Image Anal Stereol 2026;45:103-122 doi: 105566/ias.3950
Original Research Paper

AN IMPROVED EVIT NETWORK FOR SEMANTIC SEGMENTATION OF HIGH-RESOLUTION REMOTE SENSING IMAGERY

RUI XU, YIHUI YANG✉, RENZHONG MAO, YINING ZHANG, YITENG LIN AND WEIPING ZHANG
School of Computing and Data Science, Fujian University of Technology, Fuzhou 350118, Fujian, China
e-mail: xurui@fjut.edu.cn, 18876228035@163.com, mao_renzhong@126.com, yn3504456536@163.com, w2814118594@163.com, z1912528153@gmail.com
(Received March 14, 2026; revised June 14, 2026; accepted June 14, 2026)

ABSTRACT

To address the issues of blurred building boundaries, small-object omission, and severe background interference in the semantic segmentation of high-resolution remote sensing imagery, this study proposes an improved method based on the Enhanced Vision Transformer Network (EViT). Specifically, this paper introduces a Grouped Cross-Cascaded Multi-Head Self-Attention (GCC-MSA) module to enhance feature diversity while maintaining linear complexity, and a Local-Global Feature Calibration (LGC) module to fuse CNN local details with Transformer global context. Coordinate Attention (CoAt) replaces conventional channel attention to strengthen channel-spatial feature representation. Additionally, Semantic-Guided Spatial Pyramid Pooling (SGSPP) and a GCC-MSA-guided Edge Perception (GEP) module reinforce multi-scale semantic perception and boundary extraction, while a Spatial Perception Gating Mechanism (SPGM) adaptively fuses dual-branch features. On the WHU Aerial, Massachusetts, and GF-7 Building Datasets, the model achieves Intersection-over-Union (IoU) scores of 92.33%, 77.81%, and 78.29%, respectively. These represent improvements of 0.57, 0.67, and 0.62 percentage points over the original EViT. The model demonstrates superior performance in small-building extraction, complex boundary segmentation, and background noise suppression, thereby providing a robust solution for precise surface object information extraction from high-resolution remote sensing imagery.

Keywords: Attention mechanism; CNN-Transformer fusion; high-resolution remote sensing imagery; Local-Global Feature Calibration; semantic segmentation; Spatial Perception Gating Mechanism.

INTRODUCTION

Accurate building extraction from high-resolution remote sensing imagery is essential for urban planning, disaster response, and geographic information systems. However, this task faces persistent challenges, including blurred building boundaries, omission of small objects, and severe background interference. Early approaches predominantly relied on convolutional neural networks (CNNs). The Fully Convolutional Network (FCN) (Long *et al.*, 2015) pioneered end-to-end semantic segmentation, forming the basis for subsequent encoder-decoder architectures like U-Net (Ronneberger , 2015) and ResNet (He , 2016). While these models effectively capture local textures and edges, they often fail to model long-range spatial dependencies, leading to fragmented segmentation in dense urban scenes. To alleviate this, multi-scale context aggregation modules such as Feature Pyramid Network (FPN) (Lin , 2017) and Pyramid Scene Parsing Network (PSPNet) (Zhao , 2017) were proposed. DeepLabv3+ (Chen , 2018) further improved performance by integrating atrous separable convolutions to enlarge the receptive field. Despite these advances, CNN-based methods still exhibit limitations, including insufficient multi-scale fusion and feature redundancy when processing dense small targets.

Attention mechanisms have emerged as a key technique to alleviate background interference and blurred boundaries through adaptive feature weighting. Early studies focused on channel attention; the Squeeze-and-Excitation (SE) module (Hu , 2018) introduced adaptive channel recalibration. Subsequent work incorporated spatial attention, leading to dual attention fusion designs. For instance, MAP-Net (Zhu , 2021) models channel, spatial, and edge attention in parallel, though its static allocation limits adaptability to diverse urban structures. The Vision Transformer (ViT) (Dosovitskiy , 2021) and its variants offer a new paradigm by modeling long-range dependencies via self-attention, overcoming the local receptive field limitation of CNNs. Swin Transformer (Liu , 2021) addresses the quadratic complexity issue through a shifted window self-atten-

tion mechanism. Lightweight designs such as MobileViT (Mehta and Rastegari, 2022) incorporate convolutional priors to mitigate computational cost. In remote sensing building extraction, specialized architectures such as MSTrans (Yang , 2024) introduce multi-scale Transformer modules, while LGDB-Net (Zhang , 2024) adopts a dual-branch design combining Transformer-based global features with convolutional local details. Hybrid CNN-Transformer frameworks leverage the complementary strengths of both architectures. TransFuse (Zhang , 2021) adopts a parallel dual-branch structure with a fusion gating module, but its fixed fusion weights limit adaptability to complex remote sensing scenarios. TransUNet (Chen , 2021) attempts to integrate local detail extraction with global modeling yet relies on dense 1×1 convolutions and static fusion strategies. Despite these efforts, several challenges persist. First, many fusion strategies remain static, relying on fixed weights or simple concatenation that cannot adapt to varying feature distributions. Second, they often introduce considerable computational overhead; for instance, CoAtNet (Dai , 2021) stacks convolution and attention at significant cost. Third, boundary representation remains insufficient: conflicts between local noise and global semantics lead to blurred edges, and Swin-Unet's Transformer decoder (Cao , 2022) struggles to recover fine-grained details, exacerbating edge ambiguity in building extraction.

To address these challenges, this study proposes an improved semantic segmentation network for high-resolution remote sensing imagery based on the EViT framework (Zhang , 2024). The main contributions of this work are summarized as follows:

(1) Original Algorithmic Contributions (GCC-MSA & GEP): To overcome the feature degradation problem in standard linear attention caused by the removal of Softmax and the independence of attention heads, we propose the GCC-MSA module. By introducing a dependency chain among attention heads, the module restores stronger nonlinear hierarchical modeling capability while maintaining linear computational complexity O(N). In addition, a GEP module is designed to address boundary blurring in aerial imagery by utilizing reversed semantic attention maps to refine building boundaries.

(2) Strategic Architectural Adaptations (LGC, CoAt, SGSPP, & SPGM): To further improve feature fusion and representation, several architectural adaptations are introduced within the hybrid framework. The SPGM dynamically balances CNN and Transformer branches to replace static fusion strategies. The LGC module performs layer-wise feature alignment to mitigate the semantic gap between heterogeneous representations. In addition, SGSPP enhances multi-scale feature extraction while suppressing background noise, and CoAt preserves precise positional information during feature modeling.

MATERIALS AND METHODS

Datasets

WHU Aerial Building Dataset (Ji , 2019): Collected by the Photogrammetry and Computer Vision Research Group of Wuhan University, this dataset covers regions including New Zealand with a total area exceeding 450 km². The imagery has a spatial resolution of 0.3 m and contains 8,188 optical remote sensing images cropped to 512×512 pixels with corresponding building annotations. The dataset is divided into training, validation, and test sets containing 4,736, 1,036, and 2,416 images, respectively (approximately 6:1:3). Despite the clear annotation of building and background categories, several challenges remain, including dense building distributions that cause occlusion and blurred boundaries, shadows from illumination variations, and small-scale buildings with limited pixel representation.

Massachusetts Building Dataset (Mnih, 2013): Collected by the Massachusetts Institute of Technology (MIT), this dataset contains 151 aerial images from the Boston area, each with a size of 1500×1500 pixels and a spatial resolution of 1 m, covering approximately 340 km². In the experiments, the images were cropped into overlapping 512×512 patches and augmented using geometric transformations such as flipping and rotation. This process produced 6,520 image patches, which were divided into training, validation, and test sets containing 3,920, 1,312, and 1,288 images, respectively (approximately 6:2:2). The dataset presents challenges such as occlusion from nearby trees, variations in roof texture and color, and occasional annotation inaccuracies due to blurred building boundaries.

GF-7 Building Dataset (Chen , 2024): Derived from Gaofen-7 satellite imagery, this dataset covers six typical Chinese cities, including Tianjin, Chongqing, and Guangzhou, with a total area of 573.17 km². It consists of 5,175 images of size 512×512 pixels with a spatial resolution of 0.65 m. The dataset contains annotations for 170,015 buildings, including 84.8% urban buildings and 15.2% rural buildings. It is divided into training, validation, and test sets at a ratio of 6:2:2, containing 3,106, 1,034, and 1,035 images, respectively. Typical challenges include cloud and building shadows that interfere with feature extraction, as well as small rural buildings

that are difficult to distinguish from surrounding vegetation.

Proposed Model Architecture

The network model framework takes "dual-path feature extraction, multi-dimensional feature enhancement and fusion, spatial perception gating decision-making output" as its core process. Through a three-stage collaborative design, it achieves an accurate balance between the local details and global semantics of buildings in high-resolution remote sensing imagery. The structure is illustrated in Fig. 1. In the initial encoding stage, input images of size 512×512×3 are first processed by the Initial Feature Extraction and Downsampling Module (IFEADM) for preliminary feature extraction and resolution reduction. The CNN branch, built with two stacked Inverted Residual Blocks, preserves local details by maintaining spatial resolution while increasing channel depth. Concurrently, the Transformer branch enhances the IFEADM output with Relative Position Encoding (RPE) to strengthen spatial cues, followed by a four-stage backbone for global context modeling. To prevent semantic detail decoupling, a LGC module is embedded at each stage for cross-branch fusion. Multi-scale features from all stages are aggregated by the Multi-scale Feature Aggregation Module (MFAM), and then refined sequentially by CoAt, SGSPP, and the GEP module. Finally, the enhanced features are passed to dual-branch prediction heads and fused by the SPGM, which learns adaptive spatial weights to balance global semantics from the Transformer branch and local details from the CNN branch. The network outputs a pixel-level building segmentation map of size 512×512×2, maintaining full resolution consistency with the input.

Overall Mathematical Formulation

While aligning with the established hybrid CNN Transformer paradigm, our architecture is distinguished by the integration of original algorithmic contributions (GCC MSA, GEP) and strategic architectural adaptations (LGC, CoAt, SGSPP, SPGM) tailored for remote sensing challenges. Grounded in the Complementary Learning Dynamics theory, the design exploits CNNs' inductive biases (translation invariance, locality) for capturing high frequency details (boundaries) and Transformers' ability to model low frequency global context (semantic information). Rather than a mere accumulation of modules, it forms a unified system that maximizes the orthogonality of these two representations. We mathematically formulate this as a dual path interaction problem, where gradient flows are optimized to prevent feature collapse in deep layers while preserving boundary precision.

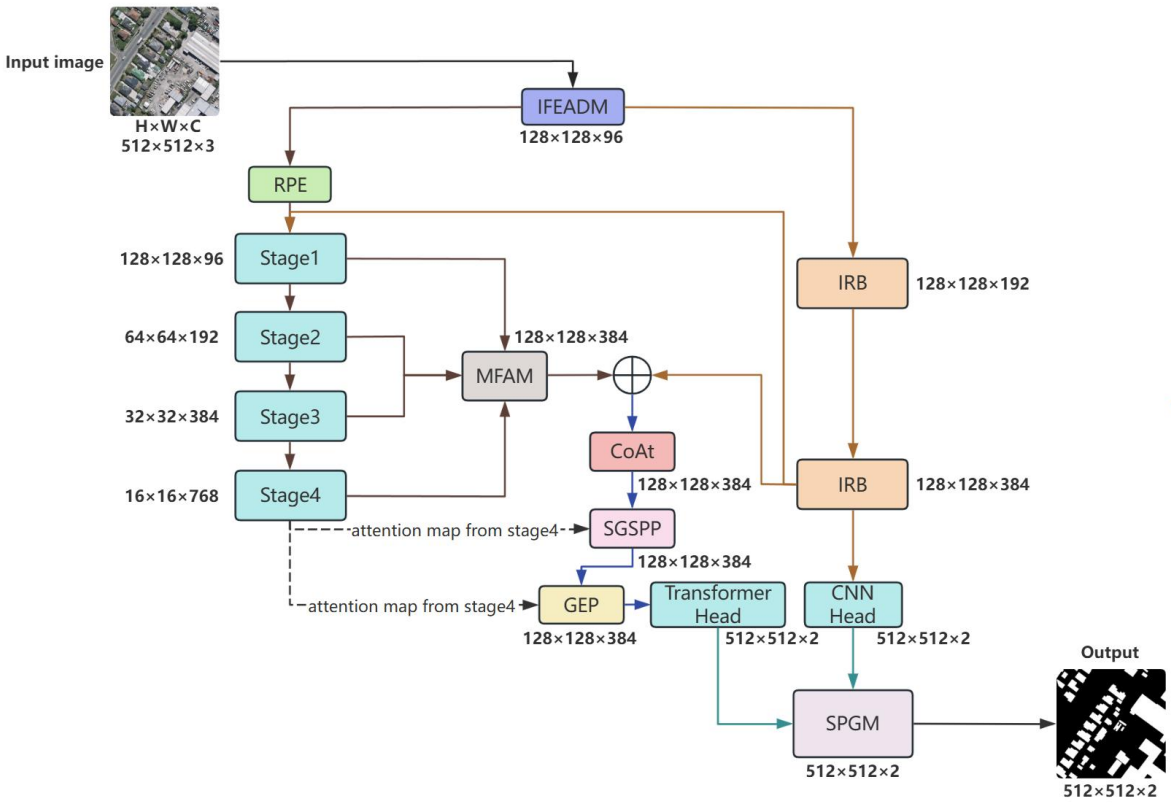

Fig. 1. *Overall Architecture of the Network.*

Let the input remote sensing image denoted as $X\epsilon\mathbb{R}^{H\times W\times C}$(with H=W=512, C=3), and the model as a mapping $\hat{Y}=\mathcal{F}_\theta(X)$, where $\hat{Y}\epsilon\mathbb{R}^{H\times W\times N_{class}}$ and θ are learnable parameters. Unlike standard single-stream networks, our dual-branch hybrid strategy is formalized as a composition of feature extraction, enhancement, and adaptive fusion. Feature extraction via the CNN branch (f_{cnn}) and the Transformer branch (f_{trans}). To address the optimization difficulty of pure ViTs, introduce a hierarchical interaction mechanism. Let F_{cnn}^0 and F_{trans}^0denote the initial feature maps generated by the IFEADM and RPE, respectively. Given the asymmetric design where the Transformer branch models deep semantics while the CNN branch preserves shallow high-frequency details. For the CNN branch, which comprises two IRB, the hierarchical features are updated for the first two stages ($k\epsilon\{1,2\}$):

$$F_{cnn}^{(k)}=\mathcal{H}_{IRB}^{(k)}(F_{cnn}^{(k-1)}) \quad (1)$$

Where $\mathcal{H}_{\mathrm{IRB}}^{(\mathrm{k})}(\cdot)$ denotes the transformation of the k-th IRB. For deeper stages ($k\epsilon\{3,4\}$), the CNN features are obtained by adaptively projecting the final output F_{cnn}^2 to the required dimensions: $F_{cnn}^{(k)}=\mathcal{P}_{proj}(F_{cnn}^2)$, where $\mathcal{P}_{proj}(\cdot)$ is implemented via a stride-based 3×3 convolution followed by Batch Normalization (BN) and ReLU. This aligns the spatial resolution and channel dimension of the shallow CNN features with the deep Transformer features at stage k, the Transformer branch evolves through all stages ($i\epsilon\{1,2,3,4\}$) and interacts with the aligned CNN features via:

$$F_{trans}^{(i)}=f_{LGC}(f_{GCC-MSA}(F_{trans}^{(i-1)}),F_{cnn}^{(i)}) \quad (2)$$

Where i denotes the stage index. The function f_{LGC} serves as a semantic calibrator, injecting local inductive biases from the CNN branch into the Transformer branch. This cross-branch interaction enforces representational orthogonality: the Transformer models long-range dependencies, while the CNN branch preserves high-frequency local details. The multi-scale semantic features are then aggregated via the MFAM module(f_MFAM) and sequentially refined through a cascade of enhancement modules:

$$F_{agg}=f_{MFAM}(\{F_{trans}^{(i)}\}_{i=1}^4) \quad (3)$$

$$F_{refined}=f_{GEP}\circ f_{SGSPP}\circ f_{CoAt}(F_{agg},F_{cnn}^{final}) \quad (4)$$

Here, $F_{agg}\epsilon\mathbb{R}^{128\times128\times384}$ denotes the aggregated global features from the MFAM, and $\mathrm{F}_{\mathrm{cnn}}^{\mathrm{final}}\epsilon\mathbb{R}^{128\times128\times384}$ represents the final detailed feature map from the CNN branch. The composition operator ∘ indicates that the global context is first spatially recalibrated by CoAt, then semantically filtered by SGSPP, and finally boundary-refined by GEP. Finally, the Spatial Perception Gating Mechanism f_SPGM serves as the decision function for fusing the dual representations:

$$\hat{Y}=f_{SPGM}(\varphi_{trans}(F_{refined}),\varphi_{cnn}(F_{cnn}^{final})) \quad (5)$$

Where φ_{trans} and φ_{cnn} denote the prediction heads that project features to the segmentation probability space R^(512×512×2) via upsampling and convolution. From an optimization perspective, this formulation ensures robust gradient propagation. The CNN branch provides a shorter gradient path akin to a macro-residual connection, allowing gradients to flow directly through f_SPGM even if they vanish in the deep Transformer branch due to attention complexity, thus stabilizing the initial feature encoding layers.

Dual-Branch Feature Encoder

The Initial Feature Extraction and Downsampling Module (IFEADM) serves as the first processing unit, extracting preliminary features and reducing spatial resolution from raw input images (Fig. 2(a)). The Inverted Residual Block (IRB), a core component of lightweight CNNs, employs an "expand-then-reduce" channel strategy to preserve feature representation while minimizing computational cost (Fig. 2(b)).

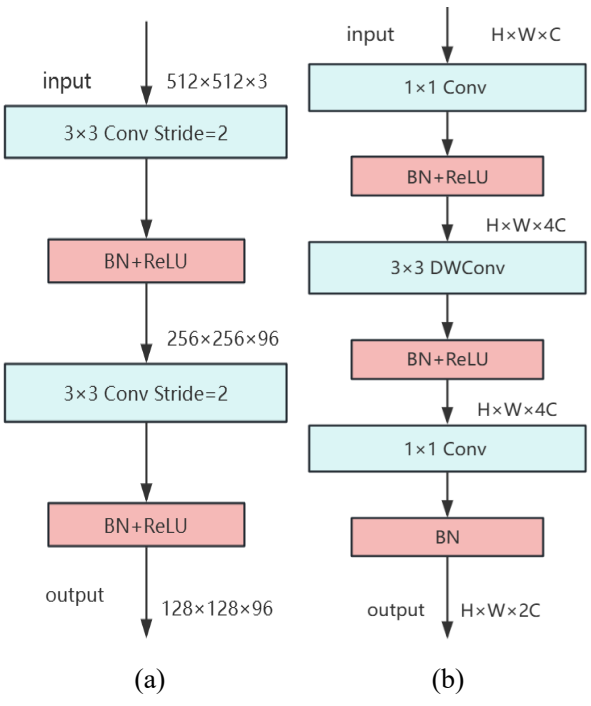

Fig. 2. *(a) Initial Feature Extraction and Downsampling Module. (b) Inverted Residual Blocks.*

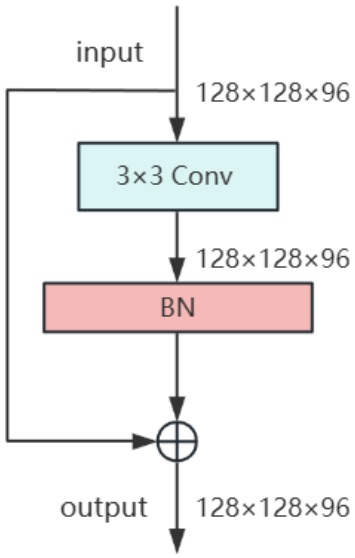

Fig. 3. *Relative Position Encoding*.

The RPE module is designed to embedding relative positional information into the feature sequence before it enters the Transformer backbone. Its structure is illustrated in Fig. 3

The RPE serves as a context injection mechanism. Its convolutional branch not only refines features but also learns and embeds the spatial layout of buildings. By injecting local spatial context into semantic features before Transformer-based global modeling, RPE compensates for the lack of spatial inductive bias in the Transformer branch. This enables subsequent modules to better capture spatial dependencies and accurately distinguish dense or complex building boundaries.

As the core modules of the Transformer branch, Stage1-Stage4 realize feature extraction and semantic aggregation from shallow to deep layers in the overall architecture. Each Stage contains multiple basic blocks, with the number of blocks being 2, 2, 6, and 2, respectively; each block comprises a GCC-MSA, LGC and Multi-Layer Perceptron(MLP). Taking Stage 4 as an example, the features of this Stage contain the most abundant semantic information. Its structure is illustrated in Fig. 4.

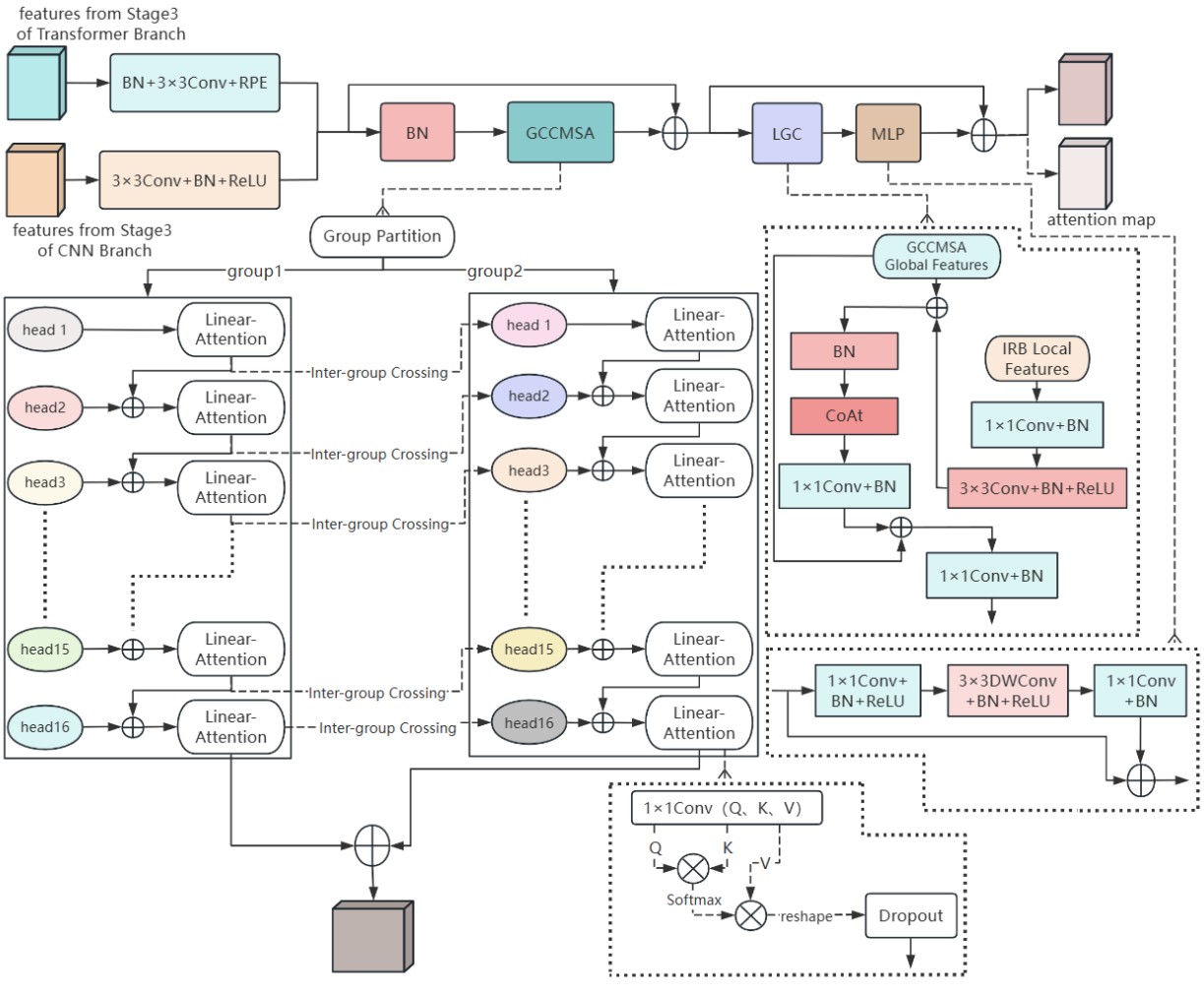

Fig. 4. *Stage 4.*

First, input feature preprocessing is performed. The global feature of Stage 3 in the Transformer branch is denoted as $F_{trans}^{3} \epsilon \mathbb{R}^{B\times H\times W\times C}$, which contains global semantic information but has relatively coarse spatial details. Here, B denotes the batch size. Downsampling and channel adjustment are implemented through BN, 3×3 convolution, and RPE operations. The local feature of Stage 3 in the CNN branch is denoted as $F_{cnn}^{3} \epsilon \mathbb{R}^{B\times H\times W\times C}$; derived from the shallow layer of the model, it retains details such as edges and textures but has weak semantic information. Through 3×3 convolution, BN, and ReLU operations, the resolution and number of channels of this local feature are adjusted to match those of the preprocessed global feature. The normalized input feature $Norm(F_{trans}^{3})$ is then linearly projected to produce the Query (Q), Key (K), and Value (V):

$$Q,K,V=Linear(Norm(F_{trans}^{3})) \tag{6}$$

Formal Definition of Grouping and Cascading Strategy: Let the input feature map be $X \epsilon \mathbb{R}^{N\times C}$, where N = H × W is the number of tokens. We define the total number of attention heads as H_{total} and the feature dimension per head as $d = C/H_{total}$. The grouping strategy partitions the heads into G parallel groups. The number of heads within each group, denoted as $H_g = H_{total}/G$, defines the cascade depth. In our implementation, $H_{total} = 32$ and G=2. This configuration results in a deep cascading chain ($H_g = 16$) within each group, significantly enhancing the feature hierarchy and nonlinear fitting capability, while the parallel groups ensure diversity in attention modeling. Let $Q_{g,k}, K_{g,k}, V_{g,k}$ denote the query, key, and value for the k-th head in the g-th group (kϵ {1,...,H_g}). We define $O_{g,k} \epsilon \mathbb{R}^{N\times d}$ as the output of the k-th attention head in the g-th group. The hierarchical interaction is mathematically defined as: first, Intra-group Cascading. For depth k>1, the query incorporates the output $O_{g,k-1}$ from the preceding head in the same group.

$$Q'_{g,k} = Q_{g,k} + Conv_{1\times 1}(O_{g,k-1}) \tag{7}$$

Then, Inter-group Crossing: For group g>1, the query interacts with the aligned head from the previous group:

$$Q''_{g,k} = Q'_{g,k} + Conv_{1\times 1}(O_{g-1,k}) \tag{8}$$

For brevity, we only present the update for Q; the key K and value V are updated using analogous cascading and crossing mechanisms to maintain feature consistency. This formulation explicitly creates a dependency chain of length H_g (cascade depth) while maintaining G parallel optimization paths. Linear Complexity Analysis: The standard Multi-Head Self-Attention (MSA) computes the attention map $A \epsilon \mathbb{R}^{N\times N}$ explicitly:

$$MSA(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{9}$$

This requires computing the N×N similarity matrix, leading to a computational complexity of $O(N^2d)$ and memory complexity of $O(N^2)$, which is prohibitive for high-resolution remote sensing imagery. To achieve linear complexity, our GCC-MSA reformulates the attention computation using the Linear Attention mechanism based on the associative property of matrix multiplication. By replacing the Softmax with a kernel feature map $\emptyset(\cdot) = Softplus(\cdot)$ and utilizing the property $(Q \cdot K^T) \cdot V = Q \cdot (K^T \cdot V)$, the kernelized attention is defined as:

$$GCC - MSA(Q,K,V) = \frac{\emptyset(Q)\cdot(\emptyset(K)^T V)}{\emptyset(Q)\cdot\sum_{j=1}^{N}\emptyset(k_j)^T} \tag{10}$$

This reformulation decomposes the computation into two steps. First, Global Context Aggregation, we first compute the global context matrix $M = \emptyset(K)^T V \epsilon \mathbb{R}^{d\times d}$ with a complexity of $O(Nd^2)$. Second, Feature Redistribution, we then multiply the query with the context matrix: $O = \emptyset(Q)M \epsilon \mathbb{R}^{N\times d}$ with a complexity of $O(Nd^2)$. Consequently, the total computational complexity of GCC-MSA is $O(Nd^2)$. Since the head dimension d is a fixed constant, the complexity is strictly linear with respect to the number of tokens N, effectively solving the scalability issue. The LGC module is defined as a learnable function $F_{cal} = \emptyset(F_l, F_g)$ that recalibrates global semantic features using local inductive biases. Given local CNN features $F_l = F_{cnn}^{(i)} \epsilon \mathbb{R}^{H\times W\times C_l}$) and global Transformer features $F_g = f_{GCC-MSA}\left(F_{trans}^{(i-1)}\right)$ $\epsilon \mathbb{R}^{H\times W\times C_g}$), the calibration process is decomposed into the following stages. First, Alignment and Refinement, F_l is first aligned in resolution and channel dimension to match F_g, followed by a depthwise convolution to extract refined local cues:

$$F_l^{ref} = \sigma(BN(DWConv_{3\times 3}(Align(F_l)))) \tag{11}$$

Second, Feature Coupling and Normalization, the features are concatenated and subjected to BN to ensure a stable distribution before attention modeling:

$$F_{cat} = BN(Concat(F_g, F_l^{ref})) \tag{12}$$

Finally, Attention-based Re-fusion, to address spatial-channel coupling, we apply CoAt followed by a dimensionality reduce and a final residual-like fusion:

$$F_{red} = Reduce(CoAt(F_{cat})) \tag{13}$$

$$F_{cal} = \varphi_{proj}(Concat(F_g, F_{red})) \tag{14}$$

The explicit use of BN after concatenation prevents internal covariate shift. $\varphi_{proj}(\cdot)$ and reduce are implemented as 1×1 convolutions with BN, which provide learnable parameters to optimize the feature interaction.

Training stability is maintained through Sigmoid gating in CoAt and Value Clipping, effectively preventing the vanishing or exploding gradient problems common in hybrid architectures. The MLP is used to enhance the non-linear expression of features, and it expands the receptive field through depthwise convolution and dilated convolution to capture long-range spatial dependencies. Finally, the output is enhanced through two residual connections, while the attention map is retained as the input for SGSPP and GEP.

Multi-Dimension Feature Enhancement

MFAM is designed to address scale discrepancies among features from different hierarchical levels. Its structure is illustrated in Fig. 5.

Feature maps from deep layers contain rich semantic information but have low spatial resolution, whereas shallow-layer features preserve higher spatial resolution and more spatial details but weaker semantics. To eliminate channel discrepancies, four 1×1 convolution layers are used to unify the channel dimensions of the four input features ($F^1_{trans}, F^2_{trans}, F^3_{trans}, F^4_{trans}$). To fuse global and local features, the spatial resolution of high-level semantic features is first aligned with lower-level features through upsampling. Starting from the highest-level feature F^4_{trans}, hierarchical fusion is performed progressively with lower-level features. Specifically, F^4_{trans} is upsampled and fused with F^3_{trans}, followed by fusion with F^2_{trans}, and finally with F^1_{trans}, producing the cross-layer fused feature ($F^4_{trans} + F^3_{trans} + F^2_{trans} + F^1_{trans}$). The fused features are further refined by convolution and upsampling operations. After resolution alignment, the processed features from different levels are denoted as p3,p2,p1, and p0. All features are upsampled to the same spatial resolution and aggregated to produce the final multi-scale fused feature map.

CoAt is a mechanism that combines Channel Attention and Spatial Attention. Its structure is illustrated in Fig. 6.

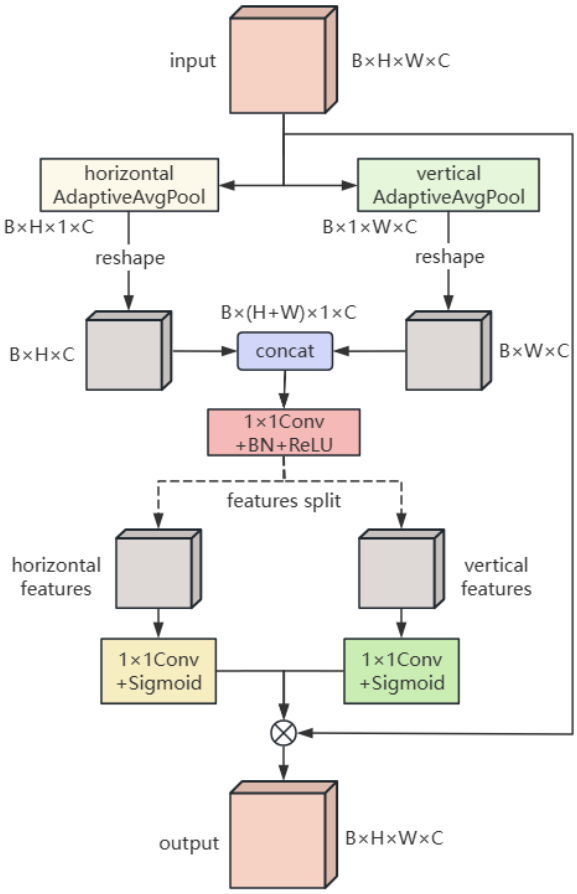

Fig. 6. *Coordinate Attention.*

To address the limitations of conventional channel attention in preserving spatial details, we introduce CoAt, which reformulates the feature aggregation process to achieve improved spatial-channel coupling. Conventional channel attention mechanisms, such as the SE block, typically employ Global Average Pooling (GAP) to encode channel-wise statistics. For an input feature map $X \epsilon \mathbb{R}^{H \times W \times C}$, the squeezed feature z_c for the c-th channel is calculated as:

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \qquad (15)$$

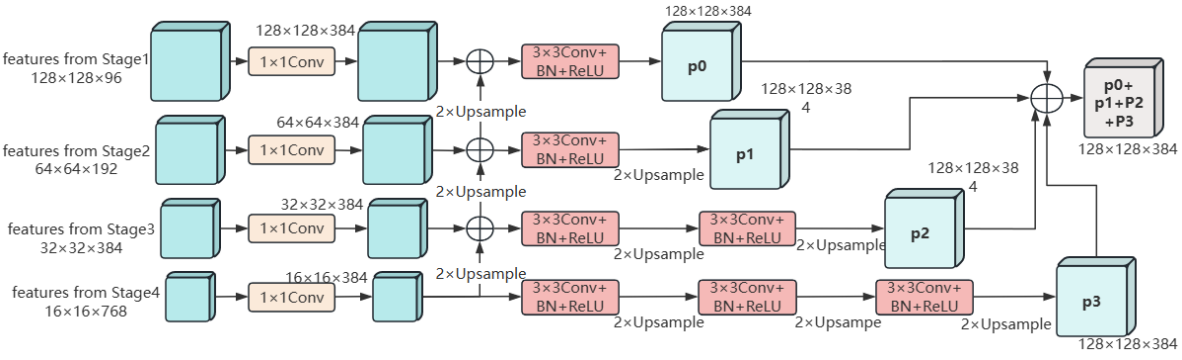

Fig. 5. *Multi-scale Feature Aggregation Module.*

While this operation captures global channel correlations via $A_c = \sigma(W_2\delta(W_1 z))$($\delta(\cdot)$ ReLU, $\sigma(\cdot)$ Sigmoid, and W_1, W_2 fully-connected weights, it compresses spatial information into a single scalar, losing spatial structure, suboptimal for dense segmentation requiring precise boundary localization. In contrast, CoAt decomposes global pooling into two orthogonal coordinatewise encoding operations, capturing long range dependencies along one direction while preserving positional information along the other. Specifically, for the input $X \in R^{B\times H\times W\times C}$, pooling kernels (H,1) and (1,W) are applied to encode vertical and horizontal features, respectively. The output of the c-th channel at height h and width w is formulated as:

$$z_c^h(h) = \frac{1}{w}\sum_{0\le j<w} x_c(h,j), \quad z_c^w(w) = \frac{1}{H}\sum_{0\le i<H} x_c(i,w) \tag{16}$$

These operations generate two direction-aware feature maps $Z^h \epsilon \mathbb{R}^{H\times 1\times C}$ and $Z^w \epsilon \mathbb{R}^{1\times W\times C}$, which preserve spatial coordinates. To model interactions, the features are concatenated along the spatial dimension to form $[Z^h, Z^w] \epsilon \mathbb{R}^{(H+W)\times 1\times C}$, and then processed by a shared 1×1 convolution F_{conv} followed by BN and a non-linear activation, generating the intermediate feature map f:

$$f = \delta(BN(F_{conv}([z^h, z^w]))) \tag{17}$$

Where $[\cdot,\cdot]$ denotes concatenation, and $f \epsilon \mathbb{R}^{(H+W)\times 1\times \frac{C}{r}}$ with reduction ratio r(r = 16). The feature f is split along the spatial dimension into $f^h \epsilon \mathbb{R}^{H\times 1\times \frac{C}{r}}$ and $f^w \epsilon \mathbb{R}^{1\times W\times \frac{C}{r}}$, which are transformed by two independent 1×1 convolutions F_h and F_w, followed by the Sigmoid activation σ, to generate attention weights:

$$g^h = \sigma(F_h(f^h)), \quad g^w = \sigma(F_w(f^w)) \tag{18}$$

The final output is obtained by re-weighting the input feature map using the outer product of the horizontal and vertical attention weights:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{19}$$

The term $g_c^h(i) \times g_c^w(j)$ forms a spatial-channel coupling mechanism. Unlike SE, which applies a uniform weight across the channel plane, CoAt produces a spatially varying weight map, enabling the model to suppress background noise and highlight building regions with precise coordinate awareness.

SGSPP is a multi-scale semantic feature enhancement module for aerial image buildings. Its structure is illustrated in Fig. 7.

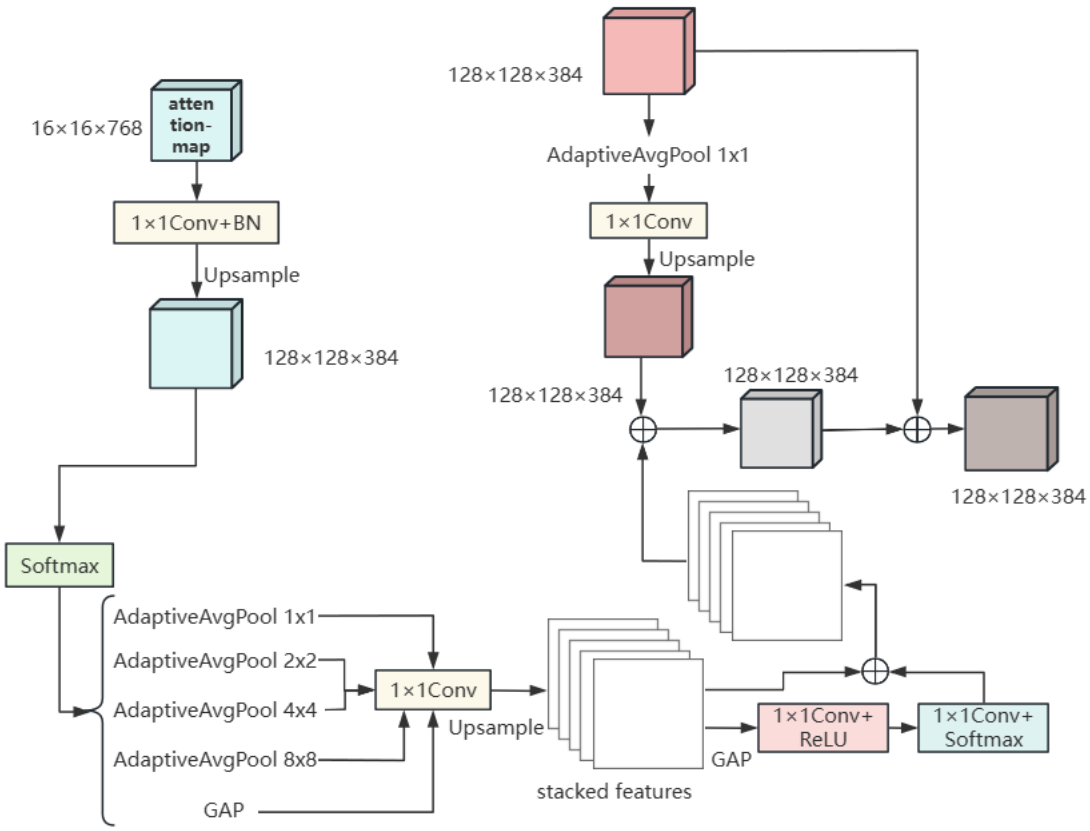

Fig. 7. *Semantic-Guided Spatial Pyramid Pooling.*

The SGSPP module enhances multi-scale semantic representation by filtering background noise and focusing on high-semantic regions. It takes two inputs: the CoAt-enhanced features and the semantic attention map from Stage 4 of the Transformer branch. The process includes three stages: semantic masking, multi-scale feature generation, and adaptive scale selection. First, the semantic attention map is processed with a 1×1 convolution to reduce the channel dimension from 768 to 384, aligning it with the input feature dimension. A Spatial Softmax normalization is then applied to improve spatial discriminability. Based on the normalized map, a binary mask is generated by retaining the top 30% of attention responses. This threshold is selected according to the typical building pixel occupancy rate in aerial imagery, enabling the module to emphasize potential building regions while suppressing most background noise. The resulting mask is applied to the input features to filter irrelevant responses. Adaptive average pooling is then performed on the masked features at four scales (1×1,2×2,4×4,8×8) to capture targets of different sizes. In addition, a GAP branch is applied to the original unmasked input to preserve global contextual information. All pooled features are then upsampled to the original resolution (128×128) via bilinear interpolation after channel adjustment.

Let $P = \{P_k\}_{k=1}^{K}$ denote the set of feature maps obtained from these K scales (where K=5 in our implementation, including four multi-scale pooling branches and one global context branch), where each $P_k \epsilon \mathbb{R}^{H\times W\times C}$ shares the same spatial resolution. Unlike conventional SPP modules that fuse features through heuristic summation or concatenation, we introduce an Adaptive Scale Selection (ASS) mechanism to dynamically assign scale weights according to the input content. The fused output PPP is defined as a convex combination of the multi-scale features:

$$P = \sum_{k=1}^{K} \alpha_k \cdot P_k \tag{20}$$

Subject to the normalization constraint:

$$\sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k \epsilon\, [0,1] \tag{21}$$

The adaptive weights α_k are generated through a channel-wise attention network. First, multi-scale features are aggregated to obtain a global descriptor $U = \sum_{k=1}^{K} P_k$. A GAP operation then compresses the spatial dimensions, followed by an MLP that produces the scale-selection score vector $z \epsilon \mathbb{R}^{K}$:

$$z = W_2 \cdot \delta(W_1 \cdot GAP(U)) \tag{22}$$

Where $W_1 \epsilon \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \epsilon \mathbb{R}^{K\times\frac{C}{r}}$ are the learnable weights of the reduction and expansion layers with reduction ratio r(r=4), and δ denotes the ReLU activation function. The final scale weights are obtained by applying the Softmax function along the scale dimension:

$$\alpha_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \tag{23}$$

Where z_k denotes the k-th element of the score vector z. This normalization ensures that the network adaptively emphasizes the most informative spatial scales. Finally, the fused feature PPP is added to the original input through a residual connection to prevent feature degradation.

Boundary-Aware Decoding and Decision

GEP is a key module designed to enhance building edge features in aerial images. Its structure is illustrated in Fig. 8.

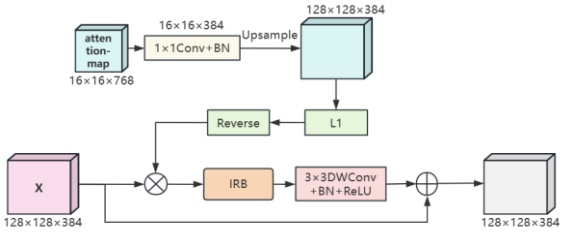

Fig. 8. *GCC-MSA-guided Edge Perception.*

First, the semantic attention map is preprocessed with a 1×1 convolution to reduce channels from 768 to 384, followed by BN and spatial alignment. The adjusted map is L1-normalized along channels to obtain a stable weight distribution, then inverted so that originally high-response interior regions become low and edge regions receive amplified weights, thus prioritizing edges. The reversed map weights the input features to select edge responses, which are refined by a single shallow IRB (chosen to preserve high-frequency details and avoid over-smoothing that deep stacks would cause). A subsequent deep convolution further strengthens the edge representation, and the enhanced edge features are merged with the original input via a residual connection to retain global semantics while adding boundary detail. Optimization of these edge features is guided by the Boundary-Aware Hybrid Loss (Focal + Tversky), which places stronger penalty on high-frequency boundary errors where the GEP operates.

The core function of SPGM is to adaptively balance the outputs of the Transformer and CNN branches. Its structure is illustrated in Fig. 9.

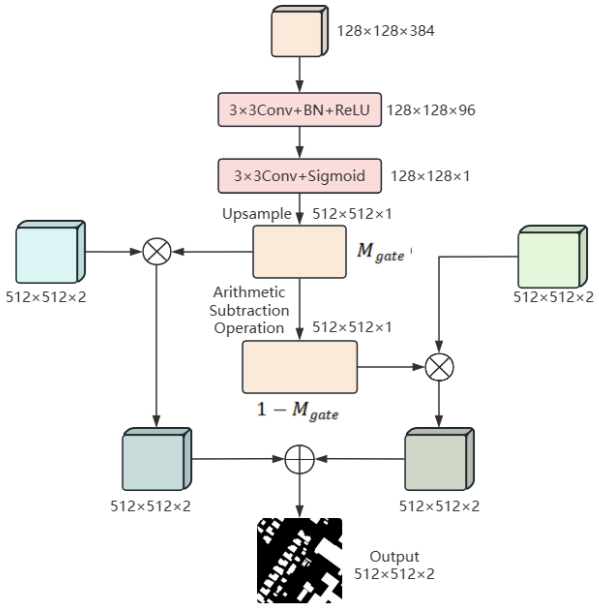

Fig. 9. *Spatial Perception Gating Mechanism.*

SPGM learns a spatially adaptive weight map from the detailed features of the CNN branch to guide the fusion of dual-branch predictions. The module takes two types of inputs: features used for weight generation and prediction results to be fused. Weight generation is based on the detail path features. A 3×3 convolution first adjusts channel dimensions while preserving the 128×128 spatial resolution, capturing local spatial patterns and retaining edge and texture information. BN and ReLU are then applied to stabilize feature distribution and enhance important responses. Another 3×3 convolution compresses the multi-channel features into a single-channel spatial weight map while maintaining the 128×128 resolution. A Sigmoid activation normalizes the weights to [0,1], after which the map is upsampled to 512×512 for spatial alignment.

$$\hat{Y} = M_{gate} \odot \varphi_{trans}(F_{refined}) + (1 - M_{gate}) \odot \varphi_{cnn}(F_{cnn}^{final}) \quad (24)$$

Where $\hat{Y}$ denotes the final prediction map, φ_trans $\phi_{\text{trans}}(\cdot)$ and $\phi_{\text{cnn}}(\cdot$ represent the prediction heads of the Transformer and CNN branches, respectively. $M_{gate} \in \mathbb{R}^{H \times W \times 1}$ is the spatial weight map generated by SPGM, and $\odot$ denotes element-wise multiplication. The term $1 - M_{gate}$ provides the complementary weight for the CNN branch.

Experimental Setup

Experimental Environment

Experiments were implemented in PyTorch 1.10.0 with Python 3.7, running on CUDA 11.3 and cuDNN 8.2.1; model training and inference used an NVIDIA GeForce RTX 4090 GPU. We compare the proposed method against several mainstream segmentation frameworks: SegNet (Badrinarayanan , 2017), U-Net, U-Net++ (Zhou , 2018), DeepLabv3+, and PSPNet on the WHU Aerial, Massachusetts, and GF-7 building datasets. SegNet, U-Net and U-Net++ are standard encoder-decoder architectures that progressively integrate local and global features during upsampling. PSPNet employs pyramid pooling for multi-scale context aggregation, while DeepLabv3+ combines atrous (dilated) convolutions with spatial pyramid pooling to fuse multi-scale features.

Training Configuration

All networks used identical training settings (same data splits and the Boundary-Aware Hybrid Loss) for fair comparison. We trained on a single GPU for 100 epochs with batch size 4, using the Lookahead+AdamW optimizer (initial learning rate 1×10^(-4), weight decay 0.0025) and CosineAnnealingWarmRestarts learning-rate schedule. To improve generalization, we applied DropPath (rate 0.3) and Dropout (rate 0.1 in the prediction head). Model selection was performed using the checkpoint with the highest validation IoU to avoid degradation from over-training.

Boundary-Aware Hybrid Loss Function

To provide strong supervision for GEP, we employ a Boundary-Aware Hybrid Loss that combines Focal Loss and Tversky Loss to emphasize hard boundary pixels and improve structural consistency:

$$L = \lambda_1 L_F + \lambda_2 L_T \quad (25)$$

Here L_F (Focal Loss) addresses class imbalance by down-weighting easy examples and focusing learning on hard samples, while L_T (Tversky Loss) flexibly balances penalties for false negatives and false positives to better preserve edges. We evaluated five weight pairs (λ_1, λ_2): (0.6, 0.4), (0.8, 0.2), (1.0, 1.0), (1.2, 0.8), (1.4, 0.6), and found $\lambda_1 = \lambda_2 = 1.0$ gives the best validation performance.

Focal Loss is defined as:

$$L_F = -\frac{1}{N}\sum_{i=1}^{N}(\alpha_F(1-p_i)^{\gamma} y_i log(p_i) + (1 - \alpha_F) p_i^{\gamma}(1-y_i) log(1-p_i)) \quad (26)$$

Where y_i and p_i are the ground truth and predicted probability for pixel i, N is the number of pixels, α_F is the class-balance factor, and γ(set to 2.0) is the focusing parameter. The modulating factor $(1-p_i)^{\gamma}$ reduces the

contribution of well-classified pixels, directing optimization toward uncertain boundary pixels.

Tversky Loss is given by:

$$L_T = 1 - \frac{\sum_{i=1}^{N} p_i\, y_i + \varepsilon}{\sum_{i=1}^{N} y_i p_i + \alpha_T \sum_{i=1}^{N} p_i(1 - y_i) + \beta_T \sum_{i=1}^{N} (1 - p_i) y_i + \varepsilon} \quad (27)$$

Where α_T and β_T weight missed detections and false detections respectively. We set $\alpha_T = 0.3$ $\beta_T = 0.7$ (slightly favoring recall to better capture thin boundaries), which empirically improves boundary alignment compared with standard IoU losses.

Evaluation Metrics

Precision, Recall, IoU, and F1-score are used to evaluate the network performance, and their formulas are defined as follows:

$$Precision = \frac{TP}{TP + FP} \quad (28)$$

$$Recall = \frac{TP}{TP + FN} \quad (29)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (30)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (31)$$

Among them, TP and FP are used to evaluate the positive sample set, while TN and FN are used to evaluate the negative sample set. While Precision and Recall measure the accuracy of positive predictions and the completeness of target detection respectively, F1-score and IoU serve as integrated indicators to evaluate the overall balanced performance and spatial overlap accuracy, ensuring a multi-dimensional assessment of the segmentation results.

RESULTS

Experimental Results and Analysis on the WHU Aerial Building Dataset

We evaluated six models, U Net, U Net++, DeepLabv3+, PSPNet, SegNet, and the proposed method on the WHU Aerial Building Dataset. As shown in Fig. 10, all networks capture the basic building outlines, yet they exhibit marked differences in fine details. For scattered small buildings (Fig. 10(a)), comparative models suffer from false and missed detections due to insufficient small object perception. In contrast, our method extracts complete buildings with well preserved edges, benefiting from its dedicated detail path and multi scale fusion. Under strong shadows (Fig. 10(b)), U Net and DeepLabv3+ produce discontinuous contours; our method (and U Net++) mitigates this via CoAt for precise spatial localization and LGC for enhanced local details, enabling context aware contour completion. When background objects resemble buildings (Fig. 10(c)), competing models introduce false alarms, while our approach suppresses such interference by focusing on key building features through CoAt. In shadow dominated areas (Fig. 10(d)), PSPNet misclassifies shadows, DeepLabv3+ shows missing edges, and U Net++ exhibits local omissions. Our model eliminates shadow artifacts using SGSPP, which selects high importance semantic regions, suppresses noise, and retains global context via multi scale pooling. For irregular building shapes (Fig. 10(e-f)), U Net and DeepLabv3+ miss some details and PSPNet fails to capture curved or polygonal boundaries. Our method accurately restores these irregular contours, closely matching the ground truth. Overall, the proposed method consistently outperforms the compared approaches across all challenging scenarios, demonstrating superior robustness to scale variation, illumination change, background clutter, and complex geometry.

Quantitative results (Table 1) show that our method achieves the best overall performance, leading in IoU, F1 score, and recall. This demonstrates its ability to accurately and comprehensively identify building regions while maintaining a low false positive rate. U Net++ attains the highest precision, benefiting from its nested encoder decoder structure and deep supervision that suppress false detections. However, its relatively lower recall limits the final IoU, likely due to insufficient sensitivity to small or partially occluded buildings. Conversely, the baseline U Net achieves the highest recall but lowest precision, indicating a tendency to over detect building pixels at the cost of increased false positives. PSPNet and DeepLabv3+ incorporate pyramid pooling and dilated convolutions for multi scale context, yielding balanced metrics. Nevertheless, their feature fusion strategies still struggle to preserve fine details, resulting in only moderate performance gains.

Table 1. *Performance Comparison of Various Models on the WHU Aerial Building Dataset.*

Method	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
U-Net	92.83	94.82	93.81	88.35
U-Net++	**96.59**	94.08	95.32	91.06
DeepLabv3+	95.47	93.96	94.71	89.95
PSPNet	95.31	94.72	95.01	90.50
SegNet	94.28	93.74	94.01	88.71
Ours	95.96	**96.06**	**96.01**	**92.33**

The bold values indicate that highest values within the corresponding evaluation indices.

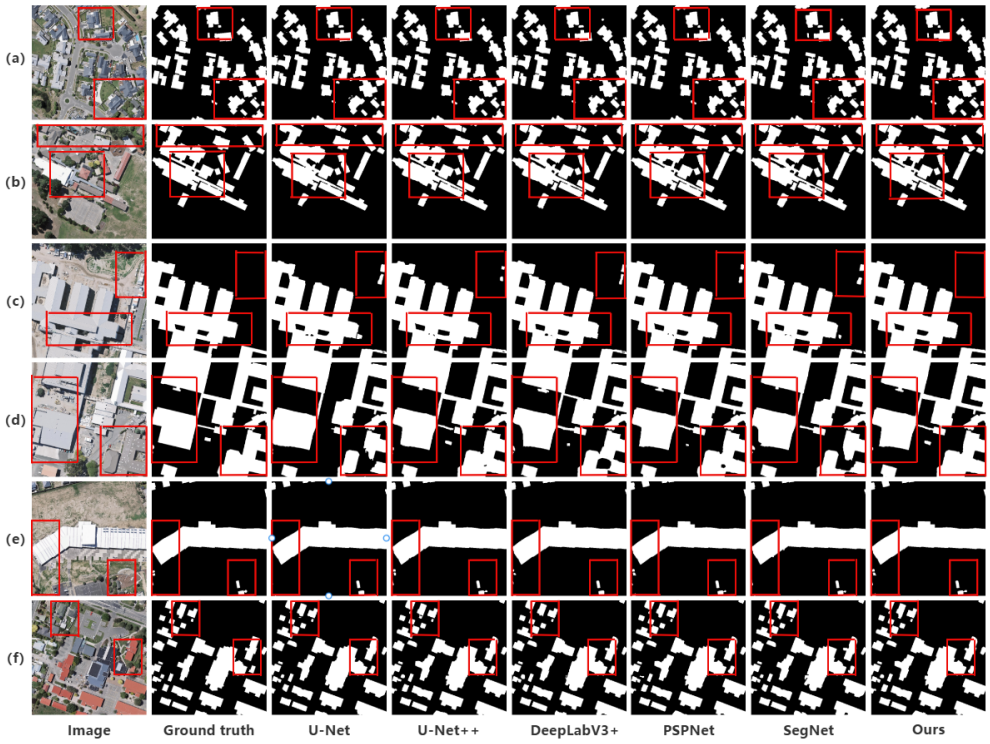

Fig. 10. *Extraction Results on the WHU Aerial Building Dataset.*

Experimental Results and Analysis on the Massachusetts Building Dataset

The six models were evaluated on the Massachusetts Building Dataset (Fig. 11). Our method achieves the best overall extraction performance, followed by DeepLabv3+, while PSPNet performs relatively weaker. Specifically, it extracts small buildings more completely with regular boundaries via MFAM's multi level fusion (Fig. 11(a-b)), restores connectivity in occluded or shadowed regions through LGC and SPGM (Fig. 11(c)), accurately separates adjacent buildings and reduces false detections via multi scale context modeling (Fig. 11(d)), and improves detection of small and low rise buildings using the dual branch structure (Fig. 11 (e-f)). Overall, the proposed method outperforms PSPNet, U Net, U Net++, and DeepLabv3+ in detail preservation, boundary clarity, and robustness to interference.

Quantitative results (Table 2) show that our method achieves the best overall performance, with precision, recall, F1 score, and IoU reaching 88.37%, 86.71%, 87.53%, and 77.81%, respectively, substantially outperforming all compared models. U Net and U Net++ attain relatively high recall, demonstrating the effectiveness of their encoder decoder structures in capturing building regions. However, their lower precision indicates a tendency to misclassify shadows and textures as buildings, likely due to dataset challenges such as occlusions and varied roof textures. DeepLabv3+ achieves relatively high precision, benefiting from atrous convolutions and spatial pyramid pooling for large scale context, yet its recall remains limited, reflecting insufficient sensitivity to occluded or small buildings. PSPNet exhibits the most balanced but modest performance; its pyramid pooling module integrates multi scale context but adapts poorly to boundary ambiguity. Compared to PSPNet, our method improves precision by 6.27%, recall by 7.48%, F1 score by 6.89%, and IoU by 10.25%, validating its strong adaptability and robustness in complex scenarios

Table 2. *Performance Comparison of Various Models on the Massachusetts Building Dataset.*

Method	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
U-Net	80.18	82.41	81.28	68.46
U-Net++	79.26	83.47	81.31	68.51
DeepLabv3+	86.66	81.10	83.79	72.10
PSPNet	82.10	79.23	80.64	67.56
SegNet	82.61	84.26	83.43	71.58
Ours	**88.37**	**86.71**	**87.53**	**77.81**

The bold values indicate the highest values for the corresponding evaluation indices.

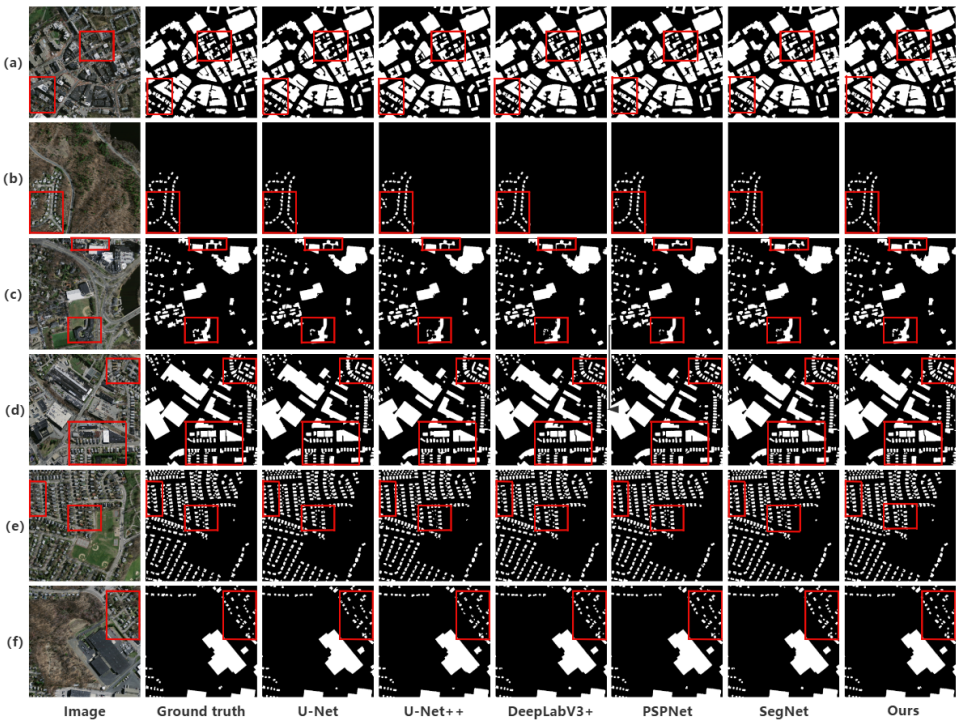

Fig. 11. *Extraction Results on the Massachusetts Building Dataset*.

Experimental Results and Analysis on the GF-7 Building Dataset

We further evaluated the six models on the GF-7 Building Dataset (Fig. 12). The proposed method achieves the best building extraction performance, followed by PSPNet and U Net++, while SegNet yields relatively weaker results. In Fig. 12(a), our method attains superior target integrity and edge accuracy through GCC MSA for global modeling and LGC for detailed information. Under large area shadows (Fig. 12(b)), all models exhibit omissions, but ours remains relatively robust. In complex dense regions (Fig. 12(c)) and against interfering textures (Fig. 12(d)), the dual path design effectively preserves details and reduces false detections. For boundary precision against similar objects (Fig. 12(e)), the GEP module enhances edge features, yielding the best detail extraction. In dense areas (Fig. 12(f)), our method excels in error control and detail preservation, outperforming U Net and SegNet, which show fragmentation or noise. Overall, the proposed approach demonstrates enhanced robustness and precision across all challenging scenarios.

Quantitative results (Table 3) show that the proposed method achieves the best performance on the GF-7 Building Dataset, attaining optimal values across precision, recall, F1 score, and IoU. This demonstrates its ability to realize more accurate and complete building region recognition. SegNet exhibits the weakest performance across all metrics, as its basic encoder decoder structure struggles with cloud shadow interference and distinguishing rural buildings from vegetation. U Net shows relatively low recall, indicating limited capability in detecting shadow covered or small scale buildings. U Net++ achieves high precision, but its recall and IoU lag substantially, suggesting that its nested deep supervision, while suppressing false detections, may also filter out real building targets, particularly rural low rise buildings with weak features. DeepLabv3+ and PSPNet display relatively balanced performance through multi scale context modeling. However, when faced with complex interference from overlapping cloud and building shadows, their convolution based methods have limited capacity for modeling long range dependencies, creating a performance bottleneck. In contrast, our hybrid architecture effectively addresses these challenges.

Table 3. *Performance Comparison of Various Models on the GF-7 Building Dataset.*

Method	Precision(%)	Recall(%)	F1-score(%)	IoU(%)
U-Net	84.13	81.88	82.97	70.92
U-Net++	89.12	82.12	85.48	74.64
DeepLabv3+	86.73	82.54	84.55	73.28
PSPNet	87.55	82.86	85.13	74.13
SegNet	85.94	82.19	84.01	72.45
Ours	**91.83**	**84.15**	**87.82**	**78.29**

The bold values indicate the highest values for the corresponding evaluation indices.

Fig. 12. *Extraction Results on the GF-7 Building Dataset.*

Comparison With State-of-the-Art Methods

We further compared our method with several state of the art approaches on the WHU Aerial and Massachusetts building datasets, including TCNet (Xiang , 2024), EU Net (Kang , 2019), BuildFormer (Wang , 2022), DE Net (Liu , 2019), MA FCN (Shrestha and Vanneschi 2018), BOMSC Net (Zhou , 2022), HD Net (Li , 2024), MAP Net, DC Swin (Wang , 2022), and the baseline EViT. For the WHU Aerial Building Dataset (Table 4). Our method achieves the best comprehensive performance, ranking first in all four metrics. Compared with EViT, it improves precision by 0.11%, recall by 0.53%, F1 score by 0.32%, and IoU by 0.57%, demonstrating stable performance optimization while maintaining high accuracy. Among other methods, TCNet ranks second in recall but its F1 score is substantially lower due to metric imbalance. BuildFormer, MAP Net, and DC Swin exhibit relatively balanced metrics, yet still lag behind EViT and our approach.

For the Massachusetts Building Dataset (Table 5), our method ranks first in precision, F1 score, and IoU, and second in recall. Compared with EViT, it improves precision by 0.40%, recall by 0.47%, F1 score by 0.43%, and IoU by 0.67%, demonstrating steady gains in both accuracy and completeness. Among other methods, TCNet achieves the highest recall but lowest precision, resulting in substantially lower F1 score and IoU. BuildFormer, HD Net, and DC Swin show relatively balanced metrics but still underperform overall.

Table 4. *Quantitative Comparison with State-of-the-Art Methods on the WHU Aerial Building Dataset.*

Method	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
TCNet	95.15	95.55	93.95	91.16
EU-Net	94.98	95.10	95.04	90.56
Build-Former	95.65	95.40	94.97	91.44
DE-Net	95.16	94.79	94.98	90.36
MA-FCN	95.20	95.10	95.15	90.70
BOMSC-Net	95.14	94.50	94.80	90.15
HD-Net	95.00	94.68	94.84	90.19
MAP-Net	95.62	94.81	95.21	90.86
DC-Swin	95.68	95.24	95.46	91.32
EViT	95.85	95.53	95.69	91.76
Ours	**95.96**	**96.06**	**96.01**	**92.33**

The bold values indicate the highest values for the corresponding evaluation indices.

For the GF-7 Building Dataset (Table 6), our method achieves the best comprehensive performance. Compared with EViT, it improves precision by 0.81%, recall by 0.21%, F1-score by 0.47%, and IoU by 0.74%, demonstrating consistent gains in both accuracy and

completeness. Among other methods, BuildFormer attains the highest recall (84.60%) but its precision and IoU are lower than ours, resulting in an overall inferior performance. EU-Net, BOMSC-Net, and DC-Swin show relatively balanced metrics, yet still lag behind our approach.

Table 5. *Quantitative Comparison with State-of-the-Art Methods on the Massachusetts Building Dataset.*

Method	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
TCNet	85.17	**86.82**	84.29	76.21
EU-Net	86.70	83.40	85.01	73.93
BuildFormer	87.52	84.90	86.19	75.74
MA-FCN	87.07	82.89	84.93	73.80
BOMSC-Net	86.64	83.68	85.13	74.71
HD-Net	85.98	86.13	86.06	75.53
MAP-Net	85.21	81.28	83.20	71.23
DC-Swin	86.74	84.97	85.86	75.22
EViT	87.97	86.24	87.10	77.14
Ours	**88.37**	86.71	**87.53**	**77.81**

The bold values indicate the highest values for the corresponding evaluation indices.

Table 6. *Quantitative Comparison with State-of-the-Art Methods on the GF-7 Building Dataset.*

Method	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
TCNet	89.92	82.37	85.99	75.42
EU-Net	90.10	82.85	86.32	75.95
Build-Former	90.45	**84.60**	87.44	77.69
MA-FCN	89.35	81.96	85.47	74.63
BOMSC-Net	90.28	83.11	86.55	76.28
HD-Net	89.78	83.54	86.53	76.25
MAP-Net	88.94	81.32	84.96	73.87
DC-Swin	90.47	83.68	86.94	76.88
EViT	91.02	83.94	87.35	77.55
Ours	**91.83**	84.15	**87.82**	**78.29**

The bold values indicate the highest values for the corresponding evaluation indices.

Model Efficiency Analysis

We evaluate computational cost using parameters (Params) and FLOPs on the Massachusetts Building Dataset (Table 7). Compared with the baseline EViT, our method reduces parameters by 2.19M and FLOPs by 17.26G, while improving IoU by≈0.70%. Against heavier architectures, our model requires 19.89M fewer parameters than DeepLabv3+ yet achieves 5.71% higher IoU; it is also lighter and more accurate than PSPNet. Overall, the proposed method strikes a favorable balance, minimizing computational burden without excessive parameter growth while remaining competitive with advanced approaches.

To address whether the proposed architectural complexity is necessary, we further compare our method with three representative relatively lightweight segmentation networks: PIDNet (Xu , 2023), SwiftFormer (Shaker , 2023), and MobileNetV4 (Qin , 2024). The results are reported in Table 8. As shown, these lightweight models achieve IoU scores between 59.94% and 62.67%, which are 15.14-17.87 percentage points lower than ours (77.81%). Their Precision and Recall are also substantially inferior. While these networks have significantly fewer parameters and lower FLOPs, their limited capacity fails to capture fine building boundaries, detect small structures, and suppress shadow/background interference, which are critical requirements for high-resolution remote sensing building extraction. For precision-demanding applications such as urban planning and disaster response, such large accuracy gaps are unacceptable. Therefore, the proposed complexity is well-justified by the task requirements and represents a favorable tradeoff within the high-accuracy regime.

Statistical Significance Verification

To verify that the observed IoU improvements (0.57-0.67% in Tables 4 and 5) stem from architectural en-

Table 7. *Efficiency Comparison of Different Methods on the Massachusetts Building Dataset.*

Method	Precision(%)	Recall(%)	F1-score(%)	IoU(%)	FLOPs(G)	Params(M)
PSPNet	82.10	79.23	80.64	67.56	256.62	67.95
DeepLabv3+	86.66	81.10	83.79	72.10	254.53	62.57
EU-Net	86.70	83.40	85.01	73.93	135.97	60.28
HD-Net	85.98	86.13	86.06	75.53	**67.56**	58.11
BOMSC-Net	86.64	83.68	85.13	74.71	73.75	46.37
BuildFormer	87.52	84.90	86.19	75.74	117.12	**40.52**
EViT	87.97	86.24	87.10	77.14	141.61	44.87
Ours	**88.37**	**86.71**	**87.53**	**77.81**	124.35	42.68

The bold values indicate the highest values for the corresponding evaluation indices.

hancements rather than stochastic variation, we conducted hypothesis testing ($H_0: E[IoU_{Ours}] = E[IoU_{baseline}]$). While Tables 4 and 5 report peak performance, here we present statistics from ten independent runs with different random seeds. As shown in Table 9, mean IoU scores closely match the peak values, with low standard deviations (0.08 and 0.10), indicating robustness to initialization. Paired t-tests yield $p < 0.05$ for both datasets, rejecting the null hypothesis. These results confirm that the performance gains are systematic and reproducible.

Feature Map Visualization Analysis

Fig. 13 provides a qualitative visualization of feature maps extracted from different stages (Stage 1 to Stage 4) of our proposed network, illustrating the progressive evolution of features. The heatmaps demonstrate a clear transition from local details to global semantics. As shown in columns (b) and (c), the feature maps in the early stages exhibit strong activation responses primarily along building boundaries, corners, and high-frequency texture regions. This behavior is consistent with the characteristics of shallow CNN layers, which specialize in extracting local structural details. In column (d), as the network deepens and the receptive field expands, the activation regions begin to shift from mere boundaries to cover the main bodies of the buildings. The distinction between foreground objects and the background starts to become clearer. Column (e) represents the deepest, most abstract features. Notably, the activations yield highly uniform and complete responses across entire building instances, while irrelevant background areas are effectively suppressed. This strong semantic consistency attests to the crucial role of the Transformer branch and the GCC-MSA module in capturing long-range dependencies and global context.

DISCUSSION

Ablation Study

We conduct ablation experiments on the Massachusetts Building Dataset to validate the importance of the dual branch structure (Table 10). The proposed method achieves optimal performance across all metrics. Removing the CNN branch decreases IoU by 1.27%, while removing the Transformer branch causes a more substantial drop of 2.65%. This confirms that the Transformer branch dominates global context modeling, and its absence more severely impacts performance. The CNN

Table 8. *Comparison with relatively lightweight segmentation networks on the Massachusetts Building Dataset.*

Method	Precision (%)	Recall (%)	F1-score (%)	IoU (%)	FLOPs (G)	Params (M)
PIDNet	76.17	75.84	76.00	61.29	**25.64**	**7.85**
SwiftFormer	77.43	76.68	77.06	62.67	32.45	15.37
MobileNetV4	75.66	74.29	74.96	59.94	36.92	18.43
Ours	**88.37**	**86.71**	**87.53**	**77.81**	124.35	42.68

The bold values indicate the highest values for the corresponding evaluation indices.

Table 9. *Statistical Stability Analysis on WHU Aerial and Massachusetts Building Datasets.*

Dataset	Mean IoU(%)	Std Dev(σ)	95% CI	P-value
WHU Aerial	92.30	0.08	[92.24,92.36]	< 0.001
Massachusetts	77.78	0.10	[77.71,77.85]	< 0.001

The p-value is calculated using a paired t-test against the baseline EViT model results. A p-value < 0.05 indicates statistical significance.

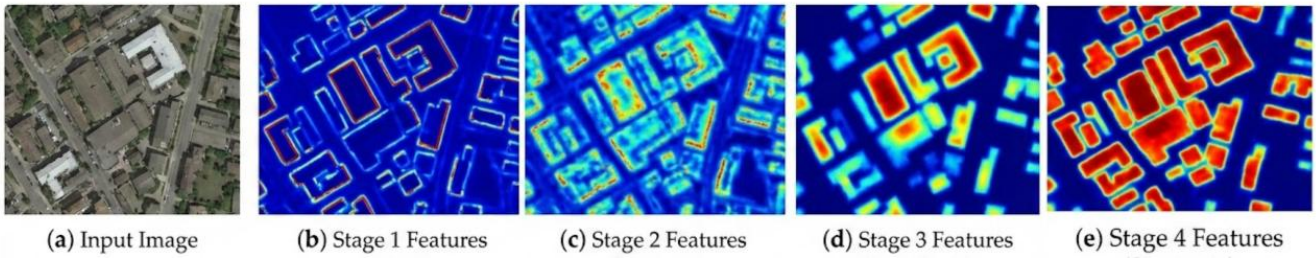

Fig. 13. *Visualization of feature maps for each stage.*

branch serves as a necessary detail supplement, providing local texture and edge information that the Transformer branch lacks. Their synergy yields a balanced tradeoff between precision and recall, demonstrating that the hybrid architecture effectively improves segmentation accuracy by fusing global semantics and local features.

We conducted extensive ablation experiments on the Massachusetts Building Dataset to evaluate each proposed module (Table 11). Removing GCC MSA causes an IoU drop of 0.92%, and replacing it with standard multi head self attention (MSA) further reduces IoU by 1.15%. This underscores the benefit of GCC MSA's intra group cascading and inter group crossing, which enhance feature hierarchy and diversity. Removing CoAt decreases IoU by 0.52%; substituting it with a standard SE block lowers IoU by 0.88%. CoAt's coordinate wise attention preserves spatial details crucial for fine boundaries, whereas SE's global pooling loses positional information. SGSPP contributes a 0.48% IoU gain; replacing it with conventional ASPP reduces IoU by 0.86%. SGSPP's semantic guided multi scale pooling selectively emphasizes important regions while suppressing background, unlike ASPP's fixed dilated convolutions. For fusion, removing SPGM drops IoU by 0.37%, and using simple concatenation instead of adaptive weighting causes a 0.73% decrease. SPGM's pixel wise re weighting effectively bridges the semantic gap between CNN and Transformer branches. LGC adds 0.40% IoU by injecting local CNN details into Transformer stages, alleviating spatial blurring. GEP contributes 0.25% IoU; although modest, it targets hard boundary pixels, improving morphological quality as seen in qualitative results. Overall, each module plays a distinct and necessary role, together achieving state of the art performance.

Impact of Spatial Resolution Discrepancy

The datasets employed in this study exhibit significant variations in spatial resolution, ranging from 0.3 m to 1.0 m. This discrepancy introduces specific challenges and limitations that affect implementation and performance analysis.

The variation in resolution alters the physical scale of objects within the fixed input size (512 × 512). For high-resolution data (0.3 m), the model must focus on

Table 10. *Ablation Study on the Dual-Branch Structure on Massachusetts Building Dataset.*

Method	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
Ours without CNN branch	87.63	85.77	86.69	76.54
Ours with CNN branch	88.37	86.71	87.53	77.81
Ours without Transformer branch	86.95	84.71	85.82	75.16
Ours with Transformer branch	88.37	86.71	87.53	77.81

Table 11. *Ablation Study on the Key Modules on the Massachusetts Building Dataset.*

Method	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
Ours	88.37	86.71	87.53	77.81
Ours without SGSPP	87.96	86.49	87.22	77.33
ASPP(replace SGSPP)	87.93	86.05	86.98	76.95
Ours without SPGM	88.16	86.44	87.29	77.44
fixed (replace SPGM)	87.89	86.23	87.05	77.08
Ours without GEP	88.23	86.50	87.36	77.56
Ours without CoAt	87.88	86.51	87.19	77.29
SE (replace CoAt)	87.96	85.98	86.96	76.93
Ours without LGC	88.21	86.35	87.27	77.41
Ours without GCC-MSA	87.84	86.06	86.94	76.89
MSA (replace GCC-MSA)	87.45	85.77	86.80	76.66

fine-grained boundary reconstruction, whereas for lower-resolution data (1.0 m), it must rely more on global semantic context to resolve blurred edges. Our architecture addresses this through the SGSPP module, which aggregates multi-scale context, allowing the model to adaptively capture features regardless of the resolution. It is important to note that performance metrics should be compared within the same dataset. The experimental results demonstrate that our method consistently outperforms baselines across all resolutions. This confirms that the proposed hybrid architecture possesses strong generalization capabilities, effectively handling both the detailed textures of high-resolution imagery and the semantic ambiguities of low-resolution imagery. Despite these designs, the resolution discrepancy imposes an upper bound on performance. In 1.0 m resolution imagery, the 'mixed pixel' effect at building boundaries is unavoidable, which inherently limits the achievable IoU compared to 0.3 m imagery.

Limitations

While the proposed EViT-based architecture demonstrates strong performance, a few limitations remain for future exploration.

First, regarding the experimental design, our study strictly utilizes the official data splits (training, validation, and testing) provided by standard benchmarks (WHU Aerial, Massachusetts, GF-7). While adhering to these splits is essential to ensure fair comparisons with existing state-of-the-art methods, we acknowledge that remote sensing imagery inherently contains strong spatial autocorrelation. Because nearby patches may share similar textures and roof types, such evaluations can sometimes produce overly optimistic metrics. This represents a broader limitation in current benchmarks for fully assessing zero-shot generalization across entirely unseen urban morphologies. Second, practical application boundaries exist. Performance is inherently constrained by mixed-pixel effects in lower-resolution images. Furthermore, the model is currently optimized for optical imagery, and its specific parameter complexity may pose challenges for deployment in strictly resource-constrained or real-time edge devices. Generalization to multimodal data and the development of lightweight variants will be the focus of future work.

CONCLUSION

This paper presents an improved EViT-based semantic segmentation network for high-resolution remote sensing imagery. By integrating GCC-MSA, LGC, CoAt, SGSPP, GEP, and SPGM, the hybrid architecture simultaneously models local details and global semantics. Experiments on three building datasets (WHU Aerial, Massachusetts, GF-7) show that the proposed method outperforms most existing models in metrics such as IoU. Ablation studies confirm the importance of the dual-branch structure and the contribution of each key module. The work validates the effectiveness of CNN-Transformer hybrid architectures for building extraction and provides an extensible solution for precise surface object extraction from high-resolution remote sensing imagery. Building upon the limitations discussed, future work will focus on exploring geographically separated data splits to rigorously evaluate zero-shot generalization, extending the framework to multimodal remote sensing data, and developing lightweight architectures for efficient real-time deployment.

Funding

This research was funded by the Scientific Research Startup Foundation of Fujian University of Technology, grant number GY-Z24009.

Data Availability Statement

The data supporting the findings of this study are available in Zenodo at https://doi.org/10.5281/zenodo.17735828.

REFERENCES

Badrinarayanan V, Kendall A, Cipolla R (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39:2481–95. https://doi.org/10.1109/TPAMI.2016.2644615

Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M (2023). Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky L, Michaeli T, Nishino K, eds. Computer Vision – ECCV 2022 Workshops. Proceedings of the 17th European Conference on Computer Vision Workshops, 2022 Oct 23–27; Tel Aviv, Israel. Cham: Springer, 205–18.

Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y (2021). TransUNet: Transformers make strong encoders for medical image segmentation. Retrieved 2021 Feb 8, from https://arxiv.org/abs/2102.04306.

Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. Computer Vision - ECCV 2018: 15th European Conference. 2018 Sep 8–14; Munich, Germany. Cham: Springer, 801–18.

Chen P, Huang H, Ye F, Liu L, Li X, Liu M, Zhang L (2024). A benchmark GaoFen-7 dataset for building extraction from satellite images. Sci Data 11:187. https://doi.org/10.1038/s41597-024-03009-5

Dai Z, Liu H, Le QV, Tan M (2021). CoAtNet: Marrying convolution and attention for all data sizes. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, eds. Advances in Neural Information Processing Systems 34 (NeurIPS 2021). 2021 Dec 6–14; Virtual. Red Hook, NY: Curran Associates, 3965–77.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR 2021). 2021 May 3-7; Virtual. Available from: https://openreview.net/ forum?id=YicbFdNTTy.

He K, Zhang X, Ren S, Sun J (2016). Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). 2016 Jun 26–Jul 1; Las Vegas, NV, USA. New York: IEEE, 770–8.

Hu J, Shen L, Sun G (2018). Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). 2018 Jun 18–22; Salt Lake City, UT, USA. New York: IEEE, 7132–41.

Ji S, Wei S, Lu M (2019). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery dataset. IEEE Trans Geosci Remote Sens 57:574–86. https://doi.org/10.1109/ TGRS.2018.2858817

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 Sens 11(23):2813. https://doi.org/10.3390/rs11232813

Li Y, Hong D, Li C, Yao J, Chanussot J (2024). HD-Net: High-resolution decoupled network for building footprint extraction via deeply supervised body and boundary decomposition. ISPRS J Photogramm Remote Sens 209:51–65. https://doi.org/10.1016/j.isprsjprs. 2024.01.022

Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017). Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). 2017 Jul 21–26; Honolulu, HI, USA. New York: IEEE, 2117–25.

Liu H, Luo J, (2019). DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery. Remote Sens 11(20):2380. https://doi.org/10.3390/rs11202380

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021). 2021 Oct 10–17; Montreal, QC, Canada. New York: IEEE, 10012–22.

Long J, Shelhamer E, Darrell T (2015). Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). 2015 Jun 7-12; Boston, MA, USA. New York: IEEE, 3431-40.

Mehta S, Rastegari M (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In: Proceedings of the International Conference on Learning Representations (ICLR 2022). 2022 Apr 25–29; Virtual. Available from: https://openreview.net/forum?id=vh-0n7s7sEx.

Mnih V (2013). Machine learning for aerial image labeling. Ph.D. dissertation. University of Toronto, Toronto, ON, Canada. Retrieved 2013, from https://api.semanticscholar.org/CorpusID:114890196.

Qin D, Leichner C, Delakis M, Marcin M, Wang J, Adam G, Howard A (2024). MobileNetV4: Universal models for the mobile ecosystem. In: Proceedings of the European Conference on Computer Vision (ECCV 2024). 2024 Sep 29–Oct 4; Milan, Italy. Cham: Springer. arXiv:2404.10518.

Ronneberger O, Fischer P, Brox T (2015). U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference. 2015 Oct 5–9; Munich, Germany. Cham: Springer, 234–41.

Shrestha S, Vanneschi L (2018). Improved fully convolutional network with conditional random fields for building extraction. Remote Sens 10(7):1135. https://doi.org/10.3390/rs10071135

Shaker A, Maaz M, Rasheed H, Khan S, Yang MH, Khan FS (2023). SwiftFormer: Efficient additive attention for transformer-based real-time mobile vision applications. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023). 2023 Oct 2–6; Paris, France. New York: IEEE, 17425–36.

Wang L, Fang S, Meng X, Li R (2022a). Building extraction with vision transformer. IEEE Trans Geosci Remote Sens 60:5625711. https://doi.org/10.1109/ TGRS.2022.3186634

Wang L, Li R, Duan C, Zhang C, Meng X, Fang S (2022b). A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. IEEE Geosci Remote Sens Lett 19:1–5.

Xiang X, Gong W, Li S, Chen J, Ren T (2024). TCNet: Multiscale fusion of transformer and CNN for semantic segmentation of remote sensing images. IEEE J Sel

Top Appl Earth Obs Remote Sens 17:3123–36. https://doi.org/10.1109/JSTARS.2024.3349625

Xu J, Xiong Z, Bhattacharyya SP (2023). PIDNet: A real-time semantic segmentation network inspired by PID controllers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). 2023 Jun 18–22; Vancouver, BC, Canada. New York: IEEE, 19529–39.

Yang F, Jiang FL, Li JZ, Lu L (2024). MSTrans: Multi-scale transformer for building extraction from HR remote sensing images. Electronics 13(23):4610. https://doi.org/10.3390/electronics13234610

Zhang H, Wang Y, Li Q, Xu L, Yang M-H (2024). Extracting building footprint from remote sensing images by an enhanced vision transformer network. IEEE Trans Geosci Remote Sens 62:5602315. https://doi.org/ 10.1109/TGRS.2024.3421651

Zhang R, Zhao J, Li M, Zou Q (2024). LGDB-Net: Dual-branch path for building extraction from remote sensing image. In: Proceedings of the 30th IEEE International Conference on Parallel and Distributed Systems (ICPADS 2024). 2024 Oct 10–14; Belgrade, Serbia. New York: IEEE, 452–61.

Zhang Y, Liu H, Hu Q (2021). TransFuse: Fusing transformers and CNNs for medical image segmentation. In: de Bruijne M, , eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference. 2021 Sep 27–Oct 1; Strasbourg, France (virtual). Cham: Springer, 3–11.

Zhao H, Shi J, Qi X, Wang X, Jia J (2017). Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). 2017 Jul 21–26; Honolulu, HI, USA. New York: IEEE, 2881–90.

Zhou Y, Chen Z, Wang B, Li S, Liu H, Xu C (2022). BOMSC-Net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery. IEEE Trans Geosci Remote Sens 60:5618617. https://doi.org/ 10.1109/TGRS.2022.3152575

Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2018). UNet++: A nested U-Net architecture for medical image segmentation. In: Stoyanov D, , eds. 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. 2018 Sep 16–20; Granada, Spain. Cham: Springer, 3–11.

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 Trans Geosci Remote Sens 59:6169–81. https://doi.org/10.1109/TGRS.2020.3026051.