**Journal of Computing and Electronic Information Management**
ISSN: 2413-1660 | Vol. 22, No. 1, 2026

# A Hybrid Pyramid and Strip Pooling Network for Accurate Building Extraction from Remote Sensing Images

**Hamdoun Youssef, Xingyuan Li, Yongtao Yu, Haiyan Guan**

School of Computer Science, Huai'an University, Meicheng Road Campus, Huai'an, Jiangsu, China

**Abstract:** Accurate extraction of building footprints from remote sensing imagery is important for urban planning, disaster management, and geographic information systems. However, complex building shapes, occlusions, and scale variation continue to challenge conventional segmentation models. This paper presents SRB-Net, a U-Net-based framework that combines three complementary components: (1) strip pooling (SP) for long-range horizontal and vertical context; (2) residual multi-scale atrous spatial pyramid pooling (RMASPP) with squeeze-and-excitation (SE) blocks for multi-scale and channel-aware feature learning; and (3) a bottleneck attention module (BAM) for refining skip-connection features. The model was trained with the Adam optimizer and evaluated on the aerial and Satellite Dataset II subsets of the WHU Building Dataset. Among the evaluated baselines, SRB-Net achieved the best overall performance, reaching 98.83% accuracy and 90.12% Intersection over Union (IoU) on the aerial dataset and 98.28% accuracy and 70.89% IoU on Satellite Dataset II. These results show consistent performance improvements across the two evaluated WHU subsets while avoiding claims beyond the within-dataset experimental setting.

**Keywords:** Building Extraction; Remote Sensing Imagery; Semantic Segmentation; Multi-Scale Feature Fusion; Strip Pooling; Attention Mechanism.

## 1. Introduction

Accurate extraction of building footprints from high-resolution remote sensing imagery is a core task in urban analysis, yet it remains challenging due to scale variation, occlusions, and complex urban textures. Deep learning-based methods have demonstrated remarkable capabilities in addressing these issues. Ji et al. [1] developed a fully convolutional network that performs pixel-level segmentation across multisource imagery, laying the groundwork for deep building extraction models. To improve robustness against scale diversity, Ji et al. [2] introduced a scale-adaptive architecture that enhances feature learning for buildings of varying sizes. Lin et al. [3] contributed ESFNet, a lightweight and efficient network optimized for high-resolution imagery, offering a practical balance between speed and accuracy. Attention-based methods have further improved segmentation precision; MAP-Net, proposed by Zhu et al. [4], integrates multiple attention paths to strengthen contextual understanding and boundary detection. Expanding on this, Cai and Chen [5] introduced MHA-Net, which fuses hybrid attention mechanisms to capture both global semantics and fine-grained details, resulting in improved segmentation performance under complex scene conditions. These advancements form the basis for further innovations in building extraction, particularly models that enhance multi-scale feature fusion and contextual awareness while maintaining computational efficiency.

Over the past decade, deep learning has significantly advanced the accuracy and efficiency of building extraction from high-resolution aerial and satellite imagery. Fully convolutional networks (FCNs) have laid the foundation for dense prediction tasks, including semantic segmentation of urban structures. Wu et al. [6] proposed a boundary-regulated FCN that improves roof segmentation by enforcing edge continuity and shape regularity. In a follow-up, Wu et al. [7] introduced a multi-constraint FCN framework that integrates spectral, spatial, and structural cues to enhance aerial building segmentation. Wei et al. [8] combined convolutional neural networks with regularization techniques, refining building footprint delineation and improving cross-domain generalization. Kang et al. [9] developed EU-Net, an efficient FCN designed to reduce computational cost while maintaining competitive accuracy in optical remote sensing scenarios. Deng et al. [10] incorporated attention gates into an encoder–decoder network, selectively focusing on building-relevant features to improve segmentation precision. Jing et al. [11] introduced a selective pyramid dilated network, which captures multi-scale context for fine-grained segmentation, particularly in high-resolution SAR imagery. Zhou et al. [12] proposed a large-scale hierarchical mapping approach that applies deep learning models to Gaofen-2 satellite imagery for urban-scale building extraction. Li et al. [13] presented Building-A-Nets, an adversarial training framework that enhances model robustness in complex urban environments. Xia et al. [14] applied semi-supervised learning using semantic edge detection to reduce the dependency on extensive manual annotations. Chen et al. [15] proposed SPMF-Net, which utilizes superpixel pooling and multi-scale feature fusion for improved building boundary preservation. Liu et al. [16] introduced DE-Net, a deep encoding architecture that effectively models spatial hierarchies in building features. Chen et al. [17] developed DR-Net, focusing on reducing noise segmentation and improving edge clarity through architectural refinements. Guo et al. [18] integrated attention blocks and multiple loss functions into U-Net to enhance feature discrimination. Hou et al. [19] introduced a multi-scale residual network that captures structural information across different spatial resolutions. Wang and Miao [20] implemented a deep residual U-Net, combining residual learning and skip connections to maintain spatial coherence in segmentation outputs. These models

collectively emphasize feature fusion, multi-scale processing, and boundary accuracy, reflecting the ongoing evolution toward more contextualized and structurally aware building extraction networks.

Beyond architectural innovations, recent research has explored auxiliary mechanisms to enhance semantic segmentation in remote sensing. Weakly supervised learning has emerged as a promising strategy for reducing annotation costs; Li et al. [21] introduced a framework that leverages weak supervision to guide semantic segmentation, showing effectiveness in extracting building footprints from high-resolution imagery. Yuan et al. [22] conducted a comprehensive survey on deep learning methods for building extraction, highlighting the importance of integrating vision techniques with spatial reasoning. Li et al. [23] further emphasized the role of combining semantic and geometric information, advocating for hybrid approaches that improve both recognition and delineation accuracy. Xu et al. [24] proposed HiSup, a hierarchical supervision network that aligns predicted polygons more closely with real building structures, improving structural consistency. Li et al. [25] presented a joint semantic–geometric learning model that simultaneously captures visual context and spatial regularity for enhanced polygonal segmentation. Li et al. [26] introduced the attraction field representation, which transforms buildings into structured vector fields, facilitating geometry-aware predictions. Zhang et al. [27] developed deep structured active contours that combine deep features with contour-based energy minimization to extract precise building shapes. Girard et al. [28] applied frame field learning for polygonal building extraction, enabling regularized and smooth structural outlines.

Attention mechanisms and spatial modeling techniques have also contributed to segmentation refinement. Woo et al. [29] introduced the Convolutional Block Attention Module (CBAM), which adaptively weights features along spatial and channel dimensions to enhance salient regions. Hou et al. [30] proposed strip pooling, a spatial pooling strategy that captures long-range dependencies across horizontal and vertical directions. Cheng et al. [31] proposed the Deep Active Ray Network (DARNet), which models building contours through directional rays. Architectures such as CGSANet, proposed by Chen et al. [32], incorporate contour-guided mechanisms to preserve local edge structures, while Chen et al. [33] developed a context feature enhancement network to improve segmentation in dense urban scenes. Fang et al. [34] proposed a pseudomask generation technique for weakly supervised building extraction, enabling training with minimal manual labels. Borba et al. [35] evaluated semantic segmentation under weak supervision and domain shifts, highlighting practical trade-offs in real-world applications. Miao et al. [36] introduced a feature residual analysis network that refines residual feature maps. Together, these mechanisms contribute to more reliable and structurally aware building extraction in challenging environments.

Despite substantial progress in deep learning-based building extraction, several challenges remain unresolved. Many methods struggle to maintain segmentation accuracy across varying building scales, particularly for small or elongated structures. Architectures such as U-Net [37] and its variants provide strong baselines but can lose spatial detail during downsampling, resulting in coarse boundaries and fragmented predictions. Attention and multi-scale strategies have been used to address these limitations, but they are often introduced separately without coordinated interaction between long-range context modeling, multi-scale feature aggregation, and skip-connection refinement. These limitations motivate a unified framework that strengthens feature representation, suppresses redundant activations, and preserves fine structural information.

In this paper, SRB-Net is presented as a U-Net-based framework for building extraction from high-resolution remote sensing imagery. It integrates three components: (1) strip pooling (SP) [30] to capture long-range dependencies along horizontal and vertical dimensions; (2) a residual multi-scale atrous spatial pyramid pooling (RMASPP) module that aggregates context at several receptive fields and uses SE blocks for channel recalibration; and (3) BAM embedded in skip connections to suppress redundant features and emphasize building-relevant responses. SRB-Net is evaluated on the aerial and Satellite Dataset II subsets of the WHU Building Dataset. The experimental results show that SRB-Net achieved the highest IoU, precision, and F1 score on both datasets. It also achieved the highest recall on Satellite Dataset II, whereas UNetFormer achieved the highest recall on the aerial dataset.

# 2. Methodology

## 2.1. SRB-Net Network Architecture

To segment buildings of diverse scales and shapes, SRB-Net uses an encoder–decoder architecture built on U-Net [37]. As illustrated in Figure 1, the encoder applies repeated 3×3 convolutions with ReLU activations followed by 2×2 max pooling, progressively reducing spatial resolution while increasing channel depth. SRB-Net integrates strip pooling for long-range spatial context, RMASPP with SE blocks for multi-scale and channel-aware learning, and BAM in the skip connections for feature refinement. Together, these components strengthen semantic consistency and the visual coherence of predicted building regions.

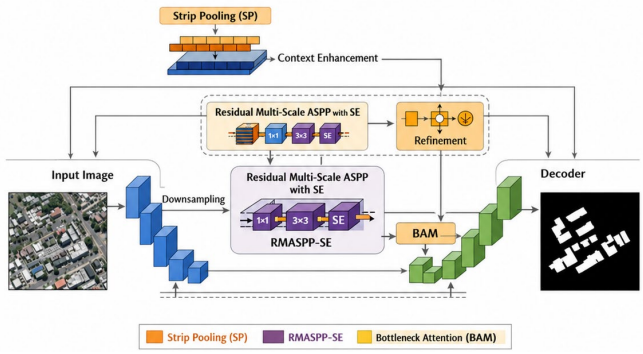


**Figure 1.** Overall SRB-Net architecture and detailed integration of Strip Pooling, RMASPP with SE, and Bottleneck Attention.

The SP module captures connections between distant spatial regions, thereby preserving contextual detail and reducing information loss. Its output is concatenated with the deepest encoder features and passed to the RMASPP module for multi-scale contextual feature extraction. The decoder then uses 2×2 transposed convolutions to progressively restore spatial resolution. At each skip connection, BAM filters low-level noise and emphasizes building-relevant features before encoder and decoder features are fused. Finally, a 1×1 convolution maps the refined features to the output classes.

## 2.2. Strip Pooling Module

The Strip Pooling (SP) module [30] captures elongated contextual dependencies along horizontal and vertical directions, as shown in Figure 2. Given an input tensor of size H×W, it aggregates features across entire rows and columns through parallel strip-pooling paths. Each output position can therefore incorporate information from distant locations that share the same row or column, enlarging the effective receptive field while retaining spatial detail.

Repeating this aggregation process establishes long-range spatial relationships and improves the representation of background context and irregular targets. For an input feature map, the horizontal and vertical strip-pooled outputs are calculated as shown in Equation (1).

$$y_{c,i}^{h} = \frac{1}{W}\sum_{j=0}^{W-1} X_{s,c,i,j}\,, \quad y_{c,j}^{v} = \frac{1}{H}\sum_{i=0}^{H-1} X_{s,c,i,j} \quad (1)$$

where $X_{s,(i,j)}$ is the input tensor, and H and W denote its height and width, respectively. The input is processed through the horizontal and vertical paths in parallel.

The two directional outputs are then combined as shown in Equation (2).

$$\mathrm{y_{c,i,j}} = \mathrm{y_{c,i}^{h}} + \mathrm{y_{c,j}^{v}} \quad (2)$$

where $y_{c,i}^{h}$ and $y_{c,j}^{v}$ denote the horizontal and vertical outputs, respectively.

The combined feature is transformed by a 1×1 convolution and sigmoid activation, then multiplied element-wise with the original input to produce the output z in Equation (3).

$$z = \mathrm{scale}(X_s, \sigma(f(y))) \quad (3)$$

where scale() denotes element-wise multiplication, σ is the sigmoid function, f is a 1×1 convolution, Xs is the input tensor, and y is the fused strip-pooled feature. By modeling anisotropic context, the SP module improves the segmentation of buildings with irregular shapes, varied orientations, and partial occlusions.

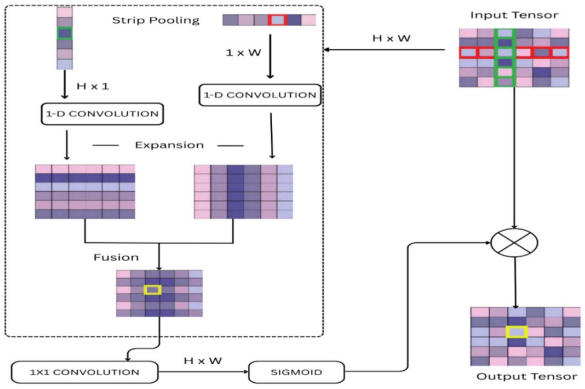


**Figure 2.** Structure of the Strip Pooling Module for Long-Range Contextual Feature Enhancement (adapted from [30]).

## 2.3. Residual Multi-Scale Atrous Spatial Pyramid Pooling Module

For building segmentation in remote sensing imagery, it is particularly important to better preserve spatial resolution features while capturing multi-scale contextual information.

Dilated (atrous) convolution expands the receptive field by inserting gaps between kernel elements while preserving feature-map resolution [38]. This allows RMASPP to aggregate multi-scale context without increasing the kernel parameter count.

For a dilation rate r, the operation is formally defined in Equation (4):

$$Y(i,j) = \sum_{m=0}^{K-1}\sum_{n=0}^{K-1} k(m,n)\cdot F(i+r\cdot m, j+r\cdot n) \quad (4)$$

where $Y(i,j)$ is the output feature at spatial location $(i,j)$, $F$ is the input feature map, $k(m,n)$ represents the convolutional kernel weights of size $K$x $K$, and $r$ denotes the dilation rate. The dilation rate $r$ controls the spacing between kernel elements, effectively enlarging the receptive field without increasing computational cost. For a standard 3×3 kernel, the effective receptive field size is given in Equation (5):

$$K_{\mathrm{eff}} = K + (K-1)(r-1) \quad (5)$$

Increasing the dilation rate from 1 to 3 to 6 progressively enlarges the receptive field, enabling the network to capture both fine local features and broader global context for building segmentation.

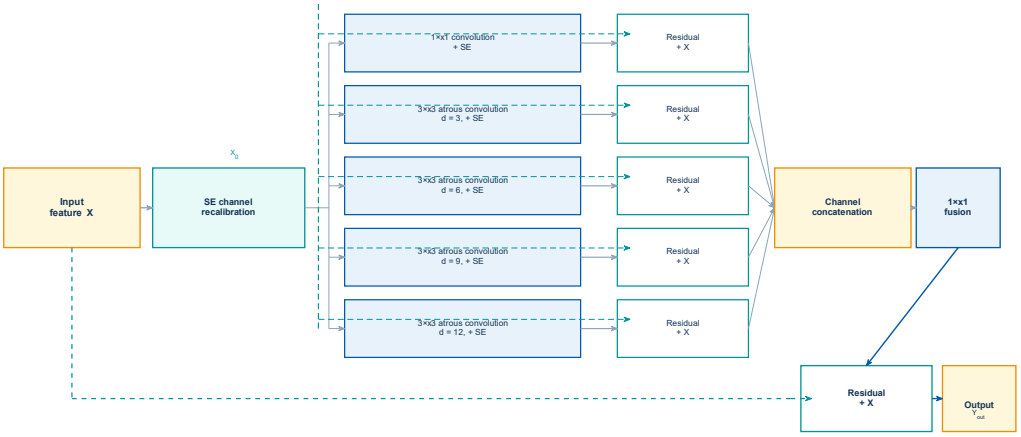


**Figure 3.** Architecture of the Residual Multi-Scale Atrous Spatial Pyramid Pooling (RMASPP) module.

To enhance multi-scale feature representation while preserving spatial detail, RMASPP combines parallel dilated convolutions, channel-wise attention, and residual learning. Let $X \in \mathbb{R}^{C\times H\times W}$ be the input feature map, where C, H, and W denote its channels, height, and width. The input is first recalibrated by an SE block [39], denoted $X_0 = SE(X)$. Four parallel 3×3 dilated-convolution branches then process $X_0$ using dilation rates D={3,6,9,12}; each branch is followed by an SE block, as shown in Equation (6). All branch convolutions preserve C channels so that the residual additions in Equation (8) are dimensionally valid. In the experiments, the SE reduction ratio was set to r=16.

$$Y_d = SE\left(f_d(X_0)\right), \text{ for } d \in D \quad (6)$$

Additionally, a standard 1×1 convolution is applied in parallel to extract local context, as shown in Equation (7):

$$Y_1 = SE(f_{1\times1}(X_0)) \quad (7)$$

Each output branch is enhanced by a residual connection from the input, as shown in Equation (8):

$$Y'_d = Y_d + X, \text{ for all } d \in \{1,3,6,9,12\} \quad (8)$$

All enhanced feature maps are concatenated channel-wise, as shown in Equation (9):

$$Y_{\text{cat}} = Y'_1 \| Y'_3 \| Y'_6 \| Y'_9 \| Y'_{12} \quad (9)$$

Finally, a 1×1 convolution is applied to fuse the aggregated features, and the original input X is added via a residual connection to form the output in Equation (10):

$$Y_{\text{out}} = f_{1\times1}(Y_{\text{cat}}) + X \quad (10)$$

This configuration enables the RMASPP module to capture both local and global contextual information across varied spatial scales, which is crucial for segmenting buildings with diverse shapes and sizes in remote sensing imagery. The complete RMASPP architecture is summarized in Figure 3.

## 2.4. Integration of the SE Module

Within RMASPP, the Squeeze-and-Excitation (SE) module [39] adaptively recalibrates channel responses so that informative building features receive greater weight.

Given an input tensor, the initial convolutional transformation is expressed in Equation (11), and the SE workflow is illustrated in Figure 4.

$$U = F_{\text{tr}}(X_{\text{se}}) \quad (11)$$

where $U$ is the generated feature map; $F_{\text{tr}}$ is the convolution operator; $X_{\text{se}}$ is the input tensor.

The squeeze stage summarizes each feature channel by global average pooling, producing one scalar descriptor per channel. This compression operation is defined in Equation (12).

$$z_c = F_{\text{sq}}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \quad (12)$$

where $z_c$ represents the $c$ component of z; $u_c$ is the $c$ component of $u$ .

The SE module explicitly models inter-channel dependencies so that weak or irrelevant responses are suppressed while channels supporting building segmentation are emphasized. Its operational flow is illustrated in Figure 4.

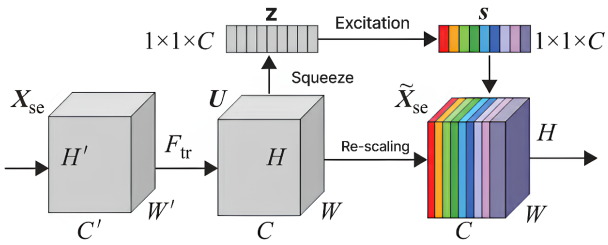


**Figure 4.** Squeeze-and-Excitation (SE) workflow for channel-wise feature recalibration (adapted from [39]).

The excitation stage learns nonlinear channel dependencies from the compressed descriptor through a gated transformation. The resulting weights rescale the feature channels. Within RMASPP, the five recalibrated branch outputs are concatenated, fused by a 1×1 convolution, and combined with the residual input.

To enhance the discriminative power of the features, the output z is applied to two fully connected layers to obtain the activation vector s. Vector s is then used to rescale the feature map, and the activation is calculated using Equation (13).

$$s = F_{\text{ex}}(z, W) = \sigma\left(W_2\delta(W_1 z)\right) \quad (13)$$

where $F_{\text{ex}}$ is the activation function; δ represents the ReLU function; σ represents the sigmoid activation function; and the hyperparameter r represents the reduction rate.

## 2.5. Bottleneck Attention Module

The Bottleneck Attention Module (BAM) [40] is inserted into the skip connections to strengthen building-related responses and suppress irrelevant background features. BAM combines channel and spatial attention before encoder and decoder features are fused. Its structure is shown in Figure 5.

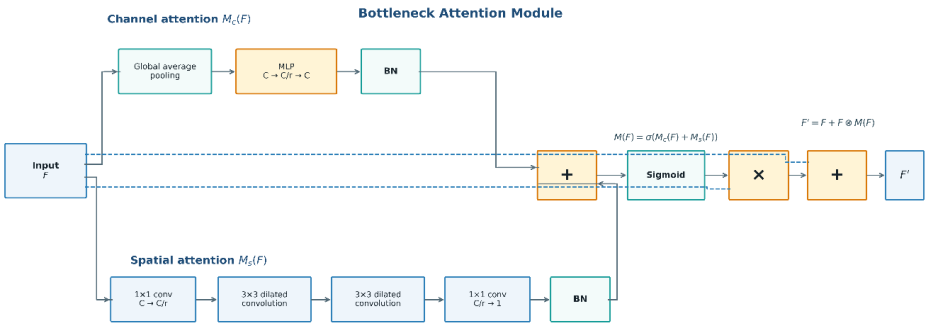


**Figure 5.** Bottleneck Attention Module used for skip-connection refinement (adapted from [40]).

For an input feature map $F \in \mathbb{R}^{C\times H\times W}$, BAM infers a three-dimensional attention map $M(F) \in \mathbb{R}^{C\times H\times W}$. The refined feature map F′ is computed using the residual-attention formulation in Equation (14).

$$F' = F + F \otimes M(F) \quad (14)$$

where M(F) is the attention map generated by BAM and ⊗ denotes element-wise multiplication.

The channel-attention branch first applies global average pooling to produce a C×1×1 channel descriptor. A two-layer multilayer perceptron with reduction ratio r=16 and batch normalization then produces the channel-attention vector $M_c(F)$, as defined in Equation (15) [40].

$$M_c(F) = \text{BN}(W_1(W_0\text{AvgPool}(F) + b_0) + b_1) \quad (15)$$

where $b_0$ and $b_1$ are bias terms, and BN denotes batch normalization.

The spatial-attention branch projects F to C/r channels using a 1×1 convolution, applies two 3×3 dilated convolutions with dilation value d=4, and then reduces the output to a 1×H×W spatial-attention map through a final 1×1 convolution and batch normalization. The spatial branch $M_s(F)$ is defined in Equation (16).

$$M_s(F) = \text{BN}\left(C_{1\times1}\left(C_{3\times3}\left(C_{3\times3}\left(C_{1\times1}(F)\right)\right)\right)\right) \quad (16)$$

where $C_{1\times1}$ and $C_{3\times3}$ denote convolutional operations with kernel sizes of $1\times1$ and $3\times3$, respectively. The first $1\times1$ convolution reduces the number of channels from $C$ to

$C/r$, where $r = 16$is the reduction ratio. The two successive $3 \times 3$dilated convolutions, each using a dilation rate of $d = 4$, aggregate contextual information over a larger receptive field without reducing the spatial resolution. The final $1 \times 1$ convolution reduces the channel dimension to one, producing a spatial-attention map of size $1 \times H \times W$. Batch normalization (BN) is then applied to adjust the scale of the resulting attention map.

$$M(F) = \sigma\big(M_c(F) + M_s(F)\big) \quad (17)$$

BAM therefore combines global channel selection with spatial context to recalibrate skip-connection features before encoder–decoder fusion. The implementation used here follows the standard BAM formulation [40]; its internal structure is summarized in Figure 5.

# 3. Results and Discussion

## 3.1. Datasets and Experimental Settings

SRB-Net was evaluated on the aerial dataset and Satellite Dataset II of the WHU Building Dataset [1]. The aerial dataset covers approximately 450 km² in Christchurch, New Zealand, and provides 8,189 image tiles of 512×512 pixels containing approximately 187,000 buildings. The official split contains 4,736 training tiles (130,500 buildings), 1,036 validation tiles (14,500 buildings), and 2,416 test tiles (42,000 buildings). Satellite Dataset II covers approximately 860 km² in East Asia and contains 17,388 tiles of 512×512 pixels with 34,085 buildings. It is divided into 13,662 training tiles containing 25,749 buildings and 3,726 test tiles containing 8,358 buildings. The diverse building scales, shapes, and scene densities make both subsets challenging benchmarks for building segmentation.

The WHU Building Dataset serves as a widely recognized benchmark for evaluating building footprint segmentation in high-resolution remote sensing imagery. It comprises a large collection of aerial and satellite images obtained from geographically diverse regions, including Christchurch, New Zealand, and selected urban areas in East Asia. These regions present a wide range of urban morphologies, from sparse rural landscapes with minimal structural presence to densely built-up environments with high variability in building scale, shape, and orientation. Each image in the dataset is provided with a corresponding binary ground truth mask, where building footprints are annotated in white and non-building background areas are represented in black. This pixel-level supervision enables robust training and reliable quantitative evaluation of segmentation models. The dataset is particularly useful for studying generalization under heterogeneous urban contexts, scale variation, and challenges such as occlusions and overlapping structures. Figure 6 illustrates representative samples from the dataset. Columns (a) and (b) show image-mask pairs from two distinct urban areas. Each column includes three rows, showcasing a variety of scene types. The top rows depict largely undeveloped or peri-urban areas, where building presence is sparse or entirely absent, providing insight into false positive susceptibility. The middle rows illustrate moderately dense residential neighborhoods with well-defined, rectangular building footprints, emphasizing typical suburban segmentation tasks. The bottom rows feature densely constructed urban zones with complex layouts, including rotated, closely packed buildings and narrow alleys, which challenge both boundary accuracy and small-object detection. These representative examples highlight the diverse visual characteristics and structural complexities captured in the WHU dataset, reinforcing the need for segmentation models that are both precise and adaptable across varying scene conditions. They serve to visually contextualize the segmentation challenges addressed in this work and motivate the design of architectures capable of robust generalization across different spatial configurations.

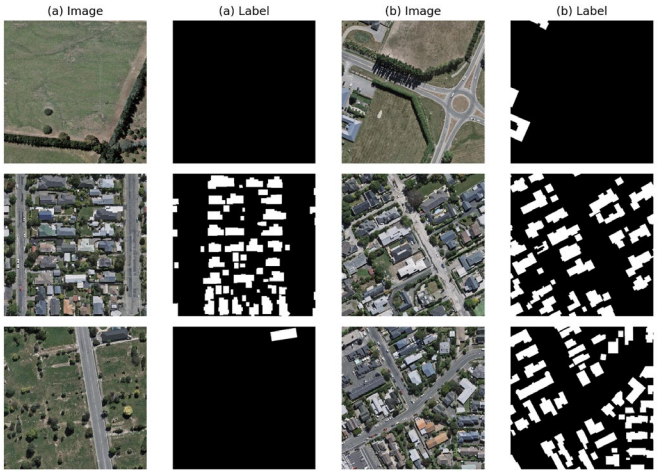


**Figure 6.** Sample Images and Ground Truth Annotations from the WHU Building Dataset

To enhance the generalization ability of the model and prevent overfitting, a diverse set of data augmentation techniques was applied to the WHU Building Dataset during training. These augmentations are designed to simulate variations in viewpoint, illumination, and image quality that commonly occur in real-world remote sensing scenarios. The transformations include both geometric and photometric modifications, targeting robustness to spatial orientation, pixel-level noise, and contrast variations.

As shown in Figure 7, the original image-label pair is presented in subfigure (a), while subfigures (b) through (d) illustrate geometric augmentations: 90° rotation, vertical flip, and horizontal flip, respectively. These transformations expose the model to various building alignments and perspectives, allowing it to learn rotational and reflection invariance. Subfigures (e) and (f) depict the application of Gaussian noise and Poisson noise, which mimic sensor noise and signal distortion, challenging the model to maintain performance under degraded image conditions. Subfigure (g) demonstrates the effect of Adaptive Histogram Equalization (AHE), a contrast enhancement technique that amplifies local image detail, particularly beneficial in scenes with low contrast or heavy shadows.

Geometric transformations were applied identically to each image and its corresponding mask, whereas Gaussian noise, Poisson noise, and AHE were applied only to the image. Masks remained binary after augmentation. Together, these transformations increased training diversity without misaligning images and labels.

Experiments were conducted on Ubuntu 20.04 using a 14-vCPU Intel Xeon Platinum 8362 processor, 45 GB of memory, and one NVIDIA RTX 3090 GPU with 24 GB of memory. The implementation used Python 3.8.0, PyTorch 1.11.0, and CUDA 11.3. Training was performed for 200 epochs with a batch size of 4, the Adam optimizer, and an initial learning rate of 0.0001.

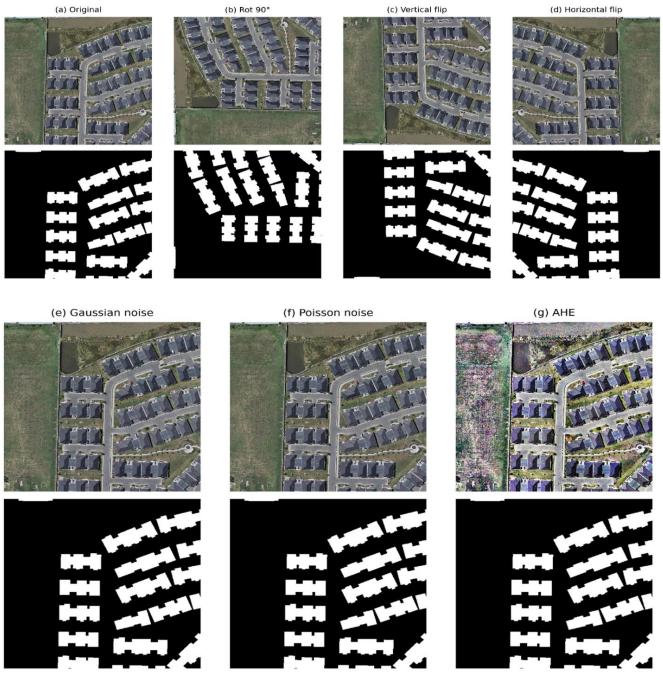


**Figure 7.** Illustration of Data Augmentation Techniques Applied to the WHU Building Dataset

Let TP, TN, FP, and FN denote the numbers of true-positive, true-negative, false-positive, and false-negative pixels, respectively. Accuracy, Intersection over Union (IoU), precision, recall, and F1 score were defined as Accuracy = (TP + TN)/(TP + TN + FP + FN), IoU = TP/(TP + FP + FN), Precision = TP/(TP + FP), Recall = TP/(TP + FN), and F1 = 2 × Precision × Recall/(Precision + Recall). Each metric was computed independently for every test image and then averaged across all images in the test set. Because IoU, precision, recall, and F1 score were averaged separately at the image level, the reported mean IoU is not required to be algebraically recoverable from the reported mean precision, recall, or F1 score. All evaluated models were assessed using the same metric implementation and aggregation procedure.

## 3.2. Ablation Experiment

To evaluate the progressive effect of the proposed components, cumulative ablation experiments were conducted on both WHU subsets. The tested sequence consisted of the U-Net baseline, U-Net with strip pooling, U-Net with strip pooling and RMASPP, the preceding configuration with SE recalibration inside RMASPP, and the complete SRB-Net with BAM. Because the components were introduced cumulatively in a fixed order, the results measure the incremental change associated with each successive configuration and should not be interpreted as an independent or factorial evaluation of every module.

The aerial-dataset ablation study evaluates the progressive integration of SRB-Net's components (Table 1). The U-Net baseline achieved an IoU of 88.09% and an F1 score of 93.56%. Adding SP increased these values to 88.15% and 93.84%, respectively. Adding RMASPP raised the IoU to 88.97% and the F1 score to 94.08%. Incorporating SE within RMASPP further improved the IoU to 89.64% and the F1 score to 94.34%. Finally, adding BAM produced the best performance, with an IoU of 90.12% and an F1 score of 94.64%. These cumulative results show incremental gains in the tested integration order; they do not constitute a full factorial independence test. Figure 8 summarizes the IoU and F1-score progression.

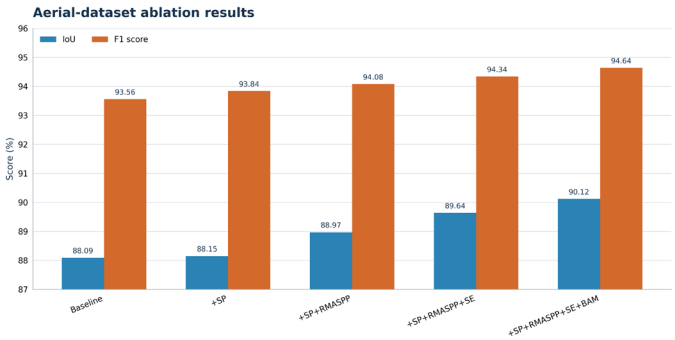


**Figure 8.** Aerial-dataset ablation results: IoU and F1-score progression.

**Table 1.** Ablation results on the aerial dataset

| No. | Strip Pooling (SP) | RMASPP | SE | BAM | Accuracy (%) | IoU (%) | Precision (%) | Recall (%) | F1 Score (%) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | 98.56 | 88.09 | 94.02 | 93.11 | 93.56 |
| 2 | ✓ | | | | 98.63 | 88.15 | 94.34 | 93.35 | 93.84 |
| 3 | ✓ | ✓ | | | 98.69 | 88.97 | 94.54 | 93.62 | 94.08 |
| 4 | ✓ | ✓ | ✓ | | 98.70 | 89.64 | 94.67 | 94.01 | 94.34 |
| 5 | ✓ | ✓ | ✓ | ✓ | 98.83 | 90.12 | 94.73 | 94.56 | 94.64 |

**Table 2.** Ablation results on Satellite Dataset II

| No. | Strip Pooling (SP) | RMASPP | SE | BAM | Accuracy (%) | IoU (%) | Precision (%) | Recall (%) | F1 Score (%) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | 98.05 | 68.46 | 80.71 | 81.33 | 81.02 |
| 2 | ✓ | | | | 98.11 | 68.53 | 80.92 | 81.47 | 81.19 |
| 3 | ✓ | ✓ | | | 98.18 | 69.67 | 81.35 | 81.66 | 81.50 |
| 4 | ✓ | ✓ | ✓ | | 98.21 | 70.35 | 81.54 | 81.80 | 81.67 |
| 5 | ✓ | ✓ | ✓ | ✓ | 98.28 | 70.89 | 81.84 | 82.12 | 81.98 |

Table 1 reports the complete aerial-dataset ablation results. Across the cumulative sequence, IoU increased from 88.09% for U-Net to 88.15% with SP, 88.97% with RMASPP, 89.64% with SE, and 90.12% with BAM. F1 score increased from

93.56% to 94.64%. These monotonic gains support the complementary value of the components in the tested configuration.

Table 2 presents the cumulative ablation experiment on Satellite Dataset II. IoU increased from 68.46% for U-Net to 68.53% with SP, 69.67% with RMASPP, 70.35% with SE, and 70.89% with BAM; F1 score increased from 81.02% to 81.98%. These results quantify incremental gains in the tested cumulative order rather than statistically independent module effects.

## 3.3. Comparative Experiment

SegNet [41], U-Net++ [42], DeepLabV3+ [43], A2FPN [44], UNetFormer [45], and U-Net [37] were selected for comparison on the aerial dataset (Table 3). SRB-Net achieved the highest accuracy (98.83%), IoU (90.12%), precision (94.73%), and F1 score (94.64%) among the evaluated models. Relative to the strongest competing result for each metric, SRB-Net improved accuracy by 0.16 percentage points over U-Net++, IoU by 1.15 points over UNetFormer, precision by 0.04 points over U-Net++, and F1 score by 0.47 points over U-Net++. UNetFormer achieved the highest recall at 95.41%, 0.85 points above SRB-Net's 94.56%.

**Table 3.** Quantitative comparison on the aerial dataset

| Model | Accuracy (%) | IoU (%) | Precision (%) | Recall (%) | F1 Score (%) |
|---|---|---|---|---|---|
| SegNet | 98.62 | 88.35 | 94.43 | 91.87 | 93.13 |
| U-Net++ | 98.67 | 88.77 | 94.69 | 93.65 | 94.17 |
| DeepLabV3+ | 97.98 | 87.86 | 93.75 | 91.84 | 92.79 |
| A2FPN | 98.60 | 88.29 | 94.41 | 91.82 | 93.10 |
| UNetFormer | 98.63 | 88.97 | 91.92 | 95.41 | 93.63 |
| U-Net | 98.56 | 88.09 | 94.02 | 93.11 | 93.56 |
| SRB-Net | 98.83 | 90.12 | 94.73 | 94.56 | 94.64 |

Figure 9 compares predictions on three aerial scenes with different scales and background complexity. In the first row, large buildings are adjacent to vehicles and vegetation; several baselines produce false detections or irregular boundaries, whereas SRB-Net yields smoother and more complete footprints. The second row contains mixed-scale buildings near walls, roads, and vehicles. Here, the strip-pooling and multi-scale components help SRB-Net preserve elongated structures while reducing background confusion. The third row contains small buildings partially obscured by vegetation. SRB-Net retains more of these fine structures and produces more coherent boundaries, consistent with the complementary contributions observed in the ablation study.

Table 4 presents the comparison on Satellite Dataset II. SRB-Net achieved the highest reported value for all five metrics: 98.28% accuracy, 70.89% IoU, 81.84% precision, 82.12% recall, and 81.98% F1 score. Compared with the next-best evaluated model for each metric, SRB-Net improved accuracy by 0.06 percentage points over UNetFormer, IoU by 1.85 points over UNetFormer, precision by 0.59 points over DeepLabV3+, recall by 0.56 points over UNetFormer, and F1 score by 0.96 points over U-Net. These results show consistent within-dataset improvements on both evaluated WHU subsets; they are not presented as a cross-dataset transfer experiment.

**Table 4.** Quantitative comparison on Satellite Dataset II

| Model | Accuracy (%) | IoU (%) | Precision (%) | Recall (%) | F1 Score (%) |
|---|---|---|---|---|---|
| SegNet | 98.10 | 68.52 | 79.93 | 80.01 | 79.97 |
| U-Net++ | 98.16 | 68.61 | 80.94 | 79.91 | 80.42 |
| DeepLabV3+ | 98.03 | 68.17 | 81.25 | 79.94 | 80.59 |
| A2FPN | 98.11 | 68.38 | 80.42 | 79.89 | 80.15 |
| UNetFormer | 98.22 | 69.04 | 79.31 | 81.56 | 80.42 |
| U-Net | 98.05 | 68.46 | 80.71 | 81.33 | 81.02 |
| SRB-Net | 98.28 | 70.89 | 81.84 | 82.12 | 81.98 |

Figure 10 presents a qualitative comparison of the segmentation results on three representative scenes from Satellite Dataset II. The green regions indicate predicted building pixels, while the yellow boxes highlight challenging areas containing large structures, elongated buildings, and small irregular footprints. In Example 1, several comparison models produce internal gaps or incomplete boundaries within the large building, whereas SRB-Net generates a more complete and spatially coherent footprint. In Example 2, SRB-Net better preserves the continuity of the elongated structures and produces boundaries that more closely follow the ground-truth annotation. In last example, which contains smaller and more irregular buildings, SRB-Net maintains the principal building shapes while reducing fragmentation in the highlighted region. These qualitative observations are consistent with the quantitative results in Table 4, particularly the improved IoU and F1 score obtained by SRB-Net.

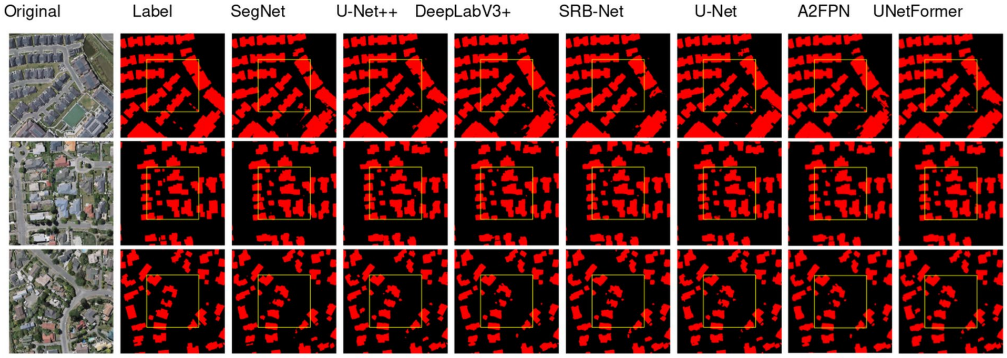


**Figure 9.** Qualitative Comparison of Building Segmentation Results Across Models on the Aerial Image Dataset

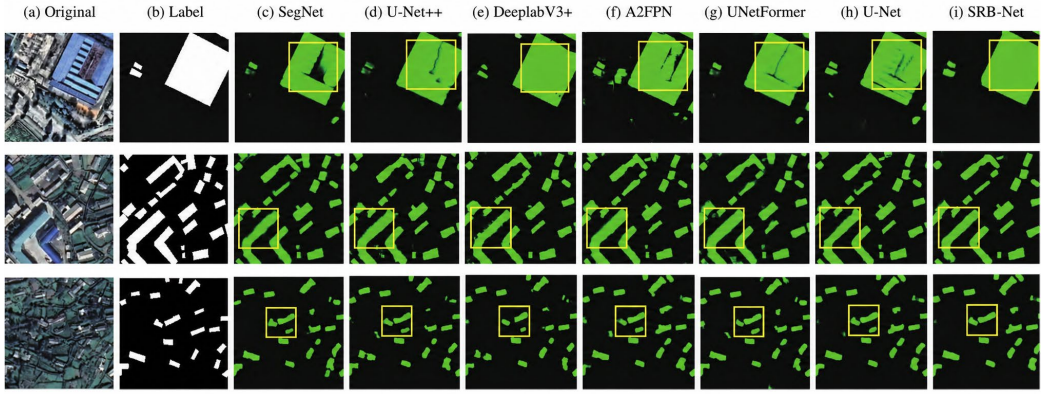


**Figure 10.** Qualitative Comparison of Building Segmentation Results Across Models on Satellite Dataset II

### 3.4. Discussion and Limitations

The larger IoU gain on Satellite Dataset II (+2.43 percentage points over U-Net) suggests that the combined context modules are particularly useful in the lower-performing satellite setting, where reduced spatial detail and heterogeneous building patterns make overlap more difficult. Accuracy remains high for every model because non-building pixels dominate many tiles; IoU and F1 score therefore provide more informative measures of building-region quality. The ablation studies are cumulative and establish incremental gains in one integration order, not fully independent causal effects. In addition, the manuscript reports no repeated-run variability and trains separate models within each dataset; consequently, the evidence supports performance consistency across the evaluated WHU subsets but not statistical significance or cross-domain transfer. Future work should include repeated-seed reporting, remove-one or factorial ablations, explicit complexity measurements, and broader failure-case analysis across additional satellite scenes.

## 4. Conclusion

This paper presented SRB-Net, a U-Net-based framework for building extraction from high-resolution remote sensing images. The architecture integrates strip pooling, RMASPP with SE blocks, and BAM to strengthen long-range context modeling, multi-scale feature representation, channel recalibration, and skip-connection refinement. Relative to U-Net, SRB-Net improved accuracy, IoU, precision, recall, and F1 score on the aerial dataset by 0.27, 2.03, 0.71, 1.45, and 1.08 percentage points, respectively. On Satellite Dataset II, the corresponding gains were 0.23, 2.43, 1.13, 0.79, and 0.96 points. The cumulative ablation results show consistent incremental gains as the components are integrated. The principal contributions are as follows:

Developed a U-Net-based building extraction framework that improves contextual representation and region-overlap performance.

Adapted and integrated strip pooling [30] to model long-range horizontal and vertical dependencies, supporting elongated and large building regions.

Introduced RMASPP with residual connections and SE recalibration to aggregate multi-scale contextual information.

Integrated standard BAM [40] in the skip connections to suppress irrelevant features before encoder–decoder fusion.

Conducted cumulative ablation and comparative experiments on two WHU subsets, obtaining the best overall performance among the evaluated baselines.

Several directions remain for future work. Multi-class urban-object extraction could extend the framework to roads and vegetation, while semi-supervised or self-supervised learning could reduce dependence on dense labels. Repeated-seed experiments, cross-dataset transfer tests, boundary-specific metrics, and remove-one ablations would strengthen the statistical and causal evidence. Finally, parameter, FLOP, latency, and memory analyses are needed before considering deployment on edge or embedded platforms.

## References

[1] Ji, S., Wei, S., & Lu, M. (2019). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on Geoscience and Remote Sensing, 57(1), 574-586. https://doi.org/10.1109/tgrs.2018.2858817

[2] Ji, S., Wei, S., & Lu, M. (2019). A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. International Journal of Remote Sensing, 40(9), 3308-3322. https://doi.org/10.1080/01431161.2018.1528024

[3] Lin, J., Jing, W., Song, H., & Chen, G. (2019). ESFNet: Efficient network for building extraction from high-resolution aerial images. IEEE Access, 7, 54285-54294. https://doi.org/10.1109/access.2019.2912822

[4] Zhu, Q., Liao, C., Hu, H., Mei, X., & Li, H. (2021). MAP-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. IEEE Transactions on Geoscience and Remote Sensing, 59(7), 6169-6181. https://doi.org/10.1109/tgrs.2020.3026051

[5] Cai, J., & Chen, Y. (2021). MHA-net: Multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 5807-5817. https://doi.org/10.1109/jstars.2021.3084805

[6] Wu, G., Guo, Z., Shi, X., Chen, Q., Xu, Y., Shibasaki, R., & Shao, X. (2018). A boundary regulated network for accurate roof segmentation and outline extraction. Remote Sensing, 10(8), 1195. https://doi.org/10.3390/rs10081195

[7] Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., & Shibasaki, R. (2018). Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. Remote Sensing, 10(3), 407. https://doi.org/10.3390/rs10030407

[8] Wei, S., Ji, S., & Lu, M. (2020). Toward automatic building footprint delineation from aerial images using CNN and regularization. IEEE Transactions on Geoscience and Remote Sensing, 58(3), 2178-2189. https://doi.org/10.1109/tgrs.2019.2954461

[9] Kang, W., Xiang, Y., Wang, F., & You, H. (2019). EU-net: An efficient fully convolutional network for building extraction

from optical remote sensing images. Remote Sensing, 11(23), 2813. https://doi.org/10.3390/rs11232813

[10] Deng, W., Shi, Q., & Li, J. (2021). Attention-gate-based encoder–decoder network for automatical building extraction. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 2611-2620. https://doi.org/10.1109/jstars.2021.3058097

[11] Jing, H., Sun, X., Wang, Z., Chen, K., Diao, W., & Fu, K. (2021). Fine building segmentation in high-resolution SAR images via selective pyramid dilated network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 6608-6623. https://doi.org/10.1109/jstars.2021.3076085

[12] Zhou, D., Wang, G., He, G., Yin, R., Long, T., Zhang, Z., Chen, S., & Luo, B. (2021). A large-scale mapping scheme for urban building from gaofen-2 images using deep learning and hierarchical approach. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 11530-11545. https://doi.org/10.1109/jstars.2021.3123398

[13] Li, X., Yao, X., & Fang, Y. (2018). Building-A-Nets: Robust building extraction from high-resolution remote sensing images with adversarial networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(10), 3680-3687. https://doi.org/10.1109/jstars.2018.2865187

[14] Xia, L., Zhang, X., Zhang, J., Yang, H., & Chen, T. (2021). Building extraction from very-high-resolution remote sensing images using semi-supervised semantic edge detection. Remote Sensing, 13(11), 2187. https://doi.org/10.3390/rs13112187

[15] Chen, J., He, F., Zhang, Y., Sun, G., & Deng, M. (2020). SPMF-net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion. Remote Sensing, 12(6), 1049. https://doi.org/10.3390/rs12061049

[16] Liu, H., Luo, J., Huang, B., Hu, X., Sun, Y., Yang, Y., Xu, N., & Zhou, N. (2019). DE-net: Deep encoding network for building extraction from high-resolution remote sensing imagery. Remote Sensing, 11(20), 2380. https://doi.org/10.3390/rs11202380

[17] Chen, M., Wu, J., Liu, L., Zhao, W., Tian, F., Shen, Q., Zhao, B., & Du, R. (2021). DR-net: An improved network for building extraction from high resolution remote sensing image. Remote Sensing, 13(2), 294. https://doi.org/10.3390/rs13020294

[18] Guo, M., Liu, H., Xu, Y., & Huang, Y. (2020). Building extraction based on U-net with an attention block and multiple losses. Remote Sensing, 12(9), 1400. https://doi.org/10.3390/rs12091400

[19] Hou, X., Wang, P., & An, W. (2022). Multi-scale residual network for building extraction from satellite remote sensing images. In IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium (pp. 1348-1351). IEEE. https://doi.org/10.1109/igarss46834.2022.9883509

[20] Wang, H., & Miao, F. (2022). Building extraction from remote sensing images using deep residual U-net. European Journal of Remote Sensing, 55(1), 71-85. https://doi.org/10.1080/22797254.2021.2018944

[21] Li, Z., Zhang, X., Xiao, P., & Zheng, Z. (2021). On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 3266-3281. https://doi.org/10.1109/jstars.2021.3063788

[22] Yuan, Y., Shi, X., & Gao, J. (2025). Building extraction from remote sensing images with deep learning: A survey on vision techniques. Computer Vision and Image Understanding, 251, 104253. https://doi.org/10.1016/j.cviu.2024.104253

[23] Li, Q., Mou, L., Sun, Y., Hua, Y., Shi, Y., & Zhu, X. X. (2024). A review of building extraction from remote sensing imagery: Geometrical structures and semantic attributes. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-15. https://doi.org/10.1109/tgrs.2024.3369723

[24] Xu, B., Xu, J., Xue, N., & Xia, G. (2023). HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision. ISPRS Journal of Photogrammetry and Remote Sensing, 198, 284-296. https://doi.org/10.1016/j.isprsjprs.2023.03.006

[25] Li, W., Zhao, W., Yu, J., Zheng, J., He, C., Fu, H., & Lin, D. (2023). Joint semantic–geometric learning for polygonal building segmentation from high-resolution remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 201, 26-37. https://doi.org/10.1016/j.isprsjprs.2023.05.010

[26] Li, Q., Mou, L., Hua, Y., Shi, Y., & Zhu, X. X. (2022). Building footprint generation through convolutional neural networks with attraction field representation. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-17. https://doi.org/10.1109/tgrs.2021.3109844

[27] Zhang, L., Bai, M., Liao, R., Urtasun, R., Marcos, D., Tuia, D., & Kellenberger, B. (2018). Learning deep structured active contours end-to-end. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 8877-8885). IEEE. https://doi.org/10.1109/cvpr.2018.00925

[28] Girard, N., Smirnov, D., Solomon, J., & Tarabalka, Y. (2021). Polygonal building extraction by frame field learning. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5887-5896). IEEE. https://doi.org/10.1109/cvpr46437.2021.00583

[29] Woo, S., Park, J., Lee, J., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19). Springer. https://doi.org/10.1007/978-3-030-01234-2_1

[30] Hou, Q., Zhang, L., Cheng, M., & Feng, J. (2020). Strip pooling: Rethinking spatial pooling for scene parsing. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4002-4011). IEEE. https://doi.org/10.1109/cvpr42600.2020.00406

[31] Cheng, D., Liao, R., Fidler, S., & Urtasun, R. (2019). DARNet: Deep active ray network for building segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7431-7439). IEEE.

[32] Chen, S., Shi, W., Zhou, M., Zhang, M., & Xuan, Z. (2022). CGSANet: 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, 15, 1526-1542. https://doi.org/10.1109/jstars.2021.3139017

[33] Chen, J., Zhang, D., Wu, Y., Chen, Y., & Yan, X. (2022). A context feature enhancement network for building extraction from high-resolution remote sensing imagery. Remote Sensing, 14(9), 2276. https://doi.org/10.3390/rs14092276

[34] Fang, F., Zheng, D., Li, S., Liu, Y., Zeng, L., Zhang, J., & Wan, B. (2022). Improved pseudomasks generation for weakly supervised building extraction from high-resolution remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1629-1642. https://doi.org/10.1109/jstars.2022.3144176

[35] Borba, P., De Carvalho Diniz, F., Da Silva, N. C., & De Souza Bias, E. (2021). Building footprint extraction using deep learning semantic segmentation techniques: Experiments and

results. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS (pp. 4708-4711). IEEE. https://doi.org/10.1109/igarss47720.2021.9553855

[36] Miao, Y., Jiang, S., Xu, Y., & Wang, D. (2022). Feature residual analysis network for building extraction from remote sensing images. Applied Sciences, 12(10), 5095. https://doi.org/10.3390/app12105095

[37] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 234-241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[38] Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1511.07122

[39] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7132-7141). IEEE. https://doi.org/10.1109/CVPR.2018.00745

[40] Park, J., Woo, S., Lee, J.-Y., & Kweon, I. S. (2018). BAM: Bottleneck attention module. In British Machine Vision Conference (BMVC). https://arxiv.org/abs/1807.06514

[41] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder–decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[42] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2020). UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging, 39(6), 1856-1867. https://doi.org/10.1109/TMI.2019.2959609

[43] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder–decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 801-818). Springer. https://doi.org/10.1007/978-3-030-01234-2_49

[44] Li, R., Zheng, S., Zhang, C., Duan, C., & Wang, L. (2022). A2-FPN for semantic segmentation of fine-resolution remotely sensed images. International Journal of Remote Sensing, 43(3), 1131-1155. https://doi.org/10.1080/01431161.2022.2030071

[45] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 190, 196-214. https://doi.org/10.1016/j.isprsjprs.2022.06.008