2024 IEEE 2nd International Conference on Information and Communication Technology (ICICT 2024)
21-22 October 2024, Dhaka, Bangladesh

Semantic Segmentation in Satellite Imagery: An Attentive U-Net Approach

Nafisa Islam[1], Md. Farhad Hossain[2], Md. Azad Hossain[3]
Department of Electronics and Telecommunication Engineering (ETE)[1,2,3]
Chittagong University of Engineering & Technology (CUET), Chattogram 4349, Bangladesh[1,2,3]
u1808006@student.cuet.ac.bd[1], farhad.hossain@cuet.ac.bd[2], azad@cuet.ac.bd[3]

2024 2nd International Conference on Information and Communication Technology (ICICT) | 979-8-3315-0822-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICICT64387.2024.10839725

***Abstract*—Semantic segmentation is a computer vision technique that assigns a distinct class label to every pixel in an image, generating a segmentation map where each pixel belongs to a specific category. This enables various applications, such as automatic and in-depth land cover and land use analysis, tracking urbanization, deforestation, road, and building detection, disaster response, traffic management, city planning, and environmental monitoring. Encoder-decoder-based architectures excel in semantic segmentation tasks than traditional CNN-based methods because they can extract hierarchical features, combine contextual and spatial information, and accurately reconstruct segmentation maps. In this paper, Attention U-Net, where the attention mechanism is introduced to the modified version of U-Net, is implemented for semantic segmentation in satellite images. Instead of just focusing on extracting features, the proposed method emphasizes making accurate segmentation output. The segmentation quality of each model was assessed using performance metrics, including accuracy, precision, recall, F1-score, and IoU score. The findings suggest that the Attention U-Net model surpasses other variations, attaining a higher IoU score of 0.7959 and exhibiting strong segmentation abilities across several land cover categories, such as water bodies, land, roads, buildings, and vegetation. Furthermore, by visual analysis of the results, it becomes apparent that the Attention U-Net is remarkably proficient at accurately capturing intricate details in satellite imagery.**

***Index Terms*—Satellite, Computer vision, Semantic Segmentation, Deep Learning, Attention U-Net, Intersection over Union(IoU)**

I. Introduction

Semantic segmentation is a technique that categorizes images at the pixel level, making them accessible to computers, robots, and other devices. As urbanization and environmental changes accelerate, the need for precise and automated analysis of satellite images becomes increasingly important for applications such as urban planning, deforestation monitoring, and disaster management. Pixel-wise semantic segmentation accurately discerns and interprets object appearance and dividing lines, demonstrating spatial correlation between elements within a single image, which can be visualized in Fig. 1.

Complex images can be difficult to segment due to localized information that the model may not understand. Satellite image segmentation divides an image into pixels by identifying boundaries like straight lines and curves. Image segmentation labels each pixel, assigning the same label to pixels with related properties. This paper focused on encoder-decoder architectures, which have shown good accuracy and performance. More sophisticated and advanced deep learning models, such as U-Net [1] and DeepLabV3 [2], are useful in this context. While U-Net was initially developed for the purpose of segmenting biological images [3], it has subsequently been utilized for several other tasks, including satellite image segmentation in numerous instances [1].

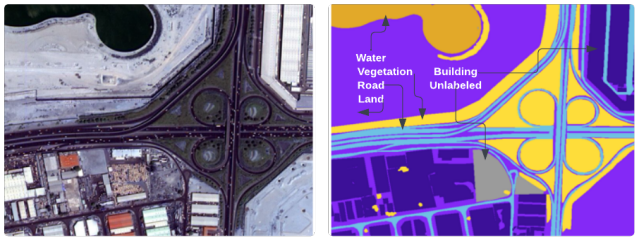

Fig. 1: Example of semantic segmentation of RGB satellite image.

The primary contributions of the suggested methodology can be stated as follows:

1) Development of a simple yet effective U-Net-based model that is efficient enough to properly segment satellite images where all classes are detected accurately.
2) Enhance the IoU score by utilizing the Attention mechanism and compare the performance of the proposed method with the existing methods.

This manuscript systematically unfolds our proposition. Section 2 dissects existing semantic segmentation methods, Section 3 explores used datasets, and Section 4 unveils our framework. Section 5 assesses its efficacy, while Section 6 briefly summarizes our findings.

II. Related Works

Various endeavors have been undertaken to create a semantic segmentation technique for satellite images, but only a few have produced satisfactory results. Therefore, this area of research is now in its early stages and has gained significant attention as a highly investigated subject in the field of computer vision.

Deep learning-based semantic segmentation techniques have shown superior performance to conventional machine learning and statistical approaches mentioned in the study [4]. Deep

learning models can acquire intricate and hierarchical representations, allowing structures to be extracted from high-resolution satellite images. Out of other deep learning models, U-Net shows promising results, acquiring the highest IoU score. The authors in [1] presented deep learning techniques for semantic segmentation, demonstrating that U-Net outperformed the PSPNet architecture with a Mean-IoU score of 0.51.

In the research [5], a modified U-Net architecture is used for segmentation problems where 1×1 convolution is applied in several stages to improve the accuracy. Ming et al. [6] introduce a new segmentation model using AD-LinkNet, a network based on U-Net that introduces an attention mechanism and a series-parallel combination of dilated convolution. The study [7] proposes a new U-Net structure incorporating self-attention features and separate convolutions. The study [8] introduces repurposed Robust U-Net architecture, incorporating an attention mechanism using the Squeeze-and-Excitation block for segmentation tasks. The SE block is employed in the bottleneck section of the U-Net design to extract the most significant features.

Pre-trained models have been incorporated as the backbone of U-Net architecture to extract features extensively. Patil et al. [9] and Maithil et al. [10] have utilized this feature to achieve higher accuracy in their research. In the cases of application of semantic segmentation, such as monitoring deforestation [11], ship detection [12], etc., also used some pre-trained models to detect the targeted class accurately. Given that segmentation in satellite images is a prerequisite for an encoder-decoder-based architecture, U-Net appears to meet this requirement. Additionally, U-Net excels in adopting finer details due to its easy implementation and phase-simple architecture, allowing easy modification.

III. Dataset

The dataset used for this paper consists of aerial images of Dubai acquired from the Mohammed Bin Rashid Space Centre (MBRSC) satellites. It comprises 72 images organized into six larger tiles, thoroughly depicting Dubai's metropolitan landscape. Each image tile includes corresponding mask tiles, as given in Fig. 2. The semantic segmentation labels for each image include the following classes: Building, Land (unpaved area), Road, Vegetation, Water, and an additional Unlabeled class. The dataset provides intuitive color representations for each class to improve its convenience.

IV. Proposed Method

In this section, we explain our proposed deep learning-based semantic segmentation technique in satellite images and its preprocessing, as illustrated in Fig. 3, which leverages the attention mechanism in a modified U-Net architecture.

A. Image Pre-processing

The image and mask tiles were too large to use in a model, so they were resized to 256×256 using Patchify. Then, we carefully created color representations for every class in the dataset. The RGB format represented colors and was an intuitive identifier for six semantic classes. The mask patches were labeled using one hot encoding that mapped colors to corresponding labels. Then, the image dataset and labeled mask dataset were created, which then split into 80% for training and 20% for testing.

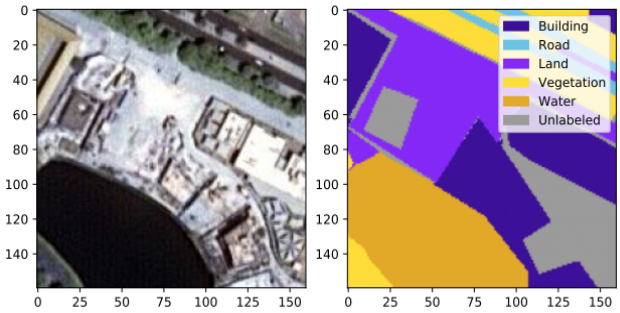

Fig. 2: RGB input image tile along with its corresponding mask containing six unique labels.

B. Encoder Layer

The modified U-Net architecture, where the attention mechanism was implemented, consists of three parts- encoder, bottleneck, and decoder. This encoder part reduces input image size and extracts hierarchical features by downsampling. Our framework has four blocks with significant layers of two convolutional layers, ReLU activation layers, and batch normalization layers. Max-pooling is used to encode feature maps and reduce their size.

C. Bottleneck Layer

The architecture has a bottleneck-like structure in the center, minimizing spatial resolution before progressively increasing it in the decoder path. It consists of a convolutional layer with highest number of filter.

D. Decoder Layer

The decoder path recovers the spatial resolution of features downsampled in the contracting path by upsampling. First, our architecture constructs an upsampled feature map using a transposed convolution. By doubling feature spatial resolution, this technique enhances detail recovery. Concatenation establishes a skip connection, enabling the model to upsample both low-level and high-level data. This allows the model to include contextual information while maintaining local details.

E. Attention gate

Attention gates are introduced in the skip connection of our modified U-Net architecture. It receives duplicate feature maps (represented by x_1) along with gating signals (represented by g_1) from the upscaling path. To effectively merge low-level and high-level features, intermediate layers with reduced feature maps, labeled as F_{int}, are generated through 1x1 convolutions. Subsequently, a stride convolution halves the spatial dimensions of x_1 to align with the coarser

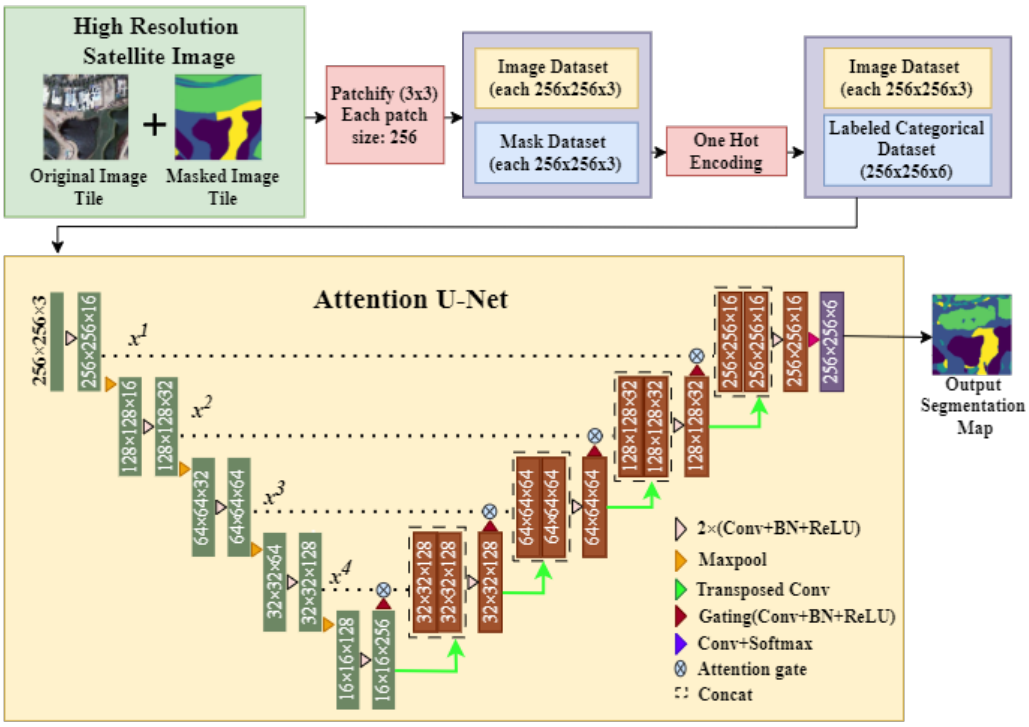

Fig. 3: The figure illustrates an overall workflow for semantic segmentation in satellite images.

scale of the gating signal. These intermediate layers are fused using ReLU activation, denoted as σ_1, followed by a 1x1 convolution and the ψ operation to compute attention coefficients. Normalization of attention coefficients

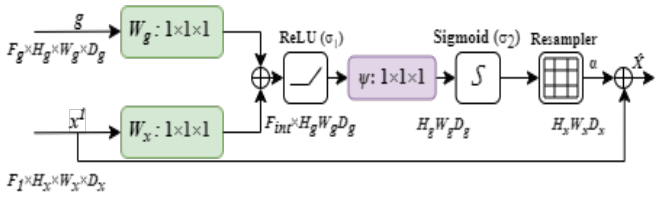

Fig. 4: Attention gate mechanism in detail. [13]

is achieved through a Sigmoid activation function, denoted as σ_2. This meticulous process enables the extraction of attention coefficients for each pixel (i) and each layer (l), enhancing the model's capacity for semantic segmentation tasks. The equations denoting this are given in equation (1)-(5).

$$u = W_x^T \times x_i^l \tag{1}$$

$$v = W_g^T \times g_i^l \tag{2}$$

$$U = \sigma_1 \times (u + v + b_g) \tag{3}$$

$$V = \sigma_2 \times (\psi^T \times (U) + b_\psi) \tag{4}$$

$$\alpha_i^l = V \tag{5}$$

After sigmoid activation, attention coefficients (α) appear. The attention coefficients are utilized to restructure the input feature map (x_l). This is achieved by calculating the product of each element in x with its corresponding attention coefficient α. The outcome is the feature map that has been weighted, denoted as (x_i^l). The attention gate enhances the network's ability to concentrate on the most significant characteristics inside the skip connections, resulting in an enhanced segmentation performance. It allows the network to prioritize important features by assigning them higher weights, suppressing irrelevant background noise, and emphasizing significant features that lead to precise segmentation.

F. Final Layer

The final output layer uses a 1x1 convolutional layer with a kernel size of (1,1), generating feature maps with pixel probability distributions. The softmax activation function normalizes output across classes, allowing the model to forecast pixels. The output layer integrates predictions with output formats and provides segmentation maps with 6 classes.

V. RESULT ANALYSIS

A. Evaluation Metrics

We evaluate the quality of output segmentation maps using Intersection over Union score (IoU), accuracy, precision, recall, and f1 score.

1) Intersection over Union (IoU): The Intersection over Union (IoU) is an important metric for evaluating segmentation accuracy. It calculates the difference between predicted

and ground truth segmentation maps by comparing their intersection to their union. Intersection refers to the overlapping area where two segments meet or cross one another and union covers the entirety of land from both distinct regions. IoU simply measures the similarities between actual and predicted masks. The mathematical expression for IoU is as follows:

$$IoU = \frac{Area\ of\ Overlap\ (Intersection)}{Area\ of\ Union} \tag{6}$$

B. Results of proposed methodology

The proposed methodology's effectiveness is evaluated through quantitative metrics and visual inspection. Here, we present the results of the proposed semantic segmentation technique applied to test images and compare them with different methods.

TABLE I: Performance comparison of different U-Net models on the test set

Models	Accuracy	IoU	Precision	Recall	f_1-score
Modified U-Net	0.8366	0.6609	0.8481	0.8251	0.8365
U-Net + ResNet34	0.8651	0.7351	0.8729	0.8585	0.8656
U-Net + DenseNet121	0.8820	0.7430	0.8920	0.8687	0.8805
U-Net + VGGNet16	0.9126	0.7846	0.9327	0.8933	0.9124
Attention U-Net (proposed)	0.8966	**0.7959**	0.9017	0.8921	0.8969

To thoroughly assess semantic segmentation models used in satellite image processing, we compared our proposed method with other implemented methods: Modified U-Net, U-Net with- ResNet34, DenseNet121, and VGGNet16. Table I comprehensively analyzes the performance indicators derived from the experiments. The Attention U-Net model demonstrated exceptional boundary definition accuracy from the table compared to other U-Nets, as evidenced by its IoU score of 0.7959. This model utilizes attention mechanisms to concentrate on pertinent parts within the image, hence improving its capacity to capture intricate characteristics and minimizing instances of incorrect identifications. The image Fig.5 displays the segmentation results generated by different U-Net models and compared with our proposed model. From the visual analysis, it can be seen that Attention U-Net has outperformed other variants of U-Net. When other variants failed to detect some of the classes properly, attention U-Net detected all the classes nearly perfectly.

C. Comparison of Related Works

We have developed a modified U-Net model with an attention mechanism, as our proposed model. Its performance is compared to the methods used by several researchers. Table II compares proposed methods with the existing ones. The U-Net architecture was modified in the works of authors cited in [9] and [10] by changing its encoder path. Pre-trained models were specifically used as the encoder to improve the accuracy of feature extraction. Similarly, Aburaed et al. [8] incorporated SE blocks, often referred to as attention mechanisms, into the bottleneck section of the U-Net. These techniques prioritize improving feature extraction rather than achieving precise segmentation outputs, which is essential for

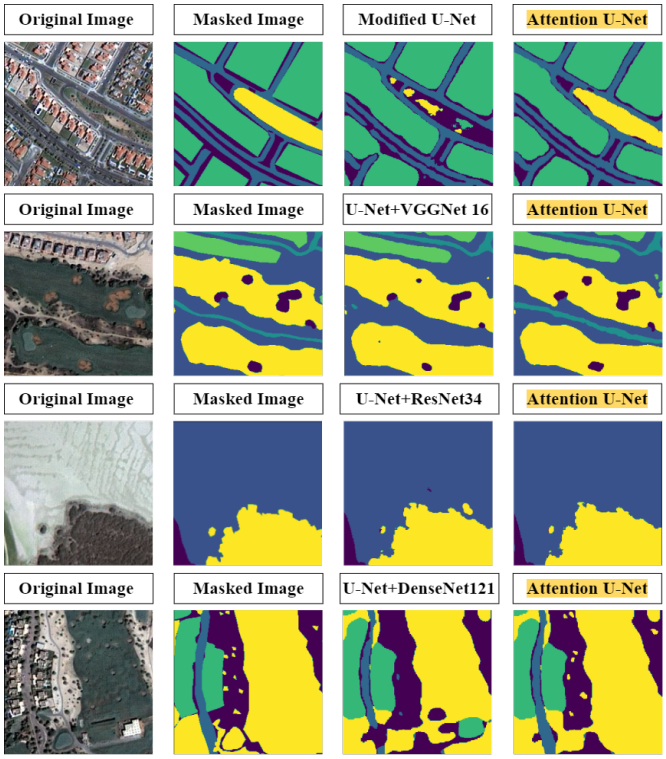

Fig. 5: Comparison of the visual segmentation results between different U-Net models vs the Proposed model for different satellite Images with their corresponding true masks.

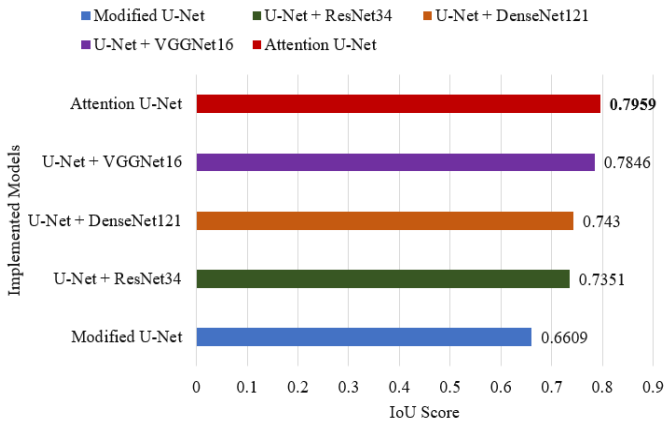

Fig. 6: Comparative analysis of the proposed method with other U-Net methods based on IoU score.

semantic segmentation. The technique described in reference [7] combines self-attention and separable convolution. This method yielded a praiseworthy IoU score, but it also resulted in notable mistakes and attenuation in the resulting segmentation map. Furthermore, it exhibited a lack of precision in identifying specific categories within the segmentation map. Our proposed method, in contrast, achieves a higher level of overall accuracy of 86.99% and an IoU score of 79.59%, which are the greatest among the existing methods described in

TABLE II: Comparative analysis of the proposed method with the existing models based on Accuracy and IoU score.

Methods	Accuracy	IoU
U-Net+InceptionResNetV2 [9]	87.00%	-
DensePlusU-Net [10]	86.11%	-
FCN [7]	71.40%	61.80%
CNN [7]	79.40%	70.20%
DeepLabV3 [7]	81.50%	71.90%
Dense+U-Net [7]	81.76%	72.43%
SA-SC U-Net [7]	87.34%	77.45%
RU-Net [8]	86.68%	48.42%
SEU-Net [8]	88.49%	49.00%
SERU-Net [8]	87.06%	47.81%
Attention U-Net (Proposed)	**89.66%**	**79.59%**

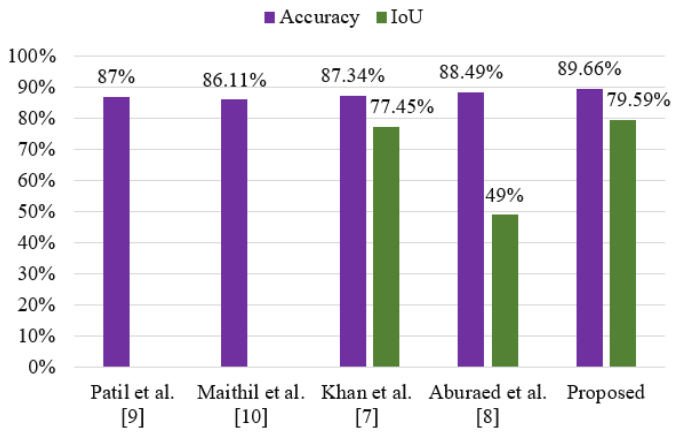

Fig. 7: Comparative analysis of the proposed method with other research approaches.

this context. This technique successfully identified all classes by prioritizing the creation of accurate output segmentation maps. The improvement of skip connections by introducing attention gates is credited. Skip connections are crucial in semantic segmentation as they enable the reconstruction of the segmented output. A visual representation of this comparison is shown in Fig.7.

VI. Conclusion and Future work

Semantic segmentation is a crucial tool in remote sensing applications, enabling precise mapping of various land use types and land cover. The increasing resolution, frequency, and satellite imagery coverage highlight the need for reliable and precise semantic segmentation methods. In this study, the Attention U-Net model performed better than the other variations of U-Net: modified U-Net, U-Net+ResNet34, U-Net+DenseNet121, U-Net+VGGNet16, obtaining the greatest IoU score. The attention mechanism in the Attention U-Net model not only improved feature extraction but also produced almost accurate segmentation maps. It effectively captured delicate details and nuances. Compared to other existing methods, Attention U-Net consistently performed better, as indicated by its higher IoU score and accuracy. This study highlights the importance of attention mechanisms in enhancing the precision and effectiveness of semantic segmentation models for satellite data, providing vital insights into remote sensing and laying the foundation for further developments in semantic segmentation approaches for satellite imagery analysis.

References

[1] Chaurasia, Kuldeep, Nandy, Rijul, Pawar, Omkar, Singh, Ravi Ranjan, and Ahire, Meghana, *Semantic segmentation of high-resolution satellite images using deep learning*, *Earth Science Informatics*, vol. 14, no. 4, pp. 2161–2170, 2021. Springer. Available at: https://link.springer.com/article/10.1007/s12145-021-00674-7.

[2] Chen, Liang-Chieh, Papandreou, George, Kokkinos, Iasonas, Murphy, Kevin, and Yuille, Alan L., *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017. IEEE. Available at:https://ieeexplore.ieee.org/abstract/document/7913730.

[3] Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas, *U-net: Convolutional networks for biomedical image segmentation*, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015, pp. 234–241. Springer. Available at:https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28.

[4] Singh, Tarun Pratap, Singh, Ravi Ranjan, Himanshu, A Mishra, and Sharma, Nidhi, *Semantic segmentation of satellite images: A survey*, *International Research Journal of Engineering and Technology (IRJET)*, vol. 7, no. 12, 2020.

[5] Singh, Ningthoujam Johny, and Nongmeikapam, Kishorjit, *Semantic segmentation of satellite images using deep-unet*, *Arabian Journal for Science and Engineering*, vol. 48, no. 2, 2023, pp. 1193–1205. Springer. Available at: https://link.springer.com/article/10.1007/s13369-022-06734-4.

[6] Wu, Ming, Zhang, Chuang, Liu, Jiaming, Zhou, Lichen, and Li, Xiaoqi, *Towards accurate high resolution satellite image semantic segmentation*, *IEEE Access*, vol. 7, 2019, pp. 55609–55619. IEEE. Available at: https://ieeexplore.ieee.org/abstract/document/8700168.

[7] Khan, Bakht Alam and Jung, Jin-Woo, *Semantic Segmentation of Aerial Imagery Using U-Net with Self-Attention and Separable Convolutions*, *Applied Sciences*, vol. 14, no. 9, pp. 3712, 2024. MDPI. Available at: https://www.mdpi.com/2076-3417/14/9/3712.

[8] Aburaed, N., Al-Saad, M., Alkhatib, M.Q., Zitouni, M.S., Almansoori, S., and Al-Ahmad, H., *Semantic Segmentation of Remote Sensing Imagery Using An Enhanced Encoder-Decoder Architecture*, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 10, pp. 1015–1020, 2023. Copernicus Publications, Göttingen, Germany.Available at: https://isprs-annals.copernicus.org/articles/X-1-W1-2023/1015/2023/isprs-annals-X-1-W1-2023-1015-2023.pdf.

[9] Patil, Dhanishtha, Patil, Komal, Nale, Rutuja, and Chaudhari, Sangita, *Semantic Segmentation of Satellite Images using Modified U-Net*, *2022 IEEE Region 10 Symposium (TENSYMP)*, 2022, pp. 1–6. IEEE. [Online; accessed 15-May-2024]. Available at: https://doi.org/10.1109/TENSYMP54529.2022.9864504.

[10] Maithil, Keerti and Rehman, Tasneem Bano, *Semantic Segmentation of Urban Area Satellite Imagery Using DensePlusU-Net*, *2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)*, 2022, pp. 1–6. IEEE. [Online; accessed 15-May-2024]. Available at: https://doi.org/10.1109/CCET56606.2022.10080484.

[11] Alzu'bi, Ahmad, and Alsmadi, Lujain, *Monitoring deforestation in Jordan using deep semantic segmentation with satellite imagery*, *Ecological Informatics*, vol. 70, pp. 101745, 2022. Elsevier. [Online; accessed 15-May-2024]. Available at: https://www.elsevier.com.

[12] Hordiiuk, Dariia, Oliinyk, Ievgenii, Hnatushenko, Volodymyr, and Maksymov, Kostiantyn, *Semantic Segmentation for Ships Detection from Satellite Imagery*, *2019 IEEE 39th International Conference on Electronics and Nanotechnology (ELNANO)*, 2019, pp. 454–457. IEEE. [Online; accessed 15-May-2024]. Available at: https://doi.org/10.1109/ELNANO.2019.8783822.

[13] Oktay, Ozan, Schlemper, Jo, Folgoc, Loic Le, Lee, Matthew, Heinrich, Mattias, Misawa, Kazunari, Mori, Kensaku, McDonagh, Steven, Hammerla, Nils Y, Kainz, Bernhard, and others, *Attention U-Net: Learning where to look for the pancreas*, *arXiv preprint arXiv:1804.03999*, 2018. arXiv. Available at: https://arxiv.org/abs/1804.03999. s