International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.2 (2025) pp. 2298-2312
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Semantic Segmentation of Satellite Images using various U-Net Architectures: A Comparison Study

Nagamani Gonthina[1]*, L. V. Narasimha Prasad[2]

[1]Department of CSE, Jawaharlal Nehru Technological University, Hyderabad
* **Corresponding Author Email:** gnvsk1986@gmail.com - **ORCID:** 0000-0001-9559-8030

[2]Department of CSE, Institute of Aeronautical Engineering
Email: lvnprasad@iare.ac.in - **ORCID:** 0000-0001-6514-1064

Article Info:

DOI: 10.22399/ijcesen.1360
Received : 05 January 2025
Accepted : 10 March 2025

Keywords :

Aerial imagery,
Attention mechanisms,
Semantic segmentation,
U-Net architecture,
Residual block.

Abstract:

Image segmentation has been a challenging issue in computer vision for years. In contrast to image classification and object detection, semantic segmentation is considered the top tier of the image analysis approach, which gives detailed details of the scene for a given input image. Analysis of aerial images without human intervention has developed a keen interest in research because of its vital importance in various domains. Different applications like disaster response and urban planning depend mostly on semantic segmentation of aerial imagery for their analysis. In a wide range of image processing tasks, convolutional neural networks (CNNs) have manifested their tremendous performance, and the transformation of the computer vision domain is achieved by deep learning. Amongst multiple varieties of CNN, U-Net has proved its efficiency in segmenting aerial images and the medical domain. Nonetheless, U-Net can't extract potential spatial features from aerial images because of insufficient layers and may output inaccurate boundaries, particularly for objects with compound structures. To circumvent these deficiencies, different varieties of U-Net are experimented with for aerial image segmentation using U-Net, Attention U-Net, Attention Res U-Net, and Recurrent Residual U-Net. We evaluated all these models on a publicly available dataset named semantic segmentation of aerial imagery. Extensive experimental results conclude that Attention Res U-Net and Recurrent Residual U-Net perform better than other U-Net architectures.

1. Introduction

Image segmentation was a challenging issue from the beginning of the computer vision domain. Image segmentation is considered pixel-level classification, which focuses on diverging an image into meaningful sections by categorizing individual pixels into a distinct entity [1]. This helps in processing only the important segments of an image instead of the whole image. The partition of similar sections in a scenario that has identical structure or texture, like the computable objects titled as things, whereas the incomputable regions, like sky and water titled stuff, is achieved [2,3]. i.e., the whole scene is partitioned into things and stuff. Image segmentation is categorized into three distinct divisions based on identifying things and stuff. The initial division is semantic segmentation. This segmentation categorizes various areas in the pictures that belong to the same group of objects or stuff. This segmentation method identifies both things and the stuff of the scene. The second category is instance segmentation. This segmentation method isolates and analyzes those scene elements with an object recognized and marked with a bounding box or segmentation mask. The third category is panoptic segmentation, an amalgamation of both semantic and instance segmentation [4]. The difference between these three segmentation types is shown with a sample image in figure 1. In contrast to object detection and image classification, semantic segmentation is considered the top tier of the image analysis approach that gives thorough details of the scene for a given input image [5]. Semantic segmentation is implemented in a variety of real-world applications, like therapy planning, self-driving vehicles, defect detection, pedestrian detection, computer-aided diagnosis,

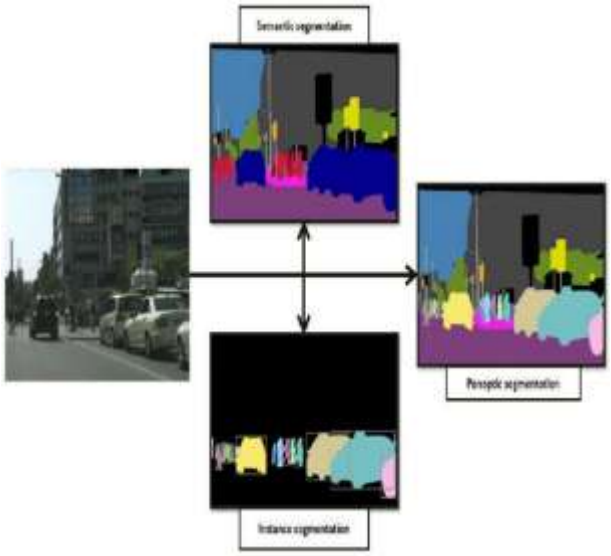


Figure 1. *Sample image depicting the variation among semantic segmentation, instance segmentation and panoptic segmentation.*

robotic systems, augmented reality, video surveillance, intelligent military systems etc., [6-8]. The huge advancement in satellite sensors allows individuals to capture vast aerial images. Aerial image analysis without human intervention has developed keen importance in research because of its vital significance in navigation, agriculture, urban planning, disaster response, land cover change detection, environmental management [9,10] etc. Semantic segmentation in aerial images is carried out to identify a cluster of pixels which have different categories. The conventional techniques in semantic segmentation focus mostly on the manual features which fail to achieve acceptable results and are confined by the representation ability of features [11]. Convolutional neural networks (CNNs) have manifested their tremendous achievement in a wide range of applications related to image processing [12] as well as deep learning has achieved the transformation of computer vision domain [13]. Semantic segmentation using deep learning has gained huge attention in the past decade [14,15]. Various researchers propose various architectures to enhance the efficiency of the resultant segmented image [16,17]. However, each architecture also has its limitations. So, extensive research is still being done in digital image processing towards architectural designs used in image segmentation to achieve better results in a specific domain. The accuracy achieved in the segmentation results is crucial in making decisions or analyzing the prevailing scenario. Hence, based on the advantages and limitations of a specific architecture, the researchers modify or ensemble various architectures to meet the needs and domain-specific challenges. The U-Net architecture gained significant attention for its capability to capture the exact boundaries of objects when it was first introduced for biological image segmentation [18]. Besides the medical domain, U-Net has also proved its efficiency in segmenting aerial images [19]. But still, it could not extract potential spatial information from satellite data because of the inadequate numbers of layers [20], and may output inaccurate boundaries, particularly for objects with compound structures [21]. A customized approach is required to get out of these challenges. To address these limitations, U-Net architecture is integrated with supplementary techniques such as attention mechanisms, skip connections, and post-processing steps to upgrade the performance of aerial imagery segmentation. The aforementioned methods suffer from high computational time and could not perform better on large datasets [22,23]. Attention mechanisms have demonstrated their efficacy in improving feature extraction and segmentation accuracy in various computer vision applications [24,25] but still suffer from low segmentation accuracy for the noisy and blurred images [26]. To examine the aforementioned issues, we experimented with varieties of U-Net and compared their performance. The contributions of the proposed work are summarized below:

- Suggested a framework to efficiently compare U-Net variations to segment the aerial images using various performance metrics.
- Preprocessing techniques like cropping and patchifying are used on the dataset to standardize the resolution of all pictures.
- Experimented the suggested work on publicly available dataset and compared it with conventional approaches.

2. Related Work

Semantic segmentation is considered the top tier of the image analysis process that gives thorough information about the scene for a given input image. The huge advancement in satellite sensors allows individuals to capture vast aerial images. The aerial image analysis without human intervention has gained a lot of significance in research. Semantic segmentation using deep learning has gained huge attention in the past decade. A vast amount of research is being carried out on various techniques used for semantic segmentation, such as deep learning in aerial images, to mitigate the existing challenges like low segmentation accuracy, high computational cost, and low performance on large datasets. Tang, Maofeng et al. implemented contrastive learning (CL) methodologies with the help of self-supervised learning [27] to segment aerial images semantically. Abdollahi et al. [28] built a Seg–unet model, combining Segnet and Unet

architectures, and achieved 92.73% accuracy in building extraction from high-resolution aerial images. A novel framework was suggested by Wang et al. [29] known as an Efficient Non-local Residual U-shape Network (ENRU-Net), that combines a well- structured encoder-decoder in U-shape and an improved non-local network known as asymmetric pyramid non-local block (APNB). The encoder-decoder is used to extract and recapitulate the feature maps effectively, and APNB efficiently retrieves global contextual information through the self-attention process. The multiRes-UNet architecture was built by Abdollahi et al. [30], which reinforces the basic UNet. MultiRes block was used in this architecture to incorporate the learned features at different scales from the data and consist of additional spatial information. Benjdira, Bilel, et al. tackled the issue of domain adaptation in aerial images for semantic segmentation [31] with Generative Adversarial Networks (GANs) assistance. A new attention-based architecture known as a hybrid multiple attention network (HMANet) was built by Ruigang Niu et al. [32] to effectively and flexibly extract global correlations in space, channel and category views. A framework for enhancing U-Net was showcased by Su Zhongbin et al. [33]. They created a deep convolutional neural network (DCNN) by integrating the dilated convolution, U-Net, and DenseNet efficiencies. Marmanis, Dimitrios, et al. [34] designed a model which is an ensemble of Fully Convolutional Networks (FCNs) that inputs range data and intensity and converts them to a full resolution pixel level classification using efficient de-convolution as well as recycling of initial network layers. Li, Weijia, et al. [35] proposed an ensemble architecture by combining U-Net with DeepLabv3+ for semantic segmentation and building extraction of satellite images with high resolution. An ensemble model called WaterNet was designed by Erdem, Firat, et al. [36] by integrating Fractal U-Net, Dilated U-Net, U-Net, Pix2Pix and FC-DenseNet to obtain shorelines from satellite images automatically. Yuan, Kunhao, et al. [37] designed a new architecture called multichannel water body detection network (MC-WBDN) with components, a multichannel fusion module, an enhanced Atrous Spatial Pyramid Pooling module, and Space-to-Depth/Depth-to-Space operations, to detect water body in satellite images.

Saifi et al. [38] designed an automated framework for semantic map extraction from satellite imagery using FCN and U-Net to track the growth in urban cities. Avenash et al. [39] suggested a framework by modifying the CNN known as U-HardNet with a new activation function known as Hard-Swish for remote sensing imagery segmentation. A 3D-2D CNN framework that uses spectral and spatial details was designed by Saralioglu et al. [40] to output precise land cover information from very high-resolution satellite imagery. A.K. Brand et al. [41] designed a framework with U-Net and CNN to semantically segment burned areas in satellite imagery. Singh, Ningthoujam Johny et al. [42] proposed an architecture named Deep Unet for semantic segmentation with pre-processing of the image based on fast and automatically adjustable Gaussian radial basis function kernel-based fuzzy C-means (FAAGKFCM) and simple linear iterative clustering (SLIC) Superpixel to establish mapping for classifying different landfills based on satellite imagery and outperformed conventional methods with accuracy of 90.63 and mIoU of 89.51%. Wang, Xiaolei, et al. [43] proposed an improved UNet called Adaptive Feature Fusion UNet (AFF-UNet) to optimize the semantic segmentation accuracy of remote sensing images. Their model obtained an improvement of 1.09% over DeepLabv3+ for the average F1 score and a 0.99% improvement in overall accuracy. Maurya Abhishek et al. [44] designed a modified U-net-based architecture to segment satellite images on a novel dataset. Even though many architectures have been designed to date, much research is still being carried out to enhance further the performance metrics in the semantic segmentation of satellite images.

3. Methodology

The workflow of the suggested model is showcased in Figure 2. This model principally consists of two phases. The training phase is the first, and the testing phase is the second. During the training phase, preprocessing techniques are implemented on the dataset to standardize the resolution of all pictures. The standardized dataset is divided into 80% -20% ratio of training and testing data. Then, the preprocessed images are trained with variations of U-Net, such as U-Net, Attention U-Net, Attention Res U-Net, and R2 U-Net. An un-trained image is input to the trained model in the prediction phase. The output is the segmented image.

3.1 Data Preprocessing

The considered dataset has aerial photos of Dubai obtained using MBRSC satellites and labelled using semantic segmentation at the pixel level for six different classes: Building, Land, Road, Vegetation, Water, and Unlabeled. It contains eight large tiles, each with nine images and their corresponding masks, which are segregated into 72 images and their corresponding masks. This dataset, which provides various aerial images extracted in different

conditions, seasons, and places, is renowned for work in aerial imagery analysis. The high-quality annotations in this dataset, which included pixel-level ground truth labels for semantic segmentation, were the deciding factor in its selection. These labels were very important for training and assessing our

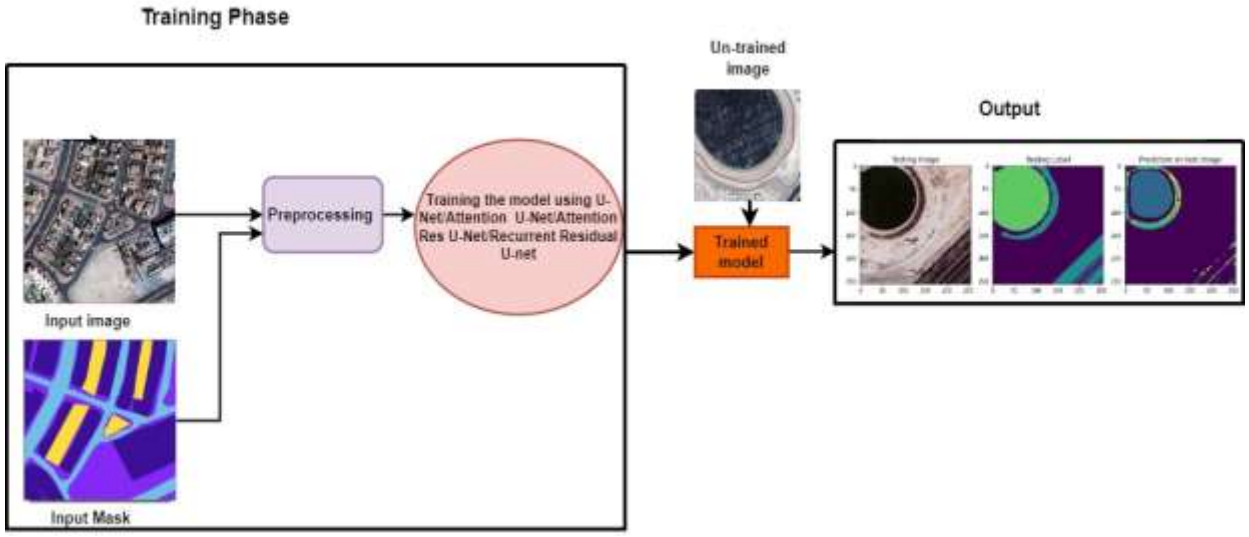


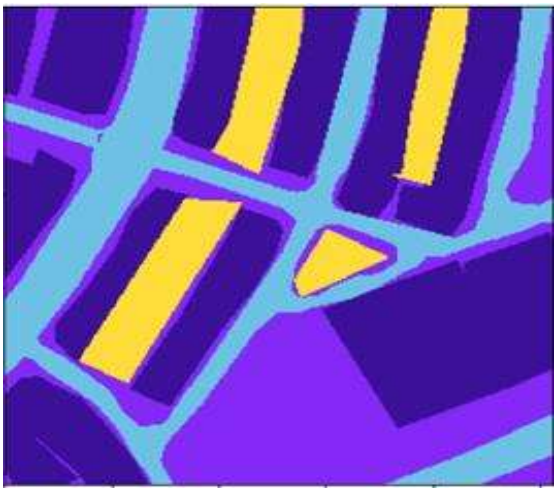
Figure 2. *Flow of the proposed model.*

. *Sample image before patchifying.*

Corresponding mask of sample image.

Figure 3. *Sample image and mask before patchifying.*

Sample patch of an image after



Corresponding mask of sample patchifying. Patch.

Figure 4. *Sample image and mask after patchifying.*

segmentation model precisely. Images in the dataset are of various sizes: 797x644, 859x838, 1817x2061, 509x544, 682x658, 1099x846, 1126x1058, and 2149x1479 each belonging to different tiles from tile1 to tile8. We standardized all picture resolutions to ensure consistency and make integration into our model easier. The images and masks are preprocessed by cropping them to a magnitude divisible by 256, retrieving the patches with the patchify library, and obtaining 1305 patches of images as well as their corresponding masks so that each image and mask can be captured into numpy arrays. The sample image and its corresponding mask before preprocessing and not divided into patches are depicted in Figure 3a, 3b. 4a, 4b depicts the sample patch and its corresponding mask after dividing the image into patches using the patchify library. The masks in the data set are in RGB and their details are provided as HEX color codes. So, we converted HEX to RGB values, then converted RGB labels to integers, and performed one hot encoding. After preprocessing, the data set is split as 20% of the data designated for testing and validation, and 80% was allocated for training.

3.2 Training the model

After preprocessing the dataset, the model is trained for 10,100,150 epochs using various architectures like U-Net, Attention U-Net, Attention Res U-Net, and R2 U-Net. Prediction Phase. After training, we continued with the prediction phase for performance evaluation of the U-Net architectures. An image from test data (20 % of the dataset, which is split using train-test split) is taken randomly and given as input to the trained model. The output is finally the segmented image. Below is a brief overview of the U-Net architectures differentiated in the proposed work.U-Net is mainly designed to segment the images in the medical field. Besides the medical domain, it has proved its supervision towards segmentation of images in other domains like surveillance, agriculture, land cover, land usage, etc. Nevertheless, U-Net has limitations like overfitting with small datasets, improper segmentation of edges or boundaries, and loss of spatial information due to excessive down sampling, etc. We can overcome the above limitations by making some improvements to traditional U-Net, such as Attention U-Net by adding attention blocks, ResU-Net by adding residual blocks, and Attention Res U-Net by adding attention, as well as residual blocks. In this section, we will discuss all these architectures in detail.

Attention U-Net

Attention is a technique that focuses solely on the admissible activations during training. This saves the resources for computation spent unnecessarily on unrelated activations and enhances the network's capacity for generalization. Two categories of attention exist:

- Hard attention Focuses relevant regions through cropping and crops a single region at a time within an image, which implies it is non-differentiable and needs reinforcement learning. The network determines whether it finds the area of interest, and there is nothing in between. Backpropagation is not supported here.
- Soft attention Various Weights are assigned to different regions of an image like relevant regions are assigned more weights than irrelevant regions. The network can be trained using backpropagation. The weights here also get trained while training, allowing the model to concentrate more towards the relevant regions. In U-net, the skip connection concatenates the spatial details of the down-sampling path with the corresponding up-sampling path to withhold proper spatial information. However, this method carries over the inadequate feature representation from the first few layers. The issue above is addressed with the implementation of soft attention at the shortcut connections, which helps to restrain activations at insignificant parts efficiently, i.e. attention is used to supply extra weightage to features of interest [45].

Figure 5 shows the Attention U-Net framework, and Figure 6 shows the block diagram of the attention gate used in Attention U-Net. The attention gate combines two inputs, x and g, and gives input to the corresponding layer. The gating signal, g, comes from the following lowest layer and has enhanced feature representation as it emerges from the network's deeper layers. x emerges from the shortcut connection and has enhanced spatial representation as it comes from the network's early layer.

Attention Res U-Net

Attention Res U-Net is an integration of attention and residual modules that is implemented in the actual U-Net model. The foremost use of the residual module is to facilitate the training for very deep networks by addressing the vanishing gradient challenge. Attention blocks help improve segmentation accuracy by capturing fine details and handling complex scenes. The full pre-activation residual block is used in this architecture as it gives the best percentage of classification error compared to other combinations. This architecture holds all the benefits of attention and residual blocks and gives better results than Attention U-Net and Res U-Net alone. Figure 7 depicts the structure of Attention Res U-net [46].

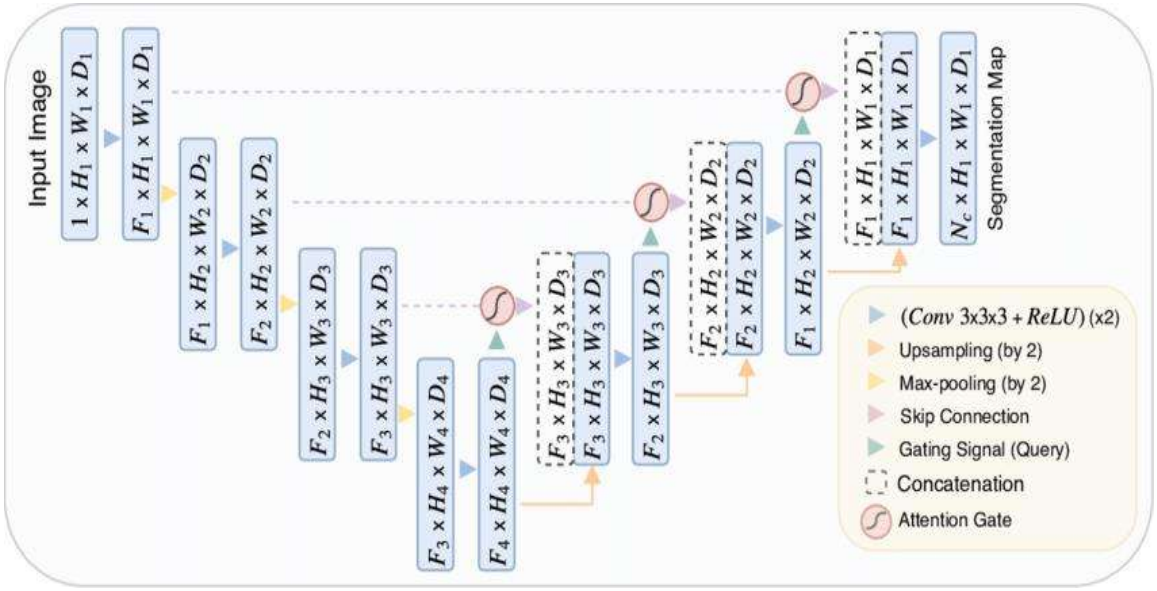


Figure 5. *Structure of the Attention U-Net segmentation model.*

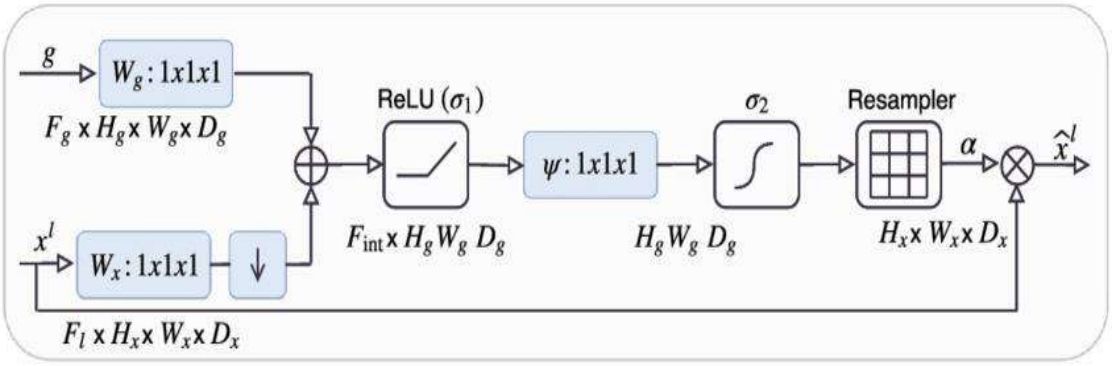


Figure 6. *Block diagram of the attention gate (AG). Input features (xl) are scaled with attention coefficients (α) evaluated in AG. Spatial regions are chosen by investigating activations as well as contextual details produced by the gating signal (g) that is gathered from a coarser scale.*

Residual block

Deeper networks can learn more compared to models with fewer layers. However, these deeper networks suffer from a serious issue called vanishing gradient, where the weights may not be updated properly during backpropagation and may ultimately tend to zero. Deeper networks also encounter the degradation problem, so the networks cannot learn identity functions. Residual blocks are mainly designed to overcome the aforementioned problems using shortcut connections. The difference between the architecture of a conventional convolution block and the residual block is shown in Figure 8 [47,48]. Training of some layers can be skipped with skip connections, shortcut connections, or residual connections, as depicted in Figure 8. The figure illustrates the direct learning capacity of an identity function by merely using skip connections. That's the reason skip connections are even known as identity shortcut connections. Due to these skip connections, deeper networks can be trained by propagating higher gradients to the initial layers, which may learn as quickly as the final layers. Figure 9 [49] depicts the arrangement of the residual block and identity connections to achieve optimal gradient flow. Figure 9(a) illustrates the actual Residual Unit. Here, Xl represents the lth Residual Unit's input feature. Batch Normalization (BN) is done following every weight layer, and ReLU is implemented after BN, except that element-wise addition comes after the final ReLU in a residual unit. Figure 9(b–e) shows other examined possibilities and are discussed as follows.

BN After Addition. Prior to converting f(x) into an identity mapping, BN is adopted after addition (figure 9(b)). Here, f(x) includes BN and ReLU. The output obtained is considerably poor compared to the baseline. In contrast to the actual model, the BN layer now changes the signal, which propagates via the shortcut and obstructs information propagation, as evidenced by the challenges in minimizing training loss at the inception of training.

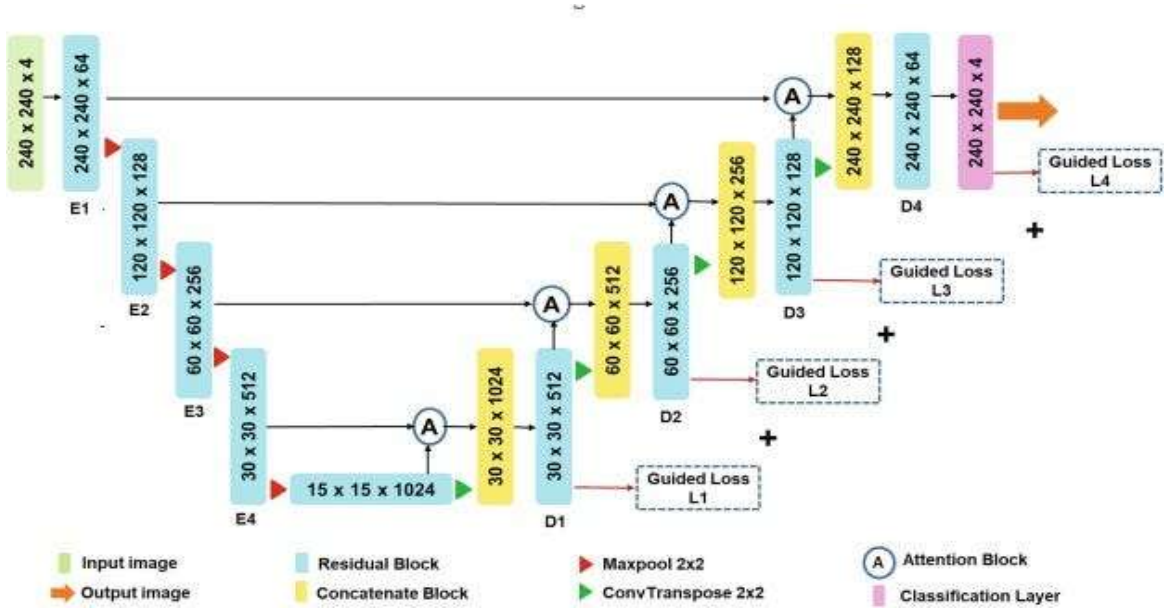


Figure 7. *Structure of the Attention Res U-Net segmentation model.*

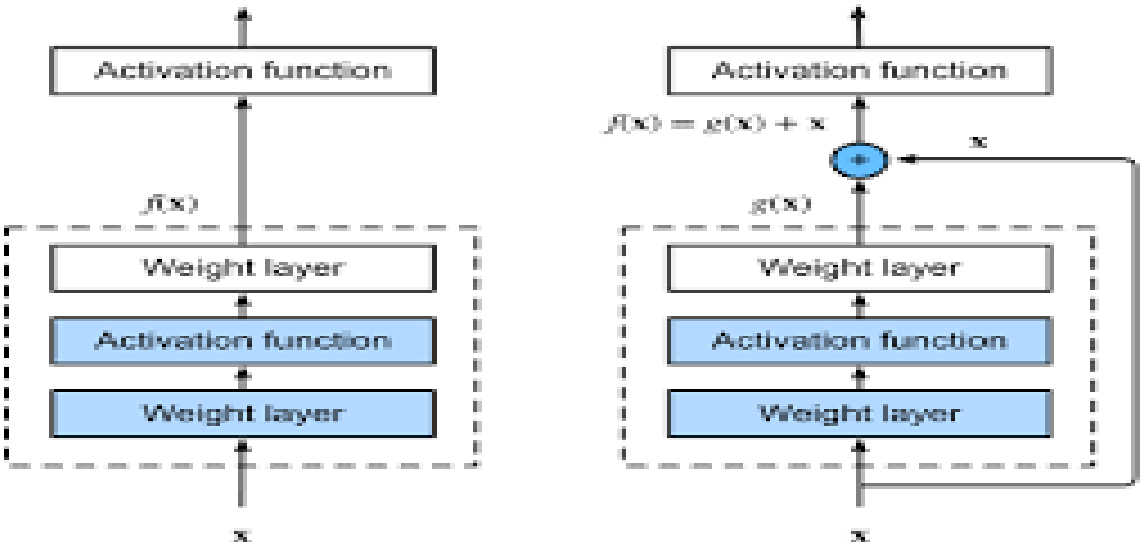


Figure 8. *In the Regular Convolution block (left), the module representing the dotted-line should learn the mapping f(x) directly. In the Residual block (right), the module representing the dotted-line should learn the residual mapping g(x) = f(x) - x, simplifying the identity mapping f(x) = x for learning.*

ReLU Before Addition. An immature option of implementing f(x) into an identity mapping is to forward the ReLU prior to addition (figure 9(c)). Anyways, this arises a non-negative result from the transform, where a "residual" function must automatically consider values in (-∞, +∞). This may impact the representational ability, and the result is worse than the original residual unit.

Post-activation or Pre-activation. The element-wise addition here produces the difference between post-activation and pre-activation. A basic network of N layers has N-1 activations (BN/ReLU), which doesn't matter if we consider them pre- or post-activations. However, the
position of activation matters in the case of branched layers combined through addition. We evaluated the structures: (i) ReLU-only pre-activation (figure 9(d)) and (ii) full pre-activation (figure 9(e)) where BN as well as ReLU are together implemented before weight layers. Table 1 shows that the ReLU-only pre-activation behaves much closer to the original residual unit. This ReLU layer is not combined with a BN layer and may not benefit from BN. Remarkably, the outcomes improve when BN and ReLU are utilized as pre-activation. The best output came from pre-activations using batch normalizations (i.e., the most promising results are seen in the right-most residual block in Figure 9).

Recurrent-residual (R2) U-Net

The R2 U-Net is derived from the conventional U-Net model, which embeds residual connections and

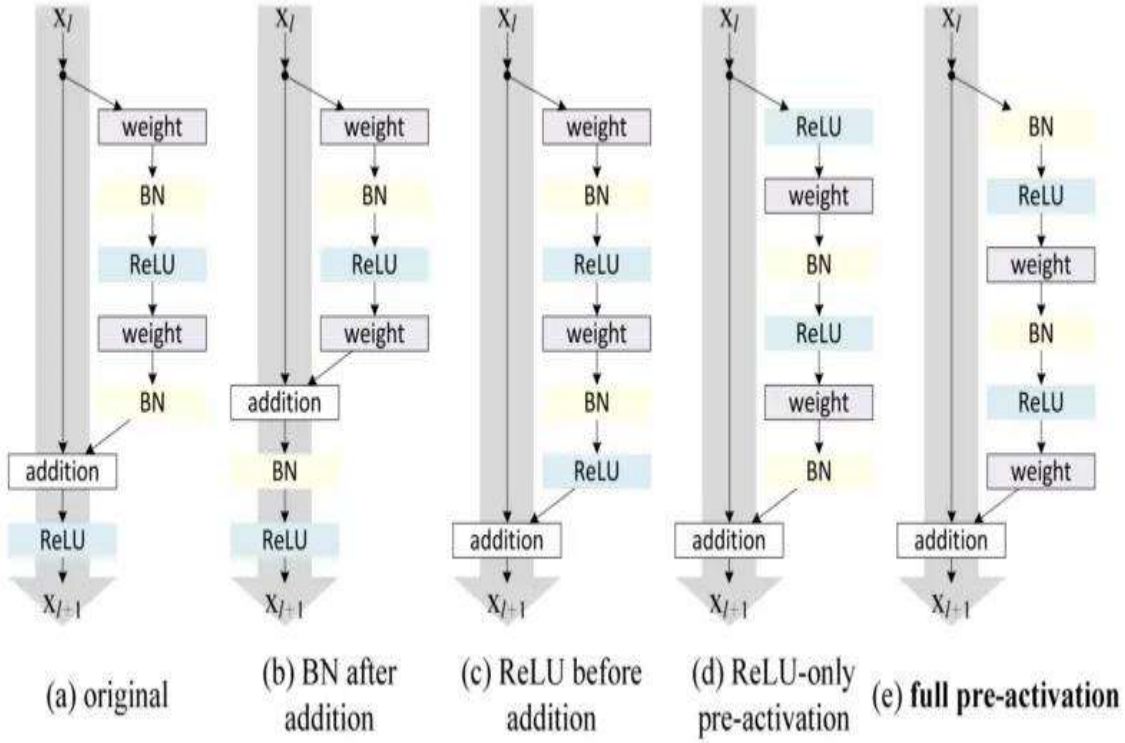


Figure 9. *Types of residual blocks.*

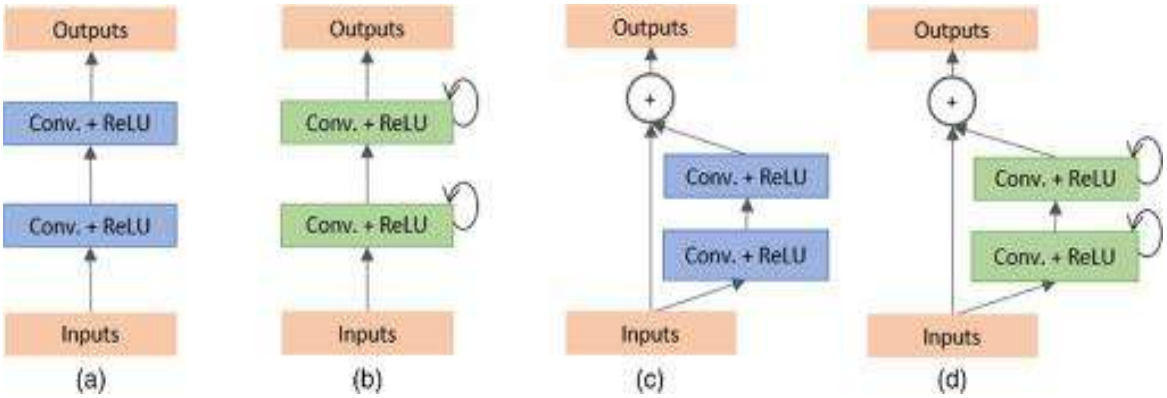


Figure 10. *Variations of the convolutional and recurrent convolutional units (RCUs) comprising (a) the forward convolutional unit, (b) the recurrent convolutional block, (c) the residual convolutional unit, and (d) the recurrent residual convolutional unit.*

Table 1. *Classification error (%) on the Semantic segmentation of aerial imagery test set with distinct activation functions.*

Type of Residual Block	**Classification Error (in %**
Actual Residual Unit	6.62
BN after addition	8.17
ReLU before addition	7.84
Full pre-activation	6.71
ReLU-only pre-activation	**6.37**

recurrent layers. Feature aggregation with recurrent, residual convolutional layers guarantees enhanced feature representation in the segmentation process. R2 U-Net integrates the advantages of residual learning and recurrent connections. Recurrent layers can capture long-range dependencies, which is important in tasks where information from distant regions of the input is relevant. Residual connections permit the reuse of features learned in earlier layers, which helps retain important information and gradients during training. This can lead to faster convergence and better overall performance. Using recurrent connections and residual blocks, you can potentially build a network with fewer parameters than non-recurrent deep networks, which can benefit model efficiency and training on limited hardware.

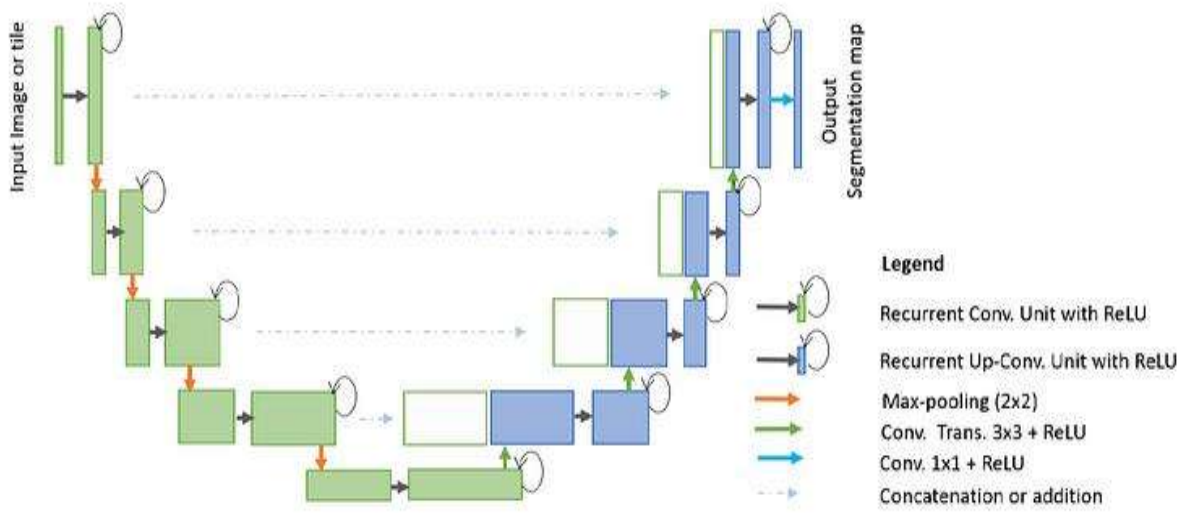


Figure 11. *Architecture of the Recurrent Residual U-Net model for segmentation*

However, the R2 U-Net exhibits marginally superior performance and has only a few more parameters, but because of the layer recurrences that require extra computation at each step, it is much slower to train and evaluate [50]. Figure 10 illustrates the variations of the convolutional and recurrent convolutional units. Figure 10(a) represents the convolutional unit with a forward pass, and Figure 10(b) illustrates the recurrent convolutional block. This block combines convolutional and recurrent layers to capture both spatial features through convolution and temporal dependencies through recurrent connections. Figure 10(c) represents the residual block, and Figure 10(d) represents the recurrent residual convolutional unit, which is a combination of residual block and recurrent connections. Figure 11 illustrates the structure of the Recurrent Residual U-Net model.

4. Results and Discussion

The initial phase of our proposed model is the training phase. In this phase, the model is preprocessed and trained with various variations of U-Net architectures. Initially, the dataset is downloaded from Kaggle and is called semantic segmentation of aerial imagery. This dataset was considered from repositories that are accessible publicly at:
https://www.kaggle.com/datasets/humansintheloop/semantic-segmentation-of-aerial-imagery

4.1 Setup for the Experiment

The software libraries like Keras and TensorFlow are considered for building the model. Our training phase was carried out with careful examination of software, hardware, and training setups. The proposed model is trained using computational infrastructure with a CPU of 8GB, GPU of 7.9 GB and an Intel i7 processor. This hybrid setup efficiently used the GPUs' parallel processing power to guarantee the model's successful training.
During the training phase, the data is initially preprocessed using techniques like cropping and patchifying to standardize the resolution of all pictures. The data is split into (80-20)\% of training and testing data, followed by preprocessing. We considered the ratio of (80-20) \% of train–test split as a higher percentage of data dedicated to training helps the model to generalize better and avoid overfitting. After the train test split, the model is trained for various numbers of epochs, like 10,100,150.

4.2 Performance Analysis

To confirm the efficiency of various architectures in attaining precise semantic segmentation of aerial pictures, different metrics like mIoU (Mean Intersection over Union) and accuracy across multiple architectures considered in this work are experimented.
True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are terms generally utilized to assess the accuracy of predictions and to examine the efficiency of a classifier.
True Positive (TP)
When the model accurately predicts a pixel to be a part of an object when it is a part of the object, is called a True Positive.
True Negative (TN)Type equation here.
A True Negative is when the model incorrectly predicts a pixel to be a part of one object when it is, infact a part of another object.

False Positive (FP)

A False Positive is when the model mispredicts a pixel in the background as part of an object.

False Negative (FN)

A False Negative is when the model misidentifies a positive example or condition as negative. It misclassifies an instance that should have been belonging to a specific class.

Accuracy

Pixel accuracy quantifies the probability of accurately identified pixels in the aerial images. It provides a pixel-by-pixel assessment of our model's accuracy. It can be mathematically expressed using the equation (1)

Accuracy = correctly predicted pixels / total number of pixels in the image i.e.

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{FP}} \quad (1)$$

Intersection over Union (IoU)

IoU finds the proportion between the intersection and union areas of the predicted and ground truth regions. It is a crucial measure in the evaluation of segmentation models as it evaluates to which extent the model can separate objects from their background in an image. For each given class, this metric details how well the model functions in every class. It can be mathematically expressed using the equation (2)

$$IoU = \frac{TP}{TP+FP+FN} \quad (2)$$

Mean Intersection over Union (mIoU)

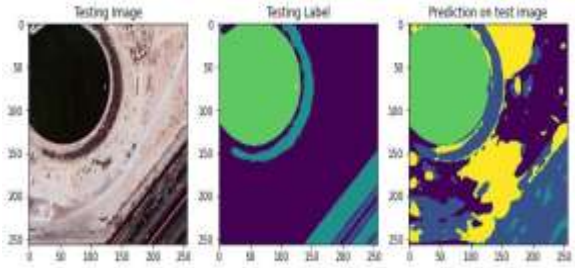
The measure mIoU combines the IoU results from all classes and provides a comprehensive analysis of the over all segmentation accuracy. The mean of the IoU values for each class in the dataset is called mIoU. It can be mathematically expressed using the equation (3)

$$mIoU = \frac{\text{IoU}_{\text{class1}} + \text{IoU}_{\text{class2}} + \cdots + \text{IoU}_{\text{classN}}}{\text{N}} \quad (3)$$

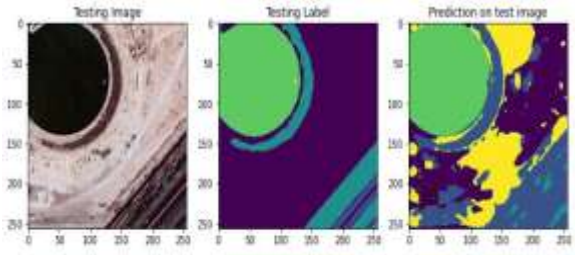
where N is the number of classes. The higher mIoU score indicates better segmentation accuracy. It gives a complete assessment of the performance of the model across every class in the dataset.

A modified categorical cross-entropy loss, which accounts for class weights, was considered the loss function. By assigning a higher priority to the underrepresented classes, this loss function made it possible for the model to learn efficiently across all classes. The activation function considered is softmax. The output images that are segmented for various U-Net architectures at 100 epochs on a given test image are presented in Figure 12.

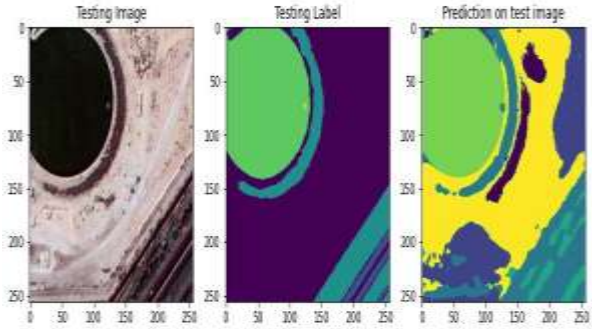
Figure 12a shows the segmentation result for the U-Net model. Figure 12b shows the segmented output for the Attention U-Net model. Figure 12c represents the prediction of the Attention Res U-Net model, and Figure 12d shows the segmented result for the R2 U-Net model. Figure 13a, 13b, 13c, 13d depicts the curves of training loss and validation loss at 100 epochs for U-Net model, Attention U-Net model, Attention Res U-Net model, Recurrent Residual U-Net model respectively. The experimental part played a crucial role in examining various architectures' efficacy in obtaining accurate segmentation results. A comparison of different metrics like training loss, validation loss, accuracy and meanIoU across various architectures used in this work is shown in table 2. Values from table 2 illustrates that Attention Res U-Net and R2 U-net performs better compared to other architectures. From the above-predicted segmentation output in Figure 12 as well as the table 2, it is apparent that Attention Res U-Net



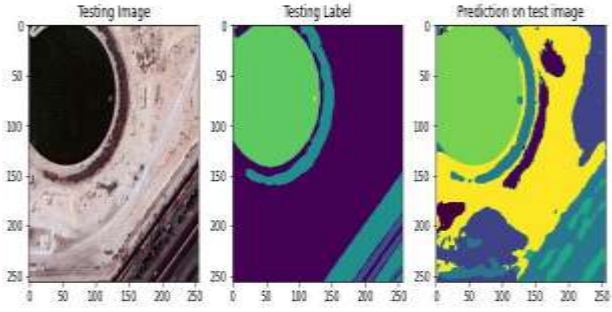
Prediction of U-Net model.



Prediction of Attention U-Net model.

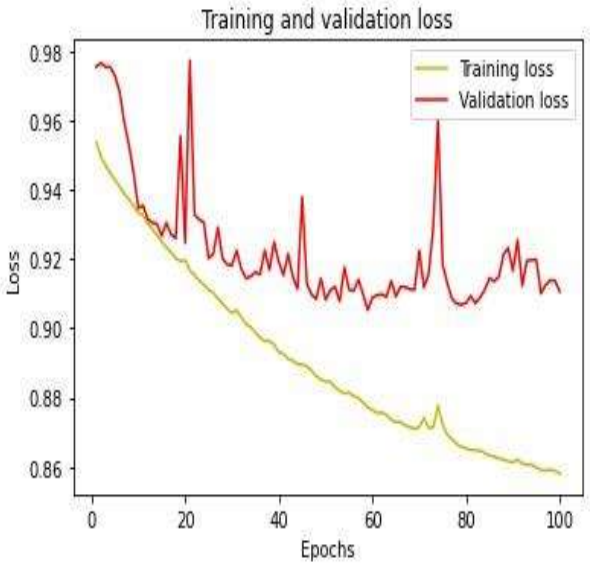


Prediction of Attention Res U-Net model.

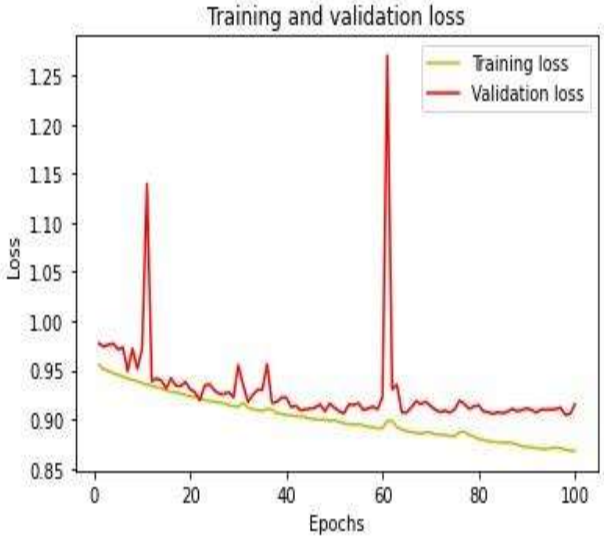


Prediction of R2 U-Net model.

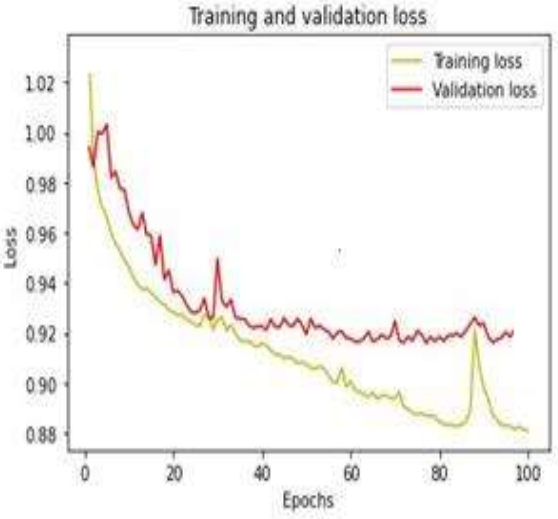
Figure 12. *Prediction of output for various U-Net models.*



Prediction of Training loss Validation loss for U-Net model.



Prediction of Training loss and Validation loss for Attention U-Net model.



Prediction of Training loss and Validation loss for Attention Res U-Net model.

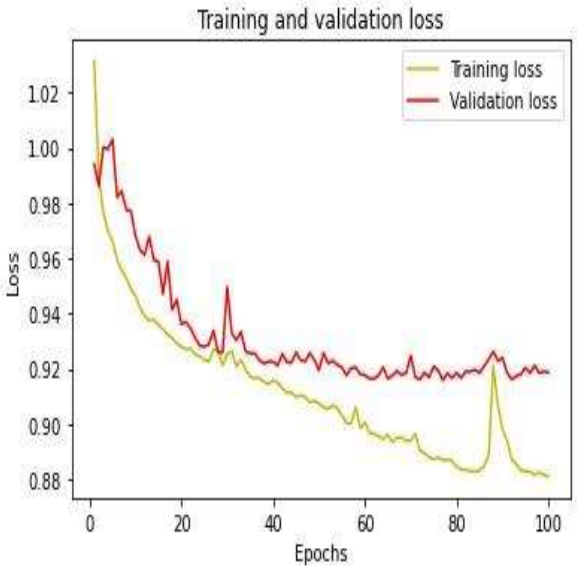


Figure 13. *Prediction of output for various U-Net models.*

and R2 U-Net perform better than other architectures, noting a very negligible difference in improvement of R2 U-Net compared to Attention Res U-Net.

4.3 Experimental Analysis

Validation accuracy and mIoU are two metrics mostly used in connection with deep learning and semantic segmentation to estimate the efficiency of a model. If the validation accuracy increases with the number of epochs but the mean IoU remains unchanged, it indicates an improvement in overall pixel-wise classification accuracy but no improvement in the spatial alignment between the segmentation result predicted and ground truth. Generally, if the gap between training loss and validation loss is low, it specifies that our model is efficiently generalizing untrained data.

The proposed model is executed for 10,100 and 150 epochs. Table 2 shows that the difference between training loss and validation loss is low for the model at 100 epochs, demonstrating that our model is efficient in generalizing untrained data at 100 epochs. It is also evident that the percentage of accuracy and meanIoU of the model is improved from 10 epochs to 100 epochs.

For 150 epochs, even though the model's accuracy has progressed, the mean is almost the same and even slightly decreased, indicating no further improvement in the spatial alignment between the ground truth and predicted segmentation. So, as the accuracy achieved at 150 epochs is maximum, and at the same time, the behaviour of meanIoU is stagnant, we conclude our evaluation at 150 epochs. Figure 14 represents the training and validation loss graph. Figure 15 illustrates the accuracy and

meanIoU graphs at 100 epochs across various U-Net architectures. The graphs show that the training and validation loss of Attention Res U-Net and R2 U-Net is less, and the accuracy and meanIoU of Attention Res U-Net and R2 U-Net is better in contrast to other architectures.

5. Conclusion

Semantic segmentation is considered the top tier of the image analysis process that gives thorough information about the scene for a given input image. In various domains, semantic segmentation of aerial images plays a challenging role. As CNNs have proved their prominence in deep learning, various deep learning architectures have been proposed by various researchers. In this work, different varieties of U-Net architectures like Res U-Net, Attention U-Net, Attention Res U-Net, Recurrent Residual U-Net are experimented and concluded that Attention Res U-Net, Recurrent Residual U-Net performs better in contrast to other architectures. Recurrent Residual U-Net has marginally superior performance to Attention Res U-Net but takes more execution time than Attention Res U-Net.

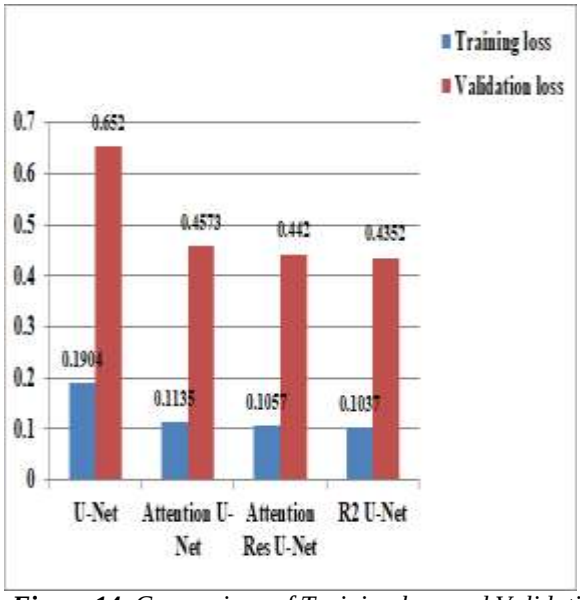


Figure 14. *Comparison of Training loss and Validation loss for various architectures.*

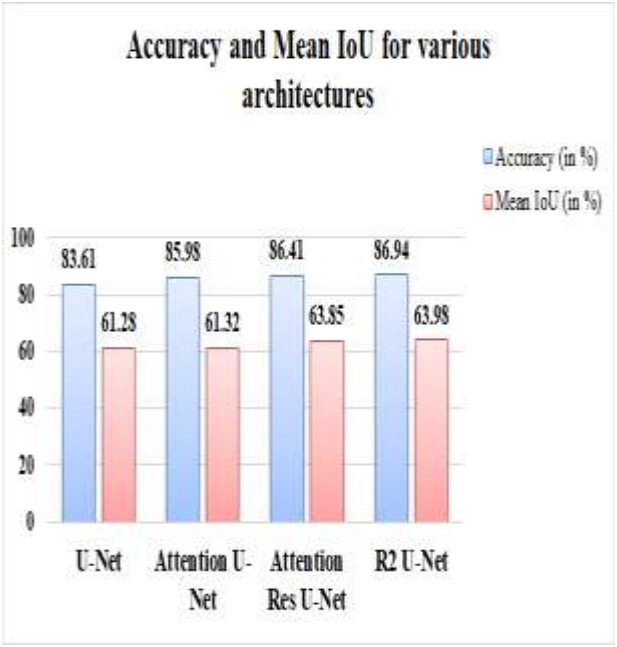


Figure 15. *Comparison of accuracy and meanIoU for various architectures.*

Table 2. *Metric comparison across various U-Net architectures*

Name of the architecture	No. of Epochs	Training loss	Validation loss	% of Accuracy	% of Mean IoU (mIoU)
U-Net	10	0.4754	1.905	53.64	39.84
Attention U-Net		0.4594	2.124	53.68	39.98
Attention Res U-Net		0.4215	1.786	53. 71	42.97
Recurrent Residual U-Net		0.4229	1.734	53.89	43.39
U-Net	100	0.1904	0.6520	83.61	61.28
Attention U-Net		0.1135	0.4573	85.98	61.32
Attention Res U-Net		0.1057	0.4420	86.41	63.85
Recurrent Residual U-Net		0.1037	0.4352	86.94	63.98
U-Net	150	0.0925	0.5854	85.43	62.12
Attention U-Net		0.0889	0.5354	86.27	61.30
Attention Res U-Net		0.0839	0.5054	87.75	63.72
Recurrent Residual U-Net		0.0825	0.5016	87.92	63.86

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Feng, X., Jiang, Y., Yang, X., Du, M., Li, X. (2019) Computer vision algorithms and hardware implementations: A survey. *Integration 69*, 309–320

[2] Vu, H., Le, T.L., Nguyen, V.G., Dinh, T.H. (2018) Semantic regions segmentation using a spatio-temporal model from an uav image sequence with an optimal configuration for data acquisition. *Journal of Information and Telecommunication* 2(2), 126–146

[3] Sardooi, E.R., Azareh, A., Choubin, B., Barkhori, S., Singh, V.P., Shamshirband, S. (2019) Applying the remotely sensed data to identify homogeneous regions of watersheds using a pixel-based classification approach. *Applied Geography* 111, 102071

[4] Elharrouss, O., Al-Maadeed, S., Subramanian, N., Ottakath, N., Almaadeed, N., Himeur, Y. (2021) *Panoptic segmentation: A review. arXiv preprint* arXiv:2111.10250

[5] Manickam, R., Kumar Rajan, S., Subramanian, C., Xavi, A., Eanoch, G.J., Yesudhas, H.R. (2020) Person identification with aerial imaginary using segnet based semantic segmentation. *Earth Science Informatics* 13, 1293–1304

[6] Hao, S., Zhou, Y., Guo, Y. (2020) A brief survey on semantic segmentation with deep learning. *Neurocomputing 406*, 302–321

[7] Ulku, I., Akag¨und¨uz, E. (2022) A survey on deep learning-based architectures for semantic segmentation on 2d images. *Applied Artificial Intelligence* 36(1), 2032924

[8] Mo, Y., Wu, Y., Yang, X., Liu, F., Liao, Y. (2022) Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing 493*, 626–646

[9] Lv, Z., Wang, F., Cui, G., Benediktsson, J.A., Lei, T., Sun, W. (2022) Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 60, 1–12

[10] Safavi, F., Rahnemoonfar, M. (2022) Comparative study of real-time semantic segmentation networks in aerial images during flooding events. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16, 15–31

[11] Zhang, X., Ma, W., Li, C., Wu, J., Tang, X., Jiao, L. (2019) Fully convolutional network-based ensemble method for road extraction from aerial images. *IEEE Geoscience and Remote Sensing Letters* 17(10), 1777–1781

[12] Javanmardi, S., Ashtiani, S.-H.M., Verbeek, F.J., Martynenko, A. (2021) Computer-vision classification of corn seed varieties using deep convolutional neural network. *Journal of Stored Products Research 92*, 101800

[13] Wang, J., Fu, P., Gao, R.X. (2019) Machine vision intelligence for product defect inspection based on deep learning and hough transform. *Journal of Manufacturing Systems 51*, 52–60

[14] Zhou, T., Ruan, S., Canu, S. (2019) A review: Deep learning for medical image segmentation using multi-modality fusion. *Array* 3, 100004

[15] Yuan, X., Shi, J., Gu, L. (2021) A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications* 169, 114417

[16] Lateef, F., Ruichek, Y. (2019) Survey on semantic segmentation using deep learning techniques. *Neurocomputing 338*, 321–348

[17] Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K. (2022) Medical image segmentation using deep learning: A survey. *IET Image Processing* 16(5), 1243–1267

[18] Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and *Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 Springer

[19] Erdem, F., Avdan, U. (2020) Comparison of different u-net models for building extraction from high-resolution aerial imagery. International Journal of *Environment and Geoinformatics* 7(3), 221–227

[20] Soni, A., Koner, R., Villuri, V.G.K. (2020). M-unet: Modified u-net segmentation framework with satellite imagery. In: *Proceedings of the Global AI Congress* 2019, pp.47–59 Springer

[21] Zhang, H., Liu, M., Wang, Y., Shang, J., Liu, X., Li, B., Song, A., Li, Q. (2021) Automated delineation of agricultural field boundaries from sentinel-2 images using recurrent residual u-net. *International Journal of Applied Earth Observation and Geoinformation* 105, 102557

[22] Wang, Y., Peng, Y., Li, W., Alexandropoulos, G.C., Yu, J., Ge, D., Xiang, W. (2022) Ddu-net: Dual-decoder-u-net for road extraction using high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 60, 1–12

[23] Hou, J., Hou, B., Zhu, M., Zhou, J., Tian, Q. (2023) Sensitivity analysis of parameters of u-net model for semantic segmentation of silt storage dams from remote sensing images. *Canadian Journal of Remote Sensing 49(*1), 2178834

[24] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086

[25] Jetley, S., Lord, N.A., Lee, N., Torr, P.H. (2018) Learn to pay attention. *arXiv preprint arXiv*:1804.02391

[26] Sun, Y., Bi, F., Gao, Y., Chen, L., Feng, S. (2022) A multi-attention unet for semantic segmentation in remote sensing images. *Symmetry* 14(5), 906

[27] Tang, M., Georgiou, K., Qi, H., Champion, C., Bosch, M. (2023) Semantic segmentation in aerial imagery using multi-level contrastive learning with local consistency. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3798–3807

[28] Abdollahi, A., Pradhan, B., Alamri, A.M. (2022) An ensemble architecture of deep convolutional segnet and unet networks for building semantic segmentation from high-resolution aerial images. *Geocarto International* 37(12), 3355–3370

[29] Wang, S., Hou, X., Zhao, X. (2020) Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block. *IEEE Access 8*, 7313–7322

[30] Abdollahi, A., Pradhan, B. (2021) Integrating semantic edges and segmentation information for building extraction from aerial images using unet. *Machine Learning with Applications* 6, 100194

[31] Benjdira, B., Bazi, Y., Koubaa, A., Ouni, K. (2019) Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing* 11(11), 1369

[32] Niu, R., Sun, X., Tian, Y., Diao, W., Chen, K., Fu, K. (2021) Hybrid multiple attention network for semantic segmentation in aerial images. *IEEE Transactions on Geoscience and Remote Sensing* 60, 1–18

[33] Su, Z., Li, W., Ma, Z., Gao, R. (2022) An improved u-net method for the semantic segmentation of remote sensing images. *Applied Intelligence* 52(3), 3276–3288

[34] Marmanis, D., Wegner, J.D., Galliani, S., Schindler, K., Datcu, M., Stilla, U. (2016) Semantic segmentation of aerial images with an ensemble of cnns. ISPRS Annals of the Photogrammetry, *Remote Sensing and Spatial Information Science*s 3, 473–480

[35] Li, W., He, C., Fang, J., Fu, H. (2018) Semantic segmentation based building extraction method using multi-source gis map datasets and satellite imagery. In: Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 238–241

[36] Erdem, F., Bayram, B., Bakirman, T., Bayrak, O.C., 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

[37] Yuan, K., Zhuang, X., Schaefer, G., Feng, J., Guan, L., Fang, H. (2021) Deep-learning-based multispectral satellite image segmentation for water body detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14*, 7422–7434

[38] Saifi, M.Y., Singla, J., et al. (2019) Deep learning based framework for semantic segmentation of satellite images. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 369–374 (2020). IEEE

[39] Avenash, R., Viswanath, P.: Semantic segmentation of satellite images using a modified cnn with hard-swish activation function. In: *VISIGRAPP* (4: VISAPP), pp. 413–420

[40] Saralioglu, E., Gungor, O. (2022) Semantic segmentation of land cover from high resolution multispectral satellite images by spectral-spatial convolutional neural network. *Geocarto International* 37(2), 657–677

[41] Brand, A., Manandhar, A. (2021) Semantic segmentation of burned areas in satellite images using a u-net-based convolutional neural network. The International Archives of the Photogrammetry, *Remote Sensing and Spatial Information Sciences* 43, 47–53

[42] Singh, N.J., Nongmeikapam, K. (2023) Semantic segmentation of satellite images using deep-unet. *Arabian Journal for Science and Engineering* 48(2), 1193–1205

[43] Wang, X., Hu, Z., Shi, S., Hou, M., Xu, L., Zhang, X. (2023) A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved unet. *Scientific Reports* 13(1), 7600

[44] Maurya, A., Mittal, P., Kumar, R., et al. (2023) A modified u-net-based architecture for segmentation of satellite images on a novel dataset. *Ecological Informatics 7*5, 102078

[45] 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

[46] Maji, D., Sigedar, P., Singh, M. (2022) Attention res-unet with guided decoder for semantic segmentation of brain tumors. *Biomedical Signal Processing and Control* 71, 103077

[47]Resnet:. https://d2l.ai/chapter convolutional-modern/resnet.html A*ccessed accessed 05-August-2022*

[48] Xiang, L., Li, Y., Lin, W., Wang, Q., Shen, D. (2018). Unpaired deep cross-modality synthesis with fast training. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th *International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings* 4, pp. 155–164 Springer

[49] He, K., Zhang, X., Ren, S., Sun, J. (2016). Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: *14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645 Springer

[50] Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K. (2019) Recurrent residual u-net for medical image segmentation. *Journal of Medical Imaging* 6(1), 014006–014006