

Article

Edge-Attentive Dual-Branch Frame Field Network for High-Precision Building Polygon Extraction

Ruijie Han ^{1,2,3}, Xiangtao Fan ^{1,2}, Jian Liu ^{1,2,*} , Weijia Bei ^{1,2,3} , Qifeng Ge ^{1,2,3} , Jianhao Xu ^{1,2,3} and Ruijie Yao ³

¹ Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; hanruijie23@mails.ucas.ac.cn (R.H.); fanxt@aircas.ac.cn (X.F.); beiweijia21@mails.ucas.ac.cn (W.B.); geqifeng21@mails.ucas.ac.cn (Q.G.); xujianhao22@mails.ucas.ac.cn (J.X.)

² International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China

³ College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China; yaoruijie23@mails.ucas.ac.cn

* Correspondence: liujian@aircas.ac.cn

Highlights

What are the main findings?

- Proposal of the Edge-Attentive Dual-Branch Frame Field Network (EA-DBFFN): A novel architecture is proposed that explicitly separates regional consistency and boundary precision for high-fidelity building polygon extraction.
- Development of a Dual-Task Decoder with Mutual Guidance: A mechanism is introduced that facilitates information exchange between mask and edge branches, significantly improving geometric consistency.
- Integration of Edge-Enhancement and Dynamic Receptive Fields: Fixed Sobel/Laplacian filters and spatial attention are leveraged to capture fine-grained structural details while maintaining global context.
- Achievement of Superior Accuracy and Efficiency: The proposed method outperforms state-of-the-art (SOTA) frame field methods with higher segmentation scores and lower tangent angle errors while simultaneously maintaining a faster inference speed.

What are the implications of the main findings?

- This study provides a robust and computationally efficient framework for generating high-quality vectorized building footprints for large-scale geospatial analysis.
- The framework improves the reliability of automated urban mapping by effectively addressing challenges like vegetation occlusion and blurred boundaries in remote sensing imagery.

Abstract

Efficient extraction of building footprints from aerial and satellite imagery is essential for urban planning, infrastructure management, and large-scale geospatial analysis. Traditional raster-based approaches provide limited geometric precision, while existing polygon-generation methods often rely on detecting and ordering small-scale building vertices, which can lead to incomplete structures, distorted shapes, and high computational cost. To address these limitations, this study proposes an Edge-Attentive Dual-Branch Frame Field Network (EA-DBFFN) for automated and high-precision building polygon extraction. The method is built upon frame field learning and introduces a dual-branch architecture that separately predicts building masks and edges. A Dual-Task Decoder enlarges and adapts receptive fields while applying spatial attention to enhance the representation of structural



Academic Editor: Massimiliano Pepe

Received: 2 March 2026

Revised: 9 April 2026

Accepted: 10 April 2026

Published: 13 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

details. Fixed Sobel and Laplacian filters are incorporated to strengthen boundary detection. In addition, a Dual-Task Mutual Guidance Module promotes the exchange of complementary information between the mask and edge branches, improving geometric consistency and reducing boundary errors. Experiments conducted on the Inria Aerial dataset and the CrowdAI dataset demonstrate that EA-DBFFN achieves superior performance in region-based metrics, with an AP75 of 72.9% on CrowdAI, representing a 2.3% improvement over competing methods. Furthermore, EA-DBFFN produces geometrically higher-quality polygons, with the Max Tangent Angle error reduced by 6.4%, the Invalid Polygon Ratio reduced by 66.3%, and Edge Smoothness improved by 72.7% compared to the best competing method. The results show that EA-DBFFN provides an effective and computationally efficient framework for generating high-quality vectorized building footprints suitable for large-scale urban analysis.

Keywords: building mapping; regular building contour extraction; remote sensing images; frame field

1. Introduction

Generating accurate and well-structured building footprints from remote sensing data is fundamental for understanding and managing the complex spatial dynamics of urban environments. Building footprints serve as fundamental spatial units in Geographic Information Systems (GISs), enabling downstream applications such as cartography, 3D city modeling [1], and cadastral mapping [2–4]. Compared with raster-based representations, vectorized building polygons offer compact and topologically consistent descriptions that better capture the geometric and semantic structure of urban areas. Consequently, developing accurate and computationally efficient building polygon extraction methods has become a crucial research frontier. While deep learning has significantly advanced this field, existing methods generally fall into four categories, each with distinct limitations.

The first category, segmentation-and-post-processing approaches [5–8], predicts binary building masks using neural networks, followed by geometric post-processing (e.g., polygon simplification) to regularize the output. A representative pipeline involves extracting building contours from segmentation networks such as Mask R-CNN [9] or U-Net [10], simplifying the contours using algorithms like Douglas-Peucker algorithm [11], and further optimizing the polygons through descriptor-based strategies such as the Minimum Description Length (MDL) principle [12]. For example, Zhao et al. [6] combined Mask R-CNN with contour simplification and MDL optimization to refine polygon representations, while Daranagama et al. [13] directly generated shapefile-format polygons from U-Net predictions. Other studies [7,8] further enhanced polygon precision by improving the segmentation backbone or incorporating additional geometric constraints such as convex hull regularization. While straightforward and easy to implement, these methods often suffer from a lack of coherence between prediction and post-processing stages, since geometric regularization is performed independently from feature learning. As a result, the final polygons may deviate from actual building contours, particularly in complex urban scenes.

To overcome this, a second class of methods [14–22] aims to directly predict polygon vertices or vector representations in an end-to-end fashion. Early vertex prediction methods, such as PolygonRNN [23] and PolygonRNN++ [24], were originally proposed for semantic image annotation and employed recurrent neural networks (RNNs) to sequentially generate polygon vertices. Building upon this idea, PolyMapper [14] extended the framework to building boundary regularization by introducing a bounding-box-based mechanism for extracting multiple building polygons and directly connecting sequentially predicted vertices.

To improve vertex prediction accuracy, Zhao et al. [15] incorporated spatial attention and Conv-GRU [25] modules. To reduce training complexity, subsequent studies [19] replaced RNNs with Graph Convolutional Networks (GCNs) [26]. BuildingVec [21] enhances building polygon prediction accuracy by jointly predicting the vertex positions and classifying whether each vertex represents a corner. In addition, a technique called ROIPoly [20] reduces computational costs by representing vertices as queries within a Region-of-Interest (ROI) framework, thereby limiting attention computation to pertinent areas.

Another approach adopted a two-stage strategy: first detecting an initial contour segmentation map, then extracting polygons and vertices from these contours, and subsequently deriving building polygons through vertex optimization and fine-tuning [22,27–30]. Other works directly detect vertex locations from images, refine them via Graph Neural Networks (GNNs), and establish edge connections to generate polygons [17,18].

With the emergence of the Transformer [31] architecture, global attention mechanisms have been increasingly adopted for modeling long-range dependencies in vertex prediction tasks. Related studies have shown that Transformer's attention mechanisms (as used by Zhang et al. [32]) and the DETR [33] architecture (as used by Hu et al. [34]) offer an effective approach for building vertex sequence prediction.

Another line of research seeks to directly generate regularized building polygons by incorporating geometric constraints into the network training process [35]. Instead of relying on explicit post-processing or sequential vertex prediction, these methods integrate regularization objectives within the learning framework to encourage structured polygon outputs. For instance, Generative Adversarial Networks (GANs) [36] have been employed to regularize segmentation maps through adversarial supervision, guiding the network toward more geometrically consistent building boundaries [37,38]. Similarly, regularization loss functions [39,40] have been introduced to impose structural constraints during training, enabling the network to produce polygons with desired geometric properties.

More recently, end-to-end architectures have attempted to combine contour refinement and vectorized prediction within a unified framework. For example, Sun et al. [41] proposed a Transformer-based contour refinement head that jointly performs roof extraction and fine-grained classification to enhance polygon accuracy. While these direct extraction approaches demonstrate promising improvements in boundary regularization, they often involve adversarial optimization or tightly coupled network designs. Such training schemes tend to be computationally expensive and less stable than conventional supervised learning, and their structural complexity may further increase optimization difficulty, which limits their scalability and practical applicability in large-scale remote sensing scenarios.

Frame field-driven methods [42–44] offer a compelling alternative by introducing geometric priors into the learning process to facilitate boundary alignment and polygon regularization. Inspired by Marcos et al. [45], who combined the Active Contour Model (ACM) with CNNs, Girard et al. [42] introduced frame fields, shifting contour optimization from direct boundary evolution to skeleton-based graph representations.

Subsequent studies advanced the framework. Nguyen et al. [43] combined super-resolution, frame field learning, and polygonization to extract building footprints in dense building areas. Sun et al. [44] fused Digital Surface Models (DSMs) and RGB imagery, performing post-processing by comparing segmentation maps obtained from a U-Net with frame fields to derive regularized building polygons, which resulted in high-precision outputs. Xu et al. [46] proposed the Alignment-Free Module (AFM), where the network simultaneously predicts vertices, masks, and AFM, matching them to obtain the final result. These studies collectively validate the effectiveness of frame field methods in building polygon extraction tasks and their potential in boundary regularization.

The existing frame field network architecture typically consists of a feature encoder for multi-scale image features and a decoder that predicts frame fields, binary masks, and edges. Despite their overall effectiveness, conventional decoders in such networks typically employ simple stacked convolution layers. Consequently, while these methods achieve good performance in regularized building extraction, their efficacy is dependent on the accuracy and completeness of the intermediate binary mask and edge predictions. Remote sensing images often contain vegetation occlusions, complex textures, and blurred boundaries, which make it difficult for simple decoders to simultaneously capture global structure and local details. This fundamental limitation poses significant challenges for further improving the performance of frame field methods.

To address these limitations, we propose an Edge-Attentive Dual-Branch Frame Field Network (EA-DBFFN), building upon existing frame field networks for high-precision building polygon extraction. Our method's key insight is to separately optimize for regional consistency and boundary precision. We implement this through a Dual-Task Decoder, structured with a shared encoder, dual task-specific branches, and a dedicated Dual-Task Mutual Guidance Module. The mask branch integrates a dynamic receptive field module and a spatial attention mechanism to improve the consistency of regional representations. The edge branch leverages fixed Sobel and Laplacian filters to amplify high-frequency features, improving boundary localization accuracy. The mutual guidance module further facilitates residual-based cross-task optimization, enabling effective collaboration between mask and edge predictions. This strategy enhances the alignment between masks and frame fields, ultimately leading to higher-precision building polygons.

The main contributions of this work are as follows:

- We propose a dual-branch decoding structure that explicitly separates regional and boundary modeling, addressing the limitation of conventional unified decoders in simultaneously capturing global structure and local boundary details, thereby enhancing the geometric precision of building boundaries.
- We design lightweight and efficient sub-modules, including a dynamic receptive field module to handle multi-scale building variations, and an edge enhancement module incorporating fixed Sobel and Laplacian filters to mitigate boundary ambiguity caused by vegetation occlusion and cluttered backgrounds.
- We introduce a dual-task mutual guidance mechanism that enables residual-based information exchange between mask and edge predictions, addressing structural misalignment and prediction inconsistency across the two branches, and ultimately improving the alignment between masks and frame fields for higher-precision polygon extraction.

2. Materials and Methods

2.1. Network Structure

We propose an Edge-Attentive Dual-Branch Frame Field Network (EA-DBFFN) for precise building polygon extraction, whose overall architecture is shown in Figure 1. Our end-to-end workflow is clearly divided into four main stages: input images first undergo feature extraction via a backbone network; subsequently, extracted features are fed into a dual-task decoder to concurrently generate refined building masks and edge predictions, which mutually enhance each other through collaborative optimization. Simultaneously, the network performs frame field generation, which is then fused with the aforementioned predictions in a frame field alignment module to enhance boundary orientation and geometric consistency. Finally, leveraging the fused information, a polygonization step yields high-precision vector building polygons. This explicit decoupling and synergistic enhancement of area and edge modeling effectively mitigates the limitations of traditional unified

decoding strategies, significantly improving the geometric accuracy and topological quality of building boundaries. This section details the core components of the network.

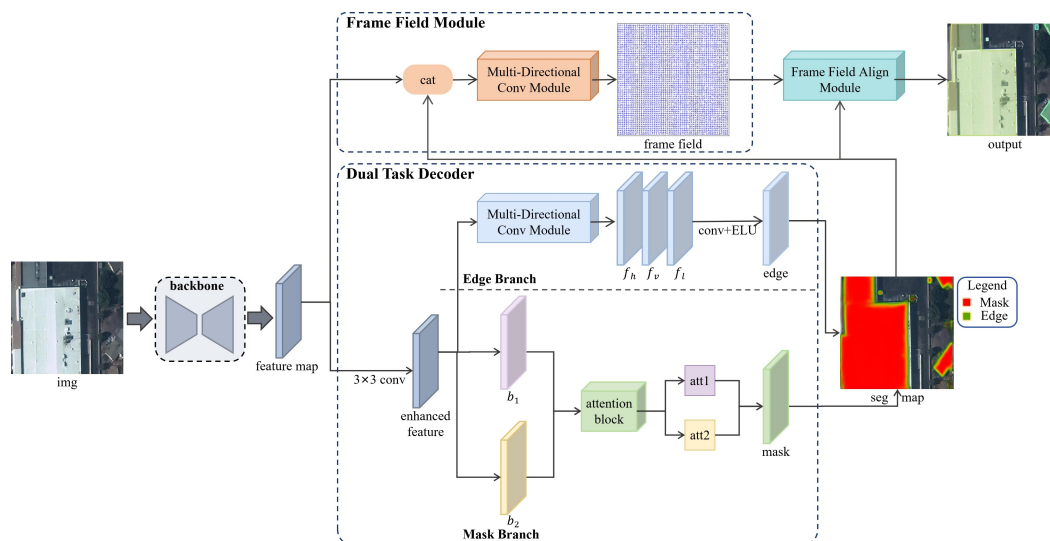


Figure 1. Overview of EA-DBFFN.

2.2. Dual-Task Decoder

To address structural discontinuities and tracking errors commonly encountered in corners or occluded regions, we introduce a lightweight dual-branch decoder. It incorporates a mutual guidance mechanism that promotes joint optimization between mask and edge predictions. As illustrated in Figure 1, the decoder comprises a shared feature encoder and two task-specific branches: the mask branch employs a Dynamic Receptive Field module and spatial attention mechanism to enhance regional consistency, while the edge branch utilizes fixed Sobel and Laplacian filters to strengthen boundary detection. These outputs are further refined via a Dual-Task Mutual Guidance Module using cross-task residual learning.

2.2.1. Shared Feature Encoder

To provide high-quality input to both branches, a shared feature encoder enhances shallow features from the backbone. It is built from two sequential convolutional modules, where each module includes a 3×3 convolution layer followed by batch normalization and an ELU activation function. The first module projects the backbone output from 32 to 64 channels to enrich feature representation, while the second refines the expanded features to improve spatial consistency. This two-layer design strikes a balance between representational capacity and computational efficiency: a single layer is insufficient to bridge the gap between the backbone output and the task-specific branches, while deeper stacking would introduce unnecessary parameters without meaningful performance gain.

2.2.2. Mask Branch

At the core of the mask branch lies an advanced attention module designed to efficiently fuse global contextual information with fine-grained local structural details—a critical factor for generating high-quality segmentation masks. As illustrated in Figure 2, this module achieves this goal through a dual mechanism combining the Dynamic Receptive Module and Spatial Gating.

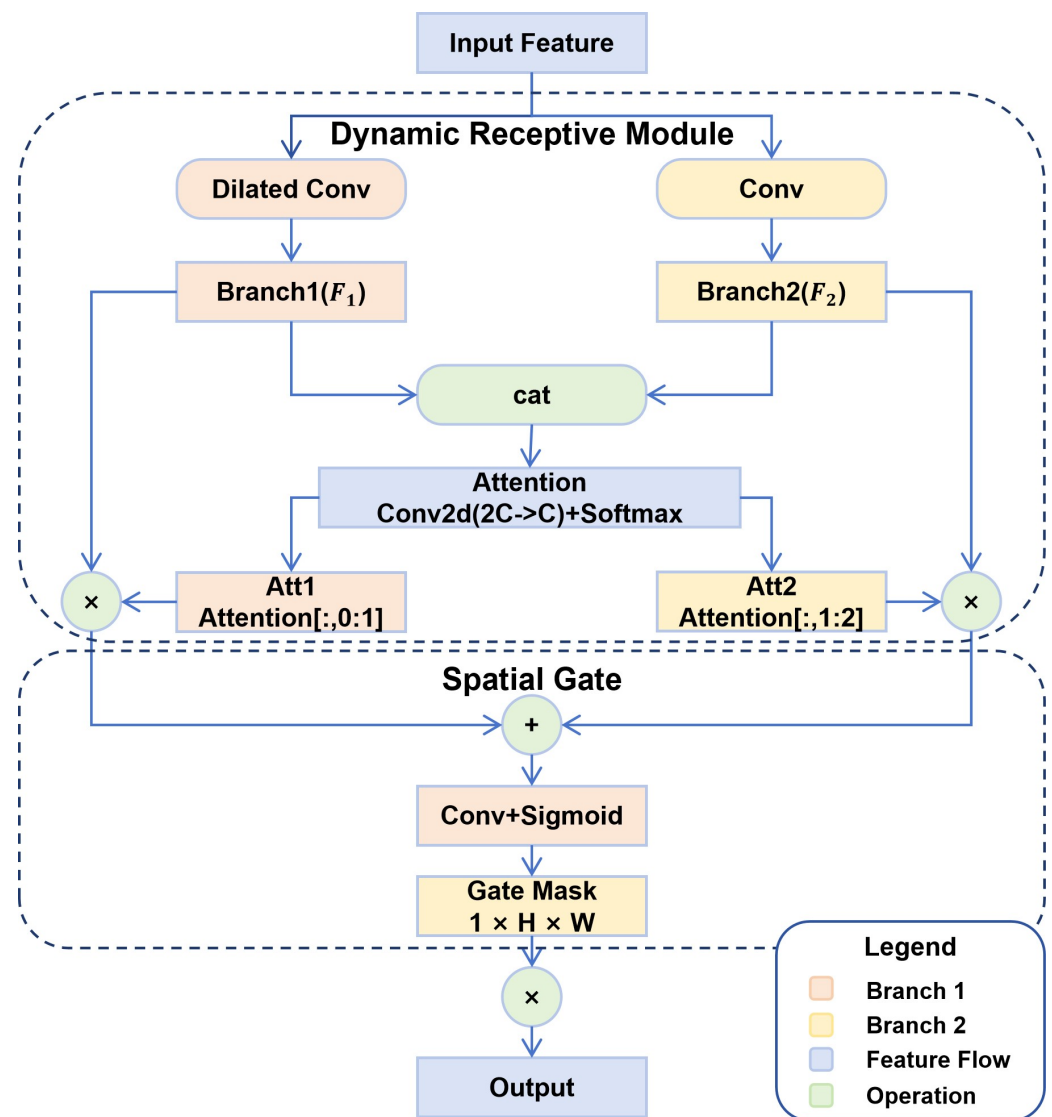


Figure 2. The structure of the Mask Branch.

First, the Dynamic Receptive Block captures multi-scale features through parallel branches F_1 (hollow convolution) and F_2 (original receptive field). The attention weights A_1 and A_2 are dynamically generated by concatenating F_1 and F_2 along the channel dimension, followed by a 1×1 convolution to reduce the dimension to 2 channels and a channel-wise Softmax operation, ensuring that the attention weights at each spatial location are normalized:

$$[A_1, A_2] = \text{Softmax}(\text{Conv}_{1 \times 1}([F_1; F_2])), \quad (1)$$

where $[F_1; F_2]$ denotes channel-wise concatenation and the Softmax operation normalizes the attention weights at each spatial location across the two branches. The channel attention mechanism then adaptively fuses the two branches, yielding the channel-weighted output F_{att} , where $\odot$ denotes element-wise multiplication:

$$F_{att} = A_1 \odot F_1 + A_2 \odot F_2, \quad (2)$$

Next, the spatial gating module employs lightweight 3×3 convolutions and Sigmoid operations to generate a spatial attention mask G . This mask G suppresses background

noise while enhancing features in the target region. The final masked feature F_{out} is obtained through gating and 3×3 convolution:

$$F_{out} = \sigma(\text{Conv}_{3 \times 3}(F_{att} \odot G)), \quad (3)$$

where σ denotes the Sigmoid operation.

This dual attention mechanism effectively fuses multi-scale information while suppressing spatial noise, significantly enhancing the masking branch's perception of complex structural boundaries.

2.2.3. Edge Branch

To enhance boundary representation capability, we introduce an Edge Enhancement Block specifically for high-frequency feature modeling, whose structure is shown in Figure 3.

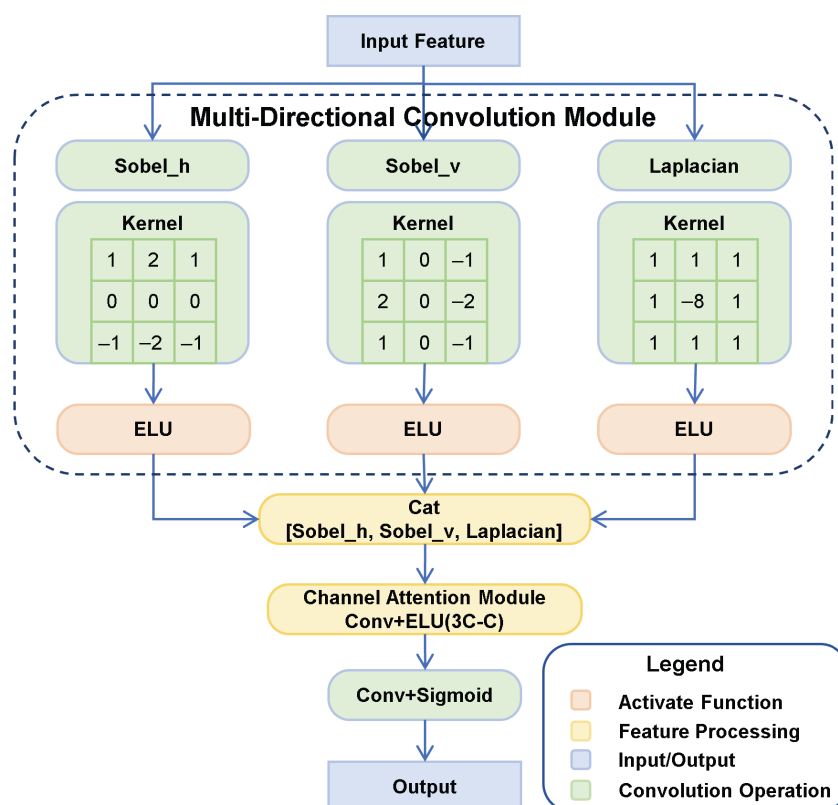


Figure 3. The structure of the Edge Branch.

This module first convolves the input feature map $F \in R^{C \times H \times W}$ using three classical edge detection operators—horizontal Sobel (S_h), vertical Sobel (S_v), and the Laplacian operator (L)—and extracts the initial multi-directional edge features F_h , F_v , and F_l . These features are then concatenated along the channel dimension to form $F_{edge} \in R^{3C \times H \times W}$. To fuse multi-directional edge information and reduce dimensionality, we apply a 1×1 convolution combined with ELU activation to obtain the mixed edge feature F_{mix} .

Finally, applying a set of 3×3 convolutions followed by Sigmoid activation, the module generates a single-channel edge attention map $A_{edge} \in R^{1 \times H \times W}$ to guide feature learning in the main branch:

$$A_{edge} = \sigma(\text{Conv}_{3 \times 3}(F_{mix})), \quad (4)$$

where σ denotes the Sigmoid operation.

The choice of fixed Sobel and Laplacian operators is motivated by their complementary frequency characteristics. Sobel captures first-order directional gradients, while the Laplacian responds to second-order intensity variations, together providing a diverse representation of boundary structures. Compared to learnable convolutional kernels, fixed operators introduce no additional trainable parameters, which helps control model complexity and may reduce the risk of overfitting, especially under limited training data. In this work, they are used as lightweight priors rather than replacements for learnable components. Although the Laplacian is sensitive to noise, its output in our design is not used as a direct prediction but instead serves as an intermediate feature that is subsequently fused and refined through learnable 1×1 and 3×3 convolutions, effectively suppressing noise-induced artifacts before the features influence the final edge attention map. Moreover, building boundaries in high-resolution remote sensing imagery typically exhibit strong and consistent gradients, allowing classical edge operators to provide useful structural cues. The overall performance improvements observed in our experiments also suggest the effectiveness of this design in practice. This design ensures the network explicitly models high-frequency boundary information, significantly enhancing edge accuracy in segmentation results.

2.2.4. Mutual Guidance Module

To exploit the complementarity between region and boundary predictions, we design a Mutual Guidance Module. Given the initial mask prediction M and edge prediction E , both paths share the same joint input formed by channel-wise concatenation $F_{joint} = [M; E]$. Each path applies two 3×3 convolutional layers with an ELU activation in between to produce a residual correction, which is then added back to the original prediction and normalized via Sigmoid:

$$\hat{M} = \text{Sigmoid}(M + R_{mask}(F_{joint})) \quad (5)$$

$$\hat{E} = \text{Sigmoid}(E + R_{edge}(F_{joint})), \quad (6)$$

where R_{mask} and R_{edge} denote the two independent refiners. This residual formulation ensures that each path corrects prediction errors rather than overwriting the original outputs, mitigating structural misalignment and enhancing consistency across branches.

2.3. Frame Field Alignment and Polygonization

To achieve the transition from pixel-level prediction to structured vector polygons for building contours, we introduce the Frame Field representation into the network. The Frame Field characterizes the principal direction and geometric consistency of local boundaries within an image. It is a direction field with second-order symmetry, described by a pair of complex coefficients $c_0(x)$, $c_2(x)$ at each pixel location, representing the local principal direction θ_x and its relative stability. In our network, the frame field is predicted from feature maps shared between the backbone network and decoder, serving as directional priors for edge structures.

After producing the frame field, we combine it with the mask and edge predictions to refine the contour. First, our algorithm employs Active Skeleton Model (ASM) optimization to refine the initial contour for precise alignment with the predicted frame field. The ASM constructs a skeleton graph from the predicted edge map and optimizes it by minimizing an energy function that jointly considers boundary probability, frame field alignment, and path length regularity. Subsequently, a critical corner detection step utilizes frame field information to identify key geometric vertices of buildings. Corner-aware simplification then minimizes non-corner vertices while preserving essential geometric structures. Fi-

nally, polygonization is performed, followed by polygon filtering to remove redundant or minute polygons that do not conform to building geometry standards, yielding the final high-precision vector building outline. Both the ASM optimization and corner detection procedures follow the implementation of Girard et al. [42].

2.4. Loss Function

To support accurate segmentation and stable geometric modeling, the network employs a composite loss comprising Segmentation Loss, Frame Field Loss, Output Coupling Loss, and a final weighted aggregation. This design jointly optimizes semantic prediction and structural consistency for vectorized building contour extraction.

2.4.1. Segmentation Loss

To achieve fine-grained modeling of regions and boundaries, we introduce a dual-branch segmentation head structure on top of the backbone network to separately predict the building interior regions $\hat{y}_{int}$ and boundary regions $\hat{y}_{edge}$. A composite loss function, comprising a weighted blend of cross-entropy and Dice losses, is employed during training:

$$\mathcal{L}_{int} = CE(\hat{y}_{int}, y_{int}) + \lambda_1 \cdot Dice(\hat{y}_{int}, y_{int}), \quad (7)$$

$$\mathcal{L}_{edge} = CE(\hat{y}_{edge}, y_{edge}) + \lambda_2 \cdot Dice(\hat{y}_{edge}, y_{edge}), \quad (8)$$

where y_{int} and y_{edge} denote the ground-truth masks for the building interior and boundary regions, respectively, and λ_1 and λ_2 are the balance coefficients.

2.4.2. Frame Field Loss

Building upon the semantic segmentation prediction, we introduce a frame field prediction branch to encode the directional field information of building boundaries. This branch takes $[\hat{y}_{backbone}, \hat{y}_{seg}]$ as input and outputs a complex symmetric tensor field composed of $\hat{c}_0$ and $\hat{c}_2$, which fits the principal orientation of the building contours. We design three constraint terms:

Alignment Loss:

$$\mathcal{L}_{align} = \frac{1}{HW} \sum_x y_{edge}(x) \cdot \left| f(e^{i\theta_\tau}; \hat{c}_0(x), \hat{c}_2(x)) \right|^2, \quad (9)$$

ensuring alignment with contour tangential direction.

Orthogonal Alignment:

$$\mathcal{L}_{align90} = \frac{1}{HW} \sum_x y_{edge}(x) \cdot \left| f(e^{i\theta_\tau^\perp}; \hat{c}_0(x), \hat{c}_2(x)) \right|^2, \quad (10)$$

preventing field degeneration into a linear field.

Smoothness Regularization:

$$\mathcal{L}_{smooth} = \frac{1}{HW} \sum_x (||\nabla \hat{c}_0(x)||^2 + ||\nabla \hat{c}_2(x)||^2), \quad (11)$$

enforcing spatial continuity to avoid local oscillations.

2.4.3. Output Coupling Loss

To enhance the collaborative relationship between different prediction branches, we introduce three output coupling terms:

Internal Segmentation and Frame Field Coupling:

$$\mathcal{L}_{int-align} = \frac{1}{HW} \sum_x f(\nabla \hat{y}_{int}(x); \hat{c}_0(x), \hat{c}_2(x))^2, \quad (12)$$

aligning the internal segmentation map's gradient direction with the frame field for geometric consistency.

Edge Segmentation and Frame Field Coupling:

$$\mathcal{L}_{edge-align} = \frac{1}{HW} \sum_x f(\nabla \hat{y}_{edge}(x); \hat{c}_0(x), \hat{c}_2(x))^2, \quad (13)$$

applying frame field alignment to the edge segmentation map's gradient for enhanced edge-direction coordination.

Internal Segmentation and Boundary Consistency:

$$\mathcal{L}_{int-edge} = \frac{1}{HW} \sum_x \max \left(1 - \hat{y}_{int}(x), \|\nabla \hat{y}_{int}(x)\|^2 \right) \cdot \left| \|\nabla \hat{y}_{int}(x)\|^2 - \hat{y}_{edge}(x) \right|, \quad (14)$$

encouraging the edge map's strength to match the internal map's gradient magnitude at building boundaries, boosting outer contour response.

2.4.4. Final Loss Composition

To prevent training instability due to differing numerical scales, each loss term undergoes statistical normalization with respect to the initial network parameters. These normalized losses are then linearly weighted and combined to form the final loss function:

$$\mathcal{L}_{final} = \alpha_1 \cdot \tilde{\mathcal{L}}_{int} + \alpha_2 \cdot \tilde{\mathcal{L}}_{edge} + \dots + \alpha_8 \cdot \tilde{\mathcal{L}}_{int-edge}, \quad (15)$$

where α_i represents the weight for each individual loss term, and $\mathcal{L}$ represents the loss term after normalization. The weights α_i are set empirically based on the relative importance of each loss term. Specifically, $\alpha_1 = 10$ for the segmentation loss, $\alpha_2 = 1.5$ for edge loss, $\alpha_3 = 1$ for crossfield alignment, $\alpha_4 = 0.2$ for orthogonal alignment, and $\alpha_5 = 0.005$ for smoothness regularization. For the coupling losses, α_6 , α_7 and α_8 are gradually increased from 0 to 0.2 during the first 5 training epochs to stabilize early training, after which they remain fixed.

2.5. Evaluation Metrics

Standard segmentation performance indicators, such as MS COCO's [47] Average Precision (AP) and Average Recall (AR) as well as their variants (AP₅₀, AP₇₅, AR₅₀, AR₇₅), quantify regional overlap and matching performance. However, building contour extraction requires a stronger emphasis on geometric regularity and structural fidelity. Overlap-based measures often favor smooth or blurred boundaries and may not adequately capture the sharp angles and rectilinear structures characteristic of buildings, particularly under annotation noise. Therefore, we complement the standard metrics with geometry-oriented measures to more comprehensively evaluate contour quality.

2.5.1. Max Tangent Angle Error (MTA)

This metric [42] measures the maximum deviation between the tangent directions of the predicted and ground-truth contours, reflecting the preservation of local geometric morphology. Given the contour curves $C_{gt}(t)$ and $C_{pred}(t)$, the tangent angle difference is:

$$\theta(t) = \arctan 2 \left(\|T_{gt}(t) \times T_{pred}(t)\|, T_{gt}(t) \cdot T_{pred}(t) \right). \quad (16)$$

The MTA is then defined as the maximum angle value over all sampled points.

2.5.2. Invalid Polygon Ratio (IPR)

This measures [48] the proportion of predictions that fail to form valid closed polygons (e.g., self-intersections or missing edges), indicating the structural legality and robustness of the reconstructed contours.

2.5.3. Edge Smoothness (ES)

Computed as the standard deviation of internal polygon angles, this metric reflects boundary regularity. Lower values correspond to cleaner, more rectilinear edges and stronger corner preservation.

2.5.4. Abnormal Convex Hull Ratio (ACHR > 1.5)

This evaluates the proportion of polygons whose convex hull area exceeds the polygon area by a factor greater than 1.5, capturing anomalies such as excessive concavity or incomplete reconstruction.

Together, these geometric metrics offer a multi-dimensional assessment of contour quality—covering structural correctness, edge sharpness, legality, and smoothness—thereby complementing traditional region-based accuracy measures.

2.6. Datasets

We compare the performance of our method with that of the state-of-the-art methods on the two publicly available datasets.

2.6.1. Inria Dataset

The Inria dataset [49] is a comprehensive benchmark for building extraction, comprising orthorectified RGB aerial imagery with a spatial resolution of 0.3 m and a size of 5000×5000 pixels. The imagery is collected from different cities, covering diverse urban morphologies, building densities, and architectural styles. This diversity makes the dataset well suited for evaluating model generalization. Given its high spatial resolution, the dataset requires models to accurately delineate both large building footprints and small, detailed structures. Variations in building shape, size, and orientation, along with background clutter, further increase the difficulty of achieving precise segmentation, making Inria a challenging benchmark for fine-grained building extraction. Following the official benchmark protocol, the dataset is split into disjoint training and testing areas. The training subset is used to learn model parameters, while all quantitative evaluations are conducted on the official test set only.

2.6.2. CrowdAI Dataset

The CrowdAI dataset [50] consists of high-resolution satellite imagery paired with annotated building outlines. Owing to its geographical diversity and the high fidelity of its labels, it serves as a benchmark for evaluating building extraction methods. The dataset is available in two variants: a full version and a small version. Both provide RGB images with a resolution of 0.3 m per pixel and a fixed spatial size of 300×300 pixels. The full version is a large-scale benchmark, containing 280,741 training images and 60,317 test images. The small version is utilized for resource-efficient experimentation, particularly for ablation studies, with 8366 training images and 1820 test images. For both versions,

models are trained on the official training split, and evaluation metrics are computed on the corresponding test set.

2.7. Experimental Setup

All experiments were conducted under a unified hardware and software environment to comprehensively validate the proposed method. The hardware platform is equipped with an NVIDIA GeForce RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA), and the software environment is based on Windows, using PyTorch 2.3.0 and Python 3.9. All models are trained using the Adam optimizer with a base learning rate of 1×10^{-3} , scaled by the effective batch size, and capped at a maximum learning rate of 1×10^{-1} . An exponential learning rate scheduler with a decay factor of $\gamma = 0.95$ is applied throughout training. The batch size is set to 2 per GPU, and all models are trained for 150 epochs. No weight decay or dropout is applied.

To verify the effectiveness and generalization capability of EA-DBFFN, experiments were conducted on two datasets: Inria and CrowdAI. Firstly, a series of ablation studies were designed to evaluate the contribution of each module within the dual-task decoder, all conducted on the small version of CrowdAI for efficiency. Secondly, to validate the representativeness and reliability of the results, a comparative analysis was performed between our proposed network and mainstream benchmarks (e.g., Mask RCNN [9], PANet [51], TransBuilding [52], PolyMapper [14], Frame Field Network [42], and PolyWorld [17]) across the Inria and full CrowdAI datasets.

To maintain fair and consistent evaluation, slightly different strategies were adopted for the two datasets. On the Inria dataset, we report only COCO-style segmentation metrics, as the dataset focuses on large-scale region extraction and most existing methods provide only raster-level results without publicly available polygon predictions, making polygon-level comparison impractical. In contrast, the CrowdAI dataset provides richer annotations and is a mainstream dataset for polygon-based building extraction, so we conducted a more comprehensive evaluation, including COCO metrics, polygon quality indicators, and computational efficiency. Finally, visual comparisons in complex urban scenes are provided to highlight our method's advantages in detailed boundary depiction, structural clarity, and overall polygon regularity.

3. Results

This section provides a thorough analysis of the results obtained using EA-DBFFN. We first report results from ablation studies to evaluate the contribution of each module within the dual-task decoder, followed by comparisons with several advanced networks on the Inria and CrowdAI datasets. To ensure fairness and consistency, COCO-style segmentation metrics are used for both datasets, while polygon-level and efficiency evaluations are conducted only on CrowdAI due to its richer annotations. Finally, qualitative visualizations further demonstrate the advantages of our approach in extracting building polygons in complex urban environments. Overall, the findings demonstrate the reliability and practical utility of the proposed method for precise building polygon extraction.

3.1. Ablation Studies

To assess the contribution of each component, we performed three sets of ablation experiments, focusing on the architectural design and task collaboration mechanisms. The configurations for these studies are illustrated in Figure 4.

We first evaluated each branch of the dual-task decoder independently. Two ablated models were built: an edge-only branch (using baseline mask prediction, Figure 4a) and a mask-only branch (using baseline edge prediction, Figure 4b), isolating semantic and

geometric modeling effects. We then removed the inter-branch feedback mechanism (Figure 4c) to assess mutual guidance importance. This tests how cross-task interaction improves structural consistency and prediction accuracy.

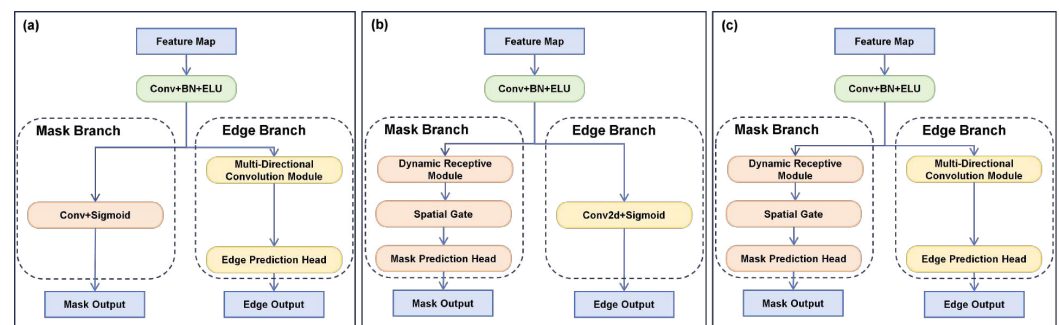


Figure 4. Comparison schematic of different model structures in ablation studies. From left to right: (a) Edge-Only; (b) Mask-Only; (c) The complete dual-branch structure, but with the mutual guidance mechanism between the mask and edge branches removed.

All ablation experiments were conducted on the small-scale CrowdAI dataset for 150 epochs. Models are trained on the corresponding training split, and performance is evaluated on the test set. Table 1 presents the training accuracy results for each model.

Table 1. Ablation experiment settings and corresponding performance results on the small version of the CrowdAI dataset. The table lists three key modules: Mask Branch (Mask), Edge Branch (Edge), and Mutual Guidance (Mutual) Module, along with the number of parameters (Params) introduced by each variant. Checkmarks (✓) and crosses (×) indicate whether each module is enabled, forming different model variants.

Branch Configuration				Performance Metrics					
Mask	Edge	Mutual	Params	AP	AP ₅₀	AP ₇₅	AR	AR ₅₀	AR ₇₅
×	×	×	55,680	46.3	74.3	51.9	56.9	84.0	64.5
×	✓	×	68,609	47.0	74.7	53.2	56.4	83.9	64.1
✓	×	×	130,436	47.5	75.2	53.3	57.6	84.8	65.9
✓	✓	×	143,365	47.7	74.7	54.1	58.5	84.7	66.7
✓	✓	✓	143,818	52.0	80.7	59.8	58.8	85.4	66.9

The results show that enabling a single branch yields slight performance gains over the baseline, with 1% AP improvement and marginal AR differences. The dual-branch structure (row 4) improves AP by 1.4% and AR by 1.6%, confirming the synergistic benefit of modeling both region and edge features. Finally, introducing the mutual guidance module (row 5) leads to the best performance, especially on stricter metrics like AP₇₅ (59.8%) and AR₇₅ (66.9%), validating its effectiveness in promoting joint feature refinement. The parameter counts listed in Table 1 provide clear evidence that the performance gains are driven by the proposed architectural design rather than by increased model capacity. Despite the Mask and Edge branches together introducing 87,685 additional parameters, the AP improvement from the baseline to the dual-branch variant without mutual guidance is only 1.4%. In contrast, the Mutual Guidance Module introduces merely 453 parameters, accounting for 0.31% of the total, yet produces the largest single performance jump of 4.3%. This disproportionate relationship between parameter count and performance gain strongly indicates that the improvements stem from the effectiveness of the proposed architecture rather than from increased model capacity.

Table 2 further presents the geometric quality metrics for each model variant, revealing a consistent improvement trend as modules are progressively added. The baseline model yields an MTA of 33.3, an IPR of 7.42%, an ACHR of 4.88%, and an ES of 145.1. Introducing the edge branch alone improves boundary localization, reducing MTA to 32.6 and IPR

to 4.32‰. The mask branch contributes more notably to polygon regularity, with ACHR dropping to 2.45‰ and ES to 139.2. Combining both branches without mutual guidance achieves further improvement, particularly in ACHR (1.09‰). The full EA-DBFFN achieves the best results across all geometric indicators, with an MTA of 30.8, IPR of 4.07‰, ACHR of 1.08‰, and ES of 130.3, demonstrating that the mutual guidance module plays a critical role in refining both boundary precision and overall polygon regularity.

Table 2. Ablation study results on polygon geometric quality metrics for different model variants on the small-scale CrowdAI dataset.

Method	MTA	IPR (‰)	ACHR (‰)	ES
Baseline	33.3	7.42	4.88	145.1
Edge_only	32.6	4.32	2.24	142.3
Mask_only	33.1	5.03	2.45	139.2
No_mutual	32.1	4.09	1.09	134.4
EA-DBFFN	30.8	4.07	1.08	130.3

Figure 5 presents visual comparisons of different model variants across typical regions, with each row showing a scene and each column corresponding to a model configuration. Visually, the baseline model trained without any of the proposed modules exhibits contour blurring, edge discontinuities, and missed detections in complex or densely arranged buildings. Introducing an improved single branch helps enhance geometric consistency and sharpen edge details; for instance, in the first row, the “Only-Edge” model shows better contour closure than the dual-branch variant, reflecting its strength in localized detail refinement. Nevertheless, from an overall perspective, the dual-branch structure more comprehensively captures semantic shapes and geometric boundaries, producing clearer contours and more complete segmentation, with generally more regular predicted edges compared to single-branch alternatives. When the mutual guidance module is incorporated, inter-branch coordination is further strengthened, resulting in continuous boundaries, higher contour closure, and more regular shapes across most highlighted areas. This demonstrates stronger generalization, particularly in complex clusters or heavily occluded regions.



Figure 5. Schematic diagram of prediction results for different models on the CrowdAI dataset.

In summary, despite single-branch models sometimes exhibiting finer local details, the overall advantages of the dual-branch structure combined with the mutual guidance mechanism are more significant. This combination enhances object completeness while maintaining edge precision, substantially boosting the stability and robustness of building contour extraction.

3.2. Results on the Inria Dataset

This section presents a detailed evaluation of our method on the Inria dataset, with comparisons against several leading approaches, including Mask RCNN, PANet, PolyMapper and the original Frame Field Network.

Table 3 presents the quantitative comparison of several models under standard instance segmentation metrics. Our EA-DBFFN demonstrates superior overall performance, with AP and AP₇₅ values of 46.8% and 49.5%, respectively, outperforming all other methods. The improvement at higher IoU thresholds highlights the model's superior boundary localization and geometric precision in building extraction.

Table 3. Quantitative comparison of different networks on the Inria dataset under instance-level evaluation on the test set. The best results are shown in bold.

Method	AP	AP ₅₀	AP ₇₅	AR	AR ₅₀	AR ₇₅
Mask RCNN	40.0	79.2	35.3	51.5	87.3	54.4
PANet	39.6	79.0	35.0	51.5	87.3	54.2
PolyMapper	44.9	82.5	45.4	55.4	90.8	61.7
FrameField	38.3	67.3	39.8	49.0	78.1	53.4
EA-DBFFN	46.8	79.8	49.5	64.4	92.0	71.8

Moreover, EA-DBFFN achieves a significantly higher AR (64.4%) compared to the other models, indicating enhanced detection completeness across varying building scales and densities. These improved results are a consequence of employing a dual-branch framework, which jointly optimizes edge and region representations, allowing the network to better capture detailed boundaries while maintaining region consistency. While PolyMapper and PANet exhibit competitive AP₅₀ performance, their drop at stricter IoU thresholds suggests weaker polygon refinement capability.

Figure 6 provides qualitative visualizations of the predicted building polygons on several Inria samples. It can be observed that our method is capable of generating geometrically regular and detailed polygonal structures that closely follow the true building outlines, regardless of variations in building density, size, or shape across different urban environments. This further confirms the stability and adaptability of our method on diverse aerial scenes.

3.3. Results on the CrowdAI Dataset

This section presents a detailed evaluation of the proposed method on the full CrowdAI dataset, benchmarking its performance against leading state-of-the-art approaches such as Mask RCNN, PANet, TransBuilding, PolyMapper, the original Frame Field Network, and PolyWorld. We quantitatively evaluated each model across standard segmentation metrics, polygon geometric quality, and model efficiency, supplemented by visualizations, to comprehensively demonstrate our method's superiority and effectiveness.

3.3.1. Analysis of Region-Based Standard Evaluation Metrics

Table 4 presents various methods' performance under standard instance segmentation metrics. Our EA-DBFFN achieved an AP of 63.7%, outperforming all others (PolyWorld

63.3%, FrameField 61.3%). Notably, EA-DBFFN scored highest at 72.9% for AP_{75} , indicating high overlap with ground-truth building regions even under stricter IoU thresholds.

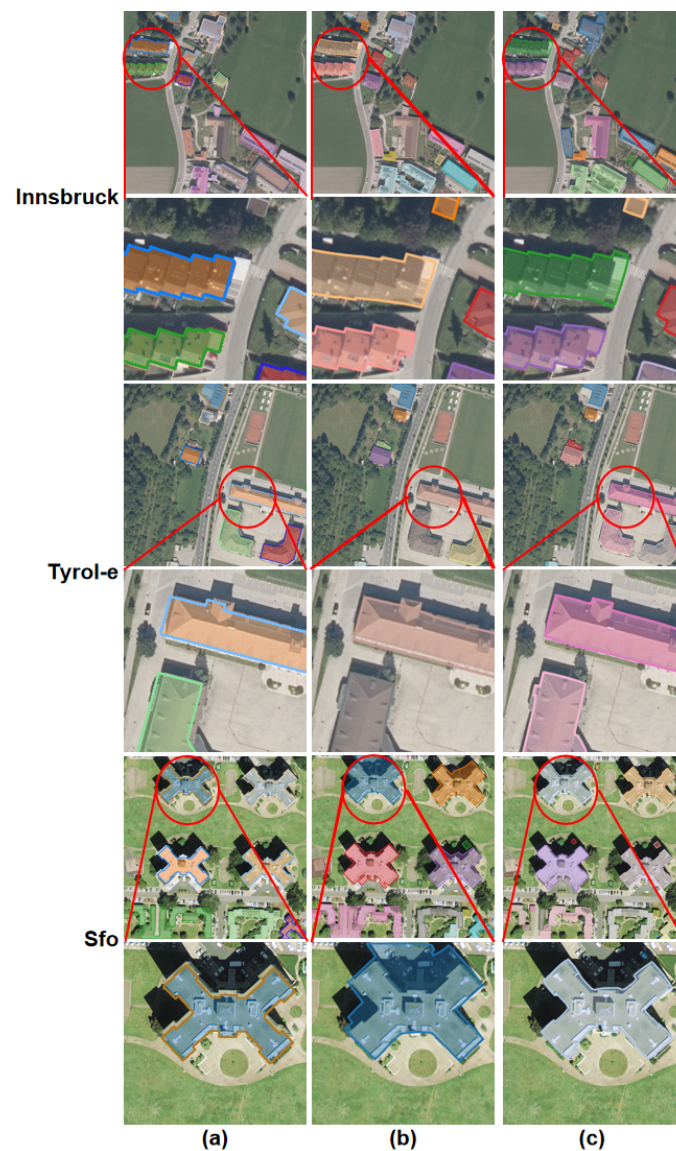


Figure 6. Schematic diagram of prediction results for different networks on the Inria dataset. From left to right: (a) Ground Truth, (b) Frame Field Network's prediction, (c) EA-DBFFN's prediction.

Table 4. Quantitative comparison of different networks on the CrowdAI dataset under instance segmentation metrics on the test set. The best results are shown in bold.

Method	AP	AP ₅₀	AP ₇₅	AR	AR ₅₀	AR ₇₅
Mask RCNN	41.9	67.5	48.8	47.6	70.8	55.5
PANet	50.7	73.9	62.6	54.4	74.5	65.2
TransBuilding	54.4	88.6	64.1	62.1	91.6	72.7
PolyMapper	55.7	86.0	65.1	62.1	88.6	71.4
FrameField	61.3	87.4	70.6	64.9	89.4	73.9
PolyWorld	63.3	88.6	70.5	75.4	93.5	83.1
EA-DBFFN	63.7	88.7	72.9	71.1	92.3	76.9

While PolyWorld excelled in AR and its variants, EA-DBFFN's AP advantage, especially at high IoU thresholds (AP_{75}), is more critical for high-precision building polygon extraction. This suggests a trade-off between recall and geometric precision, which will be further discussed in the Discussion section.

3.3.2. Analysis of Polygon Geometric Quality Metrics

To comprehensively evaluate extracted building polygon geometric quality, we used Max Tangent Angle Error (MTA), Invalid Polygon Ratio, Abnormal Convex Hull Ratio, and Edge Smoothness. Table 5 shows results for FrameField, PolyWorld, and EA-DBFFN.

Table 5. Quantitative comparison of different networks on the CrowdAI dataset under polygon geometric quality metrics on the test set. The best results are shown in bold.

Method	MTA	IPR (‰)	ACHR (‰)	ES
PolyWorld	32.9	19.44	3.45	486.7
FrameField	31.9	7.27	2.57	149.7
EA-DBFFN	30.8	6.55	2.06	133.1

Table 5 reveals EA-DBFFN's significant advantages in polygon geometric quality. Our MTA decreased to 30.8, lower than PolyWorld (32.9) and FrameField (31.9). Smaller MTA indicates better alignment of extracted polygon boundaries with ground-truth tangents, proving our advantage in capturing precise geometric shapes. EA-DBFFN's Invalid Polygon Ratio (6.55‰) and Abnormal Convex Hull Ratio (2.06‰) are significantly lower than PolyWorld (19.44‰, 3.45‰) and FrameField (7.27‰, 2.57‰). This implies better internal consistency and topological correctness in our network's masks and frame fields, substantially reducing polygon degradation or distortion and ensuring higher usability.

EA-DBFFN's Edge Smoothness metric is 133.1, significantly better than PolyWorld's 486.7 and FrameField's 149.7. A lower Edge Smoothness value implies smoother, more natural extracted building polygon boundaries, reducing jagged or irregular edges, which is crucial for high-quality GIS data production.

The geometric quality metrics also provide additional insight into the recall gap between EA-DBFFN and PolyWorld observed in Table 4. PolyWorld's higher AR of 75.4% should be interpreted with caution. Its substantially higher Invalid Polygon Ratio of 19.44‰, compared to EA-DBFFN's 6.55‰, suggests that a non-trivial proportion of its recalled instances are geometrically degraded polygons. Such polygons, despite being matched to ground truth under lower IoU thresholds, may not meet the geometric quality requirements of practical GIS applications. In this sense, PolyWorld's higher AR may partly reflect lenient matching of low-quality predictions rather than genuinely complete building detection. In contrast, EA-DBFFN applies strict geometric constraints during polygonization to filter out incomplete or irregular predictions, which contributes to its lower recall but ensures that all detected buildings are represented with accurate and regularized boundaries. This property is critical for downstream GIS tasks such as cadastral mapping and 3D city modeling, where boundary accuracy directly affects the quality of subsequent analysis.

Figure 7 intuitively demonstrates the visual effect of our method's building polygon extraction in different scenarios, further confirming the quantitative analysis results. For irregular or complex buildings, original FrameField and PolyWorld often show overly smooth or distorted boundaries (red circles). EA-DBFFN precisely captures these fine geometric features, yielding polygon boundaries that better align with ground-truth contours and sharper edges, consistent with our lower MTA and higher AP₇₅. In cases of tree canopy occlusion as shown in rows 2 and 4, vertex-prediction methods like PolyWorld often miss vertices, leading to incomplete or entirely missed buildings. Our method, however, exhibits stronger robustness, effectively identifying and completely extracting these challenging targets, thus reducing abnormal convex hulls and invalid polygons.



Figure 7. Schematic diagram of prediction results for different networks on the CrowdAI dataset. From left to right: (a) Ground Truth, (b) Frame Field Network's prediction, (c) PolyWorld's prediction, (d) EA-DBFFN's prediction.

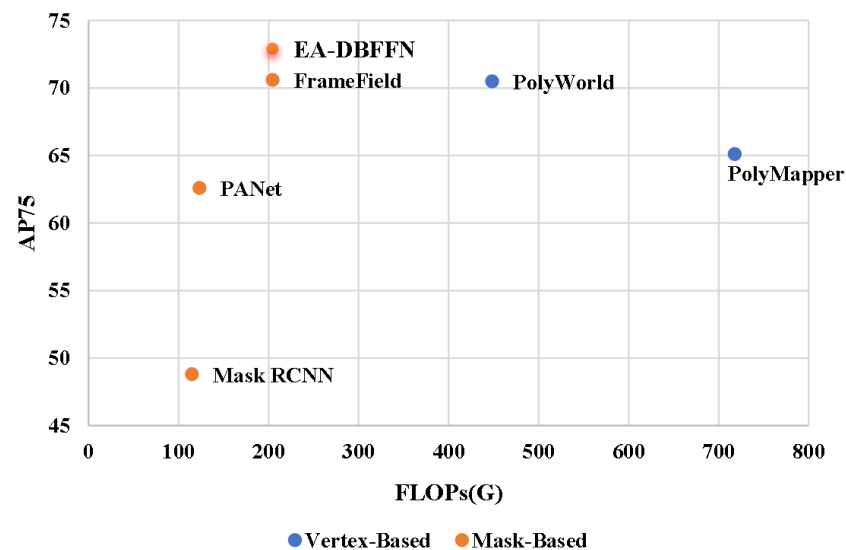
3.3.3. Analysis of Model Efficiency and Inference Speed

To comprehensively evaluate the network's computational performance, we compared its floating-point operations (FLOPs) and accuracy (AP_{75}) with representative vertex- and mask-based methods, as shown in Table 6 and Figure 8.

Vertex-based approaches such as PolyMapper and PolyWorld achieve relatively high precision but at the expense of extremely high computational costs. In contrast, mask-based methods including Mask R-CNN and PANet exhibit lower complexity but limited accuracy. FrameField achieves a good balance between these two paradigms. Building upon this foundation, our EA-DBFFN achieves the highest accuracy (AP_{75} of 72.9%) while maintaining the same FLOPs (204.4G) as FrameField, clearly standing out in the upper-left region of Figure 8. Notably, despite PolyMapper and PolyWorld consuming $3.5\times$ and $2.2\times$ more FLOPs than EA-DBFFN, respectively, and PolyMapper employing a larger VGG-16 backbone, both methods yield lower AP_{75} scores of 65.1% and 70.5%. These results demonstrate that the performance gains of EA-DBFFN are attributable to the effectiveness of the proposed architectural design rather than a larger computational budget or more powerful backbone, confirming that EA-DBFFN achieves a favorable balance between accuracy and efficiency through architectural innovation alone.

Table 6. Comparison of network computational complexity and performance on the CrowdAI test set. The best results are shown in bold.

Category	Method	Backbone	Params	FLOPs (G)	AP ₇₅
Vertex-Based	PolyMapper	VGG-16	138M	717.6	65.1
	PolyWorld	R2U-Net	39M	448.3	70.5
Mask-Based	MaskRCNN	—	—	114.7	48.8
	PANet	ResNeXt-101	86M	123.1	62.6
	FrameField	ResNet101	44M	204.4	70.6
	EA-DBFFN	ResNet101	44M	204.4	72.9

**Figure 8.** Comparison of Computational Complexity and AP₇₅ Performance Across Different Models. The top-performing model is highlighted with shading.

In summary, our EA-DBFFN achieved excellent performance on the CrowdAI dataset. Its innovative dual-branch mask prediction and dual-task mutual guidance modules significantly enhanced mask and frame field registration accuracy. Quantitatively, the method demonstrated superior instance segmentation and notably improved polygon geometric quality by reducing MTA, invalid polygon, and abnormal convex hull ratios while substantially enhancing edge smoothness. Furthermore, its high computational efficiency and exceptionally fast inference speed underscore its practicality for high-precision mapping tasks requiring efficient processing of large image data.

4. Discussion

This section will first discuss the impact of EA-DBFFN on the building polygon extraction results and then the limitations of the method and the outlook of future work.

4.1. Interpretation of Results and Comparison with Related Work

The ablation results demonstrate the effectiveness of the proposed dual-branch decoding architecture. By separating regional and boundary modeling, EA-DBFFN achieves consistent improvements over single-branch variants across both segmentation metrics and geometric quality indicators, indicating that the proposed design better balances global consistency and boundary precision under complex imaging conditions. In addition, the use of frame field representation contributes to improved geometric regularity, as reflected by the superior performance on polygon-based metrics such as MTA, IPR, ACHR, and ES.

A comparison with PolyWorld further illustrates the design philosophy underlying EA-DBFFN. While PolyWorld achieves a higher Average Recall of 75.4% compared to

EA-DBFFN's 71.1%, and a comparable AP of 63.3% versus EA-DBFFN's 63.7%, its substantially higher Invalid Polygon Ratio of 19.44‰ versus EA-DBFFN's 6.55‰, higher Edge Smoothness of 486.7 versus 133.1, and higher MTA of 32.9 versus 30.8 collectively suggest that a non-trivial proportion of PolyWorld's recalled instances are geometrically degraded polygons. Figure 7 directly compares the predictions of PolyWorld and EA-DBFFN against the ground truth on the CrowdAI dataset, covering representative scenes with varying levels of complexity, including irregular building shapes and tree canopy occlusion. In this sense, PolyWorld's higher recall may partly reflect lenient matching of low-quality predictions rather than genuinely complete building detection. For applications such as cadastral mapping, 3D city modeling, and infrastructure planning, where boundary accuracy directly affects downstream analysis quality, the trade-off between recall and geometric precision in EA-DBFFN represents a deliberate and justified design choice.

4.2. Methodological Limitations and Perspectives for Future Work

Despite the promising results demonstrated in this study, several limitations should be acknowledged. First, the edge branch relies on fixed Sobel and Laplacian operators to extract high-frequency boundary features. Although the outputs of these operators are subsequently refined through learnable convolutional layers, which helps mitigate noise sensitivity, the use of fixed non-learnable filters may still limit the model's robustness under challenging imaging conditions such as atmospheric noise and motion blur. The exploration of learnable gradient modules as a potential replacement is left for future work. Second, both datasets employed in this study share the same spatial resolution of 0.3 m per pixel, which limits the evaluation of the model's generalization capability across different imaging scales. A systematic evaluation on datasets with varying spatial resolutions remains an important direction for future research.

5. Conclusions

This study proposes an Edge-Attentive Dual-Branch Frame Field Network (EA-DBFFN) for accurate building polygon extraction from aerial imagery. The network effectively leverages the geometric fidelity of the Frame Field representation, while its novel dual-branch feature learning and edge-aware enhancement mechanisms precisely capture both regional context and high-fidelity boundary information. This approach achieves superior geometric precision and structural consistency compared with state-of-the-art methods. Evaluations on the Inria and CrowdAI datasets confirm the reliability and cross-scenario applicability of the proposed approach in diverse urban settings.

Furthermore, by improving the accuracy and automation of urban structure mapping, this work supports applications such as smart urban monitoring, land-use analysis, and resilient infrastructure planning, offering a reliable technical foundation for large-scale urban analysis and management.

Author Contributions: Conceptualization, R.H. and X.F.; methodology, R.H. and J.L.; software, R.H.; validation, R.H. and W.B.; formal analysis, R.H.; investigation, R.H., W.B., Q.G. and J.X.; resources, J.L.; data curation, R.H.; writing—original draft preparation, R.H.; writing—review and editing, X.F., J.L., Q.G. and R.Y.; visualization, R.H. and J.X.; supervision, X.F. and J.L.; project administration, X.F. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2021YFB3901201).

Data Availability Statement: The data that support the findings of this study are openly available from the CrowdAI Mapping Challenge dataset and the Inria Aerial Image Labeling dataset at

<https://www.crowdai.org/challenges/mapping-challenge> (accessed on 29 September 2025) and <https://project.inria.fr/aerialimagelabeling/> (accessed on 29 September 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu, W.B.; Ma, J.; Banzhaf, E.; Meadows, M.E.; Yu, Z.W.; Guo, F.X.; Sengupta, D.; Cai, X.X.; Zhao, B. A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. *Remote Sens. Environ.* **2023**, *291*, 113578. [CrossRef]
2. Esch, T.; Heldens, W.; Hirner, A.; Keil, M.; Marconcini, M.; Roth, A.; Zeidler, J.; Dech, S.; Strano, E. Breaking new ground in mapping human settlements from space—The Global Urban Footprint. *ISPRS J. Photogramm. Remote Sens.* **2017**, *134*, 30–42. [CrossRef]
3. Tian, Y.; Sun, X.; Niu, R.; Yu, H.; Zhu, Z.; Wang, P.; Fu, K. Fully-weighted HGNN: Learning efficient non-local relations with hypergraph in aerial imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 263–276. [CrossRef]
4. Zhang, Z.; Qian, Z.; Zhong, T.; Chen, M.; Zhang, K.; Yang, Y.; Zhu, R.; Zhang, F.; Zhang, H.; Zhou, F.; et al. Vectorized rooftop area data for 90 cities in China. *Sci. Data* **2022**, *9*, 66. [CrossRef] [PubMed]
5. Zhao, W.; Persello, C.; Stein, A. Building Instance Segmentation and Boundary Regularization from High-Resolution Remote Sensing Images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2020), Waikoloa, HI, USA, 26 September–2 October 2020; pp. 3916–3919.
6. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building Extraction from Satellite Images Using Mask R-CNN with Building Boundary Regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 247–251.
7. Wei, S.; Ji, S.; Lu, M. Toward Automatic Building Footprint Delineation From Aerial Images Using CNN and Regularization. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2178–2189. [CrossRef]
8. Xie, Y.; Zhu, J.; Cao, Y.; Feng, D.; Hu, M.; Li, W.; Zhang, Y.; Fu, L. Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1842–1855. [CrossRef]
9. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
11. Douglas, D.H.; Peucker, T.K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica* **1973**, *10*, 112–122. [CrossRef]
12. Sohn, G.; Jwa, Y.; Jung, J.; Kim, H. An implicit regularization for 3D building rooftop modeling using airborne LiDAR data. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *I-3*, 305–310. [CrossRef]
13. Daranagama, S.; Witayangkurn, A. Automatic building detection with polygonizing and attribute extraction from high-resolution images. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 606. [CrossRef]
14. Li, Z.; Wegner, J.D.; Lucchi, A. Topological map extraction from overhead images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1715–1724.
15. Zhao, W.; Persello, C.; Stein, A. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 119–131. [CrossRef]
16. Wei, S.; Ji, S. Graph Convolutional Networks for the Automated Production of Building Vector Maps From Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602411. [CrossRef]
17. Zorzi, S.; Bazrafkan, S.; Habenschuss, S.; Fraundorfer, F. PolyWorld: Polygonal Building Extraction with Graph Neural Networks in Satellite Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA, 19–24 June 2022; pp. 1938–1947.
18. Zhu, Y.; Huang, B.; Gao, J.; Huang, E.; Chen, H. Adaptive Polygon Generation Algorithm for Automatic Building Extraction. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4702114. [CrossRef]
19. Ling, H.; Gao, J.; Kar, A.; Chen, W.; Fidler, S. Fast Interactive Object Annotation With Curve-GCN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019; pp. 5257–5266.
20. Jiao, W.; Cheng, H.; Persello, C.; Vosselman, G. RoIPoly: Vectorized building outline extraction using vertex and logit embeddings. *ISPRS J. Photogramm. Remote Sens.* **2025**, *224*, 317–328. [CrossRef]

21. Huang, X.; Chen, K.; Wang, Z.; Sun, X. Instance-aware contour learning for vectorized building extraction from remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 12745–12759. [[CrossRef](#)]
22. Li, W.; Zhao, W.; Yu, J.; Zheng, J.; He, C.; Fu, H.; Lin, D. Joint semantic–geometric learning for polygonal building segmentation from high-resolution remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *201*, 26–37. [[CrossRef](#)]
23. Castrejón, L.; Kundu, K.; Urtasun, R.; Fidler, S. Annotating object instances with a Polygon-RNN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 4485–4493.
24. Acuna, D.; Ling, H.; Kar, A.; Fidler, S. Efficient interactive annotation of segmentation datasets with Polygon-RNN++. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 859–868.
25. Ballas, N.; Yao, L.; Pal, C.; Courville, A. Delving deeper into convolutional networks for learning video representations. *arXiv* **2016**, arXiv:1511.06432. [[CrossRef](#)]
26. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2017**, arXiv:1609.02907. [[CrossRef](#)]
27. Wei, S.; Zhang, T.; Ji, S. A concentric loop convolutional neural network for manual delineation-level building boundary segmentation from remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4407511. [[CrossRef](#)]
28. Wei, S.; Zhang, T.; Ji, S.; Luo, M.; Gong, J. BuildMapper: A fully learnable framework for vectorized building contour extraction. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 87–104. [[CrossRef](#)]
29. Chen, Q.; Wang, L.; Waslander, S.L.; Liu, X. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 114–126. [[CrossRef](#)]
30. Zhu, Y.; Huang, B.; Fan, Y.; Usman, M.; Chen, H. Iterative polygon deformation for building extraction. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4704314. [[CrossRef](#)]
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
32. Zhang, M.; Liu, Q.; Wang, Y. HiT: Building mapping with hierarchical transformers. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5606316. [[CrossRef](#)]
33. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV 2020)*, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 213–229.
34. Hu, Y.; Wang, Z.; Huang, Z.; Liu, Y. PolyBuilding: Polygon transformer for building extraction. *ISPRS J. Photogramm. Remote Sens.* **2023**, *199*, 15–27. [[CrossRef](#)]
35. Zorzi, S.; Fraundorfer, F. Regularization of building boundaries in satellite images using adversarial and regularized losses. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2019), Yokohama, Japan, 28 July–2 August 2019; pp. 5140–5143.
36. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [[CrossRef](#)]
37. Zorzi, S.; Bittner, K.; Fraundorfer, F. Machine-learned regularization and polygonization of building segmentation masks. In Proceedings of the International Conference on Pattern Recognition (ICPR 2020), Milan, Italy, 10–15 January 2021; pp. 3098–3105.
38. Zhao, K.; Kamran, M.; Sohn, G. Boundary regularized building footprint extraction from satellite images using deep neural network. *arXiv* **2020**, arXiv:2006.13176. [[CrossRef](#)]
39. Tang, M.; Perazzi, F.; Djelouah, A.; Ayed, I.B.; Schroers, C.; Boykov, Y. On regularized losses for weakly-supervised CNN segmentation. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 524–540.
40. Tang, M.; Djelouah, A.; Perazzi, F.; Boykov, Y.; Schroers, C. Normalized cut loss for weakly-supervised CNN segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1818–1827.
41. Sun, X.; Huang, X.; Mao, Y.; Sheng, T.; Li, J.; Wang, Z.; Lu, X.; Ma, X.; Tang, D.; Chen, K. GABLE: A first fine-grained 3D building model of China on a national scale from very high resolution satellite imagery. *Remote Sens. Environ.* **2024**, *305*, 114057. [[CrossRef](#)]
42. Girard, N.; Smirnov, D.; Solomon, J.; Tarabalka, Y. Polygonal building extraction by frame field learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Nashville, TN, USA, 20–25 June 2021; pp. 5887–5896.
43. Nguyen, V.; Ho, T.-A.; Vu, D.-A.; Anh, N.T.N.; Thang, T.N. Building footprint extraction in dense areas using super resolution and frame field learning. In Proceedings of the International Conference on Awareness Science and Technology (iCAST 2023), Da Nang, Vietnam, 1–3 November 2023; pp. 112–117.

44. Sun, X.; Zhao, W.; Maretto, R.V.; Persello, C. Building polygon extraction from aerial images and digital surface models with a frame field learning framework. *Remote Sens.* **2021**, *13*, 4700. [[CrossRef](#)]
45. Marcos, D.; Tuia, D.; Kellenberger, B.; Zhang, L.; Bai, M.; Liao, R.; Urtasun, R. Learning deep structured active contours end-to-end. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8877–8885.
46. Xu, B.; Xu, J.; Xue, N.; Xia, G.-S. HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 284–296. [[CrossRef](#)]
47. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV 2014)*, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
48. Wei, S.; Zhang, T.; Yu, D.; Ji, S.; Zhang, Y.; Gong, J. From Lines to Polygons: Polygonal Building Contour Extraction from High-Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2024**, *209*, 213–232. [[CrossRef](#)]
49. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2017), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
50. Mohanty, S.P.; Czakon, J.; Kaczmarek, K.A.; Pyskir, A.; Tarasiewicz, P.; Kunwar, S.; Rohrbach, J.; Luo, D.; Prasad, M.; Fleer, S.; et al. Deep learning for understanding satellite imagery: An experimental survey. *Front. Artif. Intell.* **2020**, *3*, 534696. [[CrossRef](#)] [[PubMed](#)]
51. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
52. Zhang, M.; Liu, Q.; Wang, W.; Wang, Y. Transbuilding: An end-to-end polygonal building extraction with transformers. In Proceedings of the IEEE International Conference on Image Processing (ICIP 2023), Kuala Lumpur, Malaysia, 8–11 October 2023; pp. 460–464.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.