Automatic Extraction of Buildings From UAV-Derived Orthophotos Using a Deep Neural Network Approach*

[1]R. A. Nsiah, [1]S. Mantey, and [1]Y. Y. Ziggah
[1]University of Mines and Technology, Tarkwa, Ghana

Nsiah, R. A., Mantey, S. and Ziggah, Y. Y., (2023), "Automatic Extraction of Buildings From UAV-Derived Orthophotos Using a Deep Neural Network Approach", *Ghana Journal of Technology*, Vol. 7, No. 2, pp. 19-24.

Abstract

Buildings are the most common element in urban environments, and their accurate and efficient extraction from remotely sensed data is essential for various applications, such as urban planning and monitoring, population estimation, disaster planning, management and response, and updating geographic databases. However, conventional techniques for extracting buildings from remotely sensed data pose challenges due to the complexity and difference in buildings, changes in scenery, imaging sensors, and conditions. They require expert knowledge and expertise, thus undermining the applicability of these conventional approaches and making them time-consuming. The aim of the research is to evaluate the applicability and performance of deep neural networks (DNNs) in accurately identifying and delineating buildings within the Ghanaian context. To accomplish this objective, a supervised learning approach was adopted, and the U-Net model and its variant UResNet-34 were trained on a labelled dataset. The dataset comprised labelled UAV-derived orthophotos capturing urban areas with diverse architectural styles and building patterns in Ghana. The evaluation results indicated that U-Net and UResNet-34 models achieved promising performance in building extraction tasks. Remarkably, the U-ResNet-34 model, benefiting from the residual connections of ResNet-34, exhibited improved performance compared to the original U-Net model. The implications of these findings are significant, as they contribute to identifying informal settlements and estimating population density. Additionally, it aids in disaster response planning and post-event damage assessment. In conclusion, this study highlights the effectiveness of DNNs for automatically extracting buildings from UAV orthophotos in the Ghanaian context, offering valuable insights for informed decision-making in urban and environmental domains.

Keywords: Buildings, Extraction, UAVs, Deep Neural Networks, U-Net, ResNet-34

1 Introduction

The rapid urbanisation and population growth witnessed in recent years have propelled the demand for efficient land management and urban planning processes. Accurately identifying and extracting buildings from remotely sensed data is pivotal in supporting decision-making for sustainable urban development, disaster response, and environmental monitoring (Erdem and Avdan, 2020). Unmanned Aerial Vehicles (UAVs) equipped with advanced sensors capture high-resolution images for generating orthophotos, providing a valuable data source for detailed mapping and analysis of urban environments (Dey *et al*., 2022). These UAV-derived orthophotos offer an up-to-date and cost-effective means of acquiring geospatial information, thus creating an opportune environment for automated building extraction to provide timely and accurate information for urban planners and authorities (Sumer and Turker, 2013)

Conventional techniques for building extraction from UAV-derived orthophotos rely typically on manual feature engineering and rule-based algorithms, thus presenting formidable challenges due to the complexity and diversity of buildings, variations in scenery, imaging sensors, and environmental conditions (Chuangnong Li *et al*., 2021; Liu *et al*., 2020) Furthermore, the expert-driven nature of these conventional approaches hampers their applicability and renders them time-consuming, impeding timely and large-scale building detection (Abdollahi and Pradhan, 2021). Traditional Convolutional Neural Netwroks (CNNs) can automatically learn hierarchical representations from data and capture intricate features. Thus, CNNs have shown better results in image segmentation tasks than handcrafted computer vision algorithms.

In recent years, DNNs have become a transformative force in computer vision tasks, achieving state-of-the-art object detection and segmentation performance (Guo *et al*., 2020). The remarkable capability of DNNs to automatically learn hierarchical features from data and capture intricate patterns, rather than relying on handcrafted features, has garnered significant attention in the remote sensing community (Jin *et al*., 2021; Zhang *et al*., 2020). This has led to a shift towards exploring the potential of DNNs in automating building extraction processes from remotely sensed data (Hu *et al*., 2021). CNNs have since been advocated and utilised to automate the semantic segmentation of buildings. For example, Saito *et al*. (2016) proposed using CNN-based roads and buildings segmentation algorithm using aerial images.

Similarly, Alshehhi *et al*. (2017) presented a single batch-based CNN to segment roads and buildings from high-resolution remotely sensed images.

**Manuscript received June 11, 2023*
Revised version accepted September 20, 2023

Although these studies attained remarkable results, the proposed models could not perform the dense pixel-wise predictions, resulting in relatively fuzzy object boundaries and limiting the models' suitability for building segmentation from high-resolution images. Long *et al*. (2015) put forward the fully convolutional network (FCN) architecture to address the limitations of traditional CNNs in semantic segmentation. FCNs extended the CNN structure to permit dense prediction, facilitating pixel-level classification and segmentation (Abdollahi and Pradhan, 2021). In addition, FCNs can produce output feature maps that maintain the input dimensions, thus providing precise object boundaries and enhancing the accuracy of building extraction from remote sensing imagery (Liu *et al*., 2019). These capabilities are demonstrated in the studies by Bittner *et al*. (2017) and Maggiori *et al*. (2017). However, traditional FCNs are associated with inefficient and inaccurate performance as these networks drop low-level feature maps containing significant and intricate details during the segmentation process and only utilise high-level feature maps (Yang *et al*., 2018). Various encoder-decoder networks capable of reusing low-level feature maps with rich details have been proposed to address these limitations. U-Net, proposed by Ronneberger *et al*. (2015), has gained significant attention in building segmentation tasks among the encoder-decoder networks. U-Net utilises skip concatenation connections, allowing it to harness valuable information from both the encoder and decoder components, resulting in precise and well-defined building boundaries. Furthermore, U-Net can be trained using less data and requires fewer computational resources (Pan *et al*., 2020). Thus, this study chooses U-Net as the primary network for building segmentation.

The motivation for this work arises from the absence of prior investigations that have specifically evaluated the applicability and performance of DNNs in accurately identifying and delineating buildings from UAV-derived orthophotos within the Ghanaian context, where the demand for efficient urban planning and management processes has grown exponentially due to rapid urbanisation and population growth. A supervised learning paradigm was adopted to achieve this objective by employing the U-Net model and its variant UResNet-34, designed by replacing the original U-Net's encoder path with a residual network with thirty-four (34) layers (ResNet-34).

The key contributions of this research include:

- Investigating the application of U-Net and UResNet-34 for building extraction from UAV-derived orthophotos within the Ghanaian context.
- Evaluating the generalisation capabilities of the proposed approach across diverse urban landscapes.
- Providing insights into the potential challenges and future directions for advancing automatic building extraction methods using DNNs.

The subsequent sections of this paper delve into the study area, data utilised, detailed methodology, and experimental results, followed by discussions of the findings, research implications, and future research directions.

1.1 Related Works

This section presents a comprehensive analysis of related works focusing on building extraction from high-resolution remote sensing images using U-Net and its variants. The selection of papers encompasses various research studies that have leveraged U-Net and its modifications to address the challenges of building segmentation in complex urban environments.

Pan *et al*. (2020) proposed a U-Net-based urban village mapping paradigm to characterise individual buildings in high-density urban settlements. The U-Net model achieved impressive results, with an overall accuracy of over 86% for building segmentation and over 83% for classification, demonstrating the feasibility and efficiency of deep learning for mapping unplanned urban settlements. Liu *et al*. (2020) applied U-Net with a ResNet encoder to remote sensing image segmentation for building extraction. Their model achieved a high MIoU of 0.83, demonstrating its effectiveness in accurately segmenting buildings. However, the authors acknowledged the need for further improvements, such as refining building outlines and reducing misclassifications. Guo *et al*. (2020) proposed AMUNet, a multi-loss neural network based on U-Net with an attention mechanism for building segmentation. The proposed showcased significant improvements in building segmentation results compared to other methods.

The model demonstrated superior performance on public datasets, validating its efficiency for building object extraction from aerial imagery. Wagner *et al*. (2020) introduced U-net-id, a CNN architecture for building instance segmentation. The proposed method achieved excellent semantic and individual instance segmentation results, with a mean IoU of 0.582 and an overall accuracy of 97.67% for delineating individual buildings. Erdem and Avdan (2020) compared diverse U-Net variants for building extraction, including Vgg16 U-Net, InceptionResNetV2 U-Net, DenseNet121 U-Net, and a majority voting method. Evaluation using the Inria Aerial Image Labelling Dataset showed

promising results for the models, with the majority voting approach achieving the best F1 score of 0.877. Abdollahi and Pradhan (2021) introduced MultiRes-UNet, an enhanced version of U-Net, for building extraction. The model utilised the MultiRes block and convolutional operations with skip connections to enhance feature learning and achieve better results, outperforming other state-of-the-art models.

Wei *et al*. (2021) presented a U2-net model for building outline extraction by modifying the binary cross-entropy loss function. Compared to other models, the model achieved better accuracy, precise positioning, and refined building outlines without needing postprocessing steps like non-maximum suppression. Li *et al*. (2021) introduced a robust DNN termed HAC U-net. The model leverages attention units to replace certain skip connections, capturing varying receptive fields and enhancing spatial and contextual information extraction at different scales. Experimental results demonstrated that HAC U-net outperformed other baseline models, achieving an accuracy of 93.90% and an impressive intersection over union (IoU) of 61.92%. Xu *et al*. (2021) presented HA U-Net, a holistically-nested attention U-Net, which exhibited improved segmentation accuracy for building extraction from high-resolution remote sensing images. By incorporating attention mechanisms, multi-scale nested modules, and a weighted loss function, the proposed model successfully addressed the challenge of blurry segmentation in higher-resolution images. Li *et al*. (2021) presented a simple yet powerful U-Net variant capable of efficiently extracting buildings from farmland areas using Google and Worldview images. The model effectively addressed the challenges of complex ground features and scattered buildings in farmland areas. By incorporating ResNet architecture into the U-Net framework and introducing both spatial and channel attention mechanisms, the resulting model attained remarkable performance with an accuracy of 97.47% and an F1 score of 85.61%., outperforming other semantic models.

In a separate study, El Asri *et al*. (2022) introduced a convolutional neural network (CNN)-based system that combined the strengths of the U-Net and VGG19 architectures for building extraction from high-resolution satellite imagery. Integrating these two architectural approaches resulted in notable improvements in model accuracy, underscoring the potential of hybrid CNN systems in this domain.

Temenos *et al*. (2022) investigated the utilisation of U-Net and its various adaptations, including the Residual U-Net and Attention U-Net, for the automated extraction of buildings from high-resolution satellite imagery. The study's findings highlighted the enhanced performance of U-Net architectures, particularly in accurately pinpointing building structures along their corners and edges.

The reviewed works demonstrate the widespread application and effectiveness of U-Net and its variants in building extraction from high-resolution remote sensing images. Various modifications, such as attention mechanisms, residual blocks, and multi-scale fusion, have been introduced to improve the accuracy as well as performance of U-Net-based models. However, the datasets primarily utilised aerial and satellite images from developed countries, characterised by well-laid-out buildings. In contrast, access to such imagery is limited and costly for developing regions like Ghana. Furthermore, the freely available datasets in these areas are often outdated and insufficient to keep up with the rapid pace of urbanisation in Ghana. As such, this study proposes to exploit the benefits of UAV orthomosaics and U-Nets and its variant, UResNet-34, to segment buildings within Ghana.

2 Study Areas and Data Used

The datasets used for this study were obtained mainly from southern Ghana, notably Greater Accra, Western, and Central Regions. The data for training the DNNs consisted of orthophotos of the East Legon Area (ELA) – depicted in Fig. 1(a), and UMaT Area (UA) – depicted in Fig. 1(b), located in the Accra and Tarkwa townships. These areas were chosen due to the data availability and various architectural styles and building patterns, providing a representative sample to train and evaluate the DNN models. For ELA, a WingtraOne vertical take-off and landing (VTOL) UAV was used to acquire aerial images and subsequently processed using Pix4D mapper to generate an orthomosaic, which is a georeferenced, seamless, and orthorectified representation of the area. The orthomosaic had a spatial resolution of 3 cm/p and covered an area of 148.84 ha.

For UA, aerial images were collected using a Phantom 4 Pro (P4P) UAV, and Agisoft Metashape was exploited for the orthomosaic generation. The resulting orthomosaic had a spatial resolution of 7.5 cm/p and covered an area of approximately 130.60 ha. Both orthophotos were of three red-green-blue (RGB) bands, and their structures were manually digitised. The digitising was performed using QGIS software and ensures that an accurate and detailed annotation, capturing the complex geometries of buildings in the orthomosaic, is obtained. ELA had 2221 manually digitised polygons representing building structures, while UA had 152 structures annotated as buildings.

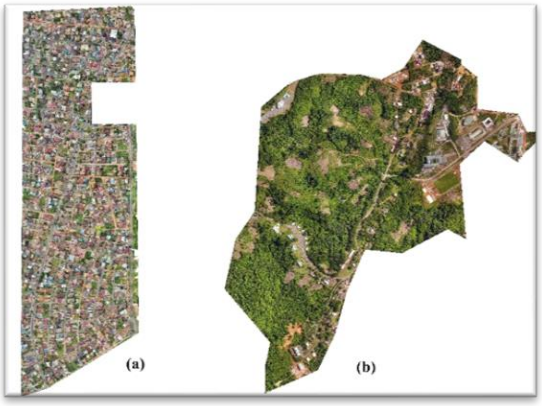

Fig. 1 (a) Orthomosaic of ELA and (b) Orthomosaic of UA

The shapefiles, obtained after the digitising process and containing the outlines of each building, are then converted to binary masks. In the binary mask, pixels representing building areas were designated as white while background pixels were set to black, each having pixel values of 1 and 0, respectively. These binary masks served as ground truth representations, precisely indicating the locations of buildings in the orthomosaics. Next, the binary masks with the original RGB images are used to create training data pairs. Each training pair is then sliced into 256×256 image dimensions, allowing the data to fit computer memory and providing ready-to-use data for DNN training. In addition, the training pair enables the DNNs to learn the spatial correspondence between building footprints and their visual appearance. Lastly, data augmentation techniques were applied to increase the generalisation potential of the DNNs and escape overfitting. These techniques include random rotations, translations, scaling, and flipping of the binary masks and corresponding RGB images. Generally, augmentation diversifies the training data and helps the DNN learn to handle building appearance, orientation, and scale variations. Fig. 2 depicts the general workflow for generating the training data.

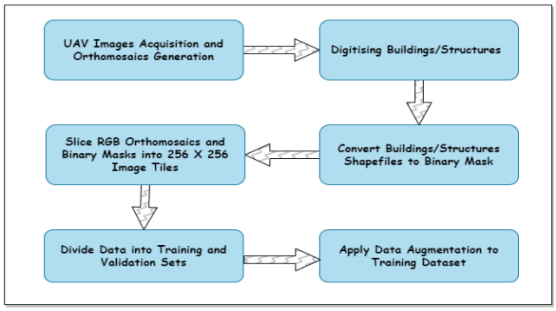

Fig. 2: Training Data Generation Workflow

To evaluate the trained DNNs, orthomosaics from four distinct localities, each characterised by different building configurations and designs, were utilised. The first locality, L1, predominantly featured buildings with well-designed and organised architectural structures. In the second locality, L2, the buildings were predominantly slums, with a few large structures interspersed. Locality 3, known as L3, was distinguished by sparse buildings and abundant vegetation, with the buildings' rooftops having a similar texture to the surrounding vegetation—lastly, the fourth locality, L4, comprised buildings of varying architectural designs and sizes. While most buildings in L4 were properly laid out, a small section exhibited a slum layout. Furthermore, the generalisability and capability of the proposed DNNs over more extensive areas were also verified using two other orthomosaics. Fig. 3 and Table 1 represent the orthomosaics of these test localities and further details regarding the training, test, and verification datasets, respectively

.

Fig. 3: Orthomosaics for Evaluating Trained DNNs

Table 1 Details of Datasets Used

S. No	Area	Region	UAV Used	Purpose	Resolution
1	East Legon	Greater Accra	WingtraOne	Training	3 cm/p
2	UMaT	Western	P4P	Training	5 cm/p
3	L1 – Spintex- Manet	Greater Accra	P4P	Evaluation	4 cm/p
4	L2 – UMaT	Western	P4P	Evaluation	5 cm/p
5	L3 – UMAT	Western	P4P	Evaluation	5 cm/p
6	L4 – East Legon	Greater Accra	P4P	Evaluation	3 cm/p
7	Abnabna	Central	P4P	Verification	4.5 cm/p

3 Proposed Methodology

This section describes the architecture of the two DNNs utilised for this study, mainly U-Net and UResNet-34.

3.1 U-Net

U-Net is a DNN introduced by Ronneberger *et al.* (2015) and was designed for semantic segmentation tasks. U-Net is known for its unique and efficient design, making it widely adopted and influential in various image segmentation applications. Its architecture is based on an FCN and is characterised by its U-shaped pattern, giving it its name. The network consists of two main parts: the contracting path (encoder) and the expansive path (decoder), as shown in Fig. 4.

The encoder's responsibility is to acquire high-level features and the broader context from the input images. It consists of a series of convolutional layers and max-pooling layers. These convolutional layers perform local feature extraction by progressively reducing the spatial dimensions of the feature maps while increasing the network's depth. The max-pooling layers, on the other hand, reduce the spatial dimensions, allowing the network to capture a larger context. The convolutional and max-pooling layers' operations enable the contracting path to extract abstract representations of the input images, enabling the network to understand the objects in the images.

The decoder works with the encoder to reconstruct the segmented output. It employs two repeated 3 × 3 convolution kernels, a down-sampling layer of 2 × 2 window size combined with a rectified linear unit (ReLU) activation function, and a 2 × 2 transpose convolutional layer to up-sample the feature maps. This configuration enables the encoder to recover the spatial resolution lost during the pooling operations in the encoder. The up-sampled feature maps are concatenated with the corresponding feature maps from the encoder via skip connections. These skip connections permit the network to fuse multi-scale information effectively and enhance the network's ability to perform precise localisation by combining local details directly from the contracting path to the global features in the expansive path. Ultimately, a convolution layer with a 1 × 1 kernel and a sigmoid function transforms every feature map into the desired outputs.

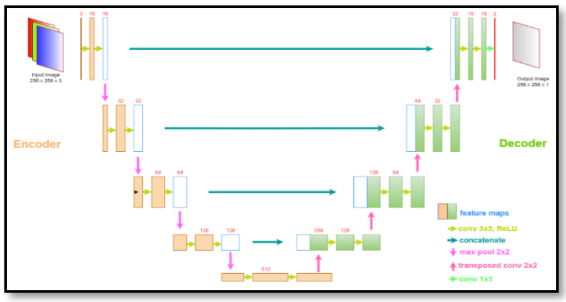

Fig. 4: Architecture of U-Net

3.2 U-Net with ResNet-34 Backbone (UResNet-34)

To further enhance the performance of the U-Net, a ResNet-34 network is used to replace the decoder of the U-Net model. ResNet is a revolutionary architecture proposed by He *et al.* (2016), which introduced the concept of residual learning. It addressed the issue of vanishing gradients during the training of DNNs. The ResNet architecture uses shortcut connections, also known as residual connections, to skip one or more layers, allowing the network to learn residual mappings directly. The ResNet-34 is a variant of the original ResNet architecture with 34 layers. The fusion of multi-scale information through skip connections and the powerful representation learning capabilities of ResNet-34 facilitate smoother and more efficient gradient flow during backpropagation, enabling feature reuse across layers, reducing the number of parameters, and balancing the model's performance and accuracy. Fig. 5 illustrates the architecture of the ResNet-34 model.

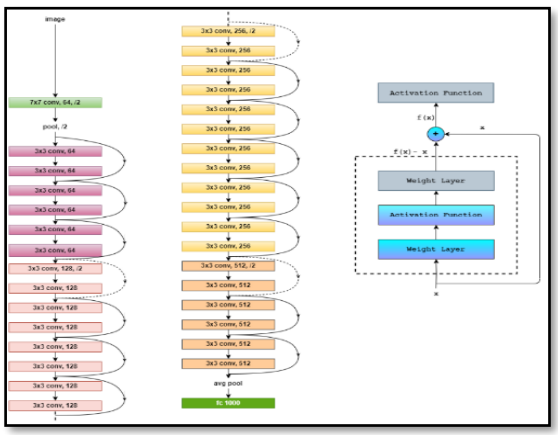

Fig. 5: Architecture of ResNet-34

3.3 Evaluation Metrics

Evaluating the performance of a DNN for a segmentation task is essential as it enables assessing its accuracy and effectiveness in accurately delineating objects of interest in an image. Several metrics are frequently employed to assess the effectiveness of a segmentation network, with each

metric offering valuable perspectives on various facets of the model's performance. This study utilised six metrics: accuracy, precision, recall, F-1 score, mean intersection over union (mIoU), and Cohen's kappa, to evaluate the U-Net and UResNet-34 models.

Accuracy is the fraction of correctly segmented pixels (foreground and background) to the sum of the number of pixels in the image. Precision measures the ability of the model to distinguish positive samples among the predicted positive samples correctly and is computed as the fraction of correctly detected positive targets by the total sum of targets detected as positive. Recall computes the ability of the model to identify positive samples among all the actual positive samples. It is determined by dividing true positives by the total actual positives. F1-score is the harmonic mean of precision and recall and provides a balanced measure of the model's performance. Intersection over union (IoU) measures the spatial alignment between the predicted and ground truth segmentation masks, evaluating the ratio of their intersection to their union. mIoU is the average of IoU, and it is calculated across all classes in the segmentation task, providing an overall indicator of the model's effectiveness in segmenting diverse objects in the image. Kappa is a statistical metric that considers the agreement between the predicted segmentation and the ground truth while also considering the agreement that might occur by chance. Equations 1 to 6 represent the mathematical formulation for six evaluation metrics.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1-score} = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

$$\text{MIoU} = \frac{1}{K}\sum_{i=1}^{K}\frac{P \cap G}{P \cup G} \quad (5)$$

$$\text{Kappa} = \frac{P_0 - P_e}{1-p_e} \quad (6)$$

where, TP, TN, FP, and FN stand for true positive (accurately predicted building pixels), true negative (correctly predicted non-building pixels), false positive (incorrectly predicted building pixels that are not), and false negative (missed non-building pixels), respectively. K represents the number of classes, which is 2 in this study. P refers to the predicted buildings, while G represents the ground truth P_0 indicates the proportion of observed agreement (Accuracy), and P_e depicts the proportion of agreement expected by chance. All the metrics, except kappa, have values ranging from 0 to 1, where 0 indicates the poorest prediction, and 1 represents the best prediction performance. Kappa has values ranging from -1 to 1, where -1 indicates no agreement, 0 indicates agreement by chance, and 1 indicates perfect agreement

4 Experimental Results

4.1 Dataset Preparation and Postprocessing

The RGB image tiles and their corresponding binary masks (5045 images and 5045 masks) were randomly divided into training (80%) and validation (20%) datasets. The purpose of the training dataset is to provide the model with the necessary information and visual properties of the buildings. The validation data, on the hand, aids in verifying and improving the model's performance during training. Data augmentation was performed on the training dataset to obtain 12000 images and 12000 masks.

The testing images were purposely chosen to have dimensions of 5000×5000 pixels, primarily to match the computer's processing capabilities. This image dimension was carefully selected to ensure the computational resources could handle the size effectively. Similarly, the verification images were initially divided into tiles with dimensions of 5000×5000 pixels each before applying the proposed models. The models were then used to predict building locations on each tile, resulting in binary images representing the presence or absence of buildings in each tile.

After the prediction phase, the postprocessing steps involved combining the predicted binary images from the tiles to reconstruct the complete image. A retiling process was applied to assemble the individual tiles into the original image dimension, and an image-to-image georeferencing technique was employed to align the retiled images with their corresponding orthomosaics. This georeferencing step ensures that the predicted binary images are accurately positioned and have appropriate geographic coordinates.

Once the images were georeferenced, they were further processed to obtain vector layers representing the building outlines. This raster-to-vector conversion allowed for the extraction of precise building boundaries. During this conversion, a simplification algorithm was applied to enhance the regularity of the building outlines, ensuring smoother and more coherent representations. The ultimate goal of these postprocessing steps was to obtain accurate and well-defined building outlines from the initial binary predictions.

4.2 Experimental Design

The experimental setup used Python programming language and open-source libraries, including TensorFlow, OpenCV, NumPy, and Segmentation Models. The overall flow for conducting the experiments for the proposed models was in three phases: model development and training, model testing, and verification. These phases were conducted on a Windows operating system, leveraging the computational power of a GeForce RTX 2060 GPU equipped with 16 GB of RAM. During the training process, a data generator with a batch size of 16 was employed to efficiently read and process the images and their corresponding masks from the training and validation datasets.

These datasets were subsequently fed into the U-Net and UResNet-34 models for further analysis and performance evaluation. During the training of the proposed models, model checkpointing and early stopping were utilised. Model checkpointing enables saving the model's weights at certain intervals during training. Thus, the model can later be restored from the point of best performance and can be helpful to prevent data loss and resume training if the process is interrupted. Early stopping works by monitoring a validation metric, in this study, was the mIoU, and stopping the training process early if the mIoU for the validation set stops improving. This step helps prevent overfitting and saves computational resources. The training details for the U-Net and UResNet-34 are depicted in Table 2.

Table 2 Training Details for Proposed DNN Models

Detail	Models	
	U-Net	UResNet-34
Total training time	105 mins	98 mins
Number of epochs	50	39
Batch size	16	16
Loss function	Binary cross entropy	Categorical focal jaccard loss
Optimiser	Adam	Adam
Input image size	256 × 256 RGB images	256 × 256 RGB images
Input mask size	256 × 256 RGB binary images	256 × 256 RGB binary images
Output image size	256 × 256 building mask	256 × 256 building mask
Training images	12000	12000
Validation images	1009	1009

4.3 Quantitative Assessment

Table 3 presents the results achieved by the U-Net and UResNet-34 for the four localities based on a quantitative assessment using the evaluation metrics. The results revealed stimulating insights into their segmentation capabilities. For L1, UResNet-34 demonstrated superiority over the U-Net model across various metrics. With an accuracy of 0.8712, UResNet-34 exhibited a higher ability to classify pixels than U-Net (0.7432) correctly. For precision, U-Net and UResNet-34 achieved similar and impressive scores of 0.9882 and 0.9881, respectively, indicating proficiency in making accurate positive predictions. The recall value of 0.6806 for UResNet-34 showed its capability to identify a more significant portion of the actual positive pixels, while U-Net's recall was 0.5136. The F1-score (0.8061), which balances precision and recall, and mIoU (0.7496), which indicates a better overall segmentation, for UResNet-34 were significantly higher compared to the F1-score (0.6759) and mIoU (0.5799) of U-Net. However, for Kappa, U-Net attained a superior value of 0.9244 compared to the 0.8855 attained by UResNet-34 in this test area. This score indicates that the U-Net model predictions agreed better with the ground truth data than that of UResNet-34.

Table 3 Quantitative Performance of DNN Models on L1

Test Area	Metric	Model	
		U-Net	UResNet-34
L1	Accuracy	0.7432	0.8712
	Precision	0.9881	0.9882
	Recall	0.5136	0.6806
	F1-score	0.6759	0.8061
	mIoU	0.5799	0.7496
	Kappa	0.9244	0.8855

For L2, UResNet-34 continued to outperform the U-Net model in specific metrics. It achieved a higher accuracy of 0.9010 compared to U-Net's 0.8192. However, U-Net exhibited a slightly better precision (0.7208) than UResNet-34 (0.6474). Remarkably, UResNet-34 excelled in recall, with a value of 0.7864, indicating its ability to identify a more significant proportion of actual positive pixels, while U-Net's recall was 0.5009. The F1-score was marginally higher for UResNet-34 (0.7363) compared to U-Net (0.5911), reflecting a better balance between precision and recall for UResNet-

34. The 0.7223 mIoU score attained by UResNet-34 was higher than that of U-Net (0.6058). U-Net showed a better kappa value (0.8667) than UResNet-34 (0.8495) in Test Area L2. This value indicates that U-Net's predicted masks agreed more with this area's ground truth.

Table 4 Quantitative Performance of DNN Models on L2

Test Area	Metric	Model	
		U-Net	UResNet-34
L2	Accuracy	0.8192	0.9010
	Precision	0.7208	0.6474
	Recall	0.5009	0.7864
	F1-score	0.5911	0.7363
	mIoU	0.6058	0.7223
	Kappa	0.8667	0.8495

In L3, UResNet-34 achieved remarkable results, excelling in various metrics. It achieved a high accuracy of 0.9786, significantly surpassing U-Net's accuracy of 0.9269. UResNet-34 also demonstrated higher precision (0.9285) and recall (0.5937) compared to U-Net's precision (0.9059) and recall (0.2810). The F1-score for UResNet-34 (0.7243) was higher, reflecting its ability to balance precision and recall. Moreover, UResNet-34 showed excellent performance in mIoU (0.7695) compared to the 0.7695 of U-Net. In this test area, U-Net beat UResNet-34 in terms of kappa value by attaining a score of 0.9020, surpassing the 0.8857 attained by UResNet-34.

Table 5 Quantitative Performance of DNN Models on L3

Test Area	Metric	Model	
		U-Net	UResNet-34
L3	Accuracy	0.9269	0.9786
	Precision	0.9059	0.9285
	Recall	0.2810	0.5937
	F1-score	0.4290	0.7243
	mIoU	0.5909	0.7695
	Kappa	0.9020	0.8857

In L4, UResNet-34 maintained its strong performance, achieving an accuracy of 0.9220, higher than U-Net's 0.8949. Both models displayed similar precision values, but UResNet-34 showcased a higher recall of 0.8093 than U-Net's recall of 0.7518. The F1-score for UResNet-34 (0.8639) was also higher, indicating its ability to balance precision and recall. The mIoU (0.8284) value of UResNet-34 also surpassed the mIoU (0.7793) obtained by UNet. U-Net continued outperforming UResNet-34 regarding the kappa value (0.9020 vs. 0.8857).

Table 6 Quantitative Performance of DNN Models on L4

Test Area	Metric	Model	
		U-Net	UResNet-34
L4	Accuracy	0.8949	0.9220
	Precision	0.9058	0.9265
	Recall	0.7518	0.8093
	F1-score	0.8216	0.8639
	mIoU	0.7793	0.8284
	Kappa	0.9020	0.8857

4.4 Qualitative Assessment

As illustrated in Fig. 6, the qualitative assessment of the U-Net and UResNet-34 models for localities 1 to 4 reveals valuable insights into their segmentation performance. In Test Area 1, both models show promising results, with U-Net achieving slightly higher accuracy in building segmentation despite some minor false positives. UResNet-34, however, segmented a portion of the road as a building, as indicated by the red markings. In L2, both U-Net and UResNet-34 struggled to identify buildings accurately in the areas with informal settlements. The qualitative evaluation of the models in L3, characterised by sparse buildings and dense vegetation, shows a remarkable performance for both U-Net and UResNet-34. Although the buildings had roofs with similar textures to the vegetation, the vegetation was not segmented as buildings or buildings mistaken for vegetation. The models exhibit similar performance for L4. Although there were false positives, most of the buildings were accurately segmented/ Overall, the U-Net and UResNet-34 models demonstrate promising performance in building segmentation, with each model excelling in all contexts.

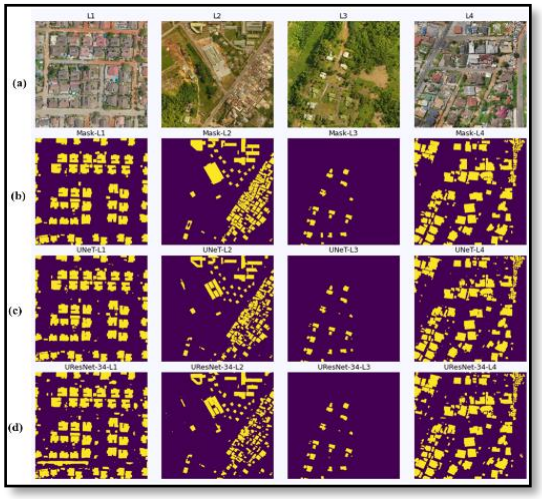

Fig. 6 Performance of DNN on Test Localities (a) Orthomosaics (b) Masks (c) U-Net Predictions (d) UResNet-34 Predictions

4.5 Verification Study

The study area chosen for this stage was the Abnabna community, covering an approximate area of 82.92 hectares. Data was collected using a P4P UAV, and the entire process took about twenty-two (22) minutes. The collected images were then processed using Agisoft Metashape to generate the orthomosaic.

For obtaining segmentation masks of the area, both U-Net and UResNet-34 models were employed; however, only one was chosen due to the similarity between the results masks. The selected mask was further tiled, georeferenced, and converted into a shapefile representing the building's outlines. Fig. 7 illustrates the orthomosaic, georeferenced mask, building shapefile, and building shapefile superimposed on the orthomosaic.

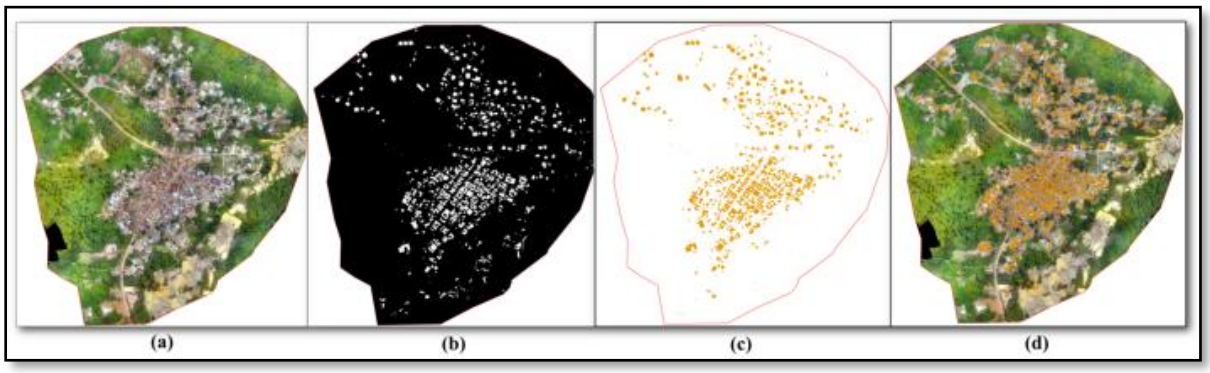

Fig. 7 DNN Models Verification Results (a) Orthomosaic (b) Tiled Mask (c) Building Outlines (d) Superimposition of Building Outlines and Orthomosaic

4.6 Discussion

The quantitative assessment demonstrated the effectiveness of U-Net and UResNet-34 in building extraction from UAV orthomosaics. In all test areas, u-Net achieved competitive performance with respectable accuracy, precision, F1-score, and mIoU values. On the other hand, UResNet-34 exhibited even higher values for these metrics, outperforming U-Net in most cases. The higher recall and F1-score values for UResNet-34 suggest its proficiency in identifying buildings accurately while maintaining a good balance between precision and recall. However, the kappa values present a different perspective. U-Net consistently achieved higher kappa values across all test areas, indicating better agreement and accuracy in its predictions than UResNet-34. The kappa metric considers the agreement by chance, making it a robust measure of model performance. U-Net's ability to maintain a higher level of agreement with the ground truth data implies its overall consistency and reliability in building segmentation tasks.

The qualitative evaluation revealed the effectiveness of U-Net and UResNet-34 architectures in accurately segmenting buildings from the UAV orthomosaics. While there were occasional false positive predictions, the models demonstrated remarkable proficiency in identifying and extracting most buildings within the images. However, in the case of the slum locality, L2, the models encountered challenges in distinguishing adjacent buildings with overlapping roofs or those situated in close proximity. Additionally, the UResNet-34 model exhibited difficulties in areas where the road textures resembled those of the buildings' roofs.

Considering the training times, U-Net and UResNet-34 exhibited relatively shorter durations of 105 minutes and 98 minutes, respectively. The models' training times are reasonable and feasible, making them suitable for various practical applications. U-Net and UResNet-34 balance training efficiency and performance, making them attractive for building extraction tasks in remote sensing and geodetic applications.

4.7 Research Implication

This study's findings have significant implications for remote sensing, geospatial analysis, and urban planning, providing valuable insights into deep learning models' application in building extraction from UAV orthomosaics. U-Net and UResNet-34 demonstrate potential in automating building identification and segmentation, benefiting various real-world applications. They enhance building extraction accuracy, supporting urban planning and development projects with reliable building footprints and better land use management. Additionally, these models streamline the process, reducing manual effort and time in data processing and analysis.

Beyond urban planning, the research contributes to various geospatial applications like land cover classification and change detection while improving environmental monitoring, natural resource management, and infrastructure assessment efficiency. In disaster management, accurate

building extraction aids prompt response efforts, assessing affected areas during disasters for targeted responses.

Furthermore, automation gains importance in geospatial applications, as deep learning models like U-Net and UResNet-34 can automate tasks, reducing reliance on manual processes. This shift improves scalability and data analysis efficiency. The research also advances deep-learning methodologies in remote sensing, providing insights into U-Net and UResNet-34's strengths and weaknesses and guiding their application for specific tasks.

5 Conclusions

In this research, a comprehensive investigation was conducted to ascertain the performance of U-Net and UResNet-34 models for building extraction from UAV orthomosaics. The experimental results demonstrated that both models achieved promising outcomes, with UResNet-34 exhibiting superior performance in various evaluation metrics, including accuracy, precision, recall, F1-score, mIoU, and kappa. Deep-learning models, particularly U-Net and UResNet-34, showcased their effectiveness in automating the building extraction process, leading to more accurate and precise building footprints. These findings have significant implications for various applications, including urban planning, disaster management, environmental monitoring, and infrastructure assessment. The automation of building extraction streamlines the data processing workflow, reducing manual effort and time and facilitating data-driven decision-making for better land use management and sustainable urban development. While U-Net and UResNet-34 demonstrated commendable performance, it is essential to highlight certain weaknesses. One of the challenges observed was the limited performance in slum areas, where the buildings had overlapping roofs. Additionally, the models faced difficulties in accurately identifying buildings in areas with irregular building patterns or the presence of shadows and occlusions. These limitations suggest that further research is necessary to enhance the models' ability to handle diverse and complex building scenarios.

Future research directions will focus on addressing the limitations of the models by exploring advanced architectural modifications and data augmentation techniques. Incorporating attention mechanisms or multi-scale fusion methods could help improve the models' capability to effectively capture fine details in densely vegetated areas and handle occlusions. Additionally, the integration of contextual information and auxiliary data sources, such as LiDAR or hyperspectral data, will be explored. These data sources could enhance the models' performance in challenging environments.

References

Abdollahi, A. and Pradhan, B. (2021), "Integrating semantic edges and segmentation information for building extraction from aerial images using UNet", *Machine Learning with Applications*, Elsevier Ltd., Vol. 6, pp. 1–10.

Alshehhi, R., Marpu, P. R., Woon, W. L. and Mura, M. D. (2017), "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 130, pp. 139–149.

El Asri, S. A., El Adib, S., Negabi, I. and Raissouni, N. (2022), "A Modular System Based on U-Net for Automatic Building Extraction from very high-resolution satellite images", *E3S Web of Conferences*, Vol. 351, p. 01071.

Bittner, K., Cui, S. and Reinartz, P. (2017), "Building extraction from remote sensing data using fully convolutional networks", *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, Vol. 42, No. 1W1, pp. 481–486.

Dey, M. S., Chaudhuri, U., Banerjee, B. and Bhattacharya, A. (2022), "Dual-Path Morph-Unet for Road and Building Segmentation from Satellite Images", *IEEE Geoscience and Remo-te Sensing Letters*, IEEE, Vol. 19, pp. 1–5.

Erdem, F. and Avdan, U. (2020), "Comparison of Different U-Net Models for Building Extraction from High-Resolution Aerial Imagery", *International Journal of Environment and Geoinformatics*, Vol. 7, No. 3, pp. 221–227.

Guo, M., Liu, H., Xu, Y. and Huang, Y. (2020), "Building extraction based on U-net with an attention block and multiple losses", *Remote Sensing*, Vol. 12, No. 9, pp. 1–17.

He, K., Zhang, X., Ren, S. and Sun, J. (2016), "Deep residual learning for image recognition", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2016-Decem, pp. 770–778.

Hu, Q., Zhen, L., Mao, Y., Zhou, X. and Zhou, G. (2021), "Automated building extraction using satellite remote sensing imagery", *Automation in Construction*, Elsevier BV, Vol. 123, No. September 2020, p. 103509.

Jin, Y., Xu, W., Zhang, C., Luo, X. and Jia, H. (2021), "Boundary-aware refined network for automatic building extraction in very high-resolution urban aerial images", *Remote Sensing*, Vol. 13, No. 4, pp. 1–20.

Li, C., Fu, L., Zhu, Q., Zhu, J., Fang, Z., Xie, Y., Guo, Y. and Gong, Y. (2021), "Attention enhanced u-net for building extraction from farmland based on google and worldview-2 remote sensing images", *Remote Sensing*, Vol. 13, No.

21, available at:https://doi.org/-10.3390/rs132-14411.
Li, C., Liu, Y., Yin, H., Li, Y., Du, P., Zhang, L. and Guo, Q. (2021), "Hybrid attention cascaded U-net for building extraction from aerial images", *Proceedings - 2021 7th International Conference on Big Data Computing and Communications, BigCom 2021*, pp. 294–301.
Liu, W., Yang, M. Y., Xie, M., Guo, Z., Li, E. Z., Zhang, L., Pei, T. and Wang, D. (2019), "Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network", *Remote Sensing*, Vol. 11, No. 24, pp. 1–18.
Liu, Z., Chen, B. and Zhang, A. (2020), "Building segmentation from satellite imagery using U-Net with ResNet encoder", *Proceedings - 2020 5th International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2020*, pp. 1967–1971.
Long, J., Shelhamer, E. and Darrell, T. (2015), "Fully Convolutional Networks for Semantic Segmentation", *IEEE Conference on Computer Vision and Pattern (CVPR), Boston, MA, USA, 7–12 June 2015*, pp. 3431–3440.
Maggiori, E., Tarabalka, Y., Charpiat, G. and Alliez, P. (2017), "Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 55, No. 2, pp. 645–657.
Pan, Z., Xu, J., Guo, Y., Hu, Y. and Wang, G. (2020), "Deep learning segmentation and classification for urban village using a worldview satellite image based on U-net", *Remote Sensing*, Vol. 12, No. 10, pp. 1–17.
Ronneberger, O., Fischer, P. and Brox, T. (2015), "U-net: Convolutional networks for biomedical image segmentation", *In International Conference on Medical Image Computing and Computer-Assisted Intervention, MICAI 2015*, pp. 234–241.
Saito, S., Yamashita, T. and Aoki, Y. (2016), "Multiple object extraction from aerial imagery with convolutional neural networks", *Journal of Imaging Science and Technology*, Vol. 60, No. 1, pp. 0104021–0104029.
Sumer, E. and Turker, M. (2013), "An adaptive fuzzy-genetic algorithm approach for building detection using high-resolution satellite images", *Computers, Environment and Urban Systems*, Vol. 39, pp. 48–62.
Temenos, A., Temenos, N., Doulamis, A. and Doulamis, N. (2022), "On the Exploration of Automatic Building Extraction from RGB Satellite Images Using Deep Learning Architectures Based on U-Net", *Technologies*, Vol. 10, No. 1, p. 19.
Wagner, F. H., Dalagnol, R., Tarabalka, Y., Segantine, T. Y. F., Thomé, R. and Hirye, M. C. M. (2020), "U-net-id, an instance segmentation model for building extraction from satellite images-Case study in the Joanopolis City, Brazil", *Remote Sensing*, Vol. 12, No. 10, pp. 1–14.
Wei, X., Li, X., Liu, W., Zhang, L., Cheng, D., Ji, H., Zhang, W. and Yuan, K. (2021), "Building outline extraction directly using the u2-net semantic segmentation model from high-resolution aerial images and a comparison study", *Remote Sensing*, Vol. 13, No. 16, pp. 1–20.
Xu, L., Liu, Y., Yang, P., Chen, H., Zhang, H., Wang, D. and Zhang, X. (2021), "HA U-Net: Improved Model for Building Extraction from High-Resolution Remote Sensing Imagery", *IEEE Access*, IEEE, Vol. 9, pp. 101972–101984.
Yang, H., Wu, P., Yao, X., Wu, Y., Wang, B. and Xu, Y. (2018), "Building extraction in very high-resolution imagery by dense-attenti networks", *Remote Sensing*, Vol. 10, No. 11, pp. 1–16.
Zhang, P., Du, P., Lin, C., Wang, X., Li, E., Xue, Z. and Bai, X. (2020), "A hybrid attention-aware fusion network (Hafnet) for building extraction from high-resolution imagery and lidar data", *Remote Sensing*, Vol. 12, No. 22, pp. 1–20.

Authors

Richmond Akwasi Nsiah is a Ph.D. candidate at the Geomatic Engineering Department of the University of Mines and Technology (UMaT), Tarkwa, Ghana. He holds a BSc degree from UMaT. His research interest includes the application of Remote Sensing, UAV Photogrammetry, Geospatial Analyses, and artificial intelligent applications in the Built Environment.

Saviour Mantey is an Associate Professor at the Department of Geomatic Engineering of the University of Mines and Technology (UMaT), Tarkwa, Ghana. He holds a Bachelor of Science degree in Geomatic Engineering from the Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. He obtained his Master of Philosophy degree and Doctor of Philosophy from the University of Cambridge and the University of Mines and Technology, respectively. He is a Professional member of the Ghana Institution of Surveyors (GhIS), a member of the Canadian Remote Sensing Society (CRSS), a member of Aerial Cartographic and Remote Sensing Association (ACRA), and a member of the Licensed Surveyors Association of Ghana (LISAG). His research interest includes the application of Remote Sensing and GIS in Health and Environmental Analysis, UAVs, and Web GIS applications.

Yao Yevenyo Ziggah is a Senior Lecturer at the Geomatic Engineering Department of the University of Mines and Technology (UMaT). He holds a BSc in Geomatic Engineering from Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. He obtained his Master of Engineering and Doctor of Philosophy degrees in Geodesy and Survey Engineering from China University of Geosciences (Wuhan), P. R. China. His research interests include artificial intelligent application in engineering, geodetic coordinate transformation, gravity field modelling, height systems and geodetic deformation modelling.