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

Batuhan Sariturk * and Dursun Zafer Seker

Department of Geomatics Engineering, Faculty of Civil Engineering, Istanbul Technical University, Istanbul 34469, Turkey
* Correspondence: sariturkb@itu.edu.tr

Abstract: Building segmentation is crucial for applications extending from map production to urban planning. Nowadays, it is still a challenge due to CNNs’ inability to model global context and Transformers’ high memory need. In this study, 10 CNN and Transformer models were generated, and comparisons were realized. Alongside our proposed Residual-Inception U-Net (RIU-Net), U-Net, Residual U-Net, and Attention Residual U-Net, four CNN architectures (Inception, Inception-ResNet, Xception, and MobileNet) were implemented as encoders to U-Net-based models. Lastly, two Transformer-based approaches (Trans U-Net and Swin U-Net) were also used. Massachusetts Buildings Dataset and Inria Aerial Image Labeling Dataset were used for training and evaluation. On Inria dataset, RIU-Net achieved the highest IoU score, F1 score, and test accuracy, with 0.6736, 0.7868, and 92.23%, respectively. On 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 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 RIU-Net was significantly successful on Inria dataset. On Massachusetts datasets, Residual U-Net, Attention Residual U-Net, and Trans U-Net provided successful results.

Keywords: building segmentation; CNN; Transformer; Inception; residual connections; satellite images

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 [2]. 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 [3–6]. Nonetheless, obtaining reliable and accurate building maps from high-resolution satellite images is still challenging due to various reasons such as complex backgrounds [7,8], similarities between the background and the buildings [9], noise in data [4], heterogeneity in data structures [7,8], diversity in roof types [10], and characteristics (size, shape, color, etc.) of buildings [11], 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 [7]. 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 [12]. 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 [13]. Transformers, which have been used frequently in recent years, are computationally inefficient and need large amounts of memory and big datasets [6].

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 [14] 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 [13]. However, extracting these features requires prior knowledge and is labor-intensive [16]. Although these methods have realized some progress, they have shortcomings, such as low accuracy, low generalization ability, and complex processing [3]. 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 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 [14] 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 [13]. However, extracting these features requires prior knowledge and is labor-intensive [16]. Although these methods have realized some progress, they have shortcomings, such as low accuracy, low generalization ability, and complex processing [3]. 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 [13]. 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 [7]. 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 [4]. 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 [4]. Many researchers have used CNN architectures such as AlexNet [17], LeNet [18], VGGNet [19], ResNet [20], and GoogleNet [21]. However, these CNNs perform region divisions and use computationally expensive fully connected layers [3]. 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 [4]. 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 [4].

To improve the performance of CNNs, Long et al. [22] 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 [16].

Consequently, FCNs have become effective methods for image segmentation studies, and various variants have emerged, such as U-Net [23] and SegNet [24], 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 [23]. SegNet has a
convolutional encoder with a pooling layer and a symmetrical decoder with transposed convolutions, which reduces the number of training parameters [24].

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 [13]. To capture the global context, the most widely used approach is to implement the attention mechanism into the networks [25]. Several studies introduced the attention mechanism to improve feature representation; thereby, they can differentiate buildings from complex backgrounds [26,27]. Deng et al. [28] proposed a grid-based attention mechanism, Guo et al. [29] developed a parallel attention module, Pan et al. [30] tried to combine channel and spatial attention to improve the accuracy for building segmentation, and Cai and Chen [31] 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 [13]. 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 [34].

The Transformer approach was first proposed to be used in natural language processing (NLP), but has recently attracted interest in the computer vision community [6]. The first self-attention-based Vision Transformer (ViT) was proposed in 2020 [35]. 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 [13]. ViT-based approaches have made a tremendous job for image segmentation studies [36–38]. Many researchers in the computer vision and remote sensing field have implemented these methods for segmentation tasks, such as urban monitoring [39,40], land cover/land use analysis [41,42], change detection [43,44], and building segmentation [12,45]. Chen et al. [6] introduced a Sparse Token Transformer to learn the global dependency of Transformer tokens, Yuan and Xu [16] implemented Swin Transformer as an encoder and designed a multi-scale adaptive decoder, and Wang et al. [13] 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 [6]. To deal with these issues, the Swin Transformer approach [46] 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 [34]. 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 10 CNN and Transformer models were generated, and building segmentation from high-resolution satellite images was carried out. Alongside our proposed approach Residual-Inception U-Net (RIU-Net), 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. [48], 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 [49], Inria Aerial Image Labeling Dataset [50], and WHU Building Dataset [51]. 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.

2. Materials and Methods

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.

2.1.1. Inria Aerial Image Labeling Dataset

Inria Aerial Image Labeling Dataset [50] features RGB, orthorectified images with a 30 cm spatial resolution. It covers a total area of 810 km² 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) [50]. Each image and mask are 5000 × 5000 pixels and cover an area of 1500 m × 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.

2.1.2. Massachusetts Buildings Dataset

Massachusetts Buildings Datasets were introduced by Volodymyr Mnih [49] 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 × 1500 pixels in size, has a spatial resolution of 1 m, and spans a 2.25 km² 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 [49].

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 1).

![Figure 1](image1.png)

(a)

(b)

**Figure 1.** Sample 256 × 256 pixel image and mask: (a) Inria dataset, (b) Massachusetts dataset.
2.2. Methodology

A total of 10 CNN and Transformer models were generated and used for building segmentation from high-resolution satellite images. Alongside our proposed approach Residual-Inception U-Net (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.

2.2.1. U-Net

U-Net is an FCN-based architecture proposed by Ronneberger et al. [23]. 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 $3 \times 3$ convolution layers. There are four convolution blocks in the encoder path, and each one includes two $3 \times 3$ convolution layers with Rectified Linear Unit (ReLU) activation, and a $2 \times 2$ max pooling layer. The decoder path also has of four blocks, with each block having one $2 \times 2$ transposed convolution layer, a concatenation layer to connect the extracted feature maps, and two $3 \times 3$ convolution layers [23].

2.2.2. Residual U-Net

The Residual connection was introduced by He et al. [20] 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. [52], shown in Figure 2, was chosen, and it was modified to replace the conventional convolutional block in U-Net.

![Figure 2](image_url)

Figure 2. (a) Residual connection design from the ResNet. (b) Residual connection design implemented in the study.
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 [53]. 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. [26] was implemented in the Residual U-Net model (Figure 3).

![Figure 3. Attention mechanism implemented in the study [26].](image)

2.2.4. Inception Backboned U-Net

Szegedy et al. [21] 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 $1 \times 1$ 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 [21]. Thus, the Inception architecture is restricted to $1 \times 1$, $3 \times 3$, and $5 \times 5$ 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.

2.2.5. Inception-ResNet Backboned U-Net

One of the examples of the widely used deep CNNs is the Inception architecture [21], which achieves successful results at a low computational cost. Lately, the residual connections [20] have provided accomplished results in computer vision tasks. This raised a question for Szegedy et al. [54] 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 [54].

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 $1 \times 1$ 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.
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” [55]. It is possible to think of the depthwise separable convolution as an Inception module with a lot of towers [55]. The idea prompted them to suggest a deep CNN modeled after Inception, by using separable convolutions in place of the Inception module [55]. This extreme version of the Inception module firstly maps the cross-channel correlations by using a \(1 \times 1\) 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.

2.2.7. MobileNet Backboned U-Net

Howard et al. presented mobile and efficient models called “MobileNets” in 2017 [56]. 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 [56]. 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.

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 our study, a smaller variant is used, due to the computational resource limitations. We scaled the model according to the ratios of hyperparameters between the base and the large models, and used a variant named as “Mini”, referred in Table 1.

| 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            |

2.2.9. Swin U-Net

Cao et al. [57] 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 [46], 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 [57].

2.2.10. Residual-Inception U-Net (RIU-Net)

In this study, we proposed a residually connected, Inception-based, u-shaped encoder-decoder architecture with skip connections named as Residual-Inception U-Net (RIU-Net). 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 [54] was used.

Throughout the network, as differs from the Inception-ResNet-v2, “Layer Normalization” [58] was used as a regularization technique instead of Batch Normalization, and the “GELU” activation function [59] 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 [48]. 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 [48]. 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 [48]. In our 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. [48], where they “modernize” the CNNs toward the design of ViT.

The overall design of the proposed approach is given in Figure 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 [27]. It allows Inception to acquire all the benefits of the residual connection while maintaining the computational efficiency [54].

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, $1 \times 1$ 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. 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.
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 [60]. Thus, upsampling modules have been preferred for overcoming these problems. The upsampling module was inspired by the paper from Zhang et al. [60] 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 6.
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 [61]...
was used with a $1 \times 10^{-4}$ 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 (1)) [62]. 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 (2)) [62]. 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 2.

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

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

Table 2. Number 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 3), Precision (Equation 3), IoU (Equation 4), and F1 Score (Equation 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 [63]. 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 [62]. 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.

$$Precision = \frac{TP}{(TP + FP)} \quad Recall = \frac{TP}{(TP + FN)}$$

(3)
\[ \text{IoU} = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{(TP + FP + FN)} \]  \hspace{1cm} (4)

\[ \text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]  \hspace{1cm} (5)

4. Results and Discussion

Evaluation results on the Inria test set are shown in Figure 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.

![Inria Evaluation Metric Results](image)

**Figure 7.** Evaluation metric results on Inria test set.

Evaluation results on Massachusetts Small test set are shown in Figure 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.

According to the results on Massachusetts Large test set, shown in Figure 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.
RGB images from test sets, corresponding masks, and example segmented images are shown in Figure 10, Figure 11, and Figure 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 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 11 and 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 10. Inria test set image no. 1036 segmentation results.

Figure 11. Massachusetts Small test set image no. 192 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. 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.

Author Contributions: Conceptualization, B.S. and D.Z.S.; methodology, B.S. and D.Z.S.; software, B.S.; validation, D.Z.S.; formal analysis, B.S.; investigation, D.Z.S.; writing—original draft preparation, B.S.; writing—review and editing, B.S. and D.Z.S.; visualization, B.S.; supervision, D.Z.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.
Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, J.; Jiang, Y.; Luo, L.; Gu, Y.; Wu, K. Building footprint generation by integrating U-Net with deepened space module. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 3847–3851.

2. Zhang, Y.; Gong, W.; Sun, J.; Li, W. Web-Net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries. Remote Sens. 2019, 11, 1897. [CrossRef]

3. Yu, M.; Chen, X.; Zhang, W.; Liu, Y. AGs-Unet: Building Extraction Model for High Resolution Remote Sensing Images Based on Attention Gates U Network. Sensors 2022, 22, 2932. [CrossRef]

4. Wang, H.; Miao, F. Building extraction from remote sensing images using deep residual U-Net. Eur. J. Remote Sens. 2022, 55, 71–85. [CrossRef]

5. Sun, X.; Zhao, W.; Maretto, R.V.; Persello, C. Building outline extraction from aerial imagery and digital surface model with a frame field learning framework. Int. Arch. Photogramm. Remote Sens. Spat. Inf. 2021, 43, 487–493. [CrossRef]

6. Chen, K.; Zou, Z.; Shi, Z. Building extraction from remote sensing images with sparse token transformers. Remote Sens. 2021, 13, 4441. [CrossRef]

7. Li, Q.; Shi, Y.; Zhu, X.X. Semi-supervised building footprint generation with feature and output consistency training. IEEE Trans. Geosci. Remote Sens. 2022, 60. [CrossRef]

8. Bakirman, T.; Komurcu, I.; Sertel, E. Comparative analysis of deep learning based building extraction methods with the new VHR Istanbul dataset. Expert Syst. Appl. 2022, 202, 117346. [CrossRef]

9. Liu, S.; Ye, H.; Jin, H.; Cheng, H. CT-UNet: Context-Transfer-UNet for Building Segmentation in Remote Sensing Images. Neural Process. Lett. 2021, 53, 4257–4277. [CrossRef]

10. Ok, A.O. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. ISPRS J. Photogramm. Remote Sens. 2013, 86, 21–40. [CrossRef]

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

12. Chen, X.; Qiu, C.; Guo, W.; Yu, A.; Tong, X.; Schmitt, M. Multiscale feature learning by transformer for building extraction from satellite images. IEEE Geosci. Remote Sens. 2022, 19, 1–5. [CrossRef]

13. Wang, L.; Fang, S.; Li, R.; Meng, X. Building extraction with vision transformer. IEEE Trans. Geosci. Remote Sens. 2022, 60, 1–11. [CrossRef]

14. Tang, Y.; Li, L.; Wang, C.; Chen, M.; Feng, W.; Zou, X.; Huang, K. Real-time detection of surface deformation and strain in recycled aggregate concrete-filled steel tubular columns via four-ocular vision. Rob. Comput. Integr. Manuf. 2019, 59, 36–46. [CrossRef]

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

16. Yuan, W.; Xu, W. MSST-Net: A Multi-Scale Adaptive Network for Building Extraction from Remote Sensing Images Based on Swin Transformer. Remote Sens. 2021, 13, 4743. [CrossRef]

17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 2012, 60, 84–90. [CrossRef]

18. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278–2324. [CrossRef]

19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.

20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 770–778.

21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.

22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

23. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.

24. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 2481–2495. [CrossRef] [PubMed]

25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.

26. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; Glocker, B.; Rueckert, D. Attention u-net: Learning where to look for the pancreas. arXiv 2018, arXiv:1804.03999.

27. Sariturk, B.; Seker, D.Z. Comparison of Residual and Dense Neural Network Approaches for Building Extraction from High-Resolution Aerial Images. Adv. Space Res. 2022. [CrossRef]
28. Deng, W.; Shi, Q.; Li, J. Attention-gate-based encoder–decoder network for automatic building extraction. IEEE J. Sel. Top. Appl. 2021, 14, 2611–2620. [CrossRef]
29. Guo, H.; Shi, Q.; Du, B.; Zhang, L.; Wang, D.; Ding, H. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. IEEE T. Geosci. Remote 2020, 59, 4287–4306. [CrossRef]
30. Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Zhang, B.; Fan, H.; Ren, J. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. Remote Sens. 2019, 11, 917. [CrossRef]
31. Cai, J.; Chen, Y. MHA-Net: Multipath Hybrid Attention Network for building footprint extraction from high-resolution remote sensing imagery. IEEE J. Sel. Top. Appl. 2021, 14, 5807–5817. [CrossRef]
32. Sariturk, B.; Seker, D.Z.; Ozturk, O.; Bayram, B. Performance evaluation of shallow and deep CNN architectures on building segmentation from high-resolution images. Earth Sci. Inf. 2022, 15, 1801–1823. [CrossRef]
33. Ozturk, O.; Isik, M.S.; Sariturk, B.; Seker, D.Z. Generation of Istanbul road data set using Google Map API for deep learning-based segmentation. Int. J. Remote Sens. 2022, 43, 2793–2812. [CrossRef]
34. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Strasbourg, France, 27 September–1 October 2021; pp. 14–24.
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.
36. Strudel, R.; García, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 7262–7272.
37. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neur. In. 2021, 34, 12077–12090.
38. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 6681–6690.
39. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. Remote Sens. 2021, 13, 3065. [CrossRef]
40. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote. Sens. 2022, 190, 196–214. [CrossRef]
41. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking hyperspectral image classification with transformers. IEEE Trans. Geosci. Remote Sens. 2021, 60, 1–15. [CrossRef]
42. He, X.; Chen, Y.; Lin, Z. Spatial-spectral transformer for hyperspectral image classification. Remote Sens. 2021, 13, 498. [CrossRef]
43. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. IEEE Trans. Geosci. Remote Sens. 2021, 60, 1–14. [CrossRef]
44. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. IEEE Trans. Geosci. Remote Sens. 2022, 60, 1–13. [CrossRef]
45. Qiu, C.; Li, H.; Guo, W.; Chen, X.; Yu, A.; Tong, X.; Schmitt, M. Transferring transformer-based models for cross-area building extraction from remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2022, 15, 4104–4116. [CrossRef]
46. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021, Virtual, 11–17 October 2021; pp. 10012–10022.
47. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. arXiv 2021, arXiv:2102.04306.
48. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the 2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 11976–11986.
49. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
50. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
51. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 2018, 57, 574–586. [CrossRef]
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 630–645.
53. Fang, Z.; Chen, Y.; Nie, D.; Lin, W.; Shen, D. Rca-u-net: Residual channel attention u-net for fast tissue quantification in magnetic resonance fingerprinting. In Proceedings of the 2022 Medical Image Computing and Computer-Assisted Intervention (MICCAI), Shenzhen, China, 13–17 October 2019; pp. 101–109.
54. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
55. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.

56. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv 2017, arXiv:1704.04861.

57. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv 2021, arXiv:2105.05537.

58. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. arXiv 2016, arXiv:1607.06450.

59. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). arXiv 2016, arXiv:1606.08415.

60. Zhang, Z.; Wu, C.; Coleman, S.; Kerr, D. DENSE-INception U-net for medical image segmentation. Comput. Methods Programs Biomed. 2020, 192, 105395. [CrossRef]

61. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv 2014, arXiv:1412.6980.

62. Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd ed.; O’Reilly Media, Inc.: Sebastopol, CA, USA, 2019; pp. 470–471.

63. Patterson, J.; Gibson, A. Deep Learning: A Practitioner’s Approach, 2nd ed.; O’Reilly Media, Inc.: Sebastopol, CA, USA, 2017; p. 39.