

BEAU-Net: Boundary And Edge Aware U-Net For Improved Building Segmentation

Akil Ahmad Taki^{1*}, Md. Rifat Ur Rahman², Mohammed Imamul Hassan Bhuiyan³, Md Sultan Mahmud⁴

^{1,4}Department of EEE, University of Asia Pacific, Dhaka, Bangladesh

^{2,3}Department of EEE, Bangladesh University of Engineering and Technology, Dhaka Bangladesh

¹<https://orcid.org/0009-0004-6966-5956>; ² <http://orcid.org/0009-0002-0241-0962>,

³<https://orcid.org/0009-0009-6767-9357>, ⁴<https://orcid.org/0000-0002-1651-9806>

*Corresponding Author: Akil Ahmad Taki

Abstract— Extracting building footprints automatically from high-resolution remote sensing imagery has emerged as a focal area of active investigation. Over the past few years, deep learning techniques, especially convolutional neural networks (CNNs), have delivered encouraging outcomes in this domain. Nevertheless, existing leading-edge architectures continue to face difficulties when dealing with building boundaries, abrupt edge transitions, and intricate geometric configurations. To tackle these issues, we introduce Boundary and Edge Aware U-Net (BEAU-Net), an attention-driven segmentation framework built upon the UNet backbone. The BAM component concentrates on features containing abundant boundary cues. This strengthens the network's ability to separate buildings from neighboring structures, which proves especially valuable in dense urban settings where buildings are closely spaced. The EAM component accentuates fine-grained features including precise edges and elaborate geometric patterns. This allows the network to represent intricate building geometries and identify sudden shifts in edge characteristics. Comprehensive experiments were carried out on the Massachusetts Building, WHU Building, and INRIA Aerial Image Labeling datasets to confirm the efficacy of our proposed framework. The findings reveal notable gains in both IoU and F1 metrics, particularly in scenarios involving tight boundaries and complex structural forms.

Keywords— Building footprint extraction, U-Net architecture, Remote sensing imagery, High-resolution satellite data, Boundary delineation, Edge recognition, Boundary attention module (BAM), Edge attention module (EAM).

I. INTRODUCTION

Owing to swift progress in earth observation disciplines, aerospace technology, geographic information systems, and sensor engineering, the availability of high-resolution remote sensing imagery has grown substantially. Applications such as urban planning, demographic estimation, smart city construction, disaster response, and transportation management all demand automated extraction of structural information from high-resolution remote sensing images (HRSIs). Consequently, accurate retrieval of building information from such high-resolution imagery is of critical importance. The heterogeneity and complexity of building environments pose significant obstacles to automated extraction procedures. Structures situated adjacent to roads and public spaces that share comparable spatial and textural characteristics are prone to misclassification. Within HRSIs, vegetation canopy and shadow effects can conceal building details, thereby degrading extraction accuracy. Furthermore, buildings exhibiting varied and elaborate forms along with minimal inter-building gaps add further complexity to the extraction process. Automated building extraction thus continues to be a demanding undertaking. Numerous early studies attempted to devise distinctive feature representations. Various color and color-invariant feature spaces were employed to establish corner-based descriptors. \cite{4682675} employed entropy and color cues to differentiate buildings from trees, incorporating textural descriptors. Because these hand-engineered features lacked sufficient robustness and representational power, they were unable to handle the intricate relationships between buildings and their surrounding backgrounds.

In the contemporary era, deep learning methods, particularly convolutional neural networks (CNNs), have risen to prominence across a wide range of computer vision challenges due to their powerful feature extraction capabilities. CNNs efficiently encode spatial detail and local contextual relationships while progressively acquiring richer, higher-order semantic representations as depth increases. The Fully Convolutional Network (FCN) was originally conceived to reformulate general image segmentation as a per-pixel classification task. It adopts an encoder-decoder architectural paradigm. The encoder branch processes the input image to generate feature representations, whereas the decoder branch utilizes various strategies to produce semantic predictions.

The down sampling operations inherent in early FCN designs reduce spatial resolution and discard fine-grained spatial detail, which complicates the segmentation of small objects and precise boundary delineation. In an effort to preserve spatial detail, SegNet \cite{badrinarayanan2017segnet} introduced an unpooling mechanism within the decoder stage that leverages stored max-pooling indices. A further influential architecture developed for segmentation purposes is UNet \cite{ronneberger2015u}. Originally designed for biomedical image analysis, UNet has demonstrated broad applicability across diverse tasks, including the extraction of building footprints from remote sensing imagery. By merging both semantic and spatial features via skip connections, UNet achieves strong segmentation performance. Yet, remote sensing imagery frequently lacks the clarity found in medical images owing to complicated boundary patterns and abrupt edge variations. This shortcoming stems from the spatial information loss caused by UNet's downsampling layers, which hampers the discrimination of small and geometrically complex structures. To overcome these shortcomings, architectures such as UNet++ \cite{zhou2018unet++} and UNet3+ \cite{huang2020unet} \cite{li2024method} have been put forward. Through additional skip connections and supplementary decoder layers, these designs more effectively preserve spatial information and localize fine-grained details. While these models represent an advancement over the standard UNet, they continue to face difficulties in capturing complex boundary patterns and subtle edge features. Res-UNet \cite{diakogiannis2020resunet} attempts to further retain spatial details by using a backbone network like ResNet, enhancing the encoder's ability to capture spatial information. Nonetheless, detecting intricate boundaries remains problematic because of analogous constraints in the downsampling and upsampling processes. In a comparable fashion, multi-encoder U-Net \cite{el2024enhancing} strives to maintain spatial information by processing dual views and merging them, which helps address misclassification of smaller structures. However, when both encoders extract comparable high-level features while diverging in low-level details, the merging process may introduce contradictory information, resulting in indistinct boundaries.

A significant breakthrough arrived with the adoption of attention mechanisms, enabling models to selectively concentrate on salient regions within an image. Attention UNet \cite{yu2022ags}, for example, uses gated attention to highlight important foreground regions while suppressing background noise, which aids in capturing larger structures' geometries and improves edge detection compared to UNet or UNet++. Despite these gains, it still encounters challenges with recurring sharp edge patterns within compact spatial regions. Attention UNet++ \cite{zhao2022multiscale} extends this strategy by retaining spatial information more efficiently, although difficulties with complex geometries remain. The Convolutional Block Attention Module (CBAM) offers an alternative strategy by applying channel and spatial attention in a sequential manner.

CBAM improves feature selection across both channel and spatial dimensions, thereby boosting building localization precision. Nevertheless, its channel attention component, despite enhancing feature emphasis, can still encounter difficulties with highly elaborate boundary configurations. Additional efforts have explored integrating other attention strategies alongside CBAM \cite{lei2024dual}.

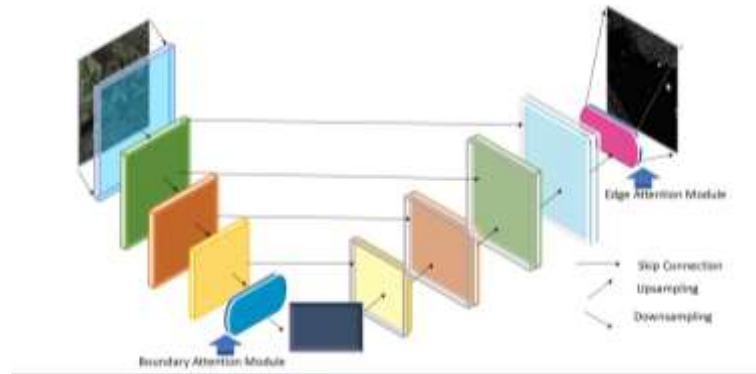
Attention-driven methods that specifically target object boundaries have also been explored in segmentation tasks. Boundary Attention-UNet \cite{wang2022boundary} was developed for medical image segmentation applications. When adapted for building segmentation, these networks successfully capture and distinguish tight boundaries but continue to struggle with complex building forms and occasionally miss small structures. So, to utilize the advantage of Boundary Attention Module (BAM), researchers have worked to combine it with components like multi-scale feature extraction modules \cite{sun2020bas} or combination of spatial and channel attention \cite{nong2021boundary} for the purpose of land cover classification. Networks incorporating BAM have likewise been introduced for land cover classification tasks \cite{li2021multitask}. A separate research direction centers on boundary-aware loss functions that seek to directly enhance boundary predictions. As an example, the Dice Boundary Loss merges the conventional Dice loss with a boundary-oriented penalty term, discouraging misclassifications near edges to improve accuracy. This strategy has proven effective in both medical and satellite image segmentation contexts \cite{milletari2016v} \cite{omati2024efficiency}. The Boundary IoU loss formulation encourages intersection over union specifically at object boundaries, improving edge detail preservation \cite{9157598}. The Contour Loss, supplying additional supervisory signals for object contours, has been investigated to refine segmentation edges \cite{chen2017rethinking}. However, this method may not transfer effectively to varied datasets or tasks, particularly when object boundaries are poorly defined or highly irregular.

Accurately capturing building edges represents a further challenge that contributes to segmentation inaccuracies. Multiple strategies have been proposed to tackle this issue. The Holistically-nested Edge Detection (HED) network \cite{xie2015holistically} enhances edge precision through deep supervision applied across multiple network stages, capturing boundaries at both coarse and fine scales. Nevertheless, HED is limited in its ability to manage complex textures and subtle boundary details. It additionally has difficulty with particular edge orientations. Richer Convolutional Features (RCF) \cite{liu2017richer} strengthens edge identification through the fusion of hierarchical features from multiple convolutional layers, performing well in tasks such as building segmentation from satellite imagery. Yet it faces challenges related to scale variability and inconsistent feature fusion, which may diminish edge sharpness along complex boundaries. Multi-task learning paradigms, including Edge-aware Networks (EANet) \cite{chen2021cgsanet}, perform edge detection and segmentation simultaneously, sharing a feature extractor for improved accuracy in both tasks. However, EANet faces generalization difficulties when deployed across different datasets.

More recently, investigators have examined models that go beyond UNet-based designs, including PSPNet \cite{zhao2017pyramid} and DeepLabV3+ \cite{chen2018encoder}, which harness pyramid attention mechanisms and multi-scale aggregation \cite{wu2024prior} approaches to strengthen feature representation across different spatial resolutions. PSPNet's Pyramid Pooling Module (PPM) encodes both global and local contextual information through multi-scale feature map pooling, enabling improved discrimination of small buildings in aerial imagery. However, PSPNet's

computational intensity can limit its efficiency in large-scale applications. DeepLabV3+ employs dilated convolutions in conjunction with the Atrous Spatial Pyramid Pooling (ASPP) module to balance broad contextual awareness with fine detail preservation. ASPP's ability to aggregate information across multiple dilation rates makes DeepLabV3+ effective for capturing complex boundaries and fine textures, though its higher computational demands can challenge scalability for real-time applications.

Fig.1 : Proposed BEAU-Net Architecture



The principal contributions of this work are summarized below.

- We integrate a Sobel filter-driven Edge Attention Module (EAM) within U-Net to identify building shapes and edge characteristics. This module is particularly effective at highlighting abrupt edges and delineating elaborate geometric forms.
- Additionally, we employ a Boundary Attention Module (BAM), drawing inspiration from land cover classification research, to target building boundaries and enable the model to differentiate buildings from adjacent regions.
- We merge BAM and EAM to capitalize on the complementary strengths of each module. To our knowledge, this represents the first integration of BAM and EAM specifically for building segmentation. This combination empowers the model to capture both boundary and shape cues simultaneously, especially in areas characterized by complex geometric configurations and tightly packed buildings.

The remainder of this article is organized as follows: Section II offers a thorough description of the proposed methodology. Section III presents the experimental configuration and evaluates our approach through detailed results. Finally, Section IV provides concluding remarks and discusses potential future directions.

II. METHODOLOGY

This section begins with an overview of the proposed network architecture, which is tailored to improve the semantic segmentation of buildings in remote sensing imagery. Following this overview, we present detailed descriptions of both the Boundary Attention Module (BAM) and the Edge Attention Module (EAM).

Network Architecture

As depicted in Fig. 1, the proposed network extends the U-Net architecture by incorporating boundary and edge attention mechanisms aimed at elevating segmentation performance. U-Net is broadly recognized for its effectiveness in pixel-level segmentation tasks, particularly within medical imaging and satellite image analysis. Our design introduces additional attention mechanisms to focus the network's learning on salient boundaries and edges, which are critical for precise segmentation. The network comprises three core components: an encoder, a bottleneck, and a decoder, organized symmetrically around the central bottleneck layer.

The encoder pathway gradually diminishes the spatial dimensions of the input while expanding the depth of the feature maps. The input initially traverses four convolutional blocks, each comprising two 3×3 convolution layers, succeeded by batch normalization and a ReLU activation. This procedure extracts a hierarchy of features from the input image. Following each block, a 2×2 max pooling operation halves the spatial resolution, enabling the network to acquire broader contextual information.

Let x represent the input image, and let f_{enc}^i denote the output of the i -th encoder block. Then, the encoder operations can be expressed as:

$$f^i\{enc\} = ConvBlock(f^{i-1}\{enc\}), \quad \text{where } i = 1, 2, 3, 4, \quad f^0\{enc\} = x$$

Here, each convolutional block is succeeded by a downsampling step via max pooling:

$$f^{i+1}\{enc\} = \{MaxPool\}(f^i\{enc\})$$

At the deepest encoder block, a Boundary Attention Module (BAM) is applied to improve object boundary detection. This module enriches feature representation by refining the features obtained at this level. A comprehensive description of this module appears in Section II-B.

Within the bottleneck, the network encodes the most abstract and globally informative features. A 3×3 convolutional block incorporating batch normalization and ReLU activation extracts high-level feature representations. This stage constitutes the deepest level of the U-Net, where spatial resolution reaches its minimum and feature map depth is at its maximum.

Before reaching the bottleneck, a Boundary Attention Module (BAM) is applied to the feature map f_{enc}^4 . Let $f_{enc-attn}^4$ represent the output after applying BAM:

$$f_{\{enc-attn\}}^4 = BAM\left(f_{\{enc\}}^4\right)$$

The resulting feature map is subsequently forwarded to the bottleneck for additional processing:

$$f_{\{bottom\}} = \{ConvBlock\}\left(\{MaxPool\}\left(f_{\{enc-attn\}}^4\right)\right)$$

The decoder pathway mirrors the encoder and progressively restores spatial resolution through successive up sampling of the feature maps. Each decoder stage involves a transposed convolution (up-convolution) that doubles the feature map resolution,

followed by concatenation with the matching encoder feature map . This enables the network to recover fine-grained details by reincorporating high-resolution features.

Let $\{f^i_{dec}\}$ denote the i -th decoder block. Upsampling step can be expressed as:

$$f^i_{dec} = \{UpConv\}(f^{\{i+1\}}_{dec}) \quad \{where\} i = 4, 3, 2, 1$$

Upon upsampling, the feature map is merged with the corresponding encoder output:

$$f^{\{i\}}_{cat} = \{Concat\}(f^{\{i\}}_{dec}, f^{\{i\}}_{enc})$$

Each merged feature map then passes through a convolutional block for feature refinement:

$$f^{\{i\}}_{refined} = \{ConvBlock\}(f^{\{i\}}_{cat})$$

This process progressively reconstructs the spatial dimensions of the input while fusing high-level and low-level features to achieve accurate segmentation.

Following the final decoder block, an edge attention mechanism is applied to further refine the feature maps, with particular emphasis on enhancing edge information. This module is instrumental in achieving precise segmentation of fine details and is described in depth in Section II-C.

After the final feature map undergoes edge attention refinement, it is processed through a 1×1 convolution to generate the segmentation map. This step reduces the feature map's depth to the desired number of output channels, corresponding to the number of classes for segmentation.

Let $\{f^1_{refined-attn}\}$ represent the output after applying EAM:

$$f^{\{1\}}_{\{refined-attn\}} = \{EAM\}(f^{\{1\}}_{refined})$$

The feature map $\{f^1_{refined-attn}\}$ is then passed through a 1×1 convolution to produce the segmentation map:

$$f_{out} = \{Conv\}(f^{\{1\}}_{\{refined-attn\}})$$

Finally, a sigmoid activation function is used to compute the probability map for each individual pixel :

$$y = \sigma(f_{out})$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Boundary Attention Module

The Boundary Attention Module (BAM) is engineered to strengthen boundary feature representation within the segmentation network. It functions by producing an attention map, depicted in Fig. 2, that accentuates boundary regions across the feature maps. This directs the model's focus toward these essential areas, thereby improving segmentation accuracy.

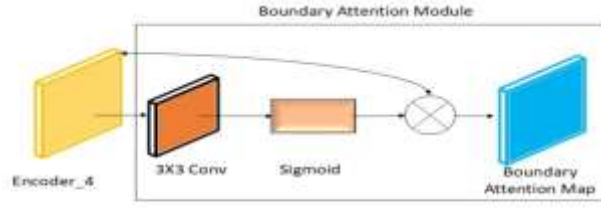


Fig. 2: Boundary Attention Module(BAM)

Let f_{enc} represent the input feature map to the BAM, the boundary map f_{bmap} is computed as:

$$f_{bmap} = \sigma(\text{Conv}(f_{enc}))$$

where $\text{Conv}(\cdot)$ denotes the 3×3 convolution operation, and $\sigma(\cdot)$ represents the sigmoid activation function. The resulting boundary attention map f_{bmap} is subsequently multiplied element-wise with the original input feature map f_{enc} , generating the final refined output f_{bam_out} :

$$f_{bam_out} = f_{enc} \odot f_{bmap}$$

Here, $\odot$ denotes element-wise multiplication. This refinement step amplifies boundary regions, enabling downstream network layers to concentrate more effectively on these pivotal areas for precise segmentation.

Edge Attention Module

The Edge Attention Module (EAM) is constructed to sharpen the model's focus on edges present in the feature maps, consequently improving segmentation precision along object boundaries. As illustrated in Fig. 3, this module utilizes Sobel filters to detect edges within the input feature maps, facilitating the identification of regions exhibiting significant gradient changes—which is vital for effective segmentation.

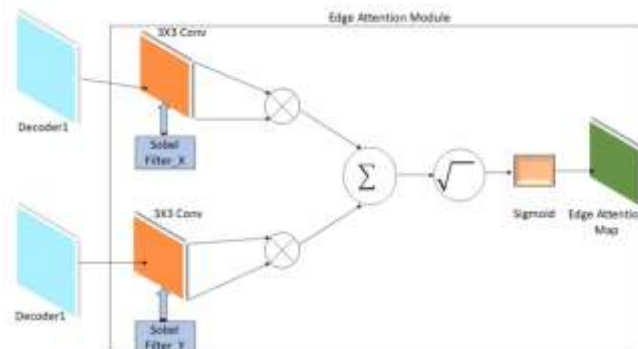


Fig. 3: Edge Attention Module(EAM)

Within the EAM, two distinct convolutional layers are used to apply Sobel filters along the horizontal and vertical axes. The Sobel filter in the x-direction $\{Sobel\}_x$ identifies horizontal edges, whereas the Sobel filter in the y-direction $\{Sobel\}_y$ detects vertical edges.

The Sobel filter in the x-direction is defined as:

$$\{Sobel\}_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

The Sobel filter in the y-direction is defined as:

$$\{Sobel\}_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

The incoming feature map f_{enc} is convolved with both Sobel kernels to derive the edge responses:

$$\begin{aligned} f_{edge_x} &= \{Sobel\}_x(f_{enc}) \\ f_{edge_y} &= \{Sobel\}_y(f_{enc}) \end{aligned}$$

The edge responses from both orientations are merged to form an edge map that highlights areas exhibiting substantial gradients:

$$f_{\mathrm{edge_map}} = \sqrt{f_{\mathrm{edge}_x}^2 + f_{\mathrm{edge}_y}^2}$$

A sigmoid activation function is then applied to the edge map to produce the edge attention map

$$f_{\mathrm{edge_attention}} = \sigma(f_{\mathrm{edge_map}})$$

where σ denotes the sigmoid function.

The edge attention map $f_{\mathrm{edge_attention}}$ is subsequently employed to refine the output feature map by prioritizing areas with strong edge responses, directing the model toward critical boundaries in the segmentation process. Through the incorporation of this module into the architecture, the network achieves improved differentiation between objects and background, yielding enhanced segmentation accuracy.

III. RESULT & ANALYSIS

Dataset

For our experiments, we employed the Massachusetts Building Dataset\cite{mnih2013machine}, WHU Building Dataset\cite{ji2018fully} and INRIA Aerial Image Labeling\cite{maggiori2017dataset} dataset. Table I summarizes the detailed configurations of these datasets.

Massachusetts Building Dataset

This dataset contains 151 high-resolution satellite images divided into training, testing, and validation subsets. Each image measures 1500×1500 pixels with a ground sampling distance of 1 meter per pixel. The imagery covers diverse urban, suburban, and rural regions across the United States. All images were randomly cropped to dimensions of 256×256 , since the original 1500×1500 pixel resolution contained sparse building information.

WHU Building Dataset

The WHU Building Dataset comprises 8189 images partitioned into training, validation, and testing subsets. The images are of size 512×512 pixels at a resolution of 0.075 meters per pixel, spanning approximately 450 km^2 . This dataset covers a variety of environments, encompassing both urban and suburban areas, and has been widely utilized for building detection and segmentation research. For our experiments, the images were resized to 256×256 pixels.

INRIA Aerial Image Labeling Dataset

This dataset provides 180 training images. The test set is deliberately withheld. Every image measures 5000×5000 pixels. The imagery was acquired at a spatial resolution of 0.3 meters per pixel. It comprises high-resolution aerial photographs spanning diverse urban locations across several cities in the United States and Austria. Accordingly, we partitioned the 180 images into training, testing, and validation subsets following a 70:20:10 ratio. The images were also randomly cropped to 256×256 pixels to enhance building detail.

All datasets underwent a thorough preprocessing pipeline as part of the experimental workflow. Data augmentation strategies were employed to promote model generalization across diverse scenarios.

TABLE I DATASET CONFIGURATION

Dataset	Train	Validation	Test	Size (pixel)
Massachusetts Building Dataset	137	10	4	256×256 (cropped)
WHU Building Dataset	5732	1228	1228	256×256 (resized)
INRIA Aerial Image Labeling Dataset	125	36	19	256×256 (cropped)

Loss Function

In this study, we employ a combined Dice Loss and Binary Cross-Entropy (BCE) loss to exploit the strong class imbalance mitigation properties of Dice Loss alongside the stable gradient characteristics of BCE loss. Given that buildings frequently occupy only a small fraction of an image, Dice Loss is especially suitable for managing this imbalance, while BCE Loss guarantees stable gradient flow throughout training. The total loss function $\mathcal{L}_{total}$ is formulated as follows:

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{BCE} + \beta \cdot \mathcal{L}_{Dice}$$

where $\mathcal{L}_{BCE}$ is the Binary Cross-Entropy loss, $\mathcal{L}_{Dice}$ is the Dice loss, and α and β are weighting coefficients that balance the contributions of the two losses. The Binary Cross-Entropy loss $\mathcal{L}_{BCE}$ is defined as:

$$\mathcal{L}_{BCE} = - (1/N) \sum (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

where N is the number of pixels, y_i represents the ground truth label for pixel i , and p_i is the predicted probability for pixel i .

The Dice loss $\mathcal{L}_{Dice}$ is a measure of overlap between the predicted segmentation and the ground truth. It is defined as:

$$\mathcal{L}_{Dice} = 1 - (2 \sum y_i p_i) / (\sum y_i + \sum p_i)$$

where y_i and p_i are the ground truth and predicted probability, respectively.

Evaluation Metrics

The segmentation model performance is assessed using several widely adopted metrics: Overall Accuracy (OA), Intersection over Union (IoU), Precision, Recall, and F1 Score.

Overall accuracy is formulated as:

$$OA = (TP + TN) / (TP + TN + FP + FN)$$

where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively.

IoU quantifies the overlap between predicted and ground truth segmentation masks:

$$IoU = TP / (TP + FP + FN)$$

Precision represents the ratio of true positives to all predicted positives:

$$Precision = TP / (TP + FP)$$

Recall captures the ratio of true positives to all actual positives:

$$Recall = TP / (TP + FN)$$

The F1 score computes the harmonic mean of precision and recall, expressed as:

$$F1 = (2 * Precision * Recall) / (Precision + Recall)$$

These metrics provide a comprehensive evaluation of the model's results, focusing on different aspects of prediction quality.

Implementation Details

The proposed model and all compared baseline architectures were implemented within the PyTorch deep learning framework, with data preprocessing handled through the Albumentations library. The model was trained in Kaggle's notebook environment, which utilizes an NVIDIA Tesla P100 GPU with 16 GB of memory.

Hyperparameters underwent multiple rounds of tuning, and the optimal results were achieved with the following configuration.

TABLE III HYPERPARAMETER TABLE

Hyperparameter	Fine Tuned Value
Optimizer	Adam
Batch Size (Massachusetts Dataset)	8
Batch Size (WHU Dataset)	32
Batch Size (INRIA Dataset)	8
Learning Rate	0.0001
Epoch (Massachusetts Dataset)	210
Epoch (WHU Dataset)	60
Epoch (INRIA Dataset)	240

Result Analysis

This section provides a comparative analysis of the performance metrics obtained by our proposed model against several leading architectures. We benchmark our model's performance against architectures including UNet \cite{ronneberger2015u}, Unet++ \cite{zhou2018unet++}, Attention-Unet \cite{yu2022ags}, Attention-Unet++ \cite{zhao2022multiscale}, PSPNet \cite{zhao2017pyramid}, and DeepLabV3+ \cite{chen2018encoder}.

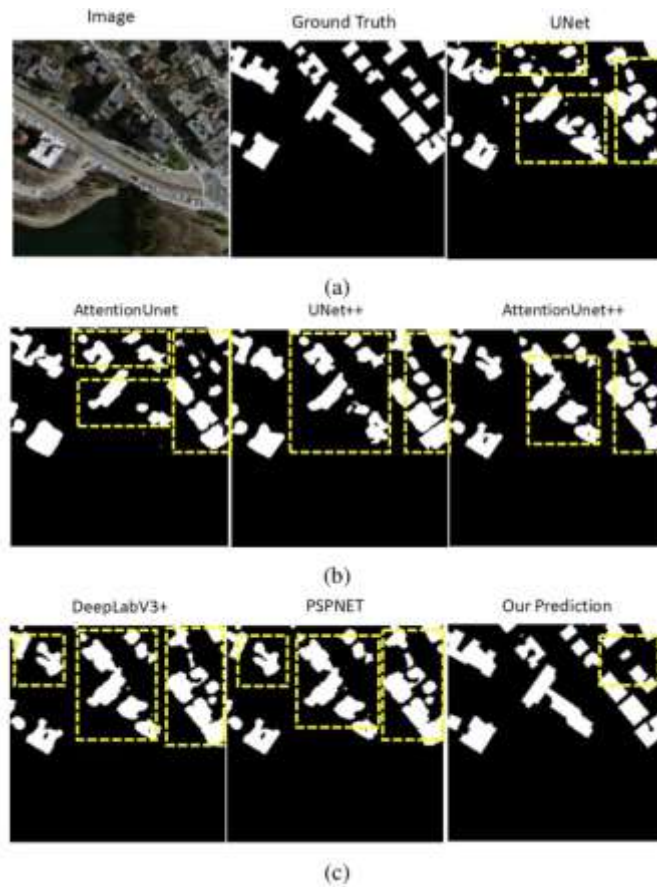
Massachusetts Building Dataset

Here, we present a comparison of our results on the Massachusetts Building Dataset against established baseline models. Our model achieves an OA of 93.5\%, IoU of 85.42\%, precision of 91\%, recall of 84\%, and F1 of 87.23\%, as shown in Table III. Apart from recall, it surpasses all other models across every evaluation metric. The incorporation of BAM and EAM reduces the count of erroneously predicted false negatives, yielding the highest IoU score of 85.42\%.

TABLE III PERFORMANCE METRICS FOR MASSACHUSETTS BUILDING DATASET

Model	OA	IoU	Precision	Recall	F1
UNet	88.74	57.94	87.85	94.47	90.08
Attention UNet	91.23	69.07	87.59	76.40	81.55
UNet++	91.48	67.91	82.00	75.00	78.73
Attention UNet++	91.98	81.84	92.41	80.03	86.09
PSPNet	92.10	75.33	84.00	74.50	78.97
DeepLabV3+	92.50	78.97	85.50	77.00	81.10
Ours	93.50	85.42	91.00	84.00	87.23

Fig.4: Comparison of our model with state of the art model for Massachusetts Building Dataset



Attention UNet++ does exhibit marginally higher precision than our model, as it suppresses false positives slightly more effectively. However, our model compensates for this through a stronger recall score driven by fewer false negatives, as reflected in the F1 metric. Moreover, our model outperforms multi-scale architectures such as DeepLabV3+ and PSPNet, which typically excel at identifying smaller buildings, as demonstrated by its leading OA of 93.5%. Overall, the integration of both BAM and EAM modules translates into consistently superior performance scores.

As illustrated in Fig. 4(a) and Fig. 4(b), our proposed model exhibits substantially reduced spatial detail loss relative to UNet and Attention-UNet. It also demonstrates superior detection of intricate boundary information, exceeding the performance of UNet, UNet++, Attention-UNet, and PSPNet, as shown in Fig. 4(c). Fig. 4(b) and Fig. 4(c) reveal that Attention-UNet++ and DeepLabV3+ have difficulty with sharp edges, whereas our model handles them effectively.

WHU Building Dataset

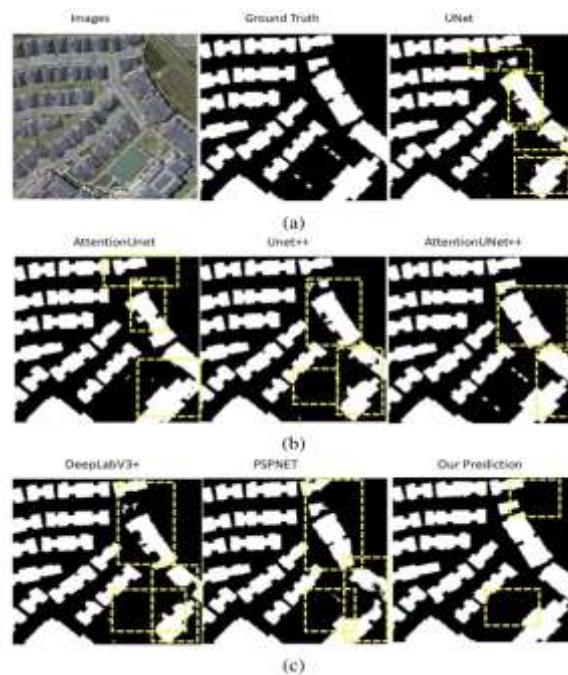
This subsection compares our results with leading architectures on the WHU Building Dataset. As shown in Table IV, our model surpasses all competing methods with a remarkable overall accuracy (OA) of 98.50%. It also achieves a leading IoU of 92.33%, demonstrating its effectiveness in classifying true positives, especially along building boundaries. Our model's F1 score of 94.80%, the highest among all models, reflects an excellent balance, even though our precision of 95.00% is slightly below the best precision of 95.20% from DeepLabV3+. This is counterbalanced by a robust recall of 93.00%. The elevated IoU demonstrates the model's success in reducing false negatives,

particularly in identifying sharper building edge transitions. Achieving a notable OA of 98.50%, our model exceeds the performance of other approaches, including multi-scale architectures like PSPNet, in terms of true positive classification. Collectively, the incorporation of attention modules substantially elevates overall model performance.

TABLE IV FONT SIZES FOR PAPERS

Model	OA	IoU	Precision	Recall	F1
UNet	97.20	83.51	87.85	94.47	90.89
Attention UNet	97.90	88.00	94.00	93.00	93.50
UNet++	98.13	88.45	94.65	93.20	93.85
Attention UNet++	98.02	87.69	93.96	92.32	93.12
PSPNet	98.20	89.00	95.00	93.50	94.00
DeepLabV3+	98.30	89.50	95.20	93.00	94.00
Our Model	98.50	92.33	95.00	93.00	94.80

Fig.5: Comparison of our model with state of the art model for WHU Building Dataset



In contrast to the Massachusetts Building Dataset, both our proposed model and the baseline architectures performed considerably better on the WHU Building Dataset, largely due to its substantially larger training set. Even amid these near-saturated results, our model continues to deliver measurable improvements. Attention-UNet and UNet++ both handle standard geometric shapes adequately, but they encounter difficulty with particularly complex geometric configurations, as shown in Fig. 5(a), where UNet in Fig. 5(a) incorrectly labeled several structures. PSPNet and DeepLabV3+ incorrectly classified the sharp edges, as seen in Fig. 5(c). Attention UNet++ was the architecture whose performance most closely approached ours, as evident from Fig. 5(b). However, unlike

our model, it struggled with very small and detailed structural forms. Notably, neither model successfully captured the smallest structures.

INRIA Aerial Image Labeling Dataset

In this subsection, we compare our results on the INRIA Aerial Image Labeling Dataset with those of established baseline architectures. As indicated in Table V, our model outperforms the other methods on the majority of evaluation metrics, with the highest Overall Accuracy (OA) of 93.5%, Intersection over Union (IoU) of 83.12%, recall of 84%, and F1 score of 87.34%. While Attention UNet++ achieves a marginally higher precision of 92.41%, our model offsets this through a stronger recall score, indicating better true positive detection. Our model's greatest IoU score of 83.12% results from the BAM and EAM modules' effectiveness in minimizing false negatives. We observe that our model delivers superior overall performance compared to multi-scale architectures like DeepLabV3+ and PSPNet, which are typically competent at detecting smaller buildings. The combination of BAM and EAM modules yields enhanced performance metrics across the board.

TABLE V FONT SIZES FOR PAPERS

Model	OA	IoU	Precision	Recall	F1
UNet	88.74	58.19	87.85	94.47	90.08
Attention UNet	91.23	66.19	87.59	76.40	81.55
UNet++	91.48	67.12	82.00	75.00	78.73
Attention UNet++	91.98	76.12	92.41	80.03	86.09
PSPNet	92.10	75.33	84.00	74.50	78.97
DeepLabV3+	92.50	78.97	85.50	77.00	81.10
Ours	93.50	83.12	91.00	84.00	87.34

Fig.6: Comparison of our model with state of the art model for INRIA Aerial Image Labeling Dataset

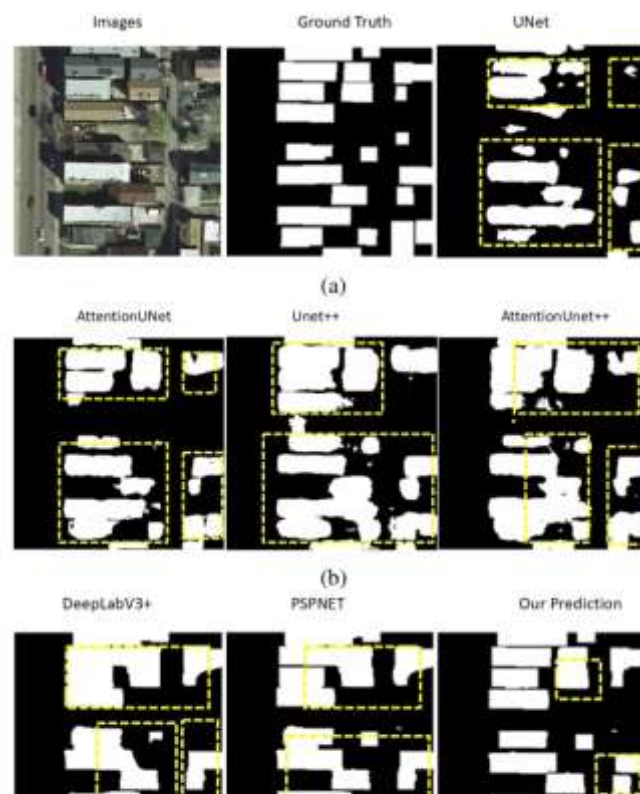


Fig. 6(a) shows that UNet exhibits notably poor performance. Foreground regions were substantially mislabeled as background. Both Attention-UNet and UNet++ struggled to identify building boundaries, as seen from Fig. 6(b). As shown in Fig. 6(b) and Fig. 6(c), boundary detection presents additional challenges for both Attention UNet++ and DeepLabV3+. While PSPNet in Fig. 6(c) performs well in boundary detection, it frequently misclassifies background regions as foreground. Because our model reliably identifies boundaries without frequent misclassification, it outperforms all compared architectures.

IV. ABLATION STUDY

This section illustrates the incremental improvement achieved by our model and highlights the contribution of each attention module. We demonstrate how the sequential inclusion of each module led to performance gains, emphasizing the impact of BAM and EAM on segmentation metrics including IoU and F1 score.

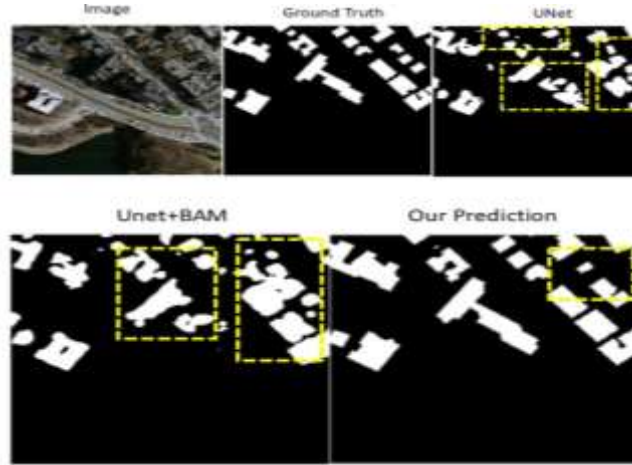
TABLE VII ABLATION STUDY

Dataset	Model	OA	IoU	Precision	Recall	F1
Massachusetts	UNet	88.74	57.94	87.85	94.47	90.08
Massachusetts	UNet + BAM	93.88	68.11	90.13	78.62	83.98
Massachusetts	UNet + BAM + EAM	93.50	85.42	91.00	84.00	87.23
WHU	UNet	97.20	83.51	87.85	94.47	90.89
WHU	UNet + BAM	98.15	88.50	94.10	93.72	93.88
WHU	UNet + BAM + EAM	98.50	92.33	95.00	93.00	94.80
INRIA	UNet	88.74	58.19	87.85	94.47	90.08
INRIA	UNet + BAM	91.23	76.83	71.09	82.50	76.37
INRIA	UNet + BAM + EAM	93.50	83.12	91.00	84.00	87.34

- UNet Baseline:** The baseline UNet model provides a solid starting point for our task across all datasets, though performance exhibits some variation. On the Massachusetts Dataset, UNet achieves an Overall Accuracy (OA) of 88.74\% with an IoU of 57.94\% and an F1 score of 90.08\%. The model effectively captures the majority of building pixels (Recall 94.47\%), but its lower Precision (87.85\%) indicates some misclassification of non-building areas, leading to over-segmentation and reduced IoU. For the WHU Dataset however, UNet achieves higher metrics, with an OA of 97.20\%, IoU of 83.51\%, Precision of 87.85\%, Recall of 94.47\%, and F1 score of 90.89\%. This improvement is likely attributable to the larger number of training samples available. Again, on the INRIA Dataset, UNet attains an OA of 88.74\% and an IoU of 58.19\%, with Precision and Recall at 87.85\% and 94.47\%, respectively, yielding an F1 score of 90.85\%. Across all three datasets, the UNet baseline exhibits strong building pixel detection capability but faces difficulty with

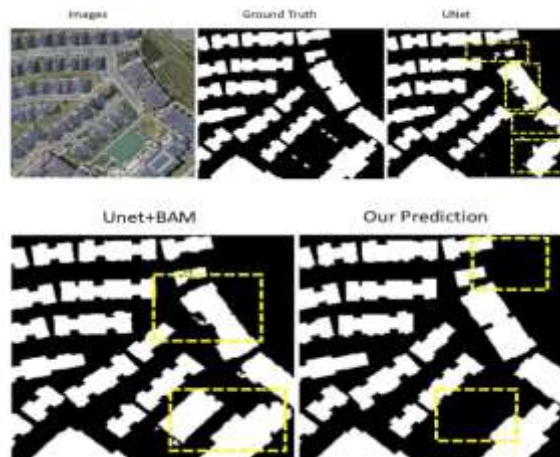
boundary delineation, thereby establishing a performance baseline for improvement through boundary-focused modules.

Fig7: Impact of Gradual Addition of Attention Modules on Massachusetts Building Dataset



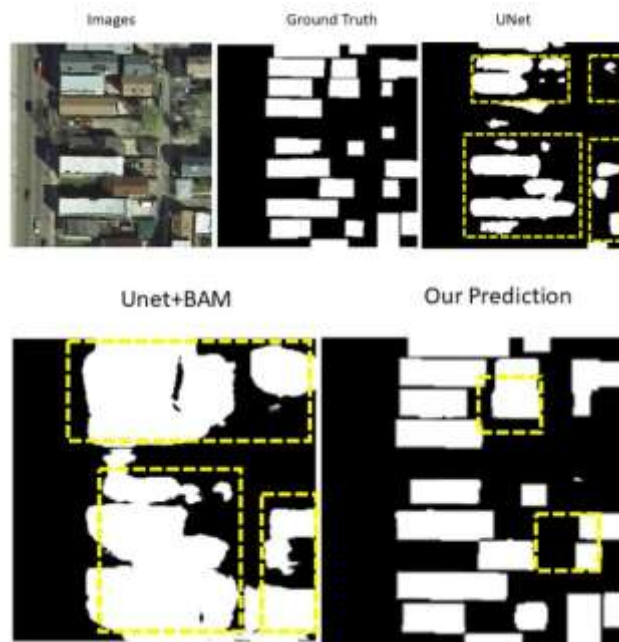
- Contribution of BAM:** Incorporating the Boundary Attention Module (BAM) heightens boundary awareness and markedly improves segmentation performance across all datasets. On the Massachusetts Dataset, BAM elevates the IoU from 57.94\% to 68.11\% and raises OA to 93.88\%. Precision improves to 90.13\%, though the increased boundary focus slightly reduces Recall to 78.62\%. On the WHU Dataset, BAM raises OA to 98.15\% and increases IoU to 88.50\%, with Precision rising significantly to 94.10\%. Recall slightly decreases to 93.72\%, resulting in an F1 score of 93.88\%, as the model becomes more discerning at boundaries. For the INRIA Dataset, BAM achieves an OA of 91.23\% and an IoU of 76.83\%, with Precision at 71.09\% and Recall at 82.50\% (F1 score 76.43\%). Across datasets, BAM improves boundary detection by focusing on transition regions, enhancing the model's ability to distinguish building edges, and reducing over-segmentation, though sometimes at the cost of Recall.

Fig 8: Impact of Gradual Addition of Attention Modules on WHU Building Dataset



- Contribution of EAM:** Layering the Edge Attention Module (EAM) over BAM produces additional improvements, especially in IoU and precision. On the Massachusetts Dataset, the addition of EAM raises IoU significantly to 85.42\% and boosts F1 score to 87.23\%, while OA remains high at 93.50\%. Recall and precision both reach 84.00\% and 91.00\%, respectively. The earlier issue of compromised recall is resolved through precise detection of sharper edges and complex geometric shapes. In the WHU Dataset, EAM improves OA to 98.50\% and IoU to 92.33\%, with high precision of 95.00\% and recall of 93.00\%, resulting in an F1 score of 94.80\%. On the INRIA Dataset, EAM raises OA to 93.50\% and IoU to 83.12\%, with precision and recall at 91.00\% and 84.00\%, resulting in an F1 score of 87.34\%. The dual attention strategy (BAM and EAM) enhances edge detection, enabling the model to capture finer boundary details and achieve an effective precision-recall balance for accurate building segmentation across varied datasets.

Fig.10: Impact of Gradual Addition of Attention Modules on Inria Aerial Image Labeling dataset



From the Fig. 7,8 and 9, we can observe enhanced boundary detection following the addition of BAM, though challenges with complex shapes and edges persist. Upon integrating EAM, the model more effectively detects these sharp edges and structural shapes.

Statistical Similarity Analysis

In this subsection, we analyzed the distributional differences between established baseline models and our model using KL divergence (Kullback-Leibler divergence) loss. Lower values indicate that a model's prediction histogram is closer to the ground truth histogram, reflecting better alignment in terms of intensity distribution and finer preservation of structural details.

TABLE VVIII KL DIVERGENCE LOSS OF THE MODELS FOR MASSACHUSSETS BUILDING DATASET

Model	KL-Divergence Loss
UNet	0.17114
Attention-UNet	0.17066
UNet++	0.16449
Attention-UNet++	0.17884
PSPNet	0.18114
DeepLabV3+	0.00169
Ours	0.00083

Table VII reports the KL-Divergence Loss for each model on the Massachusetts Building Dataset; our model records the lowest loss at 0.00083. The reason for this low divergence is that our model incorporates EAM and BAM, which focus on details such as edges and geometrical shapes and boundaries, respectively. This facilitates more precise alignment between the true and predicted distributions. DeepLabV3+ also performs well with a loss of 0.00169, benefiting from its multi-scale context capture but still lacking our model's finer feature discrimination. Models such as UNet and Attention-UNet++ exhibit higher losses, suggesting that although they detect general building structures effectively, they have difficulty achieving fine-grained distributional alignment. The comparatively high KL-Divergence of PSPNet suggests limitations in its boundary precision, emphasizing the effectiveness of our model's boundary-focused enhancements.

TABLE VII VIII KL DIVERGENCE LOSS OF THE MODELS FOR WHU BUILDING DATASET

Model	KL-Divergence Loss
UNet	0.00365
Attention-UNet	0.01132
UNet++	0.00438
Attention-UNet++	0.00091
PSPNet	0.01332
DeepLabV3+	0.01218
Ours	0.00083

Table VIII shows the KL-Divergence Loss for all models evaluated on the WHU Building Dataset, with our model attaining the lowest loss at 0.00083. Our model derives particular benefit from EAM on this dataset. It demonstrates strong capability in capturing complex geometric shapes, leading to close alignment between predictions and true distributions. While Attention-UNet++ also exhibits a comparatively low divergence at 0.00091, it does not match the precise feature discrimination attained by our model. Elevated losses for models such as PSPNet (0.01332) and DeepLabV3+ (0.01218) suggest challenges in boundary precision and distributional alignment, likely stemming from insufficient attention to critical spatial features. The performance of UNet and UNet++ further underscores the effectiveness of our model's targeted approach, as their higher divergence losses reflect reduced capability in capturing nuanced building structures.

TABLE IXx KL DIVERGENCE LOSS OF THE MODELS FOR INRIA AERIAL IMAGE LABELING DATASET

Model	KL-Divergence Loss
UNet	0.41127
Attention-UNet	0.34263
UNet++	0.32458

Attention-UNet++	0.34150
PSPNet	0.11719
DeepLabV3+	0.21927
Ours	0.03390

Table IX presents the KL-Divergence Loss for each model on the INRIA Aerial Image Labeling Dataset, where our model achieves the lowest loss of 0.03390. This superior performance on this dataset is largely attributable to the inclusion of BAM. BAM effectively differentiates narrow boundaries, resulting in closer alignment between our model's predictions and the true distribution. PSPNet, recording a loss of 0.11719, performs reasonably well but lacks the refined boundary detection and feature alignment demonstrated by our model. Elevated losses observed for UNet (0.41127) and Attention-UNet (0.34263) point to limitations in their ability to accurately distinguish between building and non-building regions. The results emphasize the effectiveness of our model's focused feature representation, enabling more precise segmentation compared to other architectures.

V. CONCLUSIONS

In this study, we have presented two attention modules designed for UNet: the Boundary Attention Module (BAM) for capturing building boundaries, and the Edge Attention Module (EAM) for identifying sharp edges and complex geometric patterns in buildings. To verify the effectiveness of our model, experiments were performed on the Massachusetts Building Dataset, the WHU Building Dataset, and the INRIA Aerial Image Labeling Dataset. Experimental results confirm that our model surpasses established baseline architectures on the majority of evaluation metrics. BAM assists in delineating building outlines, which is especially beneficial in urban areas where buildings are typically densely clustered. EAM offers a mechanism to represent complex geometric forms and abrupt transitions in building architecture. Nevertheless, the model faces challenges with very small buildings, indicating some loss of local spatial information. Future work could explore multi-scale fusion mechanisms, spatially-aware encoders, or additional skip connections to mitigate this limitation.

ACKNOWLEDGEMENT

THE AUTHORS EXPRESS THEIR GRATITUDE TO THE UNIVERSITY OF ASIA PACIFIC FOR PROVIDING RESEARCH FUNDING AND FACILITIES. THE AUTHORS ALSO ACKNOWLEDGE THE GUIDANCE OF DR. MOHAMMED IMAMUL HASSAN BHUIYAN .

REFERENCES

- [1] C. Iovan, D. Boldo, and M. Cord, "Detection, characterization, and modeling vegetation in urban areas from high-resolution aerial imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 1, no. 3, pp. 206–213, 2008.
- [2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. Springer, 2015, pp. 234–241.

- [4] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.
- [5] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.
- [6] Y. Li, Y. Li, X. Zhu, H. Fang, and L. Ye, "A method for extracting buildings from remote sensing images based on 3DJA-UNet3+," *Scientific Reports*, vol. 14, no. 1, p. 19067, 2024.
- [7] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.
- [8] S. El Hajjar, H. Kassem, F. Abdallah, and H. Omrani, "Enhancing building segmentation by deep multiview classification for advancing sustainable urban development," *Journal of Building Engineering*, vol. 83, p. 108421, 2024.
- [9] M. Yu, X. Chen, W. Zhang, and Y. Liu, "AGs-Unet: Building Extraction Model for High Resolution Remote Sensing Images Based on Attention Gates U Network," *Sensors*, vol. 22, no. 8, p. 2932, 2022.
- [10] H. Zhao, H. Zhang, and X. Zheng, "A multiscale attention-guided UNet++ with edge constraint for building extraction from high spatial resolution imagery," *Applied Sciences*, vol. 12, no. 12, p. 5960, 2022.
- [11] J. Lei, X. Liu, H. Yang, Z. Zeng, and J. Feng, "Dual Hybrid Attention Mechanism-Based U-Net for Building Segmentation in Remote Sensing Images," *Applied Sciences*, vol. 14, no. 3, p. 1293, 2024.
- [12] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," *Medical image analysis*, vol. 78, p. 102395, 2022.
- [13] X. Sun, A. Shi, H. Huang, and H. Mayer, "BAS4 Net: BoundaryAware Semi-Supervised Semantic Segmentation Network for Very High Resolution Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5398–5413, 2020.
- [14] Z. Nong, X. Su, Y. Liu, Z. Zhan, and Q. Yuan, "Boundary-aware dualstream network for vhr remote sensing images semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5260–5268, 2021.
- [15] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.
- [16] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.
- [17] M. E. Omati and F. T. Mahmoudi, "Efficiency analysis of itn loss function for deep semantic building segmentation," *Earth Science Informatics*, pp. 1–15, 2024.
- [18] L. Liu, J. Zhang, R. He, Y. Liu, Y. Wang, Y. Tai, D. Luo, C. Wang, J. Li, and F. Huang, "Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation," in 2020

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6488–6497.

[19] L.-C. Chen, “Rethinking atrous convolution for semantic image segmentation,” arXiv preprint arXiv:1706.05587, 2017.

[20] S. Xie and Z. Tu, “Holistically-nested edge detection,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1395–1403.

[21] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, “Richer convolutional features for edge detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3000–3009.

[22] S. Chen, W. Shi, M. Zhou, M. Zhang, and Z. Xuan, “CGSNet: A contour-guided and local structure-aware encoder-decoder network for accurate building extraction from very high-resolution remote sensing imagery,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 1526–1542, 2021.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.

[24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoderdecoder with atrous separable convolution for semantic image segmentation,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

[25] Y. Wu, P. Zhao, F. Wang, M. Zhou, S. Geng, and D. Zhang, “A prior-guided dual branch multi-feature fusion network for building segmentation in remote sensing images,” Buildings, vol. 14, no. 7, p.2006, 2024.

[26] V. Mnih, “Machine learning for aerial image labeling,” in PhD Thesis, University of Toronto, 2013.

[27] S. Ji, S. Wei, and M. Lu, “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.

[28] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “A large-scale benchmark for semantic segmentation of very high resolution satellite images,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 161–169.