International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, Delaware, USA
Volume 13– No. 3, July 2026 – www.ijais.org

An Enhanced U-Net Architecture with Attention Gates and Atrous Spatial Pyramid Pooling for Building Segmentation in Aerial Imagery

Joseph Ngwa
Faculty of Engineering and Technology (FET)
Department of Computer Engineering, University of Buea
P.O. Box 63, Buea, Cameroon

Elie Fute Tagne
Faculty of Engineering and Technology (FET)
Department of Computer Engineering, University of Buea
P.O. Box 63, Buea, Cameroon

Nde Nguti
Faculty of Engineering and Technology (FET)
Department of Computer Engineering, University of Buea
P.O. Box 63, Buea, Cameroon

ABSTRACT
Accurate building segmentation from high-resolution aerial imagery remains challenging due to variations in building size, shape, background clutter, and class imbalance. This study proposes an Enhanced U-Net architecture for automatic building extraction using the Massachusetts Buildings Dataset and the Inria Aerial Image Labeling Dataset. The proposed model extends the conventional U-Net by incorporating a deeper encoder with Batch Normalization, Attention Gates (AGs) in decoder skip connections, and an Atrous Spatial Pyramid Pooling (ASPP) module for multi-scale contextual feature extraction. Different ASPP dilation-rate configurations were investigated, with dilation rates of 3, 6, and 9 yielding the best performance. To handle the high spatial resolution of the imagery, a patch-based training strategy with 50% overlap was adopted, using patch sizes of 256 × 256 and 512 × 512 for the Massachusetts and Inria datasets, respectively. The model was optimized using a hybrid Binary Cross-Entropy (BCE) and Dice loss function to balance pixel-level classification and region-overlap accuracy. Experimental results demonstrate that the proposed Enhanced U-Net outperformed U-Net, DeepLabv3+, and HRNet. On the Massachusetts Buildings Dataset, it achieved an IoU of 0.737 and an F1-score of 0.848. On the Inria dataset, it achieved an IoU of 0.799 and an F1-score of 0.888. These results demonstrate the effectiveness and generalization capability of the proposed architecture for aerial building segmentation.

Keywords
Building Segmentation; Aerial Imagery; Enhanced U-Net; Attention Gates; Atrous Spatial Pyramid Pooling; Patch-Based Training.

1. INTRODUCTION
Accurate and efficient building segmentation from aerial imagery plays a crucial role in various applications, including urban planning, disaster management, and environmental monitoring [1]. With the increasing availability of high-resolution aerial imagery, automated building extraction has gained significant attention in remote sensing. Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have emerged as powerful tools for image segmentation tasks.

Among various CNN architectures, U-Net, with its encoder-decoder structure and skip connections, has demonstrated remarkable success in biomedical image segmentation and has been widely adopted for building segmentation from aerial imagery [2]. However, the complexity and variability inherent in aerial images, which are characterized by diverse building structures, scales, and occlusions, pose significant challenges for standard U-Net models.

To address these challenges, this study proposes an enhanced U-Net architecture tailored for building segmentation in aerial images. This study focuses on five key improvements as follows:

1. Deepening the encoder for a richer feature representation.
2. Incorporating Batch Normalization to stabilize training.
3. Adding Attention Gates for focused feature fusion.
4. Employing Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale context.
5. Using loss functions tailored to class imbalance.

The effectiveness of the proposed model was evaluated using the Massachusetts Buildings Dataset and the Inria Aerial Image Labeling Dataset. These are large-scale benchmark datasets comprising high-resolution aerial images with diverse building types and densities. The model's efficiency was also compared with other standard models (U-Net, DeepLabv3+, and HRNet).

The remainder of this paper is structured as follows: Section 2 provides a concise overview of related work on building segmentation using U-Net deep learning. Section 3 details the proposed enhanced U-Net architecture and its components. Section 4 describes the experimental setup, including the dataset, evaluation metrics, and training procedure. Section 5 presents the experimental results and discusses the performance of the enhanced model compared to other standard models. Finally, Section 6 concludes the paper and outlines the future research directions.

2. RELATED WORK
Semantic segmentation, particularly for building footprint extraction in aerial and satellite imagery, has witnessed significant advancements, primarily driven by convolutional neural networks (CNNs) such as U-Net and its variants. This section outlines the evolution of segmentation architectures and the incorporation of enhancements, such as attention mechanisms and atrous convolutions. It also highlights key

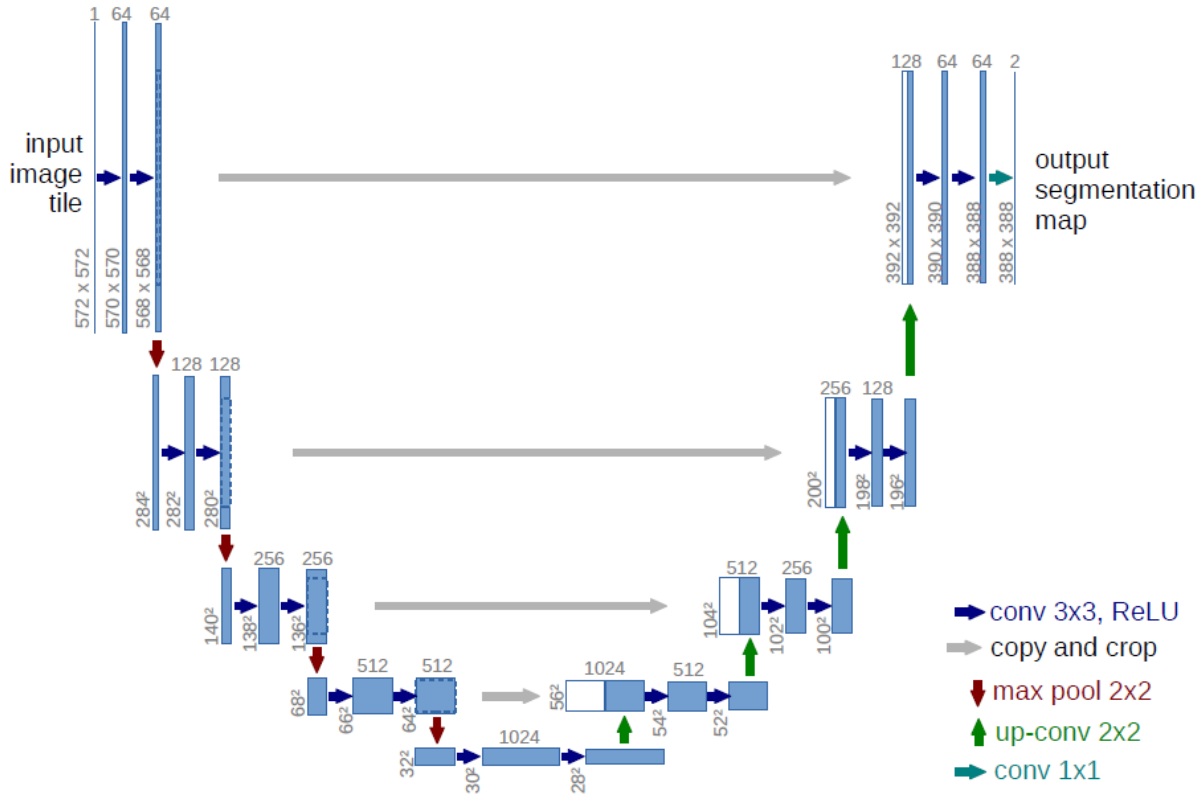


Fig. 1: Original U-Net Model: Source [3]

studies that have explored the Massachusetts Buildings Dataset and the Inria Aerial Image Labeling Dataset and identifies the gaps addressed by this research.

2.1 Foundations of U-Net in Image Segmentation

U-Net was originally developed by the Computer Science Department of the University of Freiburg for biomedical

image segmentation tasks involving limited training data [3]. The model builds upon the Fully Convolutional Network (FCN) architecture [4], but introduces a symmetric encoder–decoder structure with skip connections. These connections preserve high-resolution spatial information that is lost during down-sampling and facilitate more accurate localization of target features. An image of the original U-Net is shown in Figure 1.

The encoder path of the U-Net progressively reduces the spatial dimensions while extracting deep contextual features. Conversely, the decoder restores the spatial resolution through up sampling and integrates information from the encoder via skip connections. Owing to its effective architectural design, U-Net has been widely adopted in satellite and aerial imagery segmentation, where the ability to delineate fine boundaries is crucial. Its success in building segmentation is further supported by its robustness, simplicity, and strong performance, even with limited training data [5] [6] [7].

2.2 Applications of U-Net in Aerial and Satellite Imagery

Multiple studies have validated the applicability of U-Net in the domain of automatic building detection from aerial and satellite imagery. For instance, a U-Net-based model achieved a remarkable F1 score of 0.93 and an accuracy of 97.67% on the Joanópolis dataset in Brazil, which comprises over 7,500 manually annotated building instances [5]. In another study focusing on Casablanca, binary masking using U-Net reached a 90% F1 score after 25 training epochs [7]. Similarly, U-Net achieved 60% accuracy after 20 epochs using binary cross-entropy loss function and the Adam algorithm for optimization in detecting buildings from satellite images [8].

Further supporting its generalizability, U-Net models applied to Worldview satellite imagery for urban mapping tasks have demonstrated segmentation accuracies exceeding 86% for buildings and 83% for land-use classification [5]. In Jakarta, aerial photographs segmented using U-Net yielded training and testing accuracies of 0.83 and 0.87, respectively, highlighting their practical value for municipal planning and urban infrastructure monitoring [9].

2.3 U-Net Enhancements for Improved Performance

Numerous enhancements have been proposed for the base U-Net architecture to address challenges such as vanishing gradients, spatial resolution loss, and class imbalance. For example, incorporating batch-normalization layers improved model convergence and segmentation accuracy in the DeepGlobe SpaceNet challenge [10]. Siamese U-Net models, which utilize shared weights for dual input streams (e.g., original and down sampled images), were shown to improve segmentation of large-scale structures by learning complementary feature representations [11]

More advanced U-Net variants include Residual-Inception U-Net and the Attention Residual U-Net. These models achieved Intersection over Union (IoU) scores of approximately 0.62-

0.63 and F1 scores close to 0.76 on the Massachusetts Buildings Dataset, demonstrating their effectiveness for detailed urban segmentation tasks [12]. Another notable model, BuildFormer, integrates a transformer-based encoder-decoder structure and achieves a state-of-the-art intersection over union (IoU) of 0.7574 [13]. Similarly, the EfficientNet-based U-Net++, which employs compound scaling and nested skip connections, reported an impressive IoU of 0.8832 [14]. Recently, another transformer-based method, known as Super Token Vision Transformer with multi-scale (SMC) convolution, obtained a reported IoU of 73.38% on the same dataset [15].

DeepLabv3+ has demonstrated competitive performance for aerial building segmentation. On the Inria Aerial Image Labeling Dataset, it achieved an accuracy of 96.77% when benchmarked against other deep learning approaches [16]. In a subsequent study, an improved DeepLabv3+ model further achieved a precision of 91.54%, confirming the effectiveness of advanced contextual feature aggregation techniques in remote sensing applications [17].

2.4 Attention Mechanisms and ASPP in U-Net Variants

The integration of attention and multi-scale context modules has proven effective in various segmentation tasks. Chowdhury and Mirzaei (2025) combined Attention U-Net with Atrous Spatial Pyramid Pooling (ASPP), leading to significant improvements in Dice and IoU metrics for brain tumor segmentation [18]. Similarly, SAR-U-Net incorporated ASPP, residual connections, and squeeze-and-excitation (SE) blocks for liver segmentation, outperforming traditional U-Net models on the LiTS17 and SLiver07 datasets [19].

The use of Coordinate Attention with ASPP in a 2024 study further enhanced fine boundary detection in brain tumor segmentation [20]. In the field of remote sensing, SA-U-Net increased land cover classification accuracy from 93.3% to 96.2% on Landsat-8 imagery using spatial attention and ASPP [21]. RA-SE-ASPP-Net validated both the independent and joint contributions of attention and ASPP through ablation studies conducted in biomedical segmentation tasks [22], showing that combining spatial attention with multi-scale context is particularly beneficial in high-resolution or class-imbalanced segmentation problems.

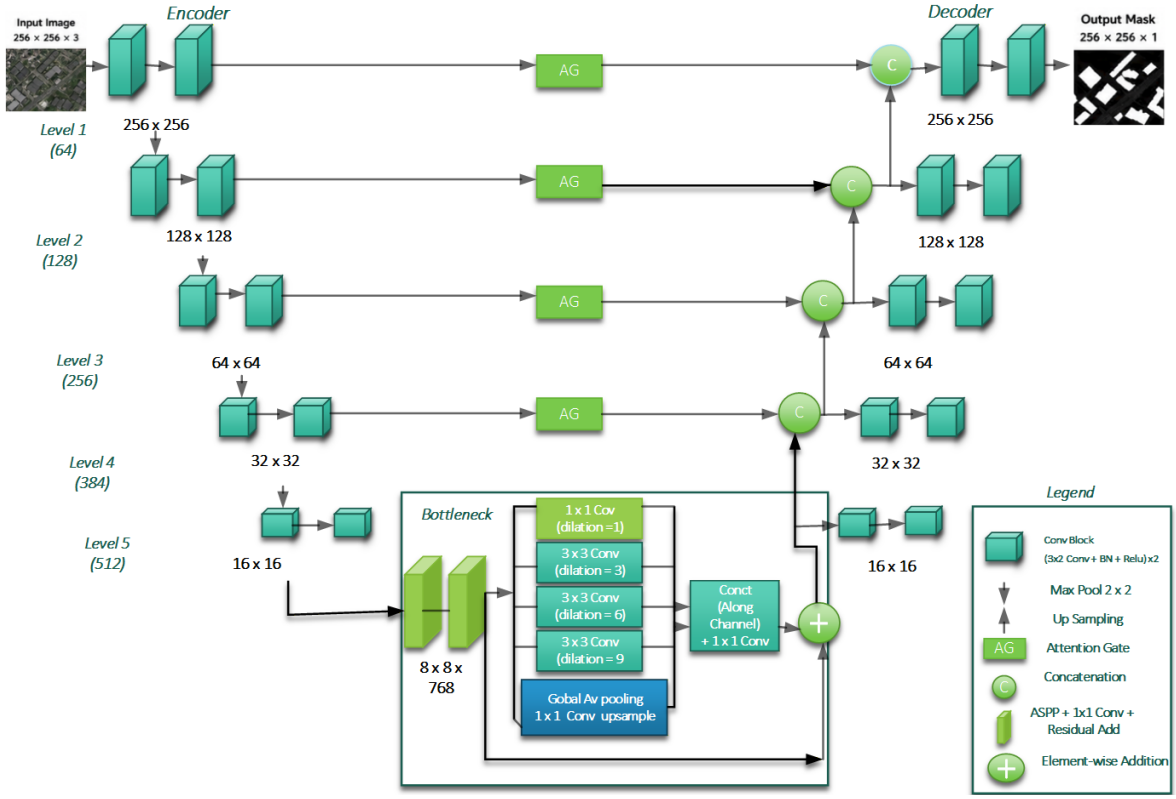


Fig. 2: Enhanced U-Net Architecture

2.5 Problem Statement and Research Gap

Accurate extraction of building footprints from high-resolution aerial imagery remains a challenging computer vision problem due to several inherent factors. Firstly, buildings in urban environments exhibit considerable variability in shape, size, texture, and orientation, making it difficult for segmentation models to generalize across diverse architectural styles. Small and densely clustered rooftops are particularly susceptible to misclassification due to their limited spatial footprint and low contrast against surrounding structures.

Secondly, the highly imbalanced nature of aerial imagery, where background pixels far outnumber building pixels, causes conventional pixel-wise loss functions to bias predictions toward the majority class, often leading to under-segmentation of small buildings and boundary imprecision. While advanced loss formulations such as Dice, Tversky, and hybrid combinations have been proposed, their effectiveness within deep encoder–decoder architectures for aerial segmentation have not been thoroughly compared under standardized conditions.

Thirdly, although several state-of-the-art architectures, including U-Net++, DeepLabv3+, HRNet, and transformer-based models, achieve competitive segmentation performance, they involve substantial computational overhead and typically require high-end GPUs to train effectively on large-scale datasets. Even with modern accelerators such as the NVIDIA A100 GPU, training full-resolution imagery (≥1500 × 1500 pixels) poses challenges related to memory consumption, batch-size limitations, gradient stability, and runtime efficiency.

Finally, despite the widespread use of patch-based training to address memory constraints and enhance local feature learning, prior work seldom investigates how patch extraction strategies interact with architectural enhancements such as deeper encoders, Attention Gates, and multi-scale dilated convolutions. The combined effect of these mechanisms on the segmentation accuracy of complex urban scenes is therefore not well understood.

Consequently, there is a need for a systematic study that evaluates the individual and combined contributions of architectural and loss-function enhancements within a unified framework, using consistent patch-based training, fixed input resolutions, and standardized evaluation metrics. Addressing this gap is essential for developing efficient and accurate models for building extraction from high-resolution aerial imagery.

3. ENHANCED U-NET ARCHITECTURE

The proposed Enhanced U-Net builds upon the classical U-Net framework while integrating a set of targeted architectural refinements designed to improve segmentation performance on high-resolution aerial imagery. These modifications address challenges such as small building structures, class imbalance, and the loss of multi-scale contextual information. The overall architecture is shown in Figure 2.

3.1 Deepened Encoder

The model architecture included an additional encoding block compared to the standard U-Net, as illustrated in Figure 2. While the original U-Net comprises four encoder blocks, the deepened version introduces a fifth block, thereby increasing the network's capacity to extract high-level semantic features. This deeper architecture is especially useful for segmenting complex and high-resolution images, such as 256×256 patches extracted from 1500×1500 and 515 x 512 patches extracted from 5000 x 5000 corresponding to the Massachusetts and Inria datasets, respectively. This enables the model to learn richer and more abstract feature representations, which are crucial for detecting small and irregular objects, such as rooftops, in urban environments.

3.2 Convolutional Block

In convolutional neural networks, intermediate representations are stored as tensors, which are multi-dimensional arrays that encode spatial and semantic information. Each tensor is denoted by H×W×C, where H and W represent the height and width of the feature map, respectively, and C denotes the number of channels generated by convolutional filters, as seen in Figure 3. This notation is widely used in deep learning literature to describe feature maps flowing through encoder-decoder architectures such as U-Net and its variants [23], [3]. For example, an input RGB image of size 256×256×3 is transformed into higher-dimensional feature tensors such as 256×256×64 after the first convolution block. As the network progresses through pooling layers, the spatial dimensions decrease while the number of channels increases, enabling the model to capture progressively richer semantic information [23]. This tensor representation is also used within the attention gates and ASPP module, where multiple feature tensors are projected, combined, and refined before being passed to subsequent layers [24].

Batch Normalization (BN) layers are inserted after each convolution operation throughout the network. BN normalizes intermediate feature distributions, thereby stabilizing gradient updates and accelerating convergence. Furthermore, BN

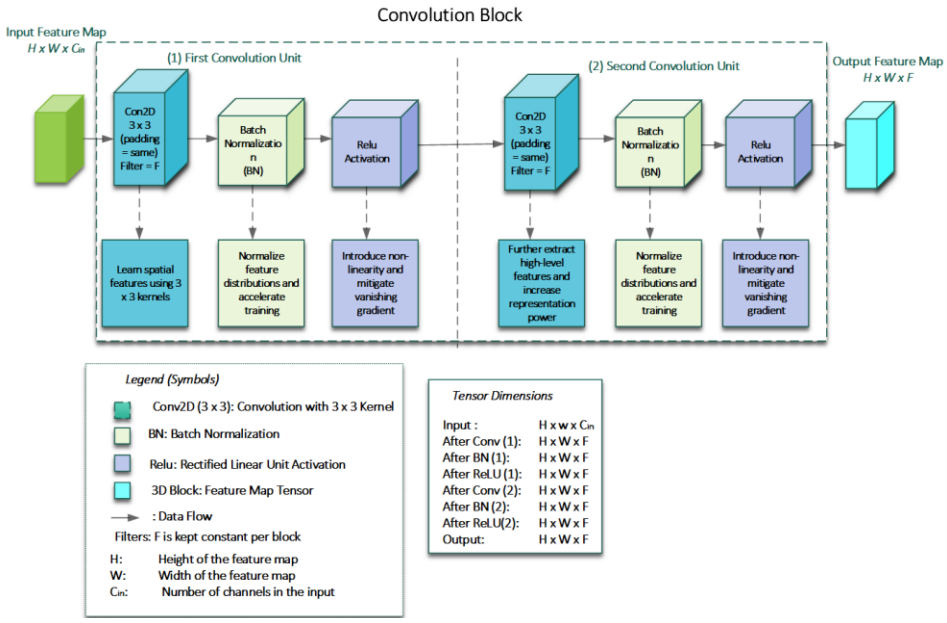


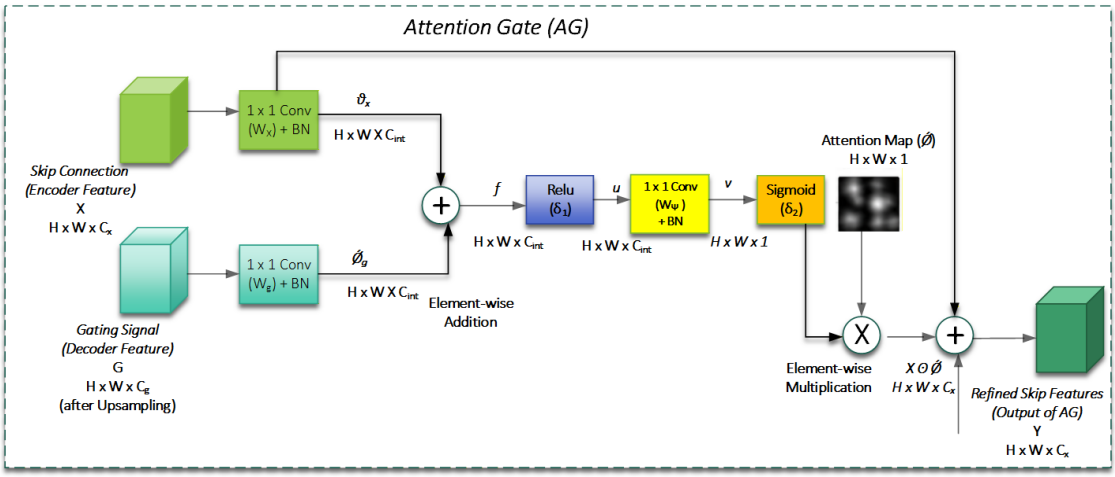
Fig. 3: Convolution Block

provides an implicit regularization effect, reducing overfitting and improving generalization without relying heavily on dropout. This leads to smoother training dynamics and more robust feature representations.

3.3 Attention Gates (AG)

The Attention Gate (AG) is integrated into the skip connections of the U-Net architecture to selectively emphasize relevant building features while suppressing irrelevant background information before feature fusion, as shown in Figure 4. The gate receives two inputs: the encoder feature map (skip connection) and the decoder feature map (gating signal). Both inputs are first projected using 1×1 convolutions and batch normalization, then combined and passed through a ReLU activation followed by another 1×1 convolution and a sigmoid activation to generate an attention coefficient map. This coefficient map is multiplied element-wise with the original encoder feature map to produce a refined feature representation that highlights salient regions. By filtering skip features in this manner, the Attention Gate improves feature localization and reduces false positives, leading to more accurate segmentation boundaries and enhanced overall performance

3.4 Atrous Spatial Pyramid Pooling



Notation:

H, W: Spatial height and width of the feature map
C_x: Number of channels in skip connection (encoder feature
C_g: Number of channels in gating signal (decoder feature)
C_{int}: Number of intermediate channels (typically $C_{int} = C_x/2$)
Ø: Attention map (values in [0, 1])
Y: Refined skip feature used in decoder
Θ Element-wise multiplication
The residual addition (x Θ Ø + x) preserves original information and stabilizes training

Fig. 4: Attentions Gate

(ASPP)

To strengthen the network's ability to capture multi-scale contextual information, an Atrous Spatial Pyramid Pooling (ASPP) module is positioned at the bottleneck.

Figure 5 illustrates the ASPP block employed in the proposed Enhanced U-Net architecture. The ASPP module is designed to capture multi-scale contextual information by applying parallel atrous (dilated) convolutions with different dilation rates to the same input feature map. In this study, dilation rates of 3, 6, and 9 were found to provide the best performance as shown in the experimental results.

The ASPP block consists of five branches: (1) a 1 x 1 convolution branch to capture local information, (2 – 4) three 3 x 3 atrous convolution branches with dilations rates of 3, 6, and 9 to capture medium-to-large scale contextual features, and (5) a global average pooling branch to encode image-level context. Each branch is followed by Batch Normalization (BN) and ReLU activation to improve training stability and feature representation. The output of all branches reduces the number of channels and enables effective feature integration. Finally, a residual connection is used to add the ASPP output to the input feature map, which facilitates gradient flow and preserves low-level information

3.5 Loss Function

To improve building segmentation performance, this study employed a hybrid loss function that combines Binary Cross-Entropy (BCE) and Dice losses. The hybrid formulation was designed to simultaneously optimize pixel-wise classification accuracy and region-level overlap between the predicted segmentation mask and the ground truth.

Binary Cross-Entropy loss evaluates segmentation performance at the pixel level by penalizing incorrect pixel classifications. It is defined as:

$$L_{BCE} = \frac{1}{N}\sum_{i=1}^{N}\{y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\} \quad (1)$$

where:

- y_i represents the ground truth label,
- p_i denotes the predicted probability,
- N is the total number of pixels.

Although BCE provides stable optimization and improves precision, it may struggle with class imbalance, commonly observed in aerial image segmentation tasks where background pixels dominate building pixels.

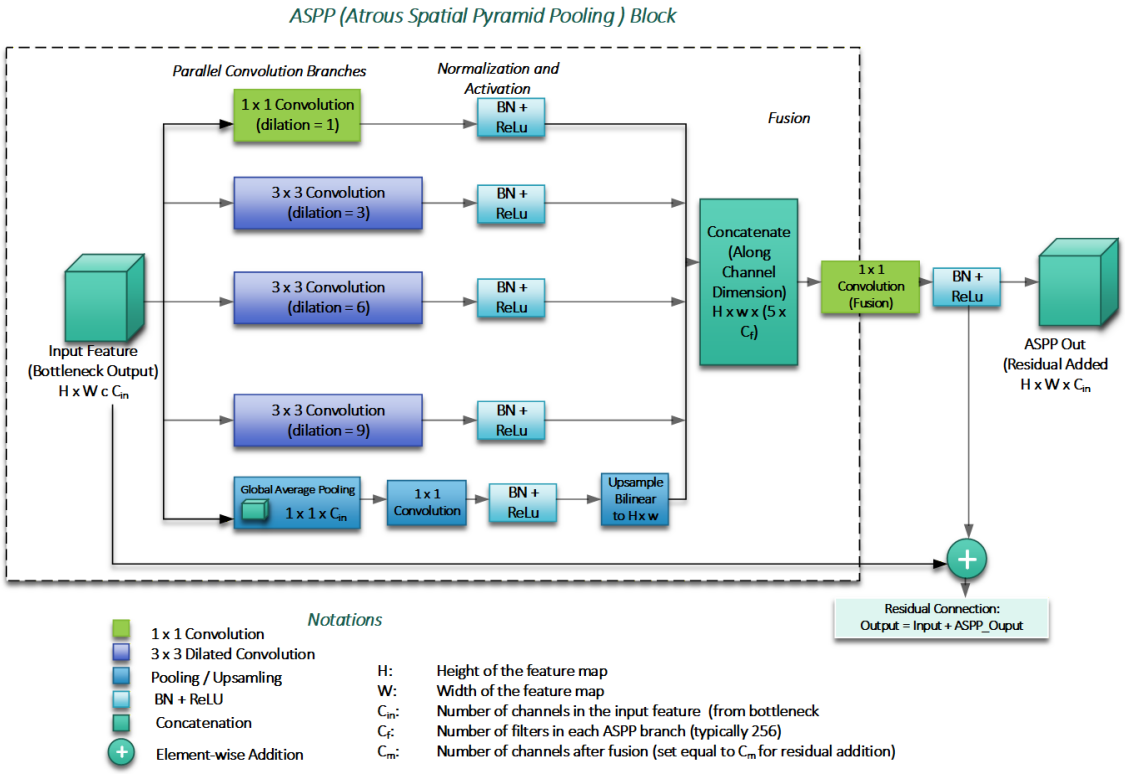


Fig. 5: ASPP Block

To address this limitation, Dice loss was incorporated to optimize spatial overlap between predicted and target regions. Dice loss is derived from the Dice Similarity Coefficient (DSC) and is expressed as

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i\, y_i + \epsilon}{\sum_{i=}^{N} p_i + \sum_{i=1}^{N} + \epsilon} \qquad (2)$$

Where:

- € is the small constant used for numerical stability. Dice loss improves segmentation consistency by directly maximizing region overlap, which is particularly important for accurately delineating building boundaries.

The final hybrid loss function used in the proposed Enhanced U-Net model combines both losses as follows:

$$L_{Hybrid} = 0.7L_{BCE} + 1.3L_{Dice} \qquad (3)$$

The weighting coefficients were empirically selected to balance precision and overlap optimization. Increasing the Dice contribution improved IoU and F1-score, while the BCE component maintained stable convergence and reduced false positive predictions.

Experimental results demonstrated that the hybrid BCE–Dice loss achieved superior segmentation performance compared to standalone BCE, Dice, and Tversky loss functions, producing the best balance among IoU, F1-score, Precision, Recall, and Accuracy on both the Massachusetts Buildings and Inria aerial image datasets

4. EXPERIMENTAL SETUP

This section describes the experimental design used to evaluate the proposed enhanced U-Net architecture. It includes an overview of the dataset, the preprocessing and augmentation procedures, the training configuration, the evaluation metrics, and the computational environment. Figure 6 presents the complete workflow of the experimental pipeline.

4.1 Patch-Based Training

Due to the high spatial resolution of aerial imagery, patch-based training was adopted to reduce GPU memory requirements and increase the number of training samples. For the Massachusetts Buildings Dataset (1500 × 1500 pixels), images were divided into overlapping patches of 256 × 256 pixels using a stride of 128 pixels. For the Inria Aerial Image Labeling Dataset (5000 × 5000 pixels), larger patches of 512 × 512 pixels with a stride of 256 pixels were used to preserve broader contextual information. A 50% overlap was employed in both datasets to mitigate boundary artifacts and improve segmentation continuity.

4.2 Massachusetts Building Dataset

The Massachusetts Buildings Dataset is a publicly available benchmark dataset widely used for building extraction and semantic segmentation in high-resolution aerial imagery. The dataset comprises 151 RGB aerial images captured over the Boston metropolitan area in the United States. Each image has a spatial resolution of 1500 × 1500 pixels and covers

approximately 2.25 km², resulting in a total mapped area of nearly 340 km² [25].

The dataset is divided into three subsets consisting of 137 training images, 4 validation images, and 10 testing images. Corresponding binary ground-truth masks are provided for all images, where building pixels are assigned a value of 1 and background pixels a value of 0.

The patch extraction process produced approximately 46,656 training patches, 5,832 validation patches, and 5,832 internal testing patches. The official unlabeled test images were used exclusively for qualitative visual assessment of model predictions and were not included in the quantitative evaluation.

The substantial variability in scene content, building morphology, and geographic location makes the Inria dataset particularly suitable for evaluating the robustness and generalization capability of deep learning models for building segmentation. Consequently, it serves as a complementary benchmark to the Massachusetts Buildings Dataset, enabling a comprehensive assessment of model performance across different urban environments

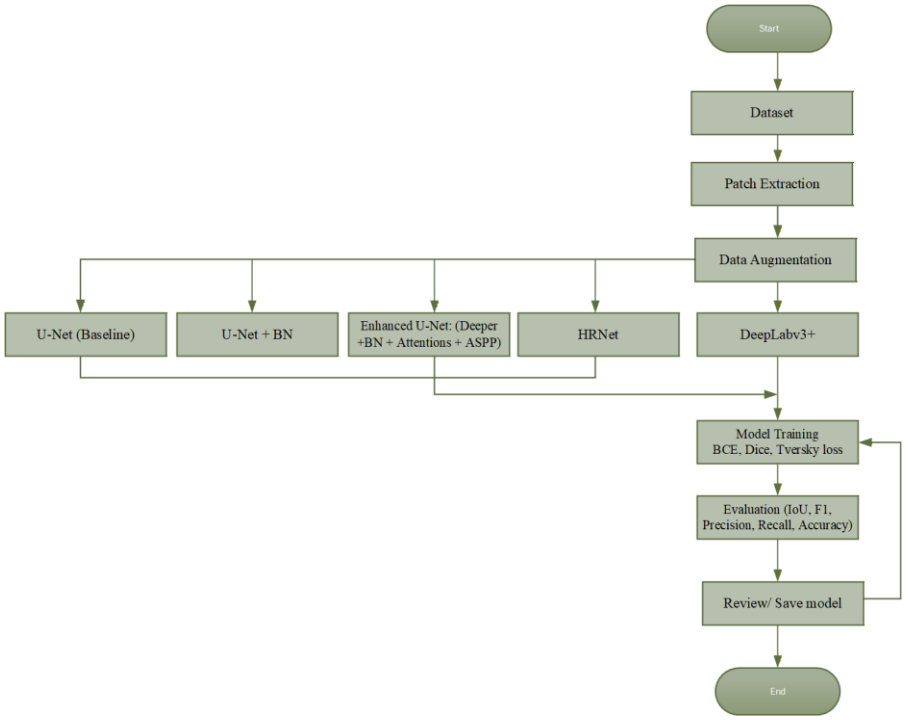


Fig. 6: Workflow Diagram

The patch extraction process generated approximately 13,700 training patches, 1,000 validation patches, and 400 testing patches. The resulting increase in sample diversity improved model generalization and enabled efficient training of deep learning architectures on high-resolution aerial imagery.

4.3 Inria Aerial Image Labeling Dataset

The Inria Aerial Image Labeling Dataset is a widely used benchmark dataset for semantic segmentation and building extraction in very high-resolution aerial imagery. The dataset contains 360 orthorectified RGB image tiles acquired over urban areas in the United States and Austria, including locations such as Austin, Chicago, Kitsap County, Tyrol, Vienna, and San Francisco. Each image tile has dimensions of 5000 × 5000 pixels with a ground sampling distance of 0.3 m per pixel, covering diverse urban environments characterized by varying building densities, architectural styles, vegetation cover, and land-use patterns [26]

The original dataset consists of 180 labeled training images accompanied by binary building masks and 180 unlabeled test images. Since ground-truth masks for the official test set are not publicly available in the Kaggle distribution used in this study, the labeled training set was randomly partitioned into 80% training (144 images), 10% validation (18 images), and 10% testing (18 images), following common practice in supervised learning studies [27].

4.4 Preprocessing and Data Augmentation

To improve model generalization and reduce overfitting, real-time augmentation was applied to each image/mask pair using Keras' ImageDataGenerator. The following transformations were used:

- Rotation: Random rotations within a range of ±10 degrees.
- Translation: Horizontal and vertical shifts up to 5% of the image width/height.
- Zooming: Random zoom within a range of ±10%.
- Flipping: Random horizontal flips applied with equal probability.
- Mask Preprocessing: After augmentation, binary masks were thresholded using a custom function to ensure values remained strictly 0 or 1.

This augmentation pipeline introduces geometric and photometric variability, enhancing robustness to real-world

scenarios such as illumination shifts, occlusions, and diverse building shapes.

4.5 Training Configuration

Model development and training were implemented in TensorFlow/Keras. The input to the network consisted of 256×256×3 patches for the Massachusetts building dataset and 512 x512 for the Inria Label Dataset. Training was performed with a batch size of 4.

The enhanced U-Net was trained for 50 epochs using the Adam optimizer, chosen for its adaptive learning rate behavior and stable gradient updates in deep convolutional architectures. The hybrid Binary Cross-Entropy (BCE) + Dice loss function was used to balance pixel-wise classification with overlap-based region accuracy.

4.6 Evaluation Metrics

Model performance was assessed using five standard metrics widely adopted in semantic segmentation tasks:

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

$$\text{F1 Score} = \frac{\text{2 x Precision x Recall}}{\text{Precision +Recall}} \quad (6)$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (7)$$

Where:

- TP = True Positives (correctly predicted building pixels)
- FP = False Positives (non-building pixels wrongly predicted as building)
- FN = False Negatives (building pixels wrongly predicted as non-building)
- TN = True Negatives (correctly predicted non-building pixels)

4.7 Computation Environment

All experiments were executed in Google Colab Pro, equipped with an NVIDIA A100 GPU (40 GB VRAM), 83.5 GB of system RAM, and 235.7 GB of storage. The A100's large memory capacity enabled efficient handling of high-resolution patches and real-time data augmentation, ensuring stable training for the deeper U-Net architecture with ASPP and Attention Gates.

5. RESULTS AND DISCUSSION

This section presents the quantitative results of the experiments. The performance of the enhanced U-Net model was compared with that of the baseline U-Net and other model variants. Tables and figures visualize the results and discuss the significance of the findings, highlighting the contribution of each architectural enhancement.

5.1 Progressive Evaluation of Model Enhancement on the Massachusetts Building Dataset

Table 1 summarizes the quantitative test performance of the different U-Net model variants. The experiments progressively integrate architectural and functional enhancements to the baseline U-Net, including Batch Normalization (BN), deeper encoder layers, Attention Gates (AG), and Atrous Spatial Pyramid Pooling (ASPP). Various loss functions, Binary Cross-Entropy (BCE), Dice, Tversky, and a hybrid BCE + Dice formulation, were also investigated to assess their influence on segmentation performance.

A threshold optimization strategy was employed on the validation dataset to determine the optimal segmentation threshold maximizing the IoU metric. The optimized threshold was subsequently applied during the final evaluation on the independent test dataset.

The baseline U-Net achieved an IoU of 0.667 and an F1-score of 0.801 using BCE loss. Introducing Batch Normalization significantly improved performance, increasing the IoU to 0.729 and the F1-score to 0.843, demonstrating improved training stability and feature generalization.

The Deep U-Net + BN architecture further improved segmentation performance, achieving an IoU of 0.734 and an F1-score of 0.847, indicating that deeper feature extraction enhanced contextual representation. However, the integration of Attention Gates alone did not produce additional improvement, suggesting that attention refinement without sufficient multi-scale context was less effective.

The ASPP module improved performance by capturing multi-scale contextual information through dilated convolutions. Among the evaluated dilation configurations, the rates of 3, 6, and 9 produced the best results, outperforming larger dilation settings such as 6, 12, and 18. This indicates that moderate dilation rates better preserve local spatial continuity and building boundaries.

The influence of different loss functions was also investigated. Dice loss improved overlap-based metrics, while Tversky loss achieved the highest precision (0.888) but significantly reduced recall, indicating overly conservative predictions. The hybrid Dice–Tversky formulation provided a better balance between precision and recall.

The proposed Enhanced U-Net employing the hybrid BCE–Dice loss achieved the best overall performance with an IoU of 0.737, F1-score of 0.848, precision of 0.848, recall of 0.849, and accuracy of 0.944. Furthermore, the proposed model outperformed the comparative HRNet and DeepLabv3+ architectures, demonstrating the effectiveness of combining deep feature extraction, lightweight ASPP, attention-guided refinement, and hybrid BCE–Dice optimization for aerial building segmentation.

Table 1: Test Result from Model Enhancement

Model Variant	IoU	F1 Score	Precision	Recall	Accuracy	Loss	Dilation
U-Net (baseline)	0.667	0.801	0.796	0.805	0.926	BCE	
U-Net (baseline) BN	0.729	0.843	0.829	0.857	0.941	BCE	
Deep U-Net + BN	0.734	0.847	0.836	0.858	0.943	BCE	
Deep U-Net + BN + AG	0.728	0.842	0.832	0.861	0.941	BCE	
Deep U-Net + BN + ASPP	0.729	0.843	0.832	0.854	0.941	BCE	6,12,18
Deep U-Net + BN + ASPP	0.731	0.845	0.829	0.861	0.941	BCE	3,6,9
Deep U-Net + BN + ASPP	0.727	0.842	0.828	0.856	0.940	BCE	4,8,12
Deep U-Net + BN + ASPP	0.720	0.837	0.821	0.853	0.938	BCE	2,4,6
Deep U-Net + BN + AG + ASPP	0.734	0.847	0.829	**0.865**	0.942	BCE	3,6,9
Deep U-Net + BN + AG + ASPP	0.736	0.848	0.837	0.859	0.943	Dice	3,6,9
Deep U-Net + BN + AG + ASPP	0.709	0.830	**0.888**	0.779	0.941	Tversky	3,6,9
Deep U-Net + BN + AG + ASPP	0.735	0.847	0.848	0.846	0.943	Dice+Tversky	3,6,9
Enhanced U-Net	**0.737**	**0.848**	0.848	0.849	**0.944**	BCE+Dice	3,6,9
HRNet	0.684	0.813	0.776	0.852	0.927	BCE+Dice	
DeepLabv3+	0.691	0.817	0.811	0.824	0.932	BCE+Dice	

Table 2: Different Dilations Evaluation on Test Set

Model Variant	IoU	F1 Score	Precision	Recall	Accuracy	Loss	Dilation
Deep U-Net + BN + ASPP	0.729	0.843	0.832	0.854	0.941	BCE	6,12,18
Deep U-Net + BN + ASPP	0.731	0.845	0.829	0.861	0.941	BCE	3,6,9
Deep U-Net + BN + ASPP	0.727	0.842	0.828	0.856	0.940	BCE	4,8,12
Deep U-Net + BN + ASPP	0.720	0.837	0.821	0.853	0.938	BCE	2,4,6

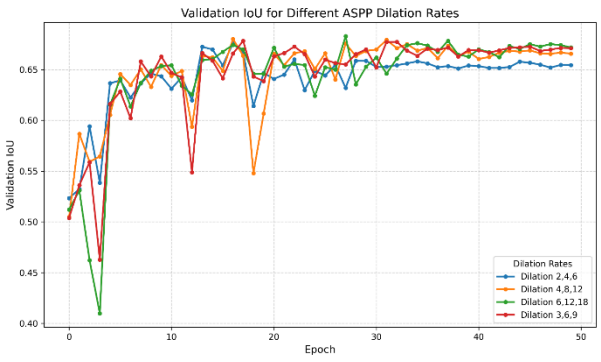


Fig. 7: Validation IoU for Different ASPP Dilation Rates

Figure 7 illustrates the influence of different ASPP dilation-rate configurations on segmentation performance. All configurations demonstrated stable convergence after approximately 20 training epochs. Larger dilation rates improved contextual feature extraction compared to smaller dilation settings such as 2,4,6. Although the 6,12,18 configuration achieved the highest validation IoU during training, the 3,6,9 configuration produced the best test performance, as illustrated in Table 2. This indicates that moderate dilation rates provided a better balance between local feature preservation and multi-scale contextual aggregation, leading to improved generalization on unseen aerial imagery. Consequently, the dilation rates of 3,6,9 were adopted in the final Enhanced U-Net architecture.

The validation results demonstrate that the proposed Enhanced U-Net achieved the most stable convergence and the best overall segmentation performance on the Massachusetts Buildings Dataset. Compared with the baseline U-Net, the integration of Batch Normalization, deeper feature extraction, Attention Gates, ASPP, and hybrid BCE–Dice optimization significantly improved validation IoU, F1-score, precision, and accuracy throughout training, as illustrated in Figure 8, 9, 10, and 11, respectively.

The Enhanced U-Net consistently maintained superior convergence stability after approximately 20 epochs and

achieved the highest overall validation performance among the evaluated architectures. The ASPP module effectively improved multi-scale contextual learning, while the hybrid BCE–Dice loss enhanced the balance between precision and recall. In contrast, HRNet and DeepLabv3+ exhibited comparatively lower validation performance and less stable convergence behavior.

Overall, the validation curves confirm that the proposed Enhanced U-Net provided improved feature representation, stronger generalization capability, and more accurate building boundary delineation for aerial building segmentation

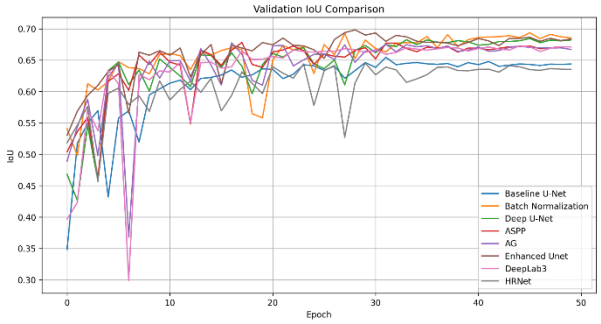


Fig. 8: Validation IoU Comparison

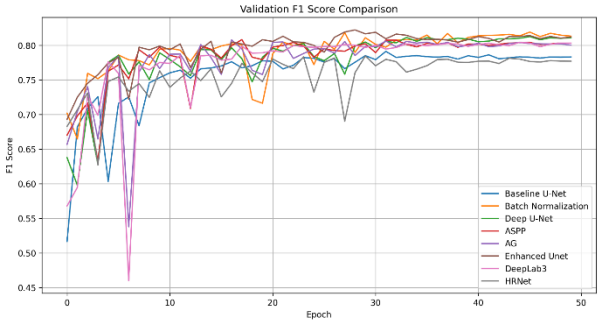


Fig. 9: Validation F1 Score Comparison

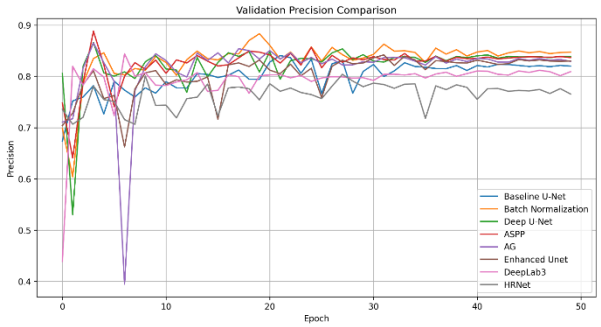


Fig. 10: Validation F1 Score Comparison

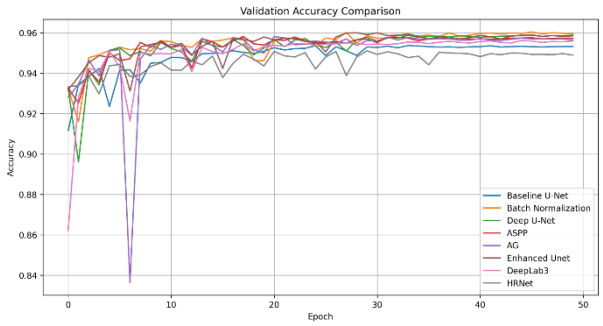


Fig. 11: Validation Precision Comparison

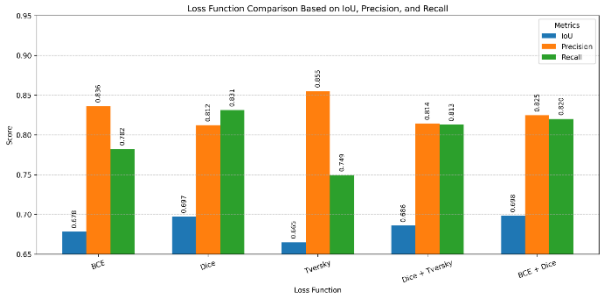


Fig. 12: Loss Function Comparison Based on IoU, Precision and Recall

Figure 12 presents the comparative performance of different loss functions evaluated on the proposed Enhanced U-Net architecture. The comparison was conducted using IoU, Precision, and Recall metrics derived from the validation results.

The results show that the Tversky loss achieved the highest precision (0.855), indicating strong suppression of false positives. However, this came at the expense of recall (0.749) and IoU (0.665), suggesting that the model became overly conservative and failed to detect several building regions.

Dice loss improved recall (0.831) and IoU (0.697) by directly optimizing spatial overlap between predicted and ground-truth masks. The hybrid Dice + Tversky loss produced a more balanced performance, achieving stable precision and recall simultaneously.

Among all evaluated configurations, the proposed BCE + Dice hybrid loss achieved the best overall balance with the highest IoU (0.698) while maintaining strong precision (0.825) and recall (0.820). This demonstrates that combining pixel-wise Binary Cross-Entropy optimization with region-overlap Dice optimization improved segmentation robustness and boundary delineation performance.

Overall, the results indicate that the BCE + Dice hybrid loss provided the most stable and balanced optimization strategy for aerial building segmentation

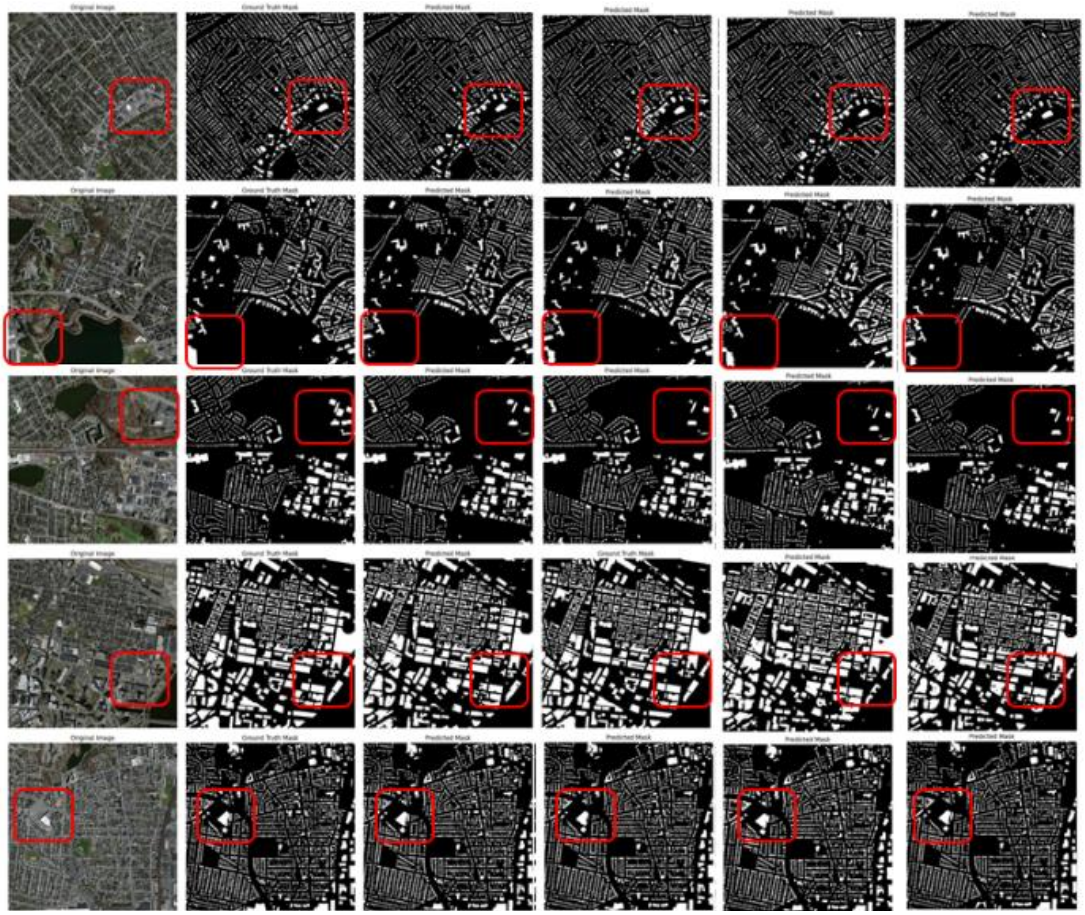


Fig. 13: Qualitative comparison of building segmentation results on representative test images from the Massachusetts Buildings Dataset. From left to right: input aerial image, ground-truth mask, predictions generated by the Enhanced U-Net, DeepLabv3+, HRNet, and U-Net + BN. Red bounding boxes highlight regions of interest. Rows 1-2 illustrate densely built urban areas, Rows 3-4 show regions affected by occlusions and complex structures, and Row 5 corresponds to sparsely populated residential zones

Figure 13 presents a qualitative comparison between the proposed Enhanced U-Net and three benchmark segmentation models on representative test samples. In densely populated urban regions (Rows 1-2), the Enhanced U-Net produces more continuous building footprints with fewer fragmented predictions compared to DeepLabv3+ and HRNet. In scenes affected by occlusions and complex background structures (Rows 3-4), the attention-guided architecture improves boundary delineation and suppresses false positives. In sparsely populated areas (Row 5), the proposed model better preserves small and isolated building structures, whereas competing models tend to miss thin rooftops or merge nearby objects. These visual observations are consistent with the quantitative improvements reported in Section 5.1, particularly in terms of IoU and recall.

5.2 Evaluation of Models on the Inria Dataset

The experimental results obtained on the Inria Aerial Labeling Dataset further validate the effectiveness and generalization capability of the proposed Enhanced U-Net architecture, as shown in Table 3. Using a patch size of 512 x 512 with stride of 256, the proposed model consistently outperformed the comparative architectures across all evaluation metrics.

Table 3: Inria Image Labeling Data Test Results from Model Variants

Model Varian	IoU	F1 Score	Precision	Recall	Accuracy	Loss
U-Net + BN	0.792	0.884	0.884	**0.884**	0.953	DICE + BCE
HRNet	0.744	0.853	0.861	0.846	0.941	DICE + BCE
DeepLab v3+	0.770	0.870	0.879	0.861	0.948	DICE + BCE
Enhanced U-Net	**0.799**	**0.888**	**0.893**	0.883	**0.955**	DICE + BCE

The Enhanced U-Net achieved the best overall performance with an IoU of 0.799, F1-score of 0.888, precision of 0.893, recall of 0.883, and accuracy of 0.955 using the hybrid Dice + BCE loss function. These results demonstrate the effectiveness of combining deep feature extraction, Batch Normalization, ASPP, Attention Gates, and hybrid overlap-aware optimization for large-scale aerial building segmentation.

Compared with the standard U-Net + BN model, the proposed architecture achieved improved precision and IoU, indicating better building boundary delineation and stronger suppression of false positives. The Enhanced U-Net also outperformed both HRNet and DeepLabv3+, confirming its superior capability in capturing multi-scale contextual information and preserving spatial consistency in complex urban scenes. Figures 14-17 illustrate the validation F1-score, precision, accuracy, and IoU curves, respectively, for all evaluated models. These curves show that the proposed Enhanced U-Net converged more consistently and achieved superior performance across all evaluation metrics.

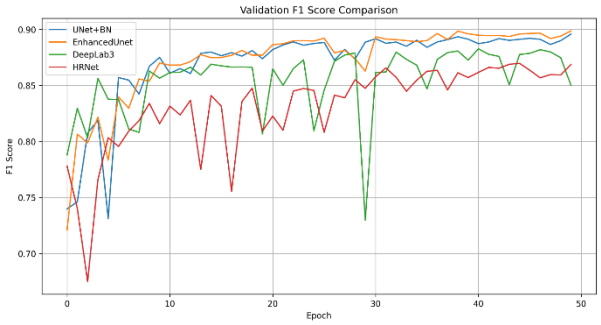


Fig. 14: Validation F1 Score Comparison

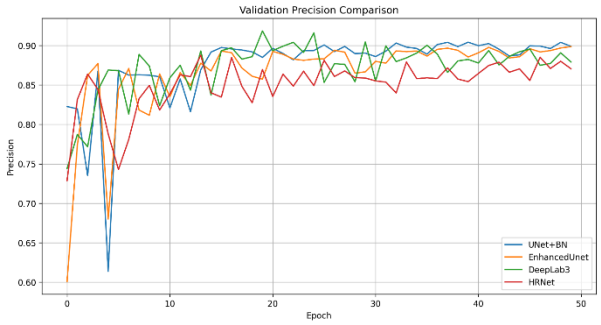


Fig. 15: Validation Precision Comparison

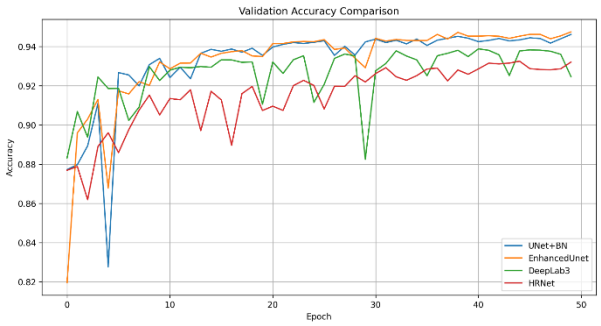


Fig. 16: Validation Accuracy Comparison

Fig. 17: Validation IoU Comparison

5.3 Parameters

Table 4 presents the architectural complexity of the evaluated deep learning models in terms of trainable parameters, non-trainable parameters, and memory footprint. HRNet is the most lightweight architecture, containing only 766,657 trainable parameters and requiring approximately 2.94 MB of storage, making it computationally efficient for deployment in resource-constrained environments. DeepLabV3 exhibits moderate complexity with 6.5 million trainable parameters and a model size of 24.89 MB. In contrast, the U-Net architecture contains over 31 million trainable parameters, resulting in a larger memory footprint of 119.79 MB. The proposed Enhanced U-Net has the highest complexity, with 36.18 million trainable parameters and a model size of 138.01 MB. This increase in model capacity is attributed to the incorporation of a deeper encoder, Batch Normalization layers, Attention Gates, and the Atrous Spatial Pyramid Pooling (ASPP) module. Although the Enhanced U-Net requires more computational resources, the additional parameters enable richer feature extraction and multi-scale contextual learning, contributing to the improved segmentation performance observed in the experimental results.

Table 4: Parameters

Model	Trainable Parameters	Non-Trainable parameters	Total Size (MB)
DeepLapV3	6,504,481	21,472	24.89
HRNet	766,657	2,944	2.94
U-Net	31,390,721	11,776	119.79
Enhanced U-Net	36,179,622	20,096	138.01

The results indicate a trade-off between model complexity and segmentation accuracy. While HRNet offers the smallest memory footprint and computational cost, its reduced parameter count limits its representational capacity. Conversely, the Enhanced U-Net leverages a larger number of trainable parameters to capture more discriminative spatial and contextual features, resulting in superior segmentation performance on both the Massachusetts Buildings and Inria datasets. These findings suggest that the increase in model

complexity is justified by the corresponding gains in building extraction accuracy.

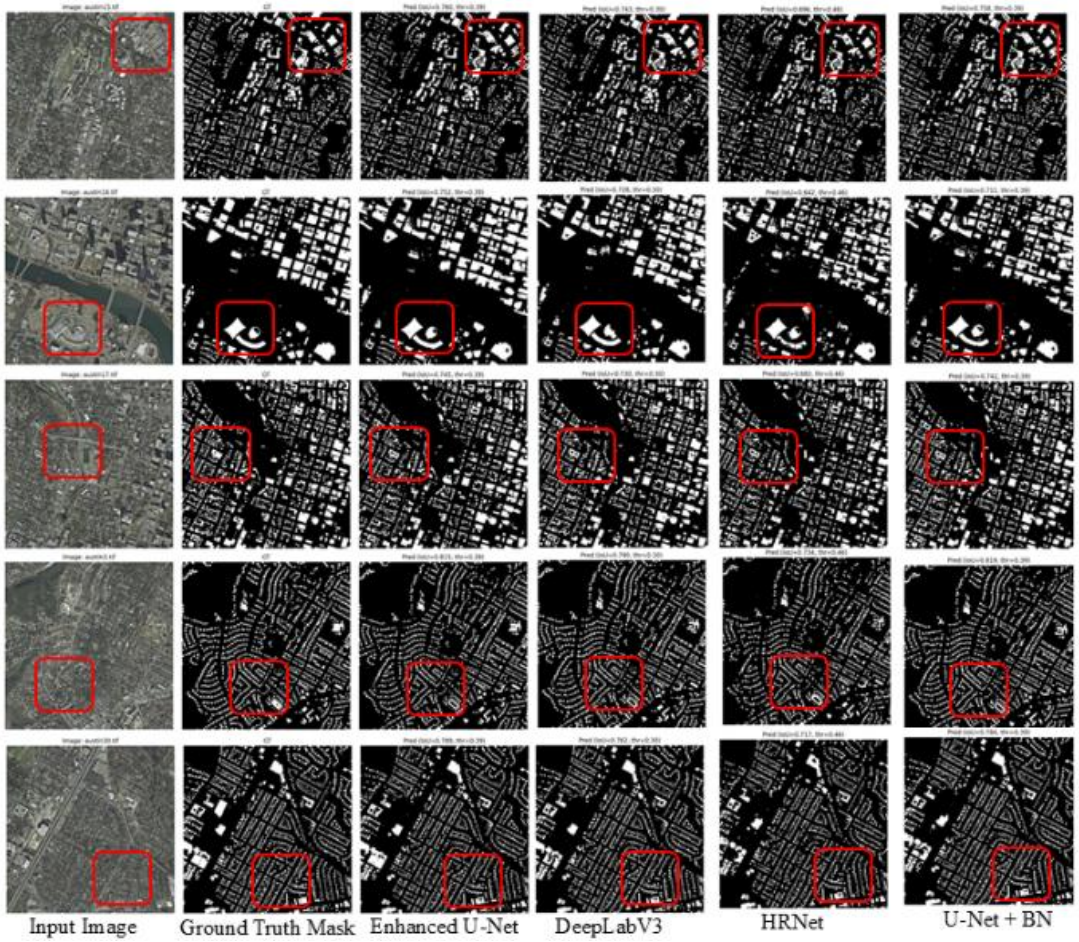


Fig. 18: Qualitative comparison of building segmentation results on representative test images from the Inria Image Label Dataset. From left to right: input aerial image, ground-truth mask, predictions generated by the Enhanced U-Net, DeepLabv3+, HRNet, and U-Net + BN. Red bounding boxes highlight regions of interest. Rows 4-5 illustrate densely built urban areas, Rows 1 and 3 show regions affected by occlusions and complex structures, and Row 2 corresponds to sparsely populated residential zones

6. CONCLUSION AND FUTURE WORK

This section presents the conclusions drawn from the study and identifies key directions for future research. Based on the experimental results and comparative analysis, the principal findings regarding the performance of the proposed model and the effectiveness of the architectural enhancements are summarized. In addition, potential research directions are discussed to further improve model generalization, robustness, and applicability for building segmentation across diverse aerial imagery datasets.

6.1 Conclusion

This study presented an enhanced U-Net architecture for aerial building segmentation by integrating Batch Normalization (BN), a deeper encoder–decoder network, Attention Gates (AG), Atrous Spatial Pyramid Pooling (ASPP), and a hybrid Binary Cross-Entropy (BCE)–Dice loss function. The proposed model was evaluated on the Massachusetts Buildings Dataset using 256 × 256 image patches and further validated on the INRIA Aerial Image Labeling Dataset using 512 × 512 image patches. Comparative experiments demonstrated that the proposed Enhanced U-Net outperformed U-Net, DeepLabv3+, and HRNet, achieving the best performance on both datasets. These findings demonstrate that the integration of spatial attention and multi-scale contextual feature extraction significantly improves building segmentation accuracy and generalization, making the proposed architecture a robust solution for automated building footprint extraction from high-resolution aerial imagery

6.2 Future Work

Although the proposed Enhanced U-Net demonstrated strong performance on both the Massachusetts Buildings Dataset and the INRIA Aerial Image Labeling Dataset, several opportunities remain for further research.

6.2.1 Evaluation on Additional Benchmark Datasets

Future studies should evaluate the proposed architecture on additional publicly available building segmentation datasets, such as the SpaceNet Building Footprints Dataset and the DeepGlobe Building Dataset. Evaluation across multiple datasets would provide a more comprehensive assessment of the model's robustness, scalability, and generalization capability under varying imaging conditions, urban morphologies, and spatial resolutions.

6.2.2 Development of a Local Building Segmentation Dataset

An important direction for future research is the development of a high-resolution building segmentation dataset for Buea, Cameroon. Such a dataset would enable the evaluation of the proposed model in a real-world African urban environment characterized by diverse building structures, vegetation cover, shadows, and informal settlements. The resulting dataset could also serve as a valuable benchmark for future research and support applications in urban planning, land administration, disaster risk management, and smart city development.

7. DECLARATION

7.1 Funding

The authors declare that no external funding was received for this research.

7.2 Code Availability

The source code supporting the findings of this study is publicly available in a GitHub repository. The repository contains the implementation of the proposed Enhanced U-Net architecture, data preprocessing scripts, model training procedures, and evaluation code required to reproduce the experimental results. The repository is available at: https://github.com/joejoedat-collab/enhanced-unet-building-segmentation [28]

8. REFERENCES

[1] Ayala et al., "A Deep Learning Approach to an Enhanced Building Footprint and Road Detection in High-Resolution Satellite Imagery," Remote Sensing, vol. 13, no. 16, 2021. .

[2] Wagner et al., "U-Net-Id, an Instance Segmentation Model for Building Extraction from Satellite Images—Case Study in the Joanópolis City, Brazil," Remote Sensing, vol. 12, no. 10, 2020.

[3] R. Olaf, F. Philipp and B. Thomas, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Computer Vision and Pattern Recognition, vol. 9351, p. 234–241, 2015.

[4] Shelhamer et al, "Fully Convolutional Networks for Semantic Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, 2016.

[5] Zhuokun et al., "Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net," Remote Sensing, vol. 12, no. 1574, pp. 2-17, 2000.

[6] W. Alsabhan, T. Alotaiby and B. Dudin, "Detecting Buildings and Nonbuildings from Satellite Images Using U-Net," Hindawi, vol. 2022, pp. 1-13, 2022.

[7] Najjaj et al., "Deep Learning Approach for Uban Mapping," in The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Gottingen, 2022.

[8] E. Wang and D. Wang, "Using U-Net to Detect Buildings in Satellite Images," Journal of Computer and Communication, vol. 10, pp. 132-138., 2022.

[9] E. Irwansyah, Y. Heryadi and A. A. S. Gunawan, "Semantic Image Segmentation for Building Detection in Urban Area with Aerial Photograph Image using U-Net Models," IEEE Explore, pp. 48-51, 2020.

[10] Weijia et al., "Semantic Segmentation-Based Building Footprint Extraction Using Very High-Resolution Satellite Images and Multi-Source GIS Data," Remote Sensing, vol. 11, no. 4, 2019.

[11] J. Shunping, W. Shiqing and L. Meng, "Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set," IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, vol. 57, no. 1, pp. 574-586, 2018.

[12] S. Batuhan and D. Z. Seker, "A Residual-Inception U-Net (RIU-Net) Approach and Comparisons with U-Shaped CNN and Transformer Models for Building Segmentation from High-Resolution Satellite Images," Senors, vol. 22, no. 19, pp. 1-20, 2022.

[13] Wang et al., "Building extraction with vision transformer," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, no. 5625711, pp. 1-11, 2022.

[14] Wang et al., "Performance Analysis of Various EfficientNet Based U-Net++ Architecture for Automatic Building Extraction from High Resolution Satellite Images," Springe, vol. 376, pp. 385-399, 2024.

[15] Zhou et al., "Enhanced Building Extraction via STMC-UNet: Integrating Super Token Transformer and Multi-scale Convolution," Signal, Image and Video Processing, vol. 19, no. 1019, 2025.

[16] S. Ekiz and U. Acar, "Improving building extraction from high-resolution aerial images: Error correction and performance enhancement using deep learning on the Inria dataset," Sage Journals, vol. 108(I), pp. 1-22, 2025.

[17] Hui et al., "An improved DeepLabv3+ lightweight network for remote-sensing image semantic segmentation," Springer: Complex & Intelligent Systems, vol. 10, pp. 2839-2849, 2024.

[18] R. C. Satyaki and G. Mirzaei, "Hybridization of Attention UNet with Repeated Atrous Spatial Pyramid Pooling for Improved Brain Tumour Segmentation," in 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), Houston, TX, USA, 2025.

[19] Wang et al., "SAR-U-Net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver segmentation in Computed Tomography," Elsevier, vol. 208, 2021.

[20] Wang et al., "Improved Unet model for brain tumor image segmentation based on ASPP-coordinate attention mechanism," in 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE, Wenzhou, China, 2024.

[21] Xiangsuo et al., "Improved U-Net Remote Sensing Classification Algorithm Fusing Attention and Multiscale Features," Remote Sensing, vol. 14, no. 15, pp. 1-24, 2022.

[22] Jovial et al., "Attention-Guided Residual U-Net with SE Connection and ASPP for Watershed-Based Cell Segmentation in Microscopy Images," in National Library of Medeicine, 2025.

[23] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press , 2016.

[24] Ozan et al., "Attention U-Net: Learning Where to Look for the Pancreas," Conrnell University, 2018.

[25] B. Ashwath, "Kaggle," 2020. [Online]. Available: https://www.kaggle.com/datasets/balraj98/massachusetts-buildings-dataset. [Accessed 15 November 2024].

[26] Maggiori et al., "The Inria Aerial Image Labeling Benchmark," in IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Texas, United States, 2017.

[27] R. Sagar, "Inria Aerial Image Labeling Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/sagar100rathod/inria-aerial-image-labeling-dataset. [Accessed 20 May 2026].

[28] N. Joseph, "Enhanced U-Net for Building Segmentation," GitHub, 12 26 2026. [Online]. Available: https://github.com/joejoedat-collab/enhanced-unet-building-segmentation. [Accessed 26 January 2026].