Research Article GU J Sci, Part A, 12(2): 479-502 (2025) 10.54287/gujsa.1664093

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Comparison of Deep Learning Algorithms for Image Segmentation on Satellite Images

Huseyin ACEMLI[1] Nida KUMBASAR[2*]

[1] *Ozyegin University, İstanbul, Türkiye*
[2] *TÜBİTAK, Informatics and Information Security Research Center (BİLGEM), Kocaeli, Türkiye*

Keywords

Remote Sensing
High-Spatial Resolution
Semantic Segmentation
Deep Learning

Abstract

Recent advancements in deep learning have significantly contributed to the development of high spatial resolution (HSR) land cover mapping. However, the distinct geographic patterns between urban and rural areas have limited the generalizability of deep learning algorithms across these domains. To address this challenge, separate datasets for rural and urban environments have been proposed in the literature, aiming to achieve more reliable results in real-world applications. In this study, we utilize the publicly available LoveDA HSR dataset for model and parameter comparison. Experiments were conducted on two distinct scenarios: rural and urban areas. The combination of the Adam optimizer, Dice loss function, and UNet++ architecture exhibited the highest performance in both datasets. A weighted average of this combination, based on the number of test samples, was calculated for both groups, yielding a final performance score of 62.14% in terms of mean Intersection over Union (IoU).

Cite

Acemli, H., & Kumbasar, N. (2025). Comparison of Deep Learning Algorithms for Image Segmentation on Satellite Images. *GU J Sci, Part A*, *12*(2), 479-502. doi:10.54287/gujsa.1664093

Author ID (ORCID Number)

0009-0004-1804-4834 Huseyın ACEMLI
0000-0001-5497-4618 Nida KUMBASAR

Article Process

Submission Date 25.03.2025
Revision Date 17.04.2025
Accepted Date 26.05.2025
Published Date 30.06.2025

1. INTRODUCTION

Image analysis in remote sensing is becoming more and more widespread as artificial intelligence advances in the field of computer vision. In addition, with the help of various sensors, satellite vehicles and platforms, remote sensing technology is renewing itself day by day. Remote sensing image segmentation aims to segment images with semantic labels. Deep learning (DL) algorithms with a representation learning approach have been used for many segmentation problems including remote sensing in recent years and these algorithms have been observed to be very useful.

High spatial resolution (HSR) land-cover datasets are land-cover data created using high-resolution satellite imagery or aerial photography. HSR data allow the study of fine details and small areas, usually on images with a resolution of 1 meter or higher. Such data play an important role in various fields such as urban planning (Zou et al., 2024), environmental management(Zhang et al., 2023), agricultural traceability (Aksoy et al., 2023), climate change (Abunnasr & Mhawej, 2023), natural disaster assessment (Xia et al., 2023). HSR data can be collected from spaceborne sensors such as Sentinel 2, Landsat or airbone sensors such as LiDAR.

*Corresponding Author, e-mail: nida.kumbasar@tubitak.gov.tr

Spaceborne sensors are devices usually located on satellites and collect data by observing the Earth from space, while airborne sensors are sensors usually located on aircraft such as airplanes, helicopters or drones. Spaceborne sensors are in space, usually on orbiting satellite systems. Because they are well positioned on fixed platforms, spaceborne sensors have fewer problems with degradation than airbone sensors and offer continuous monitoring of large areas. Visiting a given area at regular intervals, these sensors can be affected by the state of the atmosphere, but are weather independent. On the downside, they sometimes provide coarse resolution data as an image can cover several hundred square kilometers.

In the literature, it has been emphasized that spatial resolution is more important than spectral resolution when extracting urban land cover information from remote sensing image (RSI) data (Neupane et al., 2021). It has been concluded that an image pixel with good resolution is more useful for analyzing RSI data than a larger number of spectral bands or a narrower wavelength range (Myint et al., 2011). The rapid accumulation and availability of high-resolution RSI and the advancement of DL methods have shifted HSR data from traditional pixel-based and object-based methods to DL-based semantic segmentation. Semantic segmentation is an approach based on assigning a class label to each image pixel in order to make the image highly interpretable. Land cover semantic segmentation in remote sensing aims to identify the type of land cover in each image pixel. The main objective of the remote sensing image segmentation process is to divide the image into a homogeneous set of segments according to certain criteria and map the separate regions to real world objects such as buildings, rivers, fields, roads, etc.(Long et al., 2015).

In traditional pixel-based approaches, pixel size may be insufficient to identify an object in an image. Although significant progress has been made with convolutional neural network (CNN)-based DL, the high intra-class variation and low sensor resolution of remote sensing images make segmentation difficult (Chan et al., 2021). To overcome these challenges, various strategies such as hierarchical feature structures (Tao et al., 2020), polymorphism (Peng et al., 2019) and fusion schemes (Yu et al., 2018) have been used to extend the pipeline. Kemker et al. (Kemker et al., 2018) improved performance with a hybrid architecture based on SharpMask and RefineNet for six-band multi-spectral image segmentation. Kampffmeyer et al. (Kampffmeyer et al., 2016) performed land cover mapping segmentation in urban areas by combining Monte Carlo dropout uncertainty maps with three CNN models. Sun and Wang (Sun & Wang, 2018) utilized a digital surface model based on geometry information to improve the segmentation results of HSR images with a fully convolutional network (FCN). Yuan et al. (Yuan et al., 2021) designed a multichannel fusion module for water body detection using RGB, NIR and SWIR bands of Sentinel-2. To cope with different image resolutions, they proposed that the multichannel water body detection network (MC-WBDN) is more robust to changes in light and weather conditions and can better distinguish small water bodies compared to other models. Baek et al. (Baek et al., 2024) implemented a modified U-Net architecture, called SiU-Net, with two separate inputs for Sentinel-2's RGB and NIR data. They also performed a comparison with DeepLabV3+ and U-Net to evaluate the performance of their model. Hossain et al. (Hossain & Chen, 2022) introduced a hybrid segmentation algorithm that integrates homogeneity and heterogeneity simultaneously to identify buildings. In that

algorithm, since no prior knowledge about the shape of building footprints is required, all building shapes are included in the analysis and a donut-filling technique is introduced to extract roof elements. The algorithm resulted in homogeneity within segments and heterogeneity between segments. Finally, the study segmented small and large buildings without using scale or object size parameters.

The development of convolutional networks with existing HSR land cover datasets such as Gaofen Image Dataset (Kampffmeyer et al., 2016), DeepGlobe (Demir et al., 2018), which contain pixel-wise information for remote sensing applications, has been encouraged. However, these datasets ignored the different styles between the geographical areas of urban and rural areas. Wang et al. (J. Wang et al., 2021)prepared the LoveDA dataset based on the knowledge that urban and rural land covers have large differences in class distributions, object scales and pixel spectra, and aimed to improve model generalizability separately. In the HSR images of different cities in China, they observed different patterns, especially in buildings, roads and wetlands. For example, they emphasized that buildings in rural areas are more irregularly arranged than in urban areas, and roads in rural areas are narrower than in urban areas. They noted that agricultural areas are large-scale and continuous in rural areas, while in urban areas they are found in the spaces between buildings. Wetlands are located in the form of small-scale ponds and ditches in rural areas and large-scale rivers and lakes in urban areas. In their study, they emphasized that UNet++ performs better than UNet due to its nested structure.

In this study, the LoveDA dataset was used to compare the most preferred models and parameters in the literature for urban and rural areas. UNet, UNet++, MANet and DeepLabv3+ were used as architectures; Dice Loss, Cross Entropy Loss and Weighted Cross Entropy Loss were used as loss functions; Adam and Stochastic Gradient Descent (SGD) were used as optimizers.

The rest of the paper is organized as follows: Section 2 describes the dataset, DL architectures, loss functions and optimizers used in the study. Section 3 presents the performance evaluation criteria, the experimental results and the comparison between the experimental groups. Section 4 discusses the limitations, with a comparison to the literature, and emphasizes the importance of the work. Finally, conclusions and future work are presented in Section 5.

2. MATERIAL AND METHOD

This section describes the dataset, DL approaches, loss functions and optimizers used in the study.

2.1. Dataset

In this study, we utilize the publicly available LoveDA dataset (J. Wang et al., 2021), introduced in the literature by Wang et al. in 2021. The LoveDA dataset was collected in July 2016 from 18 diverse urban and rural scenes in Nanjing, Changzhou, and Wuhan, covering a total area of 536.15 km². The dataset comprises 5,987 high spatial resolution (HSR) images, each with a size of 1024x1204 pixels and a spatial resolution of

0.3 meters. The original dataset contains 166,768 objects and is divided into training, validation, and test sets. However, since ground truth (GT) masks for the test data were unavailable for performance evaluation, this study focuses on the training and validation data. To this end, the training and validation dataset was shuffled and then randomly split into 80% for training and 20% for testing. As summarized in Table 1, a total of 4,191 RGB remote sensing image-mask pairs were used in this study.

Table 1. *Image distribution in training and test sets*

	Train	Test	**Total**
Rural Dataset	1886	472	2358
Urban Dataset	1466	367	1833
Total	3352	839	4191

Sample images and corresponding masks from rural and urban scenes, considering the segmentation task with a total of 8 classes, including the "No Data" class, are presented in Figure 1.

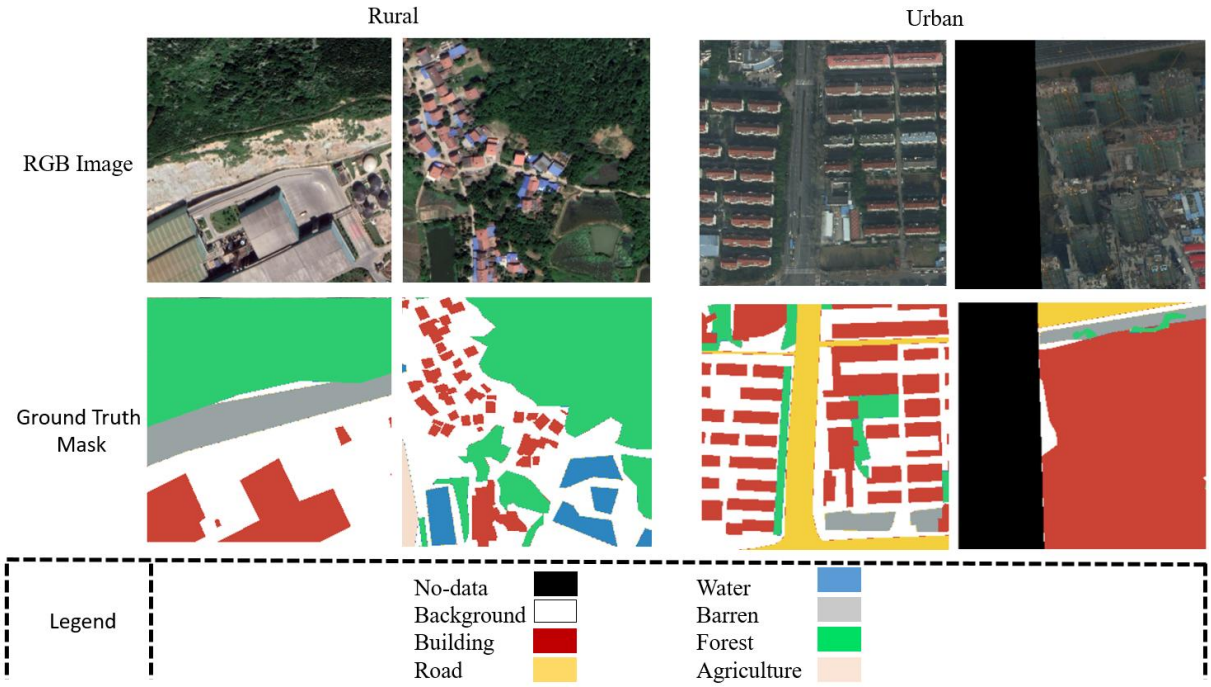

Figure 1. *Example images and corresponding segmentation masks from the LoveDA Dataset*

Wang et al. (J. Wang et al., 2021) emphasized that objects of the same category have completely different patterns in different scenes. In their dataset, urban and rural scenes have different class distributions. Urban scenes with high population density contain many artificial objects such as buildings and roads, while rural scenes contain more natural elements such as forests and water.

2.2. Method

The models, loss functions and optimization algorithms used for comparison in the study are briefly mentioned in this section.

2.2.1. Models

CNN architectures are able to distinguish land covers by classifying pixels thanks to their ability to learn spatial relationships in images. UNet, UNet++, MANet, and DeepLabv3+ DL models, which are frequently preferred in the literature for segmentation in HSR data, are briefly described in this section.

UNet:

The UNet architecture (Ronneberger et al., 2015) consists of an encoder block that coarsens the image resolution and a decoder block that increases the image resolution. The UNet architecture consists of convolution layers, pooling layers, feature fusion layer, upsampling layer and softmax layer. The structure of the encoder and decoder parts is symmetric with jump connections between them, which positively affects fine-grained segmentation. At the same time, UNet is able to preserve the feature maps with the same size as the original image.

UNet++:

UNet++ (Zhou et al., 2018) is developed by integrating multi-depth UNet(Ronneberger et al., 2015) models and linking all encoder and decoder blocks with the same resolution. The decoders in the UNet++ architecture combine multi-scale feature maps at the same and different image resolutions. Interwoven with jump links, this architecture exploits the UNet structure at different depths. Thus, low-level attributes are preserved in complex tasks. Furthermore, these jump links help to bridge the semantic gap between the attribute maps of encoders and decoders in the UNet model.

MANet:

Multiscale attention network, MANet (Fan et al., 2020), is a multiscale network that can capture local attributes with their global dependencies based on the attention mechanism. It consists of two blocks: position-wise attention block (PAB) and multiscale fusion attention block (MFAB). The detection of spatial and channel dependencies of the feature maps of the PAB and MFAB blocks is based on the self-attention mechanism. While the PAB is used to model attribute dependencies capturing spatial dependencies between pixels, the MFAB handles channel dependencies between any attribute map with multiscale semantic attribute fusion. In addition to high-level, MFAB also takes into account the channel dependencies of low-level attribute maps. Finally, the channel dependencies of high and low level attribute maps are fused to obtain multiscale semantic information and improve network performance.

DeepLabv3+:

DeepLabv3+ (Chen et al., 2018) is roughly based on adding a decoder module to DeepLabv3 (Chen et al., 2017) and extending DeepLabv3 to improve segmentation results along object boundaries. DL uses a spatial pyramid pooling module or an encoding-decoder structure for semantic segmentation. The spatial pyramid

pooling module can multi-scale encode incoming features at multiple rates with filters or pooling operations, while the encoder-decoder architecture can capture sharper object boundaries by gradually recovering spatial information. DeepLabv3+ combines the advantages of both approaches. By applying depth-separable convolution to both atrous spatial pyramid pooling and decoder modules, DeepLabv3+ achieves a faster and more powerful encoder-decoder network.

2.2.2. Losses

Loss functions play an important role in determining model performance. However, it is not possible to decide on a single universal loss function, especially for multi-class segmentation problems (Jadon, 2020). In this study, Dice Loss, Cross Entropy Loss and Weighted Cross Entropy Loss, which are widely used in segmentation applications, are compared.

Dice Loss

The Dice coefficient is a widely used metric for calculating the similarity between two images. It was later adapted as a loss function known as Dice Loss (Sudre et al., 2017). For multi-class problems, the calculation of Dice Loss over the i-th instance in the dataset is presented in Equation 1:

$$L_{Dice}(y,\hat{y}) = \frac{1}{C}\sum_{k=1}^{C}\left(1 - \frac{2\sum_i (y_{i,k}\hat{y}_{i,k})}{\sum_i y_{i,k} + \sum_i \hat{y}_{i,k}}\right) \quad (1)$$

In Equation 1, y is the true value, $\hat{y}$ is the predicted result and C is the number of classes.

Cross Entropy (CE) Loss

Cross entropy (CE) is a measure of the difference between two probability distributions for a given random variable or set of events (Yi-de et al., 2004). CE loss is widely used especially in classification problems and is the most preferred loss function in segmentation, since segmentation is considered as pixel-level classification. The CEL calculation is presented in Equation 2 for the i-th sample in the dataset,

$$L_{CE}(y,\hat{y}) = -\sum_{k=1}^{C} y_{i,k} log\left(\hat{y}_{i,k}\right) \quad (2)$$

where, y is the true value, $\hat{y}$ is the predicted result and C is the number of classes.

Weighted Cross Entropy (WCE) Loss

Weighted cross entropy (WCE) is a variant of binary cross entropy in which positive samples are weighted by coefficients (Pihur et al., 2007). WCE loss is calculated over the i-th sample in the dataset as in Equation 3,

$$L_{WCE}(y,\hat{y}) = -\sum_{k=1}^{C} w_k\, y_{i,k} log\left(\hat{y}_{i,k}\right) \quad (3)$$

where y is the actual value, $\hat{y}$ is the predicted result and w_k is the weighting coefficient for the k-th class. In case of class imbalance, larger weights are assigned for rare classes.

2.2.3. Optimizer approaches

In DL algorithms, optimizers are used to obtain the best match between the actual values and the estimated outputs (Bottou, 2010). The difficulties in satellite imagery can be taken into account in the optimization of CNN-based segmentation, allowing the development of more accurate and generalizable models (Pan et al., 2020) (H. Wang et al., 2022). In this study, experiments were conducted with Adam and SGD approaches. An overview of these optimizers is presented in this section.

Adam

The Adam optimizer (Kingma & Ba, 2014) is an iterative optimization algorithm used to minimize the loss function during the training of neural networks. In Adam optimizer, faster convergence is guaranteed by bias in the early stages of training and the training process is stabilized. Adam helps faster convergence by adjusting the learning rate for each parameter and is therefore preferred for problems with sparse gradients or noisy data (Goodfellow, 2016).

Stochastic Gradient Descent (SGD)

Gradient descent (GD) is one of the basic optimization algorithms used to minimize the loss of the model. GD tries to minimize the loss function in the learning process. The model parameters are updated and the learning process is performed. SGD (Bottou, 2010) introduces randomness into the optimization process by randomly selecting a data point at each step instead of using all data points when calculating the gradient. The size of the batch size, the frequency of parameter updates and the convergence process are determined in the gradient descent (Kingma & Ba, 2014).

3. RESULTS

In this section, evaluation criteria and experimental results are presented. Resnet50 (He,2016) was used as the backbone in all models and the input size was set to 256x256 to reduce computational complexity. In all scenarios pretrained Resnet50 weights were used and fine-tuned according to the datasets.

3.1. Evaluation criteria

In segmentation tasks, accuracy is a commonly used metric to evaluate the proportion of correctly classified pixels to the total number of pixels in the image. However, accuracy can be misleading in imbalanced datasets, where the model might classify the majority class correctly while neglecting smaller, less frequent regions. So such as IoU (Intersection over Union) to get a more comprehensive understanding of model performance, especially in cases of class imbalance.

IoU metric is commonly used to evaluate the performance of predicted regions in segmentation tasks (J. Wang et al., 2021). IoU is computed as the ratio of the intersection of the predicted and GT segmentations to the union of these areas. This metric provides a quantitative measure of how well the predicted masks align with the GT masks.

In this study, detailed results are presented on iou and accuracy is used as a support.

The formulas for Accuracy and IoU are presented in Equation 4-5, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

The definitions of True Positive (TP), False Positive (FP), and False Negative (FN) in Equation 4-5 are as follows: TP are pixels for which the model correctly predicts the target region. FP are pixels where the model positively predicts a region that does not actually exist. FN are pixels where the model incorrectly predicts a region with a positive true label.

3.2. Experimental results

In this section, the results of the experimental studies conducted separately for rural and urban categories are presented and compared using Intersection over Union (IoU) as the evaluation metric.

3.2.1. Rural Dataset

This section presents the results for the rural regions of the dataset. Figure 2 illustrates the IoU values of four models and three loss functions on the test data, using the Adam optimizer, while Figure 3 presents the IoU values obtained with the SGD optimizer.

Rural landscapes exhibit gradual transitions between land cover types such as forest, agriculture, and barren land. These transitions often lack clear geometric boundaries, requiring models to generalize broader spatial patterns rather than focus on sharply defined features. In this context, as illustrated in Figure2, UNet++ consistently outperforms the other models across all loss functions, particularly under Dice Loss (IoU = 0.6601). This suggests that UNet++'s nested architecture and dense skip connections are well-suited for modeling the multi-scale, continuous transitions often found in rural regions such as irregular forest edges or variable agricultural fields. Conversely, models like DeepLabv3+ and MANet, which rely on dilated convolutions and attention mechanisms respectively, may struggle with such gradual class boundaries, potentially overemphasizing texture or failing to capture context over large areas.

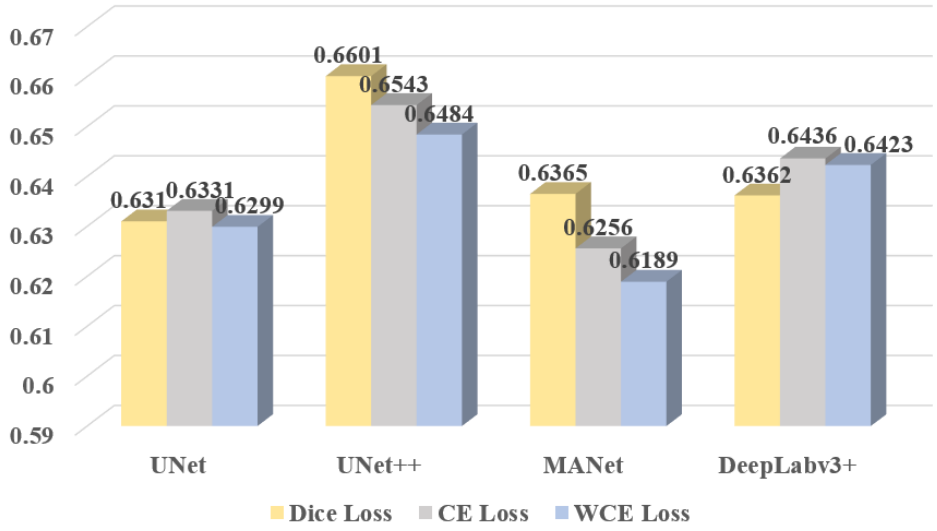

***Figure 2**. IoU performance on the rural dataset using the Adam optimizer*

Dice Loss outperforms CE and WCE, particularly for UNet++ and UNet, likely due to its sensitivity to class imbalance and ability to focus on region-level overlap. This is important in rural areas where classes like barren or roads are often underrepresented and exhibit ambiguous visual features.

Overall, model architectures that allow for deep, context-aware feature fusion and fine-scale localization perform better in rural domains, where semantic boundaries are not as crisp, but broader regional patterns carry strong class cues.

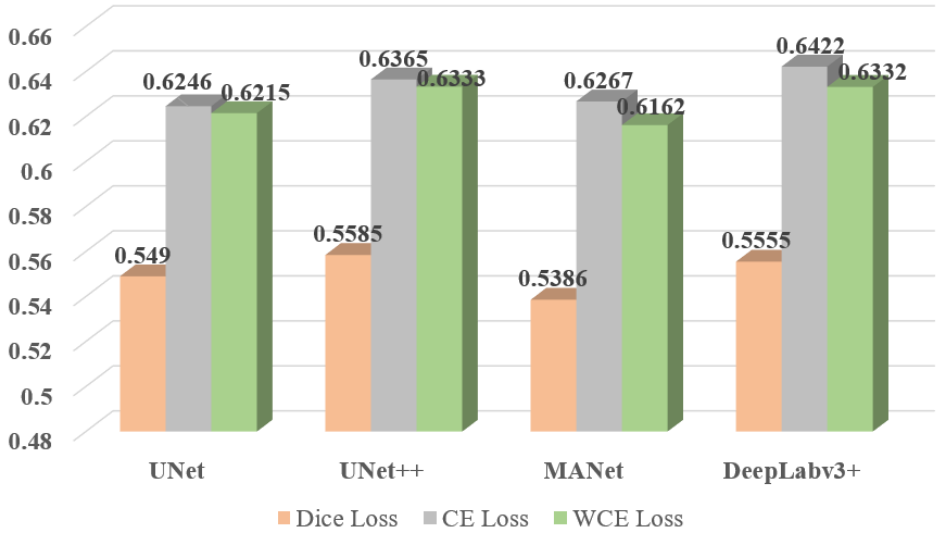

***Figure 3**. IoU performance on the rural dataset using the SGD optimizer*

In rural area segmentation tasks, the performance of models trained with the SGD optimizer shows (Figure 3) notable variation depending on both architecture and loss function. Among the evaluated models, DeepLabv3+ achieved the highest IoU score of 0.6422 when trained with the standard CE loss, indicating its strength in handling broad and visually ambiguous rural land classes. UNet++ closely follows, particularly excelling under WCE and Dice Loss, which highlights its robustness in managing class imbalance and capturing multi-scale

contextual features. MANet, on the other hand, consistently underperforms across all loss types, suggesting that its attention mechanisms may be less effective in rural settings where class boundaries are less distinct and visual patterns are more heterogeneous.

While SGD provides reasonable performance, especially with CE-based training, it appears less capable of handling the nuanced spatial transitions and imbalanced class distributions characteristic of rural environments compared to adaptive optimizers like Adam. Rural landscapes pose unique challenges for segmentation due to their class ambiguity, spatial variability, and class imbalance. Our experiments reveal that optimizer choice has a significant impact on model performance in such settings. Specifically, Adam optimizer consistently yields higher IoU scores across all models and loss functions, with UNet++ showing the largest gain (10% improvement in Dice Loss). This suggests that adaptive optimization strategies are better suited to handling the complex, diffuse class boundaries and imbalanced distributions typical of rural regions. In contrast, SGD results in lower Dice-based IoU scores, possibly due to its fixed learning rate and sensitivity to sparse gradients, which are common in rural image segmentation tasks.

In all experiments; for Adam optimizer, we use 0.003 as learning rate and 0.0001 as weight decay while For SGD optimizer, we use 0.05 as learning rate and 0.9 as momentum.

When the aforementioned figures are evaluated, it is clear that the highest performance of 0.66 is obtained in the UNet++, Dice loss, Adam optimizer combination. The GT and predict masks of this combination on the test data are presented in Figure 4 along with the RGB images.

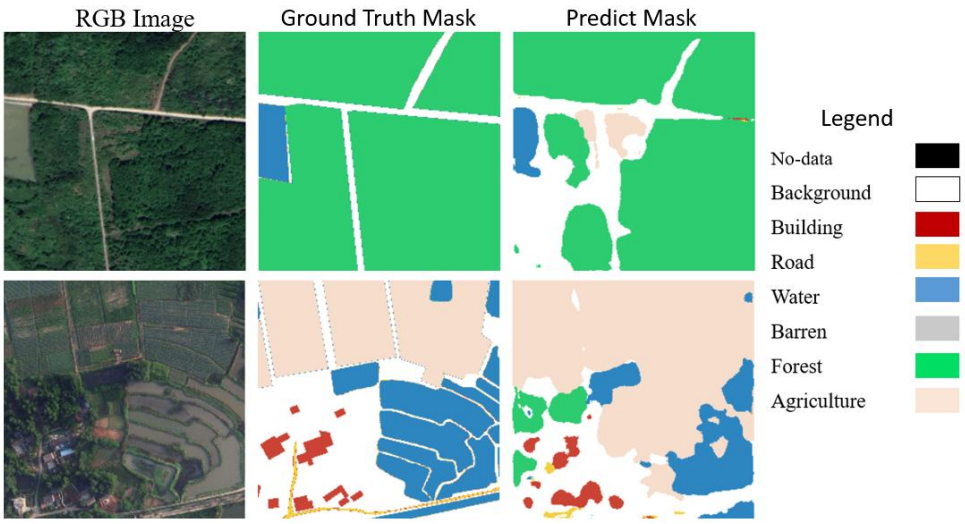

***Figure 4**. RGB images, GT, and predicted segmentations from the rural dataset*

Figure 5 displays the confidence interval (CI) plots for both the IoU and accuracy metrics on the rural test set. These plots offer a statistical perspective on the model's performance, illustrating the range of values that can be expected for each metric with a specified level of confidence. To enhance visual interpretability, the graphs are summarized. Each point in the plot corresponds to a batch, with each batch representing 16 test samples.

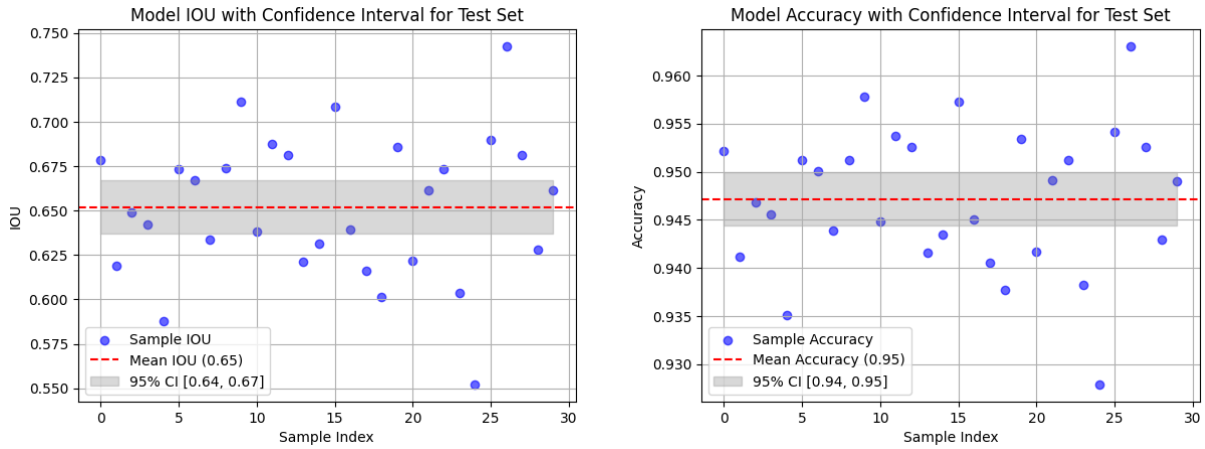

***Figure 5**. CI analysis using different metrics on the rural test set*

The complexity matrix for analyzing the pixel-based correct and incorrect decisions of the proposed model is presented in Figure 6.

As shown in Figure 7, the class-based accuracy values indicate that the segmentation model performs at a relatively high level across most categories. Notably, the "no-data" class exhibits exceptional performance with an accuracy of 99.67%, highlighting the model's proficiency in handling missing or unlabeled areas. Furthermore, the "forest" class, with an accuracy of 83.37%, stands out as one of the best-performing classes, suggesting effective segmentation of forested regions. The "water" and "agriculture" classes also demonstrate solid results, with accuracy rates of 83.08% and 82.25%, respectively, reflecting the model's capability to accurately identify these land cover types. In contrast, the "barren" class shows the weakest performance, with an accuracy of only 60.66%, indicating that further improvements are needed in this area. The "building" (80.20%) and "road" (69.35%) classes exhibit moderate accuracy, revealing room for improvement in distinguishing these features from other land types. Overall, while the model achieves high accuracy for most classes, there is clear potential for enhancement, particularly for more challenging categories such as “barren”.

Grad-CAM (Gradient-weighted Class Activation Mapping) visualizes which areas the model is focusing on. The areas marked with vibrant colors represent the regions that have the most influence on the model's decisions. This map helps us understand which features the model is concentrating on while making its predictions.

As shown in Figure 8 the rural test image, the Grad-CAM heatmaps show distinct and well-localized attention patterns for high-performing classes such as forest, water, and agriculture. The model's attention in these cases aligns closely with semantically meaningful regions for instance, forest areas activate dense regions with tree coverage, and agriculture areas correspond to regular, patch-like field textures.

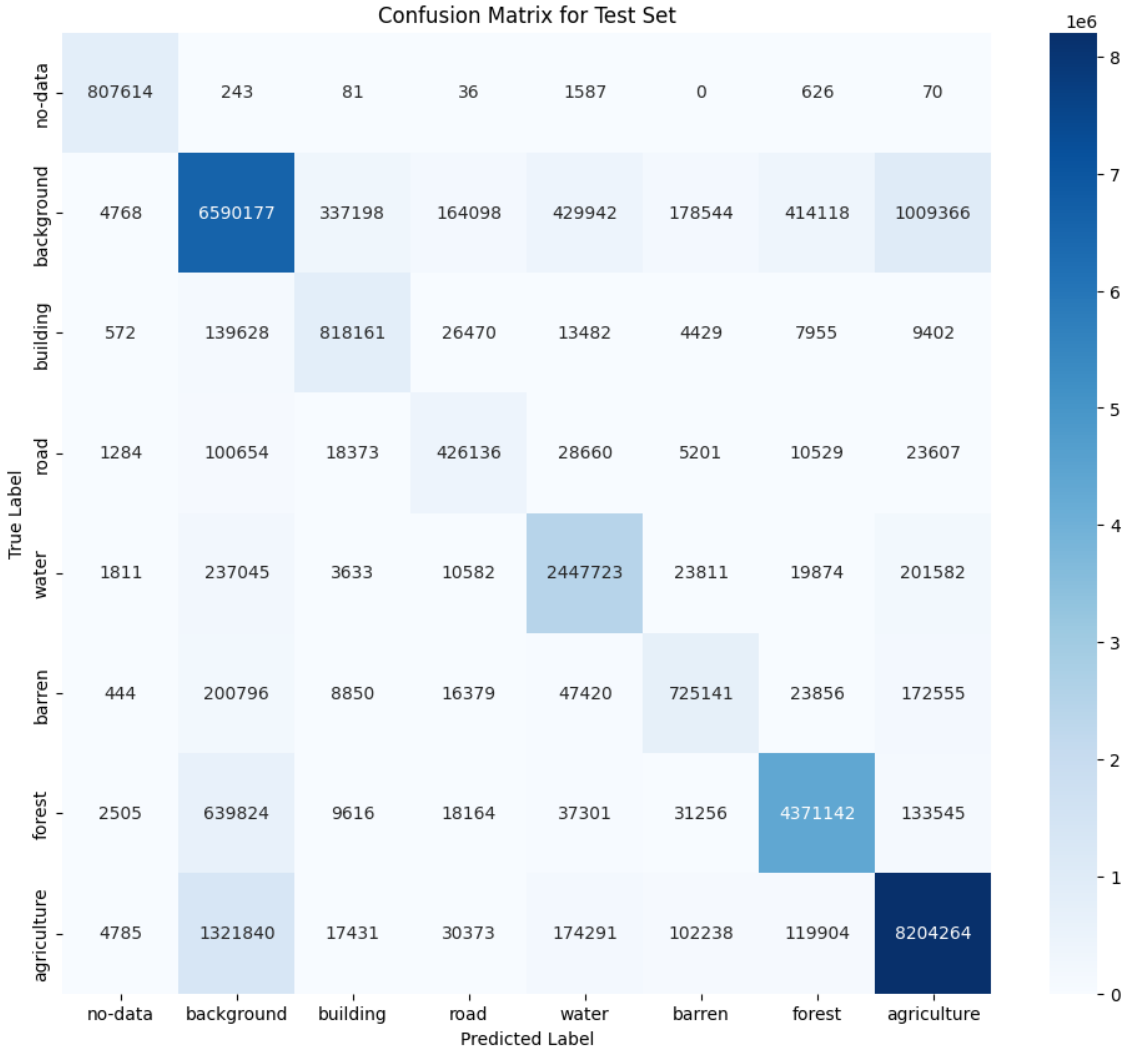

Figure 6. *Confusion matrix of the rural test set*

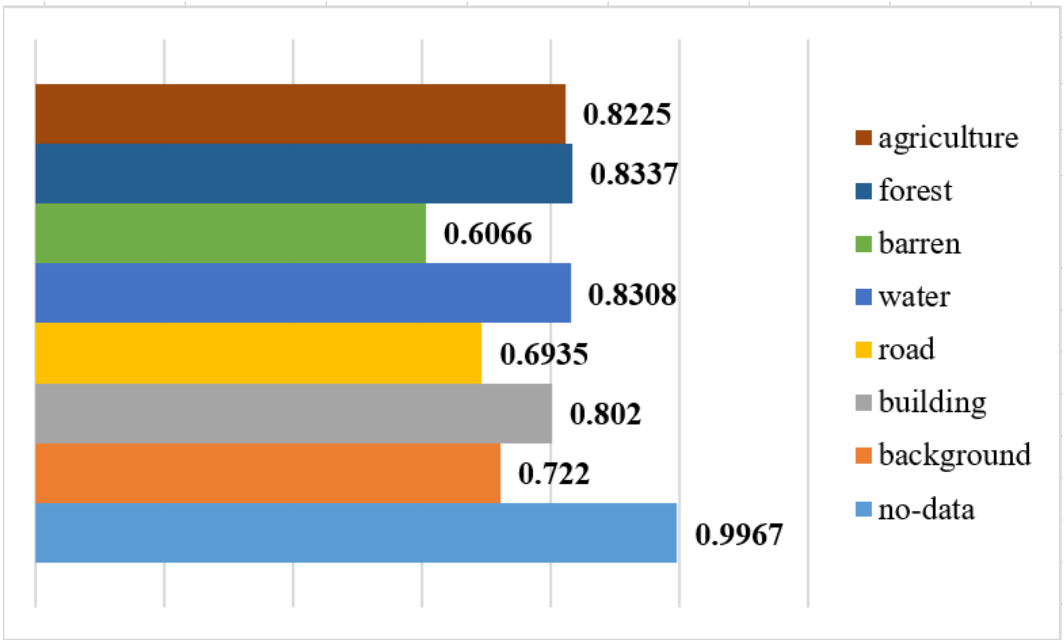

Figure 7. *Class-wise Accuracy performance of the rural segmentation model*

Classes with lower accuracy, such as barren and road, exhibit more scattered or ambiguous Grad-CAM activations. For example, road segments in rural areas may be less defined (e.g., dirt paths), leading to broader or less confident attention distributions in the visualizations. Similarly, barren regions often overlap visually with agricultural or background areas, making it harder for the model to consistently focus on them.

3.2.2. Urban Dataset

After rural dataset models are trained, it is obvious some models give worse than other results. Therefore, we omit some models in order to decrease time spent for training and GPU resources. According to result, we prefer only use Adam optimizer instead of Adam and SGD because most models having SGD as optimizer has imbalance class prediction or poor focal loss result (increasing gradually instead of decreasing). Also, SGD has slow converge compared to Adam. Also, we prefer not to use UNet because we have already used UNet++, and UNet++ is more complex and advanced version of UNet. Figure 9 shows the IoU values of three models and three loss approaches on the test data according to the Adam optimizer.

According to Figure 9, in the segmentation of urban regions, the models trained with the Adam optimizer display moderate performance, with IoU scores ranging between 0.5172 and 0.5717. Among the tested architectures, UNet++ achieves the highest IoU across all three loss functions, with its best performance observed under Dice Loss (0.5717). This result reflects the architecture's ability to capture multi-scale spatial features and preserve fine-grained details essential characteristics in urban settings where class boundaries such as roads, buildings, and water bodies are relatively well-defined. DeepLabv3+ and MANet follow closely, though their scores remain slightly lower, especially when trained with WCE loss, which may overcompensate for class imbalance in well-structured urban environments. The marginal differences in performance also suggest that while urban areas provide clearer visual cues for segmentation, accurately delineating narrow or adjacent classes (e.g., roads vs. buildings) remains challenging. Adam’s adaptive learning capabilities contribute to stable convergence and generally consistent results across architectures, though the overall IoU levels indicate room for further enhancement, particularly through architectural improvements or targeted post-processing. As a result, the combination of UNet++, Dice loss, and Adam optimizer was the most successful model with 0.5717 IoU in the urban dataset as in the rural dataset.

The GT and predict masks of this combination on the test data are presented in Figure 10 along with the RGB images.

Figure 11 presents the CI plots for both the IoU and accuracy metrics on the urban test set. As in Figure 6, each point in the graph represents a batch, with each batch corresponding to 16 test samples.

The complexity matrix for analyzing the pixel-based correct and incorrect decisions of the proposed model is presented in Figure 12.

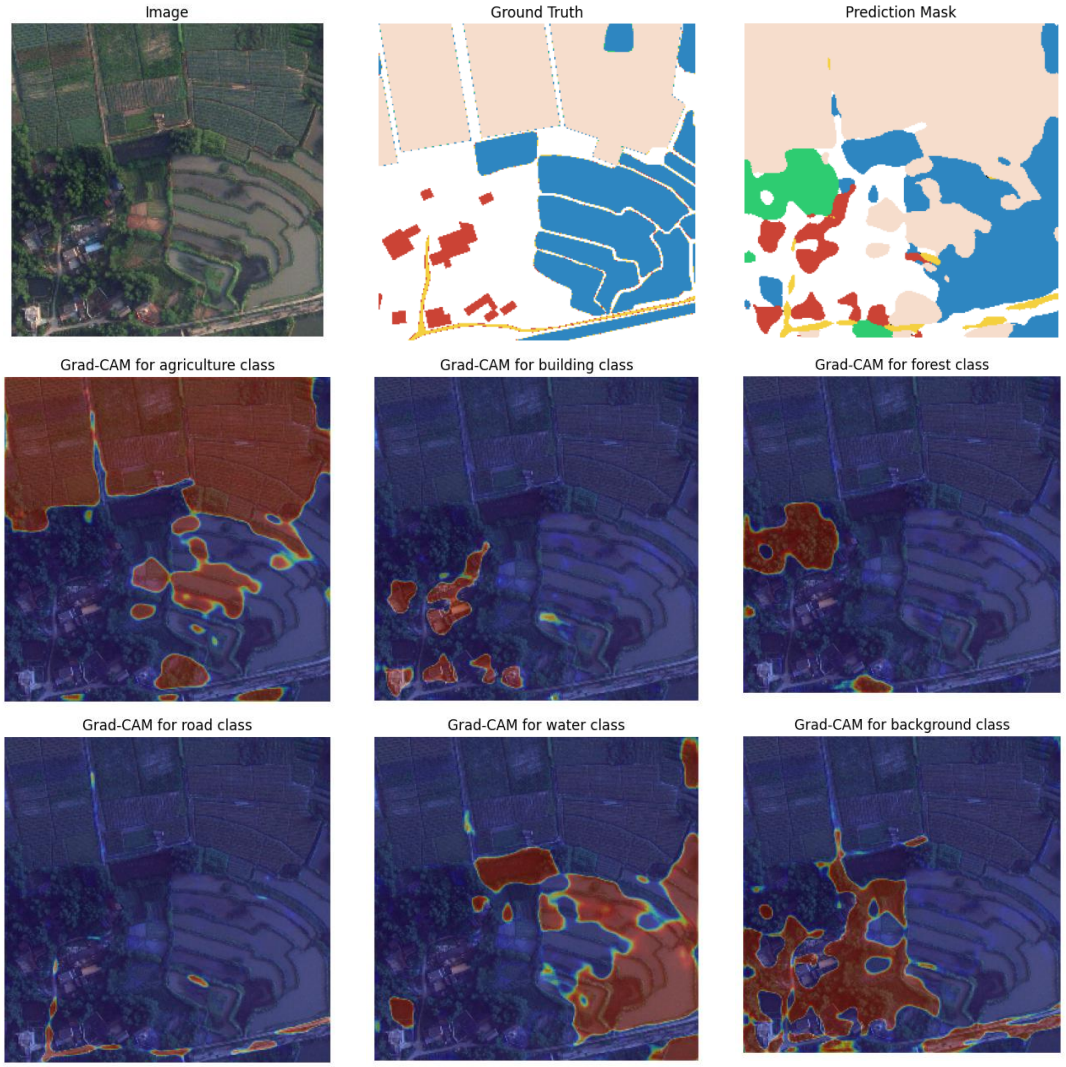

Figure 8. *Grad-CAM visualization for a rural test image*

As presented in Figure13, the per-class accuracy results for the urban region reveal considerable variation in model performance across land cover types. The model achieves relatively high accuracy for water (84.56%), road (76.83%), building (75.21%), and forest (75.13%), suggesting that these classes possess distinct spectral or structural characteristics that the model can effectively learn and distinguish. The no-data class (99.88%) is classified with near-perfect accuracy, likely due to its clearly separable visual traits from other categories. In contrast, accuracy is notably lower for the background class (64.75%) and dramatically low for agriculture (6.25%). The background class likely suffers from high intra-class variability and semantic overlap with adjacent classes, leading to increased confusion. The extremely low performance for agriculture may indicate several challenges: agricultural regions in urban environments are often fragmented, spectrally similar to other vegetative classes (e.g., forest or gardens), and underrepresented in the training dataset. This underperformance points to the difficulty of learning robust representations for agriculture in heterogeneous urban landscapes.

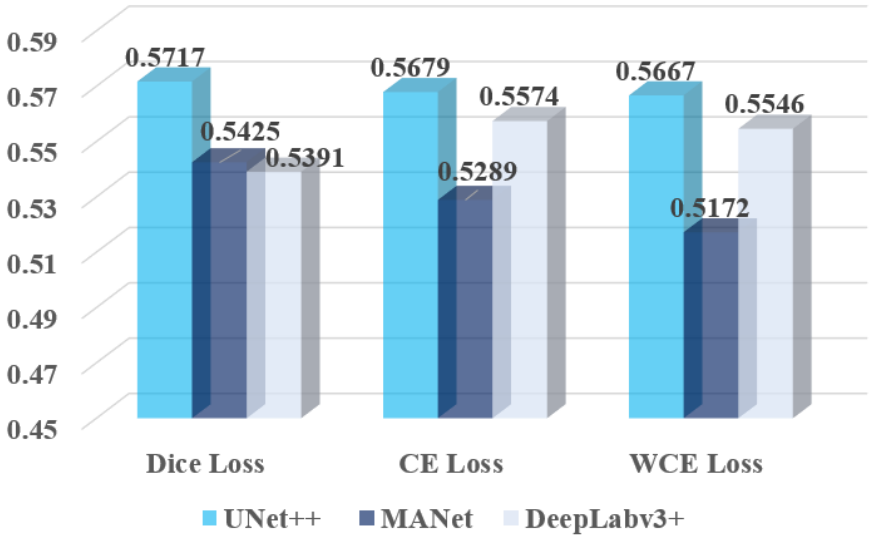

Figure 9. *IoU performance on the urban dataset using the Adam optimizer*

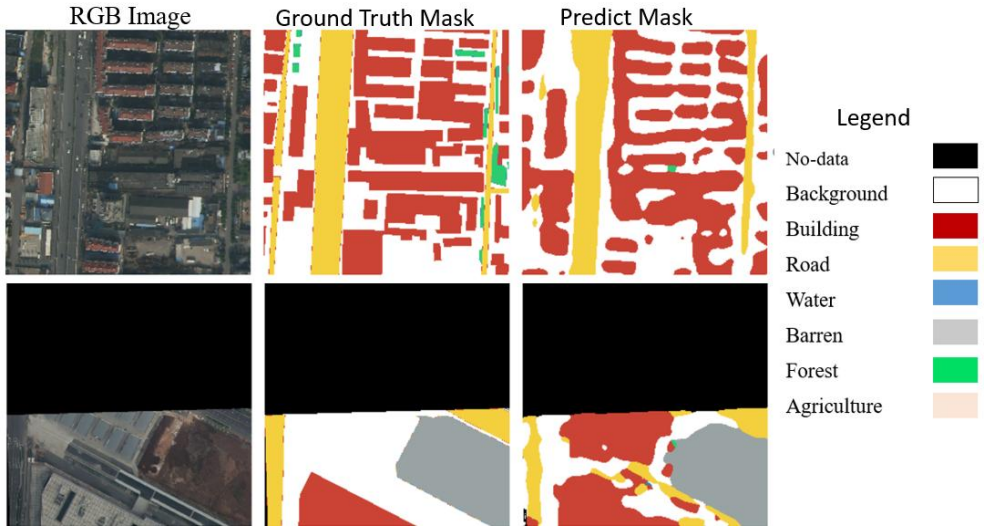

Figure 10. *RGB images, GT, and predicted segmentations from the urban dataset*

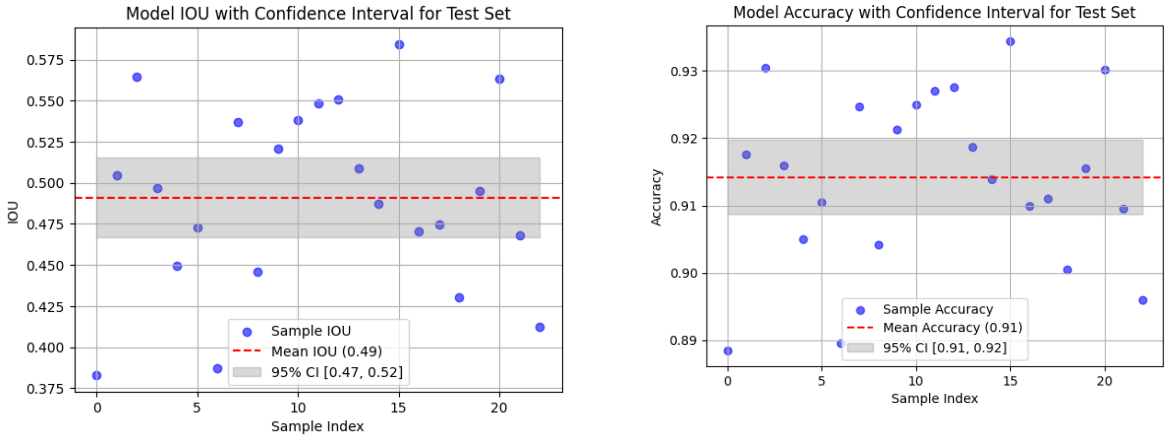

Figure 11. *CI analysis using different metrics on the urban test set*

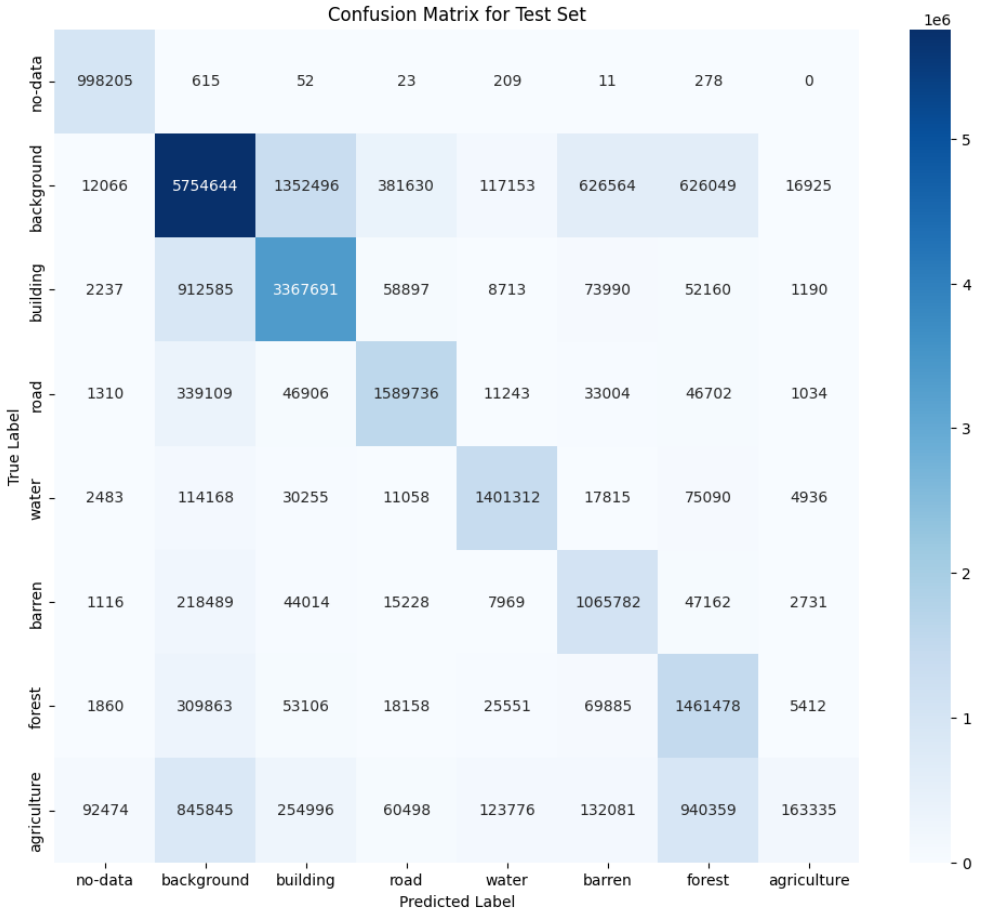

Figure 12. *Confusion matrix of the urban test set*

Overall, the model demonstrates strong classification performance for well-defined, structurally distinct classes, while accuracy declines sharply for classes with ambiguous boundaries or sparse representation. These findings highlight the importance of addressing data imbalance and intra-class variability when mapping high-resolution urban environments.

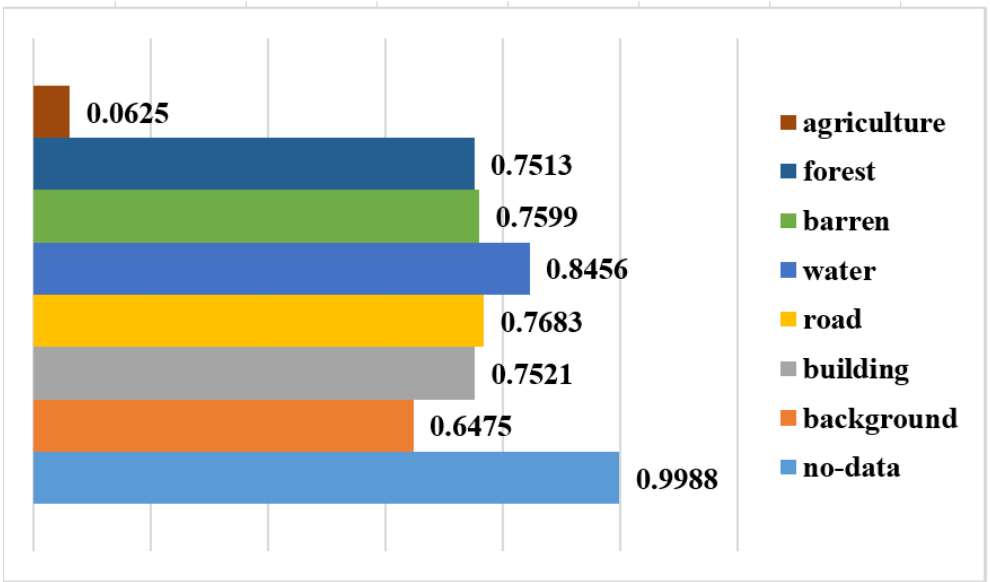

Figure 13. *Class-wise Accuracy performance of the urban segmentation model*

In urban areas, the model's attention was strongly focused on man-made structures such as roads and buildings, indicating that it successfully learns discriminative features relevant to urban classification. Generally, the Grad-CAM outputs for the urban test images show stronger, more concentrated activations for building, road, and water classes. Buildings and roads tend to have clear geometric edges and consistent textures in urban settings, which the model can more easily detect. This is visible in the Grad-CAM maps, where high-attention areas align closely with structured city blocks and well-defined roads.

Interestingly, despite the relatively high accuracy for barren and water classes, agriculture shows a drastically low performance. The corresponding Grad-CAM output reveals that the model does not effectively focus on agricultural zones, likely due to their visual similarity with barren or undeveloped land in urban fringe areas. This lack of focused attention is a key reason behind the low classification accuracy and supports the need for further domain-specific refinement. Figure 14 illustrates the activation map generated using Grad-CAM for a urban image.

4. DISCUSSION

Timely and accurate information on urban and rural land covers is critical for authorities. Urban and rural area analysis plays an important role in land cover change, population forecasting, environmental management, disaster management. (Guo & Du, 2017) HSR land cover data contribute to our interpretation of the geographical and ecological environment through remote sensing technology. These data often provide detailed information on the use of a particular area. Datasets in the literature are used to classify various types of land cover such as forests, agricultural areas, water bodies, settlements, etc.

In this study, we utilize the publicly available LoveDA HSR dataset, which presents challenges such as multi-scale objects, complex backgrounds, and inconsistent class distributions. The study focuses on comparing models and parameters across different geographical environments, specifically urban and rural areas. A total of 5,987 high-resolution (0.3 m) remote sensing images from Nanjing, Changzhou, and Wuhan cities are used for semantic segmentation tasks. Various models, loss functions, and optimization techniques are applied to perform land cover segmentation in both rural and urban settings. The study investigates the impact of different model configurations and parameter combinations on segmentation performance. Additionally, the influence of the geographic differences between rural and urban areas on segmentation success is explored, along with how the selected parameters reflect these differences.

Significant differences were observed between the segmentation results obtained from rural and urban areas using the models employed in this study. It was found that the type of area (rural or urban) influenced the performance of the DL models; however, the same model (UNet++) achieved higher performance in rural areas. This highlights the importance of optimizing the same model for different geographical contexts. The effect of optimization algorithms (Adam and SGD) on segmentation performance demonstrated notable differences in both rural and urban areas. The Adam optimizer exhibited faster and more stable convergence

in rural areas, which proved advantageous for optimizing the more complex and irregular land cover structures typically found in rural environments. The choice of loss function (Dice loss, CE loss, and WCE loss) also had a significant impact on segmentation success. Specifically, it was observed that the Dice loss function, when combined with the Adam optimizer, outperformed the SGD optimizer, particularly for smaller area segments. As a result, experiments in urban areas were conducted exclusively using the Adam optimizer. For both rural and urban areas, the combination of UNet++, the Adam optimizer, and Dice loss yielded the highest performance.

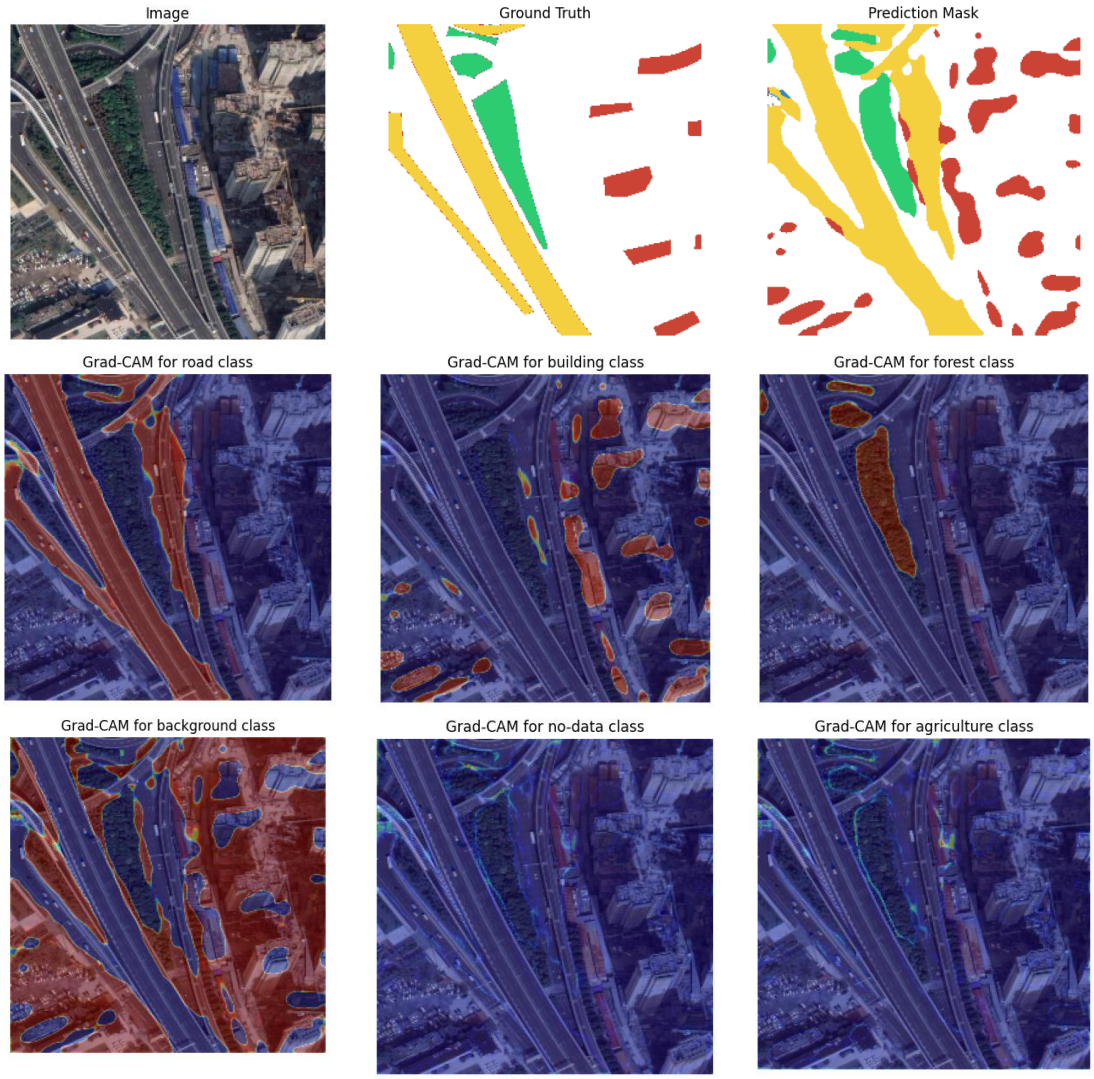

Figure 14. *Grad-CAM visualization for an urban test image*

In this study, ResNet50 was selected as the backbone architecture due to its strong balance between depth and computational efficiency, making it a widely adopted standard in various remote sensing and semantic segmentation tasks. Its residual connections enhance gradient flow and facilitate learning deeper representations, which are essential for capturing hierarchical features in complex scenes. Furthermore, all images were resized to 256×256 pixels prior to training. This resizing was a practical necessity to ensure feasible training under limited GPU resources and to maintain a consistent input size across batches. While

this transformation may introduce some degree of spatial information loss—particularly affecting small or elongated objects—visual inspection and empirical performance indicate that key semantic structures remain sufficiently preserved. Nevertheless, this trade-off is acknowledged as a limitation, and future work may explore adaptive tiling or patch-wise strategies to retain higher spatial fidelity without compromising computational feasibility.

To compare the performance of the LoveDA dataset with the existing literature, a weighted average of the combination (UNet++, Adam optimizer, and Dice loss) was calculated based on the number of test samples for both rural and urban groups. The final performance, in terms of mean Intersection over Union (IoU), was 62.14%. Table 2 presents a comparison of the proposed method with UNet-based techniques from the literature.

Table 2. *Comparison of the proposed method with existing methods in the literature*

Literature	Model	mean IoU (%)
(J. Wang et al., 2021)	UNet	47.84
(L. Wang et al., 2022)	UNetFormer	52.40
(Dimitrovski et al., 2024)	UNet Ensemble	57.36
The proposed method	**UNet++**	**62.14**

Wang et al. (J. Wang et al., 2021) obtained 47.84% mean IoU with UNet when introducing the LoveDA dataset. Wang et al. (L. Wang et al., 2022) achieved a mean IoU score of 52.40% with UNetFormer, a version of U-Net combined with transformer attention mechanisms. Dimitri et al. (Dimitrovski et al., 2024) utilized a U-Net ensemble model with three different backbones (Multi-Axis Vision Transformer, ConvFormer, and EfficientNet) to achieve a mean IoU of 57.36%. In this study, the combination of UNet++, which is a more advanced version of UNet with a hopping link structure, with Dice score and Adam optimizer proposed a mean IoU of 62.14%.

The comparative analysis between urban and rural regions highlights the sensitivity of classification performance to contextual landscape characteristics, particularly in the context of high spatial resolution (HSR) imagery. Rural areas exhibit notably higher accuracy in classes such as agriculture (82.25%), forest (83.33%), and background (72.20%), which can be attributed to the relative spectral homogeneity and spatial consistency of these land cover types in non-urban settings. In contrast, agriculture in urban regions is classified with extremely low accuracy (6.25%), underscoring one of the key challenges in HSR mapping the difficulty of detecting fragmented, spatially sparse, and spectrally ambiguous land cover patches. In urban environments, agricultural areas are often interspersed with other vegetated classes (e.g., parks, roadside greenery), and their small patch size further complicates the model's ability to learn distinctive representations. Likewise, building (80.20%) and water (83.08%) classes achieve better performance in rural settings compared to their urban

counterparts (75.21% and 84.56%, respectively). On the other hand, roads are more accurately classified in urban regions (76.83%) than in rural ones (69.35%), likely due to their more regular, linear structure and higher frequency of representation in urban high-resolution imagery. Interestingly, the barren class performs substantially better in urban areas (75.99%) than in rural ones (60.66%), which may reflect the more visually coherent and structured nature of urban barren areas such as construction sites or vacant lots, as opposed to the more heterogeneous barren landscapes in rural zones.

These findings emphasize the specific challenges of HSR mapping, especially in urban environments. Class confusion due to spectral overlap, high intra-class variability, fragmented spatial patterns, and limited class representation all contribute to reduced model performance for certain land cover types. Therefore, it is crucial to consider these HSR-specific limitations when interpreting classification results, and future work may benefit from incorporating class-aware sampling strategies, multi-modal data sources, or architecture-level enhancements to better address these issues.

The limitations of this study include the following: GT labels for the test data are not available in the dataset, necessitating the random generation of the test partition from the available data. Additionally, some of the images in the dataset exhibit spatial correlation, and the scenario was executed hierarchically due to the large number of experiments across all possible combinations.

5. CONCLUSION

DL approaches for land cover segmentation with HSR data yield more accurate, precise, and efficient results compared to manual methods such as visual interpretation or fieldwork. HSR data, with its high resolution and rich visual details, serves as a crucial data source for accurately distinguishing land cover classes. This study demonstrates that in land cover segmentation using HSR imagery, factors such as the differences between rural and urban patterns, as well as variations in model architectures, loss functions, and optimization techniques, significantly impact segmentation performance. It was observed that the geographic distinctions between rural and urban areas notably influence the optimization of model parameters and the overall success of the segmentation task. Future research should focus on further refining model configurations, incorporating larger datasets, and evaluating the developed models on diverse datasets to enhance generalizability.

AUTHOR CONTRIBUTIONS

Conceptualization: H.A., N.K.; Methodology: H.A., N.K.; Fieldwork and Data Collection: N.K.; Software and Implementation: H.A.; Title and Framing of the Study: N.K.; Validation and Experimentation: H.A.; Investigation and Literature Review: N.K.; Data Curation: H.A.; Writing – Original Draft Preparation: N.K.; Writing – Review and Editing: H.A.; Visualization and Figure Preparation: H.A., N.K..;Supervision and Project Administration: N.K. All authors have read and legally accepted the final version of the article published in the journal.

ACKNOWLEDGEMENTS

This study was conducted in the TÜBİTAK-BİLGEM We would like to express our profound gratitude to TÜBİTAK-BİLGEM.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Abunnasr, Y., & Mhawej, M. (2023). Fully automated land surface temperature downscaling based on RGB very high spatial resolution images. City and Environment Interactions, 19, 100110.

Aksoy, B., Çakmak, B., & Kumbasar, N. (2023). Muğla Wildfires Burn Severity and Vegetation Difference Analysis with Remote Sensing Techniques. 2023 31st Signal Processing and Communications Applications Conference (SIU), 1–4. https://doi.org/10.1109/SIU59756.2023.10223948

Baek, W.-K., Lee, M.-J., & Jung, H.-S. (2024). Land Cover Classification From RGB and NIR Satellite Images Using Modified U-Net Model. IEEE Access, 12, 69445–69455. https://doi.org/10.1109/ACCESS.2024.3401416

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers, 177–186.

Chan, L., Hosseini, M. S., & Plataniotis, K. N. (2021). A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. International Journal of Computer Vision, 129(2), 361–384.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. ArXiv, abs/1706.05587. https://api.semanticscholar.org/CorpusID:22655199

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European Conference on Computer Vision (ECCV), 801–818.

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

Dimitrovski, I., Spasev, V., Loshkovska, S., & Kitanovski, I. (2024). U-Net Ensemble for Enhanced Semantic Segmentation in Remote Sensing Imagery. Remote Sensing, 16(12), 2077.

Fan, T., Wang, G., Li, Y., & Wang, H. (2020). MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation. IEEE Access, 8, 179656–179665. https://doi.org/10.1109/ACCESS.2020.3025372

Goodfellow, I. (2016). Deep learning. MIT press.

Guo, Z., & Du, S. (2017). Mining parameter information for building extraction and change detection with very high-resolution imagery and GIS data. GIScience & Remote Sensing, 54(1), 38–63.

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

Hossain, M. D., & Chen, D. (2022). A hybrid image segmentation method for building extraction from high-resolution RGB images. ISPRS Journal of Photogrammetry and Remote Sensing, 192, 299–314.

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

Kampffmeyer, M., Salberg, A.-B., & Jenssen, R. (2016). Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 1–9.

Kemker, R., Salvaggio, C., & Kanan, C. (2018). Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS Journal of Photogrammetry and Remote Sensing, 145, 60–77.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. CoRR, abs/1412.6980. https://api.semanticscholar.org/CorpusID:6628106

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440.

Myint, S. W., Gober, P., Brazel, A., Grossman-Clarke, S., & Weng, Q. (2011). Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sensing of Environment, 115(5), 1145–1161.

Neupane, B., Horanont, T., & Aryal, J. (2021). Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. Remote Sensing, 13(4), 808.

Pan, Z., Xu, J., Guo, Y., Hu, Y., & Wang, G. (2020). Deep learning segmentation and classification for urban village using a worldview satellite image based on U-Net. Remote Sensing, 12(10), 1574.

Peng, C., Li, Y., Jiao, L., Chen, Y., & Shang, R. (2019). Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(8), 2612–2626.

Pihur, V., Datta, S., & Datta, S. (2007). Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. Bioinformatics, 23(13), 1607–1615.

Ronneberger, O., Fischer, P., & Brox, T. (2015). 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, 234–241.

Sudre, C. H., Li, W., Vercauteren, T. K. M., Ourselin, S., & Cardoso, M. J. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. Deep Learning in Medical

Image Analysis and Multimodal Learning for Clinical Decision Support : Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017 Quebec City, QC,..., 2017, 240–248. https://api.semanticscholar.org/CorpusID:21957663

Sun, W., & Wang, R. (2018). Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. IEEE Geoscience and Remote Sensing Letters, 15(3), 474–478.

Tao, A., Sapra, K., & Catanzaro, B. (2020). Hierarchical multi-scale attention for semantic segmentation. ArXiv Preprint ArXiv:2005.10821.

Wang, H., Dalton, L., Fan, M., Guo, R., McClure, J., Crandall, D., & Chen, C. (2022). Deep-learning-based workflow for boundary and small target segmentation in digital rock images using UNet++ and IK-EBM. Journal of Petroleum Science and Engineering, 215, 110596.

Wang, J., Zheng, Z., Ma, A., Lu, X., & Zhong, Y. (2021). LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In J. Vanschoren & S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (Vol. 1). https://datasets-benchmarks

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

Xia, H., Wu, J., Yao, J., Zhu, H., Gong, A., Yang, J., Hu, L., & Mo, F. (2023). A Deep Learning Application for Building Damage Assessment Using Ultra-High-Resolution Remote Sensing Imagery in Turkey Earthquake. International Journal of Disaster Risk Science, 14(6), 947–962.

Yi-de, M., Qing, L., & Zhi-Bai, Q. (2004). Automated image segmentation using improved PCNN model based on cross-entropy. Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004., 743–746.

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.

Yu, B., Yang, L., & Chen, F. (2018). Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(9), 3252–3261.

Zhang, H., Liu, Y., Li, X., Feng, R., Gong, Y., Jiang, Y., Guan, X., & Li, S. (2023). Combing remote sensing information entropy and machine learning for ecological environment assessment of Hefei-Nanjing-Hangzhou region, China. Journal of Environmental Management, 325, 116533.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. S. Tavares, A. Bradley, J. P. Papa, V. Belagiannis,

J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, & A. Madabhushi (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (pp. 3–11). Springer International Publishing.

Zou, S., Fan, X., Wang, L., & Cui, Y. (2024). High-speed rail new towns and their impacts on urban sustainable development: a spatial analysis based on satellite remote sensing data. Humanities and Social Sciences Communications, 11(1), 1–13.