

NATIONAL INSTITUTE OF TECHNOLOGY **UTTARAKHAND**

MAJOR PROJECT **ON**

Automatic Building Extraction on High-Resolution Remote Sensing Imagery Using Transfer Learning Based Encoder-Decoder Architecture



SUBMITTED TO:

Dr. Hariharan Muthusamy

SUBMITTED BY:

Sathyam Rajpal

BT17ECE053

ACKNOWLEDGEMENT

First and foremost, I would like to thank my guide Dr. Hariharan Muthusamy for guiding me thoughtfully and efficiently through this project, giving me an opportunity to work at my own pace along my own lines, while providing me with very useful directions whenever necessary.

I would also like to thank the project co-ordinator Mr. Nitanshu Chauhan for his valuable inputs and all other professors who directly or indirectly helped me to be able to complete this project. Also, I would like to acknowledge that all the work research and work done in this project is done by me and the results are not submitted to any other institution.

ABSTRACT

Automatic extraction of buildings from remote sensing imagery is of great significance for many applications, including urban planning, navigation, and disaster management. In recent years there have been many advancements in the field of computer vision which has led to the improvement in the capabilities of remote sensing techniques. As a result, there has been a significant increase in the availability and accessibility of high-resolution remote sensing images.

Hence, with the increased availability of high-quality data for spatially large areas, it is possible to perform accurate image segmentation targeting the extraction of buildings. With the recent availability of commercial high-resolution remote sensing multispectral imagery from sensors such as IKONOS and QuickBird, it is possible to identify small-scale features such as individual roads and buildings in the urban environment. Imagery from these sensors is an important source of timely data that can be used in a variety of urban area applications. In this paper we propose a transfer-learning based CNN architecture CNN based machine learning architecture to extract the footprints of the buildings in a given set of satellite images.

Table of Contents

ACKNOWLEDGEMENT	2
ABSTRACT	3
1. INTRODUCTION	5
1.1. Background	5
1.2. Motivation	6
1.3. Report Structure	6
2. LITERATURE REVIEW	7
2.1. Research Based on Clustering and Segmentation Algorithms	7
2.2. Research Based on CNN Architectures	8
2.3. Research Based on other novel approaches	9
3. METHODOLOGY	10
3.1. Model Architecture	11
3.2. Dataset	12
3.3. Model Training	12
4. MODEL EVALUATION	14
4.1. Experimental Setup	14
4.2. Evaluation Metrics	14
4.3. Model Testing	16
5. Results	17
5.1. TIFF images with Augmentation Scheme 1	20
5.2. TIFF images with Augmentation Scheme 2	21
5.3. PNG images with Augmentation Scheme 2	22
6. DISCUSSION	25
6.1. About the proposed model	25
6.2. Limitations	26
7. CONCLUSION	27
8. REFERENCES	28

CHAPTER – I

INTRODUCTION

I [A]-BACKGROUND

Currently, the extraction of road networks and building footprints from high-resolution imagery is done manually, and this is both times consuming and expensive. Automated and semi-automated methods for the classification of roads, buildings, and other land cover types in the urban environment are therefore of great interest. There have been numerous researches on various public datasets to extract building footprints.

Recently, there has been a tremendous increase in availability of public datasets as well as their quality of photos, so there is a growing need of efficient methods to exploit the topographical data using advanced methods. However, the diverse characteristics of buildings including color, shape, material, size, and the interference of building shadows and vegetation still make accurate and reliable building extraction of buildings is a challenging task. Land development information is very essential for urban planning and its management. This information is required for policy-making of land use, calculation of transportation infrastructure, and in the detection of the future development area. This information is required to achieve sustainable urban development. Unauthorized land development is one of the growing problems in developing countries. Unauthorized land development caused sprawl which in turn affect healthy urban development.

With the help of these building detection techniques, governments or concern authorities can prevent unauthorized land developments and stop these at their early stage if any. The datasets to be used in this project is Massachusetts Buildings dataset. This dataset is selected because it covers different imagery characteristics such as spatial resolution, object types, shapes, and sizes. Massachusetts buildings dataset consists of 137 training, 4 validation, and 10 testing images, covering a surface of 2.25 square km of urban and suburban areas of Boston (MA) in the United States of size 1500x1500 pixels covering urban and suburban regions at the area of Boston. Each image covers an area of 2.25 km at a resolution of 1 m² /pixel.

I [B] – MOTIVATION

Building detection in a remote sensing image is an instance segmentation task. We need to first classify if an object in an image is a building or not and then a mask depicting the building boundary is created. Object detection in machine learning using Convolutional Neural Networks (CNN). CNN extract high- and low-level details from the images by repeated convolution with several feature maps followed by pooling operations. Since, remote sensing images carry a lot of spatial information effective extraction of information from these images require very deep CNN's consisting of many layers of convolution and pooling that requires a lot of computation.

To deal with the potentially complex texture of buildings in general and image background, the existing methods try to perform extensive pooling and striding operations used in CNNs which reduces feature resolution causing a loss of detailed information that results in less accurate predictions. A large architecture also means that we need to use much computing power to train the model to get meaningful results. To address this issue, we intend to build a light-weight deep learning model integrating the ResNet-50 architecture as the encoder which is already been trained on ImageNet dataset and finally up-sample the images to the same size to get the final output as binary images.

I [C] – REPORT STRUCTURE

The following chapters of this report are stacked as follows: The literature review of the existing researches is done in Chapter II. The methodology of the proposed model is elaborated in Chapter III which in turn is further divided in subsections discussing the (i)Model Architecture, (ii)Dataset used, (iii)Training procedure. Chapter IV describes the testing and evaluation procedure of the trained model with subsections describing (i) Experimental setup, (ii) Evaluation Metrics, (iii) Model testing setup. In Chapter V, results are discussed along with the results obtained using different testing configurations. In Chapter VI we provide a brief discussion on the methodology of the model and throw light on the model's significance, limitations and future prospects to improve upon this research. Chapter VII presents the final thoughts and conclusion followed by the references in Chapter VIII.

CHAPTER - II

LITERATURE REVIEW

In recent years with the gradual development of computational capabilities, researchers are now properly equipped to implement abstract ideas for finding new insights and developing new methods to overcome the flaws of the previous methods.

II [A] – RESEARCH BASED ON CLUSTERING AND SEGMENTATION ALGORITHMS

Over the past few decades many methods are deployed which use segmentation and clustering methods [1-5]. In segmentation algorithms, a picture is a set of various characteristically similar and different pixels. We bunch together the pixels that have comparable properties utilizing image segmentation. We can divide or segment the picture into different parts called segments. Many times, in an image there are very low-level details that cannot be identified and analysed by analysing the image as a whole, but when the image is segmented, a particular patch can be analysed or processed separately. By isolating the image into sets, we can utilize the significant portions for analysing the image. Segmentation makes a pixel-wise mask for each item that is identified in the image. The clustering algorithms forms the segments of a given image on the basis of the similarities of the pixels. It is basically an unsupervised method to cluster the pixels together forming segments which can be processed later.

Wang et al. [1] present an approach to automatically extract the buildings from very high-resolution images using various image primitives such as lines and line intersections. It uses EDLines algorithm to find segments which are hierarchically grouped as candidate rectangular buildings since buildings are generally rectangular in shape viewing from the top.

Mirhassani et al. [2] to increase the accuracy of the existing qualitative model for Bayesian classification. The researchers use an existing classifier trained on a remote sensing dataset using Bayesian theory which creates a network of an acyclic graph which are directed in nature to classify objects as buildings based on various parameters and their conditional probabilities.

In [3], Wei et al. propose a two-step algorithm which involves unsupervised clustering by analysing histogram peaks in the image and identifying the candidate buildings using the shadows obtained then applying Canny Edge detection on Hough transform of the images.

Izadi et al. [4] propose a method in which images are first segmented using a hierarchical segmentation method based on colour and several geometrical attributes are identified in the segmented image to flag potential candidate buildings. Finally, the candidate buildings are verified based on the overlap between the predicted shadow cast and the actual shadow information analysed from the geometrical attributes.

In [5], Pan et al. proposed the segmentation of the image using the Mean Shift Algorithm and then applying SIFT (Scale Invariant Feature Transform) and finally an adaptive windowed Hough Transform is applied which extracts the straight edges, hence, approximating the building rooftop.

II [B] – RESEARCH BASED ON CNN ARCHITECTURES

In recent years with the abundance of aerial satellite images, the high-resolution images are readily available and CNN's are a very viable technique to process the images and find insights. In CNN an image is convolved with many filters, each filter designed to extract a different feature from an image. These feature maps if necessary, can be passed through another convolutional operations. As the depth of the network, we can extract more hidden features. After every convolution layer a pooling layer is present to reduce the dimensionality of the as well as only retain the most robust features. Now, the features in the form of the matrices are flattened so it can be trained as a fully connected layered and for training the neural networks, an activating function is chosen that brings non-linearity to the data. CNN are very powerful architectures and are basic building blocks for image processing tasks. Hence, many standard pre-trained CNN models are present like Inception, VGG-16, MobileNet, ResNet which are pretrained on millions of images to detect the objects. So, these standard architectures enable researchers to use them directly for their specific image segmentation or classification tasks [6-12].

Ngo et al. [6] propose a method based on the position of the shadows and then a grouping of similar regions. The similar grouped regions with a shape similar to rectangles are considered as primary candidates. The buildings which have a rectangularity score greater than a set threshold and contain more than one building segments are finally classified as buildings.

Li et al. [7] extract buildings from remote sensing images by incorporating saliency cue bases segmentation with the help of CNN.

Qi et al. [8] employ spatial pyramid pooling instead of traditional iterative pooling operations that increase the efficiency and then make use of an encoder-decoder network to reconstruct the lost image information.

Huang et al. [9] first trained their deep CNN model on the dataset and fine-tuned it by RGB and NRG colour bands. Finally, the saliency maps obtained from the two models are integrated to produce the final result.

Li et al. [10] propose a novel method for building detection in dense urban, suburban and rural areas. The proposed method is a cascaded deep neural network which is used in coherence with a multi-stage Region Proposal Network (RPN) for carrying out segmentation of the image.

In [11] Inglada et al. employ a supervised classification of the images which is followed by segmentation of the regions of interests and finally post-processing of the segments is carried out.

Chen et al. [12] developed a connected network of 27 convolution and deconvolution layers to extract the rooftops of the buildings on the pixel-level.

II [C] – RESEARCH BASED ON OTHER NOVEL APPROACHES

In [13] Unsalan et al. created a 5-layer architecture which is based on spatial voting to identify the road pixels in the images.

Xiao et al. [14] have proposed a semi-automatic method for building extraction. Instead of searching and computing over the whole image, the researches instead find other features of the high-resolution images such as roads and vegetation and finally search for the buildings in the areas that are not classified as road or vegetation, and a multi-level approach is devised to efficiently extract the buildings.

Zhang [15] propose a new postprocessing framework which is an extension of the existing Morphological Building Index (MBI).

Lee et al. [16] propose automatic building extraction method from Panchromatic (PAN) and Multi-Spectral (MS) images using a semi-automatic volumetric shadow analysis, through the analysis of the shadow casts, the height of a building is along with some spectral information is calculated to estimate a maximum cost value.

CHAPTER – III

METHODOLOGY

In this paper, an encoder-decoder based architecture is proposed for extracting the buildings from a high-resolution satellite image. The encoder-decoder is a U-Net and the encoder part is replaced with existing resnet-50 model which has been previously trained on ImageNet dataset. The U-Net is an architecture that is mostly used for biomedical image segmentation tasks. The model outline is portrayed in fig.1

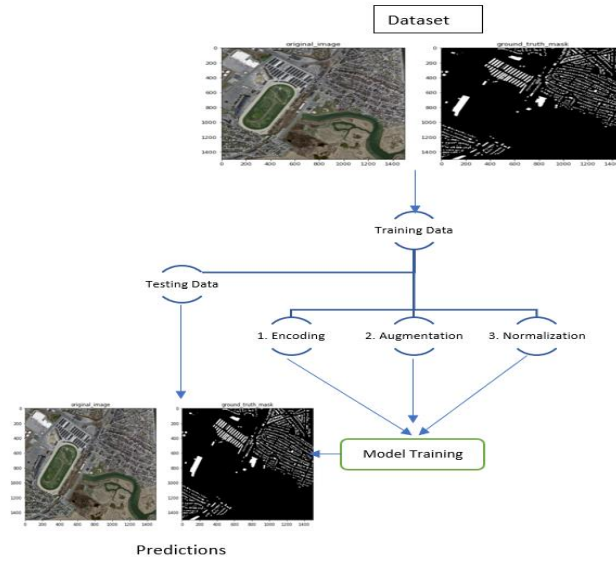


Figure.1 Model workflow.

The method to extract the buildings from the high-resolution images is based on the concept of Transfer Learning.

We aim to reduce the training time of the model by including the weights of resnet-50 model that is trained on the ImageNet dataset for object detection. The biases are stored and now used in another model with slightly different output needs. Transfer learning's fundamental concept is simple: take a model trained on a large dataset and transfer its information to a smaller dataset.

III [A] – MODEL ARCHITECTURE

The special ability of U-Net to give very good results even with less data makes it very appealing architecture for image segmentation tasks [17]. The U-Net is based of 2 stages; Encoder and Decoder. The encoder stage is the similar to a CNN in which there are multiple layers of convolution and consecutive pooling. The encoder is used to extract the features from an image as a result, the image dimensions decrease with the increase in layers and operations. Now to output the images in the same output size, the decoder is used. The decoder stage uses transpose convolutions to upsample the dimensions of the image by adding the pixels between and around the existing pixels. Since, the resolution of the satellite images is very high, there is a lot of information that can be extracted which in turn would require a very deep encoder stage with many layers. We propose a to use concepts of transfer learning to build a light weight encoder-decoder architecture by replacing the encoder part with the resnet-50 model for carrying out the encoder process. This reduces the need for creating the deep encoder layer architecture which will enable the training of the model even will less computing power and similar training times. The decoder stage ensures that the final output image has the same dimensions as that of the input image for comparison.

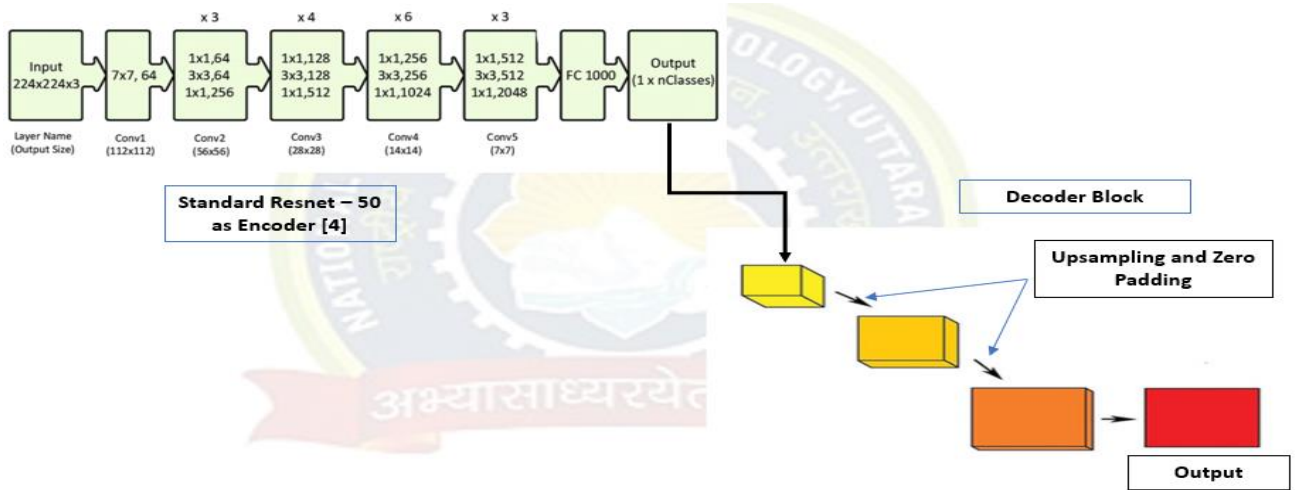


Figure 2. Sample Image from the Massachusetts

The ResNet-50 is a 50 layered convolutional network with several convolutional filter followed by batch normalisation and pooling layers [18]. There resnet architecture is preferred instead of conventional encoder due to the fact that, resnet uses skip connections between the consecutive layers unlike VGG-16/19 or conventional deep CNN's. The addition of skip

connections shows that it performs better always than the corresponding networks with no skip connections. The resnet architecture greatly improves the accuracy of the image classification problem. Finally, we will get the output images which will display buildings and their boundaries as seen from top view.

III [B] – DATASET

The dataset used in this project is Massachusetts Buildings Dataset consists of 151 remote sensing images of size 1500m x 1500m the city of Boston area, each image covers an area of 2.25km. The training set of the given dataset consists of 137 images, which can be augmented using special methods to create an even vast training set and 4 images are kept for validation and 10 images are separated for testing. A sample image is displayed in fig.3



Figure 3. Sample Image from the Massachusetts Building Dataset. (a) Satellite Image; (b) Ground truth mask.

The spatial resolution of the dataset is 1 pixel per meter square and contains photos of urban and sub-urban regions which constitutes to spatial variance. The images in the dataset are rectified for quality by omitting the images with noise levels higher than 5% and manually correcting the building footprints in the target images. The ground-truth image depicts two classes viz. buildings and non-buildings.

III [C] – MODEL TRAINING

The images from the training set are 3-channel images, since instance segmentation of images carried out by classification of pixels so we encode the 3-channel RGB image to 2 channelled images where every pixel represent sparse matrix of 0 and 1 for non-building and building classes respectively. The images and their masks of the training set are now

augmented by two schemes of augmentation parameters and results are calculated for each of the schemes.

- i. Augmentation Scheme 1:** Random application of one of (vertical, horizontal) flipping, clockwise rotation and cropping methods over an image to create 3 images per original image. In this way 411 images are now present on the training set.
- ii. Augmentation Scheme 2:** Random application of one of (vertical, horizontal) flipping, clock-wise rotation, brightness and contrast boost and gamma transformation on the single image. This enables us to create 3 images per original images increasing the training size to 3-fold (411 images) from 137 images, but this scheme has brightness/contrast boosted images as well as gamma transformed images.

The validation images are only pre-processed by application with padding to keep the size same as initial input to compensate for random cropping. The encoder of the U-Net model is selected to be resnet-50 and initial weights are taken from ImageNet dataset. Sigmoid activation function is taken to predict output since it works well for binary classification. The training set is trained in mini-batch sizes of 16 whereas the validation set is trained with batch size of 1.

CHAPTER – IV

MODEL EVALUATION

IV [A] – EXPERIMENTAL SETUP

The implementation of the proposed U-Net model is based on python library Pytorch. All experiments are carried out on computer having Intel i5 7th gen CPU (1.6Ghz), 8GB RAM with no GPU processing. The initial training has been done on training set having mini-batch size of 16 randomly selected images. The training and validation sets were trained for 50-epochs for two kind of image formats i.e., TIFF and PNG and the results are tabulated and compared with the existing state-of-the art architectures.

IV [B] – EVALUATION METRICS

To evaluate the quantitative performance of different CNN methods, the ‘Overall Accuracy’ (OA), ‘Precision’, ‘Recall’, ‘F1-score’, and mean of Intersection-over-Union (‘Mean IoU’) are used as quality metrics. ‘Overall Accuracy’ is defined as the number of correctly classified pixels divided by the total number of test pixels. ‘Precision’ is the percentage of correctly classified positive pixels amongst all pixels predicted as positive. ‘Recall’ is the percentage of correctly classified positive pixels among all true positive pixels. ‘F1-score’ is a combination of precision and recall. ‘Mean IoU’ is applied to characterize the accuracy at the segment level [19]. The values of these metrics are in the range of 0 to 1, and higher values indicate better classification performance. The five metrics can be calculated as follows:

i. *Overall Accuracy:*

$$\frac{\text{Total Correct Predictions}}{\text{Total Predictions}} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

ii. *Precision*

$$\frac{\text{Correct Positive Predictions}}{\text{Total Positive Predictions}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

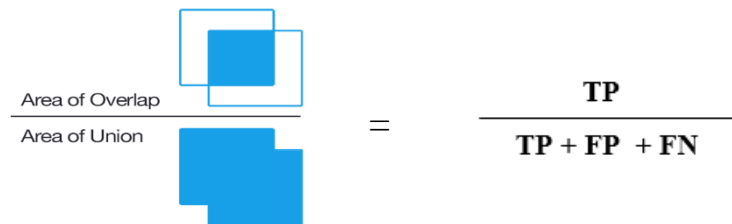
iii. *Recall*

$$\frac{\text{Correct Positive Predictions}}{\text{Total Positive Results in Test Set}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

iv. *F1– score*

$$\text{Harmonic Mean of Precision and Recall} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

v. *Mean IoU*


$$\frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

vi. *ROC – AUC Curve*

$$\frac{\text{Recall}}{\text{False Positive Rate}} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

TP: Number of pixels correctly predicted as pixels belonging to buildings.

FP: Number of pixels non-building pixels predicted as buildings.

TN: Number of pixels correctly predicted as pixels not belonging to building.

FN: Number of pixels belonging to building class wrongly predicted as non-building pixel.

IV [C] – MODEL TESTING

The proposed model is evaluated on Massachusetts Buildings Dataset and the final testing is done on two types of image formats.

(i)TIFF format is the raw format of the remote-sensing images which retains most of the spectral information.

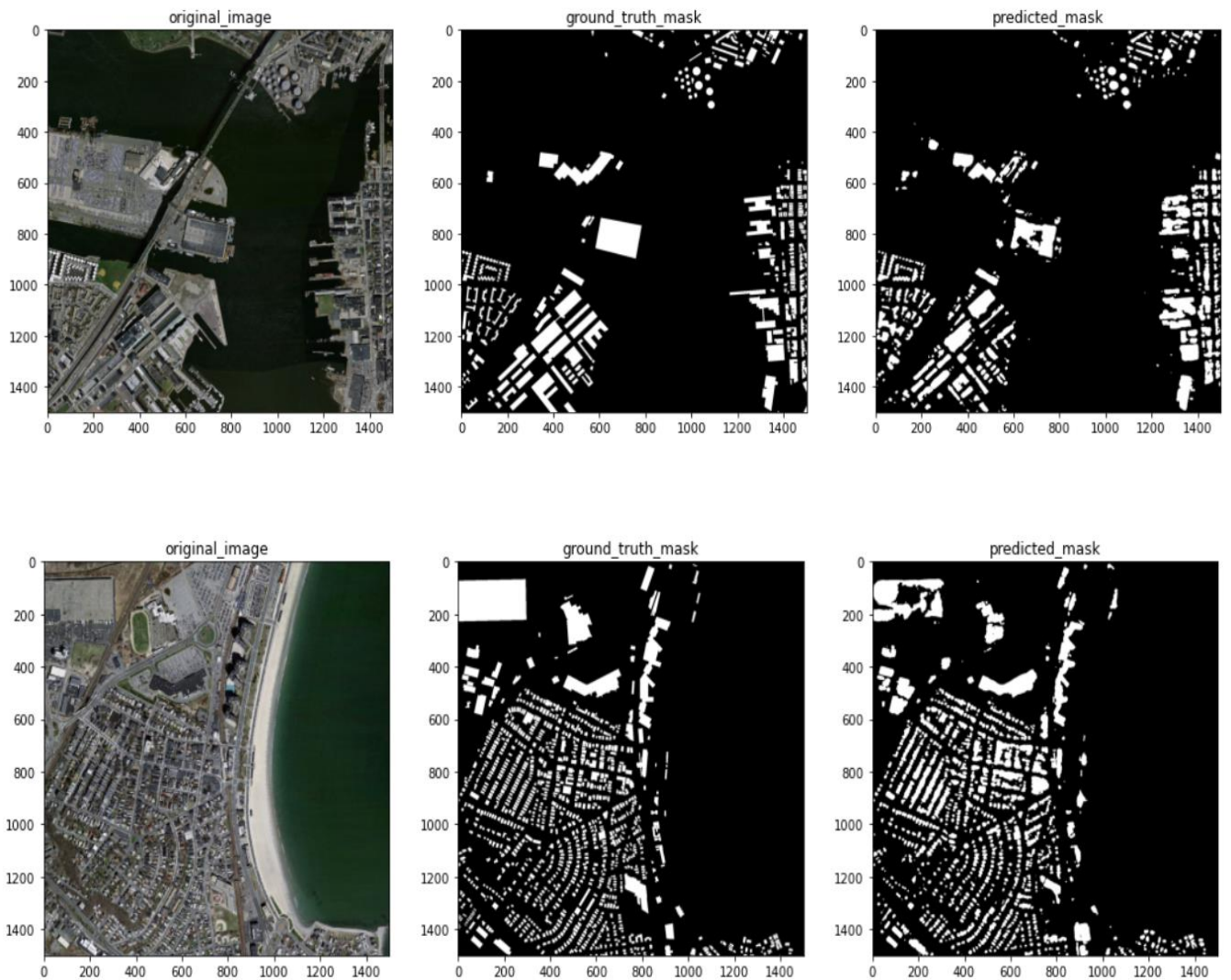
(ii)PNG format is a type of compression technique that reduces the type of the images at a cost of decrease in resolution.

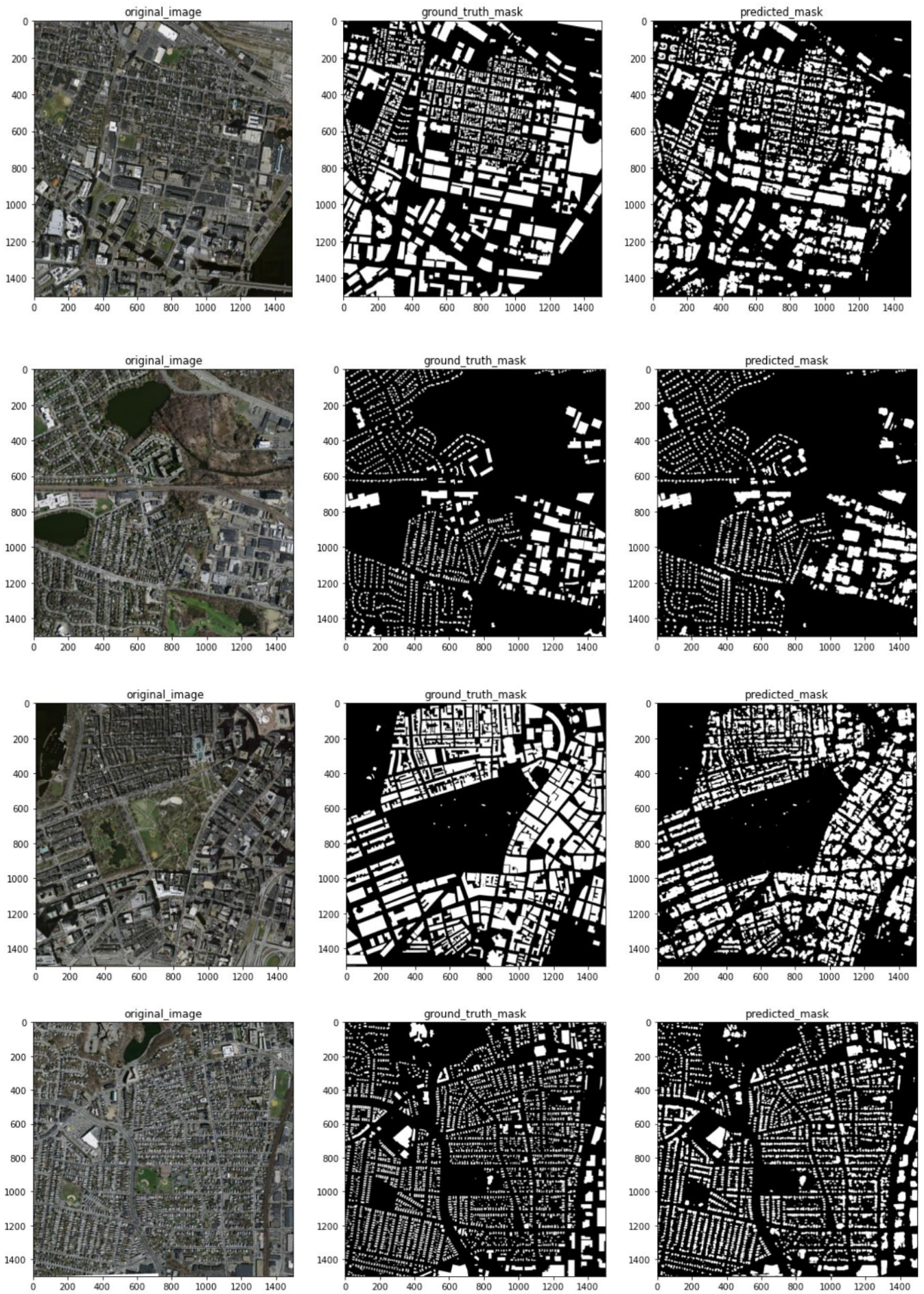
The proposed model is trained using the TIFF images and corresponding results with 2 schemes of augmentation are evaluated. The same trained model is tested on the compressed images in PNG format and the results are calculated when applied no augmentation. The ROC and AUC of the generated ROC curve is calculated.

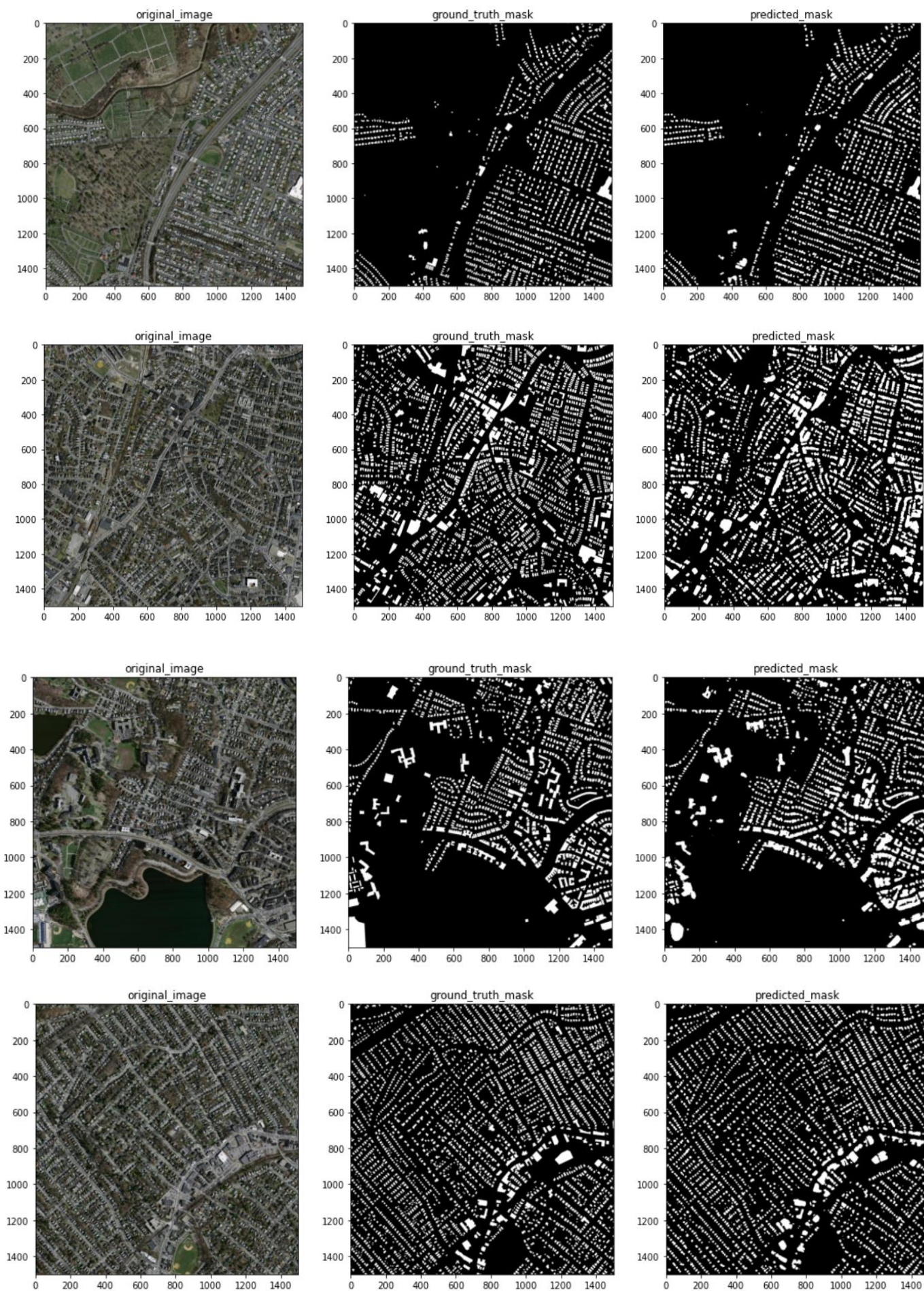
CHAPTER – V

RESULTS

The proposed model based on encoder-decoder architecture with transfer learning is trained for 50 epochs on the TIFF images and the results are evaluated on the same images in TIFF and PNG format with different augmentation schemes applied, the different schemes and the evaluated results are averaged out to give the average overall accuracy of the model and a ROC curve is plotted and AUC over the same curve is calculated. The predictions of the test set images along with corresponding predicted masks are displayed.







V [A] - TIFF images with Augmentation Scheme 1

The evaluated model got IoU Score of 0.799, F1-Score of 0.887, Precision of 0.856, a Recall of 0.920 and Overall Accuracy of 0.888.

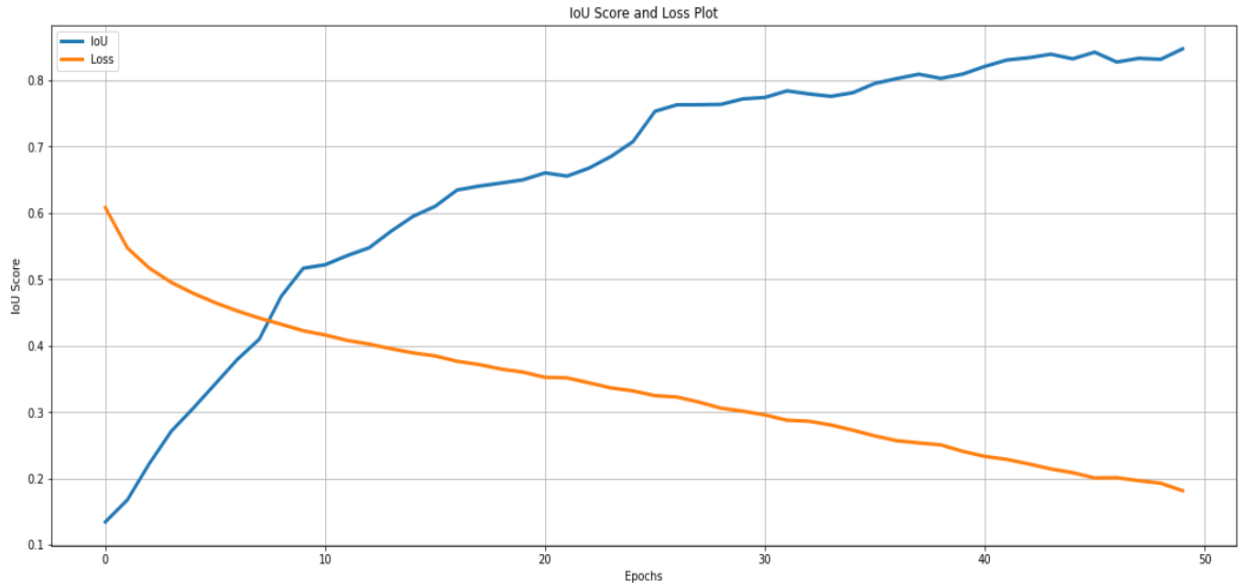


Figure 4(a). IoU and Loss graph w.r.t epoch on training set

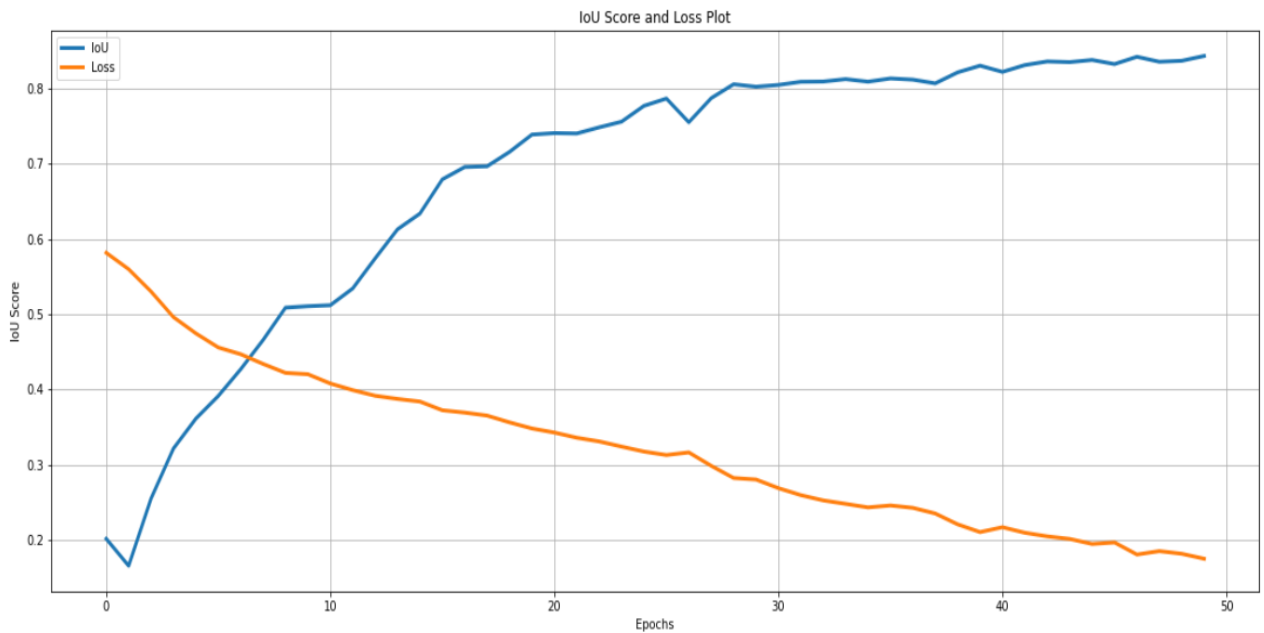


Figure 4(b). IoU and Loss graph w.r.t epoch on validation set

V [B] - TIFF images with Augmentation Scheme 2

The evaluated model got IoU Score of 0.806, F1-Score of 0.891, Precision of 0.867, a Recall of 0.917 and Overall Accuracy of 0.893.

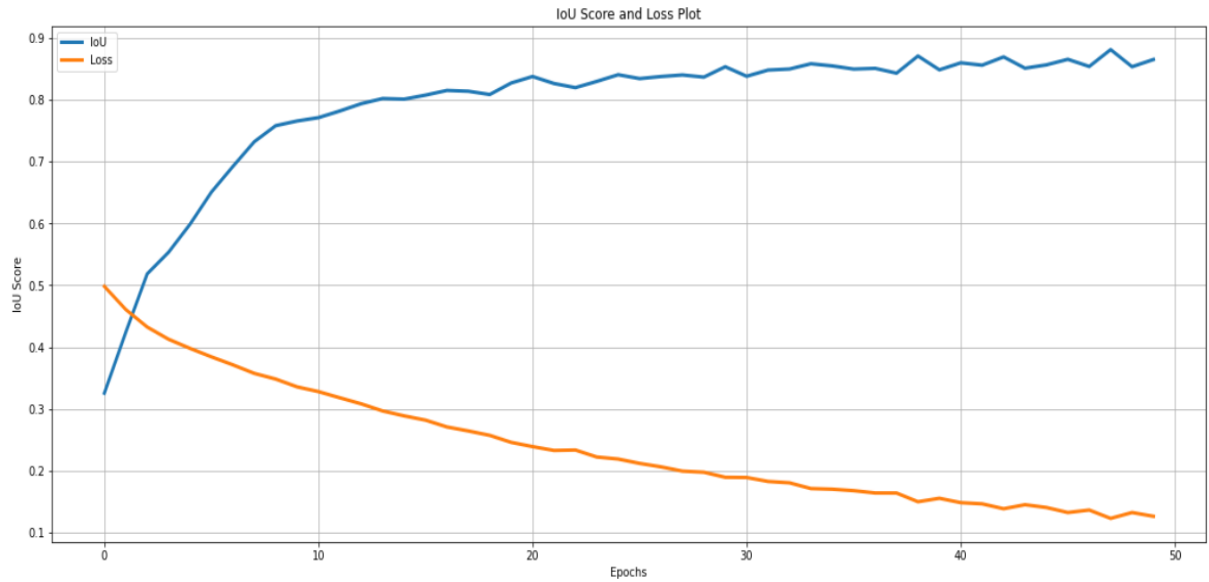


Figure 5(a). IoU and Loss graph w.r.t epoch on training set

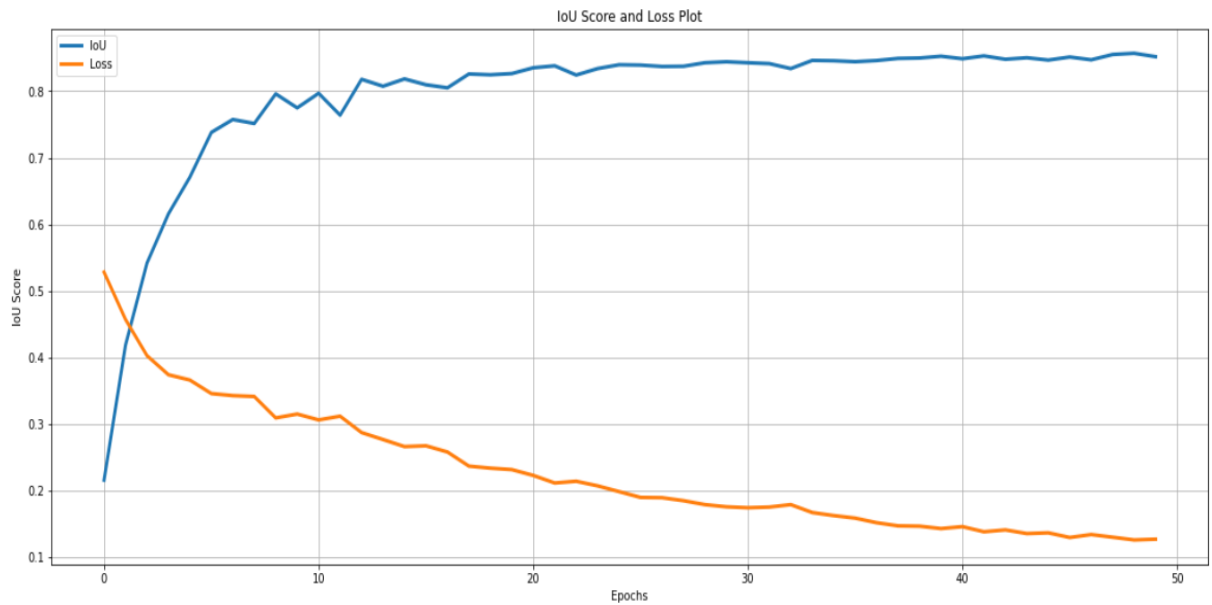


Figure 5(b). IoU and Loss graph w.r.t epoch on validation set

V [C] - PNG images with Augmentation Scheme 2

The evaluated model got IoU Score of 0.818, F1-Score of 0.899, Precision of 0.871, a Recall of 0.93 and Overall Accuracy of 0.901

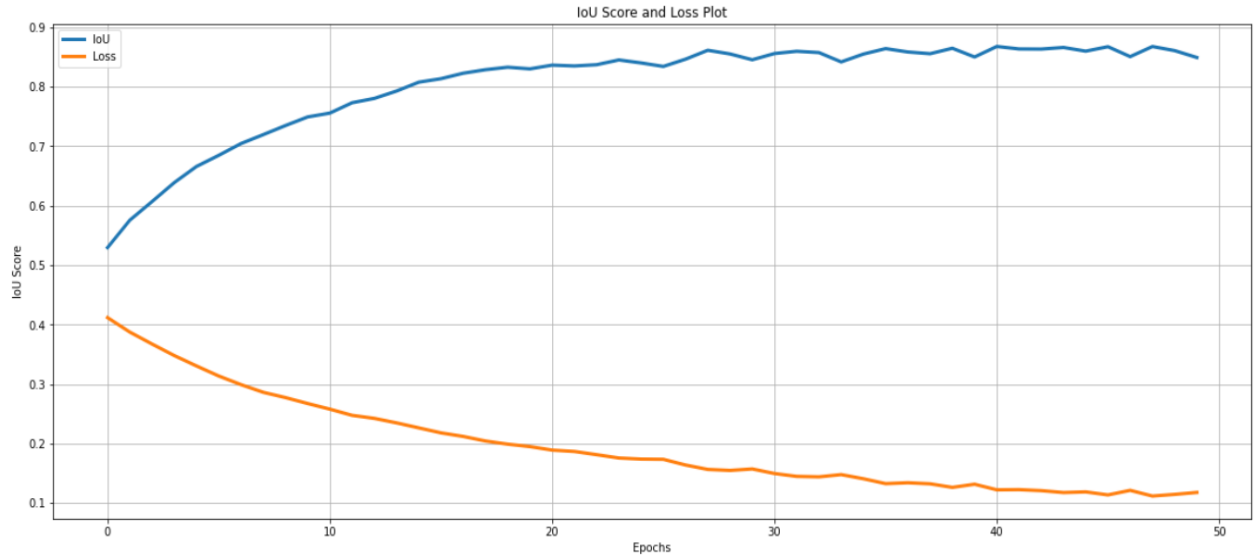


Figure 6(a). IoU and Loss graph w.r.t epoch on training set

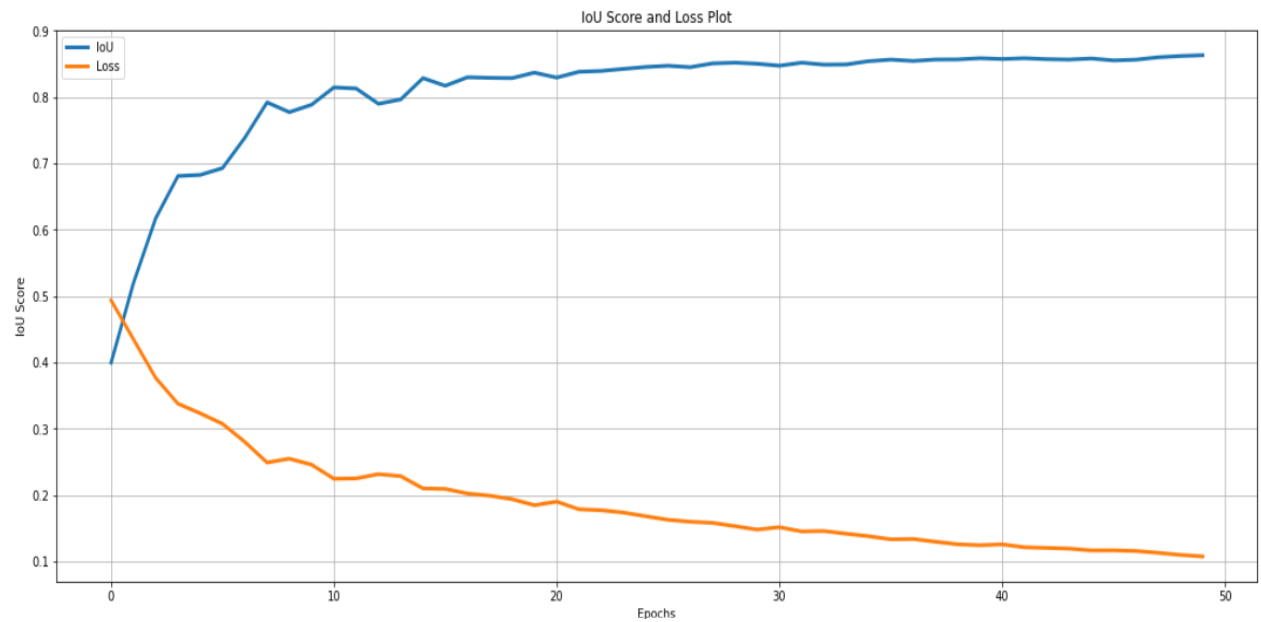


Figure 6(b). IoU and Loss graph w.r.t epoch on validation set

It is evident from the various experiments that the loss and accuracy plots diverge with the increase in number of epochs to the point to which they both become parallel.

The average of the results is calculated to generalise the model performance, the model got Average IoU Score of 0.807, Average F1-Score of 0.892, Average Precision of 0.864, Average Recall of 0.922 and Average Overall Accuracy of 0.894. A comparison between the various established models used for image segmentation is tabulated.

MODEL	Overall Accuracy	F1-Score	Precision	Recall	IoU
USPP	0.913	0.900	0.908	0.892	0.818
FRNN	0.869	0.928	0.796	0.857	0.749
U-Net	0.904	0.899	0.869	0.891	0.803
Seg-Net	0.862	0.882	0.822	0.851	0.740
Tiramisu	0.882	0.903	0.837	0.869	0.768
Proposed Model	0.894	0.892	0.864	0.922	0.807

Figure 7. Performance evaluation with the state-of-the-art image segmentation models [8] compared to proposed model.

The proposed model uses transfer learning to lay out lightweight encoder-decoder architecture and performs very well as compared to best performing models. Proposed model got average overall accuracy of 0.894 vs 0.913, average f1-score of 0.892 vs 0.928, precision of 0.864 vs 0.908, best recall of 0.922 vs 0.892 of the next best model and of IoU 0.807 vs 0.818. The ROC (Receiver Operating Curve) curve that is used to evaluate the performance of the classification model is plotted.

The ROC curve is a probability curve that depicts how much the model is capable of distinguishing between the classes. The Area under Curve (AUC) shows the area under the ROC curve. An AUC of 1 means the model is perfect in classifying the images. The ROC is the measure of True Positive Rate (TPR) vs False Positive Rate (FPR) and the ratio is different at different decision thresholds. Since, it is a binary classification, ideal scenario for any model would be a straight line upwards and horizontal at the decision threshold giving the AUC equal to 1. The proposed model as an AUC of 0.89 showing high capability of correctly segregating classes.

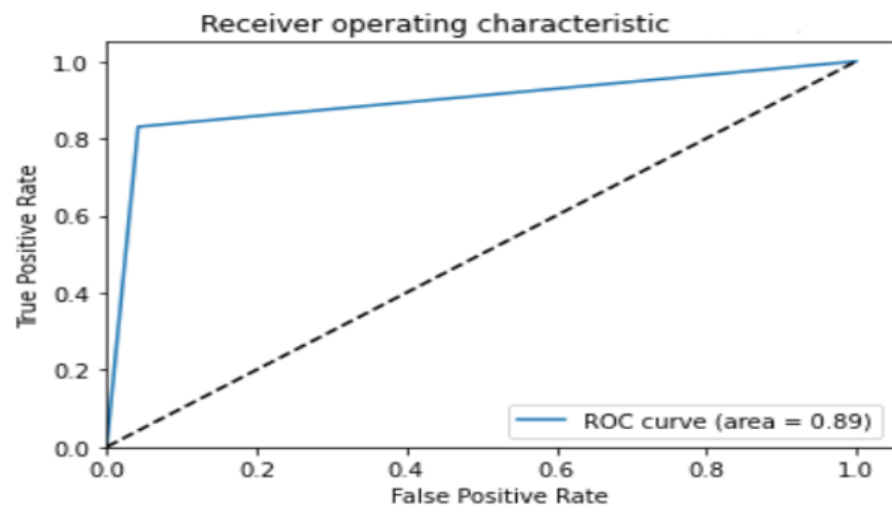


Figure 8. ROC Curve of the prediction from the test set.

CHAPTER – VI

DISCUSSION

VI [A] - ABOUT THE PROPOSED MODEL

In recent years, deep learning, especially convolutional neural networks, have been widely applied in computer vision and semantic segmentation. However, automatic building extraction from high-resolution remote sensing imagery is still a challenging task due to a large variety of appearing patterns and its spatial scale. As shown in Zhang et al. [20] accurate building extraction depends on the acquisition of the unique morphological characteristics of the building. They also pointed out that a well-performing network for building segmentation requires large receptive fields and needs to consider the multi-scale context.

To address these issues and to achieve an effective performance to extract buildings from remote sensing images, we proposed a light-weight transfer learning-based encoder-decoder model. This model follows the basic structure of U-Net, with the inclusion of the resnet-50 encoder trained on ImageNet dataset as encoder stage in the existing encoder-decoder structure. Proposed model achieves satisfactory results, proposed model achieved overall accuracy of 88.8% on TIFF images with augmentation scheme 1, an overall accuracy of 89.3% on TIFF images with augmentation scheme 2 and an overall accuracy of 90.1% on PNG images using augmentation scheme 2. The model achieves Average Overall Accuracy of 89.4% and an F1-score of 0.892 for the Massachusetts Buildings dataset. These scores are improvements over many established methods. It demonstrates that the encoder-decoder architecture for image segmentation tasks can be equally accurate as original U-Net with change in encoder backbone with benefit of reduction in computational cost.

The accuracy and loss during the training phase reported in Section V and the experimental results reported using the different augmentation schemes explained in section III-C show that the brightness and contrast boosting coupled with gamma correction on the training images increase the overall accuracy of the model over general image augmentation such as flipping, rotation and cropping. The results from section V suggests that the

proposed model also shows works very well for the images with lower resolution when the model is trained using augmentation scheme 2.

VI [B] - LIMITATIONS

Despite decent results as compared with other state-of-the-art architectures, the overall accuracy is lower as compared to USPP (0.894 vs 0.913). The proposed model is working on the images from IKONOS satellite and the model is not tested on the hyperspectral and SAR images that are now available with the development of the satellite imaging technology.

In further research, we propose to fine-tune training parameters and include a vaster data for training to improve upon the current results. The backbone encoder can be upgraded to resnet-101 and different pooling algorithms will be implemented to further improve the accuracy. The proposed model can be extended to identify roads among the buildings that will help in further urban development and planning.

CHAPTER – VII

CONCLUSION

Building Detection from remote sensing images is a very important research area in the field of computer vision. The extraction of building data is very useful for many applications such as the urban development, finding out the vegetation areas. However, the diverse characteristics of buildings including colour, shape, material, size, and the interference of building shadows and vegetation still make accurate and reliable building extraction of buildings is a challenging task.

We propose a light-weight transfer learning-based encoder-decoder architecture to extract the buildings from remote sensing images. The dataset used is Massachusetts Buildings dataset containing the raw images from the IKONOS satellite. The model is trained on 137 images and the training set is scaled 3-fold using two augmentation schemes as described in section III-C. The model is trained for 50 epochs and the results are evaluated on the two testing sets containing raw and compressed images in the TIFF and PNG formats respectively. The proposed model achieved overall accuracy of 88.8% on TIFF images with augmentation scheme 1, an overall accuracy of 89.3% on TIFF images with augmentation scheme 2 and an overall accuracy of 90.1% on PNG images using augmentation scheme 2. The proposed model achieved an average overall accuracy of 89.4 vs 91.3 of USPP model. The model performs equally well on the compressed PNG images showing robustness with decrease in resolution of the images.

The model performs better in the Augmentation Schemes 2 which incorporates image enhancing methods like brightness and contrast boosting and gamma correction. The ROC curve of the proposed model gives an AUC of 0.89 showing high classification efficiency of the model. The model is not trained on multi-spectral and SAR images which remains a prospect for further research and improving the current model. The proposed model will be further validated on other building datasets like INRIA Dataset that will help in further improving the model.

CHAPTER – VIII

REFERENCES

- [1] Jun Wang, Xiucheng Yang, Xuebin Qin, Xin Ye, & Qiming Qin. (2015). *An Efficient Approach for Automatic Rectangular Building Extraction From Very High Resolution Optical Satellite Imagery*. *IEEE Geoscience and Remote Sensing Letters*, 12(3), 487–491. doi:10.1109/lgrs.2014.2347332
- [2] Shandiz, H. T., Mirhassani, S. M., & Yousefi, B. (2008). *Hierarchical method for building extraction in urban area's images using unsharp masking [USM] and Bayesian classifier*. 2008 15th International Conference on Systems, Signals and Image Processing. doi:10.1109/iwssip.2008.4604400
- [3] Yanfeng Wei, Z. Z. (n.d.). *Urban building extraction from high-resolution satellite panchromatic image using clustering and edge detection*. *IEEE International IEEE International IEEE International Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings*. 2004. doi:10.1109/igarss.2004.1370742
- [4] Izadi, M., & Saeedi, P. (2010). *Automatic Building Detection in Aerial Images Using a Hierarchical Feature Based Image Segmentation*. 2010 20th International Conference on Pattern Recognition. doi:10.1109/icpr.2010.123
- [5] Wang, M., Yuan, S., & Pan, J. (2013). *Building detection in high resolution satellite urban image using segmentation, corner detection combined with adaptive windowed Hough Transform*. 2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS. doi:10.1109/igarss.2013.6721204
- [6] Ngo, T.-T., Collet, C., & Mazet, V. (2015). *Automatic rectangular building detection from VHR aerial imagery using shadow and image segmentation*. 2015 IEEE International Conference on Image Processing (ICIP). doi:10.1109/icip.2015.7351047
- [7] Li, E., Xu, S., Meng, W., & Zhang, X. (2017). *Building Extraction from Remotely Sensed Images by Integrating Saliency Cue*. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(3), 906–919. doi:10.1109/jstars.2016.2603184
- [8] Lui, Y., Gross, L., Li, Z., Li, X., Fan, X., Qi, W., (2019). *Automatic building extraction on high resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling*. doi: 10.1109/access.2019.2940527
- [9] Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L., & Pan, C. (2016). *Building extraction from multi-source remote sensing images via deep deconvolution neural networks*. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). doi:10.1109/igarss.2016.7729471
- [10] Li, Q., Wang, Y., Liu, Q., & Wang, W. (2018). *Hough Transform Guided Deep Feature Extraction for Dense Building Detection in Remote Sensing Images*. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/icassp.2018.8461407

- [11] Christophe, E., & Inglada, J. (2009). *Object counting in high resolution remote sensing images with OTB*. 2009 IEEE International Geoscience and Remote Sensing Symposium. doi:10.1109/igarss.2009.5417482
- [12] Chen, K., Fu, K., Gao, X., Yan, M., Sun, X., & Zhang, H. (2017). *Building extraction from remote sensing images with deep learning in a supervised manner*. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). doi:10.1109/igarss.2017.8127295
- [13] Sirmacek, B., & Unsalan, C. (2010). *Road Network Extraction Using Edge Detection and Spatial Voting*. 2010 20th International Conference on Pattern Recognition. doi:10.1109/icpr.2010.762
- [14] Yongguan Xiao, S. K. L. (n.d.). *Feature extraction using very high-resolution satellite imagery*. IEEE International IEEE International IEEE International Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. 2004. doi:10.1109/igarss.2004.1370741
- [15] Huang, X., Yuan, W., Li, J., & Zhang, L. (2017). *A New Building Extraction Postprocessing Framework for High-Spatial-Resolution Remote-Sensing Imagery*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(2), 654–668. doi:10.1109/jstars.2016.2587324
- [16] Lee, T., Kim, Y.-S., & Kim, T. (2013). *Automatic building information extraction by modified volumetric shadow analysis from high resolution multispectral data*. 2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS. doi:10.1109/igarss.2013.6723703
- [17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox (2015). *U-Net:Convolutional Networks for Biomedical Image Segmentation*. arXiv:1505.04597
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (2015). *Deep Residual Learning for Image Recognition*. arXiv:1512.03385
- [19] M. Polak, H. Zhang, and M. Pi (2009), "An evaluation metric for image segmentation of multiple objects," *Image and Vision Computing*, vol. 27, no. 8, pp. 1223-1227.
- [20] Z. Zhang and Y. Wang, "JointNet: A Common Neural Network for Road and Building Extraction," *Remote Sensing*, vol. 11, no. 6, p. 696, 2019