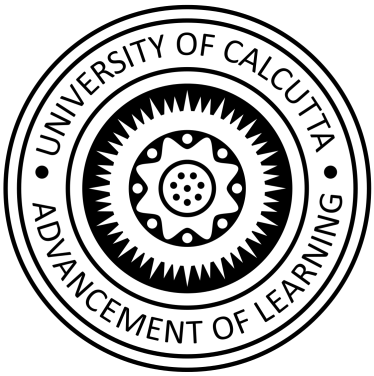
Animesh Ganai
Registration Number - D01-1114-0061-18
Roll Number - T91/CSE/184038

Soumydeep Ganguly
Registration Number - D01-1111-0074-18
Roll Number - T91/CSE/184043

Rahul Das
Registration Number - D01-1112-0005-19
Roll Number - T91/CSE/186018

Under the supervision of

Prof. (Dr.) Sunirmal Khatua



Computer Science and Engineering Department
University Of Calcutta

2018 - 2022

CERTIFICATE

This is to certify that the project report titled **"Segmentation of unbalanced aerial imagery"** submitted by Animesh Ganai and his teammates Soumydeep Ganguly and Rahul Das, in the fulfilment of the requirement of the award of the degree of Bachelor of Technology in Computer Science and Engineering in University Of Calcutta, during the academic session 2018-2022 is a bonafide record of the project work carried out by them under my supervision and guidance.

Date:

Prof. (Dr.) Sunirmal Khatua
Department of Computer Science and Engineering
University of Calcutta
Kolkata - 700106

Acknowledgment

We express our sincere gratitude to the supervisor, respected **Prof.(Dr.) Sunirmal Khatua**, under whose esteemed guidance and supervision, this work has been partially completed. We also want to thank **Swalpa Kumar**. This project work would have been impossible to carry out without their motivation and support throughout.

Animesh Ganai
T91/CSE/184038

Soumydeep Ganguly
T91/CSE/184043

Rahul das
T91/CSE/186018

Abstract

Image segmentation is a method in which a digital image is broken down into various subgroups called Image segments which helps in reducing the complexity of the image. This makes further processing or analysis of the image simpler and easier. Our task is to segment the aerial image taken from a satellite. The task is very challenging because in the dataset only a few images are there, the previous segmentation work was done on the dataset using the U-Net architecture with **77% accuracy**, in this project we are using modified U-Net for the aerial image segmentation task. We use **Vision Transformer** into the U-Net to enhance the performance of segmentation tasks. And by doing so we have achieved **80% accuracy (pixel level),** and we also achieved better precision and recall for different classes using our model over the ordinary U-Net.

Keywords

Image segmentation, U-Net, Multi Scale U-Net , ViT, Transformer Encoder, U2-net, focal loss, adam optimizer

Contents

1. Problem statement:

Our problem belongs to the class of image segmentation. In this project, we are trying to segment areal image, which contains 72 satellite images of Dubai, the UAE, and is segmented into 6 classes. The classes include water, land, road, building, vegetation, and unlabeled.

As the dataset has only 72 images, it is a very challenging task to get higher accuracy over the validation set. This segmentation is already done using U-Net. Our task is to improve the model using different models of the U-Net family and using the different loss functions.

2. What is image segmentation?

Image segmentation is a method in which a digital image is broken down into various subgroups called Image segments which helps in reducing the complexity of the image. This makes further processing or analysis of the image simpler and easier. In the below figure (Figure 1) we can see that the car is moving on a road with a footpath and houses in the background. Houses and vehicles being in the same class are painted orange, footpaths painted in green, and roads painted in black. In figure 2 we can see a similar example where the animal is painted yellow, footpath painted in green, and water bodies in blue.

Fig 1: Semantic segmentation of a car moving on a road(Orange-→ vehicles or houses, blue-→ Road, sky--→purple)

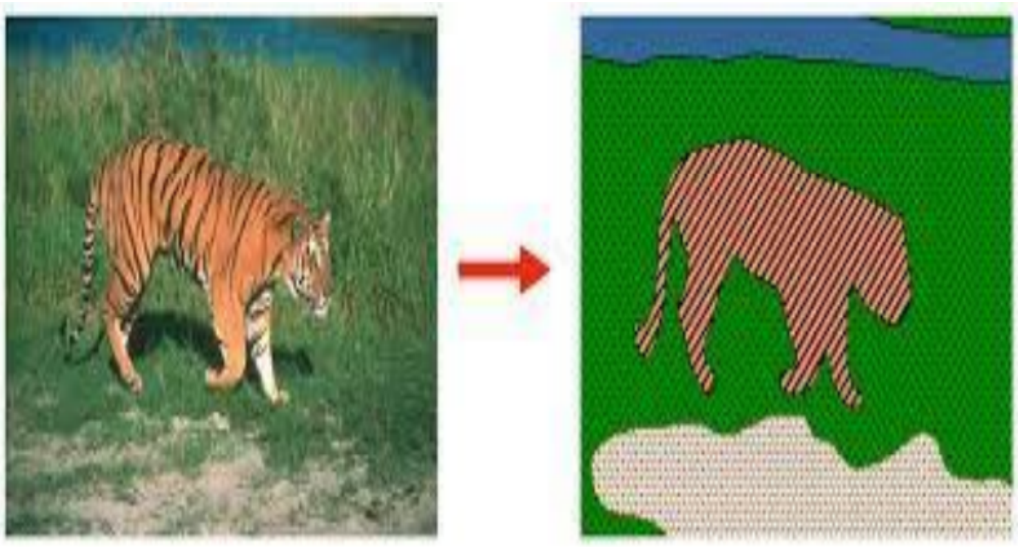


Fig 2: Semantic segmentation of a tiger in the bushes (Green→grasses, blue--→water bodies, yellow-→sand)

3. U-Net:

Convolution neural networks have outperformed the state of the art in many visual recognition tasks. It was used to classify different types of images and provided identification of images irrespective of the location of the object in the image. Their success was limited due to the size of the available training sets and the size of the considered networks. Typical use of CNN is on classification tasks, where the output to an image is a single class label.

However many such tasks like biomedical image processing and satellite image processing, class label for each pixel is required. Moreover, thousands of training images are usually beyond reach in biomedical tasks.

In this, architecture is modified and extended such that it works with very few training images yielding more precise segmentation. The main idea here is to supplement a usual contracting network by successive layers, where pooling operators are replaced by upsampling operators. Hence, these layers increase the resolution features from the contracting path combined with the upsampled output. A successive convolution layer can then learn to assemble a more precise output based on this information. One important modification in this architecture is that in the upsampling part a large number of feature channels, which allow the network to propagate context information to higher resolution layers. As a consequence, the expansive path is more or less symmetric to the contracting path and yields a u-shaped architecture. The network does not have any fully connected layers and only uses the valid part of each convolution, i.e., the segmentation map only contains the pixels, for which the full context is

available in the input image. This strategy allows the seamless segmentation of arbitrarily large images by an overlap-tile strategy (see Fig 3). To predict the pixels in the border region of the image, the missing context is extrapolated by mirroring the input image. This tiling strategy is important to apply the network to large images since otherwise the resolution would be limited by the GPU memory.[1]

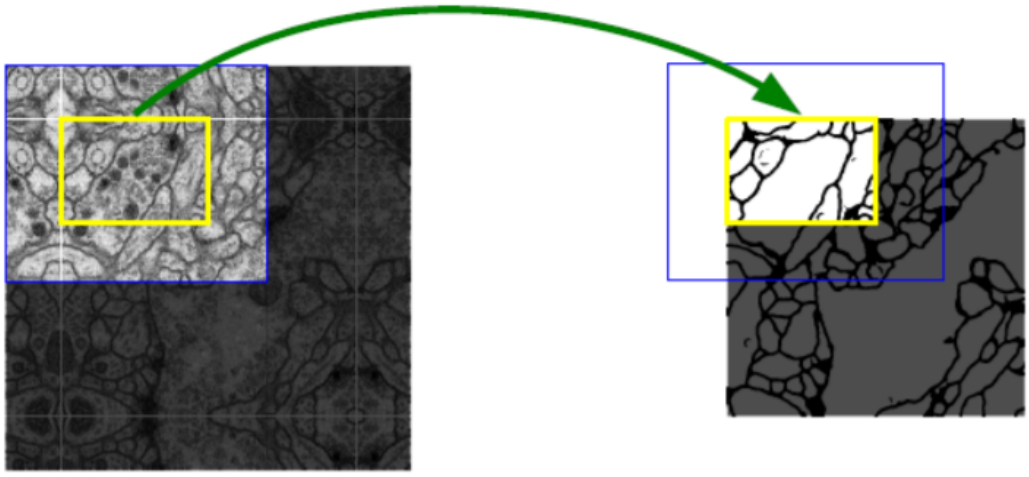


Fig 3: Overlap-tile strategy for seamless segmentation of arbitrary large images (here segmentation of neuronal structures in EM stacks). Prediction of the segmentation in the yellow area requires image data within the blue area as input. Missing input data is extrapolated by mirroring

3.1 U-Net architecture:

The network architecture is illustrated in Fig 4. It consists of a contracting path (left side) and an expansive path (right side). The contracting path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step, we double the number of feature channels. Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution (“up-convolution”) that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer, a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes. In total the network has 23 convolutional layers.[1,2,3]

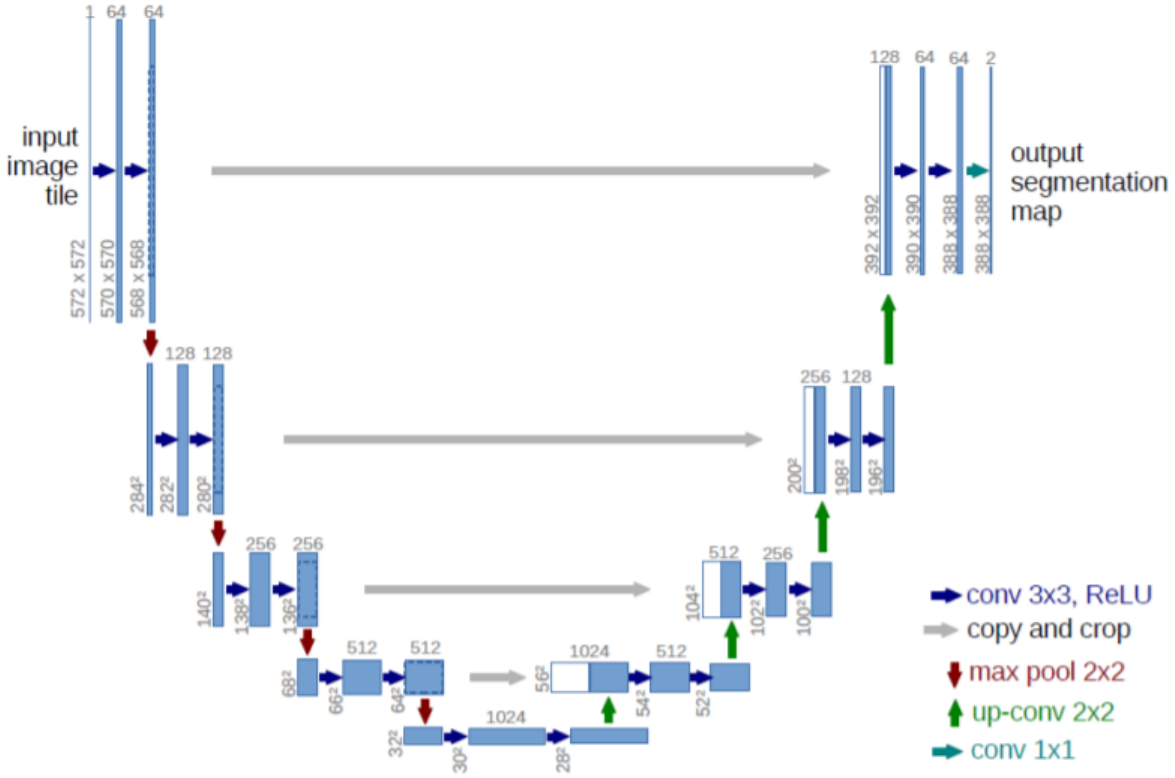


Fig 4: Architecture of UNet

Some Terms Related To The UNet model:

3.2 Convolution operation:

First, let us look at what a kernel is. A kernel is a small matrix, with its height and width smaller than the image to be convolved. It is also known as a convolution matrix or convolution mask. This kernel slides across the height and width of the image input and the dot product of the kernel and the image are computed at every spatial position. The length by which the kernel slides is known as the stride length. In the image below, the input image is of size 5X5, the kernel is of size 3X3 and the stride length is 1. The output image is also referred to as the convolved feature.

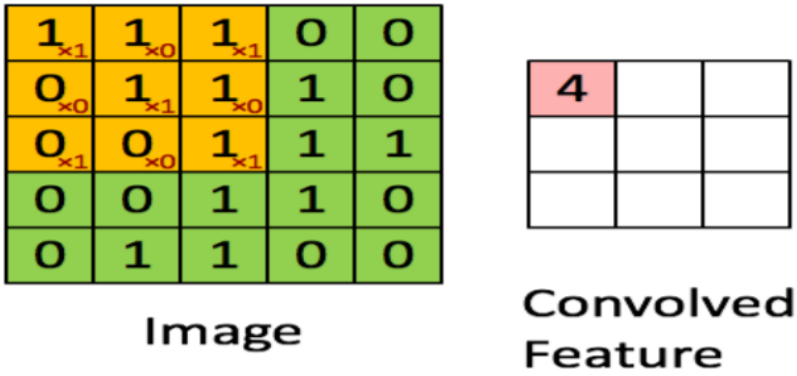


Fig 5: Example of Convolution

Pooling operation: Convolutional neural networks are typically used for image classification. However, images are high-dimensional data - so we would prefer to reduce the dimensionality to minimise the possibility of overfitting.[2] Pooling generally serves three aims: **(1)** it generally acts as a noise suppressant **(2)** makes it invariant to translation movement for image classification **(3)** helps capture essential structural features of the represented images without being bogged down by the fine details. Different types of pooling operations: Max pooling: In this maximum pixel value of the batch is selected

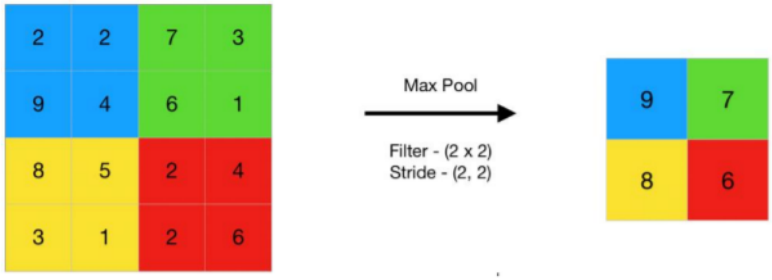


Fig 6: Max pooling operation

Min pooling: In this minimum value of the batch is selected. As shown in Fig 7

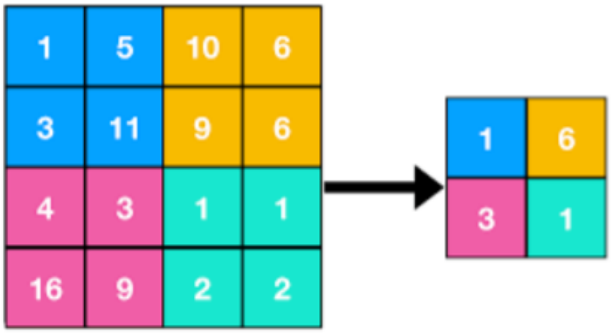


Fig 7: Min pooling operation

Average pooling: In this average value of the batch is selected. As shown in Fig 8

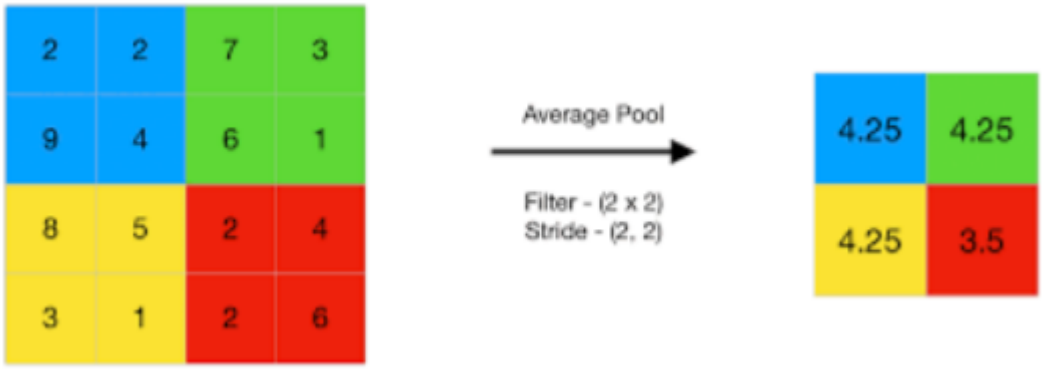


Fig 8: Average pooling operation

3.3 Upsampling:

It's the opposite of downsampling where instead of decreasing the number of rows and columns it increases the number of rows and columns. Thus creating a high-resolution matrix.[3] Below is an example of upsampling by the nearest neighbour (Fig 9).

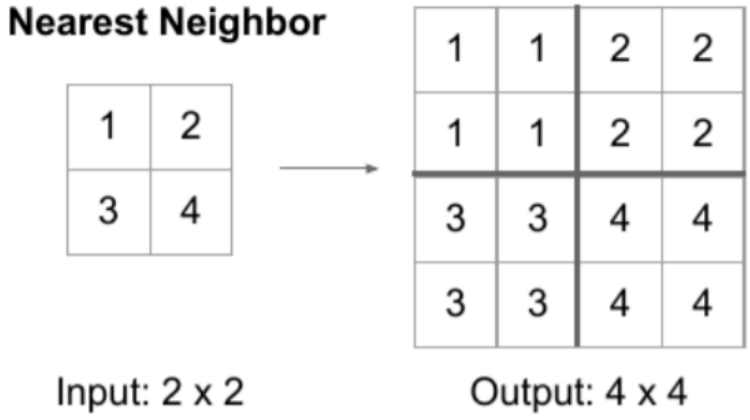


Fig 9: Upsampling by Nearest neighbour interpolation

4. OUR AIM:

Aerial image segmentation using U-Net gives 78% accuracy over the training set and 77% accuracy over the validation set after training for 100 epochs using Focal loss as well as IOU loss. Our aim is to improve the segmentation task using the same U-Net architecture with some modification after the encoder block of the U-Net.We introduced **Vision Transformers[6]** to enhance the performance of the regular U-Net model, by using this modified

architecture we are able to get better accuracy (Pixel level), in this model we are using **Focal Loss[1]** as a loss function.

5. Vision Transformers (ViT):

The Transformer architecture has become the highest standard for tasks involving natural language processing (NLP), but in the task of Computer Vision it has less uses. In Computer Vision generally Convolution Network (CNN) is used. A pure Transformer applied directly to a sequence of image patches can work very well in Computer Vision applications.

Vision Transformers (ViT) recently obtained outstanding results in benchmarks for a variety of computer vision applications, including image classification, object recognition, and semantic image segmentation.

When compared to convolutional neural networks (CNN), Vision Transformer (ViT) offers impressive outcomes while using less computer resources for pre-training. When training on fewer datasets, Vision Transformer (ViT) has a less inductive bias than convolutional neural networks (CNN), resulting in a greater dependence on model regularisation or data augmentation (AugReg).

The ViT is a visual model based on a transformer's architecture[6], which was initially created for text-based operations. The ViT model encodes an input picture as a sequence of image patches, similar to how word embeddings are represented when using text transformers, and predicts class labels for the image directly. When trained on enough data, ViT outperforms an equivalent state-of-the-art CNN using 4x less CPU resources.

ViT separates the pictures into visual tokens, whereas CNN employs pixel arrays. The visual transformer separates a picture into fixed-size patches, embeds each one appropriately, and passes positional embedding to the transformer encoder as an input. Furthermore, ViT models beat CNNs in terms of computing efficiency and accuracy by nearly four times. ViT's self-attention layer allows embedding information globally throughout the entire image. The model also uses training data to represent the relative locations of picture patches in order to recreate the image's structure.

4.1 Transformer encoder

The **transformer encoder[7]** includes:

- **Multi-Head Self Attention Layer (MSP):** This layer linearly concatenates all attention outputs to the correct dimensions. The several attention heads in a picture aid in the training of local and global dependencies.
- **Multi-Layer Perceptrons (MLP) Layer:** This layer contains a two-layer with Gaussian Error Linear Unit (GELU).
- **Layer Norm (LN):** Because it does not include any additional dependencies between the training pictures, it is added before each block. As a result, the training time and overall performance are improved.

Additionally, residual connections are provided after each block because they allow components to flow directly through the network without having to go through non-linear activations.

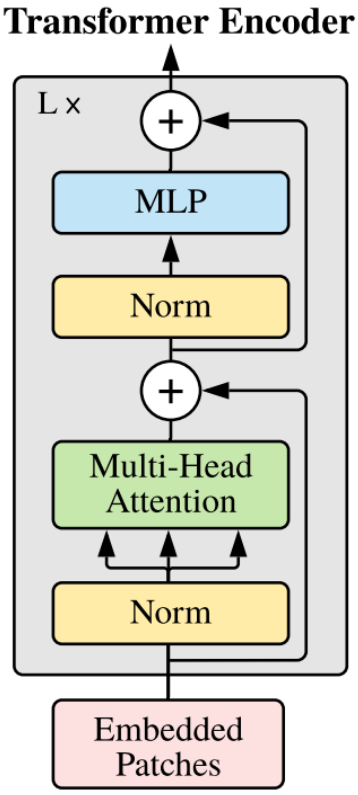


Fig 10 : Transformer encode

The MLP layer implements the classification head in the instance of image classification. At pre-training time, it uses one hidden layer and a single linear layer for fine-tuning.

4.2 Vision Transformer ViT Architecture

The overall architecture of the vision transformer[7] model is given Below step by step

1. Split an image into fixed size patches.
2. Flatten the image patches.
3. Create lower-dimensional linear embeddings from these flattened image patches.
4. Include positional embeddings.
5. Feed the sequence as an input to a state-of-the-art transformer encoder.
6. Pre-train the ViT model with image labels, which is then fully supervised on a big dataset.
7. Fine-tune on the downstream dataset for image classification.

An overview of the model is depicted in Figure 11.

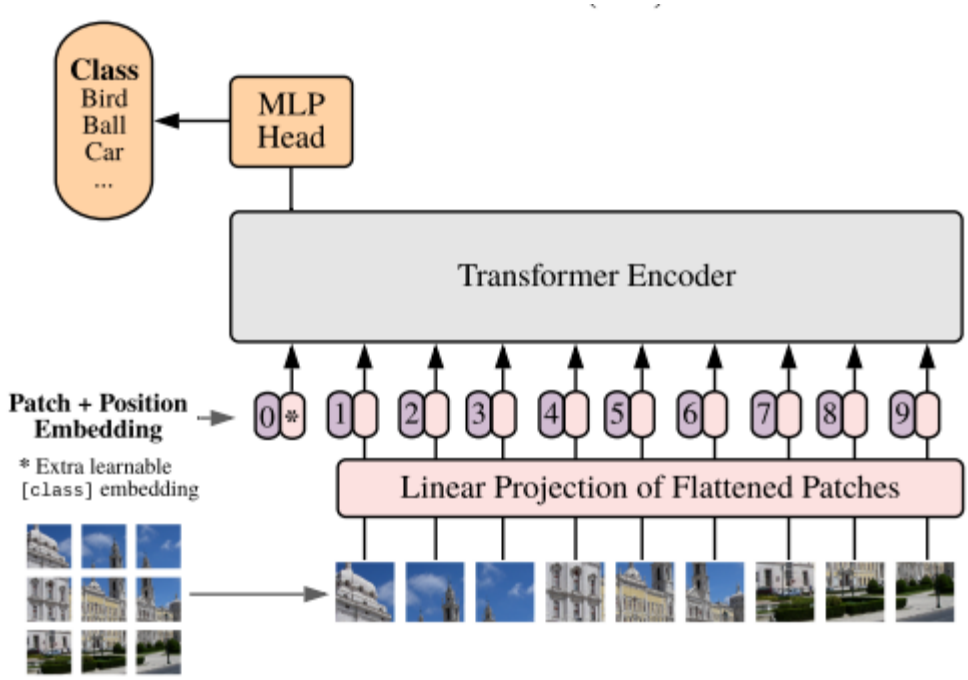


Fig 11 : overview of the vision transformer model

6. Our Work

In this project we introduce the Vision Transformer[6] (ViT) in the U-Net[1] architecture. We have attached the ViT after the encoding block of the U-Net.

6.1 Architecture

In the architecture we have 4 different phases.

1. Contraction phase
2. Vision Transformer
3. Convolution and linear operation
4. Expansion phase

6.1.1 Contraction phase

In each block, maxpooling operation is used,then convolution operation with ReLU activation function is applied two times with kernel size of 3 x 3 and stride 1. Batch normalisation is used after each convolution operation.

From fig 12, it is clear that there are 5 blocks in the contraction phase. At the block **C1** the input is an image having dimension 3×512×512 and after applying 2 convolution layer the dimension becomes 64×512×512. Similarly for the block **C2,C3,C4,C5** the input image shape is 64×512×512, 128×256×256, 256×128×128, 512×64×64 and the output image shape is 128×256×256, 256×128×128, 512×64×64 and 512×32×32 respectively.

So in the contraction phase the image size (height and width) is decreasing and the number of channels increasing.

6.1.2 Vision Transformer

Vision transformers are generally used for classification problems, but here we use in the middle of the U-Net so that our model can capture more information about the images by splitting into small patches. So we remove the last part of the vision transformer which gives the probability value for each channel. In this module we use depth 2 that means we applied a 2 time Transformer Encoder.

The input image shape is 512×32×32 and we divide each image into patch size 8. After completion of this phase the output image shape is 2048, then reshape operation is performed to reshape the image into the shape 32×8×8.

6.1.3 Convolution and linear operation

As we got 32×8×8 size image from ViT but we need an input image of size 512×32×32 for the next phase so we used a convolution layer with kernel

size 1 and padding 0 just to increase the number of channels from 32 to 512. And then we use a linear layer to increase image height and width from 8×8 to 32×32, then we did reshape operation to get the image size 512×32×32.

6.1.4 Expansion phase

In each block,upsampling operation with scale factor 2 is used, here we use 'bilinear' mode. Then, convolution operation with ReLU activation function is applied two times with kernel size of 3 x 3 and stride 1. Number of filters is now decreased. The feature maps from the block in the contraction path having similar resolution as the initial feature matrices of a block in the expansion path are concatenated.

From the fig 12 it is clear that there are 4 blocks in the expansion phase. In the **C6** block the input image comes from the convolution block and **C4** block, the shape of image comes from convolution block is 512×32×32 and the image shape comes from **C4** block is 512×64×64 so to mach the height and width of the image the upsampling method is used with scale factor 2, thus the 512×32×32 image become 512×64×64, then the two images are concatenated channelwise and the input image shape of the block **C6** become 1024×64×64. And the output image shape of this block is 256×128×128 and that goes to block **C7** along with the output image of block **C3**. All the blocks **C7,C8,C9** performs the similar operation, at the end of the block **C9** the output image shape becomes 64×512×512 then we apply one convolution layer to reduce the channels from 64 to 6, thus the shape of the output image become 6×512×512.

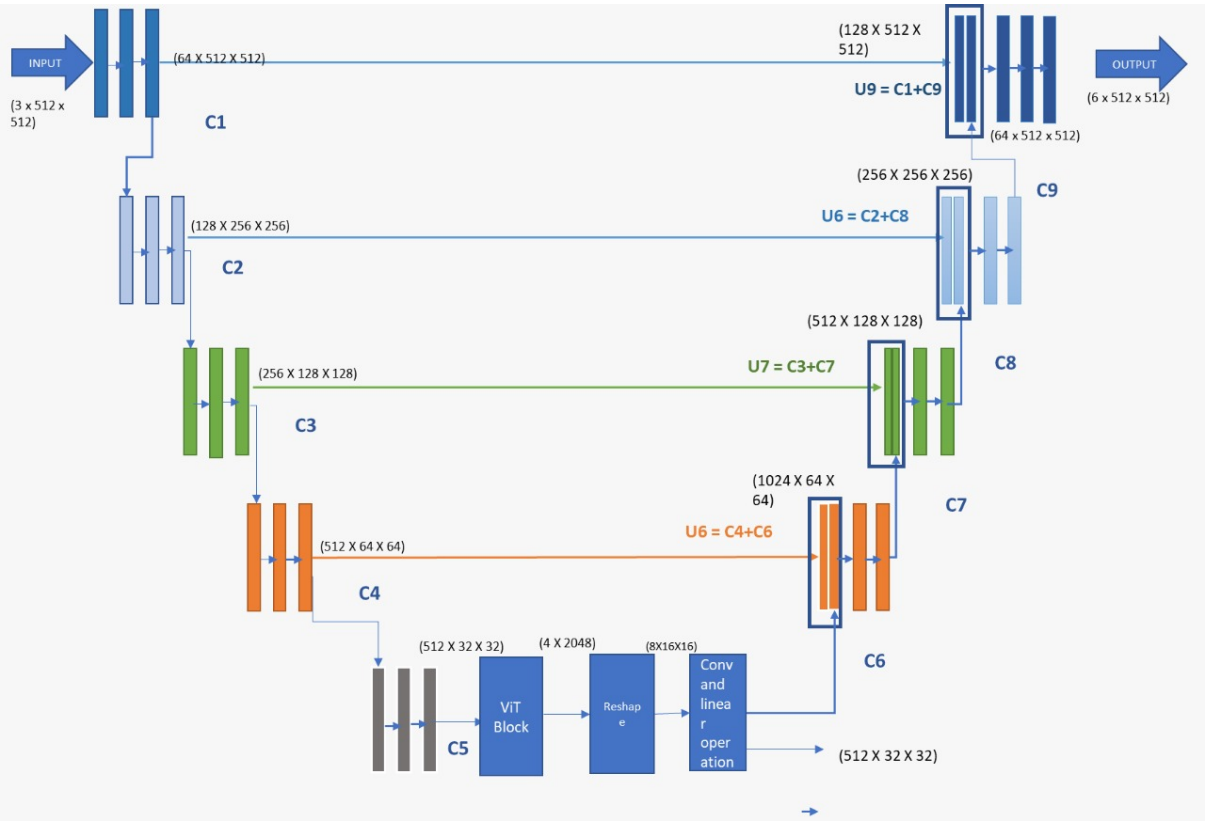


Fig 12 : U-Net with ViT architecture

6.2 Experimental setup:

1) **Training phase:**

To reduce overfitting data augmentation is needed, Data augmentation is essential to teach the network the desired invariance and robustness properties when only a few training samples are available. In the case of the satellite image, we can get more images by horizontal and vertical flip and by doing a 90-degree rotation in each image.

2) **Testing phase:**

To test an image all two types of flips are done (horizontal, vertical), and 90-degree rotation is also performed, and the result is predicted, then the average is done based on the 3 predicted results and predicts the final result.

3) **Experimental setting:**

Mini batch gradient descent is used with Adam optimizer using batch size 4.

4) **Loss Function:**

Focal loss presents a better solution to the unbalanced dataset problem. It adds an extra term to reduce the impact of correct predictions and focus on incorrect examples. The gamma is a hyperparameter that specifies how powerful

this reduction will be. This loss influences the training of a network on the unbalanced dataset and can improve segmentation results. We used this loss function to train the model using the architecture U-Net, CE-net, and U2-Net. The equation of the focal loss function is given below.

$$FL = -\alpha_c(1 - \hat{y}_i)^{\gamma}.log\hat{y}_i$$

6.3 Comparison result:

We segment the aerial images by applying different models U-Net, Multi Scale U-Net, U2-Net and also applying ViT in all models, and we observed that from the table I after applying ViT in each model the pixel level accuracy increases in all the models. For example the accuracy on validation set using **U-Net** is **77%** but using **U-Net with ViT** it becomes **80%**, and for Multi scale U-Net it is 78% but using Multi Scale U-Net with ViT the accuracy increases to 80%, similarly using U2-Net the accuracy is 80% and using U2-Net with ViT the accuracy becomes 81%.

Method	Avg Loss on the training set	Accuracy on the training set (%)	Avg Loss on the validation set	Accuracy on the validation set (%)
U-Net	0.55374	76	**0.53489**	**77**
U-Net with ViT	0.47916	80	**0.49367**	**80**
Multi scale U-Net	0.50959	79	**0.49951**	**78**
Multi scale U-Net with ViT	0.47392	80	**0.46848**	**80**
U2-Net	0.46791	80	**0.45417**	**80**
U2-Net with ViT	0.40127	82	**0.41523**	**81**

Table I : accuracy and loss function using different models

So from table I it is clear that ViT improves semantic segmentation tasks along with U-Net and different families of U-Net.

Some sample results of aerial image segmentation are given in figure 13 and figure 14.

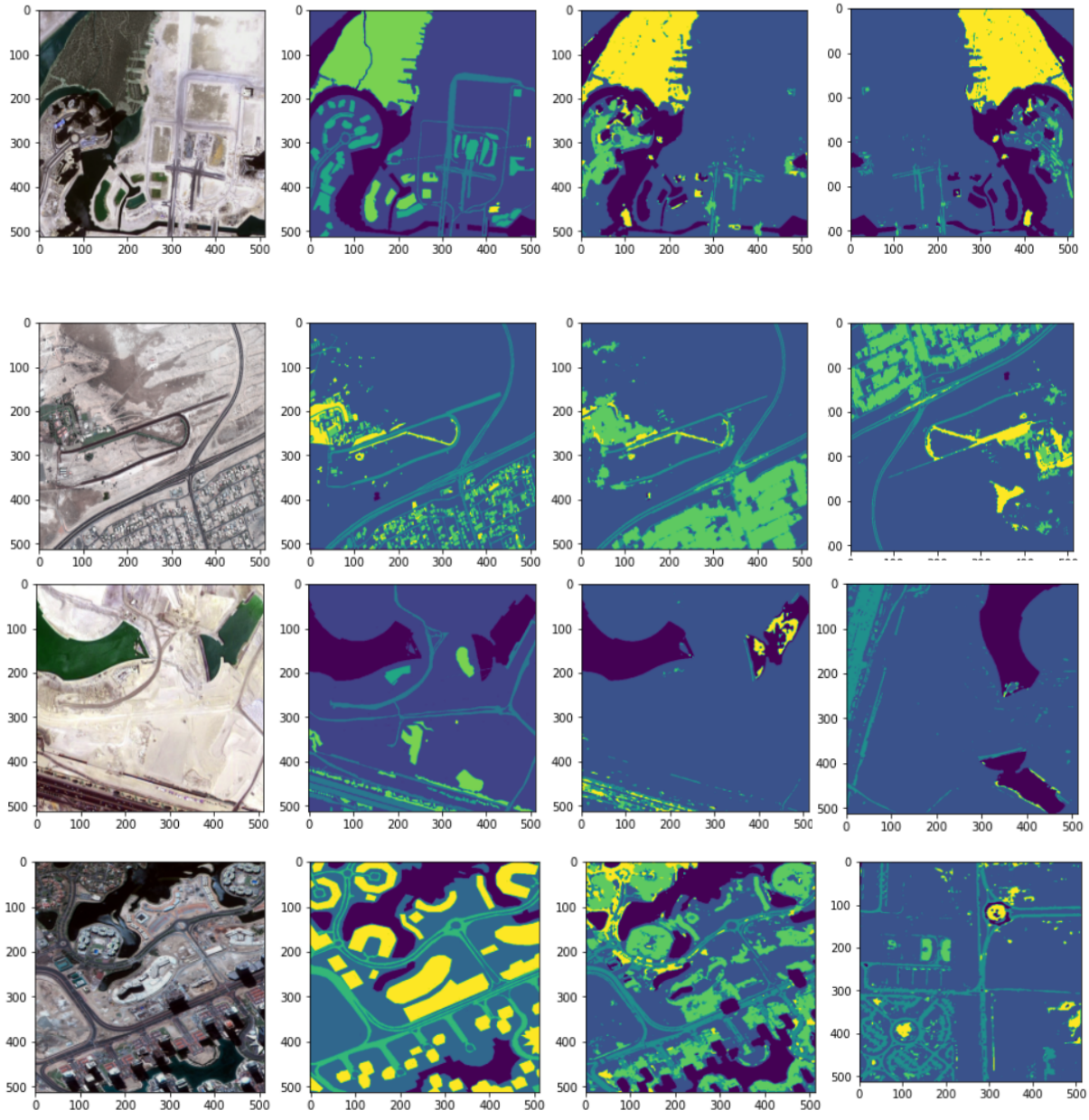
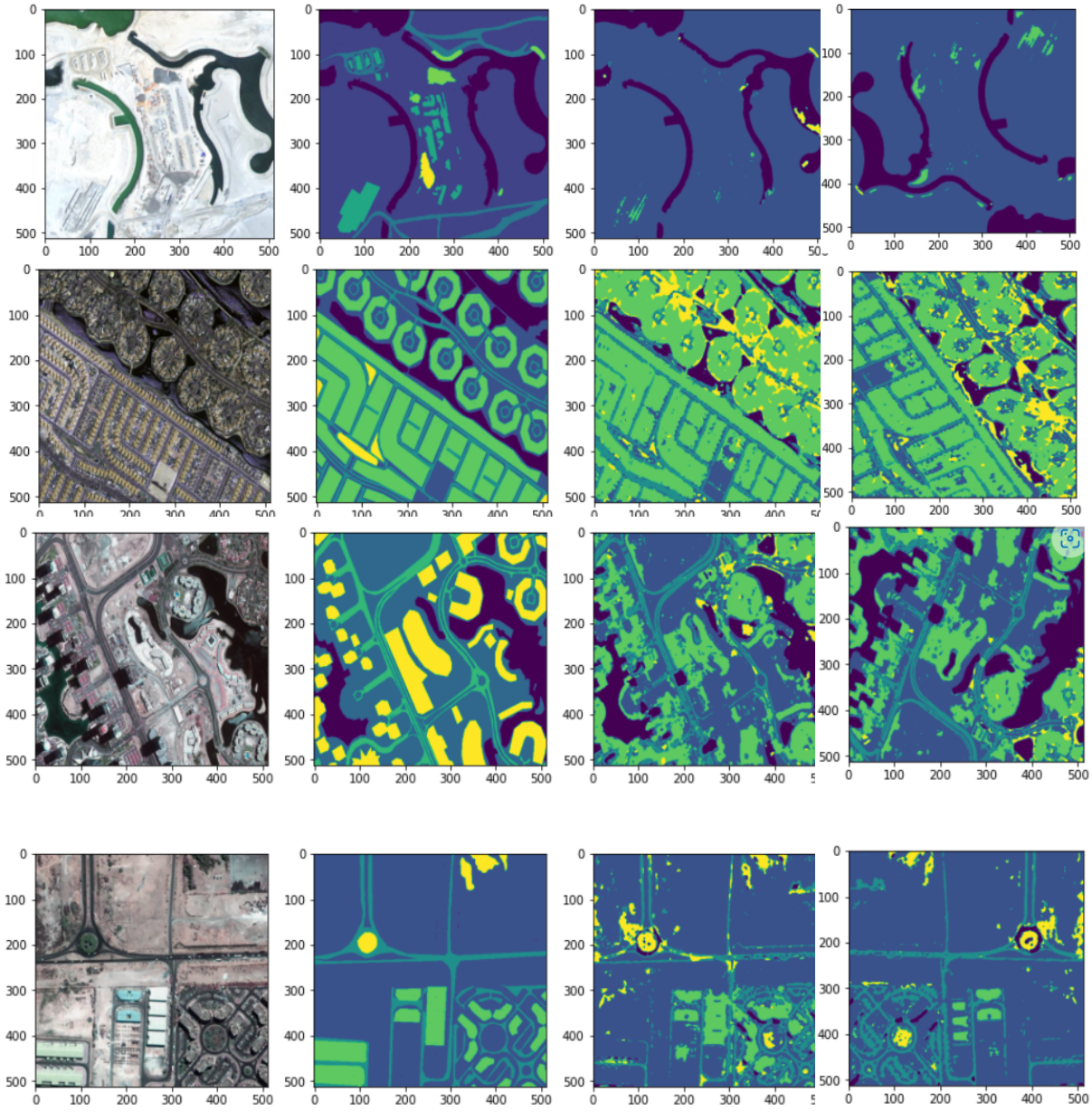


Fig 13 : Semantic segmentation results using U-Net and U-Net with ViT

From figure 13 it is clear that the roads, buildings are more clear in the U-Net[1] with the ViT[6][7] model than the U-Net model.



Original image **actual mask** **multi scale U-Net** **multi scale U-Net with ViT**

Fig 14 : Semantic segmentation results using Multi scale U-Net and Multi scale U-Net with ViT

From figure 14 it is clear that the roads, buildings are more clear in the U-Net with ViT model than the U-Net[1] model.

The table II and table III describes the precision and recall of each class in 4 different models.

Model	Water	Land	Road	Building	Vegetation	agerage
U-Net	0.76649904	0.81209689	0.544491	0.537062	0.342054	**0.60044089**
U-Net with ViT	0.7358207	0.81072872	0.58834424	0.60163444	0.32414858	**0.61213533**
Multiscale U-Net	0.78258321	0.807667	0.505644	0.573172	0.375326	**0.60887888**
Multiscale U-Net with ViT	0.696906	0.812947	0.631300	0.515502	0.418236	**0.61497871**

Table II : Precision table of each class for 4 different models.

Model	Water	Land	Road	Building	Vegetation	agerage
U-Net	0.63352362	0.79230528	0.35055865	0.46189066	0.31417442	**0.51049052**
U-Net with ViT	0.76392337	0.79320335	0.41827383	0.41233565	0.44072133	**0.56569150**
Multiscale U-Net	0.631266	0.807267	0.408021	0.440010	0.378402	**0.53299351**
Multiscale U-Net with ViT	0.708448	0.807906	0.389537	0.464898	0.311092	**0.53637663**

Table III : Recall table of each class for 4 different models.

From table II and table III we can see that the average precision and recall increases after introducing ViT in the U-Net and multi scale U-Net.

7. Conclusion

So from this project we are concluding that

1. ViT enhances the capability of U-Net and some other models in the U-Net family in the image segmentation task.
2. The overall precision and recall increases in the new models introduced using ViT.
3. From table II we can also see that for all the classes the precision and recall does not increase in the new model, but the average precision and recall increases.

8. References

[1] U-Net: Convolutional Networks for Biomedical Image Segmentation Olaf Ronneberger, Philipp Fischer, and Thomas Brox C

[2]https://machinelearningmastery.com/upsampling-and-transpose-convolution-layers-for-generative-adversarial-networks/

[3]https://stats.stackexchange.com/questions/252810/in-cnn-are-upsampling-and-transposeconvolution-the-same

[4]CE-Net: Context Encoder Network for 2D Medical Image Segmentation

[5] U2 -Net: Going Deeper with Nested U-Structure for Salient Object Detection Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane and Martin Jagersand University of Alberta, Canada

[6] AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

[7] Vision Transformers (ViT) in Image Recognition - 2022 Guide - viso.ai