MEKATRONIKA
JOURNAL OF MECHATRONICS AND INTELLIGENT MANUFACTURING
E-ISSN: 2637-0883
VOLUME 7, ISSUE 1, JANUARY 2025, 60 - 70
DOI: https://doi.org/10.15282/mekatronika.v7i1.12196

RESEARCH ARTICLE

River Segmentation Using Deep Neural Networks on Aerial Orthophoto

Abd. Hafiz Zakaria[1], Yasir M. Mustafah[1], Nor Rohaizah Jamil[2]

[1]Department of Mechatronics Engineering, Kulliyyah of Engineering International Islamic University Malaysia, P.O Box, 50728, Kuala Lumpur, Malaysia

[2]Faculty of Forestry and Environment, University Putra Malaysia, 43400, Selangor, Malaysia

ABSTRACT - This study investigates the use of deep learning for automated river segmentation from UAV-captured aerial orthophotos, addressing the limitations of traditional and labor-intensive river monitoring techniques. We introduced, annotated dataset of high-resolution river imagery and evaluated multiple semantic deep neural network architectures, including U-Net, FPN, PSPNet, and LinkNet, using ResNet50 as backbone models. To optimize performance, various image patch sizes were tested, with 768×768 pixels providing the best trade-off between segmentation accuracy (88.76%) and computational efficiency. Among the tested models, U-Net with a ResNet50 backbone achieved the highest segmentation performance, with an average Intersection over Union (IoU) of 61%, an F1-score of 73%, a precision of 74%, and a recall of 77%. These findings demonstrate the potential of UAV-based remote sensing and deep learning for enhancing the accuracy and efficiency of river monitoring.

ARTICLE HISTORY
Received : 21st March 2025
Revised : 13th April 2025
Accepted : 01st June 2025
Published : 10th June 2025

KEYWORDS
UAV
River
Segmentation
Deep Neural Networks
Remote sensing

1. INTRODUCTION

Rivers are vital components of ecosystems. River play a crucial role in maintaining biodiversity, supporting agriculture and contributing to economic activities such as hydropower generation and transportation. However, the increasing frequency and severity of flood disasters necessitate enhanced efforts in improving the inland water condition, particularly within the river system (Fu et al., 2020) . Traditional river monitoring approaches, including field surveys and manual analysis of satellite or aerial imagery, are labor-intensive, time-consuming, and often constrained by limited spatial and temporal coverage(Watanabe & Kawahara, 2016). As a result, there is a growing need for automated, cost-effective, and scalable methods for river segmentation and monitoring (La Salandra et al., 2021) .

With the advancements in remote sensing technology, Unmanned Aerial Vehicles (UAVs) have emerged as a powerful tool for capturing high-resolution aerial images of rivers for in remote sensing especially in flood risk assesment (Salmoral et al., 2020) UAVs offer a flexible, cost-effective alternative to satellite and manned aerial surveys, allowing for frequent data collection at fine spatial resolutions (Casado & Leinster, 2020). Their ability to operate in diverse environments and capture images under varying lighting and weather conditions makes them particularly suitable for river monitoring applications (Zhang et al., 2021). Moreover, with the availability of various types of UAV platforms, combined with photogrammetry computer vision algorithms there is a huge potential in providing a more efficient and low-cost approach for river flood assessment based on demand at a bigger scale (Xiang et al., 2018). Post-flood assessments utilizing UAVs for measuring high-water marks and river cross-sections have demonstrated greater accuracy and efficiency compared to traditional ground-based surveys (Forbes et al., 2020). Annis et al. (2020) conducted a comparative analysis of digital elevation models (DEMs) obtained from UAVs and LiDAR satellite data. Post-flood assessments utilizing UAVs for measuring high-water marks and river cross-sections have demonstrated greater accuracy and efficiency compared to traditional ground-based surveys (Forbes et al., 2020). Despite these advantages, extracting meaningful information from UAV-captured imagery remains a challenge due to the complexity of river environments, which include varying water turbidity, seasonal changes and occlusions caused by vegetation.

Deep learning, particularly Convolutional Neural Networks (CNNs), has demonstrated remarkable potential in addressing challenges related to flood monitoring by enabling automated image analysis with high (Muhadi et al., 2020). Researchers have leveraged CNN-based techniques to enhance water segmentation, a crucial process for detecting rising river levels and assessing flood risks. For instance, (Lopez-Fuentes et al., 2017) proposed an automated approach using water segmentation to monitor fluctuations in river levels effectively. CNN-based segmentation models have been widely employed in various remote sensing applications, including land cover classification and flood mapping, showcasing their versatility and effectiveness in analyzing geospatial data (Heffels & Vanschoren, 2020).Additionally, Zhang et al. (2021) applied deep learning techniques to river ice segmentation using UAV imagery, further expanding the scope of CNN applications in hydrological studies.

However, the implementation of deep learning models for river segmentation comes with unique challenges. These include selecting an optimal model architecture, handling variations in image resolution, and ensuring computational

***CORRESPONDING AUTHOR | Abd. Hafiz Zakaria** | ✉ abdhafizakaria@gmail.com

efficiency when processing large-scale datasets. To address these challenges, researchers have increasingly adopted advanced deep neural networks specifically designed for semantic segmentation tasks. State-of-the-art architectures such as U-Net, Feature Pyramid Network (FPN), Pyramid Scene Parsing Network (PSPNet), and LinkNet have demonstrated superior performance in extracting meaningful features from remote sensing images, making them well-suited for river segmentation and flood monitoring applications (Ronneberger et al., 2015). By leveraging these advanced models, researchers can improve segmentation accuracy, enhance flood prediction capabilities, and develop more robust monitoring systems for water-related disasters. Semantic segmentation holds crucial significance across a diverse range of applications within the field of geographic information. This technique plays a pivotal role in various geographic information applications, showcasing its multifaceted utility in extracting meaningful insights from spatial data (Lv et al., 2023). In the study of semantic segmentation, the goal is to connect or classify each pixel of an image into a certain class (Long et al., 2015)

This paper aims to leverage the potential of integrating UAV-based remote sensing with deep learning techniques to enhance river monitoring efficiency and accuracy. By automating river segmentation, this approach can significantly reduce the time and labor required for river assessments while improving monitoring precision.

2. METHODS AND MATERIAL

2.1 Study Area

The dataset for this study was collected from various river systems across Malaysia, which features an extensive network of waterways spanning Peninsular Malaysia. Specifically, data acquisition was conducted at three distinct sites along the Langat River to capture its morphological characteristics. The Langat River extends 78 km, with a catchment area of 2,350 km², originating from the Titiwangsa Range at Gunung Nuang and flowing westward into the Straits of Malacca. Each aerial survey in this study covered a 1 km reach of the river, with an approximate width of 35 m as shown in. Figure 1 shows an example of study site location at Persiaran Universiti, Bandar Baru Bangi which cover 1km long of Langat River. To enhance the diversity of training data, aerial photogrammetry was also conducted across multiple river systems, including Gombak river, Semenyih river, Semantan river, Kelantan river, and Perak river. This dataset serves as a valuable resource for comprehensive analysis, supporting research efforts toward the sustainable management of Malaysia's river networks.

Figure1. Study site of Langat river

2.2 Flight Mission

In this study, the aerial dataset was captured using a DJI Phantom 4 Pro. This quadcopter is equipped with an integrated 20MP CMOS camera (5472 × 3648 resolution) supported by a 3-axis stabilizing gimbal system. The DJI Phantom 4 Pro, a multi-rotor drone with a vertical takeoff weight of 1.4 kg, has a maximum flight endurance of approximately 25 minutes per fully charged 5870mAh Lithium Polymer (LiPo) battery. The quadcopter platform was selected over a fixed-wing alternative due to its vertical takeoff and landing capabilities, enabling deployment in confined areas without the need for a runway. Additionally, the drone is designed to withstand wind speeds of up to 10 m/s. Table 1 below shows the specification of the DJI Phantom 4 Pro.

Table 1. DJI Phantom 4 Pro specifications

Characteristics	Specifications
Sensor	1” CMOS (Effective pixels; 20M)
Lens	84° 8.8 mm/24 mm (35 mm format equivalent)
ISO range	100-3200
Electronic shutter speed	8–1/8000s
Image Size (columns x rows)	5472 x 3648

To ensure data acquisition safety and consistency, aerial imagery was collected under non-windy conditions, never exceeding the wind speed threshold. The flight path was pre-programmed using Pix4Dcapture and CTRL+DJI software (Android versions) to achieve comprehensive spatial coverage through a combination of longitudinal and cross-sectional multipasses. The waypoint navigation system followed a structured grid pattern, maintaining an 80% frontal overlap and 70% side overlap. Each waypoint represented the center of a frame, with every captured image geotagged with corresponding GPS coordinates. All flight operations were conducted at a fixed altitude of 100 m with a speed of 8.5 m/s, ensuring uniform image acquisition and resulting in a ground sampling distance (GSD) of 2.9 cm/pixel. The data collected will be processed using Agisoft Photoscan software to generate the point cloud data. Next, the point cloud data obtained will be processed to generate the Digital Elevation Model (DEM) and orthophoto image using orthorectification method as shown in Figure 2.

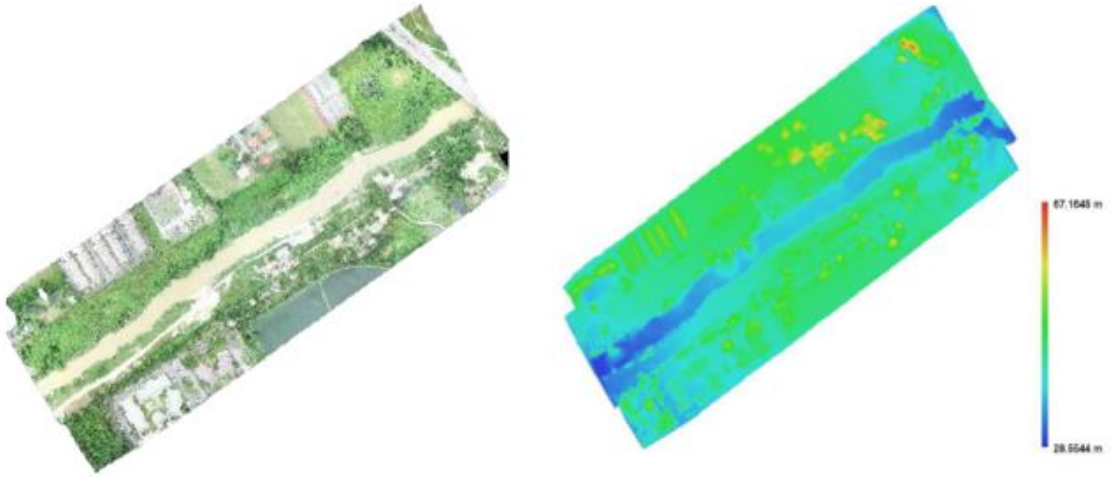

Figure 2. Example of orthophoto (left) and DEM (right) of Langat river

2.3 Data Labelling

In this study, the generated orthophoto image serves as a reference for creating ground truth (mask) images for semantic segmentation using Adobe Photoshop CC 2019. A mask layer within Adobe Photoshop CC 2019 was utilized to manually trace river features, as outlined in Table 2. Each river feature was carefully delineated using the lasso tool to classify different objects, with distinct colors assigned to differentiate between features. To enhance accuracy, a zooming technique was employed to refine feature labelling.

The labeled ground truth images were subsequently converted into one-hot encoded masks, a widely adopted format in machine learning applications, particularly in semantic segmentation and multi-class classification. In this representation, each class is assigned a separate channel, where pixels corresponding to a specific class are set to 1 while all others remain 0. To optimize computational efficiency, both the orthophoto and ground truth images were divided into smaller segments. The labeled images were then used as training data for the machine learning model. Figure 3 illustrates the semantic labeling process applied to the input image.

Table 2. River features label and description

Features	Description
Side Bars	Consolidated riverbed material along the margins of reach which is exposed at low flow
Vegetated Side Banks	Side bar presenting plant cover in more than 50% of its surface area
River	Water flows
Trees	Areas cover with trees or 50% cover with grass

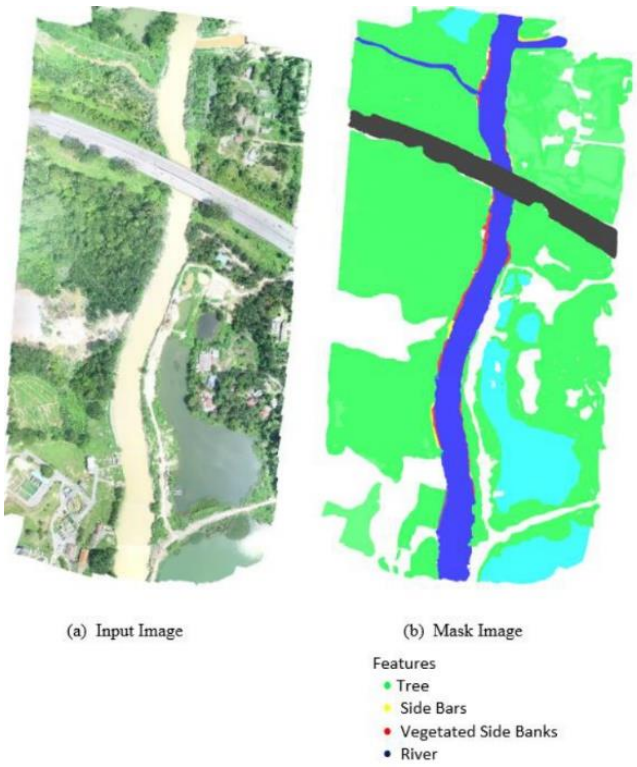

Figure 3. Example of labelled ground-truth image from orthophoto

2.4 Semantic Segmentation Network Architecture

In this study, various models and backbones were explored to develop an accurate segmentation model. The segmentation models and encoder backbones investigated are listed in Table 3. Structurally, all semantic segmentation architectures follow a similar encoder-decoder design, where the encoder extracts features while the decoder localizes spatial features (Hu et al., 2019). However, the key distinction between these architectures lies in how spatial and feature information are combined. Four models were considered for this study: (i) U-Net ((Ronneberger et al., 2015) (ii) LinkNet (Chaurasia & Culurciello, 2017), (iii) Pyramid Scene Parsing Network (PSPNet) (Zhao et al., 2017), and (iv) Feature Pyramid Networks (FPN) (Lin et al., 2017) as shown in Figure 4.

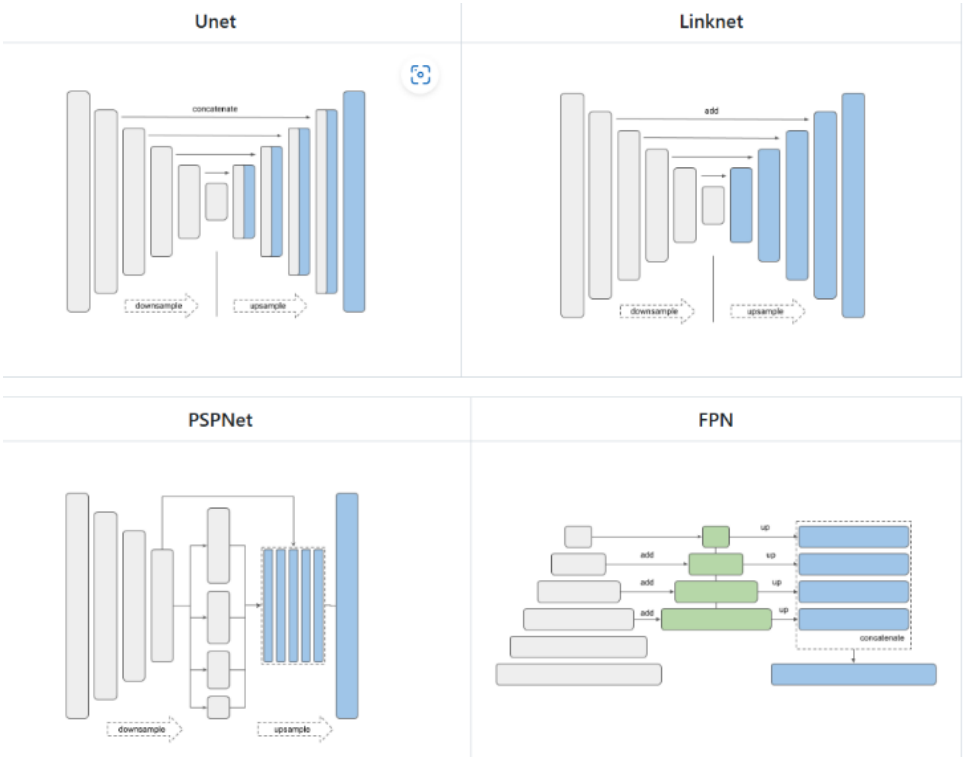

Figure 4. Segmentation models of this study

U-Net, originally developed for biomedical image segmentation, has gained widespread applicability in various domains, including remote sensing (Hu et al., 2019).It employs an encoder for multi-level feature extraction, while the decoder reconstructs resolution and learned features through a sophisticated stacking mechanism to account for both localization and feature representation (Ronneberger et al., 2015) LinkNet modifies U-Net by integrating the upsampled feature representation with resolution information instead of direct concatenation (Chaurasia & Culurciello, 2017).The two pyramid-based networks PSPNet and FPN adopt different strategies to construct a pyramid structure. PSPNet achieves this by variably pooling the lowest downsampled map to obtain multiple spatial resolutions that enrich feature representations (Zhao et al., 2017) In contrast, FPN constructs two pyramids and merges them to generate feature-rich segmentation maps at multiple levels (Lin et al., 20117).

ResNet50 was chose as encoder backbones to be evaluated in this study. ResNet50 (He et al., 2016) are pre-trained on 2012 ILSVRC ImageNet dataset. The encoder component from pre-trained model will capture semantic information, while the decoder components will facilitate the recovery of spatial details for accurate per-pixel classification. ResNet50 is a deep residual network that helps in training very deep architectures efficiently. Thus, it will leveraging its deep layers and pre-trained weights to extract robust, high-level features. ResNet50 provide a strong backbone for segmentation models by extracting meaningful features from images (Bianco et al., 2018).

Table 3. Segmentation models and backbone

Models	U-Net, FPN, LinkNet, PSPNet
Backbone (encoder)	ResNet50

2.5 Training Networks Parameters

For this study, TensorFlow will be used to train the semantic segmentation networks. Data augmentation is limited to horizontal and vertical flips to preserve the structural integrity of segmentation masks, while zooming and rotation are excluded to train each network with different encoder. The Adam optimizer is utilized with a learning rate of 0.0001. For loss optimization, categorical Focal Jaccard loss is employed, as it aligns with the objective of multi-class segmentation in aerial imagery. This combined loss function leverages Focal loss to address class imbalance issues while Jaccard loss directly optimizes segmentation quality. The integration of both losses provides a more robust optimization framework compared to using individually. Table 4 summarizes the network training parameters. For this study, the implementation was carried out using TensorFlow version 2.4.1 and Python version 3.7. The computational environment comprised a single NVIDIA GeForce RTX 2060 GPU, equipped with 6GB of VRAM and paired with 32GB of DDR4 system memory, with CUDA 11.2 (compute capability 7.5) providing GPU acceleration.

Table 4. Training parameters

Parameters	Value
Augmentation	
Loss Function	Categorical Focal Jaccard Loss
Optimizer	Adam Optimizer
Learning Rate	1×10^{-4}

2.6 Performance Evaluation Metrics

Performance evaluation is crucial especially in the segmentation process. A standard approach to model evaluation involves analyzing both the diagonal and non-diagonal elements of the confusion matrix, as illustrated in Figure 5. In this study, the trained model will be assessed using performance evaluation metrics, including Intersection-over-Union (IoU), F1 score, precision, and recall (equation 4,3,2,1) Below are performance evaluation metrics alongside their formulas. In these equations, TP (true positives), FP(false positives), FN (false negatives), and TN (true negatives) correspond to the counts of classification outcomes.

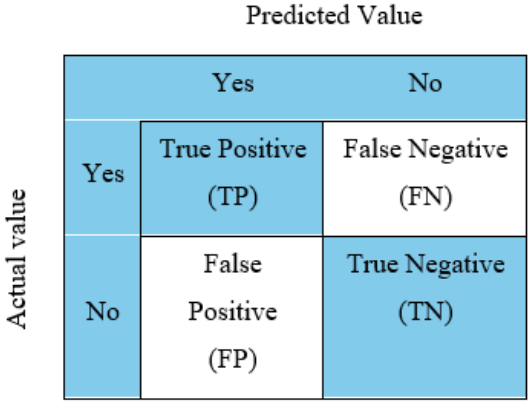

Figure 5. Confusion matrix

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2\,X\,\frac{Precision\ x\ Recall}{Precision\ x\ Recall} \tag{3}$$

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} = \frac{Area\left(B_{pred} \cap B_{groundtruth}\right)}{Area\left(B_{pred} \cup B_{groundtruth}\right)} = \frac{TP}{TP + FP + FN} \tag{4}$$

where:

TP (True Positive): The model predicted Positive and the actual class is Positive.
TN (True Negative): The model predicted Negative and the actual class is Negative.
FP (False Positive): The model predicted Positive but the actual class is Negative.
FN (False Negative): The model predicted Negative but the actual class is Positive.

3. RESULTS AND DISCUSSION

3.1 Aerial Dataset

In this study, a total of 47 high-resolution orthophoto images of rivers from various locations across Malaysia were acquired using a DJI Phantom 4. Each image size ranges between 207 MB and 730 MB, achieving a ground sampling distance (GSD) of 2.9 cm/pixel. Alongside RGB data, depth information was extracted from the Digital Elevation Model (DEM), forming a four-channel RGB-D dataset with an overall size of approximately 288 GB. Each image was manually annotated with pixel-wise semantic labels corresponding to four classes: river, side bars, trees, and vegetated banks. This detailed annotation process facilitates accurate and robust semantic segmentation.

To accommodate the substantial computational demands associated with model training, the original large-scale orthophotos (with dimensions of 17,124 × 14,214 pixels, as depicted in Figure 6) were systematically partitioned into smaller patches of predefined sizes, including 128 × 128, 256 × 256, 512 × 512, 768 × 768, and 1,024 × 1,024 pixels. Prior to cropping, image dimensions were adjusted to ensure divisibility by the selected patch size. The patches were extracted sequentially from the top-left corner without overlapping, ensuring distinct training samples. This patch-based strategy effectively reduces memory consumption and computational overhead, making it well-suited for deep learning applications. Figures 6 and 7 illustrate the original orthophoto and the corresponding patch images derived from this preprocessing approach.

Figure 6. Original orthophoto

Figure 7. Example of patches size 768x768

Following patch extraction, only images containing meaningful labeled content were retained for training. To quantify the proportion of labeled data within each patch, the ratio of unlabeled pixels (ULP) to the total pixel count (TPC) was computed based on Equation 5. If this ratio exceeded 95%, the patch was deemed to contain sufficient labeled information and was included in the training set. This selective retention process ensures that the training dataset accurately represents the target semantic classes, thereby improving the overall performance and reliability of the deep learning model. For model training, the dataset was partitioned into 80% training and 20% validation subsets. Table 4.1 presents the distribution of training and validation images across different patch sizes.

$$real\ information = \left[1 - \left(\frac{ULP}{TPC}\right)\right] x\ 100\% \tag{5}$$

3.2 Experiment 1: Optimizing Image Patch Size for Training

To achieve high segmentation accuracy during dataset processing, it is crucial to determine the optimal patch size for training. Given that orthophoto images from aerial surveys are too large to be processed in their entirety within GPU memory, the original images were partitioned into smaller tiles. This approach not only improves data manageability but also accelerates model training. Therefore, a series of experiments were conducted to identify the most effective tile size for training a U-Net model architecture, utilizing ResNet50 as the backbone network with RGB images as input. By systematically varying the dimensions of the image patches, we assessed their impact on both segmentation accuracy and computational efficiency. Table 5 provides a detailed summary of the experimental results, highlighting the trade-offs between different patch sizes in terms of model performance and resource utilization.

As shown in Table 4.3, larger patch sizes (i.e., 512×512 and above) consistently yielded higher segmentation accuracy. Accuracy improved from 73.69% at 128×128 to 88.93% at 1024×1024, as illustrated in Figure 8. However, the accuracy gains diminished with each incremental increase in patch size. Notably, the transition from 768×768 to 1024×1024 resulted in only a 0.17% accuracy improvement, while the computational time increased significantly from 458 ms/step to 1048 ms/step. This pattern suggests a point of diminishing returns, where the benefits of increased spatial context are counterbalanced by greater computational costs.

Larger patches enhance the model's ability to capture spatial structures and better align with the pretrained ResNet50 backbone, originally trained on 224×224 images. However, the increased training time and memory requirements present challenges, particularly in resource-constrained environments. While 1024×1024 yielded the highest accuracy, the 768×768 patch size emerged as a more practical choice, achieving an accuracy of 88.76% with a moderate computational demand of 458 ms/step. This balance between accuracy and efficiency is particularly beneficial in scenarios with limited computational resources, ensuring sufficient spatial coverage while effectively leveraging the pretrained ResNet50 encoder. Therefore, 768x768 patches size is selected as the training input size for following experiments.

Table 5. Table caption (no period at the end of the caption)

Patches size (WxH)	Accuracy(%)	Computing Times (ms/step)
128x128	73.69	24
256x256	80.44	65
512x512	85.16	224
768x768	88.76	458
1024x1024	88.93	1048

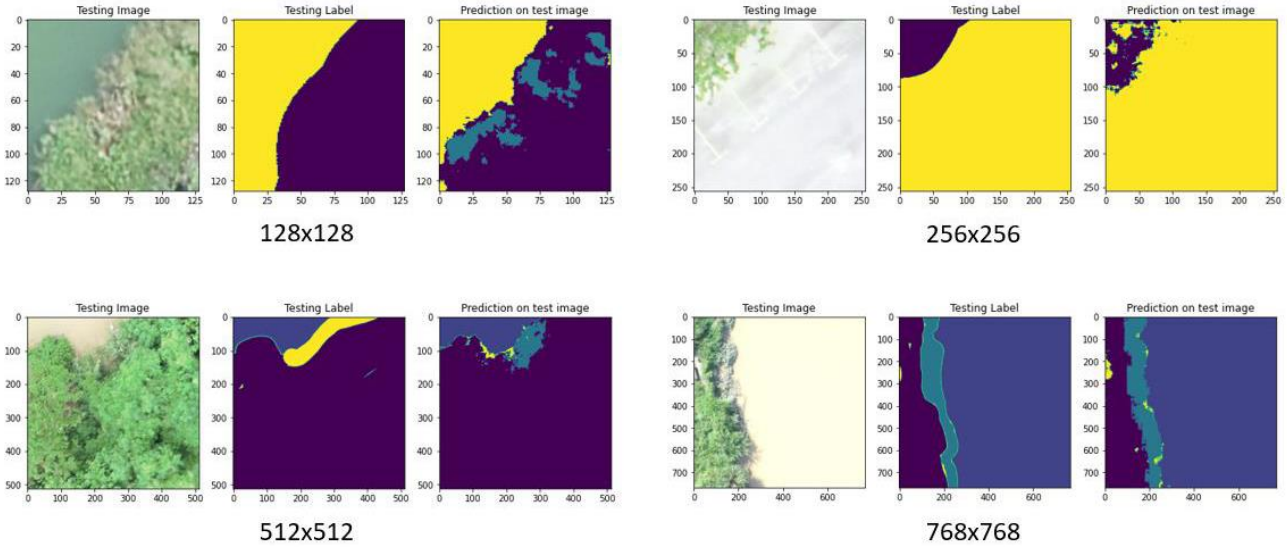

Figure 8. Figure caption (no period at the end of the caption)

3.3 Experiment 2: Segmentation Architecture with ResNet50 as a Backbone

In Experiment 2, four deep learning architectures U-Net, PSPNet, LinkNet, and FPN were evaluated for river segmentation using ResNet50 as the backbone. For consistent input dimensions across all training scenarios, 768x768 grayscale images were used. Training involved a uniform procedure of 25 epochs. Figure 9 shows the training accuracy of each tested model training for 25 epochs.

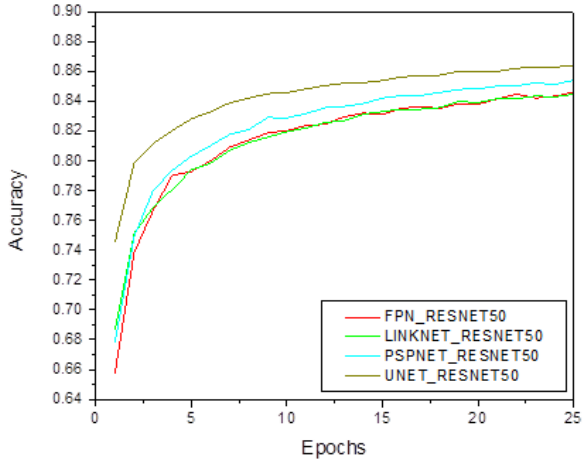

Figure 9. Training accuracy graph

The models were assessed based on their Intersection over Union (IoU), Precision, Recall, and F1-score, which provide a comprehensive evaluation of their segmentation effectiveness across different riverine features, including trees, rivers, side bars, and vegetated side bars as tabulated in Table 6. Among the tested models, U-Net demonstrated the highest overall accuracy (86.5%), achieving stable learning and strong generalization. It performed exceptionally well in segmenting rivers (IoU: 0.85, F1: 0.92) and trees (IoU: 0.84, F1: 0.91). However, it struggled with vegetated side bars (IoU: 0.29, F1: 0.45), likely due to class confusion and high visual similarity with other vegetation classes. The average IoU (0.61) and F1-score (0.73) indicate that U-Net effectively segments most classes while maintaining a balanced performance in terms of precision (0.74) and recall (0.77). Figure 10 shows the segmentation results of using U-Net.

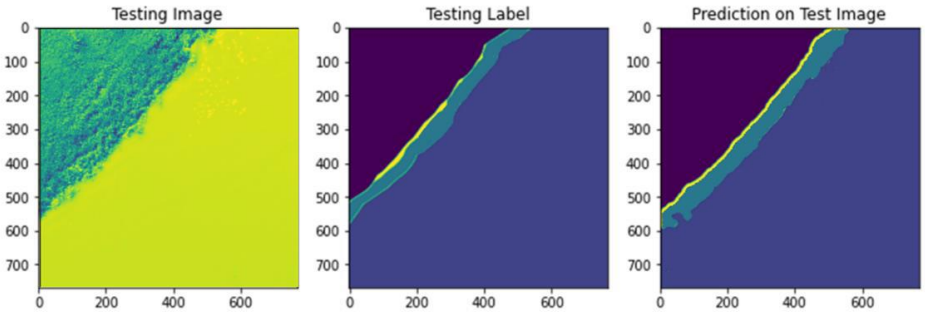

Figure 10. Segmentation results of U-Net with ResNet50

Table 6. Table caption (no period at the end of the caption)

Model	Labels	IoU	Precision	Recall	F1 Score
U-Net	Trees	0.84	0.88	0.94	0.91
	River	0.85	1.00	0.85	0.92
	Side Bars	0.46	0.76	0.54	0.63
	Vegetated Side Bars	0.29	0.32	0.74	0.45
	Average	0.61	0.74	0.77	0.73
FPN	Tress	0.83	0.94	0.88	0.9
	River	0.58	0.98	0.59	0.73
	Side Bars	0.52	0.58	0.83	0.68
	Vegetated Side Bars	0.14	0.16	0.52	0.25
	Average	0.52	0.67	0.71	0.64
PSPNet	Tress	0.84	0.94	0.89	0.91
	River	0.94	1.00	0.94	0.97
	Side Bars	0.47	0.53	0.82	0.64
	Vegetated Side Bars	0.10	0.5	0.11	0.18
	Average	0.59	0.74	0.69	0.68
Linknet	River	0.89	0.99	0.89	0.94
	Side Bars	0.45	0.52	0.78	0.62
	Vegetated Side Bars	0.1	0.24	0.14	0.18
	Average	0.49	0.57	0.69	0.61

PSPNet, with an accuracy of 85.4%, leveraged multi-scale context aggregation to enhance segmentation performance. It excelled in river segmentation (IoU: 0.94, F1: 0.97), surpassing all other models in this category. However, its effectiveness was moderate for side bars (IoU: 0.47, F1: 0.64) and limited for vegetated side bars (IoU: 0.10, F1: 0.18), likely due to challenges in distinguishing them from other terrain types. The average IoU (0.59) and F1-score (0.68) suggest that while PSPNet is highly effective for large water bodies, it requires further optimization for fine-grained terrain segmentation. Figure 11 shows the segmentation results of using PSPNet.

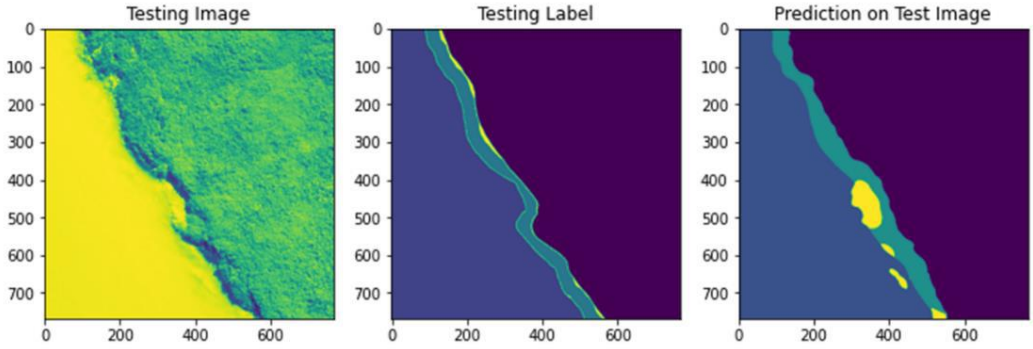

Figure 11. Segmentation results of PSPNet with ResNet50

LinkNet, achieving 84.6% accuracy, benefited from efficient feature propagation through skip connections. It performed well in river segmentation (IoU: 0.89, F1: 0.94) but exhibited low precision for side bars (0.52) and poor segmentation of vegetated side bars (IoU: 0.10, F1: 0.18). The average IoU (0.49) and F1-score (0.61) indicate that while LinkNet is computationally efficient, it lacks the robustness required for detailed segmentation tasks. In contrast, FPN recorded the lowest accuracy (84.4%), which may be attributed to multi-scale feature aggregation challenges. It demonstrated reasonable accuracy in tree segmentation (IoU: 0.83, F1: 0.90) but struggled with river segmentation (IoU: 0.58, F1: 0.73) and had low precision for vegetated side bars (0.16), highlighting its limitations in differentiating visually similar classes. The average IoU (0.52) and F1-score (0.64) suggest that although FPN is effective for segmenting tree structures, it lacks the capability for precise delineation of water bodies and vegetation. Figures 12 and 13 display the segmentation of using LinkNet and FPN architecture with ResNet50 as a backbone.

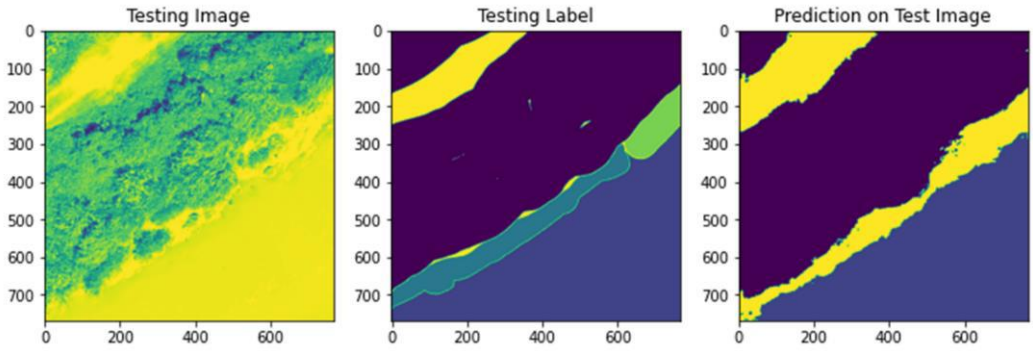

Figure 12. Segmentation results of LinkNet with ResNet50

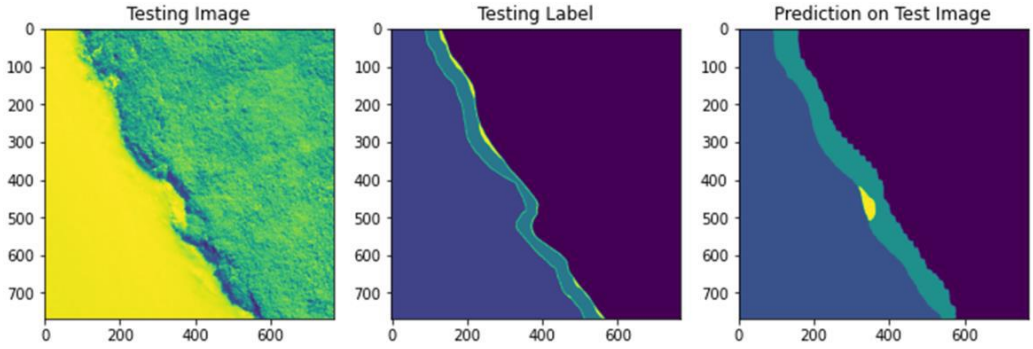

Figure 13. Segmentation results of FPN with ResNet50

Overall, the findings indicate that U-Net is the best-performing model, offering the highest segmentation accuracy across multiple classes. While PSPNet excelled in river segmentation, U-Net provided the best balance between accuracy and generalization, making it more suitable for multi-class segmentation. Both LinkNet and FPN exhibited lower accuracy, particularly struggling with side bars and vegetated areas due to class misclassification. Across all models, vegetated side bars posed the most significant segmentation challenge, primarily due to high visual similarity with trees and limited annotated training samples.

4. CONCLUSION

In conclusion, U-Net with a 768x768 patch size, achieves the most effective compromise between segmentation precision and computational resource utilization in aerial river segmentation. While PSPNet achieved superior river segmentation accuracy, U-Net's robustness across varied land cover categories positions it as the most appropriate model for this application. FPN and LinkNet yielded intermediate results, with FPN demonstrating challenges in accurately delineating intricate landscape features. These results highlight the efficacy of deep learning architectures that integrate both local and global feature representations for aerial river segmentation. In this research also introduced aerial river dataset for multi-class segmentation. Future work should consider evaluating diverse segmentation models with alternative encoder backbones, including MobileNetV2 and VGG16, and examining the impact of incorporating multi-modal input data, such as RGB and depth channels.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Ministry of Higher Education Malaysia (MOHE) for their generous financial support through the Fundamental Research Grant Scheme (FRGS) [Ref. No FRGS19-036-0644], which enabled this research..

REFERENCES

[1] A. Annis, F. Nardi, A. Petroselli, C. Apollonio E. Arcangeletti, F. Tauro, et al., "UAV-DEMs for small-scale flood hazard mapping," *Water (Switzerland)*, vol. 12, no. 6, p. 1717, 2020.

[2] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270–64277, 2018.

[3] M. R. Casado and P. Leinster, "Towards more effective strategies to reduce property level flood risk: Standardising the use of unmanned aerial vehicles," *Journal of Water Supply: Research and Technology - AQUA*, vol. 69, no. 8, pp. 807–818, 2020.

[4] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation," *2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1-4, 2017.

[5] B. T. Forbes, G. P. DeBenedetto, J. E. Dickinson, C. E. Bunch, and F. A. Fitzpatrick, "Using small unmanned aircraft systems for measuring post-flood high-water marks and streambed elevations," *Remote Sensing*, vol. 12, no. 9, p. 1437, 2020.

[6] G. Fu, F. Meng, M. Rivas Casado, and R. S. Kalawsky, "Towards integrated flood risk and resilience management," *Water (Switzerland)*, vol. 12, no. 6, p. 1789, 2020.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[8] M. R. Heffels and J. Vanschoren, "Aerial imagery pixel-level segmentation," *arXiv preprint arXiv:2012.02024*, 2020.

[9] J Hu, L Li, Y Lin, F Wu, J Zhao, "A comparison and strategy of semantic segmentation on remote sensing images," *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, pp. 21-29, 2019.

[10] M. La Salandra, G. Miniello, S. Nicotri, A. Italiano, G. Donvito, G. Maggi, et al., "Generating UAV high-resolution topographic data within a FOSS photogrammetric workflow using high-performance computing clusters," *International Journal of Applied Earth Observation and Geoinformation*, vol. 105, p. 102600, 2021.

[11] T. Y Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125. 2017

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431-3440, 2015.

[13] L. Lopez-Fuentes, C. Rossi, and H. Skinnemoen, "River segmentation for flood monitoring," *2017 IEEE International Conference on Big Data (Big Data)*, pp. 3746-3749, 2017.

[14] J. Lv, Q. Shen, M. Lv, Y. Li, L. Shi, and P. Zhang, "Deep learning-based semantic segmentation of remote sensing images: A review," In *Frontiers in Ecology and Evolution*, vol. 11, p. 1201125, 2023.

[15] N. A. Muhadi, A. F. Abdullah, S. K. Bejo, M. R. Mahadi, and A. Mijic, "Image segmentation methods for flood monitoring system," *Water (Switzerland)*, vol. 12, no. 6, p. 1825, 2020.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234-241, 2015.

[17] G. Salmoral, M. Rivas Casado, M. Muthusamy, D. Butler, P. P. Menon, and P. Leinster, "Guidelines for the use of unmanned aerial systems in flood emergency response," *Water (Switzerland)*, vol. 12, no. 2, p. 521, 2020.

[18] Y. Watanabe and Y. Kawahara, "UAV photogrammetry for monitoring changes in river topography and vegetation," *Procedia Engineering*, vol. 154, pp. 317–325, 2016.

[19] T. Z. Xiang, G. S. Xia, and L. Zhang, "Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications, and prospects," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 3, pp. 29-63, 2018.

[20] X. Zhang, Y. Zhou, J. Jin, Y. Wang, M. Fan, N. Wang, et al., "Icenetv2: A fine-grained river ice semantic segmentation network based on UAV images," *Remote Sensing*, vol. 13, no. 4, pp. 1–17, 2021.

[21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881-2890, 2017.