Towards U-Net based Semantic Segmentation for Satellite images

Chitraksh Singh
ABV-IIITM
Gwalior, Madhya Pradesh, INDIA
mtics_202306@iiitm.ac.in

Deepak Kumar Dewangan
ABV-IIITM
Gwalior, Madhya Pradesh, INDIA
deepakd@iiitm.ac.in

***Abstract*—Semantic segmentation in satellite imagery is a critical process that involves categorizing each pixel into predefined classes such as buildings and roads. The challenge lies in developing accurate automated methods capable of efficiently classifying and mapping diverse types of land cover. Traditional approaches work very well on sample data. However, with a change in the data set, performance was found to be limited. In this direction a semantic segmentation based deep learning architecture, U-Net with ResNet having versions 50, 101, and 152 has been considered which helps to find the significant feature. For this research, Landcover.ai and Dubai landcover dataset have been used, which contains high-resolution images annotated with classes buildings, roads, land, water, and forests. Adam optimizer and Jaccard loss function were considered to train our model with data augmentation. Our experiments demonstrated that this approach achieves a mean IoU score of 80.20% and 82.95% in the test data set.**

***Index Terms*—Semantic segmentation, data augmentation, ResNet, Landcover.ai dataset, Dubai landcover, Unet, IoU score.**

I. Introduction

The mapping and monitoring of land cover are critical components of geographic information systems (GIS) and environmental management [9]. Accurate land cover maps are essential for a variety of applications, such as agriculture management, urban planning, forestry, and disaster response [18]. Semantic segmentation significantly improves border security by detecting unauthorized activities, monitoring infrastructure, and identifying smuggling routes. In disaster management, it helps assess damage, flood mapping, fire monitoring, and early warning systems [22]. These applications improve decision making, optimize resource allocation, and enable rapid, cost-effective responses to security and disaster scenarios. Traditionally, these maps are generated through manual interpretation of aerial or satellite imagery, which is slow and resource-exhaustive. Recent advances in deep learning and computer vision, particularly in semantic segmentation, have opened new avenues for automating this process with high precision and efficiency [16].

Through semantic segmentation, one assigns every pixel in an image to a designated predefined category [17]. Since then, considerable advances have been made in the field due to the development of convolutional neural networks (CNNs) [14]. The U-Net architecture, works best for biomedical image segmentation and has proven highly effective in various segmentation tasks due to its symmetric encoder-decoder structure that captures both contextual and spatial information [5]. Incorporating powerful backbones, trained in large datasets, further enhances segmentation performance by leveraging rich feature representations [6].

This research exploits the LandCover.ai dataset [2], which provides a comprehensive collection of high-resolution aerial imagery that covers rural areas in Poland cover 216.27 square kilometers. The dataset is meticulously annotated with four key land cover classes buildings, roads, water, and woodlands. Additionally, the Dubai dataset [7], which includes an vegetation feature instead of woodlands. This dataset serves as an excellent dataset to evaluating the effectiveness of advanced deep learning models.

In the proposed approach an U-Net architecture with ResNet-50, 101, 152 backbones [13], trained on ImageNet [11], to perform semantic segmentation on the LandCover.ai dataset has been consider to identify the significant feature in the dataset. Adam optimizer and a categorical focal Jaccard as loss function is used, with the objective to maximize the Intersection over Union (IoU) score.

Our results indicate that the proposed approach achieves a mean IoU score of 80.20% and 82.95% on the test set, demonstrating its potential for practical deployment in land cover mapping.

II. Related works

Recent progress in semantic segmentation for satellite imagery has been largely fueled by the adoption of deep learning methods. A pivotal advancement in the field was the introduction of the Fully Convolutional Network (FCN) by Long et al. in 2015 [16], which replaced fully connected layers with convolutional ones, enabling end-to-end pixel-wise classification. This innovation laid the groundwork for the development of more sophisticated segmentation models, including those applied to satellite image analysis.

Ronneberger et al. (2015) proposed the U-Net architecture [20], initially developed for biomedical image segmentation, and it has been widely applied to a range of segmentation tasks, including satellite imagery. Its design features a symmetric encoder-decoder structure with skip connections, enabling effective capture of contextual and spatial information, which

979-8-3503-7952-5/24/$31.00 ©2024 IEEE

is particularly advantageous for high-resolution satellite data. However, the original U-Net architecture has limitations in utilizing multi scale feature maps, which are crucial for achieving detailed segmentation outputs, especially in complex landscapes.

Researchers have improved U-Net's feature extraction capabilities by integrating pretrained backbone networks such as ResNet. Introduced by He et al. (2016) [10], ResNet excels in various computer vision tasks due to its deep residual learning framework. By incorporating ResNet as the backbone, U-Net allows us to obtain feature representations from large-scale datasets such as ImageNet, leading to better segmentation accuracy in satellite images. For example, Audebert et al. (2018) demonstrated notable performance improvements by combining ResNet with U-Net for urban scene segmentation.

The LandCover.ai dataset [2], which is utilized in this research, has primarily been explored using the DeepLab architecture. However, its performance with U-Net and other architectures has not been extensively investigated. This study proposes that employing U-Net with ResNet as the backbone can be highly effective for segmentation tasks on this dataset, potentially offering enhancements over existing approaches.

III. Material and Method

This section details the methodology used for semantic segmentation of the LandCover.ai dataset, employing a U-Net architecture. The methodology consist data preparation, model architecture, data augmentation, and training procedures.

A. Dataset

The LandCover.ai dataset includes high-resolution aerial imagery as in Fig. 1 from rural areas in Poland, annotated with four land cover classes buildings, woodlands, water, and roads and there landcover area. Dubai landcover contains building, land, road, vegetation, water, unlabeled as in Fig. 2. The dataset is split into training, validation, and test sets. Before being input into the model, each image and its corresponding mask undergo preprocessing. The masks, initially in grayscale with integer values representing distinct classes, are converted into one-hot encoded format to enable training with categorical cross-entropy loss, while also maintaining a detailed count of class pixels. Various augmentation techniques are applied to increase the diversity of the training data and enhance the model's robustness, including random horizontal flips of images and saturation adjustments.

B. Model Architecture

The chosen architecture is a U-Net with a ResNet versions [15] as backbones in Fig. 3. The U-Net model is particularly well-suited for segmentation tasks due to its symmetric encoder-decoder structure [20], which captures contextual information and preserves spatial details. The encoder section of the U-Net consists of a pre-trained ResNet [19], which has been trained on the ImageNet dataset [8]. This backbone extracts high-level features from the input images through

Fig. 1: Sample image from landcover.ai dataset

Fig. 2: Sample image from dubai landcover dataset

a series of convolutional and downsampling layers. ResNet-50 balances the depth and performance, ResNet-101 variant improves extraction of complex features and ResNet-152 variant capture very fine-grained details and complex feature. The ResNet consists of an initial convolutional layer followed by multiple residual blocks consisting Relu as an activation function [23], which help in learning deep representations while avoiding the vanishing gradient problem. Max pooling is applied after the ReLU activation. The decoder section of the U-Net is composed of upsampling layers that gradually restore the spatial resolution of the feature maps [4]. After each upsampling step, we concatenate the corresponding encoder feature maps with the upsampled feature maps via skip connections to maintain fine-grained details. The final layer of the decoder performs a convolution to map the feature maps to the desired number of output classes, using a softmax activation function for pixel-wise classification. The model is optimized with the Adam optimizer, which adjusts the learning rate dynamically during training to enhance convergence. The loss function employed is a combination of categorical focal loss and Jaccard loss, which helps in handling class imbalance and improving segmentation accuracy. Jaccard Loss quantifies the overlap between predicted and true masks by directly optimizing the Intersection over Union (IoU) metric.

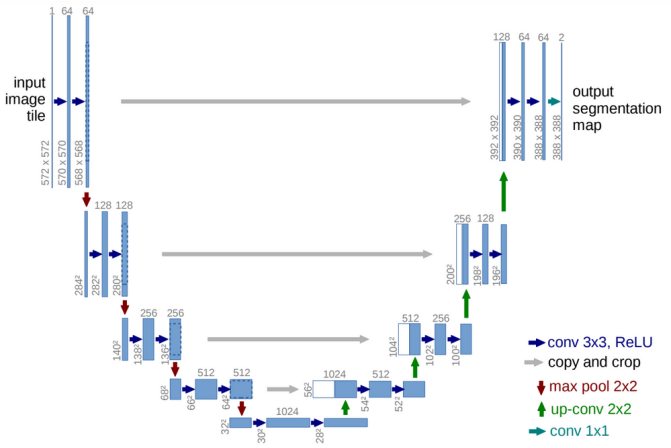


Fig. 3: Structure of U-Net architecture.

IV. Experimental Setup

A. Dataset preparation

The LandCover.ai dataset contains high resolution images of buildings, roads, water, and woodlands using aerial imagery. It includes land cover data from Poland and features three spectral bands (RGB). The dataset includes 33 orthophotos and 8 orthophotos , covering a total area of 216.27 km², were as dubai dataset contains 8 tiles each containing 9 images with there masks. The dataset categorizes land cover into four classes, labeled by pixel values: buildings (1), woodlands (2), water (3), and roads (4). The area distribution includes 1.85 km² of buildings, 3.5 km² of roads, 13.15 km² of water, and 72.02 km² of woodlands and in dubai dataset instead of woodlands, land is labelled as (2). Firstly the dimensions of images are slice into factor of 512 pixels so that it can be divided into 512*512 pixels precisely as in Fig. 4 . Secondly data have less than 5% of information were not useful as they have no background information has been removed.

B. Implementation details

We utilized the PyTorch library to perform the experiment on an NVIDIA Tesla P100 GPU. The Adam optimizer was employed with a weight decay of 0.01 and a batch size of 8, while the stride value was set to 2. The number of steps per epoch was calculated by dividing the total number of training images by the batch size. Similarly, the validation steps per epoch is also calculated and categorical focal jaccard as loss function [12]. The training uses data augmentations, given in Table I and results given in Fig. 5. Unet architecture contains backbones as ResNet-50 , ResNet-101, ResNet-152 and ImageNet is use as encoder weights for all, having 32521685, 51513813, 67157461 trainable parameter respectively, for all version base learning rate is set at $5*10^{-5}$ and trained for the 50 epochs with L2 regularization with a penalty value of 10^{-6}. The input layer specifies the height, width, and channels of an image, followed by batch normalization and zero padding, and ReLu [1] as an activation function Eq.1 defined as the

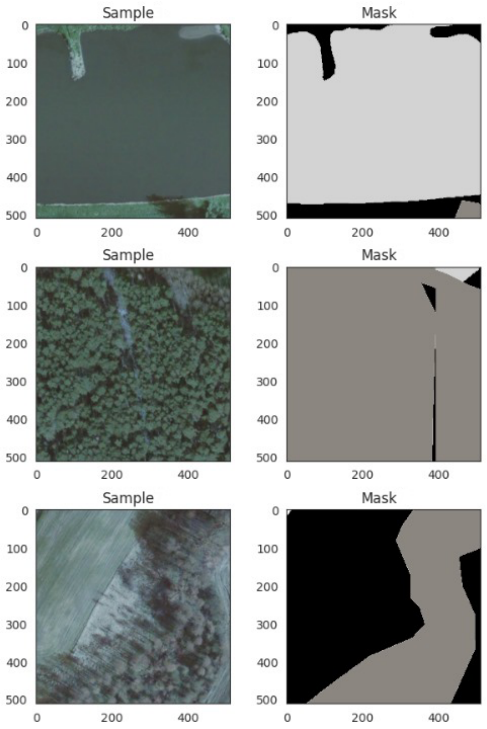


Fig. 4: Sample image and mask from training data

non-negative portion of its argument, where x represents the input to a neuron.

$$f(x) = \max(0, x) = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The final layer of the decoder employs the softmax activation function [21], as described in Eq. 2, for multiclass classification. Specifically, the standard (unit) softmax function $\sigma : \mathbb{R}^K \to (0,1)^K$, where $K \geq 1$, takes a vector $\mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$ and calculates each component of the resulting vector $\sigma(\mathbf{z}) \in (0,1)^K$ with

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2}$$

C. Results

To assess the model's performance, we will calculate the Jaccard index (IoU) [3] on the validation dataset. This metric assesses the ratio of the overlapping area between the predicted and ground truth masks to the total area covered by both masks, as shown in Eq.3. Specifically, it is calculated by dividing the number of true positive values by the total of true positives, false positives, and false negatives. The evaluation will employ a pixel-wise Jaccard index, as described.

$$IoU_j = \frac{\sum_{i=1}^{n} TP_{ij}}{\sum_{i=1}^{n} TP_{ij} + \sum_{i=1}^{n} FP_{ij} + \sum_{i=1}^{n} FN_{ij}} \tag{3}$$

where IoU_j represents the IoU score for each pixel associated with class j across a total of n images. In this

Fig. 5: Data augmentations results.

context, TP_{ij}, FP_{ij}, and FN_{ij} indicate the true positive, false positive, and false negative pixels of class j in each image i, respectively. The final score is the average IoU across all classes, as shown in Eq. 4, where k denotes the total number of classes.

$$mIoU = \frac{1}{k}\sum_{j=1}^{k} IoU_j \tag{4}$$

Fig. 6: Result of ResNet-50

Fig. 7: Result of ResNet-101

Fig. 8: Result of ResNet-152

The results obtained on the test set using IoU and loss metrics shown in Fig. 9 . In figure Fig. 8 shows close-ups of

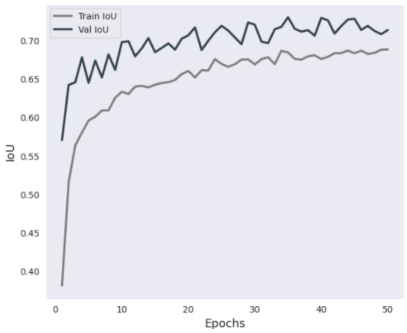
TABLE I: Data Augmentation Information Table

Data Augmentation	**Parameter(s)**	**Effect on Image**
Gaussian Blur	blur_limit=(3, 7), p=1	Smooths the image and reduces sharpness.
Elastic Transform	alpha=1, sigma=50, alpha_affine=None, p=1	Simulates real-world deformations.
Grid Distortion	p=1	Alters the grid-like structure of pixels.
Optical Distortion	distort_limit=0.5, shift_limit=0.5, p=1	Adds lens or optical distortion effects.
ShiftScaleRotate	shift_limit=0.0625, scale_limit=0.1, rotate_limit=45, p=1	Ensures spatial invariance.
Channel Shuffle	p=1	Simulates varying sensor sensitivities.
CLAHE	p=1	Improves local contrast and detail detection.
ISONoise	p=1	Simulates sensor noise in images.
Coarse Dropout	max_holes=8, max_height=16, max_width=16, p=1	Simulates missing parts or occlusions.
Motion Blur	blur_limit=7, p=1	Simulates motion blur effects.
Random Fog	fog_coef_lower=0.1, fog_coef_upper=0.3, alpha_coef=0.08, p=1	Simulates foggy weather conditions.
Random Rain	slant_lower=-10, slant_upper=10, drop_length=20, drop_color=(200, 200, 200), p=1	Adds rain streaks and effects.
Random Snow	snow_point_lower=0.1, snow_point_upper=0.3, p=1	Adds snow to simulate winter conditions.
Solarize	p=1	Alters bright regions, adding artistic distortions.
Equalize	p=1	Redistributes pixel intensity.
Invert Image	p=1	Inverts pixel intensity, dark becomes light, and vice versa.
Posterize	num_bits=4, p=1	Lowers image quality by reducing pixel color depth.
Random Sun Flare	flare_roi=(0.0, 0.0, 1.0, 0.5), angle_lower=0.0, p=1	Simulates sun flare and lighting anomalies.
Random Shadow	shadow_roi=(0.0, 0.5, 1.0, 1.0), p=1	Adds random shadows to the image.
Random Brightness & Contrast	brightness_limit=0.2, contrast_limit=0.2, p=1	Introduces lighting and exposure variations.

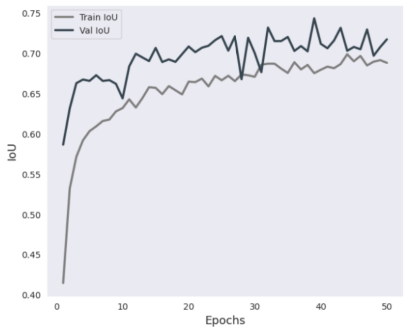
images, their corresponding labels, and the resulting accurate segmentations. Roads and buildings pose the most significant challenges for semantic segmentation due to their narrow (roads) or small (buildings) structures. Consequently, due to fewer inner pixels like imprecise edges of classes makes accurate classification more difficult. U-Net with ResNet model achieves mIoU score on the entire test set shown in Table II. Using a smaller output stride results in improved performance, and further applying augmentation techniques enhances the metrics, reaching a final mIoU score of 80.20% and accuracy Eq.5 of 92.73% shown in Fig. 10 by ResNet-152. Performance matrix having precision of 88.954% Eq.6, recall of 88.248% Eq.7 and f1-score of 88.580% Eq.8. ResNet-50 and ResNet-

TABLE II: Quantitative Comparison of Classes with performance matrix of landcover.ai dataset

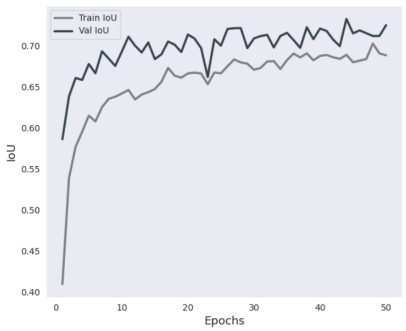
	ResNet-50			ResNet-101			ResNet-152		
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
background	0.7694	0.7452	0.7549	0.8021	0.7463	0.7732	0.7470	0.7749	0.7607
building	0.9036	0.9140	0.9088	0.9231	0.8935	0.9081	0.9101	0.9080	0.9090
woodland	0.9076	0.7990	0.8499	0.8123	0.9150	0.8606	0.8756	0.8445	0.8597
water	0.9505	0.9408	0.9457	0.9417	0.9539	0.9477	0.9425	0.9491	0.9458
road	0.9320	0.9651	0.9482	0.8999	0.9616	0.9297	0.9725	0.9359	0.9538



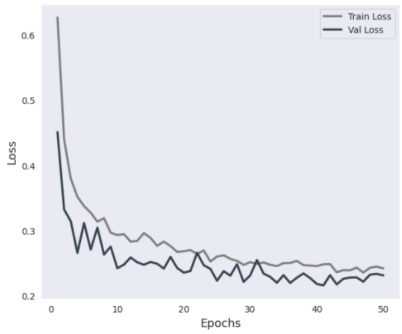
(a) Graph of Resnet-50 IoU score



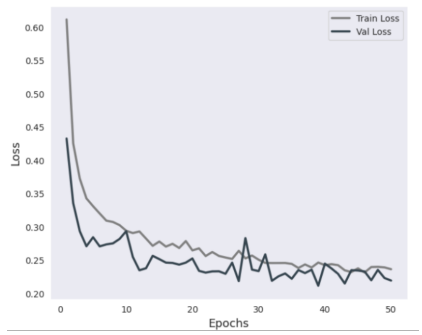
(b) Graph of Resnet-101 IoU score



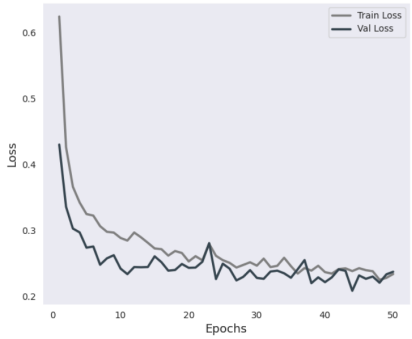
(c) Graph of Resnet-152 IoU score



(d) Graph of Resnet-50 loss



(e) Graph of Resnet-101 loss



(f) Graph of Resnet-152 loss

Fig. 9: Overall Result of ResNet version on landcover.ai dataset

101 were performing below from ResNet-152 with mIoU score of 79.53% and 79.73% respectively, prediction is shown in Fig. 7 and Fig. 8 for landcover.ai dataset and for dubai dataset mIoU score is 82.95% and accuracy is 84.95%.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

V. Conclusion

This research explores the use of a U-Net model with a ResNet variants backbones for land classification using the LandCover.ai dataset. The aim is to evaluate its effectiveness

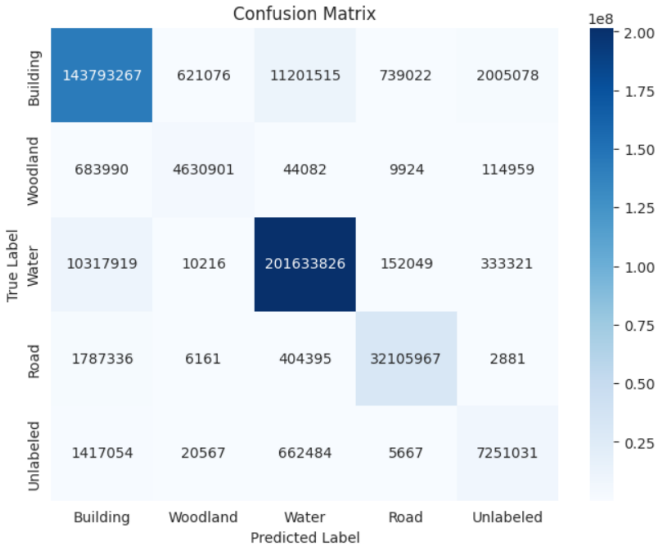


Fig. 10: Resnet 152 confusion matrix of landcover.ai dataset

in potential applications on edge devices in space. The model showed promising results, generating usable outputs. However, some limitations were observed, including inaccuracies in pixel classification for specific classes (e.g., misjudging the size of a road area) and errors in assigning the correct terrain type to certain pixel groups.

For future research, exploring alternative segmentation models could enhance the accuracy of the proposed system. Additionally, making the model more lightweight and efficient for deployment on edge devices in orbit is crucial. These devices have limited processing power and memory, so it's essential to explore techniques for compressing models to run effectively on smaller processors.

Utilizing diverse datasets that simulate the image capturing capabilities of satellites would be advantageous for improving the model. Addressing more complex challengeslike lower resolutions and cloud coverage would enhance the system's robustness in managing diverse inputs. Additionally, expanding the model's classification system to cover a broader range of terrain types and diverse locations is recommended for enhancing its overall performance and applicability.

Finally, using the model from this study could facilitate the development of a comprehensive solution that processes input images and dynamically manages storage based on the relevance of the currently stored images. Additionally, incorporating evaluation metrics such as identifying the presence of interesting subjects, evaluating image quality through super-resolution, and ensuring a high diversity of classes within an image could further enhance classification performance.

REFERENCES

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019.

[2] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery, 2022.

[3] Alexey Bokhovkin and Evgeny Burnaev. Boundary loss for remote sensing imagery semantic segmentation. In Huchuan Lu, Huajin Tang, and Zhanshan Wang, editors, *Advances in Neural Networks – ISNN 2019*, pages 388–401, Cham, 2019. Springer International Publishing.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

[5] Ramiel Deticio, Argel Bandala, John Anthony Jose, Ronnie Concepcion II, Mark Angelo Purio, Edwin Sybingco, and Richard Tan Ai. Application of a u-net segmentation model in land cover classification for use in automated data prefiltering onboard nanosatellites. In *Application of a U-Net Segmentation Model in Land Cover Classification for Use in Automated Data Prefiltering Onboard Nanosatellites*, pages 71–75, 10 2023.

[6] Deepak Kumar Dewangan and Yogesh Rathore. Image quality estimation of images using full reference and no reference method, 2011.

[7] Samy Ismail Elmahdy and Mohamed Mostafa Mohamed. Monitoring and analysing the emirate of dubai's land use/land cover changes: an integrated, low-cost remote sensing approach. *International Journal of Digital Earth*, 11(11):1132–1150, 2018.

[8] Muhammad Fayaz, Junyoung Nam, L. Minh Dang, Hyoung-Kyu Song, and Hyeonjoon Moon. Land-cover classification using deep learning with high-resolution remote-sensing imagery. *Applied Sciences*, 14(5), 2024.

[9] Gianluca Furano, Gabriele Meoni, Aubrey Dunne, David Moloney, Veronique Ferlet-Cavrois, Antonis Tavoularis, Jonathan Byrne, Léonie Buckley, Mihalis Psarakis, Kay-Obbe Voss, and Luca Fanucci. Towards the use of artificial intelligence on the edge in space systems: Challenges and opportunities. *IEEE Aerospace and Electronic Systems Magazine*, 35, 12 2020.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Shashikant Rangnathrao Kale, Chandrakant Madhukar Kadam, Raghunath Sambhaji Holambe, and Rajan Hari Chile. Land cover classification using modified u-net: A robust approach for satellite image analysis. In S. Manoharan, Alexandru Tugui, and Zubair Baig, editors, *Proceedings of 4th International Conference on Artificial Intelligence and Smart Energy*, pages 135–146, Cham, 2024. Springer Nature Switzerland.

[12] Yongkyu Lee, Woodam Sim, Jeongmook Park, and Jungsoo Lee. Evaluation of hyperparameter combinations of the u-net model for land cover classification. *Forests*, 13(11), 2022.

[13] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M. Atkinson. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[14] Zewen Li, Wenjie Yang, Shouheng Peng, and Fan Liu. A survey of convolutional neural networks: Analysis, applications, and prospects, 2020.

[15] Jiazhi Liang. Image classification based on resnet. *Journal of Physics: Conference Series*, 1634:012110, 09 2020.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015.

[17] Lucas Osco, José Junior, Ana Paula Ramos, Lucio Jorge, Sarah Narges Fatholahi, Jonathan Silva, Edson Matsubara, Hemerson Pistori, Wesley Gonçalves, and Jonathan Li. A review on deep learning in uav remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 102:102456, 07 2021.

[18] Pratibha Pandey, Kranti Kumar Dewangan, and Deepak Kumar Dewangan. Enhancing the quality of satellite images using fuzzy inference system. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pages 3087–3092, 2017.

[19] Nikhil Prakash, Andrea Manconi, and Simon Loew. Mapping landslides on eo data: Performance of deep learning models vs. traditional machine learning models. *Remote Sensing*, 12:346, 01 2020.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[21] Raghuram S, Anirudh S Bharadwaj, Deepika S K, Mridula S Khadabadi, and Aditya Jayaprakash. Digital implementation of the softmax activation function and the inverse softmax function. In *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*, pages 64–67, 2022.

[22] Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. Torchgeo: Deep learning with geospatial data, 2022.

[23] Libo Wang, RUI LI, Chenxi Duan, Ce Zhang, Xiaoliang Meng, and Shenghui Fang. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, PP:1–1, 01 2022.