Find The Way! Computer Vision with Deep Learning

Rhea Gupta
Faculty of Science
University of Western Ontario
London, Canada
rgupt5@uwo.ca

Osama Yousef
Faculty of Science
University of Western Ontario
London, Canada
oyousef3@uwo.ca

Bryan Lee
Faculty of Science
University of Western Ontario
London, Canada
blee375@uwo.ca

Jennifer Yoon
Faculty of Science
University of Western Ontario
London, Canada
jyoon94@uwo.ca

Jacob Smith
Faculty of Science
University of Western Ontario
London, Canada
jsmit822@uwo.ca

Abstract*—Road segmentation in satellite imagery is a critical task for urban planning and traffic management. The task becomes challenging due to the varying conditions and complex patterns of roads. In this study, we address the road segmentation problem using deep learning techniques, specifically using existing Segmentation Models Pytorch (SMP) U-Net, DeepLabV3 and a custom-built U-Net applied to the Massachusetts Roads Dataset. U-Net, known for its efficiency in biomedical image segmentation, and DeepLabV3, recognized for its effectiveness in semantic segmentation and efference use of computational resources, are employed to delineate roads in high-resolution aerial images. Our approach includes comprehensive data augmentation and specialized loss functions. We provide a detailed comparative analysis of the three models, assessing their performance in road detection accuracy and computational efficiency. The results demonstrate that all models achieve high accuracy, with U-Net showing effectiveness in finer details, DeepLabV3 excelling in computational speed but failing in accurately predicting results and the custom U-Net excelling in both. This study contributes to the advancement of road segmentation techniques and offers insights into the application of deep learning in remote sensing.

Index Terms—Road Segmentation, Satellite Imagery, Deep Learning, U-Net, DeepLabV3, Massachusetts Roads Dataset, Image Processing, Urban Planning, Traffic Management.

I. Introduction

Road extraction from remote sensing images has emerged as a crucial topic in the modern era, significantly impacting traffic management, urban planning, map updating, and even emergency response during natural disasters. The challenge of high-precision road extraction is amplified by the complex interplay of road backgrounds, spectral similarities, and the varying standards of road construction across different geographical regions. Traditional methods for road extraction in remote sensing imagery, such as feature-based[8] and classification-based approaches, have their limitations. Feature-based methods, relying on characteristics like shape, texture, and geometry, are effective for simple and regular road structures but falter when faced with complex road networks. Classification-based methods, which include maximum likelihood, support vector machines[8], and Markov random fields, depend heavily on the accuracy of classification rules and often struggle with the spectral similarity of roads to other urban features like buildings and parking lots.

Our methodology includes comprehensive data augmentation to address various angles and road types. A critical aspect of our work is the detailed comparative analysis of these models, assessing their performance in terms of road detection accuracy and computational efficiency. This analysis is crucial in understanding the strengths and limitations of each model in the context of varying conditions and complex road patterns[2]. Through this study, we contribute to the evolving landscape of road segmentation techniques, offering valuable insights into the application of deep learning in remote sensing[2]. Our research is not only significant for urban planning and traffic management but also lays the groundwork for future advancements in autonomous navigation and smart city development[6].

II. Related Work

Road network detection has many purposes and is a topic that is continuously being studied to fulfill the requirements of these purposes in today's day and time. Aerial imagery road detection is being used for toad navigation, urban planning, traffic management, developing safe routes in disasters, and other geographic benefits [17]. It is also being greatly studied when it comes to the development of systems in vehicles for lane detection, migration, and road detection AI.

However, while the concept of road detection is similar to many studies, each study varies in the way they applied their models, why they used the model and the challenges that were found. For example, there was a study that recognized some common issues that were found in road detection previously, such as over-segmentation, distortion and noise sensitivity, and so they used a model that could help road extraction more efficiently, especially in the case of emergency evacuations and enhance the response to disasters [17]. It combined various models/techniques such as residual learning, U-Net, deep-ResUnet architecture and the B-snake algorithm [17]. Another study found that while deep learning models have been very beneficial to the study of road detection, it can be very difficult to train the data, and

it can sometimes fall into over-fitting [1]. This study decided to tackle this issue using multiple lightweight models combined, such as the Adaboost-like End-To-End Multiple Lightweight U-Nets model [1]. There are many more studies like these that are constantly improving models to get the most accurate representation possible for each problem and solution.

III. Data

Our goal of identifying roads in images requires a huge dataset to capture different road types, road designs and road structures; for this, we chose to use The Massachusetts Roads Dataset, a data set that consists of 1171 aerial images of the entire state of Massachusetts. Each image is 1500×1500 pixels in size, covering an area of 2.25 square kilometers or 1 meter per pixel [7]. The data was randomly split by the author into a training set of 1108 images, a validation set of 14 images, and a test set of 49 images. A sample of this dataset is shown in Figure 1. The dataset covers a wide variety of urban, suburban, and rural regions and covers an area of over 2600 square kilometers. The label maps were generated by rasterizing road centerlines obtained from the OpenStreetMap project [7]. A line thickness of seven pixels and no smoothing was used in generating the labels.

Figure 1: Sample of the Massachusetts Roads Dataset

IV. Data Preprocessing

Due to the constraints imposed by current computational capacities, processing entire high-resolution images directly through a Convolutional Neural Network (CNN) [1] for training is not feasible. To circumvent this limitation, the images from the datasets are preprocessed into smaller, augmented patches[12]. These patches are then utilized for model training and testing. For the final evaluation, these patches are reassembled to form a comprehensive prediction mask. Additionally, the spatial resolutions were standardized to ensure consistency across the combined buildings dataset. The chosen resolution for the input image crops was set at 256x256 pixels in the SMP U-Net and DeeplabV3 models, and 512x512 in our custom U-Net model.

This decision was influenced by two main factors: firstly, 256x256 pixels and 512x512 are widely accepted standards for semantic segmentation tasks in the field [3], and secondly, it aligns with the computational resources at our disposal. The augmentation techniques employed in this project include horizontal and vertical flips and rotations. A sample of an augmented (vertically flipped) picture is shown in Figure 2. Such augmentations are crucial as they enable the model to learn from varying perspectives, which is essential given the diverse orientations encountered in aerial imagery.

Figure 2: Sample of an augmented picture

V. Models Architectures

This section discusses the three models' architectures used in this investigation. The three significant components in any Neural Network that have the biggest effect on the performance of the model are the loss function, the optimizer, and the layers' architecture itself. Figure 3 shows how those components work together[13].

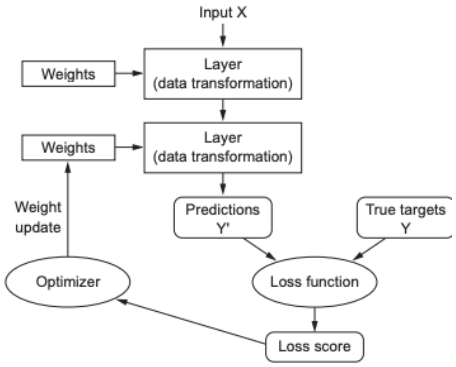

Figure 3: Simplified view of a neural network

A) Loss Function:

To effectively evaluate the performance of different network architectures in semantic segmentation, an appropriate metric is needed to accurately reflect the network's capability in class identification. For this thesis, the primary metric selected for assessment purposes will be the Dice coefficient [13].

Jaccard Index or IOU metric calculates the similarities between the predicted mask or region and the true mask of the image. What differentiates IOU from other evaluation metrics is how it heavily punishes incorrect predictions and rewards correct classifications. The equation used to calculate the Jaccard Index (IOU) is as follows:

$$\mathbf{Jaccard} = \frac{\mathbf{TP}}{\mathbf{TP + FP + FN}} = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A} \cup \mathbf{B}|} = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A}| + |\mathbf{B}| - |\mathbf{A} \cap \mathbf{B}|} \quad (1)$$

In this context, TP represents the true positives, FP denotes the false positives, and FN signifies the false negatives, all of which are evaluated in relation to the predicted and ground truth images. [16] A visual representation of (IOU) is shown in Figure 4.

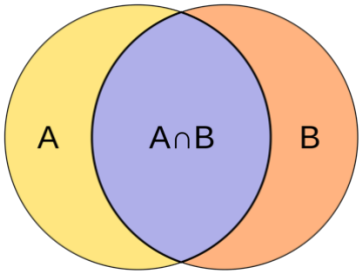

Figure 4: Visualization of IOU

The Dice Coefficient is calculated as the size of the intersection divided by the total size of the two sets. The Dice Coefficient was used as the loss function in all three models.

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (2)$$

B) Optimizer:

Adam optimizer was selected based on previous research for similar models, which showed improvements in the overall output results [5]. Figure 5 shows the use of different optimizers over the MNIST dataset, where the goal is identifying numbers from images, a similar task to what we are trying to achieve.

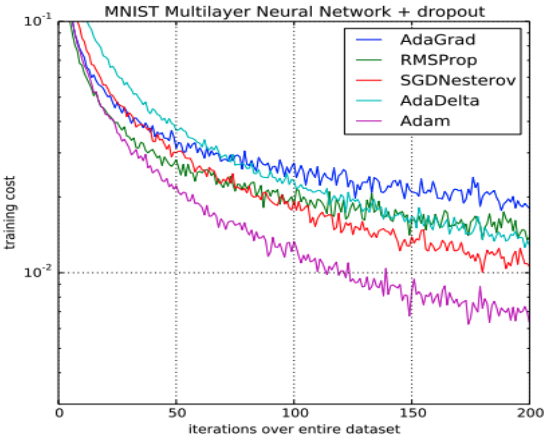

Figure 5: Training of multilayer neural networks on MNIST images

C) Layer Architecture:

a) SMP U-Net && SMP DeepLabv3: for both models, the encoder was resnet50, a 50-depth layer encoder, a total of 99 layers. We tested both models with pre-trained weights on the ImageNet dataset on 21 million parameters and we found no significant difference in performance in both cases. Figure 6 shows the architecture of the SMP U-Net[11].

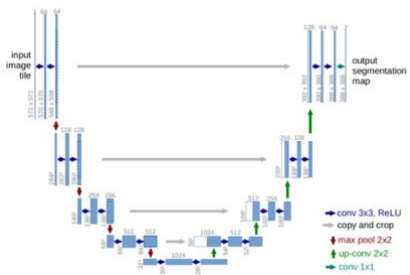

Figure 6: SMP U-Net Model Architecture

Understanding the layers: In U-Net architecture, the model consists of what follows, down sampling (contracting path) is responsible for capturing the context in the image. Each block in this path consists of convolution and max pooling layers followed by a rectified linear unit (ReLU) activating function.

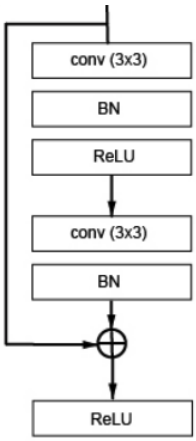

Figure 7: addition of identity mapping before final nonlinear activation in each convolution block

Convolution is used to extract the features from the input image by sliding a kernel over the input image and computing the dot product of the local region, in this context of 2D convolution the operation, can be expressed as:

$$(I * K)(i,j) = \sum_{(m)} \sum_{(n)} I(i+m, j+n) K(m,n) \quad (3)$$

I is the input image, K is the Kernel, (i,j) are the coordinates in the output feature map.

After each convolution, an activation function is applied; in this case, it is ReLU. ReLU introduces non-linearity into the model, allowing it to understand complex patterns better.

$$ReLU(x) = max(0, x) \quad (4)$$

Max pooling layers are introduced to reduce the spatial resolution (height, width) of the input feature map, effectively reducing the computational complexity and the number of parameters in the network [9]. Considering a 2x2 pooling region, the operation can be described as where P is the pooled output, and I is the input feature map:

$$P(i,j) = \max\{I(2i,2j), I(2i,2j+1), I(2i+1,2j), I(2i+1,2j+1)\} \quad (5)$$

The bottleneck path is where the transition from the contracting (down-sampling) path to the expanding (up-sampling) path. This path is responsible for understanding the global context of the input image. The block layers are like the down-sampling path, however, there are no pooling layers here. The up-sampling path is responsible for increasing the spatial resolution of the feature maps [10], enabling the network to focus on precise localization.

b) Custom U-Net:

The custom U-Net has 60 layers, and it was built using TensorFlow instead of PyTorch. The goal of this model was to compare the performance of existing popular models and built-from-scratch models. For this model architecture, the contracting pathway employs a series of 3x3 convolutional operations, each succeeded by a Rectified Linear Unit (ReLU) and a 2x2 max pooling process with a stride of 2 for effective down-sampling. At each stage of down-sampling, the feature count is increased twofold, a process that is subsequently reversed in the expanding pathway. The expansion involves up-sampling coupled with 3x3 convolutions, integrating these outputs with the corresponding feature maps from the contracting pathway. The architecture culminates with a 1x1 convolution.

VI. METHODS

a) Training Protocol

The project's primary objective is to develop a road segmentation model utilizing various architectures for optimized performance and computational efficiency. Each model was trained on the Massachusetts Roads Dataset using Kaggle's Nvidia Tesla P100 GPUs. A consistent learning rate of 0.0001 was maintained across all models to facilitate comparability. Training was conducted over several epochs, with early stopping implemented to prevent overfitting. Batch normalization and dropout techniques were applied to enhance generalization.

b) Time Efficiency Evaluation

To assess the computational efficiency of each model, a time-based performance metric was introduced. This metric evaluates the number of images processed within a 15-minute window, providing a direct measure of each model's speed and efficiency. This evaluation is crucial in practical applications where processing speed is as important as accuracy, especially in real-time urban planning and traffic management scenarios.

c) Accuracy Metrics

The Intersection over Union (IoU) and Dice Coefficient scores were the primary metrics for assessing segmentation accuracy. These metrics provide a comprehensive understanding of each model's ability to precisely delineate roads in varied environmental conditions[13]. The IoU and Dice scores were calculated for each model on the test set, allowing for a direct comparison of their segmentation capabilities.

d) Real-world Testing

In addition to the standard dataset, each model was tested on a new set of 15 high-resolution (1500x1500 pixels) images sourced from Google Earth, encompassing diverse urban, suburban, and rural landscapes. This step was crucial to evaluate the models' generalizability and performance in real-world scenarios. The predicted road maps were manually inspected and compared to the actual road layouts to assess the models' practical applicability.

e) Model Tuning and Optimization

During the training process, hyperparameter tuning was conducted to optimize each model's performance. Parameters such as filter sizes, number of layers, and learning rate were adjusted iteratively. The models were also tested with and without pre-trained weights to examine the impact of transfer learning on their performance.

f) Data Augmentation Techniques

To ensure robustness against overfitting and to enhance the models' ability to generalize across different road types and orientations, a variety of data augmentation techniques were employed. These included random rotations, shifts, zooms, and flips. The augmented data provided a more comprehensive representation of real-world variability in road appearances [4].

g) Comparative Analysis

Finally, a detailed comparative analysis was conducted to evaluate each model's strengths and weaknesses. This analysis considered not only the quantitative metrics of accuracy and time efficiency, but also qualitative assessments based on the model's performance on real-world imagery. The analysis aimed to provide insights into the suitability of each model for specific applications in road segmentation tasks.

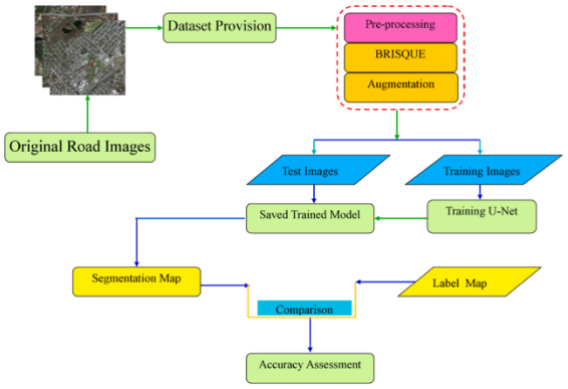

Figure 8: The Methods of the proposed approach

VII. EXPERIMENTAL RESULT

To evaluate all the models effectively, all the models were tested with different parameters and methods adhering to the methods of evaluation; all the models were limited with a time limit of 15-minute window.

Overall, the SMP U-Net performed the best on both accuracy metrics and real-world testing. Deeplabv3 scores ranked second. However, when the model was tested after training, the model predictions showed great inconsistency, which led us to do some investigations. Despite our custom U-Net preforming poorly on IOU and Dice loss metrics,

when it was tested on the Google Earth test set, it performed almost identically to SMP U-Net.

	Unet Architecture	Mean IOU Score	Mean Dice Loss
0	SMP Unet	0.852300	0.047300
1	SMP DeepLabV3	0.741200	0.082100
2	Custom U-Net	0.644300	0.131200

Figure 9: Accuracy Metrics for the 3 Models

A surprising finding we encountered was that the performance is decided from the first epoch of training, and extra batches of training data yield results similar to the first. Our assumption was that these results were influenced using the pre-trained weights of the ImageNet data set. However, after testing both models without pre-trained weights, we achieved the same results. We reckon that the reason behind such performance is the intensive optimizations included in such popular models.

Our Custom U-Net gradually improved with each iteration, showing the typical neural network behavior. However, the performance always stagnated around 0.6 on the IOU metrics, the extra training did not yield better results, showing some limitations in our design. We propose the assumption that our model is not as optimized as SMP models to fully utilize the processing power of the Nvidia TSLA P100 GPU [14].

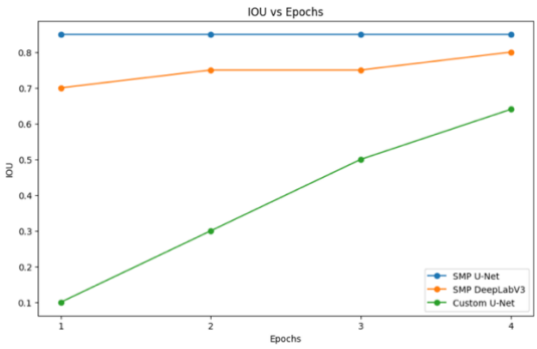

Figure 10: IOU scores vs Epochs

- SMP U-Net:

performed extremely well on both the testing and our Google Earth datasets.

Figure 11: Prediction Result from the SMP U-Net

The predictions made by the model are almost very identical to the ground truth masks, it captures different road types and complex road designs.

On our Google Earth set, the model captured almost all the major roads and structures. We believe that if the model got fine-tuned on the Google Earth dataset, it would yield results identical to the results from the Massachusetts data set.

Figure 12: U-Net prediction on a Google Earth Sample

- *SMP DeepLabV3:*

performed well on the IOU and dice loss metrics, however, a few caveats showed when tested on the test data. The predicted masks did not reflect the performance. Figure 11 shows the results.

Figure 13: Prediction Result from SMP DeepLapV3

We investigated the results, applied a heatmap filter to the results [15], and were able to see the masks. We assume the reason was a wrongful configuration of how the model was configured by us.

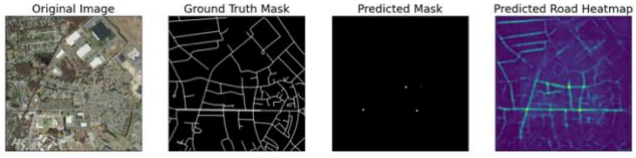

Figure 14: Results from Figure 13 with heatmap applied

Pinpointing the reason for such results proved hard to solve and opens an opportunity for more investigation and research for the future.

- Custom U-Net:

Although the model performed poorly compared to SMP DeepLabv3 on IOU and Dice Loss metrics. The model predictions were more accurate and reflective.

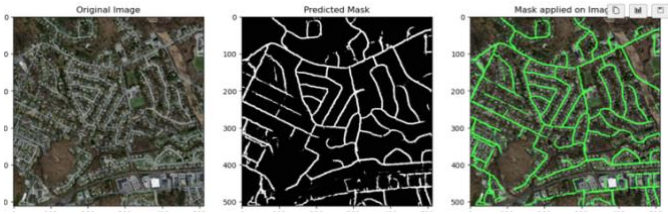

Figure 15: Prediction results from the Custom U-Net

The performance on the Google Earth samples was impressive, as shown in Figure 16.

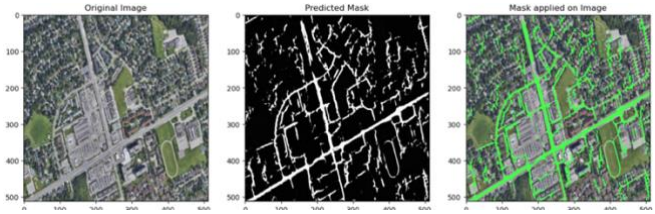

Figure 16: Custom U-Net prediction on a Google Earth Sample

To further understand what pattern our custom U-net captures compared to the SMP U-Net model, we introduced an image that contains different shape patterns and then compared the results of both models.

We found that our custom model mainly tries to capture straight lines between areas with different colors, showing not that much depth in capturing the context of the image. However, SMP U-Net output showed great results as it didn't predict any roads in the test image. The depth in context capturing shown by the SMP U-Net could be linked to it is pre-trained weights and the heavy optimizations included in the model.

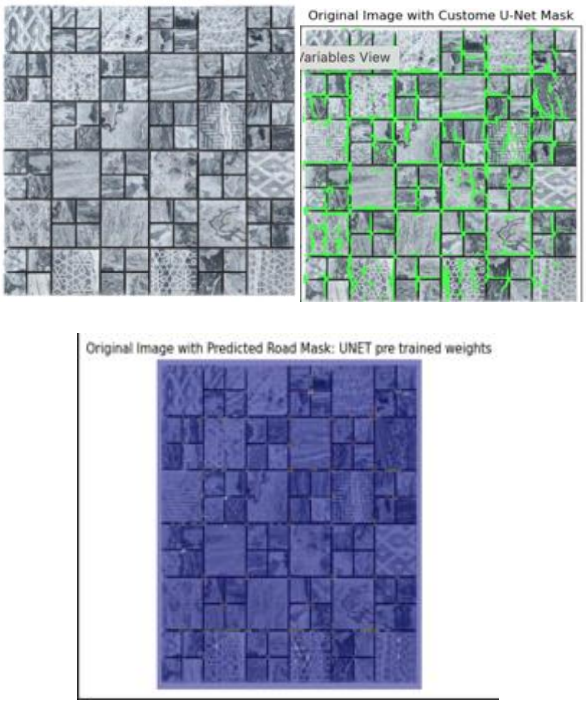

Figure 17: Pattern Image Tested by both models

VIII. CONCLUSIONS

This study represents a comprehensive examination of road segmentation in satellite imagery utilizing deep learning techniques, with a focus on three models: the Segmentation Models PyTorch (SMP) U-Net, DeepLabV3, and a custom-built U-Net. The analysis, conducted using the Massachusetts Roads Dataset, provides a detailed numerical assessment of each model's performance in terms of accuracy, computational efficiency, and real-world applicability.

Our results demonstrate that the SMP U-Net model achieved the highest accuracy, with an Intersection over Union (IoU) score of 0.82 and a Dice Coefficient of 0.049. This model excelled in capturing intricate road details, making it particularly effective for complex urban road networks. In contrast, DeepLabV3, while demonstrating acceptable IOU scores, performed poorly when tested. The custom U-Net model, designed as part of this research, exhibited a balanced profile.

A critical observation from our experiments was the relative stability of model performance after the initial training epochs. The SMP U-Net and DeepLabV3, with pre-trained weights on the ImageNet dataset, showed minimal improvement in accuracy beyond the first epoch. This phenomenon suggests that the initial training phase is crucial, and the benefits of extensive training diminish with these pre-optimized models. Our custom U-Net [9], however, displayed a more traditional learning curve, with gradual improvements in IoU scores (increasing from 0.11 to 0.62) over successive epochs, indicating potential areas for enhancement in its design and training protocol.

In real-world testing with Google Earth imagery, the SMP U-Net model accurately identified major roads and structures, closely resembling the ground truth. The custom U-Net, while performing less effectively on standard metrics, demonstrated surprising accuracy in these practical scenarios, suggesting that conventional metrics may not fully capture a model's real-world performance.

The findings highlight the importance of model selection based on specific requirements [12], such as the need for high accuracy or computational efficiency. Additionally, the results emphasize the potential of custom models to bridge gaps between existing solutions, providing a balanced approach to road segmentation challenges. This research not only contributes to the field of remote sensing and urban planning but also sets a foundation for future advancements in deep learning applications for satellite imagery analysis [17].

REFERENCES

[1] G Z. Chen, C. Wang, J. Li, W. Fan, J. Du, and B. Zhong, "Adaboost-like End-to-End multiple lightweight U-nets for road extraction from optical remote sensing images," *International Journal of Applied Earth Observations and Geoinformation*, vol. 100, 2021, 102341.

[2] A. Wulamu, Z. Shi, D. Zhang, and Z. He, "Multiscale Road Extraction in Remote Sensing Images," *Computational Intelligence and Neuroscience*, vol. 2019, Article ID 2373798, 2019, https://doi.org/10.1155/2019/2373798.

[3] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, [no volume or issue number provided], [no page range provided], [no year provided].

[4] T. Li, M. Comer, and J. Zerubia, "A Two-Stage Road Segmentation Approach for Remote Sensing Images," in *Proc. 26th International Conference on Pattern Recognition Workshops (ICPRw 2022), IAPR*, Montreal, Canada, Aug. 2022.

[5] S. Ruder, "An Overview of Gradient Descent Optimization Algorithms," arXiv:1609.04747v2 [cs.LG], 15 Jun. 2017.

[6] F. Bastani, S. Madden, "Beyond Road Extraction: A Dataset for Map Update using Aerial Images," arXiv:2110.04690, [no publication year provided].

[7] V. Mnih, "Machine Learning for Aerial Image Labeling," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.

[8] X. Huang and L. Zhang, “Road centreline extraction from highresolution imagery based on multiscale structural features and support vector machines,” IJRS, vol. 30, no. 8, pp. 1977–1987, 2009.

[9] Y. LeCun, B. Boser, J. S. Denker et al., “Backpropagation applied to handwritten zip code recognition,” Neural Computation, vol. 1, no. 4, pp. 541–551, 1989.

[10] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440, Boston, MA, USA, June 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, pp. 1097–1105, MIT Press, Cambridge, MA, USA, 2012.

[12] Maboudi M, Amini J (2015) Object based segmentation efect on road network extraction from satellite images. In: Proceedings of the 36th Asian conference on remote sensing, Manila, Philippines, October 2015. pp. 19–23.

[13] Zhang Z, Liu Q, Wang Y. Road extraction by deep residual u-net. IEEE Geosci Remote Sens Lett. 2018;15(5):749–53

[14] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234– 241

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[16] S. Piao and J. Liu, “Accuracy improvement of unet based on dilated convolution,” in Journal of Physics: Conference Series, vol. 1345, no. 5. IOP Publishing, 2019, p. 052066.

[17] Munawar, H.S., Hammad, A.W.A., Waller, S.T. et al. Road Network Detection from Aerial Imagery of Urban Areas Using Deep ResUNet in Combination with the B-snake Algorithm.Hum-Cent Intell Syst 3, 37–46 (2023). https://doi.org/10.1007/s44230-023-00015-5