ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

DEEP LEARNING-BASED BUILDING SEGMENTATION USING HIGH-RESOLUTION AERIAL IMAGES

Ph.D. THESIS

Batuhan SARITÜRK

Department of Geomatics Engineering

Geomatics Engineering Programme

OCTOBER 2022

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

DEEP LEARNING-BASED BUILDING SEGMENTATION USING HIGH-RESOLUTION AERIAL IMAGES

Ph.D. THESIS

Batuhan SARITÜRK
(501142612)

Department of Geomatics Engineering

Geomatics Engineering Programme

Thesis Advisor: Prof. Dr. Dursun Zafer ŞEKER

OCTOBER 2022

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK ÇÖZÜNÜRLÜKLÜ HAVA GÖRÜNTÜLERİ KULLANARAK DERİN ÖĞRENME TEMELLİ BİNA BÖLÜTLEMESİ

DOKTORA TEZİ

Batuhan SARITÜRK
(501142612)

Geomatik Mühendisliği Anabilim Dalı

Geomatik Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Dursun Zafer ŞEKER

EKİM 2022

Batuhan SARITÜRK, a Ph.D. student of ITU Graduate School student ID 501142612 successfully defended the thesis entitled "DEEP LEARNING-BASED BUILDING SEGMENTATION USING HIGH-RESOLUTION AERIAL IMAGES", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Dursun Zafer ŞEKER** Istanbul Technical University	.....
Jury Members :	**Prof. Dr. Bülent BAYRAM** Yıldız Technical University	.....
	Prof. Dr. Zaide DURAN Istanbul Technical University	.....
	Prof. Dr. Şinasi KAYA Istanbul Technical University	.....
	Asst. Prof. Dr. Tolga BAKIRMAN Yıldız Technical University	.....

Date of Submission : **6 September 2022**
Date of Defense : **5 October 2022**

To my family and close ones,

FOREWORD

Firstly, I would like to thank my dear supervisor, Prof. Dr. Dursun Zafer ŞEKER, for his guidance, constant support, and patience throughout my Ph.D. studies and my academic life. I would also like to thank my committee members, Prof. Dr. Zaide DURAN and Prof. Dr. Bülent BAYRAM, for their contributions and positive criticism.

I would like to thank my dear friends Mustafa Serkan IŞIK, Ozan ÖZTÜRK, Mehmet Furkan ÇELİK, Öykü KOÇ, Dr. Raziye Hale TOPALOĞLU, and Elif Rüya ÖZTÜRK for their both academic and moral support throughout my studies and my social life. I would like to thank Özlem GÜMÜŞKANAT for her everlasting support, which is very important to me, especially during the writing process of this thesis. I couldn't be here, in this position, in this state of mind, without all these dear peoples' support. I would also like to thank my late dear friends, Ömer AKIN and Beril VAROL. I hope you have found happiness, peace, and comfort.

Last but not least, I would like to thank my dear family; my mother Aytülü SARITÜRK, my father Kazım SARITÜRK, my brother Doğuhan SARITÜRK, and my sister-in-law Helin GÜMÜŞ SARITÜRK for their constant belief in me and endless support.

October 2022

Batuhan SARITÜRK
(Geomatics Engineer, M.Sc.)

TABLE OF CONTENTS

	Page
FOREWORD	**ix**
TABLE OF CONTENTS	**xii**
ABBREVIATIONS	**xiii**
LIST OF TABLES	**xv**
LIST OF FIGURES	**xviii**
SUMMARY	**xix**
ÖZET	**xxiii**
1. INTRODUCTION	**1**
1.1 Aim and Objectives of the Thesis	5
1.2 Research Theme and Structure	6
2. PERFORMANCE EVALUATION OF SHALLOW AND DEEP CNN ARCHITECTURES ON BUILDING SEGMENTATION FROM HIGH-RESOLUTION IMAGES	**7**
2.1 Introduction	7
2.2 Datasets and Methodology	11
2.2.1 Datasets	11
2.2.1.1 Inria aerial image labeling dataset	12
2.2.1.2 Massachusetts buildings dataset	14
2.2.2 Methodology	15
2.2.2.1 U-Net architecture	15
2.2.2.2 Configuration of the networks	19
2.3 Training and Evaluating the Models	20
2.4 Results and Discussions	25
2.5 Conclusions	36
3. COMPARISONS OF RESIDUAL AND DENSE NEURAL NETWORK APPROACHES FOR BUILDING EXTRACTION FROM HIGH-RESOLUTION AERIAL IMAGES	**39**
3.1 Introduction	39
3.2 Study	42
3.2.1 Datasets	42
3.2.1.1 Inria aerial image labeling dataset	42
3.2.1.2 Massachusetts buildings dataset	44
3.2.2 Methodology	45
3.2.2.1 Architectures	45
3.2.2.2 Backbones	47
3.2.2.3 Attention Residual U-Net approach	50
3.2.3 Training and testing	53
3.3 Results and Discussions	56
3.4 Conclusions	60
4. COMPARATIVE ANALYSIS OF DIFFERENT CNN MODELS FOR BUILDING SEGMENTATION FROM SATELLITE AND UAV IMAGES	**63**
4.1 Introduction	63
4.2 Methodology	66
4.2.1 Architectures	66
4.2.1.1 U-Net	66
4.2.1.2 Feature Pyramid Networks (FPN)	67
4.2.1.3 LinkNet	68
4.2.1.4 SegNet	69
4.2.1.5 Fully Convolutional Networks (FCN)	70
4.2.2 Backbone (EfficientNet)	70
4.2.3 Residual blocks and Residual U-Net approaches	71
4.3 Datasets	72

4.3.1 Inria aerial image labeling dataset 72
4.3.2 Massachusetts buildings dataset 73
4.3.3 Syedra archaeological site dataset 74
4.4 Training and Testing 74
4.5 Results and Discussions 77
4.6 Conclusions 81
5. A RESIDUAL-INCEPTION U-NET (RIU-NET) APPROACH AND COMPARISONS WITH U-SHAPED CNN AND TRANSFORMER MODELS FOR BUILDING SEGMENTATION FROM HIGH-RESOLUTION SATELLITE IMAGES 83
5.1 Introduction 83
5.2 Materials and Methods 89
5.2.1 Datasets 89
5.2.1.1 Inria Aerial Image Labeling Dataset 89
5.2.1.2 Massachusetts Buildings Dataset 89
5.2.1.3 Preparation of the Data 90
5.2.2 Methodology 91
5.2.2.1 U-Net 91
5.2.2.2 Residual U-Net 91
5.2.2.3 Attention Residual U-Net 92
5.2.2.4 Inception Backboned U-Net 93
5.2.2.5 Inception-ResNet Backboned U-Net 94
5.2.2.6 Xception Backboned U-Net 94
5.2.2.7 MobileNet Backboned U-Net 95
5.2.2.8 Trans U-Net 95
5.2.2.9 Swin U-Net 96
5.2.2.10 Residual-Inception U-Net (RIU-Net) 96
5.3 Experiments 100
5.4 Results and Discussion 103
5.5 Conclusions 109
6. CONCLUSIONS 113
REFERENCES 117
CURRICULUM VITAE 135

ABBREVIATIONS

CNN	**:** Convolutional Neural Networks
FCN	**:** Fully Convolutional Networks
FPN	**:** Feature Pyramid Networks
GIS	**:** Geographical Information Systems
DL	**:** Deep Learning
GPU	**:** Graphic Processing Unit
ReLU	**:** Rectified Linear Unit
TP	**:** True Positive
TN	**:** True Negative
FP	**:** False Positive
FN	**:** False Negative
CRF	**:** Conditional Random Fields
AIRS	**:** Aerial Imagery for Roof Segmentation
ISPRS	**:** International Society for Photogrammetry and Remote Sensing
IoU	**:** Intersection over Union
UAV	**:** Unmanned Aerial Vehicle
DNN	**:** Deep Neural Networks
NLP	**:** Natural Language Processing
ViT	**:** Vision Transformer
RIU-Net	**:** Residual Inception U-Net
Swin	**:** Shifted Windows

LIST OF TABLES

Page

Table 2.1 : The total number of parameters of the architectures. **22**
Table 3.1 : ResNet-50 architecture..... **48**
Table 3.2 : ResNeXt-50 architecture..... **48**
Table 3.3 : SE-ResNeXt-50 architecture. **49**
Table 3.4 : DenseNet-121 architecture. **50**
Table 3.5 : No. of trainable parameters and no. of training epochs of the models. **55**
Table 4.1 : The total number of parameters of the architectures. **76**
Table 4.2 : Evaluation results of the models..... **78**
Table 5.1 : Hyperparameters of Large, Base, and Mini Trans U-Net variants. **96**
Table 5.2 : No. of trainable parameters and epochs of the best models..... **102**

LIST OF FIGURES

Page

Figure 2.1 : Austin - 7 sample image and labeled image. **12**
Figure 2.2 : Sample image and labeled image from the INRIA training dataset. . **14**
Figure 2.3 : Sample image and labeled image from the Massachusetts training dataset. **15**
Figure 2.4 : U-Net architecture. **16**
Figure 2.5 : (**a**) Layers of the downsampling path, (**b**) Layers of the upsampling path. **19**
Figure 2.6 : (**a**) U-Net shallower 2 architecture, (**b**) U-Net shallower 1 architecture, (**c**) U-Net deeper 1 architecture, (**d**) U-Net deeper 2 architecture. **20**
Figure 2.7 : K-fold cross-validation. **21**
Figure 2.8 : Accuracy and loss values of the INRIA models over 50 epochs using k-fold cross validation (**a**) shallower 2, (**b**) shallower 1, (**c**) u-net, (**d**) deeper 1, (**e**) deeper 2. ..... **23**
Figure 2.9 : Accuracy and loss values of the Massachusetts models over 50 epochs using k-fold cross validation (**a**) shallower 2, (**b**) shallower 1, (**c**) u-net, (**d**) deeper 1, (**e**) deeper 2. ..... **24**
Figure 2.10 : Training, validation, and test accuracy values of the architectures on the INRIA dataset. **26**
Figure 2.11 : Training, validation, and test accuracy values of the architectures on the Massachusetts dataset. **27**
Figure 2.12 : Training, validation, and test loss values of the architectures on the INRIA dataset. **28**
Figure 2.13 : Training, validation, and test loss values of the architectures on the Massachusetts dataset. **29**
Figure 2.14 : Precision, recall, F1 score, and IoU scores of the architectures on the INRIA dataset. **30**
Figure 2.15 : Precision, recall, F1 score, and IoU scores of the architectures from the Massachusetts dataset. **31**
Figure 2.16 : Segmentation results for test image no. 61 from the INRIA test set. **32**
Figure 2.17 : Segmentation results for test image no. 156 from the INRIA test set. **33**
Figure 2.18 : Segmentation results for test image no. 60 from Massachusetts test set. **34**
Figure 2.19 : Segmentation results for test image no. 156 from Massachusetts test set. **35**
Figure 3.1 : Austin-7 RGB image and corresponding mask from the Inria Aerial Image Labeling Dataset. **43**

Figure 3.2 : Sample cropped image and corresponding mask from the prepared Inria dataset. **43**
Figure 3.3 : Sample cropped image and corresponding mask from the prepared Massachusetts dataset. **44**
Figure 3.4 : U-Net architecture. **46**
Figure 3.5 : FPN architecture and building block architecture. **47**
Figure 3.6 : (**a**) Residual block design used in ResNet, (**b**) Residual block design implemented in this study. **51**
Figure 3.7 : Schematic of the attention gate used in the study. **53**
Figure 3.8 : Evaluation metric results of the models for the Inria test set. **57**
Figure 3.9 : Evaluation metric results of the models for the Massachusetts test set. **58**
Figure 3.10 : Segmentation results for image no. 810 from the Inria test set. **59**
Figure 3.11 : Segmentation results for image no. 51 from the Massachusetts test set. **59**
Figure 4.1 : U-Net architecture. **67**
Figure 4.2 : FPN architecture. **68**
Figure 4.3 : **a**) LinkNet architecture, **b**) encoder block, **c**) decoder block. **69**
Figure 4.4 : SegNet architecture. **69**
Figure 4.5 : FCN architecture. **70**
Figure 4.6 : EfficientNet-B0 architecture. **71**
Figure 4.7 : Residual block designs were used in the study. **71**
Figure 4.8 : Sample 512×512 image and mask from Inria dataset. **73**
Figure 4.9 : Sample 512×512 image and mask from Massachusetts dataset. **73**
Figure 4.10 : Sample 256×256 image and mask from Syedra dataset. **74**
Figure 4.11 : Segmentation results of image no. 552 from the Inria test set. **79**
Figure 4.12 : Segmentation results of image no. 128 from the Massachusetts test set. **80**
Figure 4.13 : Segmentation results of image no. 329 from the Syedra test set. **80**
Figure 5.1 : Sample 256×256 pixel image and mask: (**a**) Inria dataset, (**b**) Massachusetts dataset. **90**
Figure 5.2 : (**a**) Residual connection design from the ResNet, (**b**) Residual connection design implemented in the study. **92**
Figure 5.3 : Attention mechanism implemented in the study. **93**
Figure 5.4 : Overall architecture of the RIU-Net. **98**
Figure 5.5 : The flow diagram of the modules used in the encoder path of the RIU-Net: (**a**) Module A, (**b**) Module B, (**c**) Module C, (**d**) Reduction A, and (**e**) Reduction B. **99**
Figure 5.6 : The flow diagram of the modules used in the bottleneck and decoder paths of the RIU-Net: (**a**) Module D, and (**b**) Upsampling module. **100**
Figure 5.7 : Evaluation metric results on Inria test set. **104**
Figure 5.8 : Evaluation metric results on Massachusetts Small test set. **105**
Figure 5.9 : Evaluation metric results on Massachusetts Large test set. **106**
Figure 5.10 : Inria test set image no. 1036 segmentation results. **107**
Figure 5.11 : Massachusetts Small test set image no. 192 segmentation results. **107**
Figure 5.12 : Massachusetts Large test set image no. 294 segmentation results. **108**

DEEP LEARNING-BASED BUILDING SEGMENTATION USING HIGH-RESOLUTION AERIAL IMAGES

SUMMARY

With the advancements in satellite and remote sensing technologies, and the developments in urban areas, building segmentation and extraction from high-resolution aerial images and generating building maps have become important and popular research topics. With technological developments, a large number of high-resolution images have become increasingly more accessible and convenient data sources. At the same time, due to their ability of imaging over large areas, these aerial images can be very useful for accurate building segmentation and generating building maps.

As one of the most important and key features of the urban database, buildings are the building blocks for human livelihood. Due to this importance, building maps have significant role for various geoscience-related applications such as illegal building detection, change detection, population estimation, land use/land cover analysis, disaster management, and topographic and cadastral map production.

Nonetheless, obtaining accurate and reliable building maps from high-resolution aerial images is still a challenging task due to various reasons such as complex backgrounds, differences in building size, shape, and colors, noisy data, roof type diversity, and many other topological difficulties. Therefore, improving the efficiency and accuracy of building segmentation and extraction from high-resolution aerial images is still a focus and a hot topic among researchers in the field.

Over the past years, various methods have been used to achieve automatic building segmentation from aerial images. In earlier studies, traditional image processing methods such as object-based, shadow-based, or edge-based methods were used. The low-level features and metrics that are used with these methods such as color, spectrum, length, texture, edge, shadow, and height could vary under different conditions like atmospheric state, light, scale, and sensor quality. These methods generally take manually extracted features and apply classifiers or conventional machine learning techniques to achieve building segmentation. However, manually extracting these features is costly, time-consuming, labor intensive, and requires high experience and prior knowledge. Although over time these methods have made some progress, they have some serious shortcomings such as low accuracy, low generalization ability, and complex processing.

With the technological developments and availability of large datasets, deep learning-based approaches, especially Convolutional Neural Networks (CNN), have gained a lot of attention from researchers and surpass conventional methods in terms of accuracy and efficiency. CNNs have the ability to extract relevant features directly from the input data and make predictions using fully connected layers. Many

CNN architectures such as LeNet, AlexNet, VGGNet, GoogleNet, and ResNet have been used over the years. However, CNNs perform regional divisions and use computationally expensive fully connected layers. These patch-based CNNs have achieved exceptional success, but due to their reliance on small patches around the targets to perform predictions and ignoring the relations between them, they are unable to provide accurate integrity and spatial continuity of building features. To improve the performance and overcome these problems, Long et al. proposed Fully Convolutional Networks (FCN). Instead of fully connected layers in CNNs, FCNs have convolution layers that improve the prediction accuracy greatly. FCNs output the feature maps at the size of the input images and perform pixel-based segmentation through their encoder-decoder structure. However, much of the information is lost in the decoder path, due to the FCNs having just one upsampling layer.

Despite their success, FCNs also have some limitations, such as computational complexity and a large number of parameters. To overcome these shortcomings, various variants have been proposed over the years such as SegNet, U-Net, and Feature Pyramid Networks (FPN). These CNN-based approaches have achieved successful results on image segmentation tasks, but they also have some bottlenecks. For example, the usage of fixed-size convolutions results a local receptive field. Due to their designs, they are successful at extracting local context but have a low ability to extract global context. To overcome these shortcomings, some approaches have been proposed and implemented. Such as attention mechanism, residual connections, and designing architecture in different depths.

The Transformer was first used in natural language processing (NLP), and later on, implemented to computer vision tasks. In 2020, the Vision Transformer (ViT) approach was proposed to be used in computer vision studies and obtained successful results on the ImageNet dataset. CNNs are successful in identifying local features, but they are insufficient in identifying global features due to their structure. Transformers can compensate for these shortcomings with the use of attention mechanisms. In ViT-based methods, global information can be extracted but spatially detailed context is ignored. In addition, Transformers use all the pixels in vector operations when working with large images, and therefore require large amounts of memory and are computationally inefficient.

The main objective of this thesis is to investigate, evaluate, and realize comparisons of different CNN-based and Transformer-based approaches for building segmentation from high-resolution aerial images, and propose a modernized CNN approach to deal with the mentioned shortcomings. This thesis is composed of four papers dealing with these objectives. In the first paper, four U-Net-based architectures, which are shallower and deeper versions of the U-Net, have been generated to perform building segmentation from high-resolution aerial images and they were compared with each other and the original U-Net. The models were trained and tested on datasets prepared using the Inria Aerial Image Labeling Dataset and the Massachusetts Buildings Dataset. On the INRIA test set, Deeper 1 U-Net architecture provided the highest F1 score with 0.79 and IoU score with 0.65, followed by Deeper 2 and U-Net architectures. On the Massachusetts test set, U-Net architecture provided 0.79 F1 score and 0.66 IoU score, followed by Deeper 2 and Shallower 1. Successful results were

obtained with Deeper 1 and Deeper 2 architectures show that deeper architectures can provide better results even if there is not too much data. Additionally, Shallower 1 architecture appears to have a performance not far behind deep architectures, with less computational cost, and this shows usefulness for geographic applications. In the second paper, U-Net and FPN architectures utilizing four different backbones (ResNet, ResNeXt, SE-ResNeXt, and DenseNet) and an Attention Residual U-Net approach were generated and comparisons between them were realized. Publicly available Inria Aerial Image Labeling Dataset and Massachusetts Buildings Dataset were used to train and test the models. Attention Residual U-Net model has the highest F1 score with 0.8154, IoU score with 0.7102, and test accuracy with 94.51% on the Inria test set. On the Massachusetts test set, FPN Dense-Net-121 model has the highest F1 score with 0.7565 and IoU score with 0.6188, and the Attention Residual U-Net model has the highest test accuracy with 92.43%. It has been observed that FPN with DenseNet backbone can be a better choice when working with small-size datasets. On the other hand, the Attention Residual U-Net approach achieved higher success when a sufficiently large dataset is provided. In the third paper, a total of twelve CNN-based models (U-Net, FPN, and LinkNet architectures utilizing EfficientNet-B5 backbone, original U-Net, SegNet, FCN, and six different Residual U-Net approaches) were generated, evaluated and comparisons between them were realized. Inria Aerial Image Labeling Dataset was used to train models, and three datasets (Inria Aerial Image Labeling Dataset, Massachusetts Buildings Dataset, and Syedra Archaeological Site Dataset) were used to evaluate trained models. On the Inria test set, Residual-2 U-Net has the highest F1 and IoU scores with 0.824 and 0.722, respectively. On the Syedra test set, LinkNet-EfficientNet-B5 has F1 and IoU scores of 0.336 and 0.246. On the Massachusetts test set, Residual-4 U-Net has F1 and IoU scores of 0.394 and 0.259. When the results were evaluated, it has been observed that the models using residual connections are more successful than the models using conventional convolution structures. It has also been observed that the LinkNet architecture gave good results on the Syedra test set with different characteristics than the other two datasets, and could be a good option for future studies involving archaeological sites. In the fourth paper, a total of ten CNN and Transformer models (the proposed Residual-Inception U-Net (RIU-Net), U-Net, Residual U-Net, Attention Residual U-Net, U-Net-based models implementing Inception, Inception-ResNet, Xception, and MobileNet as backbones, Trans U-Net, and Swin U-Net) were generated, and building segmentation from high-resolution satellite images was carried out. Massachusetts Buildings Dataset and Inria Aerial Image Labeling Dataset were used for training and evaluation of the models. On the Inria dataset, RIU-Net achieved the highest IoU score, F1 score, and test accuracy, with 0.6736, 0.7868, and 92.23%, respectively. On the Massachusetts Small dataset, Attention Residual U-Net achieved the highest IoU and F1 scores, with 0.6218 and 0.7606, and Trans U-Net reached the highest test accuracy, with 94.26%. On the Massachusetts Large dataset, Residual U-Net accomplished the highest IoU and F1 scores, with 0.6165 and 0.7565, and Attention Residual U-Net attained the highest test accuracy, with 93.81%. The results showed that the RIU-Net approach was significantly successful in the Inria dataset compared to other models. On the Massachusetts datasets, Residual U-Net, Attention Residual U-Net, and Trans U-Net gave more successful results.

YÜKSEK ÇÖZÜNÜRLÜKLÜ HAVA GÖRÜNTÜLERİ KULLANARAK DERİN ÖĞRENME TEMELLİ BİNA BÖLÜTLEMESİ

ÖZET

Uydu ve uzaktan algılama teknolojilerindeki ilerlemeler ve kentsel alanlardaki gelişmelerle birlikte, yüksek çözünürlüklü hava görüntülerinden bina çıkarma ve bina haritaları oluşturma, bilgisayarlı görü ve uzaktan algılama toplulukları arasında önemli ve popüler bir araştırma konusu haline gelmiştir. Teknolojik gelişmelerle birlikte çok sayıda yüksek çözünürlüklü görüntü giderek daha erişilebilir ve kullanışlı veri kaynakları haline gelmiştir. Aynı zamanda, geniş alanları görüntüleme yetenekleri nedeniyle, bu hava görüntüleri, doğru ve hassas bina bölütlemesi ve bina haritaları oluşturmak için çok faydalı olmaktadırlar.

Kentsel veritabanı için en önemli ve anahtar özelliklerden biri olan binalar, insan yaşamının temel yapı taşlarındandır. Bu önemlerinden dolayı yapı haritaları, kaçak yapı tespiti, değişim tespiti, nüfus tahmini, arazi kullanımı/arazi örtüsü analizi, afet yönetimi, topografik ve kadastro harita üretimi gibi yerbilimleri ile ilgili çeşitli uygulamalar ve çalışma alanları için büyük önem arz etmektedir.

Bununla birlikte, yüksek çözünürlüklü hava görüntülerinden doğru ve güvenilir bina haritaları elde etmek, karmaşık arka planlar, binaların boyut, şekil ve renklerindeki farklılıklar, gürültülü veriler, çatı tipi farklılıkları ve diğer birçok topolojik zorluklar gibi çeşitli nedenlerden dolayı hala zorlu bir iştir. Bu nedenle, yüksek çözünürlüklü hava görüntülerinden bina bölütlemenin ve çıkarmanın verimliliğini ve doğruluğunu artırmak, bu alandaki araştırmacılar arasında hala bir odak noktası ve ilgi çekici bir konudur.

Geçtiğimiz yıllar içinde, hava görüntülerinden otomatik bina bölütlemesi gerçekleştirmek etmek için araştırmacılar tarafından çeşitli yöntemler kullanılmıştır. Bu çalışmalarda nesne tabanlı, gölge tabanlı veya kenar tabanlı yöntemler gibi geleneksel görüntü işleme yöntemleri kullanılmaktaydı. Renk, spektrum, uzunluk, doku, kenar, gölge ve yükseklik gibi bu yöntemlerle kullanılan düşük seviyeli özellikler ve metrikler, atmosferik durum, ışık, ölçek ve sensör kalitesi gibi farklı koşullar altında değişebilmektedir. Bu geleneksel yöntemler genellikle manuel olarak çıkarılan öznitelikleri alır ve daha sonra bina bölütlemesi elde etmek için bir sınıflandırıcı veya geleneksel makine öğrenmesi tekniklerini kullanırlar. Ancak özelliklerin manuel olarak çıkarılması maliyetli, zaman alıcı, yoğun emek isteyen ve yüksek deneyim ve ön bilgi gerektiren bir işlemdir. Zaman içinde bu yöntemler bir miktar ilerleme kaydetmiş olsa da, düşük doğruluk, düşük genelleme yeteneği ve karmaşık işlemler gibi bazı ciddi eksiklikleri vardır.

Teknolojik gelişmeler ve büyük veri setlerinin mevcudiyeti ile birlikte, derin öğrenme tabanlı yaklaşımlar, özellikle Evrişimli Sinir Ağları (ESA), çok popüler hale

gelmiş, topluluklar içindeki araştırmacılardan çok fazla ilgi görmüş ve geleneksel yöntemleri doğruluk açısından geride bırakmıştır. ESA'lar, doğrudan giriş verilerinden ilgili özellikleri çıkarma ve tam bağlantılı katmanları kullanarak tahminler yapma yeteneğine sahiptir. Yıllar içinde LeNet, AlexNet, VGGNet, GoogleNet ve ResNet gibi birçok ESA mimarisi kullanılmıştır. Bununla birlikte, ESA'lar bölgesel bölünmeler gerçekleştirir ve hesaplama açısından pahalı tam bağlantılı katmanlar kullanır. Bu yama tabanlı ESA'lar, bina bölütlemesi çalışmalarında olağanüstü bir başarı elde etmiştirler, ancak tahminler yapmak için hedeflerin etrafındaki küçük yamalara güvenmeleri ve aralarındaki ilişkileri görmezden gelmeleri nedeniyle, bina özelliklerinin doğru bütünlüğünü ve mekansal sürekliliğini sağlayamamıştırlar.

ESA'ların performansını iyileştirmek ve bu problemlerin üstesinden gelmek için Long ve ark. Tam Evrişimli Ağlar'ı (TEA) önermişlerdir. ESA'lardaki tamamen bağlı katmanlar yerine, TEA, tahmin doğruluğunu büyük ölçüde artıran evrişim katmanlarına sahiptir. TEA, özellik haritalarının çıktısını girdi görüntülerinin boyutunda verir ve kodlayıcı-kod çözücü yapıları aracılığıyla piksel tabanlı bölütleme gerçekleştirir. Kodlayıcı yolunda, alt örnekleme yoluyla küçük boyutlu özellik haritaları oluşturulur ve kod çözücü yolunda, üst örnekleme yoluyla bu özellik haritaları kullanılarak nihai bölütleme haritaları oluşturulur. Ancak, TEA!nın yalnızca bir üst örnekleme katmanına sahip olması nedeniyle, bilgilerin çoğu kod çözücü yolunda kaybolur.

Elde ettikleri başarılara rağmen, TEA'nın hesaplama karmaşıklığı ve çok sayıda parametreye sahip olması gibi bazı sınırlamaları da vardır. Bu eksikliklerin üstesinden gelmek için yıllar içinde SegNet, U-Net ve Feature Pyramid Networks (FPN) gibi çeşitli varyantlar önerilmiştir. U-Net mimarisi, önceki katmanlardan sonraki katmanlara semantik bilgileri birleştirerek görüntü bölütleme doğruluğunu geliştirmek için atlama bağlantıları önermiştir. SegNet ise, parametre sayısını azaltmak için havuzlama katmanına sahip bir evrişimli kodlayıcı ve transpoze edilmiş kıvrımlara sahip simetrik bir kod çözücü önermiştir.

ESA tabanlı bu yaklaşımlar, görüntü bölütleme görevlerinde başarılı sonuçlar elde etmişlerdir, ancak bazı darboğazları da mevcuttur. Örneğin, sabit boyutlu evrişimlerin kullanımı, yerel bir alıcı alanla sonuçlanır. Tasarımları nedeniyle yerel bağlamı çıkarmada başarılıdırlar, ancak küresel bağlamı çıkarma yetenekleri düşüktür. Bu eksikliklerin üstesinden gelmek için dikkat mekanizması, artık bağlantılar ve farklı derinliklerde mimariler tasarlamak gibi bazı yaklaşımlar önerilmiş ve uygulanmıştır.

Transformer yaklaşımı ilk olarak doğal dil işleme (NLP) çalışmalarında kullanılmış ve daha sonra bilgisayarlı görü görevlerine uygulanmıştır. 2020 yılında bilgisayarlı görü çalışmalarında kullanılmak üzere Vision Transformer (ViT) yaklaşımı önerilmiş ve ImageNet veri seti üzerinde başarılı sonuçlar elde edilmiştir. CNN'ler yerel öznitelikleri belirlemede başarılıdır, ancak yapıları gereği küresel öznitelikleri belirlemede yetersizdirler. Transformer yaklaşımları, CNN'lerin bu eksikliklerini dikkat mekanizmalarının kullanımıyla telafi edebilmektedirler. ViT tabanlı yöntemler ile, CNN'lerin aksine, küresel bilgi çıkarılabilir, ancak uzamsal olarak ayrıntılı bağlam göz ardı edilir. Ayrıca, Transformer yaklaşımları büyük görüntülerle çalışırken vektör işlemlerinde tüm pikselleri kullandıklarından dolayı büyük miktarda bellek gerektirirler ve hesaplama açısından verimsizdirler.

Bu tezin temel amacı, yüksek çözünürlüklü hava görüntülerinden bina bölütlemesi gerçekleştirmek için ESA ve Transformer tabanlı farklı yaklaşımların incelenmesi, değerlendirilmesi, karşılaştırmaların gerçekleştirilmesi ve bahsedilen eksiklikleri gidermek amacıyla modernize edilmiş bir CNN yaklaşımı önermektir. İlk makalede, yüksek çözünürlüklü hava görüntülerinden bina bölütlemesi gerçekleştirmek için U-Net'in daha sığ ve daha derin versiyonları olan dört U-Net tabanlı mimari oluşturulmuş ve bunlar birbirleriyle ve orijinal U-Net ile karşılaştırılmıştır. Modeller, Inria Aerial Image Labelling Dataset ve Massachusetts Buildings Dataset kullanılarak hazırlanan veri setleri üzerinde eğitilmiş ve test edilmiştir. INRIA test setinde, Deeper 1 U-Net mimarisi 0,79 ile en yüksek F1 puanını ve 0,65 ile IoU puanını sağlamış, bunu Deeper 2 ve U-Net mimarileri izlemiştir. Massachusetts test setinde ise U-Net mimarisi 0,79 F1 puanı ve 0,66 IoU puanı elde etmiş, bunu Deeper 2 ve Shallower 1 mimarileri izlemiştir. Deeper 1 ve Deeper 2 mimarileri ile elde edilen başarılı sonuçlar, daha derin mimarilerin elde çok fazla sayıda veri olmasa bile daha iyi sonuçlar verebileceğini göstermektedir. Ek olarak, Shallower 1 mimarisi, daha az hesaplama maliyeti ile derin mimarilerin çok gerisinde olmayan bir performansa sahip görünmektedir ve bu, coğrafi uygulamalar için kullanışlılığını göstermektedir. İkinci makalede, bina bölütlemesi için, dört farklı omurga (ResNet, ResNeXt, SE-ResNeXt ve DenseNet) kullanan U-Net ve FPN mimarileri ile Attention Residual U-Net yaklaşımı oluşturulmuş ve bunlar arasında karşılaştırmalar yapılmıştır. Modelleri eğitmek ve test etmek için halka açık Inria Aerial Image Labelling Dataset ve Massachusetts Buildings Dataset kullanılmıştır. Attention Residual U-Net modeli, Inria test setinde 0,8154 ile en yüksek F1 puanına, 0,7102 ile en yüksek IoU puanına ve %94,51 ile en yüksek test doğruluğuna sahiptir. Massachusetts test setinde ise, FPN Dense-Net-121 modeli 0,7565 ile en yüksek F1 puanına ve 0,6188 ile en yüksek IoU puanına sahipken, Attention Residual U-Net modeli %92,43 ile en yüksek test doğruluğuna sahiptir. Sonuçlar incelendiğinde, küçük boyutlu veri setleri ile çalışırken DenseNet omurgasına sahip FPN'nin daha iyi bir seçim olabileceği gözlemlenmiştir. Öte yandan, Attention Residual U-Net yaklaşımının, yeterince büyük bir veri seti sağlandığında daha yüksek başarı elde ettiği gözlemlenmiştir. Üçüncü makalede, ESA tabanlı toplam on iki model oluşturulmuş, değerlendirilmiş ve aralarında karşılaştırmalar yapılmıştır. EfficientNet-B5 omurgasını kullanan U-Net, FPN ve LinkNet mimarileri, orijinal U-Net, SegNet ve FCN ile altı farklı Residual U-Net yaklaşımını kullanan U-Net mimarileri, yüksek çözünürlüklü hava görüntülerinden bina bölütlemesi oluşturmak için kullanılmıştır. Modelleri eğitmek için Inria Aerial Image Labelling Dataset ve eğitilmiş modelleri değerlendirmek için ise üç farklı veri seti (Inria Aerial Image Labelling Dataset, Massachusetts Buildings Dataset ve Syedra Archaeological Site Dataset) kullanılmıştır. Inria test setinde Residual-2 U-Net, sırasıyla 0,824 ve 0,722 ile en yüksek F1 ve IoU puanlarına sahiptir. Syedra test setinde, LinkNet-EfficientNet-B5, 0,336 ve 0,246 F1 ve IoU puanlarına sahiptir. Residual-4 U-Net, Massachusetts test setinde 0,394 ve 0,259 F1 ve IoU puanlarına sahiptir. Sonuçlar değerlendirildiğinde, artık bağlantı kullanan modellerin geleneksel evrişim yapılarını kullanan modellere göre daha başarılı olduğu görülmüştür. Ayrıca, LinkNet mimarisinin diğer iki veri setinden daha farklı detay özelliklerine sahip olan Syedra test setinde iyi sonuçlar verdiği ve arkeolojik alanları içerebilecek gelecek çalışmalar için iyi bir seçenek olabileceği gözlemlenmiştir. Dördüncü makalede, toplam on adet CNN ve Transformer modeli oluşturulmuş ve yüksek çözünürlüklü uydu görüntülerinden

bina bölütlemesi gerçekleştirilmiştir. Önerilen Residual-Inception U-Net (RIU-Net), U-Net, Residual U-Net ve Attention Residual U-Net modellerine ek olarak, dört CNN mimarisinin (Inception, Inception-ResNet, Xception ve MobileNet) kodlayıcı olarak uygulandığı U-Net tabanlı modeller de çalışma kapsamında kullanılmıştır. Son olarak, Transformer tabanlı iki yaklaşım da (Trans U-Net ve Swin U-Net) çalışmada kullanılmış ve tüm yaklaşımlar arasında karşılaştırmalar yapılmıştır. Modellerin eğitimi ve değerlendirilmesi için Massachusetts Buildings Dataset ve Inria Aerial Image Labelling Dataset kullanılmıştır. Inria veri setinde, RIU-Net sırasıyla 0,6736, 0,7868 ve %92,23 ile en yüksek IoU puanına, F1 puanına ve test doğruluğuna ulaşmıştır. Massachusetts Small veri setinde, Attention Residual U-Net 0,6218 ve 0,7606 ile en yüksek IoU ve F1 puanlarına ve Trans U-Net %94,26 ile en yüksek test doğruluğuna ulaşmıştır. Massachusetts Büyük veri setinde ise, Residual U-Net 0,6165 ve 0,7565 ile en yüksek IoU ve F1 puanlarını elde etmiş, Attention Residual U-Net ise %93,81 ile en yüksek test doğruluğuna ulaşmıştır. Sonuçlar, RIU-Net yaklaşımının diğer modellere kıyasla Inria veri setinde önemli ölçüde daha başarılı olduğunu göstermiştir. Massachusetts veri setlerinde ise, Residual U-Net, Attention Residual U-Net ve Trans U-Net yaklaşımları daha başarılı sonuçlar vermiştir.

1. INTRODUCTION

In recent years, with the developments in urbanized areas and advancements in satellite and remote sensing technologies, building segmentation and generating building maps have become important research topics [1]. With the help of remote sensing technologies, images of specific areas can be acquired repeatedly and serve as effective data sources [2]. Besides, these images can be very useful for accurate building mapping due to their ability of imaging over broad areas [3]. Due to these developments, a large amount of high-resolution satellite images have become more accessible and convenient data sources for geoscience-related applications such as segmentation and extraction of features such as buildings, roads, planes, ships, etc. [4]–[6]. As one of the most important components of the urban database, buildings are important places for human livelihood and key features for urban area mapping tasks [3]. As they are important features of the urban environment, mapping of buildings has significant importance for many different applications such as urban mapping, population estimation, land cover/land use analysis, cadastral and topographic map production, change detection, disaster management, and illegal building detection [7]–[11]. Nonetheless, obtaining reliable and accurate building maps from high-resolution satellite images is still ambitious due to various reasons such as complex backgrounds [12,13], occlusions, noise in data [8], heterogeneity in data structures [12,13], differences in building sizes, colors, and shapes, diversity in roof types [14], and other topological difficulties. Thus, improving the efficiency and accuracy of automatic building segmentation and extraction is still a focus and a challenging task for many researchers in the remote sensing and computer vision field.

Over the past years, many researchers have tried various methods to achieve automatic building segmentation [1]. In conventional methods, manually extracted features including geometrical, spatial, and spectral information, and low-level features such as shape, color, edge, texture, height, and shadow are used [15]. However, these

features vary under conditions such as light, building architectures, sensor quality, and scale. Conventional methods generally take these manually extracted features and apply classifiers or traditional machine learning techniques (e.g., Random Forests, Support Vector Machines, Boosting, Conditional Random Fields, and AdaBoost) to achieve building segmentation [16]. However, extracting these features is costly, labor-intensive, requires a lot of time, prior knowledge, and high experience [17,18]. Additionally, due to the manually extracted features, these methods have poor generalization abilities and this may cause inaccurate results [11]. Although these methods have achieved some progress, they have some shortcomings, such as low generalization ability, low accuracy, and complex processing requirement [7]. Due to the complex building structures and similarities with other ground object categories (e.g. roads, parking lots, cement floors, etc.), the predictions highly depend on the feature design and adjustment. With big data, and the developments of new algorithms and powerful hardware, deep learning methods have become very popular and gained a lot of interest in the remote sensing and computer vision communities.

Over the years, deep learning methods have gradually replaced conventional methods for image segmentation studies and achieved breakthroughs [16]. With the rapid developments in computing powers and the increasing availability of data sources, deep learning methods, particularly convolutional neural networks (CNN), have started to be used widely as they surpass conventional methods in terms of accuracy and efficiency [12]. Differently from these conventional methods, CNNs have the capacity to extract features directly from the inputs and make predictions using sequential convolutions with fully connected layers [8]. CNNs can therefore be thought of as one-step techniques that combine feature extraction and image classification into a single model, and they also possess good generalization ability [8]. Many researchers have used CNN architectures such as AlexNet [19], LeNet [20], VGGNet [21], ResNet [22], and GoogleNet [23]. However, these CNNs perform region divisions and use fully connected layers which are computationally expensive [7]. In earlier studies, CNN models that are patch-based, have achieved exceptional success in building segmentation, but due to the reliance on small patches around target features to make predictions and ignoring the relationships between them, they are unable to guarantee

the integrity and spatial continuity of building features [8]. Although the advancements in CNNs have promoted research in the area, there are still some challenges. For example; differences in building shape, size, and colors, geometric complexity, and high in-class and low inter-class variance make it hard to extract, detect, and segment the buildings from high-resolution satellite images. Therefore, traditional CNNs are not suitable for segmentation studies from satellite images [8].

To improve the performance of CNNs, Long et al. [24] presented the fully convolutional networks (FCN), which replaced fully connected layers with convolution layers, and improve the training and prediction accuracy greatly. FCNs can output feature maps at the input image size, using upsampling and achieve pixel-based segmentation through an encoder-decoder structure. Small-size feature maps are generated in the encoder path via downsampling to gather semantic information. In the decoder path, by decoding these generated feature maps, final segmentations are obtained. However, due to the FCN having only one upsampling layer, the segmentation is not very accurate, and much information is lost in the decoder [17].

Despite all the success that FCNs have achieved, they have some limitations. Firstly, the use of fully convolutional layers leads to a large number of parameters and computational complexity. Additionally, FCNs come short to provide detailed information about complex features and lose information about small objects [25]. Consequently, FCNs have become effective methods for image segmentation tasks, and various variants have emerged such as U-Net [26], Feature Pyramid Networks (FPN) [27], and SegNet [28] to improve performance and efficiency further and to overcome mentioned problems. U-Net improves the accuracy of image segmentation by concatenating the semantic information through skip connections, implementing upsampling, and integrating the same scale downsampling in each step [26]. SegNet has a convolutional encoder with pooling layer and a symmetrical decoder with transposed convolutions, which reduces the number of trainable parameters [28].

These CNN-based approaches have achieved successful results, but also have some bottlenecks, such as using a fixed-size convolution results in a local receptive field. CNNs are designed to successfully extract local context and inherently have low

capacity to extract global context [16]. To capture the global context, the most widely used approach is to implement the attention mechanism [29] into the networks. Several studies introduced the attention mechanism to improve feature representation, thereby they can differentiate buildings from complex backgrounds [30,31]. Furthermore, by concentrating on the network architectural design, numerous studies have been published to enhance the building segmentation performance of the models. Along with the attention mechanisms, approaches such as deep and shallow networks [32], multiple receptive fields and residual connections [33] have been widely used.

The Transformer approach was first proposed to be used in natural language processing (NLP) tasks, and later on, implemented to computer vision studies [10]. To be used in these computer vision tasks, a self-attention-based Vision Transformer (ViT) was proposed in 2020 [34]. ViT achieved competitive results with CNNs on the ImageNet dataset, making it an engaging approach for computer vision studies as a result. As a different approach from CNNs, ViT translates 2D images into 1D arrays and uses self-attention for feature characterization. As an advantage of this design, ViT excels over CNNs in global feature extraction [16], and has made a tremendous job on image segmentation tasks [35]–[37]. Many researchers have implemented Transformer-based approaches to segmentation tasks such as land cover/land use analysis [38,39], urban planning and monitoring [40,41], change detection [42,43], and building segmentation [37,44].

CNNs can only pay attention to small range of local features, and they are insufficient in identifying global features due to their structure. Transformer approaches can compensate for these shortcomings by using the attention mechanism. In ViT-based approaches, global information is extracted, however, spatially detailed context is ignored. In addition, when working with high-resolution, large-sized images, the vector operations of Transformers use all the pixels, and these operations need a large amount of memory and are computationally inefficient [10].

To deal with these mentioned issues, Swin Transformer [45], which utilizes shifted window-based multi-head self-attention, was proposed by Liu et al., but the complexity still increases with the window size. Transformers perform well in global context

modeling, but they also have some limitations in capturing fine-grained detail [46]. To exploit the advantages of both CNNs and Transformers, research was conducted on combining these two approaches, and the Trans U-Net architecture was proposed [47]. This approach uses CNN to extract low-level features and then Transformer to model global interactions.

In this thesis, different CNN and Transformer approaches were examined in order to propose solutions to the problems mentioned above, evaluations were made on the building segmentation from high-resolution aerial images, and a modernized CNN approach was proposed.

1.1 Aim and Objectives of the Thesis

The main aim of this Ph.D. study is to investigate and make comparisons between several different CNN and Transformer approaches, and propose a modernized CNN model towards the ViT design approach for building segmentation from high-resolution aerial images.

The objectives of the study are to assess different CNN and Transformer architectures utilizing different structures and approaches for building segmentation from high-resolution aerial images, propose a modernized CNN model that can cope with the Transformer-based models, and try to determine which approaches are more suitable in which cases. This thesis has been prepared as a combination of four academic papers on this subject. In the first paper, U-Net architecture along with its four deeper and shallower versions were generated, examined, and compared with each other. In the second paper, U-Net and FPN architectures utilizing four different backbones (ResNet, ResNeXt, SE-ResNeXt, and DenseNet) and an Attention Residual U-Net approach were generated and comparisons between them were realized. In the third paper, twelve CNN-based models (U-Net, FPN, and LinkNet architectures utilizing EfficientNet-B5 backbone, original U-Net, SegNet, FCN, and six different Residual U-Net approaches) were generated, evaluated and comparisons between them were realized. Lastly in the fourth paper, ten CNN and Transformer models (Residual-Inception U-Net (RIU-Net), U-Net, Residual U-Net, Attention Residual U-Net, Inception backboned U-Net, Inception-ResNet backboned U-Net, Xception

backboned U-Net, MobileNet backboned U-Net, Trans U-Net, Swin U-Net) were generated, building segmentation from high-resolution aerial images was carried out, and comparisons were realized.

1.2 Research Theme and Structure

The structure of this thesis is organized based on the four SCI-Expanded articles as follows:

Chapter 1, consists of the general introduction, aim and objectives of the thesis, and research theme and structure subsections.

Chapter 2 presents the journal article entitled "Performance Evaluation of Shallow and Deep CNN Architectures on Building Segmentation from High-Resolution Images".

Chapter 3 presents the journal article entitled "Comparisons of Residual and Dense Neural Network Approaches for Building Extraction from High-Resolution Aerial Images".

Chapter 4 presents the journal article entitled "Comparative Analysis of Different CNN Models for Building Segmentation from Satellite and UAV Images".

Chapter 5 presents the journal articled entitled "A Residual-Inception U-Net (RIU-Net) Approach and Comparisons with U-Shaped CNN and Transformer Models for Building Segmentation from High-Resolution Satellite Images".

Chapter 6 presents the conclusions of the presented studies.

2. PERFORMANCE EVALUATION OF SHALLOW AND DEEP CNN ARCHITECTURES ON BUILDING SEGMENTATION FROM HIGH-RESOLUTION IMAGES[1]

2.1 Introduction

Aerial images and high-resolution satellite images are one of the main data sources in geoscience related applications, such as land cover/use mapping and extraction/detection of features that can be listed as buildings, roads, ships, planes, trees, etc. [5]. Extracting features from remote sensing images and aerial photos is critical for applications such as traffic management, urban planning, and industrial development [48].

In recent years, as a result of the developments in sensor technologies, the accessibility and availability of high-resolution satellite images and aerial images have increased significantly [49]. Automatic object extraction using remote sensing images and aerial images has become a popular research topic due to advances in sensor technologies and data processing and analysis architectures [50].

Due to the frequent changes, especially in rapidly growing urban areas, it is crucial to update current databases for the aim of planning these areas [2]. One of the most important components of the urban database is buildings. Buildings, which are the most important place for human livelihood, are the key features for urban area mapping [3]. Building detection and extraction are beneficial for applications such as planning of urban areas, disaster management, change detection, military surveillance, updating geographical databases, and environmental management [5,51]. Building extraction from high-resolution satellite images is a challenging task, which has gathered attention from academia and industry for many years [52]. Remote sensing technologies can acquire images of specific areas repeatedly and serve as effective data sources for these kinds of tasks [2]. Besides, Aerial imagery has been

[1]This chapter is based on the article: Sariturk, B., Seker, D. Z., Ozturk, O., and Bayram, B. (2022). Performance evaluation of shallow and deep CNN architectures on building segmentation from high-resolution images. Earth Science Informatics, 15(3), 1801-1823.

very useful for accurate building mapping due to its ability of imaging over broad areas [3]. Along with this, based on the swift developments of sensors and imaging platforms, the increase in the accessibility and availability of high-resolution images has made this issue increasingly urgent [2]. Moreover, extracted building data from high-resolution satellite images could be used as one of the main data sources for Geographical Information Systems (GIS) [53].

As a result of the diversity of buildings in different regions according to shape, color, size, etc. and the similarity of the buildings to the other objects or background [54], the development of decent and accurate methods for the building extractions has become a challenging and important research subject [5]. Spatial distributions of the buildings are also important for monitoring urban areas and urban demographic modeling [55,56]. Over the past years, traditional methods of image processing, such as object-based [57], shadow-based, or edge-based methods were used in many studies [5]. The metrics that are commonly used such as length, color [58], spectrum [59], edge [60], shape [61], texture [59], shadow [58,62], and height could vary under different conditions such as light, atmospheric state, scale, quality of sensors and architectures of buildings. Building extraction can be done using these manual techniques by experts [58]. However, this process requires a lot of time and high experience. Recent developments in remote sensing technologies and data science studies provide new approaches to automatic building detection which can diminish manual work and contribute to the monitoring of urban dynamics [63].

As a branch of machine learning techniques, deep learning (DL) has provided great progress in studies such as image classification [19,22] and segmentation [24,26,28,64]–[66]. Researchers started to examine DL techniques to solve tasks like object extraction and achieved promising results [67]–[71]. Their studies demonstrate the great potential of these techniques to analyze high-resolution satellite images and aerial images [72].

An important point is that the progress in image classification and segmentation studies is not only the result of larger datasets and powerful hardware but mainly an outcome of new ideas, algorithms, and improved network architectures [23]. Recent developments

and researches indicate that, with the availability of large datasets and adequate computing power [73]–[75], DL techniques such as convolutional neural networks (CNN) [76] can considerably improve the performance of semantic segmentation and object detection from high-resolution imagery [19,66]. This situation shows that DL's role will be important in improving the accuracy of applications such as automatic mapping and building detection [3].

Since their introduction, numerous neural networks have been developed, like shallow and deep neural networks [77]. The terms, deep and shallow, come from the number of layers that the network has [77]. Shallow neural networks have a relatively small number of layers and deep neural networks have multiple hidden layers [77]. Both perform specific tasks better than the other and it is important to select the correct depth to create a successful model [77].

Building extraction from aerial images and high-resolution satellite images using DL techniques has been studied to be helpful with tasks like disaster monitoring and geodatabase updating [53]. Since proposed in 2015, deep CNN-based semantic segmentation algorithms [24] have been used in many studies of pixel-wise image analysis, such as road extraction [78], vehicle extraction [79], building extraction [5,54]–[56], urban land use classification [80], damage mapping [81] and other land cover mapping tasks [5].

Rapid and revolutionary changes in image processing research have been achieved, mainly driven by CNN and the availability and accessibility of large-scale training datasets [82]. With the availability of these datasets, several CNN-based semantic segmentation methods have been used to extract buildings from high-resolution images [83,84]. In the beginning, patch-based CNN methods [56,85] were used for predictions in dense urban areas [82]. These patch-based CNNs process an image patch through the neural network and label the central pixel [82]. But these networks are computationally expensive and are generally used in the detection of large objects [66,86]. Since Long et al. [24] adapted the classification network into the fully convolutional network (FCN), these networks have become the preferred method in

semantic classification and segmentation studies. The FCN is the first pixels-to-pixels and end-to-end network and most future implementations are based on it [83].

As mentioned, in the semantic segmentation of images, early non-FCN-based methods are computationally expensive [58]. Recent studies utilized FCN-based methods for building extraction [2,63]. CNN has been broadly used in segmentation and classification tasks [58]. A variety of FCN-based CNNs such as DeconvNet [87], SegNet [28,88], and U-Net have been proposed [26]. U-Net is a type of FCN that merges high-level and low-level feature maps [89]. This architecture has skip connections between decoder and encoder which prevent boundary information loss by receiving low-level feature maps from the encoder into the decoder [90]. The application of CNN, especially for building extraction tasks, makes the feature design easier and has shown encouraging results [91,92]. The output of these CNNs in image segmentation tasks is generally a one-class label. These achievements have inspired people in remote sensing society to use FCN-based CNN techniques for building extraction studies [52].

Deep CNNs show great advantages in object extraction tasks [93,94]. Nonetheless, the networks have complex structures and they include a large number of layers like convolutional layers and pooling layers [95]. Therefore, they consume valuable memory resources and computing power [95]. This might be overcome by switching to graphic processing units (GPU) which allows an increase in speed [96].

According to Gorokhovatskyi and Peredrii [96], along with those mentioned above, deep networks have two main problems. The first is to select a network structure that is effective enough and the other one is to provide the necessary computing power to train the network [96]. State-of-the-art CNNs like PatreoNet [97], AlexNet [19], CaffeeNet [98], VGGNet [21], and GoogLeNet [23] became computationally complex and with that, the usage of shallow CNNs are started to underestimate [95]. Compared with deep CNNs, shallow CNNs have a simple network structure and fewer parameters to calculate [95]. Thus, training a shallow CNN model needs less computation power and memory [95].

In this study, U-Net architecture has been chosen to be used due to the advantages of not being dependent on large datasets [99] and preventing boundary information loss [90,100], since retaining the correct boundary extraction is highly demanded in building extraction applications. Different variations of the U-Net architecture have been exploited to evaluate their performance.

For this purpose, four U-Net based architectures were exploited and they were compared with each other and the original U-Net architecture. Using these architectures, building extraction from high-resolution images was carried out. To generate these architectures, downsampling and upsampling paths were added or removed from the U-Net architecture. Thus, along with the original U-Net architecture, two shallower and two deeper versions were generated and using these architectures, models were trained from scratch (without using pre-trained backbones) using two different training datasets prepared and tested using two different test datasets prepared. To generate these datasets, a total of 26 images and their corresponding labeled images from the Inria Aerial Image Labeling Dataset and 151 images and their corresponding labeled images from the Massachusetts Buildings Dataset were divided into smaller patches of 512-pixel x 512-pixel. The contribution of this study to previous research is to investigate which depth of the selected U-Net architecture could be more useful for object extraction and segmentation tasks from high-resolution images by making a comparison between deeper and shallower architecture approaches. At the same time, this study aims to contribute to the use of shallower architectures for object extraction from high-resolution images.

2.2 Datasets and Methodology

2.2.1 Datasets

Two different public datasets were used in the study because of the differences in spatial resolution, feature shapes, and sizes to make more accurate comparisons. The datasets are Inria Aerial Image Labeling Dataset (INRIA) and Massachusetts Buildings Dataset. Both datasets include RGB images and their corresponding labeled images. Since the INRIA test dataset is not labeled, it has not been used in this study.

2.2.1.1 Inria aerial image labeling dataset

The first dataset selected to be used in the study is the "Inria Aerial Image Labeling Dataset" [101]. Dataset features:

- 810 km^2 of coverage (405 km^2 for the training set and 405 km^2 for the test set)
- Aerial images with the spatial resolution of 30 cm (Orthorectified and RGB).
- Labeled image data with two classes: building and non-building [102]

During the production of this dataset, training and test sets were created separately using images taken from different cities [102]. The images cover different urban areas, from areas with dense populations (e.g., Chicago) to rural towns (e.g., Lienz from Austrian Tyrol) [102]. All images are open-access and with corresponding official cadastral records available (Figure 2.1) [103].

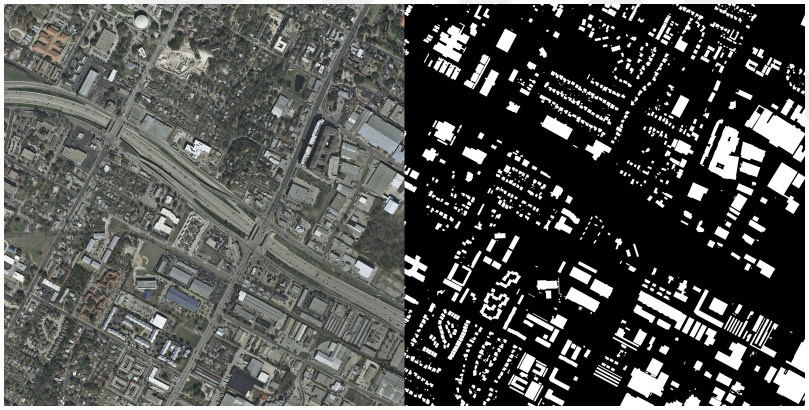


Figure 2.1 : Austin - 7 sample image and labeled image [91].

This dataset's purpose is to improve the power of generalization of the models. While images from one area are used for training, the model should be able to labeled images from other areas with different conditions [102].

In this study, only the training set was used. The training set consists of 180 RGB images of the size 5000-pixel×5000-pixel, each covering an area of 1500 m×1500 m. There are 36 images and their corresponding labeled images for each of the regions below:

- Austin (TX, USA)
- Chicago (IL, USA)
- Kitsap County (WA, USA)
- Vienna (Austria)
- Western Tyrol (Austria) [102]

The format of the data used is GeoTIFF. The pixels of labeled images have values 255 for building class and values 0 for not building class [102].

Because of the computational costs, all 180 images could not be used in the study. A total of 26 images were selected to be used to generate training and test datasets. To prepare training and test datasets, selected images and their corresponding labeled images were cropped into smaller patches of 512-pixel×512-pixel to increase the number of samples, reduce the computational cost and not lose resolution with resizing operation.

To prepare the training dataset in the size of 512-pixel×512-pixel, out of these 26 images, 21 images and their corresponding labeled images ($\approx$ 80%) were selected (selected images are Austin1, Austin2, Austin9, Austin30, Chicago3, Chicago6, Chicago25, Chicago29, Kitsap11, Kitsap18, Kitsap24, Kitsap31, Tyrol_w3, Tyrol_21, Tyrol_w24, Tyrol_w29, Tyrol_w30, Vienna13, Vienna15, Vienna21, and Vienna31). A similar distribution of urban and rural areas was considered during these selections. The images with a low amount of buildings or no buildings were removed from the dataset. Eventually, a total of 1500 images and their corresponding labeled images were prepared as the training dataset. Figure 2.2 shows samples from this dataset.

To test trained models, an independent test dataset was prepared. To prepare this dataset from INRIA, 5 images and corresponding labeled images that were not selected for training were used ($\approx$ 20%). Selected images for the independent test dataset are Austin23, Chicago23, Kitsap19, Tyrol_w23, and Vienna19. Eventually, a total of 300 images were prepared as the test dataset.

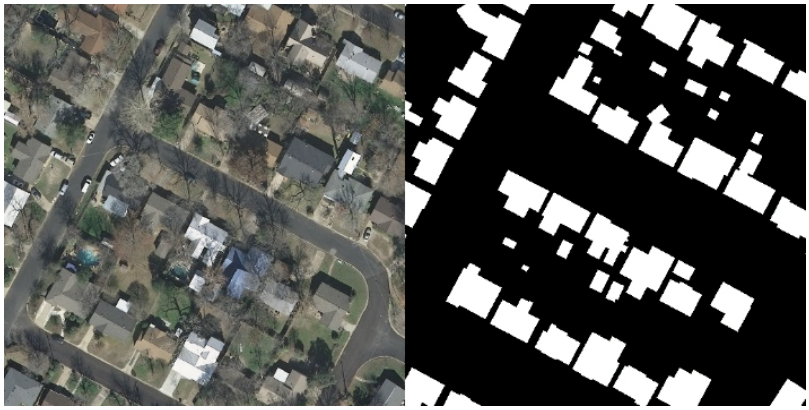


Figure 2.2 : Sample image and labeled image from the INRIA training dataset.

2.2.1.2 Massachusetts buildings dataset

The second dataset used is the "Massachusetts Buildings Dataset". The dataset is created by Volodymyr Mnih [104]. It consists of 151 RGB aerial images from Boston, MA, USA in size of 1500-pixel×1500-pixel for an area of 2.25 km^2. The labeled images were created by Mnih by rasterizing the building footprints obtained from OpenStreetMap [104]. The dataset covers mostly urban areas with buildings of all sizes.

To generate training and test datasets, images and labeled images were split into patches of 512-pixel×512-pixel, like the INRIA set. After the operation, images that have a little amount of buildings or no buildings were eliminated from the dataset. However, only a total of 488 images were able to be obtained after the operation and the elimination. Since this number of images would be insufficient to train and test the models, data augmentation methods were applied to the images.

To increase the sample images, images and labeled images were augmented using data augmentation methods rotation (90 degrees) and brightness adjustment (0.5 gamma rate). With this technique, 2 additional variations of each original image were created. As a result of this process, the dataset was tripled and 1464 images and the relevant labeled images were prepared. A sample image and corresponding labeled image are presented in Figure 2.3. Out of these prepared images, 1242 images and corresponding labeled images ($\approx 85\%$) were selected randomly to be used as training dataset, and the remaining 222 images and corresponding labeled images ($\approx 15\%$) as test dataset.

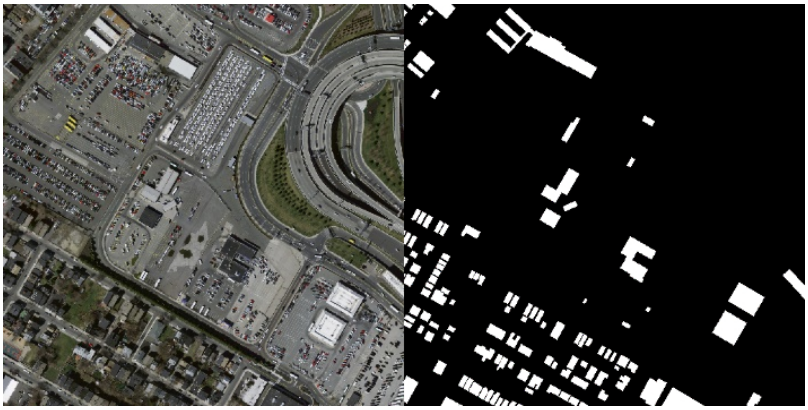


Figure 2.3 : Sample image and labeled image from the Massachusetts training dataset.

2.2.2 Methodology

2.2.2.1 U-Net architecture

U-Net was selected as a base architecture for this study. U-Net is a CNN that was developed at the Computer Science Department of the University of Freiburg, Germany [26]. U-Net is based on FCN and it was modified to work with fewer images and to provide more accurate segmentation results [26]. Compared to FCN, U-Net has two main differences:

- It has a symmetric structure
- Skip connections between downsampling and upsampling [26]

While upsampling, skip connections transfer local information to global information [26]. Because of the symmetrical structure, U-Net has a large number of feature maps in the upsampling path that allows transferring the information [26]. On the other hand, in the upsampling path, FCN architecture has 'number of classes' feature maps [26].

The U-Net architecture consists of 3 parts (Figure 2.4):

- Downsampling / contracting path
- Bottleneck

- Upsampling / expanding path [26]

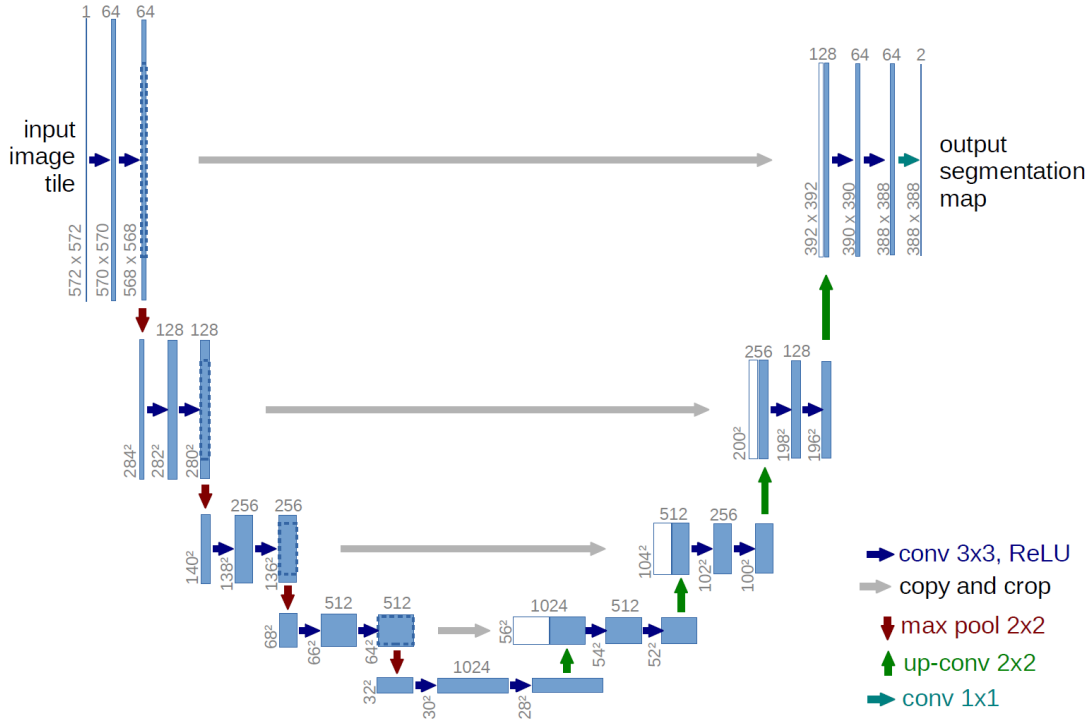


Figure 2.4 : U-Net architecture [26].

The downsampling/contracting path consists of 4 paths. Each path has:

- 3×3 convolution layer and activation function
- Another 3×3 convolution layer and activation function
- 2×2 max-pooling [26]

As in U-Net, where lower-level feature maps are combined with higher-level features, that enable precise localization [105], a CNN usually comprises convolution, max-pooling, and dropout layers.

The convolutional layer is one of the most important layers used in CNN and it provides effective extraction of the characteristics of the data [95]. In image processing and computer vision, the convolutional layer is mostly used as a layer of filtration that allows retrieving features of an image [96].

Rectified linear unit (ReLU) is a non-linear activation function used in the convolutional layer for U-Net [95]. ReLU activation function equation is defined as in (2.1) [95]:

$$f(x) = \max(0, x) \tag{2.1}$$

Max-pooling is a downsampling approach that provides aggregation to the image to leave only the most valuable values [96]. Max-pooling layer's output is given by the maximum activation over non-overlapping, rectangular-shaped regions of kernel size (K_x; K_y) [106]. Max-pooling selects superior invariant features that lead to faster convergence and it improves the performance of generalization [106].

According to the depth of a DL architecture, the supervision information gathered from the loss function will vanish because of the long back-propagation path, particularly for shallow layers [22]. Thus, the intermediate layers will receive less information and this will lead to redundancy in the model which can be resulted with overfitting [22]. Dropout layers are used to speed up the training process and at the same time to avoid overfitting [96]. They realize this by setting zero values to some of the random input neurons [96].

The bottleneck part is between upsampling and downsampling paths. Bottleneck consists of 2 convolutional layers [26].

Upsampling (expanding) path consists of 4 paths. Each path is composed of:

- Convolutional transpose layer
- Concatenation layer (with the corresponding feature map cropped in the upsampling path)
- 3×3 convolution layer, activation function
- 3×3 convolution layer, activation function [26]

The purpose of the upsampling path is to capture the context of the input image to be able to perform segmentation and allow precise localization merged with contextual

information obtained from the downsampling path [26]. After this, raw information will be transferred to the upsampling path through skip connections.

Additional to convolution and dropout layers, the downsampling path has convolutional transpose layers and concatenation.

The convolutional transpose layer (or deconvolutional layer) swaps the forward and backward passes of a convolution [107]. A convolution is defined by the kernel, but how the forward and backward passes are computed defines if it is a direct convolution or a transposed convolution [107]. It is possible to emulate the transposed convolution layer using a direct convolution layer [107]. But, it usually involves adding many columns and rows, which generate from zeros, to the input and the result is a less efficient application [107].

The main advantages of U-Net:

- It merges location information obtained from the downsampling path and the contextual information obtained from the upsampling path to acquire general information generated from localization and context which is crucial to perform good predictions
- No dense layer is used; thus, different sizes of images can be used as inputs
- The usage of data augmentation is important because the number of labeled samples is generally limited [26]

Eventually, the architecture consists of the repeated applications of two 3×3 convolutions, each one followed by a ReLU function and a 2×2 max-pooling for the downsampling path [26]. At each downsampling step, feature channel numbers double [26].

Every step in the upsampling path consists of a feature map upsampling followed by a 2×2 convolutional transpose layer which halves the feature channel numbers, a concatenation layer with the corresponding feature map cropped in the upsampling path, and two 3×3 convolutions, each one followed by a ReLU function [26]. At the final layer, a 1×1 convolution is used in the mapping of each 64-component feature

vector to the number of classes [26]. In total, the network consists of 23 convolutional layers [26].

2.2.2.2 Configuration of the networks

Structuring a successful architecture and selecting the optimal depth of layers for any specific task is a continuing discussion in DL society [108]. In this study, along with the original U-Net architecture, two shallower, and two deeper versions of the architecture were generated. To generate these architectures, downsampling and upsampling paths were added or removed (Figure 2.5).

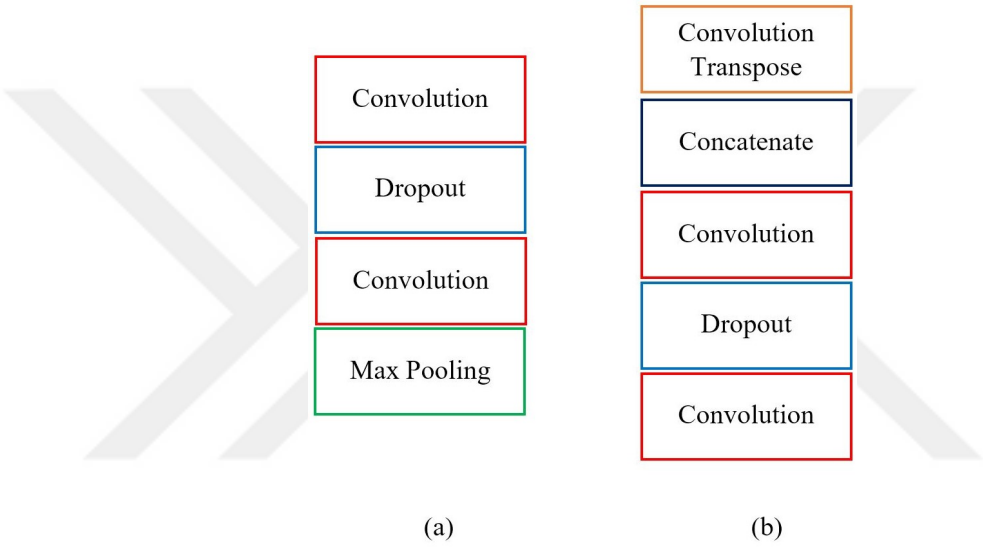


Figure 2.5 : (**a**) Layers of the downsampling path, (**b**) Layers of the upsampling path.

For the first shallow network, one downsampling path and one upsampling path were eliminated from the original one. For the second shallow network, another path from the downsampling path and another from the upsampling path were eliminated. To generate the first deep network, similarly to shallow ones, one path was added to the downsampling path, and one path was added to upsampling. Finally, for the second-deep network, another one was added to the downsampling path, and another path was added to the upsampling path. As a result, a total of five architectures have been generated to compare. Original U-Net architecture consists of four downsampling and four upsampling paths. The first shallow architecture consists of three downsampling and three upsampling paths, the second-shallow architecture consists of two downsampling paths and two upsampling paths, the first

deep architecture consists of five downsampling and five upsampling paths, and the second-deep architecture consists of six downsampling and six upsampling paths (Figure 2.6).

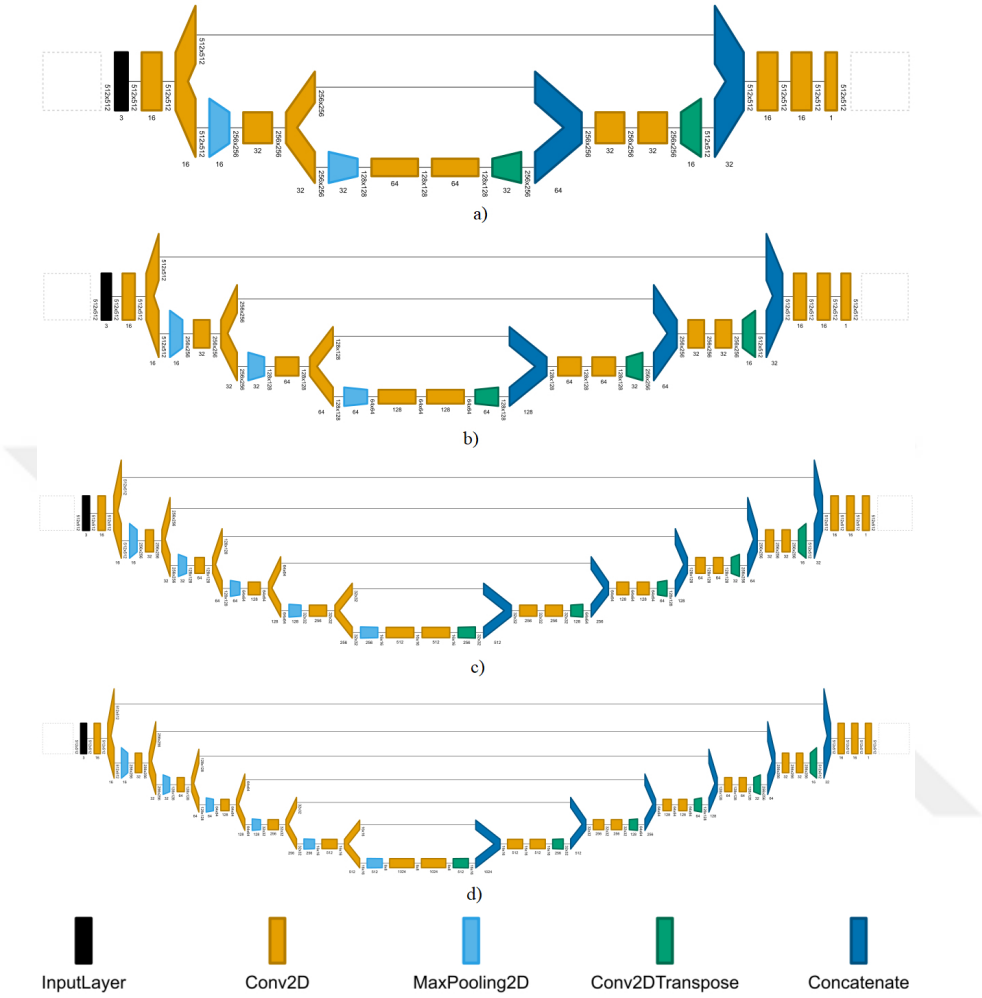


Figure 2.6 : (**a**) U-Net shallower 2 architecture, (**b**) U-Net shallower 1 architecture, (**c**) U-Net deeper 1 architecture, (**d**) U-Net deeper 2 architecture (Visualized using [109]).

2.3 Training and Evaluating the Models

The U-Net and four different shallower and deeper versions of the architecture were used to train models using prepared 512-pixel×512-pixel training datasets. In order to train the models, the K-Fold Cross-Validation method was applied to training datasets with a k value of 5 (Figure 2.7). Cross-validation is a statistical method to evaluate the generalization performance of the model as objectively and accurately as possible on data that has not been seen before. This method divides all the samples in the training

set into ‘k’ number of splits, called folds. The model is trained using k-1 folds, and the fold left out is used for validation [110].

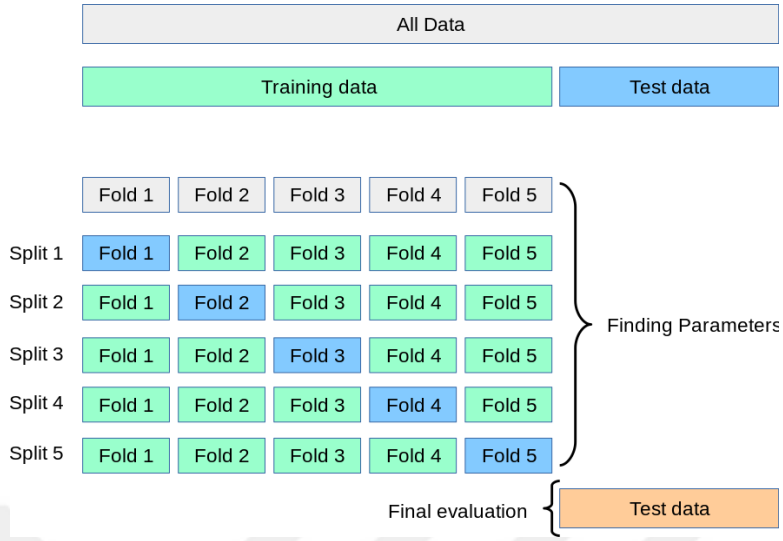


Figure 2.7 : K-fold cross-validation [111].

Adam is an adaptive learning rate optimization algorithm, it computes individual adaptive learning rate values for different parameters from estimates of first and second moments of the gradients [112]. In this study, the Adam algorithm has been chosen. An optimum learning rate was selected as 0.0001 for all implemented DL models empirically. All models are trained for 50 epochs and with a batch size of 16. It has been observed that after the 50^{th} epoch there were not any significant improvements in training processes. Because of the computational costs, the Google Colab environment could not be used with a batch size higher than 16 for all DL models. The same hyperparameters were used for the training of all models. To evaluate the training performances of the models, accuracy and loss metrics were used. Accuracy (Equation 2.2) is an intuitive performance measure and simply a ratio of correct predictions (true positives (TP) and true negatives (TN)) to the total samples (TP, TN, false positives (FP), and false negatives (FN)) [110].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.2}$$

To calculate the loss value, binary cross-entropy was used. Binary cross-entropy (Equation 2.3) is a loss function used on binary decision problems. It is a cross-entropy loss added Sigmoid activation function [113].

$$L(y,\hat{y}) = \frac{1}{N} + \sum_{i=0}^{N} (y * \log(\hat{y}_i) + (1-y) * \log(1-\hat{y}_i)) \tag{2.3}$$

All training and testing processes were performed on the Google Colab environment. All experiments were achieved with Tensorflow and Scikit-Learn libraries using Python programming language.

In terms of computational cost, as the architecture deepens, the models have more parameters to calculate. To reduce the computational cost, the numbers of filters of convolutional layers were used as 16, 32, 64, and 128, which are 64, 128, 256, and 512 in the original paper. The total number of trainable parameters are given in Table 2.1.

Table 2.1 : The total number of parameters of the architectures.

Architecture	Number of Parameters
Shallower 2	117,041
Shallower 1	482,033
U-Net	1,941,105
Deeper 1	7,775,601
Deeper 2	31,110,001

For the models that use the INRIA dataset, changes in validation accuracy, training accuracy, validation loss, and training loss over 50 epochs are shown in Figure 2.8. According to the accuracy results given in Figure 2.8, U-Net Deeper 1 architecture has the highest validation accuracy, followed by the U-Net architecture. It has been observed that the validation accuracies of the U-Net and two deeper architectures remained stable after 20 epochs. However, when shallower architectures were used, the validation accuracy increased even after 20 epochs, and the differences between training and validation accuracies were decreased.

According to the results shown on loss graphics, overfitting was observed in the models that use the U-Net architecture and deeper architectures. In these models, training and validation loss values decrease together until the 20th epoch. After the 20th epoch,

the training loss tends to decrease while validation loss remains constant. Along with these, it can be said that using a shallow architecture prevents overfitting problems.

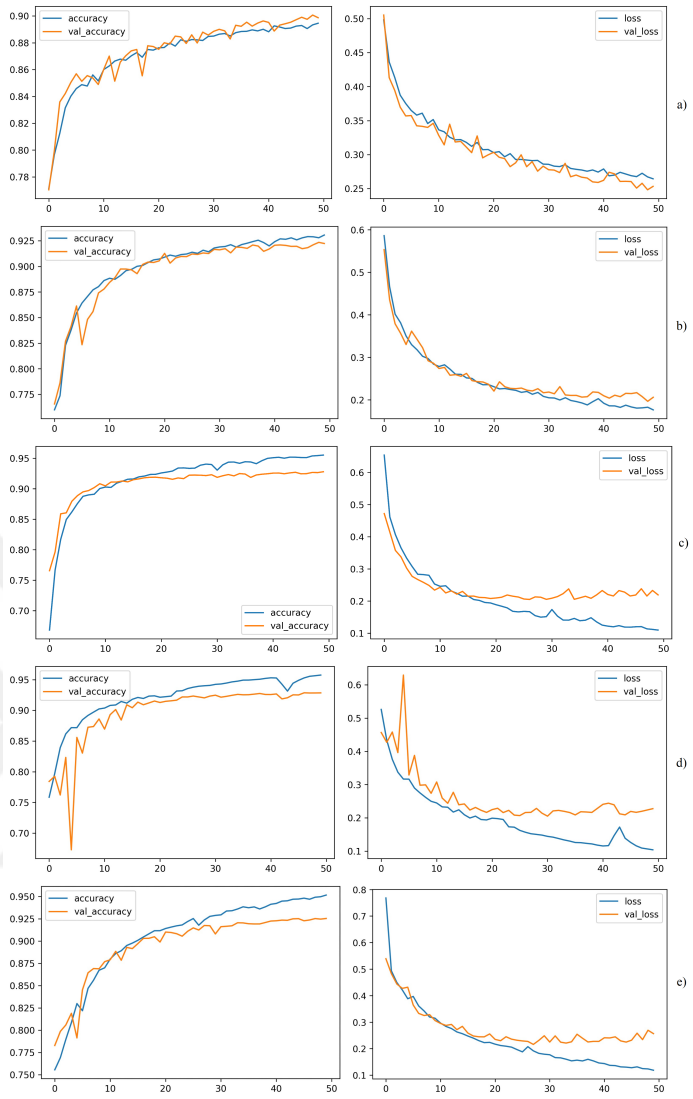


Figure 2.8 : Accuracy and loss values of the INRIA models over 50 epochs using k-fold cross validation (**a**) shallower 2, (**b**) shallower 1, (**c**) u-net, (**d**) deeper 1, (**e**) deeper 2.

For the models that use the Massachusetts dataset, validation accuracy, training accuracy, validation loss, and training loss values over 50 epochs are shown in Figure 2.9. According to the results shown on the accuracy graphics, the model that uses U-Net architecture has the highest validation accuracy, followed by U-Net Deeper 2 and U-Net Shallower 1 architectures.

According to the results shown on loss graphics, overfitting is observed in the model that uses U-Net. During the training of this model, training, and validation loss values decreased together until the 25th epoch. After the 25th epoch, training loss values continued to decrease while validation loss values remain constant. This problem disappears as architectures become shallow. According to the differences between training and validation loss values which is a sign of if there is overfitting or not, the model that uses U-Net Shallower 2 architecture has the lowest difference, but with higher loss values. U-Net Shallower 1 architecture has the second-lowest difference with a low validation accuracy value.

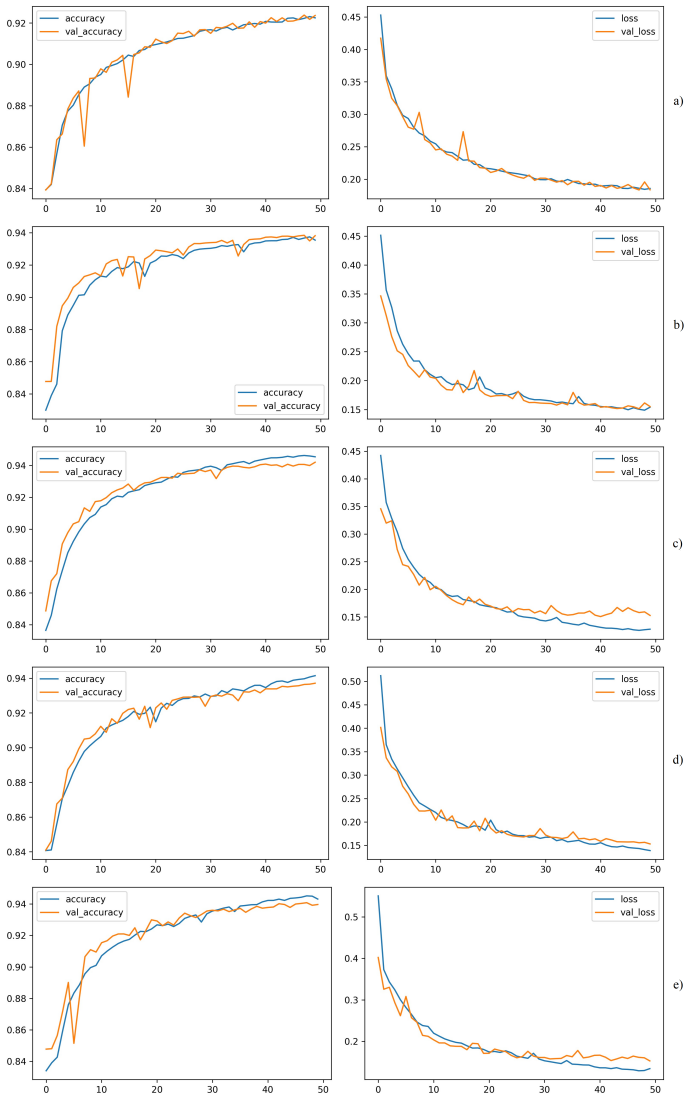


Figure 2.9 : Accuracy and loss values of the Massachusetts models over 50 epochs using k-fold cross validation (**a**) shallower 2, (**b**) shallower 1, (**c**) u-net, (**d**) deeper 1, (**e**) deeper 2.

To evaluate the models using prepared test datasets, recall, precision, F1 score, and intersection over union (IoU) evaluation metrics from the Scikit-Learn library were used. Precision (Equation 2.4) is the ratio of TP values to the total positive (TP + FP) values. Recall (Equation 2.4) is the ratio of TP values to all positive class values in the test dataset (TP + FN) [110].

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \tag{2.4}$$

F1 score (Equation 2.5) is a function calculated using precision and recall. This function is the weighted average of recall and precision. Thus, the F1 score takes both FP and FN values into account [110].

$$F1Score = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.5}$$

IoU (Equation 2.6) also referred as the Jaccard Index, is an evaluation metric used to understand the similarities between sample sets. This measurement points up the similarities between finite sample sets and it is formally defined as the size of the intersection area divided by the size of the union of the sample sets [114].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{2.6}$$

2.4 Results and Discussions

Final accuracy results of the INRIA model are shown in Figure 2.10. According to the analysis of assessment results, it has been observed that U-Net Deeper 1 architecture has the highest test accuracy with 90.53%. Followed by U-Net Deeper 2 architecture with 90.11%, the U-Net architecture with 89.95%, U-Net Shallower 1 architecture with 88.70%, and lastly U-Net Shallower 2 architecture with 87.11%.

According to the results, when the architecture was shallower than 1 path, this affected the performance negatively. Approximately 1.6% decrease was observed in the test accuracy of the U-Net Shallower 2 architecture compared to the U-Net Shallower 1 architecture. A similar decrease in test accuracy was seen when the architecture went

deeper than 1 path. When the differences between training and validation accuracies, which are an indicator of the performance of the model, are examined, it has been observed that U-Net Shallower 1 and U-Net Shallower 2 architectures have the two lowest difference values among all.

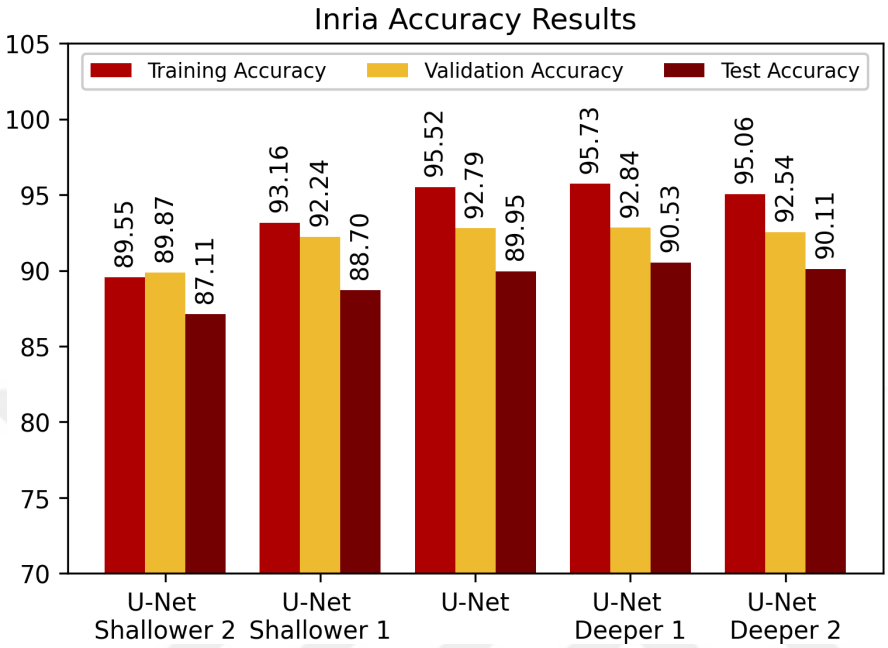


Figure 2.10 : Training, validation, and test accuracy values of the architectures on the INRIA dataset.

Final accuracy results of the Massachusetts model are shown in Figure 2.11. According to the test accuracy results, it was observed that the U-Net architecture has the highest accuracy with 92.85%. Followed by U-Net Deeper 2 architecture with 92.58%, U-Net Shallower 1 architecture with 92.42%, U-Net Deeper 1 architecture with 92.38%, and lastly U-Net Shallower 2 architecture with 90.81%.

According to the examination of the differences between training and validation accuracies to evaluate the training performances of the models, it was observed that all the models have nearly similar difference values, approximately 0.001. This shows that all approaches have similar training performances on this dataset.

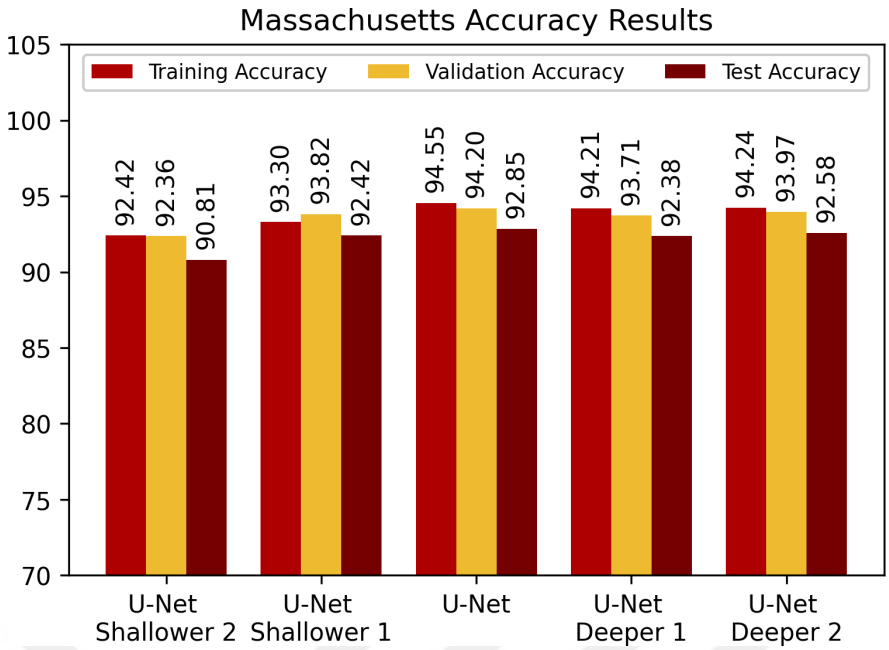


Figure 2.11 : Training, validation, and test accuracy values of the architectures on the Massachusetts dataset.

To examine loss values, the Binary cross-entropy loss function was used to calculate losses. The loss function results of the INRIA model are shown in Figure 2.12. According to the examinations of loss values on the test set, it was observed that the model which uses U-Net Shallower 2 architecture has the lowest test loss value with 0.316, followed by U-Net Shallower 1 architecture with 0.322 and U-Net Deeper 1 architecture with 0.333.

When the differences between training and validation loss values were examined, it was seen that this value increased as the architecture deepened. U-Net Deeper 2 architecture has the highest difference value with 0.136, followed by U-Net Deeper 1 architecture with 0.124 and the U-Net architecture with 0.109. It has been observed that this value decreases significantly as the architecture used becomes shallower. This is an indication that the performance of the model increases as the architecture becomes shallower in the dataset used.

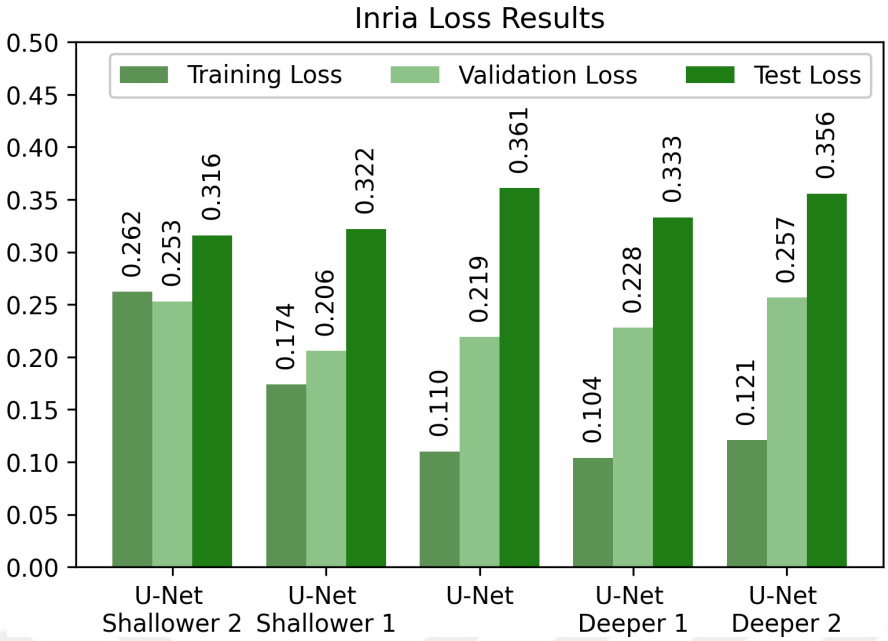


Figure 2.12 : Training, validation, and test loss values of the architectures on the INRIA dataset.

The loss function results of the Massachusetts model are shown in Figure 2.13. According to the results, it has been observed that U-Net Deeper 2 architecture has the lowest test loss value with 0.190. Followed by U-Net Deeper 1 architecture with 0.191 and U-Net Shallower 1 architecture with 0.192.

According to the differences between training and validation loss values, the U-Net architecture has the highest difference value with 0.025, followed by U-Net Deeper 2 architecture with 0.017 and U-Net Deeper 1 architecture with 0.016. Like models using the INRIA dataset, the performance of the model has been observed to increase as the architecture gets shallower.

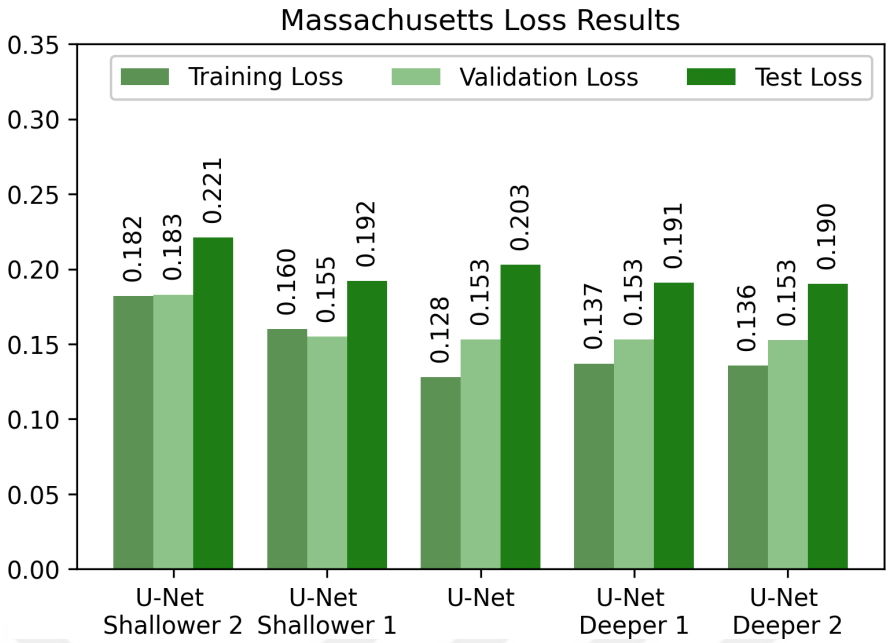


Figure 2.13 : Training, validation, and test loss values of the architectures on the Massachusetts dataset.

Recall, precision, F1, and IoU scores of all architectures using the INRIA dataset are shown in Figure 2.14. According to the F1 and IoU scores results, it has been observed that U-Net Deeper 1 architecture has the highest score among all with 0.785 F1 score and 0.659 IoU score. Followed by U-Net Deeper 2 architecture with 0.780 F1 score and 0.653 IoU score, the U-Net architecture with 0.776 F1 score and 0.646 IoU score, and U-Net Shallower 1 architecture with 0.740 F1 score and 0.604 IoU score. U-Net Shallower 2 architecture has the lowest performance among all with a 0.681 F1 score and 0.536 IoU score.

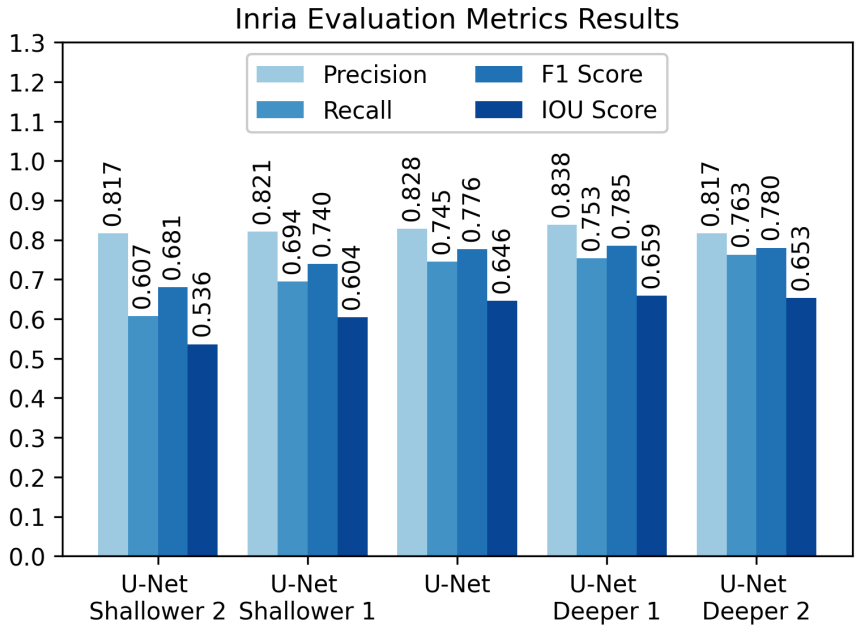


Figure 2.14 : Precision, recall, F1 score, and IoU scores of the architectures on the INRIA dataset.

Recall, precision, F1, and IoU score metrics results of models that use the Massachusetts dataset are shown in Figure 2.15. According to the examination of metric results, the U-Net architecture has the highest F1 score with 0.787 and the highest IoU score with 0.658. Followed by U-Net Deeper 2 architecture with 0.781 F1 score and 0.649 IoU score, U-Net Deeper 1 architecture with 0.772 F1 score and 0.638 IoU score, and U-Net Shallower 1 architecture with 0.771 F1 score and 0.636 IoU score. U-Net Shallower 2 architecture has the lowest F1 score with 0.729 and the lowest IoU score with 0.584.

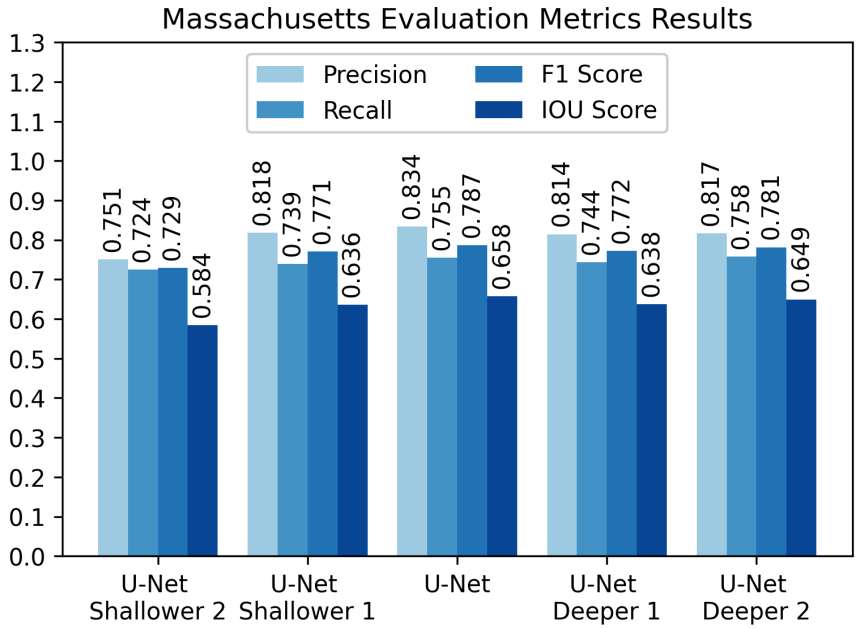


Figure 2.15 : Precision, recall, F1 score, and IoU scores of the architectures from the Massachusetts dataset.

Test, labeled, and segmented image examples obtained by mentioned models are shown in Figures 2.16, 2.17, 2.18, and 2.19. All prediction processes were performed with 0.5 threshold value. In predicted images, white pixels show segmented buildings and black pixels show the background.

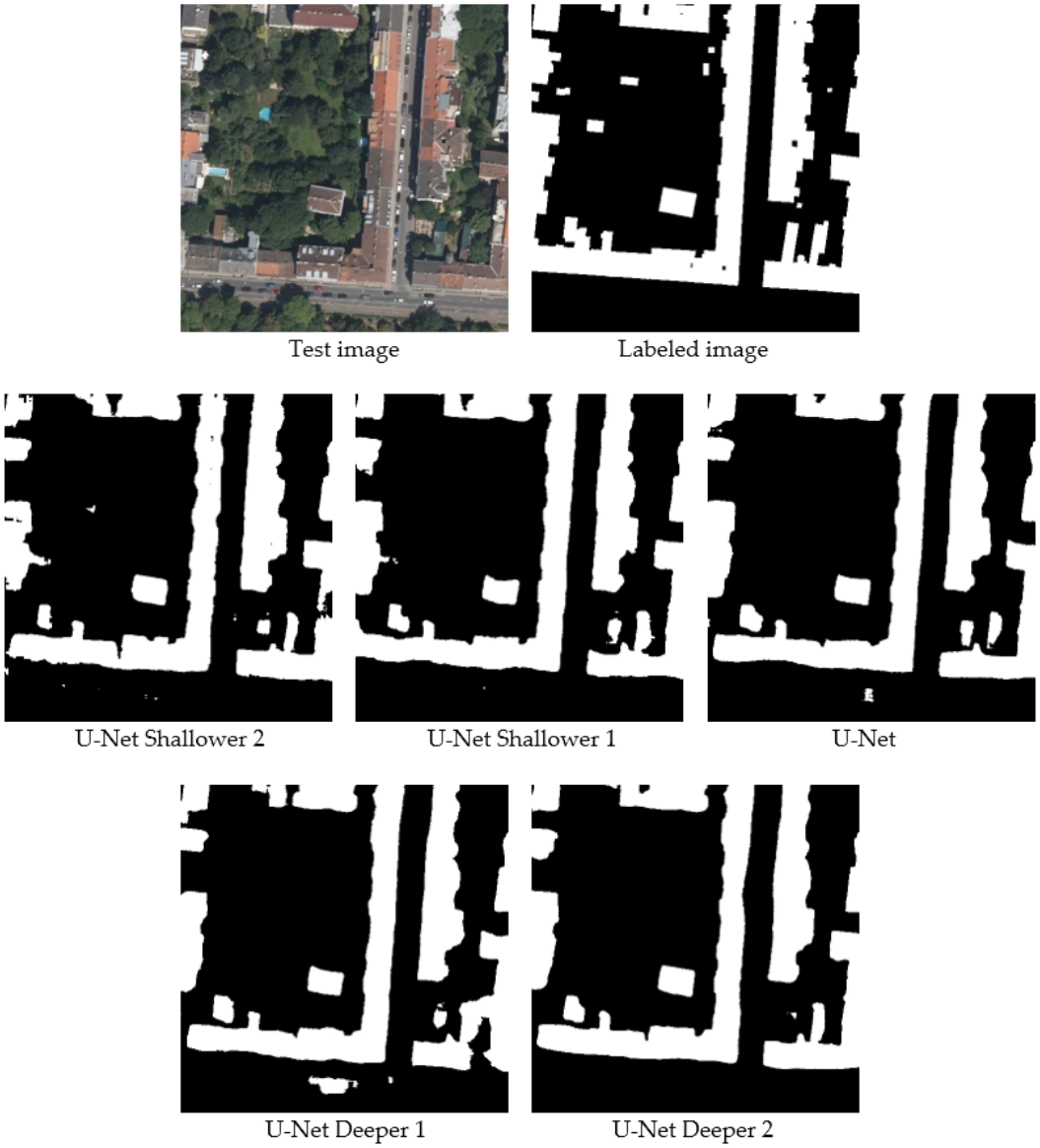


Figure 2.16 : Segmentation results for test image no. 61 from the INRIA test set.

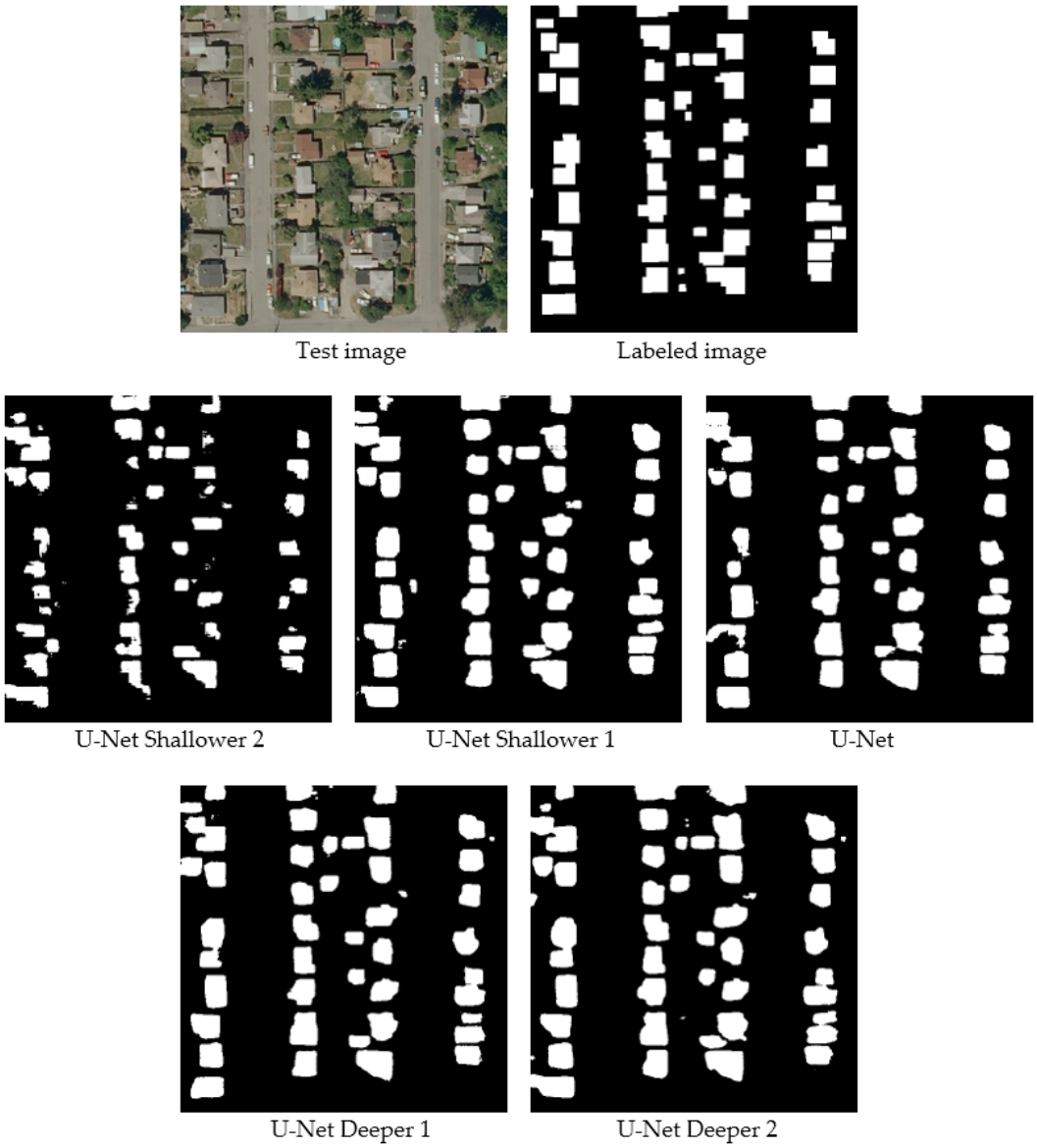


Figure 2.17 : Segmentation results for test image no. 156 from the INRIA test set.

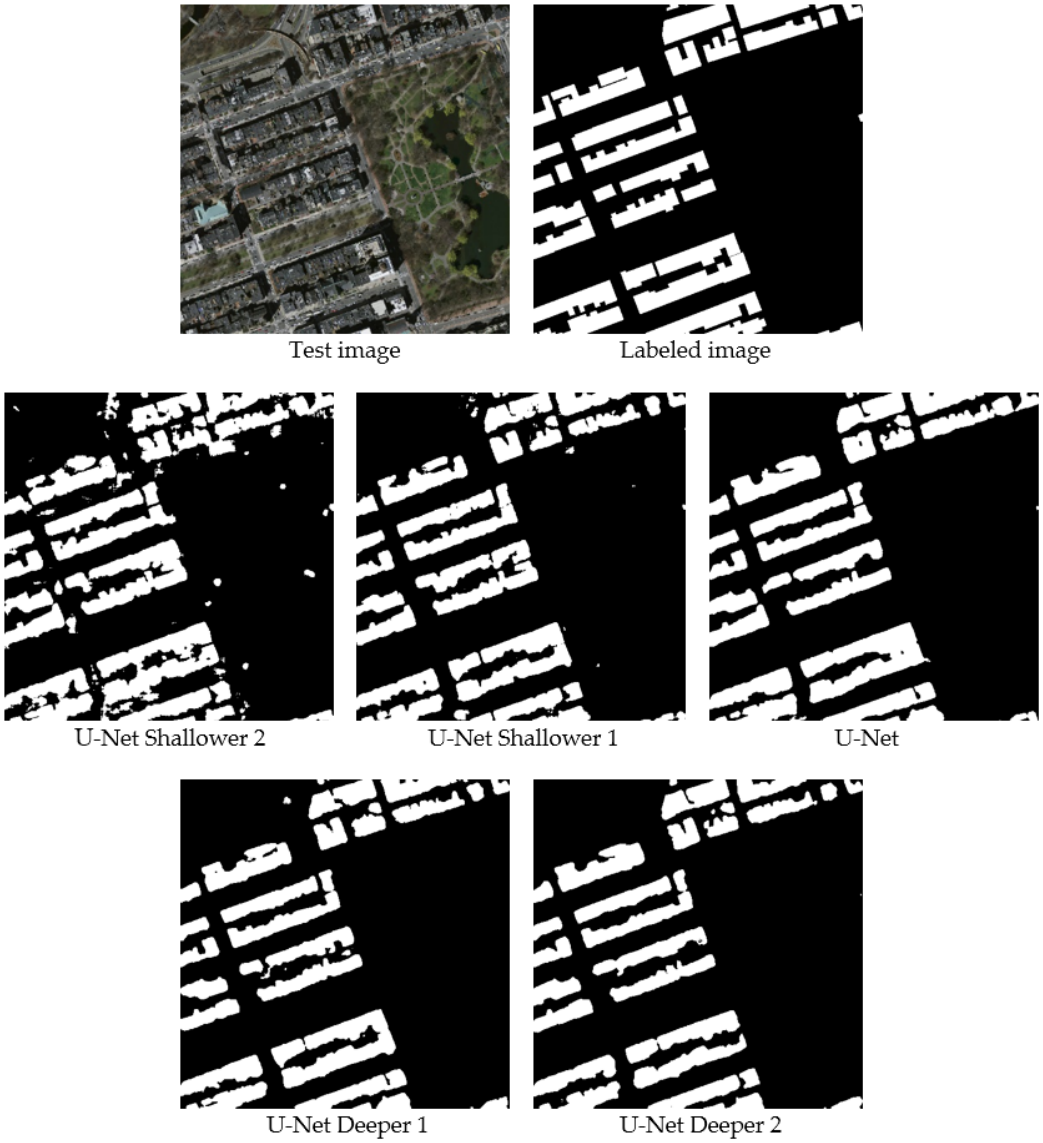


Figure 2.18 : Segmentation results for test image no. 60 from Massachusetts test set.

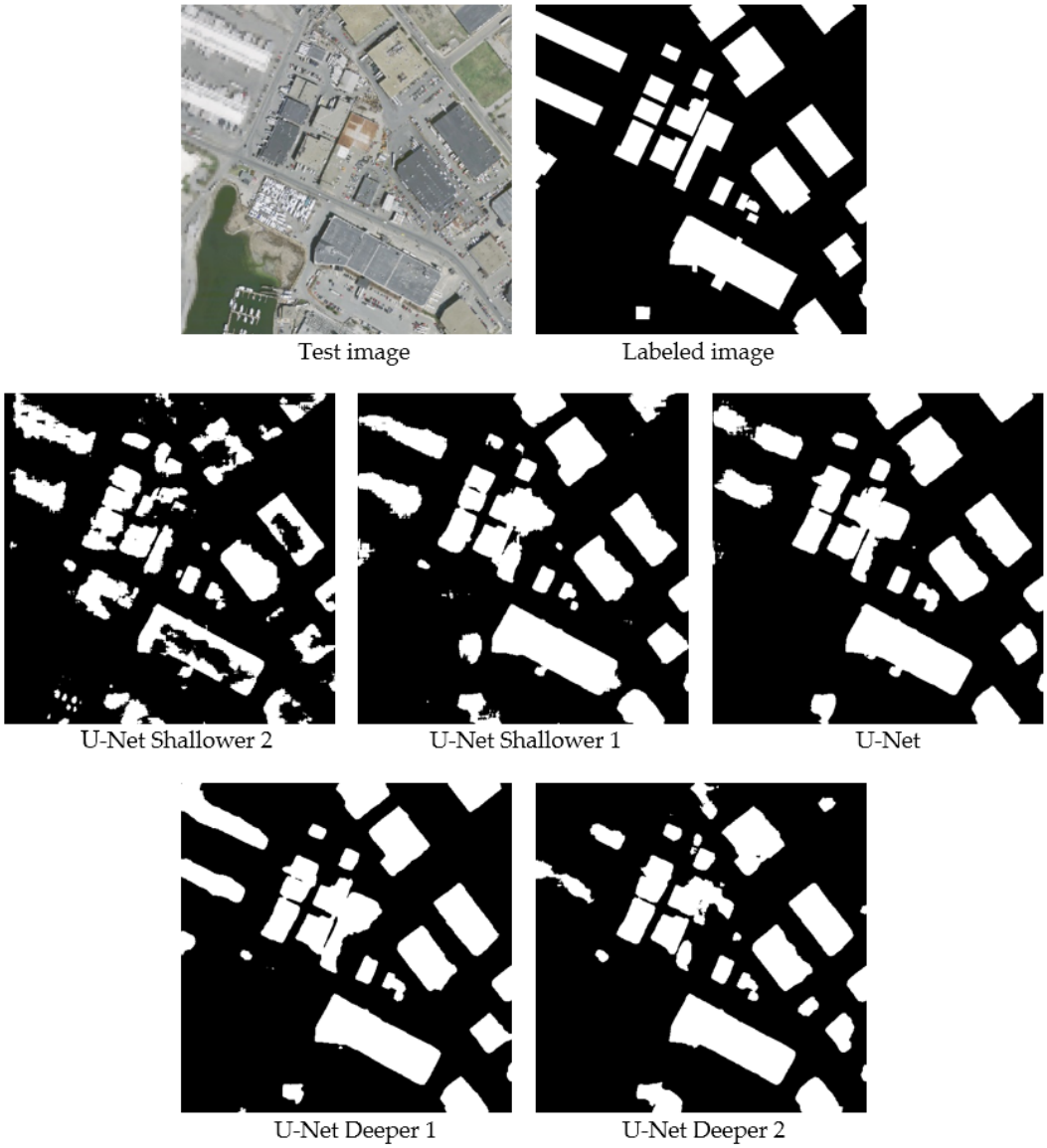


Figure 2.19 : Segmentation results for test image no. 156 from Massachusetts test set.

The results of the evaluations regarding the effects of architectural depth on accuracy have shown that the U-Net Deeper 1 architecture has the highest validation and test accuracies when using the INRIA dataset, and the U-Net architecture has the highest when the Massachusetts dataset is used. As a result of examining the models using the Massachusetts dataset, a maximum 0.5% test accuracy difference was observed between the U-Net architecture and the architectures behind the ranking, except U-Net Shallower 2 architecture. U-Net Shallower 2 architecture was found to have the lowest accuracy values in both datasets. The fact that the U-Net Deeper 1 architecture using

the INRIA dataset has higher accuracy than the U-Net Deeper 2 architecture and, the U-Net architecture using Massachusetts dataset has higher accuracy than deeper architectures indicates that the deepening of the architecture to a certain level can increase the accuracy, but further deepening shows a decrease. However, if the data is increased, it can be expected that the performance of deep architectures will increase.

When the effects of architectural depth on loss values were examined, it has been seen that U-Net Shallower 1 and U-Net Shallower 2 architectures have the lowest test loss with 0.19 when the INRIA dataset was used. For the Massachusetts dataset, U-Net Shallower 1, U-Net Deeper 1, and U-Net Deeper 2 have the lowest test loss values with 0.19. It is observed that for both datasets, U-Net Shallower 1 is the only architecture that is among the best performers.

Based on the examined evaluation metrics results, it was observed that U-Net Deeper 1 has the highest F1 and IoU metric scores when using INRIA dataset with 0.79 and 0.66 respectively. It is followed by U-Net Deeper 2 and the U-Net architectures with 0.78 and 0.65. When the Massachusetts dataset was used, the U-Net architecture provided the highest metric scores with 0.79 F1 and 0.66 IoU, followed by U-Net Deeper 2 architecture with 0.78 F1 and 0.65 IoU scores. The U-Net Shallower 2 architecture has the lowest evaluation metric results when using both datasets. It can be said that this architecture is too simple for object extraction from high-resolution images. It is also noteworthy that both the F1 and IoU scores of the most successful and successive architectures in both datasets differ by 1%.

2.5 Conclusions

In this study, U-Net based shallow and deep CNN architectures were investigated for building extraction from high-resolution images. Experimental comparisons were made between the original U-Net architecture, which was used frequently in studies, in addition to four other variants of the architecture. This paper evaluates the effects of depth and shallowness of the architecture in building extraction studies. Models were trained and tested using generated training and test datasets from two different datasets, the INRIA dataset, and the Massachusetts Buildings Dataset.

The study aimed to determine which architecture depth could be more beneficial for building extraction studies using high-resolution images by making comparisons between deep, shallow, and the original versions of U-Net architecture. It was aimed to make a comparison with the same parameters and the same amount of data, only by changing the depth of the architectures.

The fact that U-Net Deeper 1 and the original U-Net architectures yield the most successful results in INRIA and Massachusetts datasets containing a close number of images suggests that deep architectures can give better results even if there are not too much data. With regard to shallow architectures, the U-Net Shallower 2 gave the worst results in both datasets, as the structure remained too simple. U-Net Shallower 1, however, provided the same evaluation metric results as U-Net Deeper 1 in the Massachusetts dataset.

Besides the successful results of these deep architectures, an important disadvantage is that they require high computational power. U-Net Shallower 1 architecture appears to have a performance not far behind deep architectures (e.g., in Massachusetts dataset, 1-3% lower than more successful architectures according to F1 and IoU metrics) and fewer parameters to compute. It can be said that U-Net Shallower 1 can be used as an alternative architecture without sacrificing too much performance in studies where powerful hardware cannot be used. This can be interpreted as an indication of the usefulness of shallow architectures in geographic applications. With this characteristic, it can also be used as a base for future architectural development studies.

Further studies will be focusing on improving the network's performance. To achieve this, the datasets could be expanded using unused images from the sets and with the help of data augmentation approaches. Having more data to train and validate the model would increase its performance and accuracy, especially when using deep architectures. Experiments will be made on the number of layers in downsampling and upsampling paths which form the architecture to increase the performance of the model. To further improve the performance of the model, used hyperparameters will be optimized. Hyperparameter optimization is essential to achieve the maximum possible potential of the model. Hyperparameters like the number of layers, number of filters

and filter sizes, activation functions, learning rate, and number of epochs will be tuned to improve performance.

3. COMPARISONS OF RESIDUAL AND DENSE NEURAL NETWORK APPROACHES FOR BUILDING EXTRACTION FROM HIGH-RESOLUTION AERIAL IMAGES[2]

3.1 Introduction

In recent years, with the rapid developments of urban areas, automatic building extraction has become a significant research topic [115]. Generating and updating building maps from aerial images is an expensive, time-consuming, and generally manual process [18]. Applications such as urban planning [116], illegal building detection [117], change detection [118], population estimation [18], and disaster management [11] require accurate building extraction. Recently, with the help of rapid and continuous developments in sensor and satellite technology, a large number of satellite image data has become available and a preferable data source for building extraction applications due to its structural, spatial, and textural information [18,119]. But this also raises some new challenges from the diversity of characteristics of buildings (shape, size, color, etc.), complex backgrounds, and occlusions [119]–[121]. Thus, improving the accuracy and efficiency of automatic building extraction is still a challenging task and it has been a focus of many researches [5,115].

Technologies used for building extraction have also developed rapidly in recent years [115]. At first, most of the studies have used conventional machine learning techniques to process relevant objects by selecting features manually. Machine learning methods such as K-Means [122], Support Vector Machines [123], Random Forests [124], Conditional Random Fields (CRF) [125], and AdaBoost [126] have been used for pixel-wise analysis. However, all these methods require prior knowledge [119], and their poor generalization capabilities due to manual feature selection may cause inaccurate results [117]. Lately, deep learning methods, especially convolutional neural networks (CNN), have emerged as successful tools for computer vision tasks

[2]This chapter is based on the article: Sariturk, B., and Seker, D. Z. (2022). Comparison of Residual and Dense Neural Network Approaches for Building Extraction from High-Resolution Aerial Images. Advances in Space Research.

[87], due to the increase in computing power and available large datasets [119], as well as its powerful ability to extract high-level features automatically [18]. CNNs can learn semantic features directly from the input, unlike the traditional machine learning methods that need manual feature selection [119]. In the first studies, patch-wise CNNs have outperformed conventional machine learning methods [90]. Later on, based on CNNs, Long et al. [24] introduced fully convolutional networks (FCN) to be used in semantic segmentation studies. They changed fully connected layers with convolutional layers and then used transposed convolutional layers to restore the original size of the images, which change after the convolution operations. FCN is the first network proposed to efficiently transform deep CNNs for dense labeling. The transformation allows FCN to benefit from the pretrained classification CNNs and to create prediction maps, which have the size of the input [127]. With many publicly available building datasets, FCN has reached a notable performance in both accuracy and computation time [18,128].

In the last few years, several FCN based models have been introduced to achieve more successful segmentation results, such as U-Net [26] and Feature Pyramid Networks (FPN) [27]. In 2015, [26] introduced U-Net architecture. They add an upsampling stage based on FCN and add many feature channels. In this way, more original image texture information can be spread in the high-resolution layer [18]. U-Net is an FCN with symmetric encoder and decoder paths, and skip connections between them. By the connections between the corresponding layers of the encoder and the decoder, high-level and low-level features are merged [129]. Since these skip connections are used to compensate for the semantic gap, U-Net has a good performance on building extraction tasks [115]. In 2017, Lin et al. [27] introduced FPN architecture. FPNs are generally used in one-stage and two-stage object detection algorithms to solve the disturbance resulting from the scale differences of the objects [130]. It combines semantic information of deep and shallow layers and the prediction is accomplished in different feature layers [130]. As mentioned, FPN was originally developed to be used for object detection, due to its multi-scale analysis capability. But it could also be used for image segmentation tasks [129].

Recent advances of the deep learning methods provide many ideas to improve architectures, especially about U-Net which is one of the most used ones for image segmentation tasks. One of these ideas is residual blocks, which are proposed in ResNet [22]. Residual blocks provide a refined way to stack layers of CNN and show their power when used within the U-Net [131]. At the same time, an attention mechanism was also used to improve the performance of U-Net and other deep learning methods to improve their performance [30].

In this study, U-Net and FPN architectures utilizing four different backbones named ResNet, ResNeXt, SE-ResNeXt, and DenseNet were used for automatic building extraction from high-resolution aerial images. Additionally, an Attention Residual U-Net approach was used and comparisons between all models were realized. There are several datasets used in building segmentation studies, such as Inria Aerial Image Labeling Dataset [91], Massachusetts Buildings Dataset [104], WHU Building Dataset [58], Aerial Imagery for Roof Segmentation (AIRS) Dataset [3], and International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam Dataset [132]. The models were trained and tested using two of these publicly available datasets, the Inria Aerial Image Labeling Dataset and the Massachusetts Buildings Dataset. The Inria Aerial Image Labeling Dataset includes images and masks from urban areas such as Chicago, IL, USA to rural towns such as Western Tyrol, Austria, and the Massachusetts Buildings Dataset includes images and masks from Boston, MA, USA. In order to increase and test the generalization abilities of the models, these two datasets, which include images from different cities with different building characteristics, were used within the scope of the study. Models using backbones were trained using weights trained on ImageNet, while models not using backbones were trained from scratch. Accuracy assessment was performed with Intersection over Union (IoU) score and F1 score and results were discussed. The aim of this study is to realize comparisons between the proposed Attention Residual U-Net approach, and the two most commonly used neural network architectures, which are U-Net and FPN, for building extraction from high-resolution images on two different datasets. It is also

within the scope of this study to make comparisons between backbones using residual and dense connections.

3.2 Study

3.2.1 Datasets

Two different, publicly available datasets for building extraction were used in this study. These datasets are; the Inria Aerial Image Labeling Dataset and the Massachusetts Buildings Dataset.

3.2.1.1 Inria aerial image labeling dataset

The first dataset selected to be used in the study is the Inria Aerial Image Labeling Dataset [91]. The dataset covers an area of 810 km^2, features orthorectified, RGB images with 0.3 m spatial resolution, and is divided into training and testing sets (each cover 405 km^2). Each image in the training set has a mask with two classes; the building class and the nonbuilding class. Training and test sets within this dataset were created using images from different cities. The regions differ from areas with a dense population (e.g., Austin and Chicago) to rural towns (e.g., Austrian Tyrol) [91]. Using images from different locations with different types of buildings improves the generalization power of the models that use this dataset.

Since the provided test set does not contain the corresponding masks to the images, only the images and masks from the training set were used in the study. The training set consists of 180 images and corresponding masks from 5 different cities, 36 images from each (Figure 3.1). These cities are; Austin (TX, USA), Chicago (IL, USA), Kitsap County (WA, USA), Vienna (Austria), and Western Tyrol (Austria). The images and masks are in the size of 5000×5000 pixels, each one covers an area of 1500 m×1500 m and they do not overlap with each other [91].

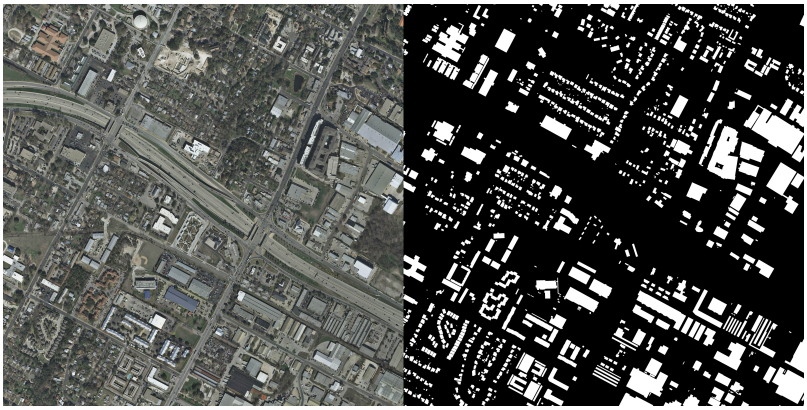


Figure 3.1 : Austin-7 RGB image and corresponding mask from the Inria Aerial Image Labeling Dataset.

To reduce the computation cost and increase the size of the dataset, the images and masks were cropped into 512×512 pixels size patches. Images with no buildings and a very low amount of buildings were eliminated from the set to deal with imbalance between classes, and 11210 images and corresponding masks were obtained as a result (Figure 3.2). Out of these images and masks, 1681 (15%) were selected to be used as test set, another 1681 (15%) were selected as validation set and the remaining 7848 (70%) were selected as training set. During this selection process, attention was paid to the distribution of urban and rural areas.

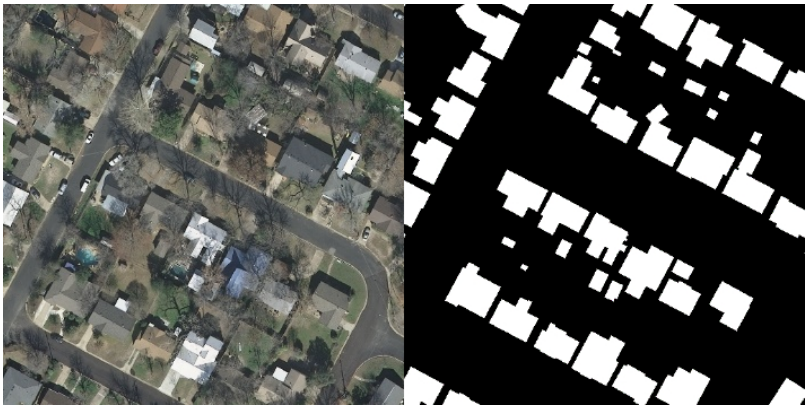


Figure 3.2 : Sample cropped image and corresponding mask from the prepared Inria dataset.

3.2.1.2 Massachusetts buildings dataset

The second dataset selected to be used in the study is the Massachusetts Buildings Dataset which was introduced by [104]. The dataset includes 151 RGB aerial images and binary masks of Boston, MA, USA. Each image is in 1500×1500 pixels size with 1 m spatial resolution and covers an area of 2.25 km^2. Images cover mostly urban and suburban areas with different sizes of buildings.

Similarly, to the Inria dataset, the images and masks were cropped into 512×512 pixels size patches and images with no building or low amount of building were eliminated. However, only 488 images could obtain after the process and this number is considered insufficient for a deep learning project. To increase the size, rotation, and brightness adjustment data augmentation methods have been applied. Images and corresponding masks were rotated 90 degrees clockwise and their brightness was increased by a 0.5 gamma rate. As a result, 1500 images and masks were obtained (Figure 3.3). Then, the dataset was split randomly into three subsets, training set (70%, 1050 images), validation set (15%, 225 images), and test set (15%, 225 images).

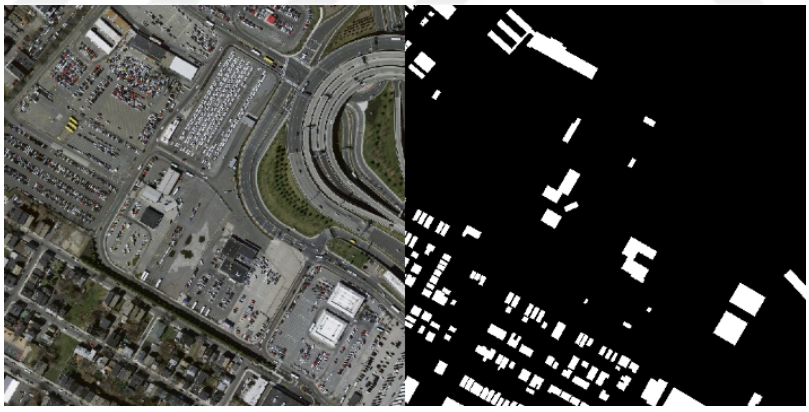


Figure 3.3 : Sample cropped image and corresponding mask from the prepared Massachusetts dataset.

3.2.2 Methodology

In this study, two CNN architectures utilizing four different backbones, and an Attention Residual U-Net approach were used. Used CNN architectures are U-Net and FPN, and used backbones are ResNet, ResNeXt, SE-ResNeXt, and DenseNet.

3.2.2.1 Architectures

The first architecture used in the study is U-Net. U-Net is a fully convolutional network (FCN) based CNN that was developed by [26]. It has become a de facto architecture for image segmentation studies due to its good representation ability and effective GPU use [133]. U-Net has a symmetrical structure of encoder and decoder parts that extract and concatenate multi-scale feature maps efficiently (Figure 3.4). With this structure, encoded features are propagated to decoder paths via skip connections and the bottleneck. The encoder path tends to extract high and low features and the decoder path follows the features learned via skip connections to reconstruct images into desired dimensions using upsampling or transposed convolutions [134]. The bottleneck is between encoder and decoder paths and includes two 3×3 convolutions. The encoder path consists of four steps; each includes two sets of 3×3 convolutions, Rectified Linear Unit (ReLU) activations, and 2×2 max-pooling. The decoder path also consists of four steps and each step includes a 2×2 convolutional transpose layer to upsample feature maps, skip connection to relevant feature map, and repeated 3×3 convolutions [26].

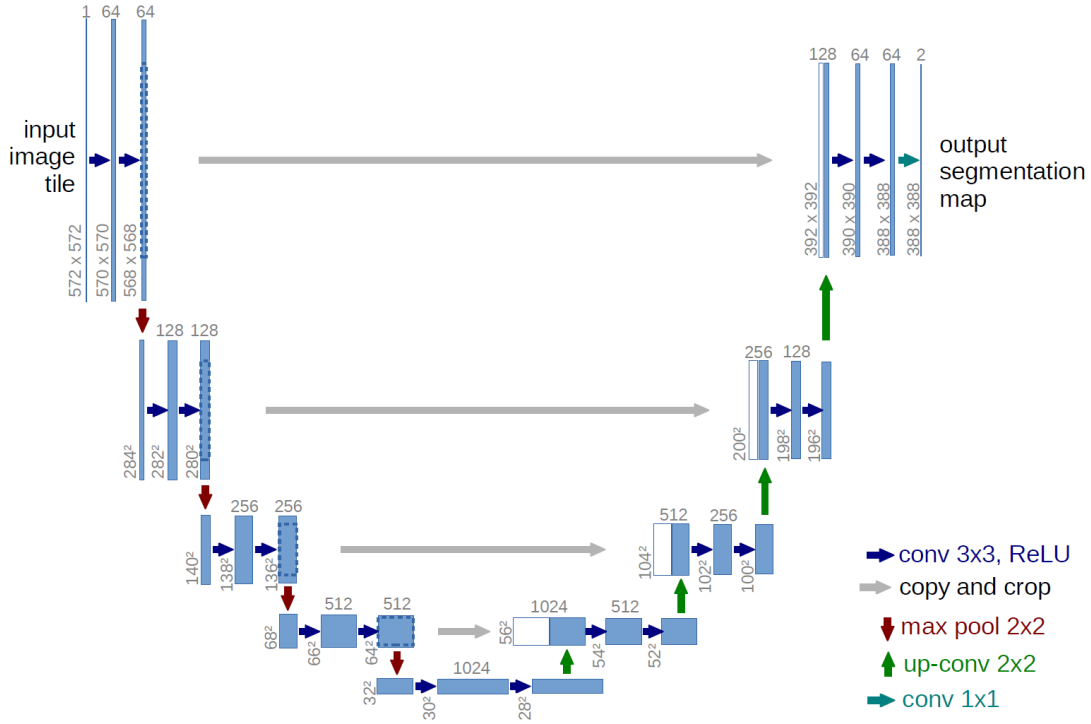


Figure 3.4 : U-Net architecture [26].

The second CNN architecture used in the study is FPN, introduced by [27]. They utilized the multi-scale and pyramidal structure of deep CNNs to generate feature pyramids. They aimed to use the pyramidal shape of CNN's feature hierarchy and generate a feature pyramid with strong semantics. To achieve this, a design was used by [27] to combine low resolution and semantically strong features with the high resolution and semantically weak features. As a result, a feature pyramid that has semantically rich and strong information was obtained.

There are two paths in FPN architecture; a bottom-up path and a top-down path (Figure 3.5). The bottom-up part is a conventional CNN that is used in image classification, and the top-down path generated pyramidal features which can be used for multi-scale object detection [135]. This path takes input from the last layer of the bottom-up path. Later on, this path up-samples the output from the preceding layer to the next layer and adds the output from the bottom-up path with the same dimension [135]. Each layer in this path creates pyramid features to separate predictors that detect objects with different scales.

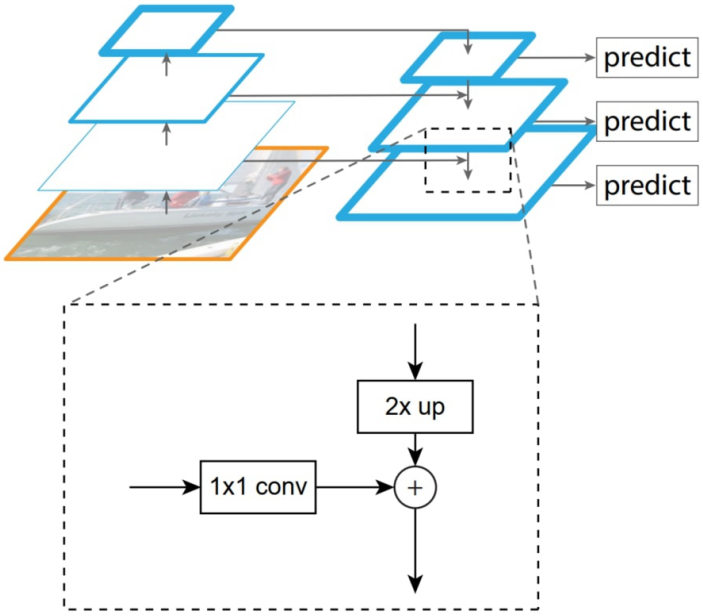


Figure 3.5 : FPN architecture and building block architecture [27].

3.2.2.2 Backbones

In the study, the mentioned architectures were used with four different backbones; ResNet, ResNeXt, SE-ResNeXt, and DenseNet. He et al. [22] introduced ResNet and residual blocks. ResNet combines shallow features to the deep features by mapping these shallow features to the deep outputs to improve the expression ability of the network [136]. Residual blocks in ResNet consist of two convolutional layers and a shortcut connection that passes the previous block's output directly to the next block. This direct connection allows a certain percentage of the output from the preceding layer to be preserved and solves the in-formation loss problem in traditional CNNs. The use of residual blocks also helps to solve the vanishing gradient problem caused by the increase in network depth [137] and the vanishing gradient problem existing in many CNNs [138]. In the study, ResNet-50 version is selected as one of the backbones (Table 3.1).

ResNeXt is a simple and highly modularized network architecture [139]. It takes the stacking and split transform merge idea from ResNet, and changes the single block with a parallel stack of blocks, which enhances the model accuracy without increasing the complexity [136]. Compare to the ResNet, ResNeXt introduces the "cardinality"

Table 3.1 : ResNet-50 architecture [22].

Layers	
Conv 1	7×7, 64, stride 2
	3×3 max-pooling, stride 2
Conv 2	$\begin{pmatrix} 1x1,64 \\ 3x3,64 \\ 1x1,256 \end{pmatrix}$ x3
Conv 3	$\begin{pmatrix} 1x1,128 \\ 3x3,128 \\ 1x1,512 \end{pmatrix}$ x4
Conv 4	$\begin{pmatrix} 1x1,256 \\ 3x3,256 \\ 1x1,1024 \end{pmatrix}$ x6
Conv 5	$\begin{pmatrix} 1x1,512 \\ 3x3,512 \\ 1x1,2048 \end{pmatrix}$ x3

concept which is the transformation set size. The width and depth of the network have a significant role in the representation power of a neural network. However, when the network goes wider or deeper, the vanishing gradient problem will occur. ResNeXt avoids this problem and searches for a way to achieve higher accuracy, which is increasing the cardinality [140]. Cardinality increase can produce better accuracy results than going wider or deeper [139]. In the study, ResNeXt-50 was selected to be used (Table 3.2).

Table 3.2 : ResNeXt-50 architecture [139].

Layers	
Conv 1	7×7, 64, stride 2
	3×3 max-pooling, stride 2
Conv 2	$\begin{pmatrix} 1x1,128 \\ 3x3,128 \\ 1x1,256 \end{pmatrix}$ x3
Conv 3	$\begin{pmatrix} 1x1,256 \\ 3x3,256 \\ 1x1,512 \end{pmatrix}$ x4
Conv 4	$\begin{pmatrix} 1x1,512 \\ 3x3,512 \\ 1x1,1024 \end{pmatrix}$ x6
Conv 5	$\begin{pmatrix} 1x1,1024 \\ 3x3,1024 \\ 1x1,2048 \end{pmatrix}$ x3

SE-ResNeXt is an architecture that integrates Squeeze and Excitation (SE) blocks [141] with ResNeXt architecture [139]. To acquire the importance of feature maps automatically, an SE block was added to the ResNeXt residual layer. [141] studied the effects of integrating SE blocks with ResNeXt and introduced additional building blocks into the base network. They observed significant performance enhancement induced by the SE blocks, without increasing the complexity significantly. The SE block models the correlation between feature channels and can obtain the importance of each of these channels automatically [136]. SE-ResNeXt-50 variant was selected as one of the backbones for this study (Table 3.3).

Table 3.3 : SE-ResNeXt-50 architecture [141].

Layers	
7×7, 64, stride 2	
3×3 max-pooling, stride 2	
Block 1	$\begin{pmatrix} 1x1conv, 128 \\ 3x3conv, 128 \\ 1x1conv, 256 \\ fc, [16, 256] \end{pmatrix}$ x3
Block 2	$\begin{pmatrix} 1x1conv, 256 \\ 3x3conv, 256 \\ 1x1conv, 512 \\ fc, [32, 512] \end{pmatrix}$ x4
Block 3	$\begin{pmatrix} 1x1conv, 512 \\ 3x3conv, 512 \\ 1x1conv, 1024 \\ fc, [64, 1024] \end{pmatrix}$ x6
Block 4	$\begin{pmatrix} 1x1conv, 1024 \\ 3x3conv, 1024 \\ 1x1conv, 2048 \\ fc, [128, 2048] \end{pmatrix}$ x3

The last backboned used in the study is DenseNet. DenseNet is a CNN that uses a modified ResNet convolutional block structure [64]. The difference between DenseNet and Resnet is that ResNet only uses one previous feature map, whereas DenseNet uses feature maps of all previous layers. The shared strategy between DenseNet and ResNet is that they both use feature maps of previous layers [142]. DenseNet has several advantages; it requires fewer parameters compared to traditional models, reduces the overfitting problem that occurs with small training sets, strengthens features propagation, and encourages feature reuse [143]. There are several DenseNets

according to their depths and DenseNet-121 is selected to be used in this study (Table 3.4).

Table 3.4 : DenseNet-121 architecture [64].

Layers	
7×7, 64, stride 2	
3×3 max-pooling, stride 2	
Dense Block 1	$\begin{pmatrix} 1x1conv \\ 3x3conv \end{pmatrix}$ x6
Transition Layer 1	$\begin{matrix} 1x1conv \\ 2x2avg-pooling, stride2 \end{matrix}$
Dense Block 2	$\begin{pmatrix} 1x1conv \\ 3x3conv \end{pmatrix}$ x12
Transition Layer 2	$\begin{matrix} 1x1conv \\ 2x2avg-pooling, stride2 \end{matrix}$
Dense Block 3	$\begin{pmatrix} 1x1conv \\ 3x3conv \end{pmatrix}$ x24
Transition Layer 3	$\begin{matrix} 1x1conv \\ 2x2avg-pooling, stride2 \end{matrix}$
Dense Block 4	$\begin{pmatrix} 1x1conv \\ 3x3conv \end{pmatrix}$ x16

3.2.2.3 Attention Residual U-Net approach

In this study, in addition to the models mentioned above, a Residual Attention U-Net approach is utilized, which combines standard U-Net architecture with residual connections and attention mechanism for further performance improvement. This combination makes the model focus on more relevant and informative features and produces better segmentation results [144].

U-Net architecture is built with repetitive convolutional blocks. But the model could be enhanced to benefit from residual blocks. To achieve this, traditional convolutional blocks in the architecture were replaced with residual blocks. Residual blocks, used in residual networks such as ResNet, were proposed to overcome the problems of deep CNNs, e.g. vanishing/exploding gradients [22]. To address this problem, ResNet architecture was introduced which adds the idea of "shortcut connections". Residual blocks add a shortcut connection to avoid vanishing gradient and speed up the convergence of the model by integrating rough local and global features [145]. In traditional convolutional layers, each layer feeds directly into the next layer. But in

residual blocks, each layer feeds into both the next layer, and additionally into the layers a few steps away. In this way, inputs can propagate faster through the residual shortcuts across layers.

In the utilized approach, instead of using the original residual block design from ResNet, a different residual connection variant was used. In addition to the original residual block design used in ResNet, different designs were proposed by He et al. [146] and implemented by Naranjo-Alcazar et al. [147]. In carried-out studies, several different residual connection variants were tried, and as a result, the design shown in Figure 3.6 was chosen to be used and integrated into the U-Net architecture to replace traditional convolutional blocks.

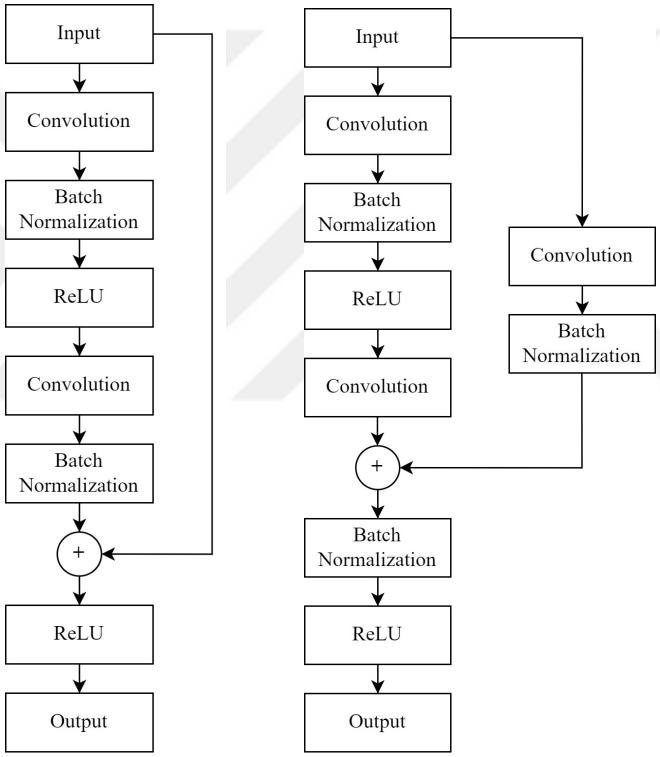


Figure 3.6 : (**a**) Residual block design used in ResNet, (**b**) Residual block design implemented in this study.

The attention mechanism is a beneficial tool used in CNNs for feature selection, which captures important local information and suppresses noisy features simultaneously [148]. Since U-Net architecture has to recover downsampling images caused by convolutions and pooling, the attention mechanism can bridge the information from

deep layers to the shallow layers and guide the upsampling learning [145]. Therefore, it could lead to higher accuracy in segmentation by enhancing the ability of the model to capture local responses [145]. Vaswani et al. [29] introduced an attention mechanism and applied it to machine translation. In time, many attention modules have emerged. Wang et al. [149] implemented an attention mechanism to computer vision and searched the effectiveness of nonlocal operations of images and videos. CNN models that utilize attention gates can be trained from scratch, and the attention gates can learn to focus on relevant features without the need for supervision [30]. Furthermore, they do not introduce additional notable computation costs, do not require calculating too many parameters, and provide better generalization [30].

There are two different attention mechanisms; hard attention and soft attention. Hard attention, for example, cropping and iterative region proposal, is generally non-differentiable and relies on reinforcement learning to update parameters [30]. Therefore, the model can either pay attention or not, and backpropagation cannot be used when it comes to hard attention. Soft attention refers to weighting different parts of the image differently. Relevant parts get large weights, and irrelevant parts get small weights. Contrarily to hard attention, soft attention is probabilistic and utilizes back-propagation [30]. During the training, weights also get trained, and that makes the model pay even more attention to relevant areas as the training goes.

At earlier layers of U-Net, there is more spatial information, and via skip connections, this information provides context to the later layers. This operation combines the spatial information from the encoder part to the decoder part to retain good spatial information. But the problem with this process is that it brings along the poor representations from the earlier layers. To overcome this problem, attention gates are applied at the skip connections to suppress activations at irrelevant regions of the image. In this study, the attention gate approach introduced by [30] was selected to be used (Figure 3.7). The proposed attention gate is integrated into the Residual U-Net approach to highlight relevant features transmitted via the skip connections.

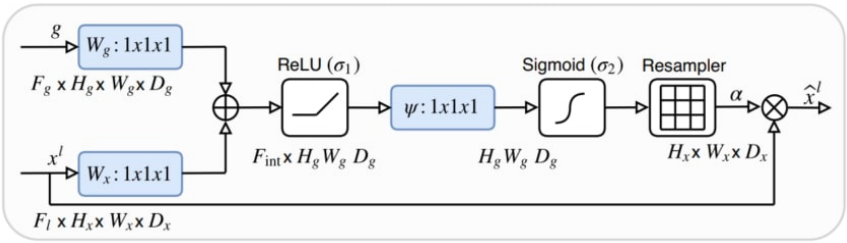


Figure 3.7 : Schematic of the attention gate used in the study [30].

3.2.3 Training and testing

In the study, U-Net and FPN architectures utilized with ResNet-50, ResNeXt-50, SE-ResNeXt-50, and DenseNet-121 backbones and the Attention Residual U-Net approach were used to train models. To train and test the models, prepared training and test sets were used. Models were built using Tensorflow framework and all models except Attention Residual U-Net are available in the GitHub repository [150]. All used backbones have weights trained on the 2012 ILSVRC ImageNet dataset and these weights were used in training the models that use these backbones. Models that use Attention Residual U-Net were trained from scratch.

As mentioned in the datasets section, for both datasets, 70% of the data were used for training and 15% were used for validation during training. To reduce the computation cost of training, prepared images and masks in size of 512×512 pixels were resized to 256×256 pixel size and these images were used as inputs in the study. Before training, all images were preprocessed with respect to the relevant backbones or when there is no backbone in use, were normalized. During the training, “Adam” algorithm was used as an optimizer [112] with 0.0001 initial learning rate. To control the decrease in the learning rate, “ReduceLROnPlateau” callback function from Tensorflow was used [151]. The function monitored the validation loss, and when the loss does not decrease for 5 epochs, the learning rate was reduced by a factor of 0.1. To calculate loss values, “Binary Cross-Entropy” loss was used. It is a loss function that is used on binary classification and segmentation problems (Equation 3.1) [113]. To determine how long the training of a model will last, another callback function from Tensorflow called "EarlyStopping" was used [152]. In this study, this function monitored the validation loss. The training was stopped when the loss value does not decrease for 10 epochs.

$$L(y,\hat{y}) = \frac{1}{N} + \sum_{i=0}^{N} (y * \log(\hat{y}_i) + (1-y) * \log(1-\hat{y}_i)) \tag{3.1}$$

In addition to these callbacks, another function called “ModelCheckpoint” was used to save the model when the validation loss decreased [153]. With the help of this function, the best models were saved and afterward used for testing and segmentation. The number of trainable parameters and the numbers of epochs for the models are presented in Table 3.5.

Table 3.5 : No. of trainable parameters and no. of training epochs of the models.

Architectures	No. of Trainable Parameters	Epoch of the Best Model (Inria)	Total No. of Epochs (Inria)	Epoch of the Best Model (Massachusetts)	Total No. of Epochs (Massachusetts)
U-Net ResNet-50	32,513,556	13	23	13	23
U-Net ResNeXt-50	31,992,980	12	22	19	29
U-Net SE-ResNeXt-50	34,523,969	7	17	14	24
U-Net DenseNet-121	12,059,345	17	27	23	33
FPN ResNet-50	26,863,940	9	19	15	25
FPN ResNeXt-50	26,343,364	10	20	17	27
FPN SE-ResNeXt-50	28,874,353	7	17	8	18
FPN DenseNet-121	9,825,793	15	25	24	34
Attention Residual U-Net	37,670,343	24	34	39	49

After the training and saving of the models, the best models were selected and applied to the test sets. Evaluation metrics were calculated and binary image segmentation was performed with a 0.5 threshold applied to the building class probabilities. To test and evaluate the models; intersection over union (IoU) score, recall, precision, and F1 score metrics were used.

Precision is the ratio of True Positive (TP) to total positive value (TP + False Positive (FP)) (Equation 3.2) [110]. Precision is generally used with another metric called Recall. Recall is the ratio of TP to all positive values (TP + False Negative (FN)) (Equation 2) [110]. F1 score is considered as a metric to measure the accuracy of a model in binary classification and segmentation [110]. It is the harmonic mean of recall and precision into a single value (Equation 3.3). F1 score varies between 0.0 and 1.0, where 1.0 is the best score and 0.0 is the worst [110]. IoU is an evaluation metric that gives the area of overlap between prediction and the mask (Equation 3.4) [154]. Along with these metrics, test accuracy and test loss values were also calculated using test sets.

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \tag{3.2}$$

$$F1Score = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3.4}$$

3.3 Results and Discussions

Evaluation metric results of the models on the Inria test set can be seen in Figure 3.8. According to these results, the Attention Residual U-Net model has the highest F1 score with 0.8154, IoU score with 0.7102, and test accuracy with 94.51%. For F1 score, FPN SE-ResNeXt-50 and FPN DenseNet-121 models follow the Attention Residual U-Net with 0.8131 and 0.8119. For IoU score, FPN DenseNet-121 and FPN SE-ResNeXt-50 models follow with 0.7076 and 0.7058, respectively. Lastly, for both

evaluation metrics, FPN ResNet-50 model has the lowest scores with a 0.7909 F1 score and a 0.6798 IoU score.

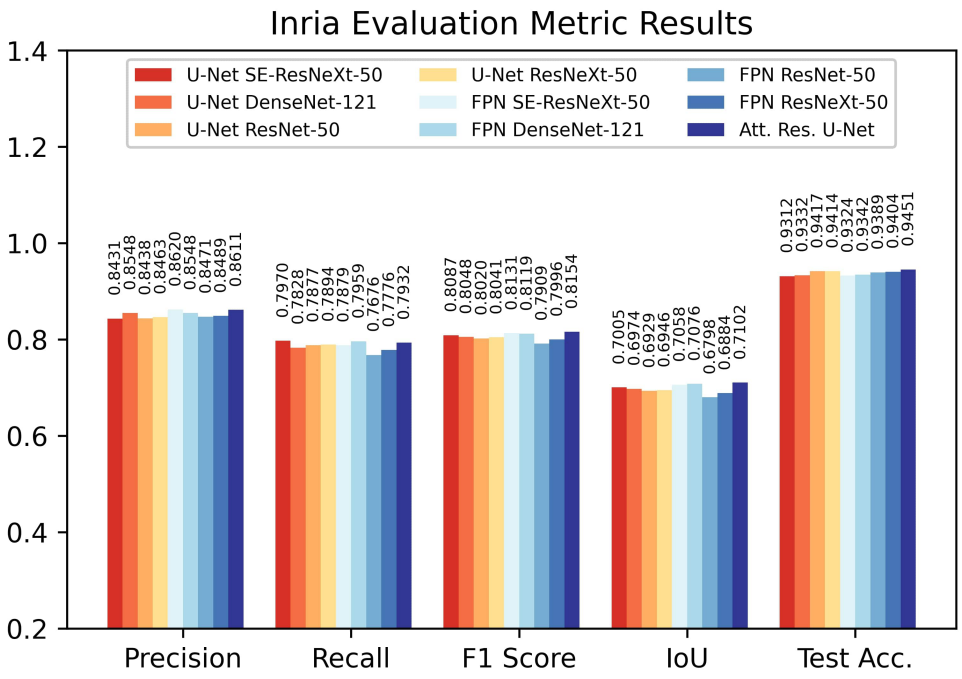


Figure 3.8 : Evaluation metric results of the models for the Inria test set.

Evaluation metric results of the models on the Massachusetts test set are can be seen in Figure 3.9. According to the results, FPN DenseNet-121 model has the highest F1 score with 0.7565 and IoU score with 0.6188, and the Attention Residual U-Net model has the highest test accuracy with 92.43%. For F1 score, FPN ResNeXt-50 and Attention Residual U-Net models follow the FPN DenseNet-121 with 0.7499 and 0.7458. For the IoU score, Attention Residual U-Net and FPN ResNeXt-50 models follow with 0.6142 and 0.6113, respectively. And finally, for both F1 score and IoU score metrics, U-Net ResNet-50 model has the lowest scores with a 0.7121 F1 score and a 0.5625 IoU score.

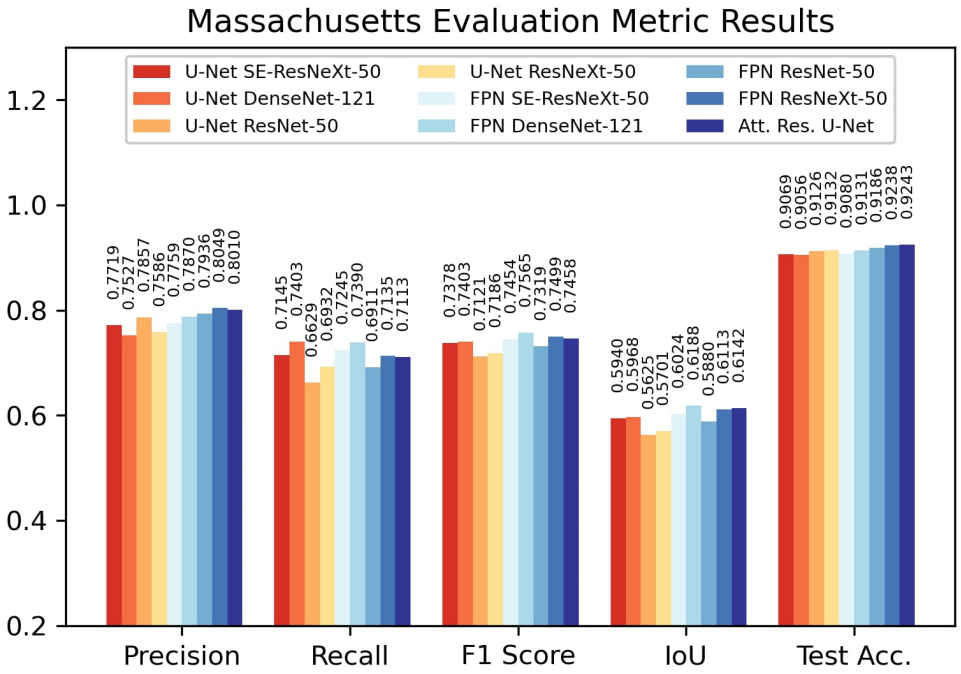


Figure 3.9 : Evaluation metric results of the models for the Massachusetts test set.

Test images, masks, and segmented image examples obtained using the models are shown in Figure 3.10 and Figure 3.11. As mentioned before, all predictions performed with a 0.5 threshold value applied to the class probabilities. In segmented images, buildings are represented by white pixels, and the background is represented by white pixels.

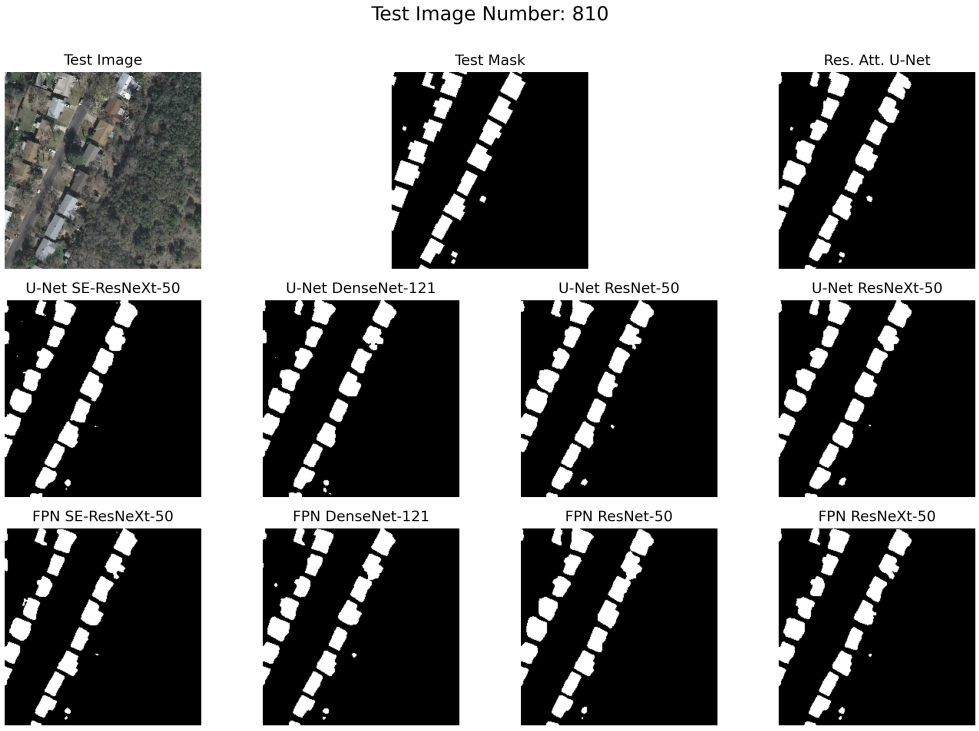


Figure 3.10 : Segmentation results for image no. 810 from the Inria test set.

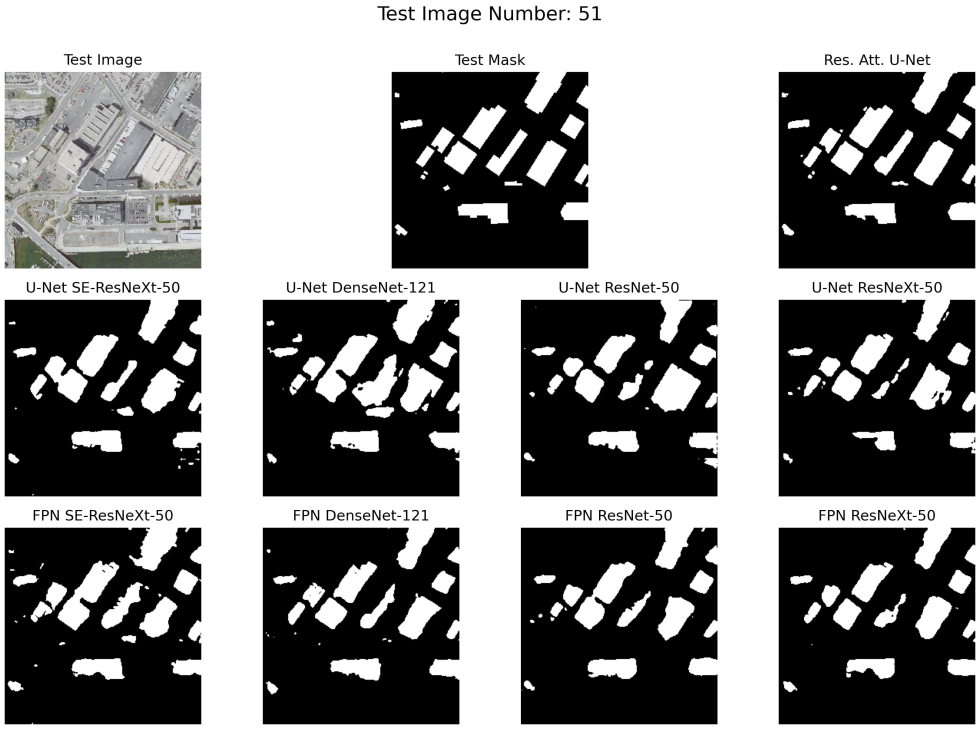


Figure 3.11 : Segmentation results for image no. 51 from the Massachusetts test set.

According to the results obtained on both datasets, it has been observed that the models using FPN architecture are more successful than the models using U-Net architecture, on the small Massachusetts dataset. Compared to U-Net, which is used widely for image segmentation applications in the literature, the results show that FPN could be a better choice when working with datasets of small size. When the used backbones are evaluated, it has been seen that SE-ResNeXt-50 and DenseNet-121 on the Inria dataset and DenseNet-121 and ResNeXt-50 on the Massachusetts dataset gave more successful results, respectively. DenseNet reduces the overfitting problem that occurs on small training sets. The success of the models using this architecture on the Massachusetts dataset, which is a smaller dataset, can be interpreted in this way.

On the other hand, on the Inria dataset, which is a larger dataset, the Attention Residual U-Net approach achieved the most successful results. Evaluations on the results show that the Attention Residual U-Net approach has a higher success than the other models used in the study when a sufficiently large dataset is used. When the models with the lowest success are examined, it has been seen that the FPN ResNet-50 model on the Inria dataset and the U-Net ResNet-50 model on the Massachusetts dataset have the lowest success according to the evaluation metrics. According to these results, it has been seen that ResNet-50 has the lowest success among the backbones used in the study, regardless of the dataset size.

3.4 Conclusions

In recent years, automatic building extraction has become an important research topic due to the developments in urban areas and the requirements in applications such as urban planning and change detection. Within the scope of this study, building extraction from high-resolution aerial images was performed using datasets generated from Inria Aerial Image Labeling Dataset and Massachusetts Buildings Dataset. A total of nine models were generated, and comparisons between U-Net and FPN architectures utilizing ResNet-50, ResNeXt-50, SE-ResNext-50, and DenseNet-121 backbones and the Attention Residual U-Net were realized.

On the smaller Massachusetts dataset, it has been seen that the FPN architecture achieves more successful results and the DenseNet backbone is the most successful

backbone among those used since it reduces the overfitting problem that can be encountered when working with small datasets. This study also includes an Attention Residual U-Net approach and comparisons of this model with other used models. Based on the evaluation metric results, this approach was found to outperform other models when a sufficiently large dataset is provided. The model using this approach achieved the most successful results on the Inria dataset, which is the larger dataset. Considering that the models using this architecture are trained from scratch without using pretrained weights, unlike other models, the approach can be said successful.

Future work on this topic will be on improving the performance of the proposed attention-residual approach, with the help of the vision transformers. Experiments will be made using different and novel approaches, and the results obtained by developing new and improved networks will be provided. Since the proposed and developed model will be trained with datasets containing different building structures from different regions, this model will also be useful for different study areas.

4. COMPARATIVE ANALYSIS OF DIFFERENT CNN MODELS FOR BUILDING SEGMENTATION FROM SATELLITE AND UAV IMAGES[3]

4.1 Introduction

In the last few years, particularly with the developments in urbanized areas, building detection and segmentation from aerial images have become challenging and important research topics within the remote sensing and computer vision community [115]. Assigning each pixel into a building or non-building class is challenging because there are big differences in sizes and shapes of buildings, strong similarities between buildings and non-buildings, and small between class and large within-class variance in pixel values in aerial images [155]. Building map generation from aerial images is a costly, laborious, and mostly a manual process [18]. However, tasks such as detection of illegal buildings [117], change detection [118], urban planning [116], and disaster management [11] all require precise building segmentation.

With continuous developments in current technologies, such as unmanned aerial vehicles (UAVs), sensors, and satellites, a large number of high-resolution aerial images have become more accessible and preferred data for building segmentation applications [18] which opened new paradigms for the community. With these high-resolution aerial images, the texture, structure, and spectral information of the buildings have become more refined [8], and these images facilitate the automatic building segmentation greatly [156]. But, this brings more challenges to arise from diversity in building characteristics to complex backgrounds [119,121]. Therefore, improving the accuracy and efficiency of building segmentation is still a challenging task and the focus of many studies [115].

Technologies used in building segmentation studies have also progressed over the years [115] and various techniques have been proposed. In earlier studies, most of these

[3]This chapter is based on the article: Sariturk, B., Kumbasar, D., and Seker, D. Z. (2022). Comparative Analysis of Different CNN Models for Building Segmentation from Satellite and UAV Images. Photogrammetric Engineering and Remote Sensing.

methods were traditional image processing methods that focus on pixels [59], lengths, edges [157], shapes [158], spectrum [159], textures [160] and shadows [59,161] as input features. Then, later on, conventional machine learning methods such as Support Vector Machines [123], K-Means [122], Random Forests [124], Conditional Random Fields (CRF) [125] and AdaBoost have been started to be used. However, the building structure complexities, similarities with other classes (such as roads and cement backgrounds), dependency on prior knowledge of the topic [119], and the need for manual feature selection lead to poor generalization capabilities and bias [117]. Thus, they are not well suited for building segmentation from high-resolution aerial images.

Recently, deep learning approaches, most popularly convolutional neural networks (CNN) have come forth as successful methods for computer vision applications [162], due to the availability of large datasets and increased computation power. CNN, which is a subclass of deep neural networks (DNN), is proposed in [76]. Differently, from previously mentioned conventional methods, deep learning methods can learn the features and semantic information directly from the input [119], perform detailed feature mapping via CNNs, and make classifications and segmentations through sequential convolutions with fully connected layers [8]. Therefore, CNN can be regarded as a method that combines feature extraction and segmentation/classification into a single model.

CNN models, which are mostly patch-based, have made notable achievements in building extraction tasks. However, they rely on small patches around targets to predict and ignore the relations between patches [8]. In addition, these patch-based models are time-consuming. To overcome these problems, fully convolutional networks (FCN) were proposed by Long et al. [24]. In FCNs, fully connected layers that are used in conventional CNNs are replaced with convolutional layers, and the upsampling layers are incorporated into the model. This allows the models to predict dense segmentation feature maps. This approach has become very popular and has achieved considerable performance on semantic segmentation tasks [33,128,163]. In addition, numerous variants based on FCN have been proposed such as SegNet [28], DeepLab [100], DeconvNet [87], and U-Net [26].

Despite FCN's success, there are some limitations. Firstly, fully convolutional layers in FCNs lead to a large number of parameters and computation complexity. Secondly, FCNs fail to provide detailed information about complex structures and lose information about small objects [25]. To overcome these disadvantages, Ronneberger et al. [26] proposed U-Net, which achieved significant success in medical image segmentation studies. U-Net, which adopts a U-shape architecture, consists of two parts. The first part is the encoder, which extracts features from the input image. The second part is the decoder, which performs the dense predictions. Nowadays, most CNNs used in segmentation studies from aerial images are based on FCNs.

In this study, twelve CNN models were generated, trained, evaluated and comparisons were realized. U-Net, Feature Pyramid Networks (FPN), and LinkNet architectures utilizing EfficientNet-B5 backbone, Original U-Net, SegNet, Fully Convolutional Networks (FCN), and six different Residual U-Net approaches that use six different residual block designs were used for building segmentation from aerial images. The main objectives of the study:

- Comparison of performance and generalization capabilities of different state-of-the-art CNN models for building segmentation.
- Comparison of different residual block designs inspired by the ResNet architecture, which has been popular in recent years and achieved successful results, with each other and other used approaches.
- Evaluating the usability of models trained with publicly available image data covering urban and rural areas, for archaeological sites, and which approaches might be more appropriate.

These models were trained using the training set prepared using the images from the Inria Aerial Image Labeling Dataset. To test the trained models, three different test sets were prepared, and evaluations were realized. These test sets were prepared using the Inria Aerial Image Labeling Dataset, the Massachusetts Buildings Dataset, and our Syedra Archaeological Site Dataset. This study aims to realize comparisons between used approaches for building segmentation tasks and test the usability of the models

trained with the data from urban and rural areas for the detection of buildings and building remains in archaeological sites.

4.2 Methodology

In this study, U-Net, FPN, and LinkNet were used as base architectures to generate the models. In addition to the original U-Net, SegNet and FCN models, EfficientNet-B5 is used as a backbone for all three base architectures. Moreover, six different residual block designs [147] were applied to the original U-Net architecture. Consequently, a total of twelve CNN models were obtained.

4.2.1 Architectures

4.2.1.1 U-Net

U-Net is a CNN that was developed originally for biomedical image segmentation [26]. Due to its effective GPU usage and good representation ability, It has become a popular architecture for image segmentation tasks [133]. U-Net is an architecture with a U-shaped, symmetrical structure. It consists of encoder, decoder, and bottleneck parts that can extract and then concatenate feature maps (Figure 4.1). The encoder path extracts feature maps and these feature maps propagate via skip connections to the decoder path. Afterward, the decoder path uses the learned features to reconstruct the images into wanted dimensions using upsampling [134]. The bottleneck, which connects encoder and decoder paths, includes two 3×3 convolutions. The encoder consists of four blocks; each one includes two sets of 3×3 convolutions, Rectified Linear Unit (ReLU) activations, and 2×2 max-pooling. The decoder also consists of four blocks and each block includes a 2×2 transposed convolution layer, skip connection to the related feature map, and two 3×3 convolutions [26].

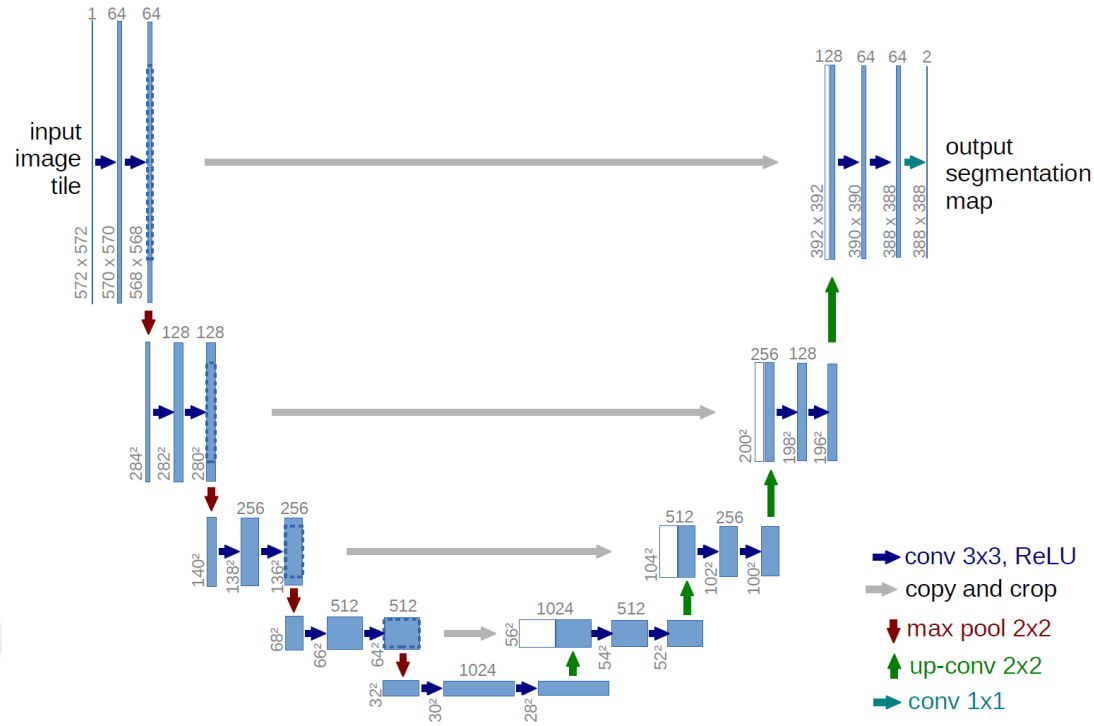


Figure 4.1 : U-Net architecture [26].

4.2.1.2 Feature Pyramid Networks (FPN)

Lin et al. introduced FPN in 2017 [27]. The authors utilized the inherent pyramidal and multi-scale hierarchy of deep CNNs to construct feature pyramids. The aim was to naturally utilize the pyramidal shape of the feature hierarchy of CNN and at the same time create a semantically strong feature pyramid [27]. To achieve this, Lin et al. used an architectural design that combines semantically strong and low-resolution features with semantically weak and high-resolution features using lateral connections and top-down path [27]. Consequently, a feature pyramid that has strong and rich semantic information was obtained.

FPN consists of two paths; a top-down and a bottom-up path (Figure 4.2). The bottom-up path is a traditional CNN used in image segmentation and classification tasks. The top-down path generates the pyramidal features to be used for object detection. This path takes the last layer of the bottom-up path, upsample the output of the previous layer, and adds the output of the bottom-up path with the same dimension.

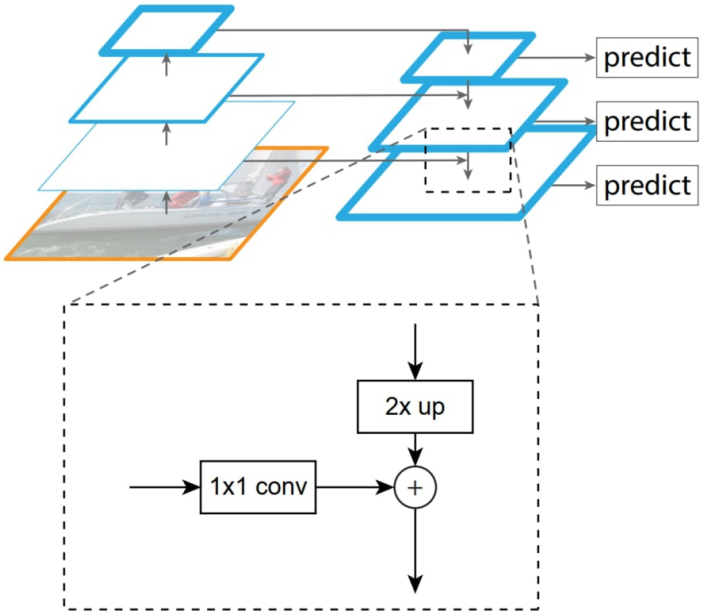
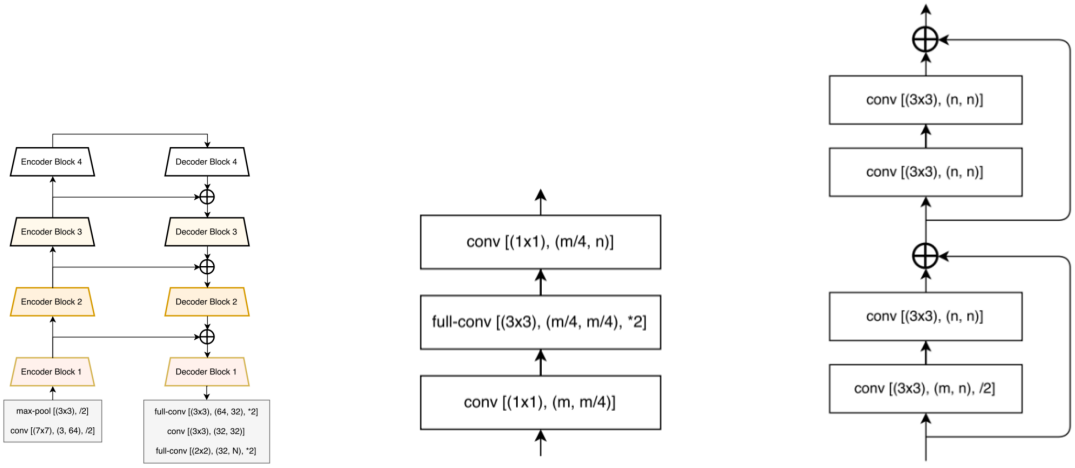


Figure 4.2 : FPN architecture [27].

4.2.1.3 LinkNet

In 2017, Chaurasia and Culurciello proposed a deep neural network called LinkNet, which allows the model to learn without any important increase in the number of parameters [164]. Unlike many CNN architectures, the novelty of LinkNet lies in the way that they link every encoder with a decoder (Figure 4.3). Due to the usage of multiple downsampling in the encoder, some spatial information is lost, and therefore it is quite hard to recover the information by using just these downsampled outputs. In LinkNet, inputs of each encoder are bypassed to the output of their corresponding decoders [164]. With this operation, lost spatial information is recovered and can be used by decoders and in upsampling operations. Additionally, since the encoder and decoder are sharing knowledge learned by the encoder at each layer, decoders can use fewer parameters. This results in a more efficient network design compared to other existing networks [164].



(a) (b) (c)

Figure 4.3 : a) LinkNet architecture, **b**) encoder block, **c**) decoder block [164].

4.2.1.4 SegNet

Badrinarayanan et al. presented a deep fully convolutional network for semantic segmentation, called SegNet [28]. The proposed architecture consists of an encoder, a decoder, and a classification layer (Figure 4.4). The topology of the encoder is identical to the first 13 convolutional layers used in the VGG16 network [28]. Every convolution layer is followed by a Batch Normalization and a ReLU. There is a corresponding decoder layer for each encoder layer; hence the network has 13 decoder layers. The decoder maps the low-resolution feature maps coming from the encoder to input resolution for pixel-wise classification. The final decoder output is a classifier to produce final class probabilities.

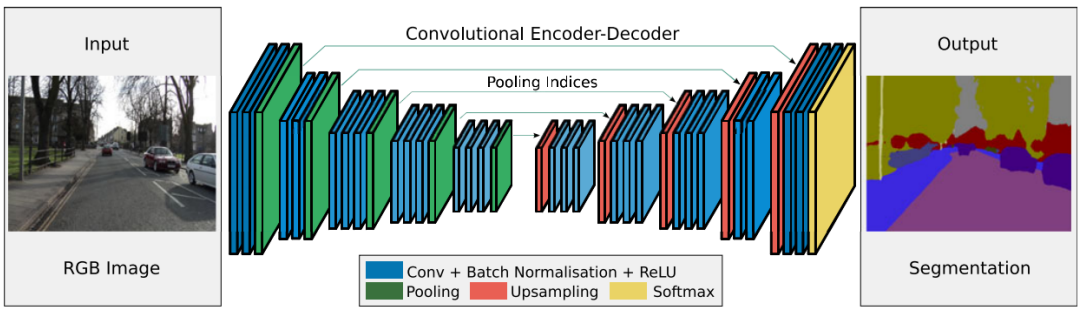


Figure 4.4 : SegNet architecture [28].

4.2.1.5 Fully Convolutional Networks (FCN)

In 2015, Long et al. [24] adapted modern networks used for classification (VGGNet, AlexNet, GoogleNet) into fully convolutional networks (FCN) and transferred their representations to image segmentation tasks by fine-tuning. Then, they defined an architecture (Figure 4.5) that combines the deep, semantic information with the shallow, appearance information by using skip connections to produce detailed and accurate segmentations [24].

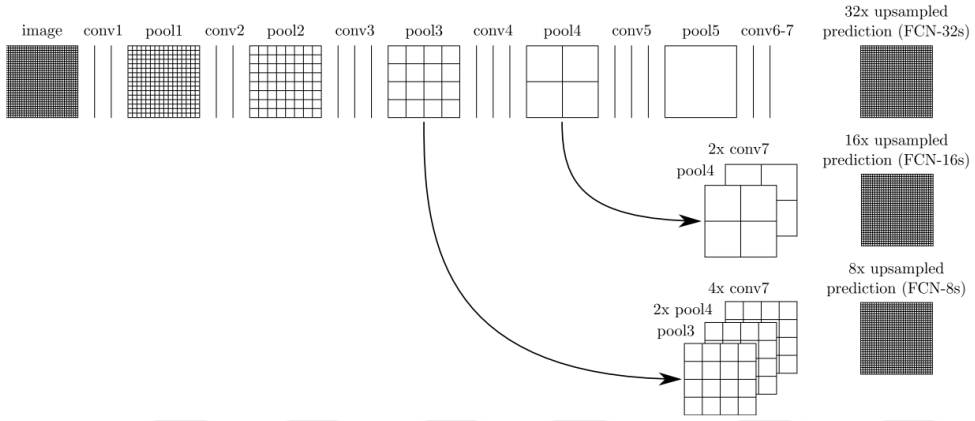


Figure 4.5 : FCN architecture [24].

4.2.2 Backbone (EfficientNet)

In the study, EfficientNet is used as a backbone for architectures U-Net, FPN, and LinkNet. Tan and Le [165] studied the model scaling and found that balancing the width, depth, and resolution of a network can lead to better performance. They proposed a scaling method that scales all dimensions uniformly, using a compound coefficient (Figure 4.6). As a result of these studies, they designed a baseline network and scaled the network up to obtain a model family called EfficientNet. The EfficientNet family consists of 8 models, named from B0 to B7 [165]. As the number in the model name increases, it is stated that the model has more parameters. For this study, the EfficientNet-B5 variant is selected as the backbone.

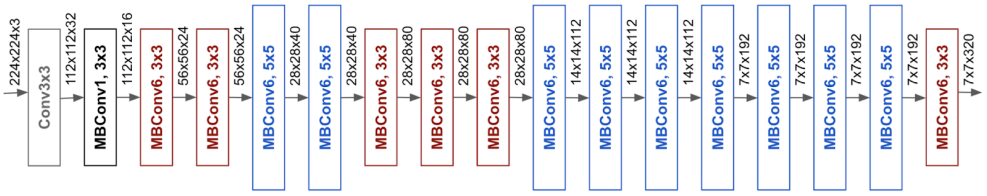


Figure 4.6 : EfficientNet-B0 architecture [165].

4.2.3 Residual blocks and Residual U-Net approaches

Residual blocks were originally proposed via ResNet architecture [22], to overcome problems of deep CNNs [22]. When the number of layers in a network continues to increase, the "vanishing gradient problem" occurs. During the training of deep models, gradients often get very small or zero, and eventually leave the weights nearly unchanged. Using residual connections, information can be directly passed many layers down [22]. In this study, additionally to the residual block (RB1) used in ResNet, four different block designs (RB2-RB5) by He et al. [146], and one (RB6) introduced by Dai et al. [166] were used to generate Residual U-Net approaches (Figure 4.7). Conventional convolution layers and convolution blocks are replaced with residually connected layers and residual blocks. As a result, six different Residual U-Net approaches have been generated.

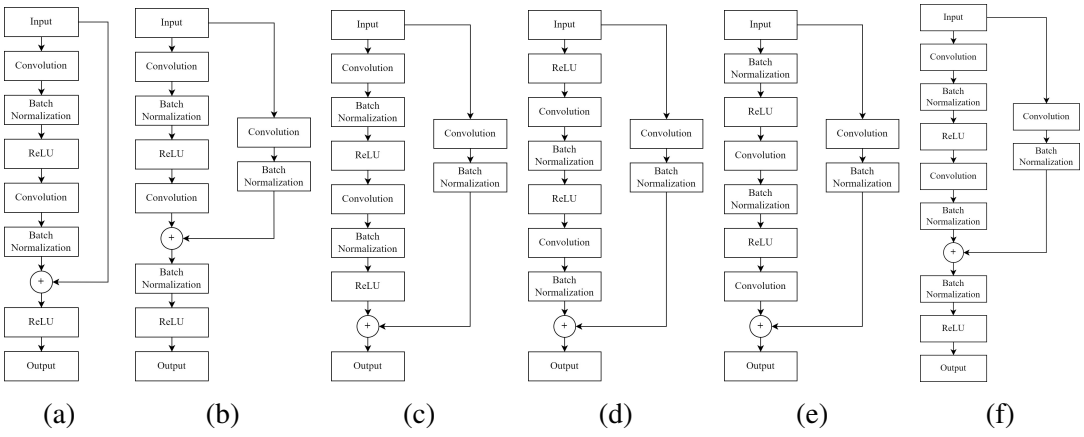


Figure 4.7 : Residual block designs were used in the study.

4.3 Datasets

In the study, a total of three different datasets were used. Inria Aerial Image Labeling Dataset was selected to train and test the models, and two additional datasets were also used to test the trained models. These additional test sets were prepared using Massachusetts Buildings Dataset and our Syedra Archaeological Site Dataset.

4.3.1 Inria aerial image labeling dataset

The Inria Aerial Image Labeling Dataset is a benchmark dataset provided by Inria for use in building segmentation studies [91]. This dataset features RGB images with 30 cm spatial resolution. The dataset is divided into two sets as training and testing [102]. Every image in the training set has a corresponding binary mask, with building and not building classes. Since the images in the provided test set don't have masks, only images and corresponding masks from the training set were used. The regions where the images are from differ from densely populated areas (e.g. Chicago and Vienna) to rural areas (e.g. Austrian Tyrol). The usage of this diverse dataset provides a high generalization ability. The training set provided by Inria consists a total of 180 images and their masks from 5 cities with different characteristics (Chicago, IL, USA; Austin, TX, USA; Kitsap County, WA, USA; Vienna and Western Tyrol, Austria). These images are in the size of 5000×5000 pixels and each covers an area of 1500×1500 m^2.

In order to augment the size of the dataset and cut the computational cost, images and masks were cropped into smaller patches in size of 512×512 pixels (Figure 4.8). Images and masks with a low amount of buildings and no buildings were eliminated from the set, and as a result, 11210 images and corresponding masks were obtained. From this dataset, 7848 images and masks (70%) were selected to be used as the training set, 1681 images and masks (15%) were selected to be the validation set, and the remaining 1681 images and masks (15%) were selected as the test set. In these selections, attention has been given to the homogeneous distribution of the rural and urban areas.

Figure 4.8 : Sample 512×512 image and mask from Inria dataset.

4.3.2 Massachusetts buildings dataset

One of two datasets used to generate additional test sets is Massachusetts Buildings Dataset. This dataset was introduced by Volodymyr Mnih in 2013 [104], and includes 151 RGB images and corresponding masks from Boston, MA, USA. Images in the dataset is the size of 5000×5000 pixels and covers an area of 2.25 km^2. Selected images and masks from the set were cropped into 512×512 pixel size patches (Figure 4.9), similar to the preparation of the Inria set, and the images with little to no amount of building classes were eliminated. As a result of all the steps, a test set that includes 225 images and corresponding masks was obtained.

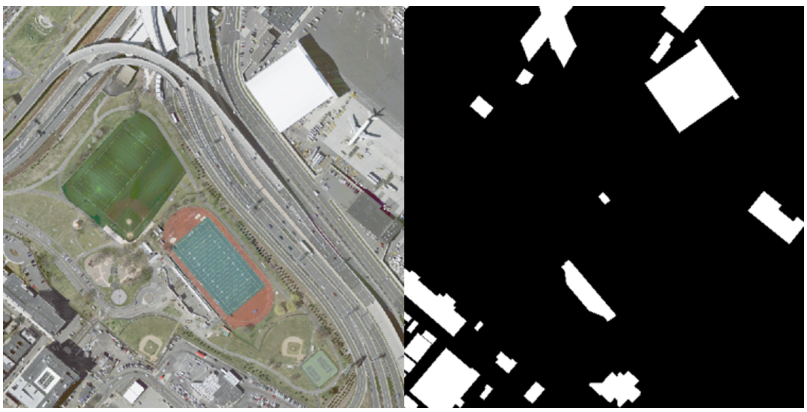


Figure 4.9 : Sample 512×512 image and mask from Massachusetts dataset.

4.3.3 Syedra archaeological site dataset

Syedra Archaeological Site Dataset was acquired over the Syedra Ancient City in November 2020. The ancient city of Syedra is located approximately 20 km southeast of Alanya, Antalya, Turkey. While the central structures of Syedra, which are spread over a very large area, are concentrated on the summit of Asar Tepe (approximately 400 m above sea level) and its south-facing high parts, there are also city-related building groups in the harbor area on the coastline and other smaller hills around Asar Tepe [167]. In the study, the images were collected from the hilly area (approximately 10 ha), which is the old city center, where the buildings that have been preserved until today are located.

The images have been acquired using a DJI Mavic 2 Pro UAV with Hasselblad L1D-20c, 20 MP aerial camera. They were obtained with two flights with fixed take-off heights of 46 m and 58 m in the south region of the city. The collected images were cropped into 256×256 pixel size patches (Figure 4.10), the images with little to no amount of building classes were eliminated, and as a result, a total of 487 images and their corresponding masks were obtained to be used as the second test set.

Figure 4.10 : Sample 256×256 image and mask from Syedra dataset.

4.4 Training and Testing

In the study, U-Net, FPN, and LinkNet architectures with EfficientNet-B5 backbone, original U-Net architecture, SegNet, FCN, and six Residual U-Net approaches with

different residual block designs were utilized to generate models and comparisons between them were realized. To train these models, the prepared Inria dataset was used. To test the trained models, in addition to the Inria dataset, Massachusetts and Syedra datasets were used. Tensorflow framework was used to build models, and all models (except Residual U-Net approaches) are available in the "Segmentation Models" repository on GitHub [150].

As mentioned earlier, 70% of the Inria dataset was used for training, and 15% was used for validation. To cut the computational cost, images and their corresponding masks were resized from 512×512 pixels to 256×256 pixels and normalized. "Adam" optimizer [112] with a 0.0001 initial learning rate was used during the training. To incrementally reduce the learning rate, with respect to training performance, the "ReduceLROnPlateau" callback was used. This callback monitored the validation loss during training, and when the value does not decrease for 5 consecutive epochs, the learning rate value was reduced by a 0.1 factor. All models are trained with a batch size of 2, and to calculate loss values, two loss functions were used together. These are Binary Cross-entropy Loss and Dice Loss. Previous studies on building detection have shown that using joint loss functions is effective [168]. Binary Cross-entropy is a loss function that is used in studies involving binary decisions (Equation 4.1) [113]. The dice coefficient is a metric widely used in computer vision studies to calculate similarities between two images [169]. In 2016, it has been adapted as Dice Loss (Equation 4.2) [170].

$$L(y,\hat{y}) = \frac{1}{N} + \sum_{i=0}^{N} (y * \log(\hat{y}_i) + (1-y) * \log(1-\hat{y}_i)) \tag{4.1}$$

$$DL(y,\hat{y}) = 1 - \frac{2y\hat{p}+1}{y+\hat{p}+1} \tag{4.2}$$

To determine the duration of the training, another callback called "EarlyStopping" was used. During the training, this callback monitored the validation loss and if the value doesn't decrease for 10 consecutive epochs, the training stopped. In addition to mentioned callbacks, the "ModelCheckpoint" callback was also used to save the model whenever the validation loss value decreased. With this callback function, the

best models were saved during the training, and later on, were used for testing and segmentation. The training epochs for all models, along with the number of trainable parameters, were shown in Table 4.1.

Table 4.1 : The total number of parameters of the architectures.

Architecture	No. of Trainable Parameters	Epoch of the Best Model	Total No. of Epochs
U-Net	31M	12	22
SegNet	37.3M	20	30
FCN	134.3M	49	50
U-Net EfficientNet-B5	37.4M	42	52
FPN EfficientNet-B5	31.8M	21	31
LinkNet EfficientNet-B5	33.7M	12	22
Residual-1 U-Net	31.7M	15	25
Residual-2 U-Net	31.7M	27	37
Residual-3 U-Net	31.7M	18	28
Residual-4 U-Net	31.7M	17	27
Residual-5 U-Net	31.7M	23	33
Residual-6 U-Net	31.7M	18	28

After training, the best models for each architecture were picked and evaluated on the test sets. To evaluate the models, several metrics were used, and image segmentation was performed with 0.5 threshold applied to the class probabilities. The metrics used for the evaluation process are Precision, Recall, F1 Score, and Intersection over Union (IoU). The precision metric is the ratio of True Positive (TP) values to total positive (TP + False Positive (FP)) (Equation 4.3) [110]. The recall is the ratio of TP values to all real positive values (TP + False Negative (FN)) (Equation 4.3) [110]. F1 Score is a metric used to measure the general performance of a model in segmentation and classification studies, and it is the harmonic mean of Precision and Recall (Equation 4.4). IoU gives the ratio of the overlapped area between the mask and the prediction, of the union of these areas (Equation 4.5) [154]. Besides these metrics, test loss and test accuracy values were calculated for all models, using each test set.

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \tag{4.3}$$

$$F1Score = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{4.5}$$

4.5 Results and Discussions

A summary of evaluation metric results is shown in Table 4.2. It has been observed that on the Inria test set, the Residual-2 U-Net model has the highest IoU score, F1 score, and test accuracy with 0.824, 0.722, and 94.7% respectively. For the F1 score, Residual-3 U-Net and U-Net follow Residual-2 U-Net with 0.824 and 0.819. For the IoU score, U-Net and Residual-6 U-Net follow with 0.712. On the Syedra test set, the LinkNet model with the EfficientNet-B5 backbone has the highest F1 and IoU scores with 0.336 and 0.246. Residual-4 U-Net and Residual-3 U-Net follow with 0.252 and 0.249 F1 scores, and 0.172 and 0.178 IoU scores, respectively.

On the Massachusetts test set, the Residual-4 U-Net model has the highest F1 and IoU scores with 0.394 and 0.259. Residual-1 U-Net and Residual-3 U-Net models follow with 0.368 and 0.336 F1 scores and 0.243 and 0.220 IoU scores. Residual-6 U-Net model has the highest test accuracy on both Massachusetts and Syedra test sets with 83.1% and 84% respectively.

Table 4.2 : Evaluation results of the models.

Architectures	Inria			Syedra			Massachusetts		
	F1	IoU	Test Acc. (%)	F1	IoU	Test Acc. (%)	F1	IoU	Test Acc. (%)
U-Net	0.819	0.712	94.1	0.188	0.134	82.6	0.312	0.200	79.3
SegNet	0.674	0.558	92.1	0.073	0.053	83.5	0.165	0.101	81.5
FCN	0.693	0.567	91.6	0.187	0.136	83.7	0.254	0.158	73.9
U-Net EfficientNet-B5	0.701	0.568	89.9	0.179	0.115	77.1	0.275	0.173	80.3
FPN EfficientNet-B5	0.661	0.527	89.8	0.140	0.086	80.0	0.299	0.187	79.6
LinkNet EfficientNet-B5	0.690	0.552	89.8	**0.336**	**0.246**	82.6	0.305	0.190	78.1
Residual-1 U-Net	0.811	0.703	94.1	0.218	0.151	83.7	0.368	0.243	83.0
Residual-2 U-Net	**0.824**	**0.722**	**94.7**	0.199	0.140	83.8	0.311	0.202	82.3
Residual-3 U-Net	0.824	0.705	94.2	0.249	0.178	83.7	0.336	0.220	82.4
Residual-4 U-Net	0.803	0.693	93.9	0.252	0.172	78.9	**0.394**	**0.259**	82.0
Residual-5 U-Net	0.797	0.686	93.	0.141	0.096	83.2	0.328	0.211	81.7
Residual-6 U-Net	0.817	0.712	94.4	0.151	0.107	**84.0**	0.313	0.204	**83.1**

Test images, their corresponding masks, and example segmentation images are shown in Figures 4.11, 4.12, and 4.13. Predictions were performed with 0.5 threshold value. In the segmented images presented, white pixels are building classes and black pixels are the background.

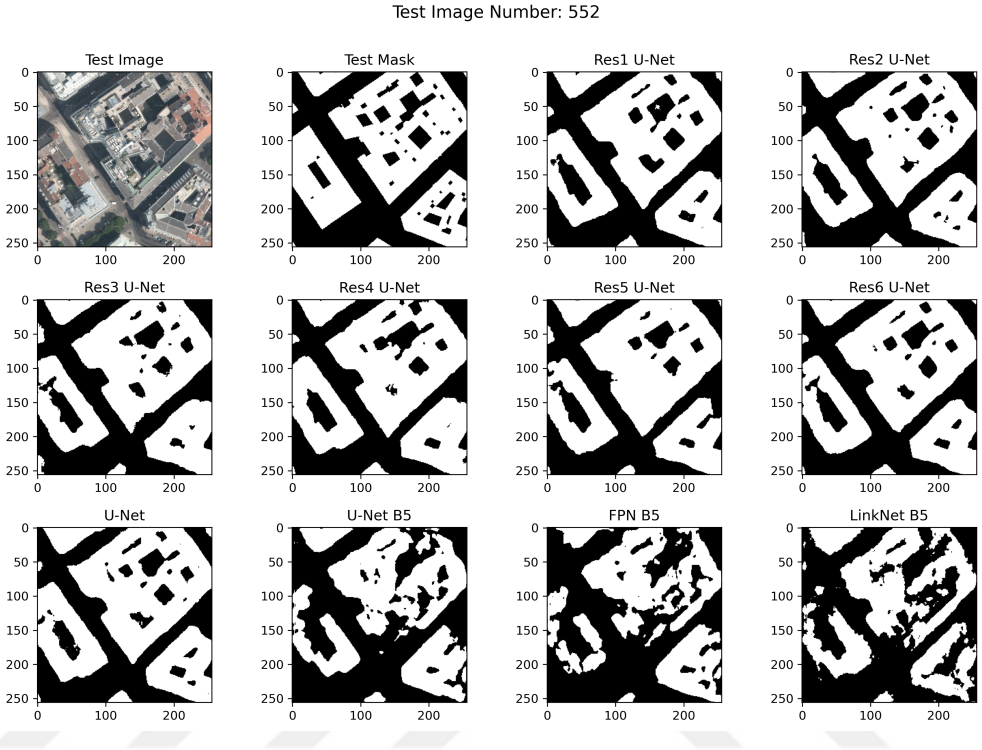


Figure 4.11 : Segmentation results of image no. 552 from the Inria test set.

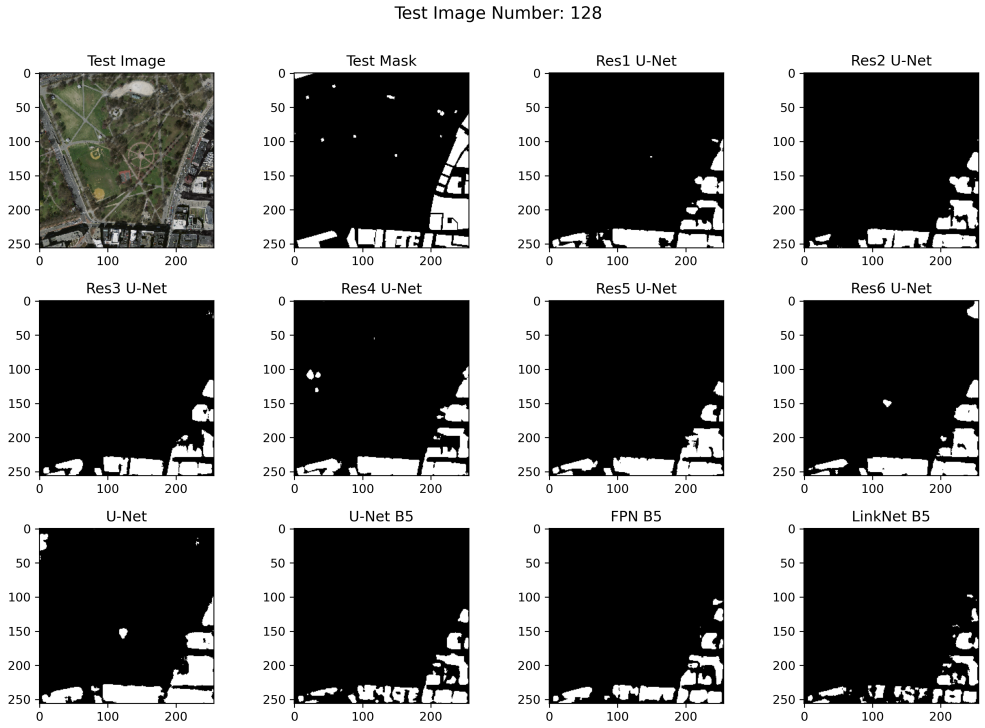


Figure 4.12 : Segmentation results of image no. 128 from the Massachusetts test set.

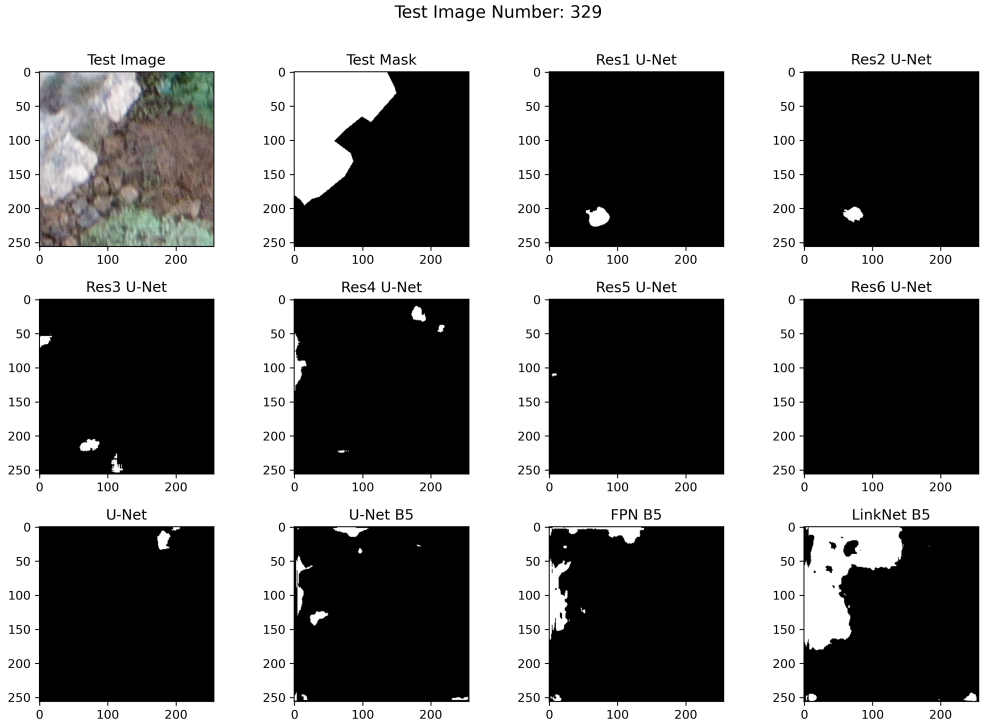


Figure 4.13 : Segmentation results of image no. 329 from the Syedra test set.

According to the results obtained on prepared test sets, the Residual-3 U-Net architecture, using the RB-3 residual block design, is in the top three for all test

sets according to the evaluations. Therefore, it can be stated that it is the most inclusive model among all used approaches. When evaluated in general, it has been observed that models using residual connections are more successful than models using conventional convolution structures. On all three test sets, at least two of the top three most successful models, according to the evaluation metric results, use architectures with residual connections. This can be interpreted as residual connections being more successful due to their ability to carry spatial information from earlier layers. The model uses LinkNet architecture with EfficientNet-B5 backbone achieved the most successful results on the Syedra test set, which is the test set that has the most different feature characteristics, which are archaeological buildings and building remains. Since this architecture achieves the most promising results compared to other models, it could be said that this approach is best suited for different feature characteristics among all models, and could be a better choice to be used for future studies on archaeological sites. On the other hand, FPN architecture with EfficientNet-B5 backbone and SegNet were the least successful models for all three test sets. According to the evaluation metric results, SegNet was in the bottom three for all three test sets and provided the lowest F1 and IoU scores for Syedra and Massachusetts test sets. Considering these results on test sets that have different types of feature characteristics than the training set, it could be said that SegNet is the approach that has the least generalization ability among the others used in the study. Therefore, it can be stated that the architectural structures of FPN and SegNet are the least suitable model approaches for building detection and segmentation study compared to other approaches used.

4.6 Conclusions

Over the years, automatic building segmentation from aerial images has become a challenging and important research topic with the rapid developments in especially urban areas. Within the study, building segmentation from satellite and UAV images was performed using the Inria Aerial Image Labeling Dataset as the training set, and the Inria Aerial Image Labeling Dataset, the Massachusetts Buildings Dataset, and the Syedra Archaeological Site Dataset as test sets. A total of twelve models were generated and comparisons were realized.

Results show that the Residual-2 U-Net model, which uses a different residual block from ResNet, performs best in all evaluation metrics on the Inria test set. According to the evaluation metrics, the LinkNet EfficientNet-B5 model on the Syedra Archaeological Site Test Set and the Residual-4 U-Net model on the Massachusetts Test Set showed the best results. Since the features in Syedra Archaeological Site Test Set have different characteristics from the Inria Training Set, it may be said that the LinkNet EfficientNet-B5 model has high generalization ability. Nevertheless, the features from Massachusetts Test Set are similar to the Inria Training set but from a different kind of sensor, and because of that it can be mentioned that the Residual-4 U-Net model has a generalization ability across similar features and different sensors.

In further studies, performance improvements will be realized on the proposed residual models, with the implementation of approaches such as attention gates and transformers. Additionally, studies to extract buildings and building remains on the archaeological sites will be continued, using Syedra Archaeological Site Dataset and different CNN approaches.

5. A RESIDUAL-INCEPTION U-NET (RIU-NET) APPROACH AND COMPARISONS WITH U-SHAPED CNN AND TRANSFORMER MODELS FOR BUILDING SEGMENTATION FROM HIGH-RESOLUTION SATELLITE IMAGES[4]

5.1 Introduction

With the developments in satellite and remote sensing technologies, building segmentation and generating building maps have become important research topics in recent years [1]. Due to these developments, high-resolution satellite images have become even more accessible and convenient data sources [4]. As an important feature in the urban environment, the mapping of buildings has significant importance for different applications such as urban mapping, population estimation, land cover/land use analysis, cadastral and topographic map production, change detection, and disaster management [7]–[10]. Nonetheless, obtaining reliable and accurate building maps from high-resolution satellite images is still challenging due to various reasons such as complex backgrounds [12,13], similarities between the background and the buildings [117], noise in data [8], heterogeneity in data structures [12,13], diversity in roof types [14], and characteristics (size, shape, color, etc.) of buildings [119], and other topological difficulties. If the problems encountered according to the methods used are briefly mentioned, the conventional methods that were used in the early studies generally use manually extracted features. In these models, the extraction process requires prior knowledge, which leads to a poor model generalization ability and is costly and time-consuming [12]. Deep learning methods that later replaced these methods in the following years have their own concerns. For example, convolutional neural networks (CNN) perform regional divisions, use expensive fully connected layers, and they lose accuracy and details due to stacked convolution layers [44]. Fully convolutional networks (FCN) use fixed-size convolutions that result in a local receptive field; thus, they lack the ability to model in the global context [16].

[4]This chapter is based on the article: Sariturk, B., and Seker, D. Z. (2022). A Residual-Inception U-Net (RIU-Net) Approach and Comparisons with U-Shaped CNN and Transformer Models for Building Segmentation from High-Resolution Satellite Images. Sensors, 22(19), 7624.

Transformers, which have been used frequently in recent years, are computationally inefficient and need large amounts of memory and big datasets [10].

Over the past few years, many researchers have tried various methods to achieve automatic building segmentation [1]. Among these methods, vision-based methods used for various studies ranging from deformation detection [171] and fraud detection to image segmentation have come to the fore. In conventional vision methods, manually extracted features including geometrical, spatial, and spectral information, and low-level features such as shape, color, edge, texture, and shadow are used [15]. These methods generally utilize these manually extracted features and apply classifiers or traditional machine learning techniques (e.g., Random Forests, Boosting, and Support Vector Machines) to achieve building segmentation [16]. However, extracting these features requires prior knowledge and is labor-intensive [17]. Although these methods have realized some progress, they have shortcomings, such as low accuracy, low generalization ability, and complex processing [7]. Due to the complex structures of buildings and similarities with other ground object categories (e.g., roads and cement floors), the predictions highly depend on the adjustment and feature design. With big data and the development of new algorithms and powerful hardware, deep learning methods have become common and gained a lot of traction in the remote sensing and computer vision communities.

Over the years, deep learning methods have gradually replaced conventional methods in image segmentation studies and achieved breakthroughs [16]. With the rapid development in computing powers and the increasing availability of data sources, deep learning methods, particularly CNNs, have started to be widely used as they surpass the traditional methods in terms of efficiency and accuracy [12]. Different from the conventional methods, CNNs have the capacity to extract features directly from the inputs and make predictions using sequential convolutions with fully connected layers [8]. CNNs can therefore be considered as one-step techniques that combine feature extraction and image classification into a single model. Hence, they also possess a good generalization ability [8]. Many researchers have used CNN architectures such as AlexNet [19], LeNet [20], VGGNet [21], ResNet [22], and GoogleNet [23]. However, these CNNs perform region divisions and use computationally expensive

fully connected layers [7]. In earlier studies, CNN models that were patch-based had achieved exceptional success in building segmentation; however, they were unable to guarantee the integrity and the spatial continuity of building features due to their reliance on small patches around target features to make predictions and ignoring the relationships between them [8]. Although the advances in CNNs have promoted research in the area, there still exist some challenges. For example, differences in building size, shape, and colors, geometric complexity, and high in-class and low inter-class variance make it hard to extract and segment buildings from high-resolution satellite images. Therefore, traditional CNNs are not suitable for segmentation studies from satellite images [8].

To improve the performance of CNNs, Long et al. [24] presented the FCN, which have convolution layers instead of fully connected layers, which improved training and prediction accuracy to a great extent. FCNs can output feature maps at the size of the input images using upsampling, and achieve pixel-based segmentation through an encoder-decoder structure. Small-sized feature maps are generated in the encoder path via downsampling to gather semantic information. In the decoder path, final segmentations are obtained by decoding the feature maps. However, the segmentation is not accurate enough, due to the FCN having only one upsampling layer, and much information is lost in the decoder [17].

Consequently, FCNs have become effective methods for image segmentation studies, and various variants have emerged, such as U-Net [26] and SegNet [28], to further improve its performance and efficiency. U-Net improves image segmentation accuracy by concatenating the semantic information through skip connections, implementing upsampling step by step, and integrating same scale downsampling in each step [26]. SegNet has a convolutional encoder with a pooling layer and a symmetrical decoder with transposed convolutions, which reduces the number of training parameters [28].

These CNN-based approaches have achieved successful results, but they also have some bottlenecks. For example, using a fixed-size convolution results in a local receptive field. CNNs are designed to successfully extract local context, and are inherently low in capacity to extract global context [16]. To capture the global context,

the most widely used approach is to implement the attention mechanism into the networks [29]. Several studies introduced the attention mechanism to improve feature representation; thereby, they can differentiate buildings from complex backgrounds [30,31]. Deng et al. [172] proposed a grid-based attention mechanism, Guo et al. [173] developed a parallel attention module, Pan et al. [174] tried to combine channel and spatial attention to improve the accuracy for building segmentation, and Cai and Chen [175] implemented a multi-path hybrid attention to improve the performance of segmenting small buildings. However, all these attention-based approaches rely on convolution operations to a great extent, and still have some limitation in global modeling [16]. Furthermore, numerous studies have been published to enhance the building segmentation performance of the models by concentrating on the network architectural design. Along with the attention mechanisms, approaches such as deep and shallow networks [32], multiple receptive fields, and residual connections [33] have been widely used.

CNNs have shown high performance on many image segmentation tasks. Despite their success, their low efficiency in capturing global context information still poses a challenge for researchers. Some studies have used the self-attention mechanism to overcome these issues; however, computational complexity grows with the spatial size, and hence, they may be suitable only for low-resolution images [46].

The Transformer approach was first proposed to be used in natural language processing (NLP), but has recently attracted interest in the computer vision community [10]. The first self-attention-based Vision Transformer (ViT) was proposed in 2020 [34]. It obtained competitive results on the ImageNet dataset and became an engaging approach in computer vision tasks. Different from CNNs, ViT takes 2D images, translates them into 1D sequences, and uses self-attention for feature characterization. Considering the advantage of this design, ViT shows superiority over CNNs in global feature extraction [16]. ViT-based approaches have made a tremendous job for image segmentation studies [35]–[37]. Many researchers in the computer vision and remote sensing field have implemented these methods for segmentation tasks, such as urban monitoring [40,41], land cover/land use analysis [38,39], change detection [42,43], and building segmentation [44,176]. Chen et al. [10] introduced a Sparse

Token Transformer to learn the global dependency of Transformer tokens, Yuan and Xu [17] implemented Swin Transformer as an encoder and designed a multi-scale adaptive decoder, and Wang et al. [16] proposed a ViT-based dual-path structure named BuildFormer.

CNNs can only pay attention to small range of neighborhood features, and they are insufficient in caring global features. Transformers can compensate for these shortcomings of CNNs by using the attention mechanism approach. In ViT-based methods, in contrast to CNNs, global information is extracted; however, spatially detailed context is ignored. In addition, when working with the large-sized, high-resolution images, their vector operations use all the pixels, and these operations are generally computationally inefficient and need a large amount of memory [10]. To deal with these issues, the Swin Transformer approach [45] utilizes shifted window-based multi-head self-attention, but the complexity still increases with the window size.

While Transformers perform well in global context modeling, they have exerted some limitations in fine-grained detail capturing [46]. To use the advantages of both sides, efforts have been paid to combine CNNs and Transformers, and Trans U-Net [47] was proposed. This approach initially utilizes CNN to extract low-level features, and then uses Transformer to model global interactions.

In this study, a total of ten CNN and Transformer models were generated, and building segmentation from high-resolution satellite images was carried out. Alongside the proposed Residual-Inception U-Net (RIU-Net) approach, U-Net, Residual U-Net, and Attention Residual U-Net models, four state-of-the-art CNN architectures (Inception, Inception-ResNet, Xception, and MobileNet) were implemented as encoders to U-Net-based CNN models. Furthermore, two Transformer-based approaches (Trans U-Net and Swin U-Net) were also used within the study, and comparisons were realized. RIU-Net is designed to cope with Transformer architectures that have achieved successful results in recent years, and deal with the issues that CNNs and Transformers are facing, by modernizing frequently used CNN approaches towards the ViT design approach. It is aimed to obtain an up-to-date CNN approach by using

modern methods such as Layer Normalization and GELU activation function inspired by the study from Liu et al. [177], together with modern and successful approaches such as Inception, residual connection, and asymmetric convolutions. There are many publicly available datasets that can be used for building segmentation, e.g., Massachusetts Buildings Dataset [104], Inria Aerial Image Labeling Dataset [91], and WHU Building Dataset [58]. The Massachusetts Buildings Dataset and the Inria Aerial Image Labeling Dataset were used in this study to train and test the created models. Performance evaluations for the models were completed using evaluation metrics F1 and Intersection over Union (IoU) scores, and the acquired results were discussed. The study aims to make comparisons between several CNN and Transformer-based models, and to see the position of our proposed RIU-Net model among the others, in terms of forming a solution to the above-mentioned problems, for building segmentation from high-resolution satellite images. The main contributions of this study can be summarized as below:

- A novel CNN architecture approach, modernized towards the ViT design, named RIU-Net, has been proposed.
- To cope with Transformers and the issues that CNNs are dealing with, our proposed architecture uses approaches and methods such as a u-shaped design, Inception, residual connections, skip connections, asymmetric convolutions, GELU, and Layer Normalization together.
- Comparisons between our proposed approach and several state-of-the-art CNN and Transformer approaches have been realized using two different publicly available building segmentation datasets that contain high-resolution satellite images.
- It has been observed that the RIU-Net approach is significantly successful when sufficient data are provided, especially on datasets containing complex buildings with different characteristics.

5.2 Materials and Methods

5.2.1 Datasets

Within the study, two publicly available building segmentation datasets were used. These datasets selected are: Inria Aerial Image Labeling Dataset and Massachusetts Buildings Dataset.

5.2.1.1 Inria Aerial Image Labeling Dataset

Inria Aerial Image Labeling Dataset [91] features RGB, orthorectified images with a 30 cm spatial resolution. It covers a total area of 810 km^2 and is split into two sets, training and testing. There is a corresponding labeled mask for each image in the training set, containing two classes, which are building and non-building. Since only the masks of the images in the training set are available publicly, these data were used in this study.

The Inria dataset includes images from different cities, extending from rural to urban areas, with various building characteristics. A total of 180 images and corresponding masks from five cities are included in the training set (36 images and masks from each city). The cities are; Western Tyrol (Austria), Vienna (Austria), Kitsap County (WA, USA), Chicago (IL, USA), and Austin (TX, USA) [91]. Each image and mask are 5000$\times$5000 pixels and cover an area of 1500 m$\times$1500 m. The use of these data provides an increase in the generalization capabilities of the models, since they contain images with different building characteristics from various different regions.

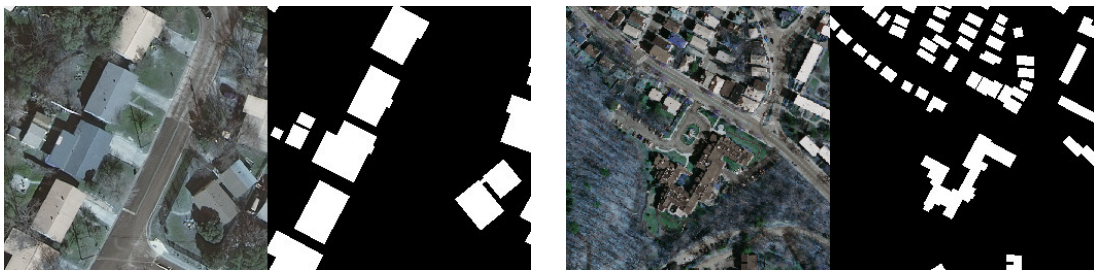
5.2.1.2 Massachusetts Buildings Dataset

Massachusetts Buildings Datasets were introduced by Volodymyr Mnih [104] in 2013, as a part of his Ph.D. thesis. This dataset consists of 151 RGB images and their labeled masks of Boston, MA, USA. Every image and mask in the dataset is 1500$\times$1500 pixels in size, has a spatial resolution of 1 m, and spans a 2.25 km^2 area. This dataset covers mostly suburban and urban areas, which include buildings with similar characteristics

and different sizes. The building labels for masks were obtained by digitizing buildings from the OpenStreetMap project [104].

5.2.1.3 Preparation of the Data

To prepare the datasets to be used in the study, the images and corresponding masks from both datasets were initially cropped into 256×256 pixel-sized patches, to reduce the computational cost. After the cropping process, images with no buildings or few buildings were removed from the datasets. From the Inria dataset, 7500 images and corresponding masks were obtained. Out of these images, 70% (5250) were selected as the training set, 15% (1125) as the validation set, and 15% as the test set. From the Massachusetts dataset, after the cropping and eliminations, only 488 images and their masks were obtained. Since the quantity of images is insufficient for this study, "brightness adjustment" and "rotation" data augmentation techniques have been used on the images. The brightness of the images was raised by a gamma rate of 0.5, and they were rotated 90° clockwise. After these operations, two datasets were obtained, named as Massachusetts Small and Massachusetts Large. Massachusetts Small dataset contains a total of 1500 images and masks, whereas Massachusetts Large dataset contains 3041 images and masks. Similar to the Inria dataset, 70% of the data for both datasets were selected as the training set, 15% as the validation set, and 15% as the test set. Consequently, three datasets were prepared using the Inria and the Massachusetts datasets (Figure 5.1).



(a) (b)

Figure 5.1 : Sample 256×256 pixel image and mask: (**a**) Inria dataset, (**b**) Massachusetts dataset.

5.2.2 Methodology

A total of ten CNN and Transformer models were generated and used for building segmentation from high-resolution satellite images. Alongside the proposed model approach RIU-Net, U-Net, Residual U-Net, and Attention Residual U-Net models, four state-of-the-art CNN architectures, which are Inception, Inception-ResNet, Xception, and MobileNet, were implemented as encoders to the U-Net-based CNN models. Moreover, two Transformer-based approaches, Trans U-Net and Swin U-Net, completed the model tree used in this study.

5.2.2.1 U-Net

U-Net is an FCN-based architecture proposed by Ronneberger et al. [26]. It has a symmetrical, u-shaped design that includes an encoder, a decoder, and bottleneck paths. The encoder extracts relevant features and these extracted features propagate to the decoder using skip connections. Afterward, the decoder reconstructs the images into desirable dimensions using these feature maps. The bottleneck path is placed between the encoder path and the decoder path, and contains two 3×3 convolution layers. There are four convolution blocks in the encoder path, and each one includes two 3×3 convolution layers with Rectified Linear Unit (ReLU) activation, and a 2×2 max pooling layer. The decoder path also has of four blocks, with each block having one 2×2 transposed convolution layer, a concatenation layer to connect the extracted feature maps, and two 3×3 convolution layers [26].

5.2.2.2 Residual U-Net

The Residual connection was introduced by He et al. [22] via ResNet architecture to cope with the problems of deep CNNs. When a model has more layers added on, it experiences the "vanishing gradient problem". With the help of the residual connections, spatial information can be passed layers down directly, to cope with this problem.

In this study, a Residual U-Net model was generated. The model combines residual connections with the U-Net architecture by replacing conventional convolutional blocks in the U-Net with residual convolutional blocks. In this model, another variant was used instead of the residual connection design used in ResNet. Several different designs have been tried and the design proposed by He et al. [146], shown in Figure 5.2, was chosen, and it was modified to replace the conventional convolutional block in U-Net.

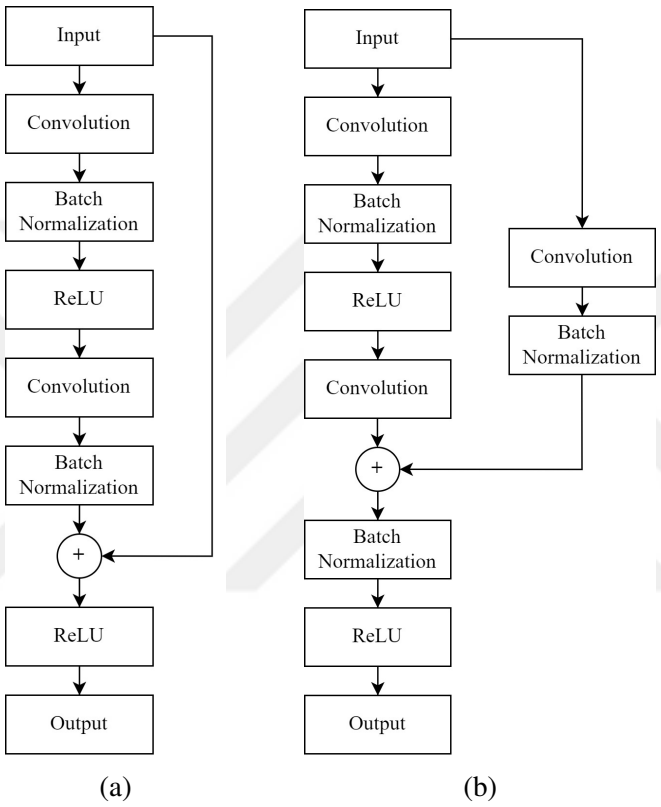


Figure 5.2 : (**a**) Residual connection design from the ResNet, (**b**) Residual connection design implemented in the study.

5.2.2.3 Attention Residual U-Net

In addition to the Residual U-Net model, an Attention Residual U-Net model was also utilized within the study. This approach combines the Residual U-Net model with the attention mechanism to achieve performance improvements. Through this implementation, the model could focus more on relevant features and produce better results [144]. U-Net has more spatial information in the earlier stages of the network, and this spatial information provides a rich context to the following stages via skip

connections. However, with this process, poor representations from the previous stages also come with good ones. To overcome the problem, the attention mechanism is implemented to skip connections to suppress activations from unrelated areas of the image. In the study, the attention mechanism proposed by Oktay et al. [30] was implemented in the Residual U-Net model (Figure 5.3).

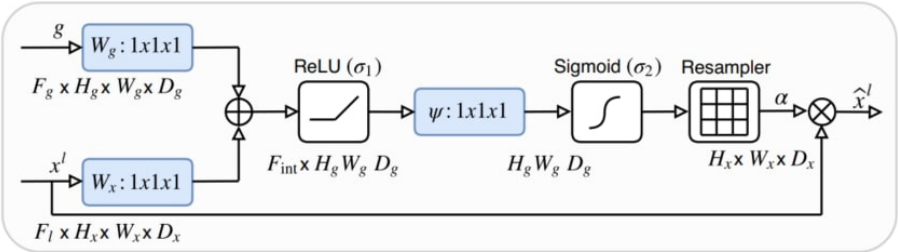


Figure 5.3 : Attention mechanism implemented in the study [30].

5.2.2.4 Inception Backboned U-Net

Szegedy et al. [23] proposed the first version of Inception in 2015 as GoogleNet. Their primary aim was to improve the computing resource usage of the model. They increased the depth of the model, along with the width, and kept the computational cost constant. They assumed that each unit of the earlier layers corresponds to some relevant region of the input image, and these are grouped into filter banks. Thus, they proposed many clusters concentrated in a specific region, and then covered them with 1×1 convolutions in the next layer. However, it would be expected that there will be spatially spread out clusters that can be covered by convolutions over large patches and a decrease in the number of patches over the larger regions [23]. Thus, the Inception architecture is restricted to 1×1, 3×3, and 5×5 filter sizes. Additionally, since pooling operations have been generally used in convolution networks, they added a pooling path in each Inception step. In the study, the encoder of the U-Net architecture was replaced with the Inception-v1 architecture, and convolution blocks in bottleneck and decoder paths were replaced with residual convolution blocks to generate the model.

5.2.2.5 Inception-ResNet Backboned U-Net

One of the examples of the widely used deep CNNs is the Inception architecture [23], which achieves successful results at a low computational cost. Lately, the residual connections [22] have provided accomplished results in computer vision tasks. This raised a question for Szegedy et al. [23] of whether there would be a benefit in combining Inception networks with residual connections or not. They studied this idea, and proposed the Inception-ResNet architecture [178].

They thought that, since Inception networks are very deep, it would be reasonable to replace the filter concatenations with residual connections. For the residual version of Inception, they used cheaper blocks than the original. Each Inception block is followed by a 1×1 convolution without the activation as a filter expansion layer to scale up the dimensionality, prior to residual addition. In the study, the encoder of the U-Net architecture was replaced with the Inception-ResNet-v2 architecture, and convolution blocks in bottleneck and decoder paths were replaced with residual convolution blocks to generate this model.

5.2.2.6 Xception Backboned U-Net

François Chollet presented an Inception interpretation as a step between conventional convolutions and depthwise separable convolutions, called "Xception" [179]. It is possible to think of the depthwise separable convolution as an Inception module with a lot of towers [179]. The idea prompted them to suggest a deep CNN modeled after Inception, by using separable convolutions in place of the Inception module [179]. This extreme version of the Inception module firstly maps the cross-channel correlations by using a 1×1 convolution; after that, it maps every output channel's spatial correlations separately. In the study, the encoder of the U-Net architecture was replaced with the Xception architecture, and convolution blocks in bottleneck and decoder paths were replaced with residual convolution blocks to generate the model.

5.2.2.7 MobileNet Backboned U-Net

Howard et al. presented mobile and efficient models called "MobileNets" in 2017 [180]. MobileNets are based on streamlined architectures that use depthwise separable convolutions to make the model lightweight. Their structure is built on depthwise separable convolutions for less expensive processing, except for the first layer, which is full convolution [180]. In the models, Batch Normalization and ReLU are placed after each convolution layer. Downsampling is performed with the strided convolution in depthwise convolutions, and in the first convolution layer. To lower the spatial resolution prior to the final output layer, an average pooling is utilized at the end of the model. In the study, the encoder of the U-Net architecture was replaced with the MobileNet-v1 architecture, and convolution blocks in bottleneck and decoder paths were replaced with residual convolution blocks to generate this model.

5.2.2.8 Trans U-Net

For various image segmentation tasks, u-shaped CNNs achieved significant success. However, due to the fundamental locality of convolutions, these architectures are demonstrating some limitations [47]. Nonetheless, Transformers have been developed as alternative approaches with self-attention mechanisms. However, due to insufficient low-level details, they can result in limited localization abilities. Chen et al. [47] proposed Trans U-Net, which utilizes both U-Net and Transformers, as an alternative for computer vision tasks. On one side, the Transformer encodes tokenized image patches from the CNN feature map as input sequences. On the other side, the decoder upsamples these encoded features, which are combined with high-resolution feature maps to achieve accurate and precise localization [47].

Chen et al. [47] conducted a study on different model sizes and investigated two configurations, as "Base" and "Large" models. These two are separated from each other by size differences due to hyperparameter differences. However, in this study, a smaller variant is used, due to the computational resource limitations. The model

was scaled according to the ratios of hyperparameters between the base and the large models, and used in the study as a variant named "Mini", referred in Table 5.1.

Table 5.1 : Hyperparameters of Large, Base, and Mini Trans U-Net variants.

Model Scale	Hidden Size (D)	No. of Layers	MLP Size	No. of Heads
Large	24	1024	4096	16
Base	12	768	3072	12
Mini	6	576	2304	9

5.2.2.9 Swin U-Net

Cao et al. [181] proposed Swin U-Net in 2021, a U-Net-like Transformer approach. In Swin U-Net, tokenized image patches are fed into a u-shaped, Transformer-based, encoder–decoder architecture that uses skip connections. They used Swin Transformers with shifted windows (swin) as an encoder for feature extraction. Motivated by the Swin Transformers [45], they proposed Swin U-Net to enhance the power of Transformers for image segmentation. Swin U-Net is the first u-shaped, Transformer-based architecture that consists of an encoder, a decoder, and a bottleneck, which are built on the Swin Transformer block, with skip connections [181].

5.2.2.10 Residual-Inception U-Net (RIU-Net)

In this study, a residually connected, Inception-based, u-shaped encoder-decoder architecture with skip connections named as Residual-Inception U-Net (RIU-Net) has been proposed. The model includes an encoder path, a bottleneck, and a decoder path. For the encoder path of this architecture, a modified Inception-ResNet-v2 network [178] was used.

Throughout the network, as differs from the Inception-ResNet-v2, "Layer Normalization" [182] was used as a regularization technique instead of Batch Normalization, and the "GELU" activation function [183] was used instead of ReLU. The ReLU activation function is widely used in CNNs due to its efficiency and simplicity. GELU, which is a smoother variant of ReLU, is used in the most advanced Transformer approaches, and most recently in ViT [177]. As a result of the experiments, it was seen that the use of

GELU instead of ReLU, which is widely used in computer vision studies, positively affects the accuracy. Batch Normalization is an important component of CNNs, since it improves convergence and reduces overfitting. However, Batch Normalization also has many complexities that can reduce the performance of the model [177]. There have been several attempts to develop an alternative, but Batch Normalization remained as the most commonly used one for many computer vision tasks. On the contrary, Layer Normalization, which is a simpler variant, has been used in Transformers, and it resulted in a good performance across many different applications [177]. In this study, the replacement of Batch Normalization with Layer Normalization resulted in a significant improvement in the accuracy of the proposed model. All convolution blocks used in the architecture include a convolution layer, Layer Normalization, and GELU activation, respectively, unless otherwise stated. These modifications are inspired by Liu et al. [177], where they "modernize" the CNNs toward the design of ViT.

The overall design of the proposed approach is given in Figure 5.4. For the first part of the encoder, STEM design from the Inception-ResNet-v2 was used, with slight modifications. All "Valid" paddings used in the STEM were replaced with "Same" paddings. For the rest of the encoder, the modules used in Inception-ResNet-v2 were modified and implemented into the architecture. The Inception approach has achieved satisfactory performance at a low computational cost. In addition, combining them with residual connections significantly accelerated the training of Inception-based models. Since the Inception networks tend to be deep, it is beneficial to use residual connections instead of conventional convolutional connections. The use of residual connections helps to solve the vanishing gradient problem that exists in deep networks [31]. It allows Inception to acquire all the benefits of the residual connection while maintaining the computational efficiency [178].

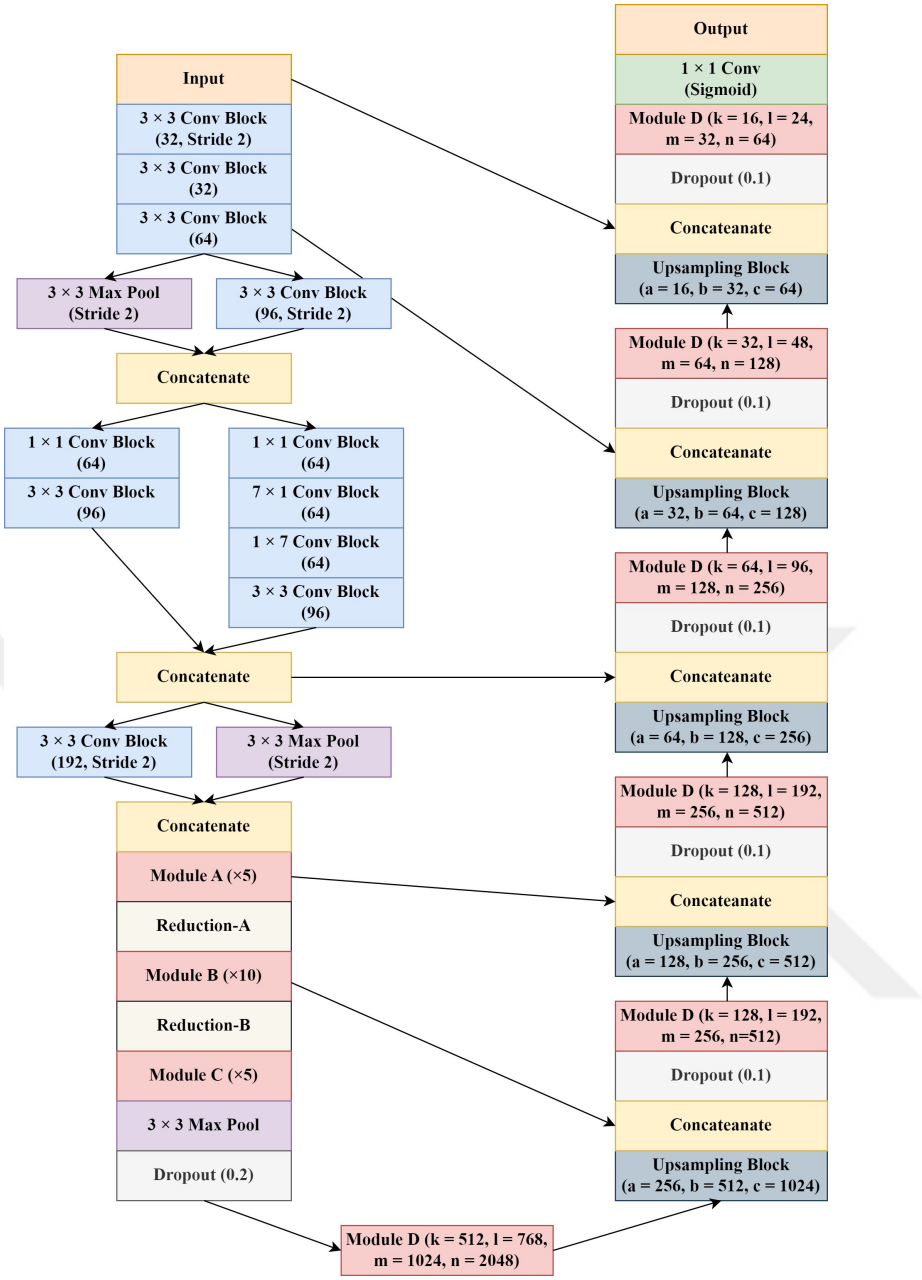


Figure 5.4 : Overall architecture of the RIU-Net.

The main modules A, B, and C were generated by modifying the Inception-ResNet-v2 modules. These modules were implemented in the encoder after the STEM. In these main modules, 1×1 convolution blocks were added and applied to the data coming from the previous layers different from the original ones. Afterward, the outputs of these shortcut blocks and the outputs of the main Inception blocks were merged to achieve the residual connection. In addition, the use of asymmetric convolutions in the STEM and the main modules resulted in significant computational cost savings.

Reduction modules A and B remained similar to the original ones, but with all "Valid" paddings being replaced with "Same" paddings. The designs of all modules used in the encoder path are shown in Figure 5.5. Lastly, at the end of the encoder path, the average pooling layer in the Inception-ResNet-v2 has been replaced with a max pooling layer.

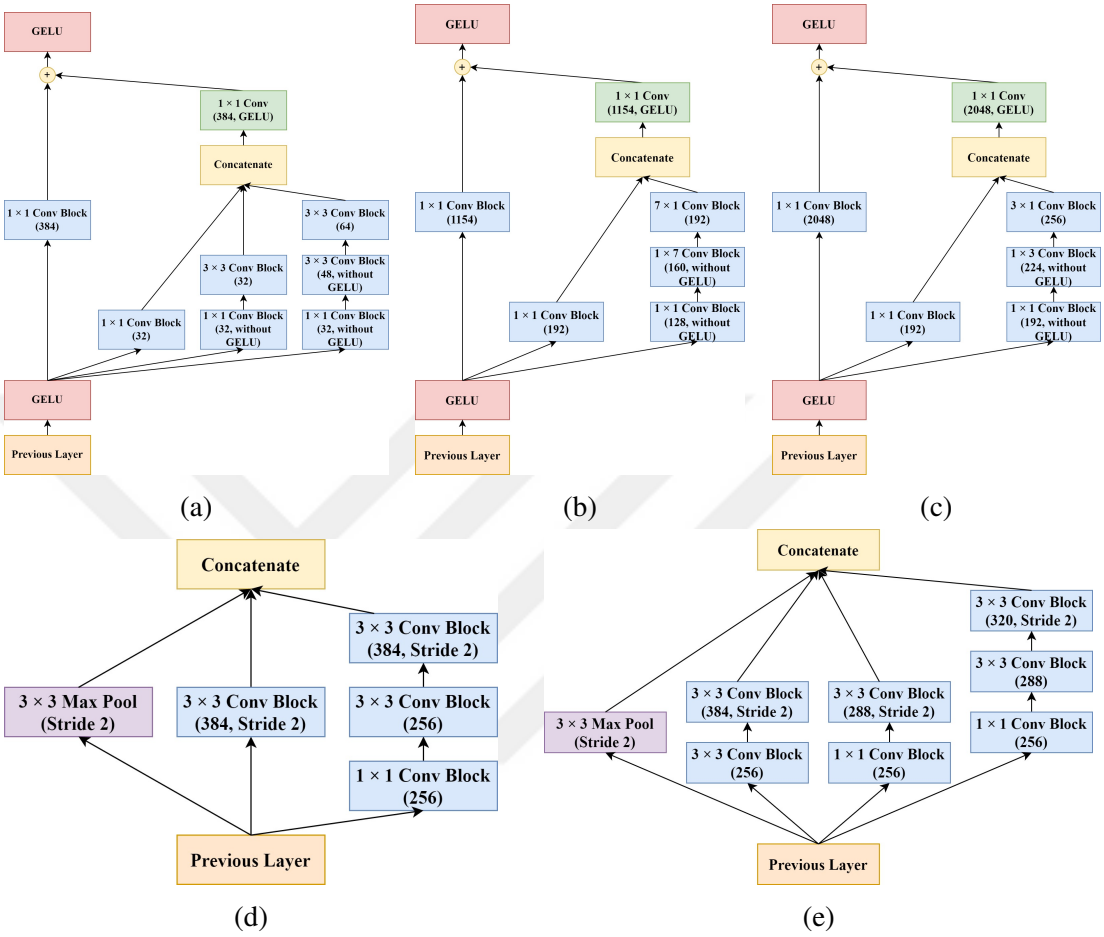


Figure 5.5 : The flow diagram of the modules used in the encoder path of the RIU-Net: (**a**) Module A, (**b**) Module B, (**c**) Module C, (**d**) Reduction A, and (**e**) Reduction B.

In the bottleneck and the decoder paths, residual Inception module D design was used with various numbers of filters. Additionally, instead of directly using a transposed convolutional layer, as was practiced in most CNN approaches, an Inception-based upsampling module was created and used in the decoder path. The direct use of max pooling and transposed convolutions to upsample the feature maps, such as in standard U-Net, may lead to accuracy reduction and feature loss [184]. Thus, upsampling modules have been preferred for overcoming these problems. The upsampling module was inspired by the paper from Zhang et al. [184] and implemented into the study

with slight modifications. Lastly, between the encoder and the decoder paths, skip connections were used to avoid spatial information loss by transferring low-level feature maps from the encoder to the decoder. With these connections established between the relevant layers, low-level and high-level details were combined. The designs of Module D and the upsampling module are shown in Figure 5.6.

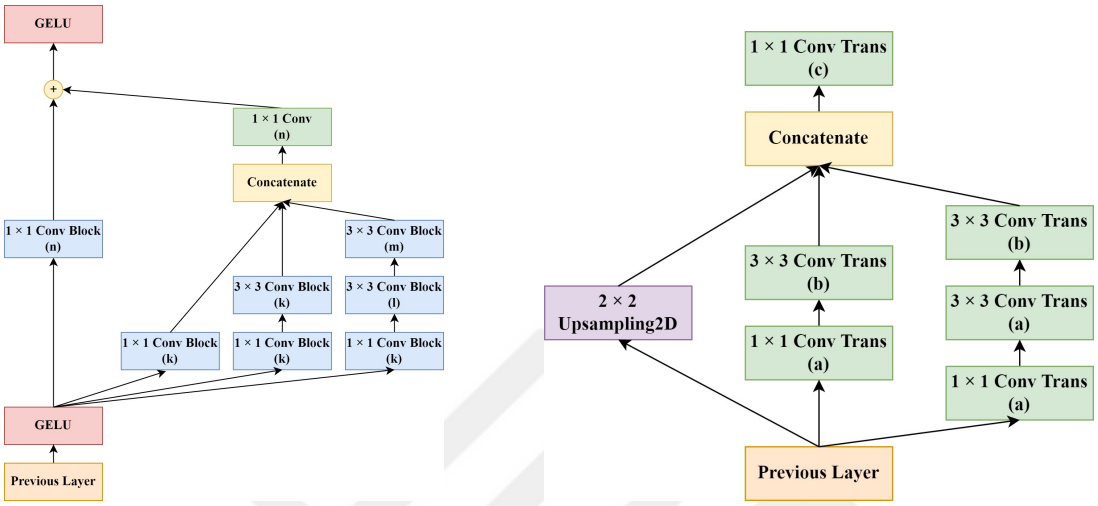


Figure 5.6 : The flow diagram of the modules used in the bottleneck and decoder paths of the RIU-Net: (**a**) Module D, and (**b**) Upsampling module.

5.3 Experiments

Prepared training sets were used to train the generated models. Every experiment within this study was run on the cloud via the Google Colaboratory platform using an NVIDIA Tesla P100 GPU, and all the experiments, model developments, and implementations were performed using the Tensorflow framework. As mentioned previously, for all datasets, 70% of the data were used for training, 15% for validation during training, and the remaining 15% for testing. During the training process, as an optimizer, Adam [112] was used with a 1e-4 initial learning rate. To reduce the learning rate during training, "ReduceLROnPlateau" callback was applied. Validation loss values were monitored by this callback, and the learning rate was decreased by a factor of 0.1 when the value did not improve for five consecutive epochs. The models were trained using a batch size of 2, and two loss functions were used together to calculate loss values. These loss functions are the Binary Cross-entropy Loss and the

Jaccard Loss. After a series of experiments, including different loss functions and their combinations, it has been observed that a combination of these two functions provided the best results, according to the evaluations made. Binary Cross-entropy is a function to calculate the loss used on binary segmentation and classification tasks, and it is frequently used to measure how well the predicted class probabilities match the correct classes (Equation 5.1) [154]. Jaccard Loss, IoU Loss, or Jaccard Index, is a ratio between the intersection of the predicted and ground truth and the union of these two (Equation 5.2) [154]. Combining two loss functions allows for diversity in the loss while taking the advantage of the stability of Binary Cross-entropy. To determine the number of epochs for the models to train, a callback called "EarlyStopping" was used with a maximum of 50 epochs. Validation loss values were monitored and if the loss value did not improve for 10 consecutive epochs, the training was stopped. Another callback, the "ModelCheckpoint", was also used during training. With this function, the model was saved each time the loss value improved during training, and then the best model was selected for testing and evaluation. The epoch of the best models and the number of trainable parameters for all models are presented in Table 5.2.

$$L(y,\hat{y}) = \frac{1}{N} + \sum_{i=0}^{N} (y * \log(\hat{y}_i) + (1-y) * \log(1-\hat{y}_i)) \tag{5.1}$$

$$Jaccard(A,B) = 1 - \frac{|\,A \cap B\,|}{|\,A \cup B\,|} = 1 - \frac{|\,A \cap B\,|}{|\,A\,| + |\,B\,| - |\,A \cap B\,|} \tag{5.2}$$

Table 5.2 : No. of trainable parameters and epochs of the best models.

Architectures	No. of Trainable Parameters	Epoch of the Best Model (Inria)	Epoch of the Best Model (Mass. Small)	Epoch of the Best Model (Mass. Large)
U-Net	31 M	27	29	25
Residual U-Net	32.4 M	10	19	21
Attention Residual U-Net	37.3 M	21	25	24
Inception backboned U-Net	373.1 M	6	16	26
Xception backboned U-Net	463.7 M	29	21	15
MobileNet backboned U-Net	102.8 M	12	16	17
Inception-ResNet backboned U-Net	164 M	10	25	17
Trans U-Net (Mini)	97.7 M	26	29	19
Swin U-Net	15 M	34	14	16
Residual-Inception U-Net (RIU-Net)	104.4 M	50	41	48

After the training, the best models were selected, and tests and evaluations were made using the prepared test sets. Metrics for evaluation were calculated, and the images were segmented using a 0.5 threshold applied to the predicted class probabilities. For the evaluation of the models; Recall (Equation 5.3), Precision (Equation 5.3), IoU (Equation 5.4), and F1 Score (Equation 5.5) evaluation metrics were used. F1 score, which is the harmonic mean of both Precision and Recall, is an evaluation metric used to assess a model's overall performance [110]. IoU is a metric that gives the ratio of the intersection of the predicted area and the labeled area to the sum of these two areas [154]. These two metrics were taken as the main evaluation metrics to evaluate the models, as they are more suitable for image segmentation tasks. Along with the mentioned metrics, test loss and test accuracy were also calculated.

$$\text{Precision} = \frac{TP}{(TP+FP)} \qquad \text{Recall} = \frac{TP}{(TP+FN)} \tag{5.3}$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{(TP+FP+FN)} \tag{5.4}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.5}$$

5.4 Results and Discussion

Evaluation results on the Inria test set are shown in Figure 5.7. According to the results, our proposed RIU-Net approach has the highest IoU score, F1 score, and test accuracy, with 0.6736, 0.7868, and 92.23%, respectively. The Transformer models Swin U-Net and Trans U-Net follow this approach. Swin U-Net has a 0.5932 IoU score, a 0.7210 F1 score, and 89.65% test accuracy. Trans U-Net follows, with 0.5375 IoU score, 0.6681 F1 score, and 89.65% test accuracy. For all three metrics, the MobileNet backboned U-Net has the lowest values, with 0.4452 IoU score, 0.5763 F1 score, and 78.13% test accuracy.

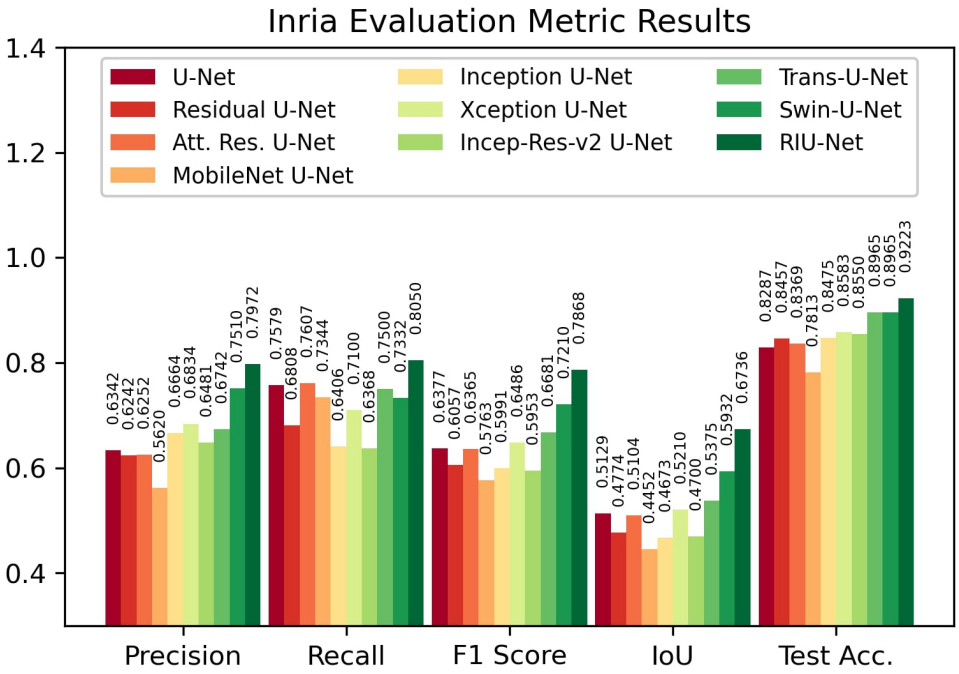


Figure 5.7 : Evaluation metric results on Inria test set.

Evaluation results on Massachusetts Small test set are shown in Figure 5.8. On this test set, the Attention Residual U-Net model provided the highest IoU score with 0.6218 and the highest F1 score with 0.7606. On the contrary, the Trans U-Net model provided the highest test accuracy with 94.26%. For the IoU score and F1 score metrics, Trans U-Net follows Attention Residual U-Net with 0.6188 and 0.7581, respectively. Residual U-Net follows Attention Residual U-Net with a 0.7566 F1 score and 0.6159 IoU score.

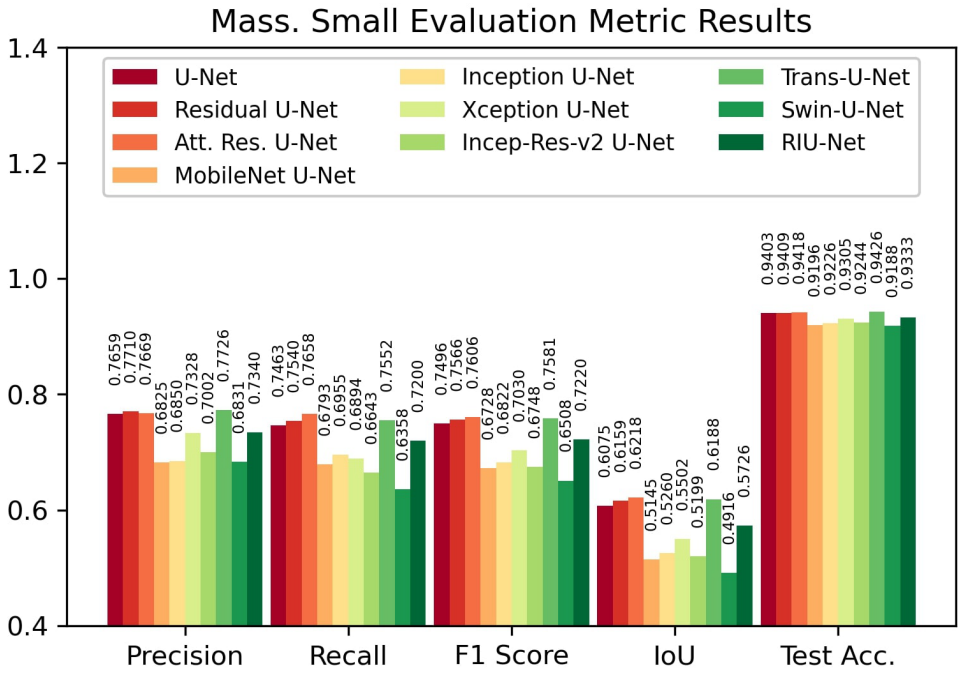


Figure 5.8 : Evaluation metric results on Massachusetts Small test set.

According to the results on Massachusetts Large test set, shown in Figure 5.9, the Residual U-Net model provided the highest IoU score with 0.6165 and F1 score with 0.7565. This is followed by Attention Residual U-Net with 0.6121 IoU score and 0.7518 F1 score, and Trans U-Net with 0.6103 IoU score and 0.7509 F1 score. On both Massachusetts test sets, Swin U-Net and MobileNet backboned U-Net models have the lowest evaluation metric values.

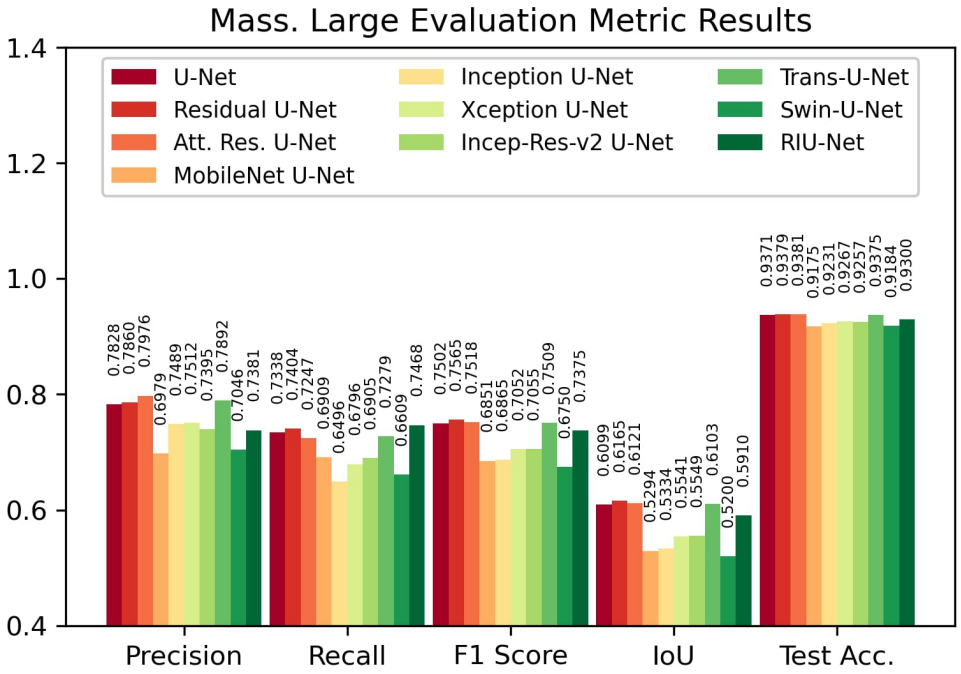


Figure 5.9 : Evaluation metric results on Massachusetts Large test set.

RGB images and corresponding masks from test sets, and example segmented images are shown in Figure 5.10, Figure 5.11, and Figure 5.12, respectively. In the segmented images, black pixels represent the background and the white pixels represent the buildings. It has been observed that RIU-Net and Transformer approaches Swin U-Net and Trans U-Net have performed well according to the predictions on the Inria test set shown in Figure 5.10. Compared to other approaches used in the study, these models, especially the RIU-Net, are successful in segmenting the buildings with different colors, shapes, and textures in the related image no. 192. Other models have been found to have difficulty in distinguishing between dark building parts and trees, which are particularly similar in hue. According to the predictions on the Massachusetts test sets shown in Figures 5.11 and 5.12, Trans U-Net, Attention Residual U-Net, and Residual U-Net have performed better than other approaches. Other models seem to classify non-buildings as buildings, or fail to classify some buildings. This problem is seen especially in areas where the shadow effect is high, or where buildings are in close color hue with the background.

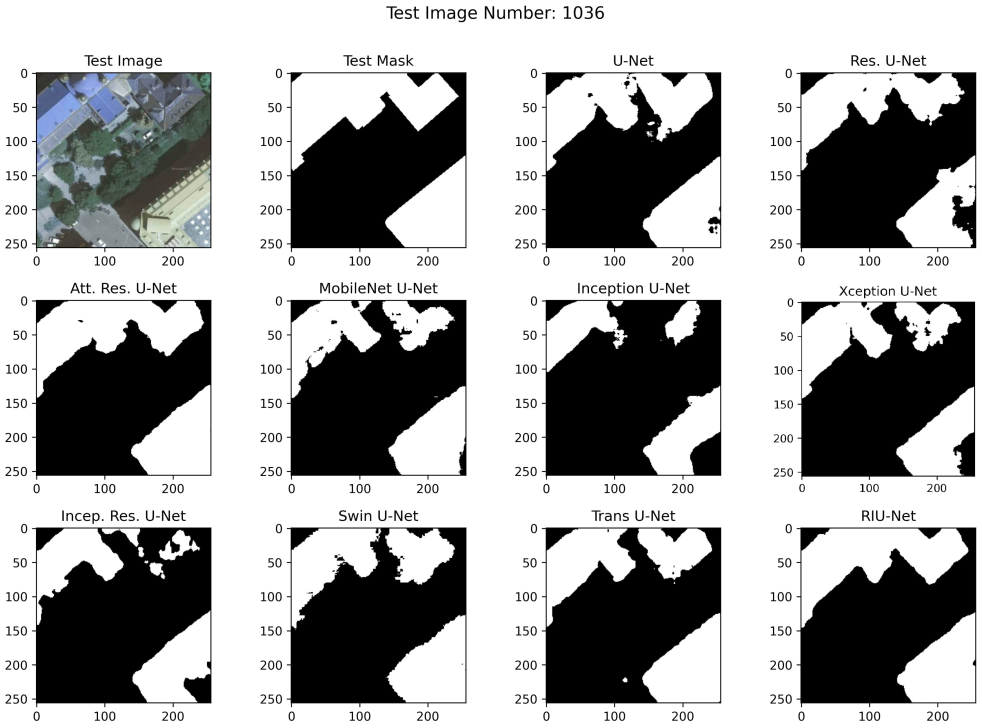


Figure 5.10 : Inria test set image no. 1036 segmentation results.

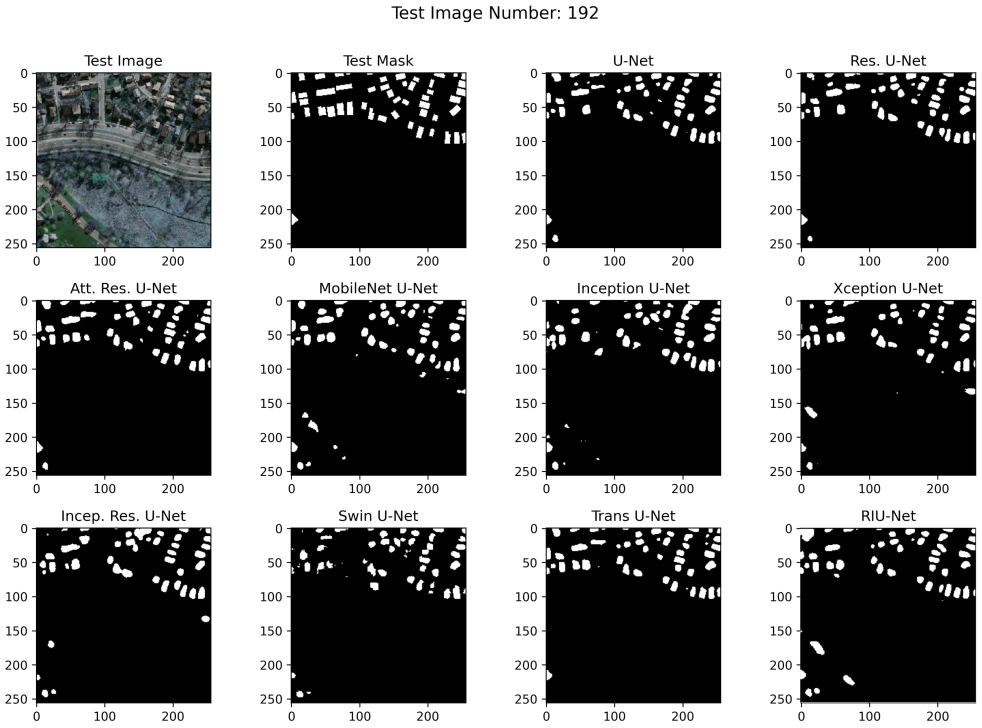


Figure 5.11 : Massachusetts Small test set image no. 192 segmentation results.

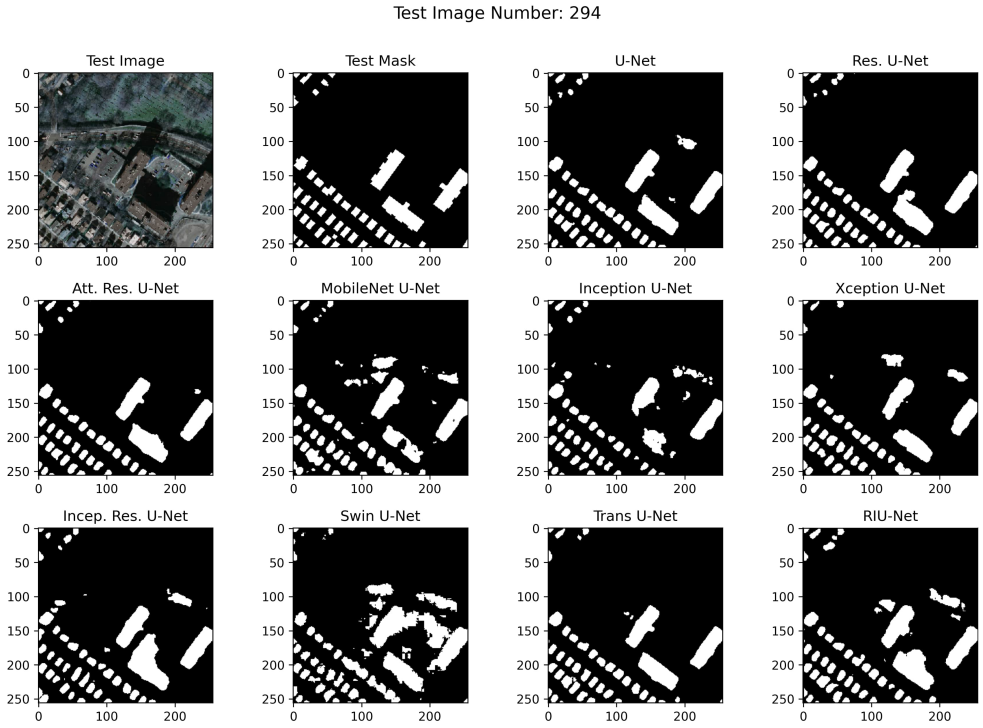


Figure 5.12 : Massachusetts Large test set image no. 294 segmentation results.

On the Inria dataset, which includes buildings with different characteristics from different cities, the proposed RIU-Net approach has achieved significantly more successful results than other models in all metrics. On Massachusetts datasets containing images of buildings with similar characteristics from a single city (Boston, MA, USA), Residual U-Net, Attention Residual U-Net, and Trans U-Net models were found to give more successful results than the proposed approach. When the results obtained from the small and large datasets generated from the Massachusetts Buildings Dataset are compared with each other, it is seen that the RIU-Net indicates a significant increase in success compared to other models as the size of the dataset increases, and the difference between the models that give more positive results than themselves decrease. In addition, while the RIU-Net continues to learn when the 50-epoch limit is reached on all datasets, it has been observed that all other models stopped training with early stopping before this limit is reached. This demonstrates that longer-term training can improve the performance of the proposed approach even further.

On all the three test sets, it was observed that the MobileNet backboned U-Net model was among the least successful models, while the Trans U-Net model was among

the top three successful models. It was seen that the Swin U-Net model gave the least successful results on Massachusetts datasets, whereas it was the second most successful model behind the proposed approach on the Inria dataset containing data with different characters. It is also seen that the proposed approach gives more effective results than the U-Net models using Inception, Xception, Inception ResNet, and MobileNet backbones in all three datasets.

It has also been seen that, in both the Inria dataset and the Massachusetts dataset, there are some inaccuracies. For example, buildings that are not in the images but are in the corresponding masks, buildings that are not fully visible in the images because they are covered with obstacles but contained in masks, and some missing labels.

In light of these results, it can be interpreted that the proposed RIU-Net approach provides more successful results than the models used in the study on data containing complex and different characteristic features, and on data containing more similar details, and that it starts to approach the performance of other models as the size of the dataset grows.

5.5 Conclusions

Over the past few years, automatic building segmentation and extraction from aerial images has become a significantly important subject due to the needs in application areas, such as city and regional planning, and change detection and disaster management, and the increase in usable data. Building segmentation was performed using generated datasets from the Massachusetts Buildings Dataset and the Inria Aerial Image Labeling Dataset within the context of the study. Ten CNN-based and Transformer-based models were generated, and their respective performance comparisons were subsequently made.

On the Inria dataset, which has buildings from various regions with different characteristics, the proposed RIU-Net model achieved the most successful results among all the models used. According to all evaluation metrics, RIU-Net performed significantly better than all other nine models, on this dataset. On the datasets generated using the Massachusetts Buildings Dataset, which includes buildings with

similar characteristics and from the same region, Residual U-Net and Attention Residual U-Net, along with the Trans U-Net, one of the Transformer-based approaches, performed better. When the results obtained on the Small and Large Massachusetts datasets are compared, it is seen that the performance of the proposed RIU-Net model significantly increases compared to other models, as the size of the dataset grows. However, when the training process is examined, it has been observed that extending the training period may increase the performance of this proposed model.

When all the results are examined in general, it can be interpreted that the proposed RIU-Net approach is expressively successful on datasets containing complex buildings with different characteristics. On the other hand, on datasets containing buildings with mostly similar characteristics, it can be predicted that the proposed model will catch up with, and perhaps even exceed, the performance of other models used in the study when sufficient data and longer periods of training times are provided.

It is predicted that the proposed architecture can be used in classification and detection studies, in addition to segmentation studies, as in other CNN architectures used in literature. This approach can also be applied to other datasets for different image segmentation studies. In this study, as a result of the use of a u-shaped design with skip connections inspired by U-Net, which has achieved successful results in segmentation studies, together with Layer Normalization, GELU, Inception, residual connection, and asymmetric convolution approaches, a CNN model that can cope with Transformers, which have achieved successful results in recent years, has been proposed. This model approach has been achieved by modernizing existing and frequently used CNN approaches toward the ViT design.

In future studies, it is planned to test these models by conducting experiments with more data and longer training times. In addition to these, the creation of a more accurate dataset containing buildings from different regions and with various characteristics is also planned. It is also premeditated to train the models used in the study with large datasets containing buildings with different characteristics from different regions, and to share these trained models with researchers and potential users. In addition, to combine the best parts of both the CNN and Transformer

approaches, it is planned to implement CNN–Transformer hybrid approaches, including our RIU-Net architecture, as the CNN part of the model.

6. CONCLUSIONS

In recent years, with the developments in satellite and sensor technologies and due to the rapid urbanization rate, automatic building segmentation from high-resolution images has become an important and popular research topic among the remote sensing and computer vision communities. With the help of these developments, large amounts of high-resolution images have become more accessible and convenient data sources.

As one the most important features of the urban database and the human livelihood, buildings are important places and key features for urban area mapping applications [3]. Within this importance, many different applications and studies, such as urban mapping, change detection, and disaster management, require building maps. Nonetheless, obtaining accurate and reliable building maps is still a challenging tasks due to various reasons and improving the accuracy and efficiency is still a focus for many researchers. In the light of this information, the studies and research presented in each paper were carried out within the scope of this thesis.

The first paper is based on the investigation of U-Net-based shallow and deep CNN architectures for building segmentation from high-resolution images. Comparisons were made between the original U-Net and four deeper and shallower variants of the architecture, and the effects of depth and shallowness of the architectures were evaluated. Models were trained and evaluated using datasets prepared from the Inria Aerial Image Labeling Dataset, and the Massachusetts Buildings Dataset. The fact that the U-Net Deeper 1 architecture provided the most successful results on the Inria dataset suggests that to a certain depth, deep architectures can give better results, even if there is not too much data. When the shallower approaches were examined, it has been observed that the U-Net shallower 2 architecture provided the worst results on both datasets, as this approach remained too simple. On the other hand, the U-Net shallower 1 approach provided similar evaluation metric results as the U-Net Deeper 1 architecture on the Massachusetts dataset. The principal aim of this first study was

to realize comparisons between deep and shallow CNN approaches. In this study, two somewhat small-sized datasets were used, and besides the depth of the networks, no further modifications were made.

Within the scope of the second paper, building segmentation from high-resolution aerial images was realized using datasets prepared using the Inria Aerial Image Labeling Dataset and the Massachusetts dataset, similar to the first paper. Nevertheless, the Inria dataset used in this study is a larger version of the Inria dataset from the first paper. In this study, a total of nine models, including U-Net and FPN which utilize four backbones named ResNet-50, ResNeXt-50, SE-ResNeXt-50, DenseNet-121, and additionally an Attention Residual U-Net approach were trained and evaluated. On the Massachusetts dataset, which is the smaller-sized dataset out of the used two, it has been observed that FPN architecture provided more successful results than U-Net. Among the used backbones, the DenseNet-121 backbone provided the most successful results since it reduces the overfitting problem which can occur when using small datasets. On the larger Inria dataset, the proposed Attention Residual U-Net approach outperformed all other models according to the evaluation metric results. Considering that the models using this approach are trained from scratch, unlike the models using backbones which used pretrained weights, this approach can be considered successful.

In the study presented in the third paper, building segmentation from high-resolution satellite and UAV images was performed using the Inria Aerial Image Labeling Dataset as training and test set, and the Massachusetts Buildings Dataset and the Syedra Archaeological Site Dataset as additional test sets. Within the scope of the study, a total of twelve CNN models (U-Net, FPN, and LinkNet architectures utilizing EfficientNet-B5 backbone, original U-Net, SegNet, FCN, and six different Residual U-Net approaches that use six different residual block designs) were generated, trained, tested, and comparisons between them were realized. According to the results, Residual-2 U-Net approach, which uses residual block-2 design, performed best in all evaluation metrics on the prepared Inria test set. On the Syedra test set, the LinkNet EfficientNet-B5 model, and on the Massachusetts test set, the Residual-4 U-Net model provided the best results. Since the features in the Syedra Archaeological Site Dataset have different characteristics from both the Inria and the Massachusetts datasets, it

could be said that the LinkNet EfficientNet-B5 model the highest generalization ability among all evaluated approaches. Nevertheless, the Inria dataset and the Massachusetts dataset have similar features but collected using different kind of sensors, it can be mentioned that the Residual-4 U-Net model has a high generalization ability across different sensors.

Building segmentation from high-resolution aerial images was performed using generated datasets from the Massachusetts Buildings Dataset and the Inria Aerial Image Labeling Dataset within the context of the study presented in paper four. Ten CNN-based and Transformer-based models (RIU-Net, U-Net, Residual U-Net, Attention Residual U-Net, Inception backboned U-Net, Inception-ResNet backboned U-Net, Xception backboned U-Net, MobileNet backboned U-Net, Trans U-Net, Swin U-Net) were generated, and their performance comparisons were subsequently made. On the Inria dataset, the proposed RIU-Net model achieved the most successful results among all the models, according to the evaluation metrics. On the datasets generated using the Massachusetts Buildings Dataset, Residual U-Net and Attention Residual U-Net, along with the Trans U-Net, performed better. When the results on the Small and Large Massachusetts datasets are compared, it is seen that the performance of the RIU-Net model significantly increases compared to other models, as the dataset size grows. Additionally, when the training process is examined, it has been observed that extending the training period may increase the performance of this proposed model. When all the results are examined, it can be interpreted that the proposed RIU-Net approach is expressively successful on datasets containing complex buildings with different characteristics. On the other hand, on datasets containing buildings with mostly similar characteristics, it can be predicted that the proposed model will catch up with, and perhaps even exceed, the performance of other models when sufficient data and longer periods of training times are provided.

This thesis was composed of four journal articles regarding with the building segmentation from high-resolution images. Regarding the scope of the thesis, all the experiments and studies presented in the thesis evaluate different deep learning-based CNN and Transformer approaches with different structures and methodologies, realize comparisons between them, and propose approaches that may be useful for the

building segmentation task, using multiple different high-resolution aerial image datasets. Along with all the studies carried out, examining the contributions of the technologies that are constantly developing in the field to the aforementioned problem and suggesting new approaches form the basis of the future studies planned.

REFERENCES

[1] **Chen, J.**, **Jiang, Y.**, **Luo, L.**, **Gu, Y. and Wu, K.** (2021). Building footprint generation by integrating U-Net with deepened space module, *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp.3847–3851.

[2] **Wu, G.**, **Shao, X.**, **Guo, Z.**, **Chen, Q.**, **Yuan, W.**, **Shi, X.**, **Xu, Y. and Shibasaki, R.** (2018). Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks, *Remote Sensing*, *10*(3), 407.

[3] **Chen, Q.**, **Wang, L.**, **Wu, Y.**, **Wu, G.**, **Guo, Z. and Waslander, S.L.** (2019). Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings, *ISPRS Journal of Photogrammetry and Remote Sensing*, *147*, 42–55.

[4] **Zhang, Y.**, **Gong, W.**, **Sun, J. and Li, W.** (2019). Web-Net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries, *Remote Sensing*, *11*(16), 1897.

[5] **Li, W.**, **He, C.**, **Fang, J.**, **Zheng, J.**, **Fu, H. and Yu, L.** (2019). Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data, *Remote Sensing*, *11*(4), 403.

[6] **Ozturk, O.**, **Sariturk, B. and Seker, D.Z.** (2020). Comparison of Fully Convolutional Networks (FCN) and U-Net for Road Segmentation from High Resolution Imageries, *International journal of environment and geoinformatics*, *7*(3), 272–279.

[7] **Yu, M.**, **Chen, X.**, **Zhang, W. and Liu, Y.** (2022). AGs-Unet: Building Extraction Model for High Resolution Remote Sensing Images Based on Attention Gates U Network, *Sensors*, *22*(8), 2932.

[8] **Wang, H. and Miao, F.** (2022). Building extraction from remote sensing images using deep residual U-Net, *European Journal of Remote Sensing*, *55*(1), 71–85.

[9] **Sun, X.**, **Zhao, W.**, **Maretto, R.V. and Persello, C.** (2021). Building outline extraction from aerial imagery and digital surface model with a frame field learning framework, *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 487–493.

[10] **Chen, K.**, **Zou, Z. and Shi, Z.** (2021). Building extraction from remote sensing images with sparse token transformers, *Remote Sensing*, *13*(21), 4441.

[11] **Liu, Y.**, **So, E.**, **Li, Z.**, **Su, G.**, **Gross, L.**, **Li, X.**, **Qi, W.**, **Yang, F.**, **Fu, B.**, **Yalikun, A.** *et al.* (2020). Scenario-based seismic vulnerability and hazard analyses to help direct disaster risk reduction in rural Weinan, China, *International Journal of Disaster Risk Reduction*, *48*, 101577.

[12] **Li, Q.**, **Shi, Y. and Zhu, X.X.** (2022). Semi-supervised building footprint generation with feature and output consistency training, *IEEE Transactions on Geoscience and Remote Sensing*.

[13] **Bakirman, T.**, **Komurcu, I. and Sertel, E.** (2022). Comparative analysis of deep learning based building extraction methods with the new VHR Istanbul dataset, *Expert Systems with Applications*, *202*, 117346.

[14] **Ok, A.O.** (2013). Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts, *ISPRS journal of photogrammetry and remote sensing*, *86*, 21–40.

[15] **Moghalles, K.**, **Li, H.C. and Alazeb, A.** (2022). Weakly Supervised Building Semantic Segmentation Based on Spot-Seeds and Refinement Process, *Entropy*, *24*(5), 741.

[16] **Wang, L.**, **Fang, S.**, **Meng, X. and Li, R.** (2022). Building extraction with vision transformer, *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–11.

[17] **Yuan, W. and Xu, W.** (2021). MSST-Net: A Multi-Scale Adaptive Network for Building Extraction from Remote Sensing Images Based on Swin Transformer, *Remote Sensing*, *13*(23), 4743.

[18] **Zhu, Y.**, **Liang, Z.**, **Yan, J.**, **Chen, G. and Wang, X.** (2021). ED-Net: Automatic building extraction from high-resolution aerial images with boundary information, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 4595–4606.

[19] **Krizhevsky, A.**, **Sutskever, I. and Hinton, G.E.** (2012). Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, *25*.

[20] **LeCun, Y.**, **Bottou, L.**, **Bengio, Y. and Haffner, P.** (1998). Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, *86*(11), 2278–2324.

[21] **Simonyan, K. and Zisserman, A.** (2014). Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.

[22] **He, K.**, **Zhang, X.**, **Ren, S. and Sun, J.** (2016). Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.770–778.

[23] **Szegedy, C.**, **Liu, W.**, **Jia, Y.**, **Sermanet, P.**, **Reed, S.**, **Anguelov, D.**, **Erhan, D.**, **Vanhoucke, V. and Rabinovich, A.** (2015). Going deeper with convolutions, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.1–9.

[24] **Long, J.**, **Shelhamer, E. and Darrell, T.** (2015). Fully convolutional networks for semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.3431–3440.

[25] **Khan, S.D.**, **Alarabi, L. and Basalamah, S.** (2022). An Encoder–Decoder Deep Learning Framework for Building Footprints Extraction from Aerial Imagery, *Arabian Journal for Science and Engineering*, 1–12.

[26] **Ronneberger, O.**, **Fischer, P. and Brox, T.** (2015). U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical image computing and computer-assisted intervention*, Springer, pp.234–241.

[27] **Lin, T.Y.**, **Dollár, P.**, **Girshick, R.**, **He, K.**, **Hariharan, B. and Belongie, S.** (2017). Feature pyramid networks for object detection, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.2117–2125.

[28] **Badrinarayanan, V.**, **Kendall, A. and Cipolla, R.** (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2481–2495.

[29] **Vaswani, A.**, **Shazeer, N.**, **Parmar, N.**, **Uszkoreit, J.**, **Jones, L.**, **Gomez, A.N.**, **Kaiser, Ł. and Polosukhin, I.** (2017). Attention is all you need, *Advances in neural information processing systems*, *30*.

[30] **Oktay, O.**, **Schlemper, J.**, **Folgoc, L.L.**, **Lee, M.**, **Heinrich, M.**, **Misawa, K.**, **Mori, K.**, **McDonagh, S.**, **Hammerla, N.Y.**, **Kainz, B.** *et al.* (2018). Attention u-net: Learning where to look for the pancreas, *arXiv preprint arXiv:1804.03999*.

[31] **Sariturk, B. and Seker, D.Z.** (2022). Comparison of Residual and Dense Neural Network Approaches for Building Extraction from High-Resolution Aerial Images, *Advances in Space Research*.

[32] **Sariturk, B.**, **Seker, D.Z.**, **Ozturk, O. and Bayram, B.** (2022). Performance evaluation of shallow and deep CNN architectures on building segmentation from high-resolution images, *Earth Science Informatics*, *15*(3), 1801–1823.

[33] **Ozturk, O.**, **Isik, M.S.**, **Sariturk, B. and Seker, D.Z.** (2022). Generation of Istanbul road data set using Google Map API for deep learning-based segmentation, *International Journal of Remote Sensing*, *43*(8), 2793–2812.

[34] **Dosovitskiy, A.**, **Beyer, L.**, **Kolesnikov, A.**, **Weissenborn, D.**, **Zhai, X.**, **Unterthiner, T.**, **Dehghani, M.**, **Minderer, M.**, **Heigold, G.**, **Gelly, S.** *et al.* (2020). An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929*.

[35] **Strudel, R.**, **Garcia, R.**, **Laptev, I. and Schmid, C.** (2021). Segmenter: Transformer for semantic segmentation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.7262–7272.

[36] **Xie, E.**, **Wang, W.**, **Yu, Z.**, **Anandkumar, A.**, **Alvarez, J.M. and Luo, P.** (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers, *Advances in Neural Information Processing Systems*, *34*, 12077–12090.

[37] **Zheng, S.**, **Lu, J.**, **Zhao, H.**, **Zhu, X.**, **Luo, Z.**, **Wang, Y.**, **Fu, Y.**, **Feng, J.**, **Xiang, T.**, **Torr, P.H.** *et al.* (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.6881–6890.

[38] **Hong, D.**, **Han, Z.**, **Yao, J.**, **Gao, L.**, **Zhang, B.**, **Plaza, A. and Chanussot, J.** (2021). SpectralFormer: Rethinking hyperspectral image classification with transformers, *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–15.

[39] **He, X.**, **Chen, Y. and Lin, Z.** (2021). Spatial-spectral transformer for hyperspectral image classification, *Remote Sensing*, *13*(3), 498.

[40] **Wang, L.**, **Li, R.**, **Wang, D.**, **Duan, C.**, **Wang, T. and Meng, X.** (2021). Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images, *Remote Sensing*, *13*(16), 3065.

[41] **Wang, L.**, **Li, R.**, **Zhang, C.**, **Fang, S.**, **Duan, C.**, **Meng, X. and Atkinson, P.M.** (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery, *ISPRS Journal of Photogrammetry and Remote Sensing*, *190*, 196–214.

[42] **Chen, H.**, **Qi, Z. and Shi, Z.** (2021). Remote sensing image change detection with transformers, *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–14.

[43] **Zhang, C.**, **Wang, L.**, **Cheng, S. and Li, Y.** (2022). SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection, *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–13.

[44] **Chen, X.**, **Qiu, C.**, **Guo, W.**, **Yu, A.**, **Tong, X. and Schmitt, M.** (2022). Multiscale feature learning by transformer for building extraction from satellite images, *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

[45] **Liu, Z.**, **Lin, Y.**, **Cao, Y.**, **Hu, H.**, **Wei, Y.**, **Zhang, Z.**, **Lin, S. and Guo, B.** (2021). Swin transformer: Hierarchical vision transformer using shifted windows, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.10012–10022.

[46] **Zhang, Y.**, **Liu, H. and Hu, Q.** (2021). Transfuse: Fusing transformers and cnns for medical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp.14–24.

[47] **Chen, J.**, **Lu, Y.**, **Yu, Q.**, **Luo, X.**, **Adeli, E.**, **Wang, Y.**, **Lu, L.**, **Yuille, A.L. and Zhou, Y.** (2021). Transunet: Transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306*.

[48] **Kim, J.H.**, **Lee, H.**, **Hong, S.J.**, **Kim, S.**, **Park, J.**, **Hwang, J.Y. and Choi, J.P.** (2018). Objects segmentation from high-resolution aerial images using U-Net with pyramid pooling layers, *IEEE Geoscience and Remote Sensing Letters*, *16*(1), 115–119.

[49] **Ma, L.**, **Li, M.**, **Ma, X.**, **Cheng, L.**, **Du, P. and Liu, Y.** (2017). A review of supervised object-based land-cover image classification, *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 277–293.

[50] **Cheng, G. and Han, J.** (2016). A survey on object detection in optical remote sensing images, *ISPRS Journal of Photogrammetry and Remote sensing*, *117*, 11–28.

[51] **Ghanea, M.**, **Moallem, P. and Momeni, M.** (2016). Building extraction from high-resolution satellite images in urban areas: Recent methods and strategies against significant challenges, *International journal of remote sensing*, *37*(21), 5234–5248.

[52] **Wen, Q.**, **Jiang, K.**, **Wang, W.**, **Liu, Q.**, **Guo, Q.**, **Li, L. and Wang, P.** (2019). Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network, *Sensors*, *19*(2), 333.

[53] **Hui, J.**, **Du, M.**, **Ye, X.**, **Qin, Q. and Sui, J.** (2018). Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network, *IEEE Geoscience and Remote Sensing Letters*, *16*(5), 786–790.

[54] **Li, L.**, **Liang, J.**, **Weng, M. and Zhu, H.** (2018). A multiple-feature reuse network to extract buildings from remote sensing imagery, *Remote Sensing*, *10*(9), 1350.

[55] **Demir, I.**, **Koperski, K.**, **Lindenbaum, D.**, **Pang, G.**, **Huang, J.**, **Basu, S.**, **Hughes, F.**, **Tuia, D. and Raskar, R.** (2018). Deepglobe 2018: A challenge to parse the earth through satellite images, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.172–181.

[56] **Xu, X.**, **Li, W.**, **Ran, Q.**, **Du, Q.**, **Gao, L. and Zhang, B.** (2017). Multisource remote sensing data classification based on convolutional neural network, *IEEE Transactions on Geoscience and Remote Sensing*, *56*(2), 937–949.

[57] **Blaschke, T.** (2010). Object based image analysis for remote sensing, *ISPRS journal of photogrammetry and remote sensing*, *65*(1), 2–16.

[58] **Ji, S.**, **Wei, S. and Lu, M.** (2018). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set, *IEEE Transactions on Geoscience and Remote Sensing*, *57*(1), 574–586.

[59] **Sirmacek, B. and Unsalan, C.** (2008). Building detection from aerial images using invariant color features and shadow information, *2008 23rd international symposium on computer and information sciences*, IEEE, pp.1–5.

[60] **Zhang, Y.** (1999). Optimisation of building detection in satellite images by combining multispectral classification and texture filtering, *ISPRS journal of photogrammetry and remote sensing*, *54*(1), 50–60.

[61] **Li, Y. and Wu, H.** (2008). Adaptive building edge detection by combining LiDAR data and aerial images, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *37*(Part B1), 197–202.

[62] **Liow, Y.T. and Pavlidis, T.** (1990). Use of shadows for extracting buildings in aerial images, *Computer Vision, Graphics, and Image Processing*, *49*(2), 242–277.

[63] **Zhong, C.**, **Xu, Q.**, **Yang, F. and Hu, L.** (2015). Building change detection for high-resolution remotely sensed images based on a semantic dependency, *2015 IEEE international geoscience and remote sensing symposium (IGARSS)*, IEEE, pp.3345–3348.

[64] **Huang, G.**, **Liu, Z.**, **Van Der Maaten, L. and Weinberger, K.Q.** (2017). Densely connected convolutional networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.4700–4708.

[65] **Li, X.**, **Chen, H.**, **Qi, X.**, **Dou, Q.**, **Fu, C.W. and Heng, P.A.** (2018). H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, *IEEE transactions on medical imaging*, *37*(12), 2663–2674.

[66] **Zhang, C.**, **Sargent, I.**, **Pan, X.**, **Li, H.**, **Gardiner, A.**, **Hare, J. and Atkinson, P.M.** (2018). An object-based convolutional neural network (OCNN) for urban land use classification, *Remote sensing of environment*, *216*, 57–70.

[67] **Mnih, V. and Hinton, G.E.** (2012). Learning to label aerial images from noisy data, *Proceedings of the 29th International conference on machine learning (ICML-12)*, pp.567–574.

[68] **Mnih, V. and Hinton, G.E.** (2010). Learning to detect roads in high-resolution aerial images, *European conference on computer vision*, Springer, pp.210–223.

[69] **Yang, X.**, **Ye, Y.**, **Li, X.**, **Lau, R.Y.**, **Zhang, X. and Huang, X.** (2018). Hyperspectral image classification with deep learning models, *IEEE Transactions on Geoscience and Remote Sensing*, *56*(9), 5408–5423.

[70] **Zhang, L.**, **Zhang, L. and Du, B.** (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art, *IEEE Geoscience and remote sensing magazine*, *4*(2), 22–40.

[71] **Cheng, G.**, **Wang, Y.**, **Xu, S.**, **Wang, H.**, **Xiang, S. and Pan, C.** (2017). Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network, *IEEE Transactions on Geoscience and Remote Sensing*, *55*(6), 3322–3337.

[72] **Yang, X.**, **Li, X.**, **Ye, Y.**, **Lau, R.Y.**, **Zhang, X. and Huang, X.** (2019). Road detection and centerline extraction via deep recurrent convolutional neural network U-Net, *IEEE Transactions on Geoscience and Remote Sensing*, *57*(9), 7209–7220.

[73] **Cordts, M.**, **Omran, M.**, **Ramos, S.**, **Rehfeld, T.**, **Enzweiler, M.**, **Benenson, R.**, **Franke, U.**, **Roth, S. and Schiele, B.** (2016). The cityscapes dataset for semantic urban scene understanding, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.3213–3223.

[74] **Deng, J.**, **Dong, W.**, **Socher, R.**, **Li, L.J.**, **Li, K. and Fei-Fei, L.** (2009). Imagenet: A large-scale hierarchical image database, *2009 IEEE conference on computer vision and pattern recognition*, Ieee, pp.248–255.

[75] **Lin, T.Y.**, **Maire, M.**, **Belongie, S.**, **Hays, J.**, **Perona, P.**, **Ramanan, D.**, **Dollár, P. and Zitnick, C.L.** (2014). Microsoft coco: Common objects in context, *European conference on computer vision*, Springer, pp.740–755.

[76] **LeCun, Y.**, **Boser, B.**, **Denker, J.S.**, **Henderson, D.**, **Howard, R.E.**, **Hubbard, W. and Jackel, L.D.** (1989). Backpropagation applied to handwritten zip code recognition, *Neural computation*, *1*(4), 541–551.

[77] **Kim, D.E. and Gofman, M.** (2018). Comparison of shallow and deep neural networks for network intrusion detection, *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, pp.204–208.

[78] **Xu, Y.**, **Wu, L.**, **Xie, Z. and Chen, Z.** (2018). Building extraction in very high resolution remote sensing imagery using deep learning and guided filters, *Remote Sensing*, *10*(1), 144.

[79] **Audebert, N.**, **Le Saux, B. and Lefèvre, S.** (2017). Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images, *Remote Sensing*, *9*(4), 368.

[80] **Cao, R.**, **Zhu, J.**, **Tu, W.**, **Li, Q.**, **Cao, J.**, **Liu, B.**, **Zhang, Q. and Qiu, G.** (2018). Integrating aerial and street view images for urban land use classification, *Remote Sensing*, *10*(10), 1553.

[81] **Bai, Y.**, **Mas, E. and Koshimura, S.** (2018). Towards operational satellite-based damage-mapping using u-net convolutional network: A case study of 2011 tohoku earthquake-tsunami, *Remote Sensing*, *10*(10), 1626.

[82] **Ye, Z.**, **Fu, Y.**, **Gan, M.**, **Deng, J.**, **Comber, A. and Wang, K.** (2019). Building extraction from very high resolution aerial imagery using joint attention deep neural network, *Remote Sensing*, *11*(24), 2970.

[83] **Fu, G.**, **Liu, C.**, **Zhou, R.**, **Sun, T. and Zhang, Q.** (2017). Classification for high resolution remote sensing imagery using a fully convolutional network, *Remote Sensing*, *9*(5), 498.

[84] **Zhao, W.**, **Du, S.**, **Wang, Q. and Emery, W.J.** (2017). Contextually guided very-high-resolution imagery classification with semantic segments, *ISPRS journal of Photogrammetry and Remote Sensing*, *132*, 48–60.

[85] **Zhao, W. and Du, S.** (2016). Learning multiscale and deep representations for classifying remotely sensed imagery, *ISPRS Journal of Photogrammetry and Remote Sensing*, *113*, 155–165.

[86] **Alshehhi, R.**, **Marpu, P.R.**, **Woon, W.L. and Dalla Mura, M.** (2017). Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks, *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 139–149.

[87] **Noh, H.**, **Hong, S. and Han, B.** (2015). Learning deconvolution network for semantic segmentation, *Proceedings of the IEEE international conference on computer vision*, pp.1520–1528.

[88] **Bischke, B.**, **Helber, P.**, **Folz, J.**, **Borth, D. and Dengel, A.** (2019). Multi-task learning for segmentation of building footprints with deep neural networks, *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp.1480–1484.

[89] **Ivanovsky, L.**, **Khryashchev, V.**, **Pavlov, V. and Ostrovskaya, A.** (2019). Building detection on aerial images using U-NET neural networks, *2019 24th Conference of Open Innovations Association (FRUCT)*, IEEE, pp.116–122.

[90] **Liu, W.**, **Yang, M.**, **Xie, M.**, **Guo, Z.**, **Li, E.**, **Zhang, L.**, **Pei, T. and Wang, D.** (2019). Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network, *Remote Sensing*, *11*(24), 2912.

[91] **Maggiori, E.**, **Tarabalka, Y.**, **Charpiat, G. and Alliez, P.** (2017). Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark, *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, pp.3226–3229.

[92] **Yuan, J.** (2017). Learning building extraction in aerial scenes with convolutional networks, *IEEE transactions on pattern analysis and machine intelligence*, *40*(11), 2793–2798.

[93] **Milosavljević, A.** (2020). Automated processing of remote sensing imagery using deep semantic segmentation: A building footprint extraction case, *ISPRS International Journal of Geo-Information*, *9*(8), 486.

[94] **Shi, Y.**, **Li, Q. and Zhu, X.X.** (2020). Building segmentation through a gated graph convolutional neural network with deep structured feature embedding, *ISPRS Journal of Photogrammetry and Remote Sensing*, *159*, 184–197.

[95] **Lei, F.**, **Liu, X.**, **Dai, Q. and Ling, B.W.K.** (2020). Shallow convolutional neural network for image classification, *SN Applied Sciences*, *2*(1), 1–8.

[96] **Gorokhovatskyi, O. and Peredrii, O.** (2018). Shallow convolutional neural networks for pattern recognition problems, *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, IEEE, pp.459–463.

[97] **Nogueira, K.**, **Miranda, W.O. and Dos Santos, J.A.** (2015). Improving spatial feature representation from aerial scenes by using convolutional networks, *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, IEEE, pp.289–296.

[98] **Nogueira, K.**, **Penatti, O.A. and Dos Santos, J.A.** (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification, *Pattern Recognition*, *61*, 539–556.

[99] **Girisha, S.**, **Pai, M.M.**, **Verma, U. and Pai, R.M.** (2019). Performance analysis of semantic segmentation algorithms for finely annotated new uav aerial video dataset (manipaluavid), *IEEE Access*, *7*, 136239–136253.

[100] **Chen, L.C.**, **Papandreou, G.**, **Kokkinos, I.**, **Murphy, K. and Yuille, A.L.** (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834–848.

[101] **Maggiori, E.**, **Tarabalka, Y.**, **Charpiat, G. and Alliez, P.** (2016). Convolutional neural networks for large-scale remote-sensing image classification, *IEEE Transactions on geoscience and remote sensing*, *55*(2), 645–657.

[102] **Inria**. `https://project.inria.fr/aerialimagelabeling/`, date retrieved: 12.05.2021.

[103] **Ohleyer, S.** (2018). https://project.inria.fr/aerialimagelabeling/files/2018/01/fp_ohleyer_compressed.pdf, date retrieved: 12.05.2021.

[104] **Mnih, V.** (2013). *Machine learning for aerial image labeling*, University of Toronto, Canada, (Doctoral dissertation).

[105] **Deepsense.ai**. https://deepsense.ai/deep-learning-for-satellite-imagery-via-image-segmentation, date retrieved: 12.05.2021.

[106] **Nagi, J.**, **Ducatelle, F.**, **Di Caro, G.A.**, **Cireşan, D.**, **Meier, U.**, **Giusti, A.**, **Nagi, F.**, **Schmidhuber, J. and Gambardella, L.M.** (2011). Max-pooling convolutional neural networks for vision-based hand gesture recognition, *2011 IEEE international conference on signal and image processing applications (ICSIPA)*, IEEE, pp.342–347.

[107] **Dumoulin, V. and Visin, F.** (2016). A guide to convolution arithmetic for deep learning, *arXiv preprint arXiv:1603.07285*.

[108] **Ghorbanzadeh, O.**, **Blaschke, T.**, **Gholamnia, K.**, **Meena, S.R.**, **Tiede, D. and Aryal, J.** (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection, *Remote Sensing*, *11*(2), 196.

[109] **Bäuerle, A.**, **van Onzenoodt, C. and Ropinski, T.** (2021). Net2vis–a visual grammar for automatically generating publication-tailored cnn architecture visualizations, *IEEE transactions on visualization and computer graphics*, *27*(6), 2980–2991.

[110] **Patterson, J. and Gibson, A.** (2017). *Deep learning: A practitioner's approach*, O'Reilly Media, Inc.

[111] **Scikit-Learn**. *3.1. Cross-validation: evaluating estimator performance*, https://scikit-learn.org/stable/modules/cross_validation.html, date retrieved: 12.05.2021.

[112] **Kingma, D.P. and Ba, J.** (2014). Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*.

[113] **Rosebrock, A.** (2017). *Deep learning for computer vision with python*, PyImageSearch.

[114] **Rezatofighi, H.**, **Tsoi, N.**, **Gwak, J.**, **Sadeghian, A.**, **Reid, I. and Savarese, S.** (2019). Generalized intersection over union: A metric and a loss for bounding box regression, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.658–666.

[115] **Ye, H.**, **Liu, S.**, **Jin, K. and Cheng, H.** (2021). CT-UNet: An improved neural network based on U-Net for building segmentation in remote sensing images, *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, pp.166–172.

[116] **Rathore, M.M.**, **Ahmad, A.**, **Paul, A. and Rho, S.** (2016). Urban planning and building smart cities based on the internet of things using big data analytics, *Computer networks*, *101*, 63–80.

[117] **Liu, S.**, **Ye, H.**, **Jin, K. and Cheng, H.** (2021). CT-UNet: Context-Transfer-UNet for Building Segmentation in Remote Sensing Images, *Neural Processing Letters*, *53*(6), 4257–4277.

[118] **Sun, Y.**, **Zhang, X.**, **Huang, J.**, **Wang, H. and Xin, Q.** (2020). Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning, *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

[119] **Liu, Y.**, **Zhou, J.**, **Qi, W.**, **Li, X.**, **Gross, L.**, **Shao, Q.**, **Zhao, Z.**, **Ni, L.**, **Fan, X. and Li, Z.** (2020). ARC-Net: An efficient network for building extraction from high-resolution aerial images, *IEEE Access*, *8*, 154997–155010.

[120] **Boonpook, W.**, **Tan, Y.**, **Ye, Y.**, **Torteeka, P.**, **Torsri, K. and Dong, S.** (2018). A deep learning approach on building detection from unmanned aerial vehicle-based images in riverbank monitoring, *Sensors*, *18*(11), 3921.

[121] **Tian, Q.**, **Zhao, Y.**, **Qin, K.**, **Li, Y. and Chen, X.** (2021). Dense feature pyramid fusion deep network for building segmentation in remote sensing image, *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*, volume11763, SPIE, pp.1361–1366.

[122] **Celik, T.** (2009). Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering, *IEEE geoscience and remote sensing letters*, *6*(4), 772–776.

[123] **Inglada, J.** (2007). Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features, *ISPRS journal of photogrammetry and remote sensing*, *62*(3), 236–248.

[124] **Dong, Y.**, **Du, B. and Zhang, L.** (2015). Target detection based on random forest metric learning, *IEEE Journal of selected topics in applied earth observations and remote sensing*, *8*(4), 1830–1838.

[125] **Li, E.**, **Femiani, J.**, **Xu, S.**, **Zhang, X. and Wonka, P.** (2015). Robust rooftop extraction from visible band images using higher order CRF, *IEEE Transactions on Geoscience and Remote Sensing*, *53*(8), 4483–4495.

[126] **Aytekin, Ö.**, **Zöngür, U. and Halici, U.** (2012). Texture-based airport runway detection, *IEEE Geoscience and Remote Sensing Letters*, *10*(3), 471–475.

[127] **Yang, H.L.**, **Yuan, J.**, **Lunga, D.**, **Laverdiere, M.**, **Rose, A. and Bhaduri, B.** (2018). Building extraction at scale using convolutional neural network: Mapping of the united states, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(8), 2600–2614.

[128] **Sariturk, B.**, **Bayram, B.**, **Duran, Z. and Seker, D.Z.** (2020). Feature extraction from satellite images using segnet and fully convolutional networks (FCN), *International Journal of Engineering and Geosciences*, *5*(3), 138–143.

[129] **Liu, X.**, **Deng, W. and Liu, Y.** (2021). Application of hybrid network of UNet and feature pyramid network in spine segmentation, *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, IEEE, pp.1–6.

[130] **Peng, F.**, **Miao, Z.**, **Li, F. and Li, Z.** (2021). S-FPN: A shortcut feature pyramid network for sea cucumber detection in underwater images, *Expert Systems with Applications*, *182*, 115306.

[131] **Chen, X.**, **Yao, L. and Zhang, Y.** (2020). Residual attention u-net for automated multi-class segmentation of covid-19 chest ct images, *arXiv preprint arXiv:2004.05645*.

[132] **ISPRS**. `https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx`, date retrieved: 11.03.2022.

[133] **Lei, M.**, **Rao, Z.**, **Wang, H.**, **Chen, Y.**, **Zou, L. and Yu, H.** (2021). Maceral groups analysis of coal based on semantic segmentation of photomicrographs via the improved U-net, *Fuel*, *294*, 120475.

[134] **Punn, N.S. and Agarwal, S.** (2022). Modality specific U-Net variants for biomedical image segmentation: a survey, *Artificial Intelligence Review*, 1–45.

[135] **Cenggoro, T.W.**, **Aslamiah, A.H. and Yunanto, A.** (2019). Feature pyramid networks for crowd counting, *Procedia Computer Science*, *157*, 175–182.

[136] **Ren, Y.**, **Huo, W.**, **Pei, J.**, **Huang, Y. and Yang, J.** (2021). Automatic Modulation Recognition for Overlapping Radar Signals based on Multi-Domain SE-ResNeXt, *2021 IEEE Radar Conference (RadarConf21)*, IEEE, pp.1–6.

[137] **Ma, L.**, **Shuai, R.**, **Ran, X.**, **Liu, W. and Ye, C.** (2020). Combining DC-GAN with ResNet for blood cell image classification, *Medical & biological engineering & computing*, *58*(6), 1251–1264.

[138] **Lu, Z.**, **Bai, Y.**, **Chen, Y.**, **Su, C.**, **Lu, S.**, **Zhan, T.**, **Hong, X. and Wang, S.** (2020). The classification of gliomas based on a pyramid dilated convolution resnet model, *Pattern Recognition Letters*, *133*, 173–179.

[139] **Xie, S.**, **Girshick, R.**, **Dollár, P.**, **Tu, Z. and He, K.** (2017). Aggregated residual transformations for deep neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.1492–1500.

[140] **Zhang, G.**, **Yang, Z.**, **Gong, L.**, **Jiang, S.**, **Wang, L. and Zhang, H.** (2020). Classification of lung nodules based on CT images using squeeze-and-excitation network and aggregated residual transformations, *La radiologia medica*, *125*(4), 374–383.

[141] **Hu, J.**, **Shen, L. and Sun, G.** (2018). Squeeze-and-excitation networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.7132–7141.

[142] **Zhang, C.**, **Benz, P.**, **Argaw, D.M.**, **Lee, S.**, **Kim, J.**, **Rameau, F.**, **Bazin, J.C. and Kweon, I.S.** (2021). Resnet or densenet? introducing dense shortcuts to resnet, *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp.3550–3559.

[143] **Hemalatha, J.**, **Roseline, S.A.**, **Geetha, S.**, **Kadry, S. and Damaševičius, R.** (2021). An efficient densenet-based deep learning model for malware detection, *Entropy*, *23*(3), 344.

[144] **Fang, Z.**, **Chen, Y.**, **Nie, D.**, **Lin, W. and Shen, D.** (2019). Rca-u-net: Residual channel attention u-net for fast tissue quantification in magnetic resonance fingerprinting, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp.101–109.

[145] **Zhang, J.**, **Lv, X.**, **Zhang, H. and Liu, B.** (2020). AResU-Net: Attention residual U-Net for brain tumor segmentation, *Symmetry*, *12*(5), 721.

[146] **He, K.**, **Zhang, X.**, **Ren, S. and Sun, J.** (2016). Identity mappings in deep residual networks, *European conference on computer vision*, Springer, pp.630–645.

[147] **Naranjo-Alcazar, J.**, **Perez-Castanos, S.**, **Martin-Morato, I.**, **Zuccarello, P. and Cobos, M.** (2019). On the performance of residual block design alternatives in convolutional neural networks for end-to-end audio classification, *arXiv preprint arXiv:1906.10891*.

[148] **Sun, J.**, **Li, J. and Liu, L.** (2021). Semantic segmentation of brain tumor with nested residual attention networks, *Multimedia Tools and Applications*, *80*(26), 34203–34220.

[149] **Wang, X.**, **Girshick, R.**, **Gupta, A. and He, K.** (2018). Non-local neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.7794–7803.

[150] **Iakubovskii, P.** (2019). *Segmentation Models*, `https://github.com/qubvel/segmentation_models`.

[151] **tf.keras.callbacks.ReduceLROnPlateau**. `https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/ReduceLROnPlateau`, date retrieved: 14.08.2021.

[152] **tf.keras.callbacks.EarlyStopping**. https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping, date retrieved: 14.08.2021.

[153] **tf.keras.callbacks.ModelCheckpoint**. https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/ModelCheckpoint, date retrieved: 14.08.2021.

[154] **Géron, A.** (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, Inc.

[155] **Chen, M.**, **Wu, J.**, **Liu, L.**, **Zhao, W.**, **Tian, F.**, **Shen, Q.**, **Zhao, B. and Du, R.** (2021). DR-Net: An improved network for building extraction from high resolution remote sensing image, *Remote Sensing*, *13*(2), 294.

[156] **Hu, Q.**, **Zhen, L.**, **Mao, Y.**, **Zhou, X. and Zhou, G.** (2021). Automated building extraction using satellite remote sensing imagery, *Automation in Construction*, *123*, 103509.

[157] **Ferraioli, G.** (2009). Multichannel InSAR building edge detection, *IEEE Transactions on Geoscience and Remote Sensing*, *48*(3), 1224–1231.

[158] **Dunaeva, A.V. and Kornilov, F.A.** (2017). Specific shape building detection from aerial imagery in infrared range, *Vestnik Yuzhno-Ural'skogo Gosudarstvennogo Universiteta. Seriya" Vychislitelnaya Matematika i Informatika"*, *6*(3), 84–100.

[159] **Zhong, S.h.**, **Huang, J.j. and Xie, W.x.** (2008). A new method of building detection from a single aerial photograph, *2008 9th international conference on signal processing*, IEEE, pp.1219–1222.

[160] **Awrangjeb, M.**, **Zhang, C. and Fraser, C.S.** (2011). Improved building detection using texture information, *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, *38*, 143–148.

[161] **Chen, D.**, **Shang, S. and Wu, C.** (2014). Shadow-based Building Detection and Segmentation in High-resolution Remote Sensing Image., *J. Multim.*, *9*(1), 181–188.

[162] **Qin, X.**, **He, S.**, **Yang, X.**, **Dehghan, M.**, **Qin, Q. and Martin, J.** (2018). Accurate outline extraction of individual building from very high-resolution optical images, *IEEE Geoscience and Remote Sensing Letters*, *15*(11), 1775–1779.

[163] **Erdem, F.**, **Bayram, B.**, **Bakirman, T.**, **Bayrak, O.C. and Akpinar, B.** (2021). An ensemble deep learning based shoreline segmentation approach (WaterNet) from Landsat 8 OLI images, *Advances in Space Research*, *67*(3), 964–974.

[164] **Chaurasia, A. and Culurciello, E.** (2017). Linknet: Exploiting encoder representations for efficient semantic segmentation, *2017 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, pp.1–4.

[165] **Tan, M. and Le, Q.** (2019). Efficientnet: Rethinking model scaling for convolutional neural networks, *International conference on machine learning*, PMLR, pp.6105–6114.

[166] **Dai, W.**, **Dai, C.**, **Qu, S.**, **Li, J. and Das, S.** (2017). Very deep convolutional neural networks for raw waveforms, *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pp.421–425.

[167] **Can, B.**, (2017). Syedra ve Teritoryumu, Barış Salman Anı Kitabı, Ege Yayınları, Istanbul, pp.263–267.

[168] **Iglovikov, V.**, **Seferbekov, S.**, **Buslaev, A. and Shvets, A.** (2018). Ternausnetv2: Fully convolutional network for instance segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.233–237.

[169] **Jadon, S.** (2020). A survey of loss functions for semantic segmentation, *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, pp.1–7.

[170] **Sudre, C.H.**, **Li, W.**, **Vercauteren, T.**, **Ourselin, S. and Jorge Cardoso, M.**, (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, pp.240–248.

[171] **Tang, Y.**, **Li, L.**, **Wang, C.**, **Chen, M.**, **Feng, W.**, **Zou, X. and Huang, K.** (2019). Real-time detection of surface deformation and strain in recycled aggregate concrete-filled steel tubular columns via four-ocular vision, *Robotics and Computer-Integrated Manufacturing*, *59*, 36–46.

[172] **Deng, W.**, **Shi, Q. and Li, J.** (2021). Attention-gate-based encoder–decoder network for automatical building extraction, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 2611–2620.

[173] **Guo, H.**, **Shi, Q.**, **Du, B.**, **Zhang, L.**, **Wang, D. and Ding, H.** (2020). Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, *59*(5), 4287–4306.

[174] **Pan, X.**, **Yang, F.**, **Gao, L.**, **Chen, Z.**, **Zhang, B.**, **Fan, H. and Ren, J.** (2019). Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms, *Remote Sensing*, *11*(8), 917.

[175] **Cai, J. and Chen, Y.** (2021). MHA-Net: Multipath Hybrid Attention Network for building footprint extraction from high-resolution remote sensing imagery, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 5807–5817.

[176] **Qiu, C.**, **Li, H.**, **Guo, W.**, **Chen, X.**, **Yu, A.**, **Tong, X. and Schmitt, M.** (2022). Transferring transformer-based models for cross-area building extraction from remote sensing images, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

[177] **Liu, Z.**, **Mao, H.**, **Wu, C.Y.**, **Feichtenhofer, C.**, **Darrell, T. and Xie, S.** (2022). A convnet for the 2020s, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11976–11986.

[178] **Szegedy, C.**, **Ioffe, S.**, **Vanhoucke, V. and Alemi, A.A.** (2017). Inception-v4, inception-resnet and the impact of residual connections on learning, *Thirty-first AAAI conference on artificial intelligence*.

[179] **Chollet, F.** (2017). Xception: Deep learning with depthwise separable convolutions, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.1251–1258.

[180] **Howard, A.G.**, **Zhu, M.**, **Chen, B.**, **Kalenichenko, D.**, **Wang, W.**, **Weyand, T.**, **Andreetto, M. and Adam, H.** (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint arXiv:1704.04861*.

[181] **Cao, H.**, **Wang, Y.**, **Chen, J.**, **Jiang, D.**, **Zhang, X.**, **Tian, Q. and Wang, M.** (2021). Swin-unet: Unet-like pure transformer for medical image segmentation, *arXiv preprint arXiv:2105.05537*.

[182] **Ba, J.L.**, **Kiros, J.R. and Hinton, G.E.** (2016). Layer normalization, *arXiv preprint arXiv:1607.06450*.

[183] **Hendrycks, D. and Gimpel, K.** (2016). Gaussian error linear units (gelus), *arXiv preprint arXiv:1606.08415*.

[184] **Zhang, Z.**, **Wu, C.**, **Coleman, S. and Kerr, D.** (2020). DENSE-INception U-net for medical image segmentation, *Computer methods and programs in biomedicine*, *192*, 105395.

CURRICULUM VITAE

Name SURNAME: Batuhan SARITÜRK

EDUCATION:

- **M.Sc.:** 2015, Istanbul Technical University, Faculty of Civil Engineering, Department of Geomatics Engineering
- **B.Sc.:** 2012, Karadeniz Technical University, Faculty of Engineering, Department of Geomatics Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2013 - 2022 Research and Teaching Assistant at Istanbul Technical University
- 2013 - GIS Specialist at ITU National Innovation and Research Center for Geographical Information Technologies

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Sariturk, B.**, and Seker, D.Z. (2022). A Residual-Inception U-Net (RIU-Net) Approach and Comparisons with U-Shaped CNN and Transformer Models for Building Segmentation from High-Resolution Satellite Images. *Sensors,* 22(19), 7624.
- **Sariturk, B.**, Kumbasar, D., and Seker, D.Z. (2022). Comparative Analysis of Different CNN Models for Building Segmentation from Satellite and UAV Images. *Photogrammetric Engineering and Remote Sensing,* (In production)
- **Sariturk, B.**, Seker, D.Z., Ozturk, O., and Bayram, B. (2022). Performance evaluation of shallow and deep CNN architectures on building segmentation from high-resolution images. *Earth Science Informatics,* 15(3), 1801-1823.
- **Sariturk, B.**, and Seker, D.Z. (2022). Comparison of Residual and Dense Neural Network Approaches for Building Extraction from High-Resolution Aerial Images. *Advances in Space Research.*
- **Sariturk, B.**, Bayram, B., Duran. Z, and Seker, D.Z. (2020). Feature extraction from satellite images using segnet and fully convolutional networks (FCN). *International Journal of Engineering and Geosciences,* 5(3), 138-143.

- **Sariturk, B.**, Ozturk, O., Seker, D.Z., and Bayram, B. (2019). Building Extraction from High-Resolution Images Using Deep Learning Approach. *International Symposium on Applied Geoinformatics (ISAG) 2019*, November 7-9, 2019 Istanbul, Turkey.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- **Sariturk, B.**, Ozturk, O., Vardar, E., and Seker, D.Z. (2022). Investigation of Potential Usage of UAV Based Photogrammetry and LiDAR for Forest Inventory. *Fresenius Environmental Bulletin,* 31(08A/2022), 8481-8487.

- Ozturk, O., **Sariturk, B.**, and Seker, D.Z. (2022). Contributions of Open Source Tools for Road Data Production in Deep Learning Applications. *Fresenius Environmental Bulletin,* 31(08A/2022), 8548-8552.

- **Sariturk, B.**, Ozturk, O., Atik, M.E., Duran, Z., and Seker, D.Z. (2022). İTÜ'de Güncel Geomatik Mühendisliği Eğitiminde Fotogrametri. *XI. Türkiye Ulusal Fotogrametri ve Uzaktan Algılama Birliği Teknik Sempozyumu (TUFUAB),* May 12-14, 2022 Mersin, Turkey.

- Ozturk, O., Isik, M.S., **Sariturk, B.**, and Seker, D.Z. (2022). Generation of Istanbul road data set using Google Map API for deep learning-based segmentation. *International Journal of Remote Sensing,* 43(8), 2793-2812.

- Ozturk, O., **Sariturk, B.**, and Seker, D.Z. (2020). Comparison of Fully Convolutional Networks (FCN) and U-Net for Road Segmentation from High Resolution Imageries. *International Journal of Environment and Geoinformatics,* 7(3), 272-279.

- Ozturk, O., **Sariturk, B.**, and Seker, D.Z. (2020). Generating Road Data Set For Deep Learning Applications. *20th International Symposium on Environmental Pollution and its Impact on Life in the Mediterranean Region (MESAEP),* October 26-27, 2020 Virtual.

- Atik, M.E., Incekara, A.H., **Sariturk, B.**, Ozturk, O., Duran, Z., and Seker, D.Z. (2019). 3D Object Recognition with Keypoint Based Algorithms. *International Journal of Environment and Geoinformatics,* 6(1), 139-142.

- Ozturk, O., **Sariturk, B.**, and Seker, D.Z (2019). Comparison of FCN and U-Net for Image Segmentation Using High Resolution Images and Aerial Orthophoto Maps. *International Symposium on Applied Geoinformatics (ISAG) 2019*, November 7-9, 2019 Istanbul, Turkey.

- Bilgilioglu, B.B., Ozturk, O., **Sariturk, B.**, and Seker, D.Z. (2019). Object Based Classification of Unmanned Aerial Vehicle (UAV) Imagery for Forest Fires Monitoring. *Fresenius Environmental Bulletin,* 1011.

- Ozturk, O., **Sariturk, B.**, Incekara, A.H, Seker, D.Z., Bayram, B., and Duran, Z. (2018). Contribution of Smartphones for Documentation of Cultural Heritage: A Case Study of Zilkale. *39th Asian Conference on Remote Sensing,* October 15-19, 2018, Kuala Lumpur, Malaysia.

- **Sariturk, B.**, Atik, M.E., Alganci, U., Ozturk, O., Duran, Z., and Seker, D.Z. (2018). Zaman Serileri Kullanılarak Doğal Arazi Örtüsü Değişiminin Analizi: İstanbul Örnek Çalışması. *Uluslararası Su ve Çevre Kongresi,* March 22-24, 2018 Bursa, Turkey.

- **Sariturk, B.**, Incekara, A.H., Ozturk, O., Duran, Z., Atik, M.E., and Seker, D.Z. (2018). Comparison of Laser Scanning and Close-Range Photogrammetry for 3D Object Modeling. *Eurosian GIS Congress,* September 4-7, 2018 Baku, Azerbaijan.

- Albeni, Y., **Sariturk, B.**, Atik, M.E., Ozturk, O., Duran, Z., Seker, D.Z., Bayram, B., Incekara, A.H., and Celik M.F. (2018). Combination of UAV Based Digital Surface Model with 3D Model of ITU Bicycle House Produced by Close-Range Photogrammetry. *ICOMOS Inter-ISC Meeting and Colloquium,* July 10-13, 2018 Kastamonu, Turkey.

- **Sariturk, B.**, Ozturk, O., Topaloglu, R.H., and Seker, D.Z. (2017). Comparison of 3D Models Generated from Images and Video File Using SFM Technique. *International Symposium on GIS Applications in Geography and Geosciences,* October 18-21, 2017 Çanakkale, Turkey.

- Ozturk, O., **Sariturk, B.**, and Seker, D.Z. (2017). Temporal Monitoring of Landslide Using Images Obtained From UAV. *International Symposium on GIS Applications in Geography and Geosciences,* October 18-21, 2017 Çanakkale, Turkey.

- Ozturk, O., **Sariturk, B.**, Seker, D.Z., and Gozel, M.M. (2017). Monitoring of Forest Fires Using Low-Budget Unmanned Aerial Vehicles. *19th International Symposium on Environmental Pollution and its Impact on Life in the Mediterranean Region (MESAEP),* October 4-6, 2017 Rome, Italy.

- **Sariturk, B.**, Ozturk, O. and Seker, D.Z. (2017). Determining the Quantification of Coastal Erosion Using UAV; A Case Study of Trabzon-Turkey. *19th International Symposium on Environmental Pollution and its Impact on Life in the Mediterranean Region (MESAEP),* October 4-6, 2017 Rome, Italy.

- **Sariturk, B.**, and Seker, D.Z. (2017). SFM Tekniği ile 3B Obje Modellemesinde Kullanılan Ticari ve Açık-kaynak Kodlu Yazılımların Karşılaştırılması. *Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi,* 17(4), 126-131.

- **Sariturk, B.**, Sivri, N., and Seker, D.Z. (2017). A Study To Determine Level Of Awareness About Health And Carbon Footprint Among Geomatics Engineers In Turkey. *Fresenius Environmental Bulletin,* 26(1), 156-161.

- Ozbas, E.E., Sivri, N., **Sariturk, B.**, Ongen, A., Ozcan, H.K., and Seker, D.Z. (2017). The relationship between income level and CFP level of the provinces in Turkey: a case study. *International Journal of Global Warming,* 11(3), 294-304.

- **Sariturk, B.**, Sivri, N., and Seker, D.Z. (2016). A Study to Determine the Relations Between Living Standards and Carbon Footprint Among Geomatics Engineers in Turkey. *FIG Working Week 2016*, May 2-6, 2016, Christchurch, New Zealand.