

Article

TBFH: A Total-Building-Focused Hybrid Dataset for Remote Sensing Image Building Detection

Lin Yi ^{1,2,3}, Feng Wang ^{1,2} , Guangyao Zhou ^{1,2,3}, Niangang Jiao ^{1,2,*}, Minglin He ^{2,3}, Jingxing Zhu ^{1,2,3} 
and Hongjian You ^{1,2,3}

- ¹ Key Laboratory of Target Cognition and Application Technology (TCAT), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; yilin23@mails.ucas.ac.cn (L.Y.); wangfeng003020@aircas.ac.cn (F.W.); zhougy@aircas.ac.cn (G.Z.); zhujingxing20@mails.ucas.ac.cn (J.Z.); hjyou@mail.ie.ac.cn (H.Y.)
- ² Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; heminglin23@mails.ucas.ac.cn
- ³ School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China
- * Correspondence: jiaong@aircas.ac.cn

Abstract

Building extraction plays a crucial role in a variety of applications, including urban planning, high-precision 3D reconstruction, and environmental monitoring. In particular, the accurate detection of tall buildings is essential for reliable modeling and analysis. However, most existing building-detection methods are primarily trained on datasets dominated by low-rise structures, resulting in degraded performance when applied to complex urban scenes with high-rise buildings and severe occlusions. To address this limitation, we propose TBFH (Total-Building-Focused Hybrid), a novel dataset specifically designed for building detection in remote sensing imagery. TBFH comprises a diverse collection of tall buildings across various urban environments and is integrated with the publicly available WHU Building dataset to enable joint training. This hybrid strategy aims to enhance model robustness and generalization across varying urban morphologies. We also propose the KTC metric to quantitatively evaluate the structural integrity and shape fidelity of building segmentation results. We evaluated the effectiveness of TBFH on multiple state-of-the-art models, including UNet, UNetFormer, ABCNet, BANet, FCN, DeepLabV3, MANet, SegFormer, and DynamicVis. Our comparative experiments conducted on the Tall Building dataset, the WHU dataset, and TBFH demonstrated that models trained with TBFH significantly outperformed those trained on individual datasets, showing notable improvements in IoU, F1, and KTC scores as well as in the accuracy of building shape delineation. These findings underscore the critical importance of incorporating tall building-focused data to improve both detection accuracy and generalization performance.

Keywords: tall building extraction; remote sensing; hybrid dataset; machine learning



Academic Editor: Wen Liu

Received: 26 May 2025

Revised: 27 June 2025

Accepted: 30 June 2025

Published: 6 July 2025

Citation: Yi, L.; Wang, F.; Zhou, G.; Jiao, N.; He, M.; Zhu, J.; You, H. TBFH: A Total-Building-Focused Hybrid Dataset for Remote Sensing Image Building Detection. *Remote Sens.* **2025**, *17*, 2316. <https://doi.org/10.3390/rs17132316>

Copyright: © 2025 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) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Background and Motivation

In recent years, the rapid advancement of global urbanization has led to increasingly complex urban spatial structures, particularly in metropolitan and megacity regions. High-rise and super-high-rise buildings have proliferated, emerging as dominant spatial elements in urban cores. These tall structures not only improve land-use efficiency but also

serve as multifunctional vertical complexes, accommodating residential, commercial, administrative, and public services [1]. As a result, the accurate and automated identification of high-rise buildings has become a critical task in the field of urban remote sensing and spatial information extraction.

Building detection refers to the automatic identification and delineation of building boundaries or footprints from remote sensing imagery [2]. Traditional approaches often rely on low-level visual cues such as edge detection, threshold segmentation, and shape matching, which are susceptible to variations in lighting, shadows, and surface textures, thereby limiting their robustness and transferability across different urban scenarios. With the advent of deep learning, particularly the development of Convolutional Neural Networks (CNNs) [3], Residual Networks (ResNet) [4], U-Net [5], and Transformer-based architectures [6], the accuracy and automation level of building detection have significantly improved. Recent research has further leveraged semantic segmentation, object-detection, and instance segmentation techniques to enable large-scale, automated extraction of building footprints from Very-High-Resolution (VHR) remote sensing images [7].

However, while general building detection has made notable progress, high-rise building detection remains a unique and unresolved challenge due to the distinct morphological and radiometric characteristics of tall structures. Specifically, high-rise buildings exhibit large projected areas, pronounced shadows, complex rooftop geometries, and strong vertical variations, which complicate accurate delineation from overhead perspectives. Existing building datasets, such as WHU [8] and Inria [9], although widely used, primarily consist of low- to mid-rise residential or commercial buildings. These datasets offer limited representation of urban high-rise scenes, resulting in data imbalance and reduced model generalization capability when applied to metropolitan environments dominated by tall buildings.

Moreover, the complex three-dimensional structure of high-rise buildings—often accompanied by rooftop elements like water tanks, solar panels, or antennas—poses substantial challenges to conventional 2D segmentation models, which lack awareness of height or projection distortion. Occlusions from adjacent buildings, shadows, and vegetation further exacerbate detection errors, leading to missed detections or False Positives. More critically, current deep learning models are predominantly trained on 2D spatial features and rarely incorporate elevation or semantic height cues that are essential for accurate high-rise building recognition.

Beyond the limitations in existing data, the construction of a high-quality high-rise building dataset itself poses several technical challenges. First, high-rise buildings are sparsely distributed in most urban areas, making it difficult to curate a sufficient number of samples and ensure label diversity. This leads to sample scarcity and exacerbates class imbalance during training. Second, although rooftop-based annotation is a practical approach, it fails to reflect the full geometry of tall structures, especially under occlusion, projection distortion, or façade ambiguity, often resulting in inconsistent or ambiguous labels. Third, dense urban environments introduce significant visual clutter—such as surrounding buildings, roads, and vegetation—which reduces the discriminative power of models and increases the risk of False Positives.

From a model-training perspective, current deep learning architectures are primarily optimized for low-rise or generic objects and lack inductive bias toward height-related features. The domain gap between low-rise and high-rise scenarios—caused by differences in scale, angle, and structural complexity—results in domain shift and degraded performance when transferring models trained on conventional datasets to tall building-detection tasks. Furthermore, most existing models are trained on purely 2D imagery, without incorpo-

rating depth, shadow context, or height priors, which are crucial for distinguishing tall structures in complex urban environments.

To address these challenges, we propose a comprehensive framework focused specifically on tall building extraction from remote sensing imagery. Our methodology includes the construction of a Total-Building-Focused Hybrid dataset (TBFH) for remote sensing image building detection, enriched with high-resolution, manually annotated masks of high-rise buildings from representative urban regions. Furthermore, we integrate a comparative experimental pipeline to train and evaluate state-of-the-art building-detection models across our dataset and existing public datasets (e.g., WHU), highlighting the unique challenges and generalization performance associated with high-rise detection.

This work not only fills a critical gap in the current data landscape but also provides a solid foundation for future advancements in high-rise building detection and urban remote sensing analysis.

1.2. Research Objectives

Tall buildings play a vital role in modern urban environments, serving as indicators of urbanization and density. However, their automatic detection in high-resolution remote sensing imagery presents significant challenges due to complex structural forms, varying heights, and interference from vertical projection and shadows. Furthermore, most publicly available building datasets lack adequate representation of tall buildings, in terms of geographic diversity and architectural complexity. This under-representation often results in limited performance and generalization capability of existing deep learning models when applied to urban scenes containing high-rise structures.

To address this issue, we first constructed a new dataset focused on tall buildings by collecting and annotating high-resolution remote sensing imagery from multiple cities. This dataset includes imagery of tall buildings from both dense urban areas and suburban regions, covering a variety of building types and spatial layouts. We then combined this newly constructed dataset with the WHU building dataset to form a Total-Building-Focused Hybrid (TBFH) dataset that maintains a balance between tall building features and general building features. This new dataset ensures that the building-detection model can not only learn to detect tall buildings with high accuracy but also maintain good generalization capabilities for common low-rise buildings, thereby enhancing the model's robustness and adaptability in complex urban environments.

To fully leverage this hybrid dataset, we designed a novel training pipeline that employs a balanced mixing strategy, combining tall building-specific samples with general building samples during training. This balanced training strategy ensures that the model is exposed to a diverse range of building features, enabling it to capture unique tall building characteristics, such as complex geometric structures, vertical projection effects, and shadow influences, while also improving its performance on low-rise buildings. This method does not require modifications to the original model architecture, but the balanced strategy strengthens the model's capability for tall building detection in complex urban settings.

We conducted extensive experiments across multiple representative urban areas to validate the effectiveness of the proposed approach. The results show that the building-detection model trained on the hybrid dataset significantly outperformed models trained on single-source datasets in various evaluation metrics, especially in tall building detection, including Intersection over Union (IoU), boundary F1 score, and Recall. Our model made significant progress in distinguishing tall buildings from the surrounding environment, particularly in dense urban areas where occlusion and shadow effects between buildings severely impact detection performance.

The main contributions of this study are as follows:

1. Construction of the TBFH dataset: We proposed the TBFH dataset by combining our tall building annotations with the WHU Building dataset. This enhances data diversity and enables more robust training across varied urban environments.
2. Keypoint Topological Consistency (KTC) metric: We propose the KTC metric to quantitatively evaluate the structural integrity and shape fidelity of building segmentation results, offering a complementary assessment to conventional pixel-wise metrics.
3. Empirical demonstration of performance gains: Our experiments show that models trained on TBFH achieve significant improvements in IoU, F1 score, Recall, and KTC. TBFH better captures the geometry of tall buildings and reveals valuable insights for urban building extraction.

In conclusion, this study proposes TBFH (Total-Building-Focused Hybrid), a novel dataset specifically curated for high-rise building detection in urban remote sensing imagery. By integrating diverse tall building samples and merging them with the widely used WHU Building dataset, TBFH aims to bridge the data gap and improve model robustness in dense urban environments. The introduction of TBFH provides a comprehensive and high-quality benchmark to support more accurate detection, segmentation, and modeling of complex urban structures.

2. Related Work

2.1. Building-Extraction Methods

Building extraction is a fundamental task in remote sensing image analysis, aiming to identify and delineate the locations and shapes of buildings from high-resolution remote sensing imagery. Accurate building extraction supports a wide range of applications, including urban planning, population estimation, and disaster monitoring. With the rapid development of remote sensing technologies and the increasing availability of high-resolution imagery, building-extraction methods have evolved considerably, transitioning from conventional image-processing techniques to advanced deep learning-based approaches.

Early building-extraction methods primarily relied on traditional image-processing techniques, such as edge detection and image segmentation. Operators like Sobel [10] and Canny [11] were commonly employed to extract building contours by detecting edges in grayscale images [12]. Although these techniques are simple and computationally efficient, they are highly sensitive to noise, illumination changes, and shadow interference, leading to poor performance in complex or cluttered environments. To improve reliability, shadow and texture-based methods were proposed [13], leveraging building shadows as auxiliary features to distinguish buildings, especially in nadir-view imagery. However, these methods still suffer from limited generalizability and robustness in diverse urban scenes, particularly when occlusions or non-uniform textures are present [14]. Rule-based classification techniques, which incorporate spectral and geometric cues, were also explored to improve discrimination between buildings and background. Yet, the handcrafted nature of these rules often results in inadequate performance in highly heterogeneous cityscapes.

With the emergence of machine learning, more robust and adaptable methods became feasible. Supervised learning algorithms such as Support Vector Machines (SVMs) [15] have been utilized to classify image pixels based on their spectral and spatial characteristics. SVMs are effective in handling high-dimensional data but are often constrained by computational costs and sensitivity to class imbalance, which limits scalability. Random Forest (RF) [16], an ensemble learning method, improved upon single classifiers by aggregating predictions from multiple decision trees, offering enhanced stability and adaptability across varying image conditions. Nonetheless, RF requires a large number of labeled samples and significant computational resources. Similarly, the K-Nearest Neighbors (KNN)

algorithm [17], while simple and interpretable, struggles with efficiency and accuracy when applied to large-scale remote sensing data.

The introduction of deep learning has led to a paradigm shift in building extraction. Convolutional Neural Networks (CNNs) [18,19] have demonstrated superior performance by learning hierarchical feature representations directly from image data, eliminating the need for handcrafted features. Among these, the U-Net architecture [5] has become a cornerstone for semantic segmentation tasks due to its symmetric encoder–decoder structure, which enables effective localization and context modeling. U-Net has been widely adopted in building extraction, particularly for delineating low-rise structures. With the integration of deep residual learning frameworks such as ResNet [20], deeper networks became trainable, addressing vanishing gradient issues and improving the extraction of complex structural patterns in high-resolution urban imagery.

More recently, Transformer-based architectures have introduced a new frontier in remote sensing image analysis. The original Transformer model [6], known for its ability to capture long-range dependencies, has been adapted for vision tasks. Vision Transformer (ViT) [21] redefined image understanding by representing images as sequences of patches, achieving competitive performance on various image-recognition tasks. Building upon this, UNetFormer [22] integrates the strengths of both U-Net and Transformer architectures. This hybrid design combines CNNs' local feature sensitivity with Transformers' global modeling capabilities, allowing for improved segmentation accuracy, particularly for high-rise buildings that are often occluded or shadowed. The model shows strong robustness in diverse urban conditions, making it well-suited for complex building-extraction tasks.

In summary, building-extraction techniques have progressed significantly, evolving from traditional rule-based systems to powerful deep learning models. CNN-based approaches remain the backbone of most modern frameworks, while Transformer-based architectures are pushing the boundaries further. Models like UNetFormer demonstrate that combining local and global feature-extraction strategies can greatly enhance both the accuracy and robustness of high-rise building detection in remote sensing imagery.

2.2. Dataset Limitations

The performance of building-extraction models is fundamentally influenced by the quality, diversity, and representativeness of the training datasets on which they are developed. High-resolution remote sensing imagery captures urban environments with inherently complex features, including varied architectural styles, dense spatial configurations, and diverse land use patterns. The generalization capability of a model—particularly its ability to perform reliably across different urban contexts—relies heavily on the extent to which such complexities are reflected in the training data.

Over the past decade, several publicly available datasets have significantly contributed to the development of building-extraction algorithms. Notable among these are the WHU Building dataset [8], the Inria Aerial Image dataset [9], and the Massachusetts Building dataset [23]. These datasets have served as foundational benchmarks, promoting methodological innovation and comparative evaluation. However, a critical examination reveals two major limitations that restrict their applicability in more complex and diverse urban scenarios.

First, the geographic and morphological coverage of these datasets is relatively narrow. A majority of the samples are derived from urban regions in developed countries—especially in Europe and North America—where city layouts tend to be highly regular, and where buildings are predominantly low- to mid-rise structures situated in well-planned residential or commercial zones. This results in an under-representation of urban areas from the Global South, including parts of Asia, Latin America, and Africa, where

rapid urbanization, informal development, and architectural heterogeneity are more prevalent. In particular, dense cityscapes in countries like China, India, and Brazil—featuring high-rise buildings, irregular block arrangements, and vertical complexity—are scarcely captured in existing benchmarks.

Second, there is a notable deficiency in the inclusion of high-rise buildings, such as skyscrapers, residential towers, and large-scale commercial complexes. These structures are characterized by substantial height, complex three-dimensional geometries, heterogeneous façade elements, and pronounced shadow effects. Such features pose significant challenges for 2D-based extraction models, which often struggle with occlusion, projection distortion, and the accurate delineation of rooftops and building boundaries. The absence of these building types from standard datasets leads to a critical performance gap when models are deployed in real-world urban settings dominated by vertical development.

In addition to dataset limitations, current evaluation methods—mainly 2D metrics like IoU and F1 score—are inadequate for capturing the structural and vertical complexity of buildings. They overlook key features such as façade details and roof geometry, leading to performance drops in high-rise scenarios, especially under shadows, occlusions, or off-nadir views [24].

To improve high-rise building extraction, it is crucial to develop diverse, high-resolution datasets that capture varied urban forms, including dense vertical developments and informal areas. Existing datasets are foundational but limited, and addressing this gap is key to enhancing model robustness and generalization in real-world urban settings.

2.3. Hybrid Training Strategies

In recent years, hybrid dataset training strategies have been developed to improve the robustness and generalization of building-extraction models. By combining datasets from multiple sources or applying multi-domain adaptation, these approaches enable models to learn diverse spatial and semantic features, enhancing performance in complex urban settings.

Hybrid training typically involves two methods: direct concatenation of standardized datasets such as WHU and Inria for joint training, and domain-adaptation techniques that reduce feature distribution discrepancies between domains [25], including the use of domain discriminators and adversarial losses to improve adaptability.

While hybrid training has been applied to low-rise building extraction, its use in high-rise scenarios remains limited due to sample scarcity and CNNs' difficulties handling occlusions and structural variations.

2.4. Tall-Building-Extraction Challenges

Tall buildings are vital in modern cities, but their automatic extraction is challenging due to occlusion, shadows, complex rooftops, and proximity to other structures. These factors make high-rise buildings hard to distinguish in remote sensing data. Moreover, existing datasets lack enough high-rise samples, limiting model performance in such scenarios. Thus, specialized datasets and advanced modeling are needed.

Shadows and occlusions from skyscrapers obscure rooftops and boundaries, especially in oblique views. High-rise buildings often cluster with complex rooftops, complicating complete boundary extraction. Low grayscale and texture contrast with the ground further hinder separation using traditional features.

Previous work has used image enhancement (e.g., shadow compensation [26]), multi-modal fusion (e.g., DSM + imagery [27]), and structural modeling (e.g., multi-scale attention networks) to improve high-rise detection. However, training is usually dominated by low-rise samples, causing accuracy drops on tall buildings. Hence, building representative high-

rise datasets and adopting hybrid training are essential to boosting accuracy and practical model use.

In summary, developing dedicated high-rise datasets alongside hybrid training strategies is crucial for enhancing extraction accuracy and model generalization.

3. Methodology

The overall workflow of our proposed approach is shown in Figure 1, comprising three main stages: image preprocessing, tall building dataset construction, and model training. First, high-resolution multispectral images are generated by fusing PAN and MS data, followed by geometric correction using DSM and DOM to ensure spatial alignment across target regions such as Shandong and Beijing. Next, large-area remote sensing images are cropped into 512×512 patches and manually annotated to create a tall building dataset with binary masks, which is split into training, validation, and testing sets (7:2:1). Finally, several building-detection models are trained on different datasets—including the proposed tall building dataset, WHU, and the combined TBFH dataset—to evaluate their generalization and adaptability for high-rise building extraction.

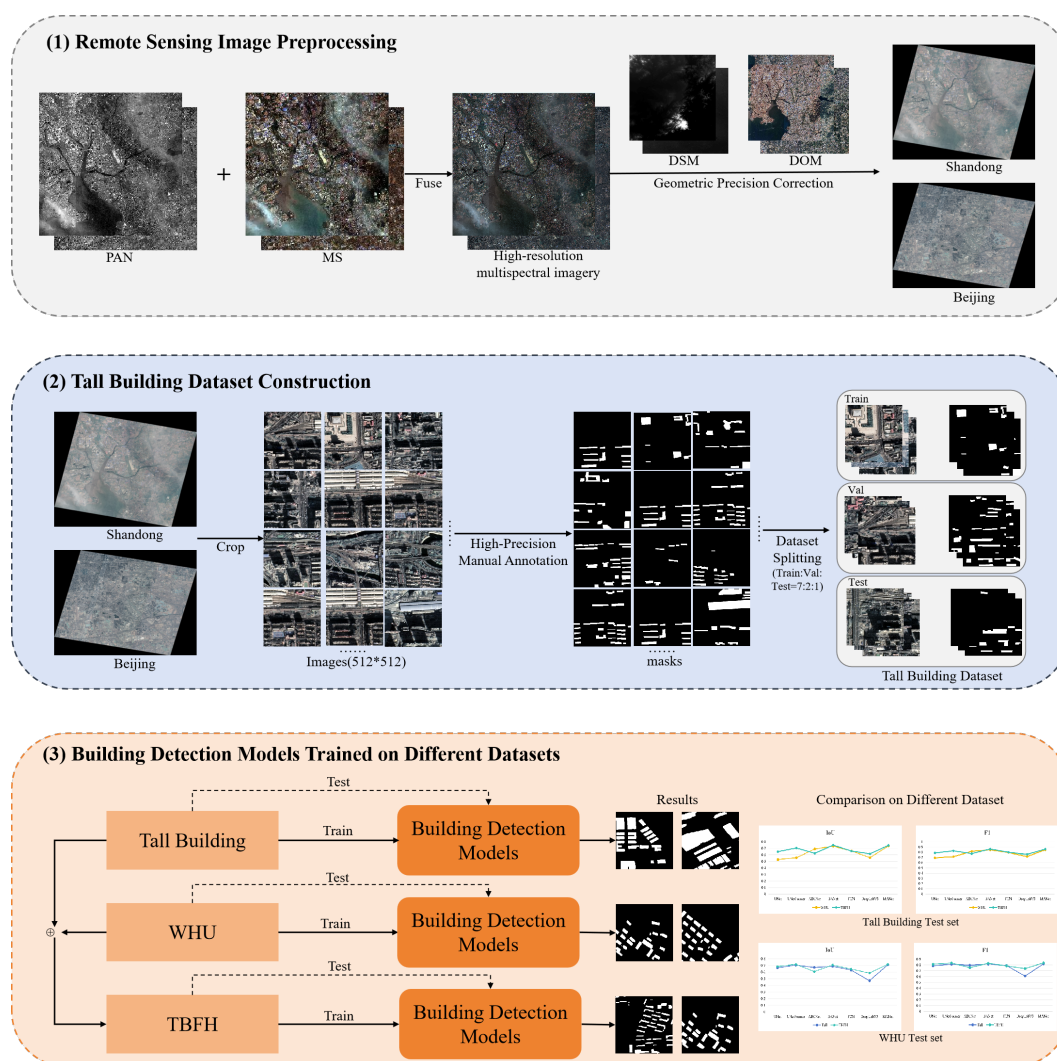


Figure 1. Overview of the proposed TBFH dataset for tall building detection from remote sensing imagery. The pipeline consists of three main stages: (1) image preprocessing through multispectral fusion and geometric correction; (2) tall building dataset construction with high-resolution annotation and dataset splitting; and (3) training and evaluation of building-detection models across multiple datasets for comparative analysis.

3.1. Dataset

3.1.1. WHU Building Dataset

The WHU Satellite dataset [8] is a high-quality, multi-source building dataset specifically designed for building-extraction tasks. Created and publicly provided by Ji et al. [8], the dataset contains building samples from both aerial and satellite imagery, covering an area of 1000 km². It provides both raster labels and vector maps. The dataset is designed to offer rich and accurate samples for building-extraction research, supporting the training and evaluation of deep learning models.

The satellite imagery portion of the WHU dataset consists of two subsets. The first subset includes images from cities around the world, with a variety of remote sensing resources, including QuickBird, Worldview series, IKONOS, and ZY-3 satellites. The resolution of these images ranges from 0.3 m to 2.5 m, covering diverse geographic environments and architectural styles (see Figure 2). The second subset is composed of six adjacent satellite images, covering an area of 550 km², with a ground resolution of 2.7 m. This subset is primarily used to evaluate and develop deep learning methods for assessing the generalization ability in geographic regions with similar architectural styles but different data sources.

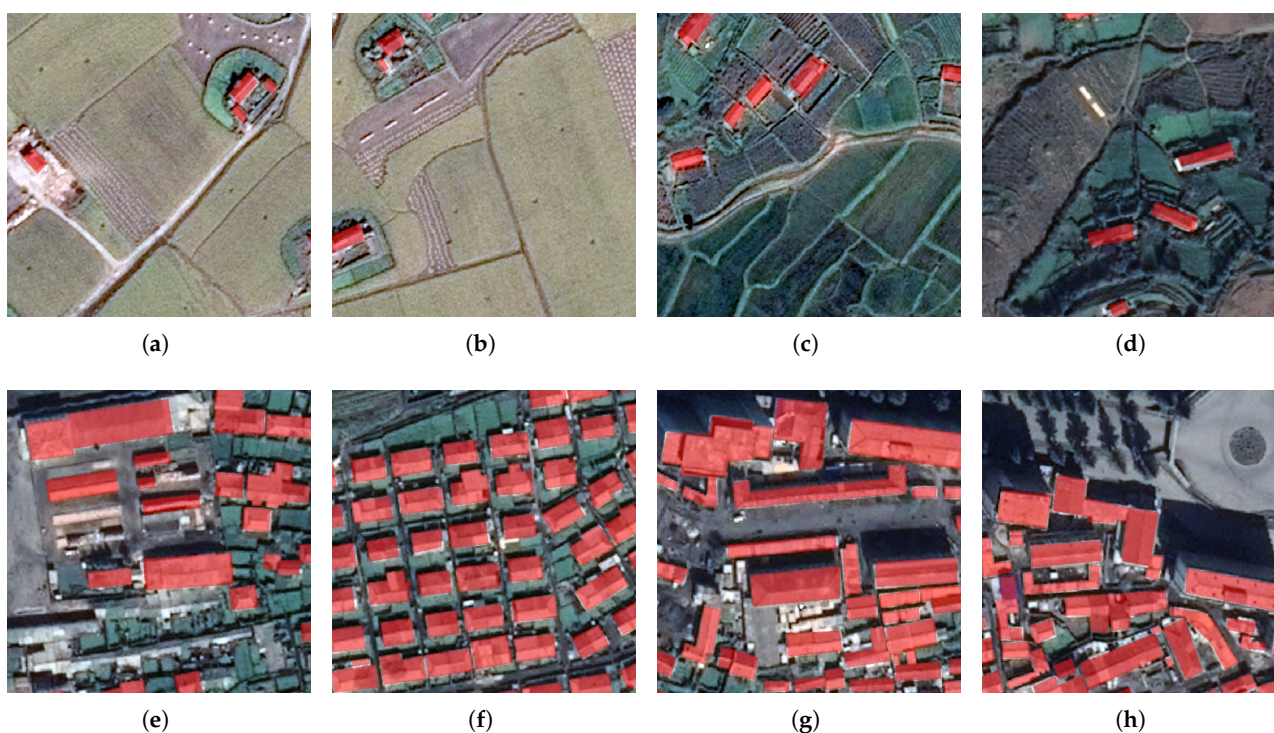


Figure 2. Examples of the WHU dataset with diverse architectures from different regions: (a,b) agricultural areas, (c,d) rural areas, (e,f) towns and villages, (g,h) urban areas.

To ensure high quality, all building vector maps were manually drawn. The researchers used ArcGIS software (10.8.1) to edit and verify the original data, producing high-precision building boundaries. In the first subset, 204 512 × 512 pixel image tiles were annotated, with resolutions ranging from 0.3 m to 2.5 m. The second subset contains 29,085 buildings, and the images were seamlessly cropped into 17,388 512 × 512 pixel tiles. For training and testing purposes, the size of these tiles is compatible with mainstream GPU graphics cards. Among these, 21,556 buildings (13,662 tiles) are used for training, and the remaining 7529 buildings (3726 tiles) are used for testing.

3.1.2. Tall Building Dataset

To better support the study of tall building extraction, we have carefully constructed a dedicated remote sensing dataset specifically focused on tall buildings. This dataset aims to address the scarcity of tall building samples in existing publicly available datasets and to provide deep learning models with more diverse and abundant training samples.

To standardize the definition of high-rise buildings in our dataset and ensure scientific rigor and applicability, we explicitly define high-rise buildings as those meeting at least one of the following criteria [28]:

1. Height exceeding 50 m;
2. More than 15 floors;
3. A building footprint area greater than 300 m² combined with noticeable vertical structures visible in the imagery (e.g., clearly visible façades, cast shadows, rooftop details).

These thresholds are aligned with common standards in urban planning and remote sensing research to ensure consistency and representativeness in identifying high-rise buildings.

The dataset was built based on high-resolution satellite imagery with meticulous manual annotation. The image data were acquired by the Chinese Gaofen-7 satellite, specifically covering typical urban scenes within China that feature a high concentration of medium- and large-scale buildings. These regions, located in Beijing and Shandong Provinces, represent characteristic cityscapes with complex architectural layouts and dense construction, making them ideal for studies focused on urban structure analysis and large building extraction. A total of 27,907 individual buildings were extracted, with a spatial resolution of 0.6 m and a total coverage area of 242 km². Among the 27,907 annotated buildings, 18,789 were identified as high-rise structures, accounting for approximately 67.3% of the total. This substantial proportion highlights the dataset's emphasis on tall buildings, making it particularly well-suited for training and evaluating models dedicated to high-rise building detection—an important aspect.

The annotation process was carried out by a team of professional remote sensing analysts who followed standardized procedures to ensure high accuracy and consistency. Building boundaries were precisely delineated at the pixel level to meet the stringent requirements of deep learning models. To align with common model input formats, all images were divided into fixed-size tiles of 512 × 512 pixels. Furthermore, the imagery underwent a series of preprocessing steps, including image fusion to generate high-resolution multispectral images and geometric correction, which improved the overall image quality and spatial consistency, providing superior input data for model training.

The dataset covers diverse urban regions, including residential, industrial, and cultural zones, ensuring both structural diversity and representative urban forms. The images encompass diverse geographical and urban environments, ensuring both the representativeness and diversity of the dataset.

Figure 3 illustrates representative samples from the constructed Tall Building dataset, showcasing diverse architectural styles and urban forms across different geographic regions. Figure 3a–d correspond to high-rise building instances in Beijing, characterized by dense urban layouts, modern skyscrapers, and complex rooftop structures. These examples illustrate the diversity and complexity of high-rise structures in dense urban settings. Figure 3e–h present examples from Shandong, where tall buildings exhibit a mixture of traditional and contemporary designs, with varied spatial distribution and rooftop morphologies.

These samples highlight the dataset's rich diversity, in terms of building height, architectural shape, and surrounding urban context. This includes a wide range of low- to high-rise structures, complex and irregular building geometries, and varied environmental settings such as dense city centers, mixed-use developments, and heterogeneous

background textures. Such diversity is essential for training deep learning models that are robust and capable of generalizing across different urban scenarios, particularly when detecting and delineating tall buildings that often exhibit complex rooftop features, cast strong shadows, and are influenced by occlusions or varying viewing angles.

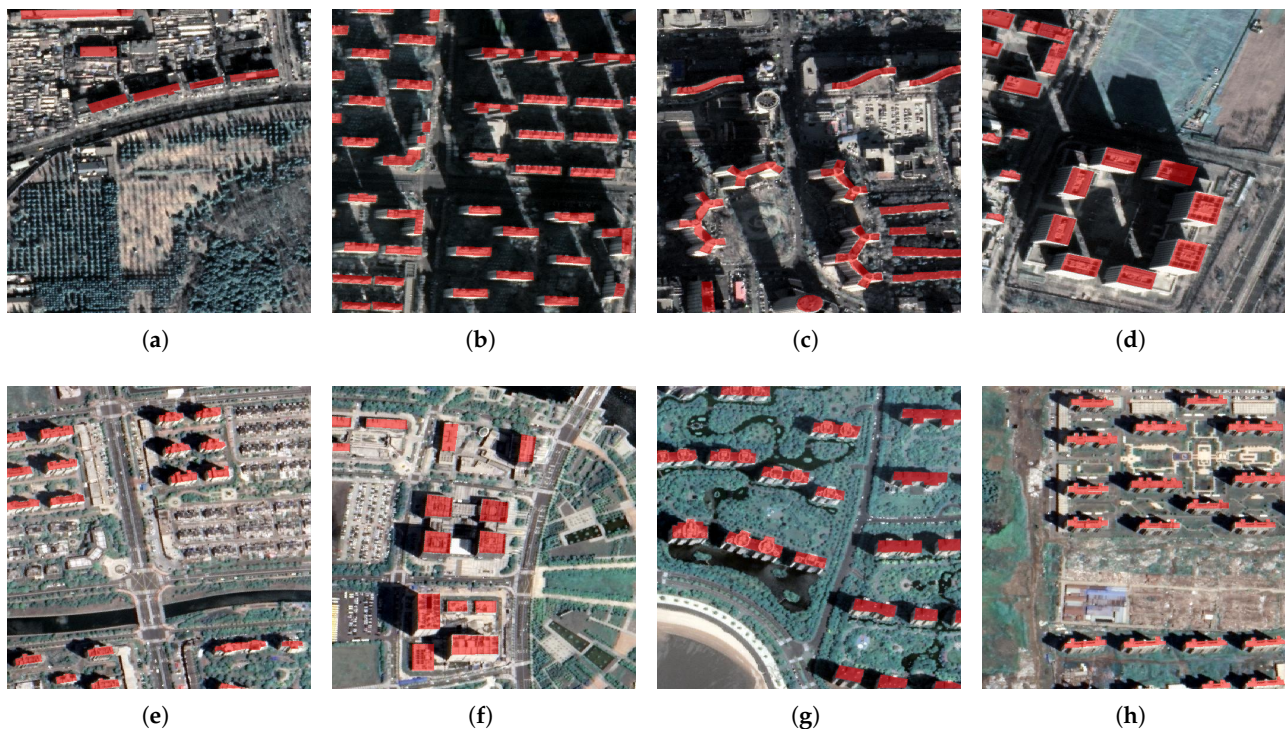


Figure 3. Examples of the Tall Building dataset with different architectures from cities: (a–d) Beijing; (e–h) Shandong.

In addition to our dataset, other widely used building extraction datasets include WHU, ISPRS (Vaihingen and Potsdam), Massachusetts, and Inria. Table 1 compares these datasets, in terms of spatial resolution, coverage area, data sources, number of image tiles, and label formats.

Table 1. General comparison between our dataset and other open-source datasets.

Datasets	GCD (m)	Area (km ²)	Sources	Tiles	Pixels	Label Format
Tall Building (ours)	0.6	242	sat	2520	512 × 512	raster
WHU	0.075/2.7	450/550	aerial/sat	8189/17,388	512 × 512	vector/raster
ISPRS	0.05/0.09	2/11	aerial	24/16	6000 × 6000/11,500 × 7500	raster
Massachusetts	1.00	340	aerial	151	1500 × 1500	raster
Inria	0.3	405 *	aerial	180	5000 × 5000	raster

* Another test dataset covering 405 km² is used for evaluating submitted algorithm with unpublished labels.

By integrating the WHU dataset with the newly constructed Tall Building dataset, we propose a novel hybrid dataset, namely TBFH, which serves as a robust and representative benchmark for evaluating the performance of building-extraction methods, particularly in the context of tall buildings. Compared to the existing datasets, TBFH provides greater structural and geographic diversity, enabling more rigorous and realistic assessments of algorithmic generalization and robustness. This benchmark dataset is expected to facilitate standardized comparisons in future research and drive the development of more effective solutions for complex urban environments.

3.2. Dataset Fusion Strategy

The TBFH dataset was constructed through a complementary integration of 8189 samples from the WHU Building dataset and 2520 high-rise samples from a self-built subset. To mitigate the low-rise dominance inherent in WHU, a diversity-aware stratified sampling strategy was employed during the selection of the high-rise samples, ensuring comprehensive coverage of buildings with varying heights, densities, and urban contexts. This approach promotes a more balanced distribution across building height categories and enhances the generalization capability of models trained on TBFH.

A two-stage data-augmentation pipeline is further applied to enrich the training data. In the first stage, geometric transformations such as random scaling and smart cropping are used to simulate variations in building size, perspective, and spatial distribution, while preserving structural integrity. The second stage introduces photometric and structural perturbations—including color jittering, flipping, rotation, and noise injection—through a custom transformation module. These augmentations expand the diversity of the training samples and improve the model's robustness under various visual conditions.

As described in Section 3.1.1, the WHU dataset comprises a wide range of satellite images with spatial resolutions ranging from 0.3 m to 2.5 m. To accommodate this variability, a simple resolution-adaptive segmentation module is introduced. During the detection process, multi-scale feature information is extracted, and hierarchical feature fusion is applied to integrate representations across different resolution levels. This design helps alleviate the adverse effects caused by resolution inconsistency. In future iterations of the dataset, additional improvements such as resolution normalization or domain adaptation strategies may be considered to further enhance cross-domain consistency.

3.3. Building-Detection Models

3.3.1. U-Net

U-Net [5] is a classic encoder–decoder architecture widely used in semantic segmentation, featuring skip connections that fuse low-level spatial details with high-level semantics. This design enhances localization and boundary precision, making U-Net suitable for extracting objects with complex shapes in high-resolution images.

The model's multi-scale convolutional structure enables it to capture detailed boundary information. The diverse shapes and textures of high-rise buildings in the TBFH dataset further enrich its learning, improving segmentation accuracy in complex scenes.

3.3.2. UNetFormer

UNetFormer [22] integrates Transformer blocks into a U-Net framework, combining local feature extraction with global self-attention for improved spatial understanding. It retains fine-grained localization through U-Net's structure while capturing long-range dependencies via the Transformer bottleneck.

The model's ability to model global context makes it well-suited for TBFH, where tall buildings exhibit complex spatial arrangements and occlusions, enhancing shape completeness and segmentation precision.

3.3.3. ABCNet

ABCNet [29] features a dual-branch architecture that jointly captures semantic context and boundary features. An attention-based fusion module adaptively balances these cues to improve edge delineation.

ABCNet combines attention mechanisms with convolution, providing strong contextual awareness. The rich tall building samples in TBFH enable it to better capture spatial correlations between façades and rooftops.

3.3.4. BANet

BANet [30] introduces a boundary-aware branch that enhances object edges via feature fusion and attention mechanisms, improving delineation in densely built environments.

BANet employs dual attention modules to enhance spatial and channel features, highlighting tall building contours and details, improving detection robustness in complex scenes.

3.3.5. FCN

Fully Convolutional Networks (FCNs) [31] replace fully connected layers with convolutional ones to achieve pixel-level predictions. Their use of skip connections allows the integration of spatial and semantic features, making them effective for dense prediction tasks in remote sensing.

As a fundamental Fully Convolutional Network, FCN has limited capacity for modeling long-range dependencies but can reliably learn rich local features supported by the diverse building instances in TBFH, serving as a stable baseline.

3.3.6. DeepLabV3

DeepLabV3 [32] utilizes atrous convolutions and an Atrous Spatial Pyramid Pooling (ASPP) module to aggregate multi-scale context while maintaining resolution. This allows effective handling of varying object sizes in remote sensing imagery.

DeepLabV3's incorporation of atrous convolution and spatial pyramid pooling to capture multi-scale context effectively addresses large-scale variations of buildings in TBFH and enables precise localization.

3.3.7. MANet

MANet [33] enhances multi-scale representations via spatial and channel attention mechanisms. It selectively emphasizes salient features and suppresses background noise, improving segmentation in cluttered scenes.

As MANet utilizes multi-scale attention to enhance perception of features across spatial scales, it adapts well to the diversity and complexity of TBFH, improving boundary and detail recognition.

3.3.8. SegFormer

SegFormer [34] combines a hierarchical Transformer encoder (MiT) with a lightweight decoder, enabling efficient capture of global and multi-scale features in remote sensing imagery. Its Transformer-based architecture operates without positional encoding, offering flexibility and generalization across domains.

SegFormer's independence from strict positional encoding imparts robustness when handling the complex and variable morphologies of high-rise buildings. The rich, high-resolution samples of tall buildings provided by the TBFH dataset synergize well with SegFormer's multi-scale modeling capabilities, thereby enhancing the model's performance in high-rise building-extraction tasks.

3.3.9. DynamicVis

DynamicVis [35] adopts dynamic token sampling and a bidirectional information flow framework to model semantic structures efficiently, improving semantic understanding of complex structures in remote sensing images. By focusing on essential spatial tokens and aggregating global context, it balances expressiveness with computational cost. Its adaptive design maintains computational efficiency while focusing on salient features, making it particularly suitable for capturing the vertical structures and spatial contextual information of high-rise buildings. Leveraging the finely annotated tall building regions in the TBFH

dataset, DynamicVis can fully exploit the data richness to improve recognition accuracy and generalization in high-rise building detection.

3.4. Performance Evaluation

In this section, we quantitatively evaluate the performance of the building-detection model on the tall-building-extraction task. Four widely used metrics are adopted for comprehensive assessment: **Intersection over Union (IoU)**, **Precision**, **Recall**, and **F1 score**. The definitions are as follows:

Precision (P) and **Recall** (R) are defined as

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

where TP (True Positive) denotes the number of correctly predicted building pixels, FP (False Positive) denotes the number of non-building pixels incorrectly predicted as buildings, and FN (False Negative) represents building pixels incorrectly classified as non-building.

The **F1 score** (F_1), which balances Precision and Recall, is given by

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (3)$$

Intersection over Union (IoU), also known as the Jaccard Index, measures the overlap between the predicted and ground-truth building regions:

$$IoU = \frac{TP}{TP + FP + FN} \quad (4)$$

In addition to traditional pixel-level evaluation metrics such as IoU and F1 score, we propose a novel structural evaluation metric named Keypoint Topological Consistency (KTC) to assess the geometric similarity between predicted and ground-truth building contours. This metric is particularly designed for high-rise buildings, whose rooftop outlines often exhibit complex, non-rectilinear shapes that cannot be sufficiently evaluated by region-based measures alone.

Given a predicted building mask and a ground-truth mask, we first extract their external contours and apply polygonal approximation to obtain sets of keypoints:

$$\mathcal{P} = \{p_1, p_2, \dots, p_m\}, \quad \mathcal{G} = \{g_1, g_2, \dots, g_n\}$$

where $\mathcal{P}$ and $\mathcal{G}$ denote the predicted and ground-truth keypoint sets, respectively.

The KTC metric is defined as a weighted sum of three components:

$$KTC = \alpha \cdot \text{Sim}_{\text{count}} + \beta \cdot \text{Sim}_{\text{topo}} + \gamma \cdot \text{Sim}_{\text{loc}}, \quad \alpha + \beta + \gamma = 1$$

where:

- $\text{Sim}_{\text{count}}$ measures the similarity in keypoint quantity:

$$\text{Sim}_{\text{count}} = \frac{\min(m, n)}{\max(m, n)}$$

- Sim_{topo} compares the structural adjacency of keypoints by constructing undirected cyclic graphs and computing the difference between their adjacency matrices:

$$\text{Sim}_{\text{topo}} = 1 - \frac{\|\mathbf{A}_{\mathcal{P}} - \mathbf{A}_{\mathcal{G}}\|_1}{k^2}$$

where $\mathbf{A}_{\mathcal{P}}$ and $\mathbf{A}_{\mathcal{G}}$ are $k \times k$ adjacency matrices (after truncation or padding), and $\|\cdot\|_1$ denotes the matrix 1-norm.

- Sim_{loc} evaluates the spatial consistency of matched keypoint pairs using a nearest-neighbor strategy:

$$\text{Sim}_{\text{loc}} = 1 - \frac{1}{K} \sum_{i=1}^K \frac{\|p_i - g_i^*\|_2}{D}$$

where g_i^* is the closest ground-truth keypoint to p_i , K is the number of matched pairs, and D is the diagonal length of the union bounding box for scale normalization:

$$D = \sqrt{(x_{\max} - x_{\min})^2 + (y_{\max} - y_{\min})^2} + \epsilon$$

with a small constant ϵ (e.g., 1×10^{-6}) to prevent division by zero.

When both $\mathcal{P}$ and $\mathcal{G}$ are empty, we define $\text{KTC} = 1.0$ to reflect complete structural agreement.

All metrics are computed on the test set. A higher value of Precision, Recall, F1 score, IoU, and KTC indicates better model performance.

Furthermore, qualitative comparisons between the predicted building masks and the ground-truth annotations are provided to visually demonstrate the effectiveness of the proposed method. Representative examples are selected from diverse urban environments to ensure a fair and comprehensive evaluation.

4. Experiment

4.1. Experiment Setup

All experiments were conducted on a workstation with an NVIDIA RTX 4090 GPU (24 GB VRAM) and a 16-core Intel® Xeon® Gold 6430 CPU, running Ubuntu 22.04 LTS. The models were implemented in PyTorch 2.1.0 with Python 3.10.

4.2. Dataset Preparation

The proposed Tall Building dataset covers approximately 242 km² at a Ground Sampling Distance (GSD) of 0.6 m. The dataset was randomly split into training (70%), validation (15%), and testing (15%) subsets.

4.3. Implementation Details

The models were trained for up to 100 epochs with early stopping based on validation performance. The training batch size was set to 8; for validation, it was 1. The base learning rate was 6×10^{-4} (backbone: 6×10^{-5}), with a weight decay of 2.5×10^{-4} . The Adam optimizer was used, with a learning rate schedule adjusted according to validation loss.

4.4. Comparison Methods

In order to evaluate the performance of the proposed method, this study compared it with several advanced building-extraction methods. The selection of these methods was based on their effectiveness and representativeness in semantic segmentation and building-extraction tasks. The methods for comparison were UNet, UNetFormer, ABCNet, BANet, FCN, DeepLabV3, MANet, SegForemr, DynamicVis.

All the models were evaluated under identical experimental conditions to ensure a fair comparison. This included using the same training, validation, and testing dataset splits, as well as employing consistent preprocessing and data augmentation strategies. To compre-

hensively assess the models' performance, we utilized several standard evaluation metrics, including Intersection over Union (IoU), Precision, Recall, and F1 score.

The experiments conducted were as follows:

1. Training on the Tall Building dataset: In this experiment, all the models were trained using only the proposed Tall Building dataset. This setup was intended to assess the performance of each model on a dataset specifically designed for tall building-extraction tasks.
2. Training on the WHU dataset: The second experiment involved training the UNetFormer model exclusively on the WHU dataset, which is a widely recognized benchmark in remote sensing tasks. This setup allowed for an evaluation of the model's performance on a more general dataset that contains a variety of building types.
3. Training on a mixed dataset: The final experiment combined both the Tall Building and WHU datasets into a mixed training set. The UNetFormer model was trained using this combined dataset to investigate the impact of using a more diverse dataset on the model's performance and its generalization ability across different scenarios.

5. Results

5.1. Performance Evaluation on Tall Building Dataset

Figure 4 shows a comparison between U-Net, UNetFormer, and several other models (such as ABCNet, BANet, FCN, DeepLabV3, MANet, SegFormer, DynamicVis). On the Tall Building dataset, the experimental results demonstrate that the UNetFormer model and MANet excel in building-detection tasks.

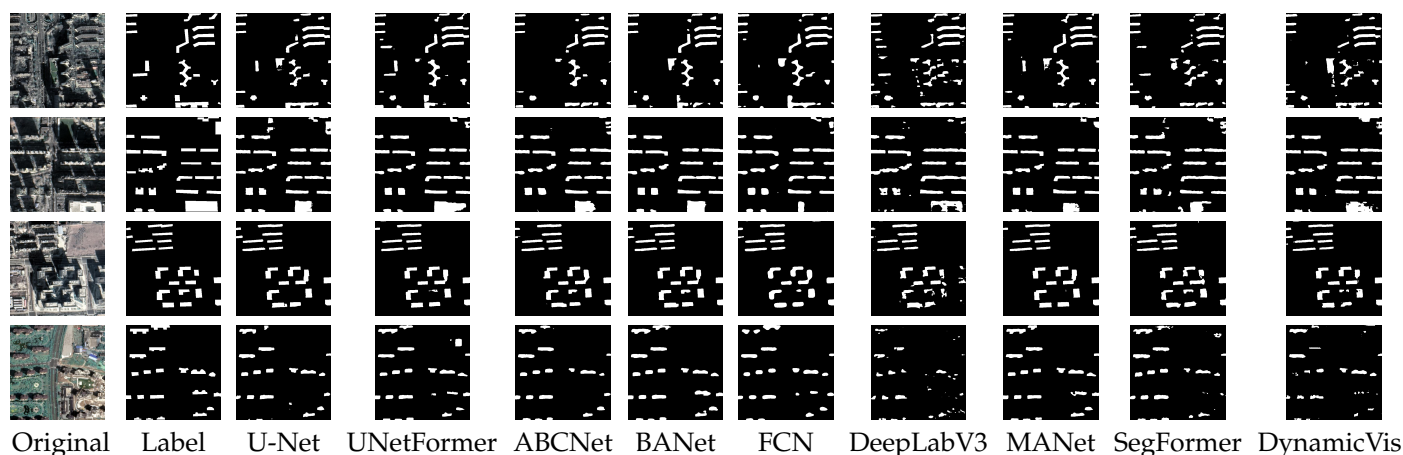


Figure 4. Comparison of building-detection models on the Tall Building dataset.

As shown in Table 2, when trained solely on the Tall Building dataset, most of the models achieved high Precision, indicating strong capability in identifying tall structures. However, the Recall values varied, reflecting differences in completeness of detection. MANet attained the highest F1 score (0.8303), followed closely by UNetFormer and BANet, highlighting their balanced accuracy. The KTC scores further validated the consistency between the predicted and ground-truth rankings of pixel-wise confidence, with UNetFormer (0.8648) and MANet (0.8608) achieving the highest correlation. These results suggest that models with better spatial awareness and multi-scale feature integration generalize more effectively to the complex morphology of tall buildings.

Table 2. Performance comparison of building-detection models trained on Tall Building dataset.

Model	IoU	Precision	Recall	F1	KTC
UNet	0.6635	0.8413	0.7584	0.7977	0.8090
UNetFormer	0.7063	0.8759	0.7849	0.8279	0.8648
ABCNet	0.6695	0.8481	0.7608	0.8021	0.8302
BANet	0.6862	0.8683	0.7660	0.8139	0.8431
FCN	0.6307	0.8384	0.7180	0.7735	0.7749
Deeplabv3	0.4692	0.6493	0.6284	0.6387	0.7214
MANet	0.7098	0.8759	0.7892	0.8303	0.8608
SegFormer	0.6565	0.8280	0.7602	0.7926	0.7992
DynamicVis	0.6426	0.8298	0.7402	0.7824	0.7784

5.2. Performance Evaluation on WHU Dataset

On the WHU dataset, MANet further demonstrated its strong modeling capabilities in the task of building extraction. As shown in Figure 5, a visual comparison of the different model outputs reveals that MANet achieved more precise delineation of building contours, especially in areas with intricate textures and dense urban backgrounds. Its ability to retain structural integrity and suppress False Positives contributed to more consistent and detailed segmentation results. These qualitative observations align with the quantitative findings, highlighting MANet’s robustness in diverse urban scenarios.

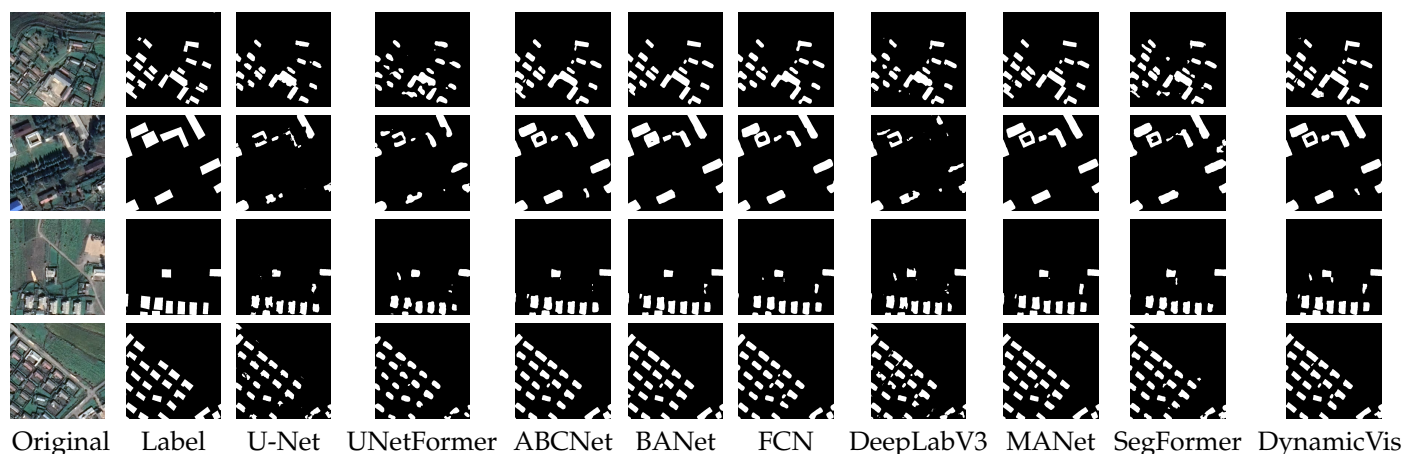
**Figure 5.** Comparison of building-detection models on the WHU dataset.

Table 3 presents the performance of all the models trained on the WHU dataset. Overall, the models achieved strong Precision scores, reflecting reliable identification of standard building regions in relatively clean urban scenes. MANet yielded the highest F1 score (0.8482), closely followed by BANet and DynamicVis, indicating superior capability in capturing building extent while maintaining boundary integrity. The Keypoint Topological Consistency (KTC) values further support this observation, with MANet (0.8714) and BANet (0.8675) achieving the highest consistency between predicted and ground-truth spatial rankings. This suggests that models with strong spatial reasoning and attention mechanisms not only segment more accurately but also align better with the underlying structural order of buildings. In contrast, models like UNet and Deeplabv3 show lower Recall and KTC, revealing limitations in capturing complete structures under varying textures. These findings highlight the WHU dataset’s adequacy for general building segmentation while underscoring its limitations in modeling complex urban morphology.

Table 3. Performance comparison of building-detection models trained on WHU dataset.

Model	IoU	Precision	Recall	F1	KTC
UNet	0.5286	0.8078	0.6047	0.6916	0.7773
UNetFormer	0.5562	0.8558	0.6137	0.7148	0.8385
ABCNet	0.6907	0.8441	0.7917	0.8171	0.8535
BANet	0.7329	0.8518	0.8400	0.8458	0.8675
FCN	0.6587	0.8845	0.7207	0.7942	0.8426
DeepLabv3	0.5572	0.8454	0.6205	0.7157	0.7236
MANet	0.7365	0.8704	0.8272	0.8482	0.8714
SegFormer	0.6741	0.8089	0.8019	0.8054	0.8367
DynamicVis	0.7137	0.8594	0.8081	0.8330	0.8401

5.3. Performance Evaluation on TBFH Dataset

The training and testing on this dataset demonstrated UNetFormer’s excellent generalization ability. Compared to traditional models such as ABCNet, FCN, and DeepLabV3, UNetFormer showed substantial improvements across multiple evaluation metrics. As illustrated in Figure 6, on the mixed dataset, UNetFormer not only captured building edges and contours accurately but also effectively ignored the interference from complex backgrounds, maintaining high prediction accuracy.

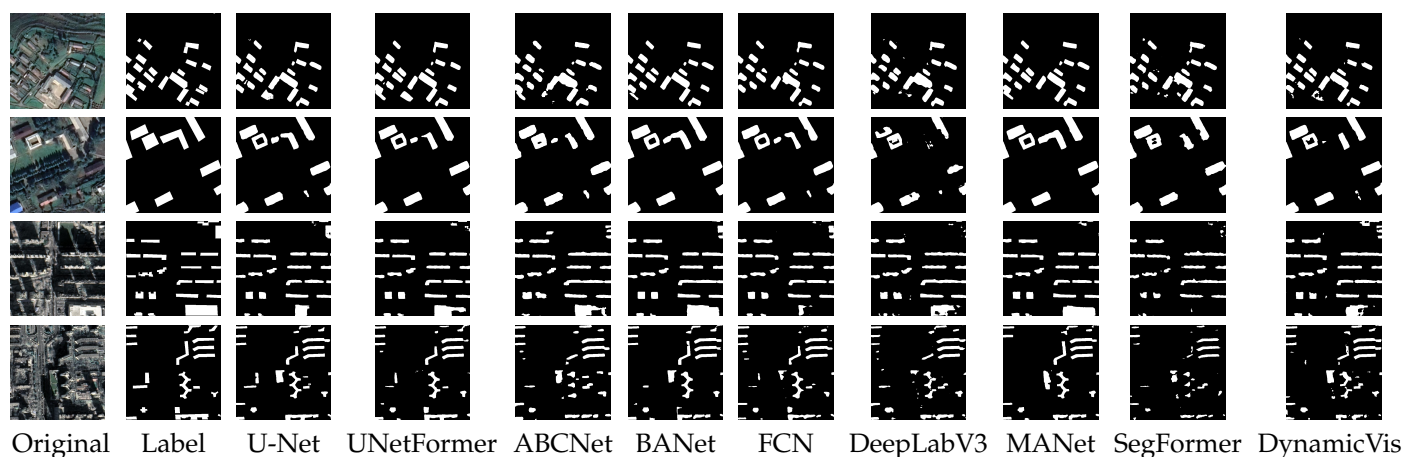
**Figure 6.** Comparison of building-detection models on the mixed dataset.

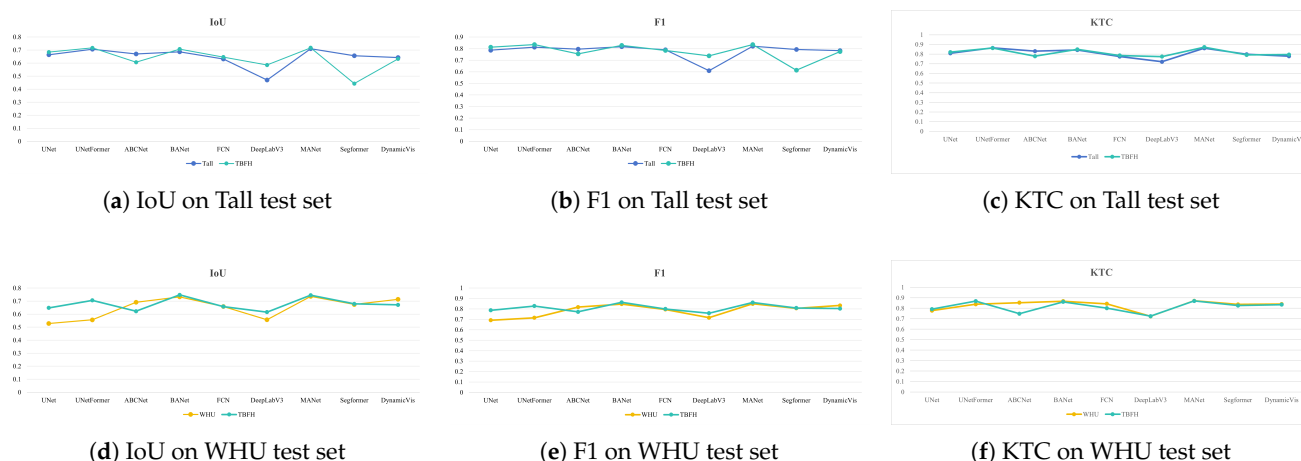
Table 4 summarizes the model performances trained on TBFH and tested on both the Tall Building and WHU datasets. The mixed training led to balanced and superior results across different scenarios. UNetFormer and MANet notably achieved high F1 scores above 0.83 on both test sets, indicating robust adaptability. Meanwhile, their high Keypoint Topological Consistency (KTC) values (above 0.86) reflected strong spatial consistency between predictions and ground truth, underscoring reliable structural ranking and boundary alignment.

This is evidence that the dataset’s diversity and complexity facilitate learning of varied building forms and urban contexts, enhancing both accuracy and generalization. Overall, TBFH serves as an effective benchmark to improve model robustness and applicability in real-world building-extraction tasks.

Table 4. Performance comparison of building-detection models trained on TBFH.

Test Data	Tall Building					WHU				
Model	IoU	Precision	Recall	F1	KTC	IoU	Precision	Recall	F1	KTC
UNet	0.6841	0.8515	0.7767	0.8124	0.8202	0.6481	0.8648	0.7212	0.7865	0.7924
UNetFormer	0.7162	0.8787	0.7948	0.8346	0.8624	0.7056	0.8680	0.7904	0.8274	0.8684
ABCNet	0.6062	0.8824	0.6976	0.7549	0.7788	0.6227	0.8633	0.7087	0.7721	0.7479
BANet	0.7067	0.8657	0.7937	0.8281	0.8489	0.7481	0.8782	0.8477	0.8627	0.8614
FCN	0.6455	0.8619	0.7200	0.7846	0.7862	0.6580	0.9068	0.7146	0.7990	0.8016
DeepLabv3	0.5851	0.8120	0.6768	0.7383	0.7744	0.6154	0.8592	0.6894	0.7585	0.7251
MANet	0.7170	0.8725	0.8010	0.8352	0.8727	0.7456	0.8870	0.8365	0.8610	0.8704
SegFormer	0.4427	0.9157	0.4615	0.6137	0.7920	0.6786	0.8122	0.8048	0.8085	0.8259
DynamicVis	0.6319	0.8464	0.7138	0.7745	0.7948	0.6713	0.8690	0.7469	0.8034	0.8350

To evaluate model generalization across domains, we compared performance on the Tall Building and WHU test sets using IoU and F1 metrics (Figure 7). The models trained on the proposed TBFH dataset consistently outperformed those trained on the traditional datasets, demonstrating better cross-domain robustness and segmentation accuracy.

**Figure 7.** Comparison of IoU, F1, and KTC scores for models trained on different datasets (Tall, WHU, and TBFH), evaluated on the Tall and WHU test sets.

Our experimental results show that the TBFH-trained models achieved notable improvements on the Tall Building test set. For example, UNetFormer’s IoU increased from 0.7063 to 0.7162, and its F1 score improved from 0.8129 to 0.8346, while KTC rose to 0.8624, indicating better spatial consistency. MANet attained the best performance with an F1 score of 0.8352 and the highest KTC of 0.8727, reflecting its strong ability to capture complex tall building structures and maintain reliable boundary alignment.

More pronounced gains appeared on the WHU test set, illustrating the enhanced generalization from mixed-domain training. UNet’s IoU improved significantly from 0.5286 to 0.6481, with F1 increasing from 0.6916 to 0.7865 and KTC reaching 0.7924. Similar improvements occurred for UNetFormer and BANet. However, ABCNet showed a slight performance drop on WHU, suggesting its architecture might be more sensitive to domain shifts introduced by tall building data and may require further adaptation.

Figure 8 shows a visual comparison of building extraction results on the Tall and WHU test sets. The models trained on the TBFH dataset produced more precise building outlines, especially for tall and irregular structures, and they reduced both missed detections and False Positives. This enhanced performance was mainly due to the TBFH dataset’s richer feature diversity and balanced sample distribution, which improved the models’ ability to generalize across different urban forms and building scales.

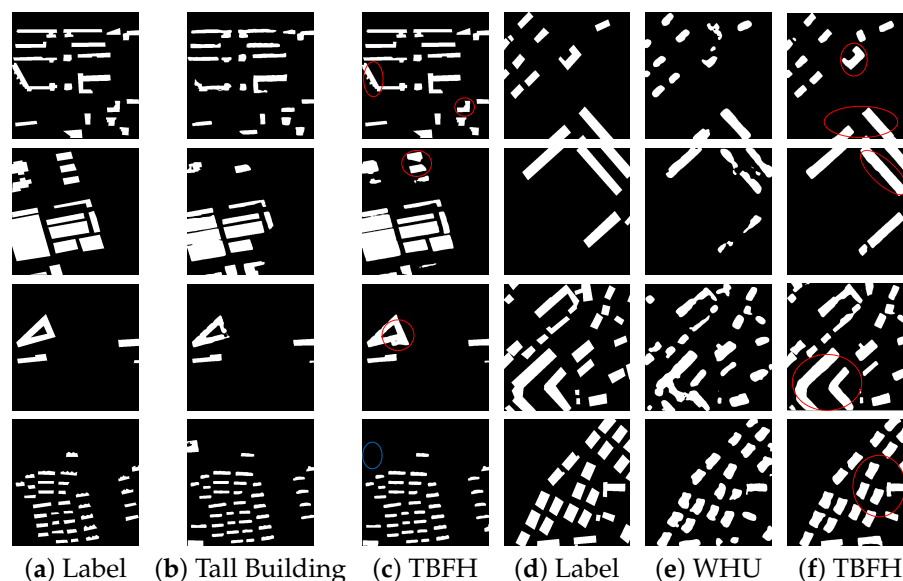


Figure 8. Visual comparison of building extraction results: (a–c) show ground truth, prediction from Tall Building-trained model, and the TBFH-trained model on the Tall Building test set, respectively; (d–f) show ground truth, prediction from WHU-trained model, and the TBFH-trained model on the WHU test set, respectively.

Overall, the TBFH-trained models demonstrated more balanced and robust performance, validating the hybrid dataset’s effectiveness in mitigating domain bias and enhancing generalization across diverse building scenes.

6. Discussion

This study presents a comprehensive experimental evaluation highlighting the effectiveness of the proposed TBFH (Total-Building-Focused Hybrid) dataset for building detection. Through extensive comparisons across the custom-built Tall Building dataset, the WHU dataset, and the proposed hybrid dataset (TBFH), the results consistently demonstrate that models trained on TBFH achieve superior generalization and detection accuracy, regardless of the specific network architecture employed.

Notably, the TBFH dataset is carefully constructed to include tall buildings from diverse geographic locations, architectural styles, and imaging conditions (e.g., varying seasons, viewing angles, and lighting). This diversity enables deep learning models to learn a broader spectrum of building features, which contributes to improved robustness when handling complex urban environments, occlusions, and variable scene contexts. Compared to single-domain datasets such as WHU or our original Tall Building set, training on TBFH significantly improves Intersection over Union (IoU) and F1 scores across multiple architectures, including U-Net, ABCNet, FCN, and UNetFormer, as evidenced by our experiments on both the Tall Building and WHU test sets.

The TBFH dataset also encompasses a wide range of building scales and morphological complexities—from isolated high-rises to dense clusters of mid-rise buildings. Such variation helps reduce model bias toward specific building types and improves performance across both low- and high-rise urban scenes. Moreover, models trained on TBFH are better equipped to reduce False Positives and False Negatives, leading to more accurate shape extraction and reduced error propagation in downstream tasks like 3D modeling or change detection.

While the proposed TBFH-based models demonstrated enhanced robustness and generalization across diverse urban environments, several challenges remain unresolved. In particular, misclassifications were still observed in complex or extreme scenarios. These in-

cluded cases of severe occlusion, where large buildings obscured neighboring structures; low-resolution imagery, which limited the visibility of fine architectural details; and oblique viewing angles, which distorted the appearance of buildings and their contours. Moreover, urban regions with highly dense and heterogeneous layouts continue to present difficulties for accurate segmentation due to overlapping shadows, irregular geometries, and limited separation between adjacent structures.

As illustrated in Figure 9a–d, all four examples exhibited segmentation errors caused by shadow artifacts resulting from tall building occlusion. These shadows often blended into the background or overlapped with nearby structures, leading to partial omission of building footprints, boundary deformation, or False Positives in shadow regions. Such issues highlight the limitations of 2D-based semantic segmentation when dealing with 3D structural complexities, and they underscore the need for incorporating contextual cues—such as shadow modeling or elevation priors—to further enhance model performance in challenging urban scenarios.

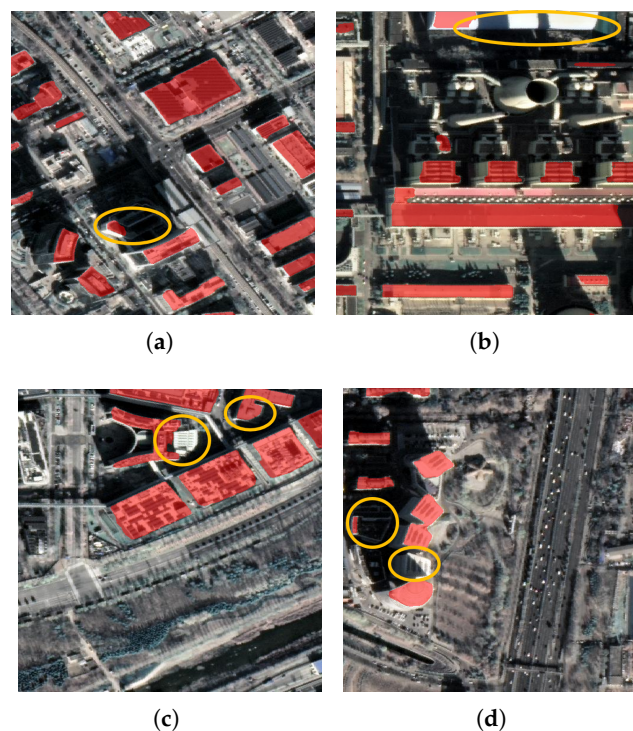


Figure 9. Representative failure cases in tall building extraction. Subfigures (a–d), the blending of shadows into the background results in partial omission of building footprints. The yellow circles indicate the regions that were not detected due to shadow interference.

In terms of computational cost, while high-capacity models like UNetFormer offer strong performance, they also demand more memory and processing power. However, the observed performance gains across various architectures suggest that the improvements are largely attributable to the dataset itself rather than to any specific network design.

Future work will focus on further expanding the TBFH dataset by incorporating data from additional regions, seasons, and imaging modalities (e.g., depth, SAR), with the goal of enhancing model adaptability to even more diverse real-world scenarios. Additionally, combining TBFH with lightweight models and exploring efficient training strategies (e.g., knowledge distillation, model pruning) may help balance detection accuracy with deployment feasibility.

7. Conclusions

In this study, we introduced the TBFH (Total-Building-Focused Hybrid) dataset, a hybrid dataset designed to improve building-detection models by capturing a diverse set of tall buildings across various geographical locations, scales, and environmental conditions. Our experiments demonstrated that models trained on the TBFH dataset consistently outperformed those trained on single-domain datasets, exhibiting superior generalization capabilities and enhanced detection accuracy on the Tall Building and WHU test sets.

The TBFH dataset's diversity in building structures, viewing angles, and environmental contexts allows deep learning models to learn a broader range of building features, improving their robustness in complex urban environments. Furthermore, the inclusion of varying building heights and scales in TBFH enables the model to effectively handle both low-rise and high-rise buildings, minimizing errors such as False Positives and False Negatives. These findings underscore the value of hybrid datasets in addressing domain bias and improving model performance across diverse scenarios.

Despite the overall success of models trained on TBFH, challenges remain in extreme situations, such as heavily occluded buildings, low-resolution images, or urban areas with high building density. Nevertheless, the TBFH dataset provides a solid foundation for further advancements in building detection and generalization.

Future work will focus on expanding the TBFH dataset by incorporating data from additional regions and imaging modalities and by optimizing models for real-time applications. Additionally, combining TBFH with lightweight model architectures will ensure that these advancements can be applied effectively in resource-constrained environments. Ultimately, the proposed TBFH dataset presents a significant step toward more accurate and robust building detection in diverse urban and rural settings.

Author Contributions: Conceptualization, H.Y.; Data curation, L.Y. and M.H.; Funding acquisition, F.W. and H.Y.; Methodology, L.Y. and N.J.; Project administration, G.Z.; Supervision, N.J. and H.Y.; Validation, L.Y.; Writing—original draft, L.Y.; Writing—review and editing, N.J., M.H. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in the GitHub repository at <https://github.com/yilin-gif/TBFH>, accessed on 1 April 2025. This repository contains the TBFH dataset used for building detection in remote sensing imagery, as described in this article.

Acknowledgments: The authors would like to express their sincere gratitude to the Group of Photogrammetry and Computer Vision (GPCV) at Wuhan University for their efforts in collecting and providing the valuable building dataset. Special thanks go to Shunping Ji for his contributions and support. This dataset has significantly enhanced our research capabilities and outcomes.

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

References

1. Dabove, P.; Daud, M.; Olivotto, L. Revolutionizing Urban Mapping: Deep Learning and Data Fusion Strategies for Accurate Building Footprint Segmentation. *Sci. Rep.* **2024**, *14*, 13510. [CrossRef] [PubMed]
2. Yuan, Q. Building Rooftop Extraction from High Resolution Aerial Images Using Multiscale Global Perceptron with Spatial Context Refinement. *Sci. Rep.* **2025**, *15*, 6499. [CrossRef]
3. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
4. Li, R.; Wang, Y.; Liu, Y. Study on the Classification of Building Based on ResNet. In Proceedings of the 11th International Conference on Information Systems and Computing Technology, ISCTech 2023, Qingdao, China, 30 July–1 August 2023; Institute of Electrical and Electronics Engineers Inc.: Qingdao, China, 2023; pp. 458–462. [CrossRef]

5. 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; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [\[CrossRef\]](#)
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
7. Khatua, A.; Bhattacharya, A.; Aithal, B.H. Automated Georeferencing and Extraction of Building Footprints from Remotely Sensed Imagery Using Deep Learning. In Proceedings of the 10th International Conference on Geographical Information Systems Theory, Applications and Management, GISTAM 2024, Angers, France, 2–4 May 2024; Science and Technology Publications, Ltd.: Angers, France, 2024; pp. 128–135. [\[CrossRef\]](#)
8. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [\[CrossRef\]](#)
9. 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), Fort Worth, TX, USA, 23–28 July 2017; IEEE: Piscataway, NJ, USA, 2017.
10. Wang, K. Edge Detection of Inner Crack Defects Based on Improved Sobel Operator and Clustering Algorithm. In *Recent Trends in Materials and Mechanical Engineering, Mechatronics and Automation, PTS 1-3*; Luo, Q., Ed.; Trans Tech Publications Ltd.: Baech, Switzerland, 2011; pp. 467–471. [\[CrossRef\]](#)
11. Yuan, L.; Xu, X. Adaptive Image Edge Detection Algorithm Based on Canny Operator. In Proceedings of the 2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS), Harbin, China, 21–23 August 2015; pp. 28–31. [\[CrossRef\]](#)
12. Wang, J.; Yang, X.; Qin, X.; Ye, X.; Qin, Q. An Efficient Approach for Automatic Rectangular Building Extraction from Very High Resolution Optical Satellite Imagery. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 487–491. [\[CrossRef\]](#)
13. Irvin, R.B.; McKeown, D.M. Methods for Exploiting the Relationship between Buildings and Their Shadows in Aerial Imagery. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 1564–1575. [\[CrossRef\]](#)
14. Ok, A.O. Automated Detection of Buildings from Single VHR Multispectral Images Using Shadow Information and Graph Cuts. *ISPRS J. Photogramm. Remote Sens.* **2013**, *86*, 21–40. [\[CrossRef\]](#)
15. Mountrakis, G.; Im, J.; Ogole, C. Support Vector Machines in Remote Sensing: A Review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [\[CrossRef\]](#)
16. Belgiu, M.; Drăguț, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [\[CrossRef\]](#)
17. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [\[CrossRef\]](#)
18. Jiao, W.; Persello, C.; Vosselman, G. PolyR-CNN: R-CNN for End-to-End Polygonal Building Outline Extraction. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 33–43. [\[CrossRef\]](#)
19. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#)
20. Vasavi, S.; Sri Somagani, H.; Sai, Y. Classification of Buildings from VHR Satellite Images Using Ensemble of U-Net and ResNet. *Egypt. J. Remote Sens. Space Sci.* **2023**, *26*, 937–953. [\[CrossRef\]](#)
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
22. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [\[CrossRef\]](#)
23. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013.
24. Wang, J.; Meng, L.; Li, W.; Yang, W.; Yu, L.; Xia, G.-S. Learning to Extract Building Footprints from Off-Nadir Aerial Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1294–1301. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Bruzzone, L.; Marconcini, M. Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 770–787. [\[CrossRef\]](#)
26. Zhou, T.; Fu, H.; Sun, C.; Wang, S. Shadow Detection and Compensation from Remote Sensing Images under Complex Urban Conditions. *Remote Sens.* **2021**, *13*, 699. [\[CrossRef\]](#)
27. Sun, G.; Chen, Y.; Huang, J.; Ma, Q.; Ge, Y. Digital Surface Model Super-Resolution by Integrating High-Resolution Remote Sensing Imagery Using Generative Adversarial Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 10636–10647. [\[CrossRef\]](#)
28. Council on Tall Buildings and Urban Habitat (CTBUH). CTBUH Official Website. Available online: <https://www.ctbuh.org/> (accessed on 23 June 2025).

29. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive Bilateral Contextual Network for Efficient Semantic Segmentation of Fine-Resolution Remotely Sensed Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [\[CrossRef\]](#)
30. Zhou, Q.; Qiang, Y.; Mo, Y.; Wu, X.; Latecki, L.J. BANet: Boundary-Assistant Encoder-Decoder Network for Semantic Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25259–25270. [\[CrossRef\]](#)
31. Tian, T.; Chu, Z.; Hu, Q.; Ma, L. Class-Wise Fully Convolutional Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 3211. [\[CrossRef\]](#)
32. Heryadi, Y.; Irwansyah, E.; Miranda, E.; Soeparno, H.; Herlawati; Hashimoto, K. The Effect of Resnet Model as Feature Extractor Network to Performance of DeepLabV3 Model for Semantic Satellite Image Segmentation. In Proceedings of the 2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS), Jakarta, Indonesia, 7–9 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 74–77. [\[CrossRef\]](#)
33. He, P.; Jiao, L.; Shang, R.; Wang, S.; Liu, X.; Quan, D.; Yang, K.; Zhao, D. MANet: Multi-Scale Aware-Relation Network for Semantic Segmentation in Aerial Scenes. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5624615. [\[CrossRef\]](#)
34. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021.
35. Chen, K.; Liu, C.; Chen, B.; Li, W.; Zou, Z.; Shi, Z. DynamicVis: An Efficient and General Visual Foundation Model for Remote Sensing Image Understanding. *arXiv* **2025**, arXiv:2503.16426.

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.