Improving Building Footprint Extraction Using NAIP and 3DEP Lidar Derived Features with Deep Learning

Jung Kuan Liu [1]*, Rongjun Qin [2,3], Samantha Arundel[1], Lexie Yang[4]

[1] U.S. Geological Survey, Center of Excellence for Geospatial Information Science, PO Box 25046, MS510, Denver 80225, USA (jliu@usgs.gov, sarundel@usgs.gov)

[2] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, USA (qin.324@osu.edu)

[3] Department of Electrical and Computer Engineering, The Ohio State University, Columbus, USA

[4] Computing and Communicational Sciences Directorate, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA (yangh@ornl.gov)

KEYWORDS: building footprints, NAIP, 3DEP, lidar, deep learning.

ABSTRACT:

Accurate building footprint extraction is critical for applications ranging from population estimation to disaster management. Although optical imagery provides detailed spectral information, it often struggles with shadows, occlusions, and background clutter in dense urban environments. Lidar data, by contrast, offer precise elevation and structural attributes but face challenges such as variable point density and noise. This study integrates multispectral imagery from the U.S. Department of Agriculture (USDA) National Agriculture Imagery Program (NAIP) with lidar-derived feature height and intensity from the U.S. Geological Survey (USGS) 3D Elevation Program (3DEP) to improve footprint extraction using a U-Net–based deep learning model. A six-band input stack (RGB, near-infrared, height, intensity) was developed, normalized, and tiled for training and evaluation against Microsoft Global Building Footprints (GBF). Results from the Houston, TX test site show that the six-band model achieved a precision of 0.86, recall of 0.88, F1 score of 0.87, and Intersection-over-Union (IoU) of 0.76, consistently outperforming four-band baselines by reducing false positives while maintaining sensitivity. Predictions on withheld Houston tiles confirmed strong within-region generalization, yielded a precision of 0.78, recall of 0.81, F1 score of 0.79, and IoU of 0.66. Qualitative analysis further revealed limitations stemming from both training label quality and vegetation–building confusion. These findings demonstrate the complementary value of integrating spectral and structural information for robust building footprint extraction and how domain adaptation strategies can be used to enhance cross-regional transferability.

1. INTRODUCTION

Building footprint extraction is essential for a wide range of geospatial applications, including population estimation, urban planning, infrastructure development, and disaster management (Che et al., 2024; Vincent and Varalakshmi, 2024). Accurate and scalable extraction of building footprints remains a challenging problem in urban remote sensing, particularly when applied across heterogeneous landscapes and recurring data acquisition cycles. High-resolution optical imagery captures detailed spatial and spectral information but is sensitive to shadows, occlusions, and vegetation cover (Liasis and Stavrou, 2016; Cao and Huang, 2021; Vincent and Varalakshmi, 2024; He et al., 2025; Ye et al., 2025), where lidar provides reliable elevation cues (Park and Guldmann, 2019; Kaplan et al., 2022; Rottmann et al., 2022; Karsli et al., 2024) but suffers from noise, variable point density, and incomplete coverage (Karsli et al., 2024). Existing approaches often rely on a single data modality or require extensive site-specific tuning, limiting their robustness and transferability across regions and time. There is therefore a need for a practical and repeatable building footprint extraction framework that effectively integrates complementary optical and lidar information while minimizing manual intervention and model reconfiguration.

Existing approaches to building footprint extraction fall into three main categories: manual interpretation or rule-based methods, machine learning (ML), and deep learning (DL). Manual interpretation can achieve high accuracy but is labor-intensive and impractical for large areas. Semi-automated rule-based methods and classical image processing techniques often fail in complex urban or densely vegetated environments, where variability in building size, shape, occlusion, and appearance limits their generalization (Vincent and Varalakshmi, 2024). ML techniques such as random forests and support vector machines (Schlosser et al., 2020; Kaplan et al., 2022) improve automation but rely heavily on hand-crafted features, making them less effective for capturing complex spatial patterns across diverse geographies (Karsli et al., 2024).

DL methods, particularly convolutional neural networks (CNNs), have transformed building footprint extraction by learning hierarchical spatial features directly from imagery (Luo et al., 2021). Among these, U-Net (Ronneberger et al., 2015) is widely adopted for semantic segmentation tasks. With its encoder–decoder structure and skip connections, U-Net preserves fine spatial details while capturing contextual information, making it effective for delineating buildings even in challenging urban scenes. Numerous studies have shown U-Net’s robustness across various remote sensing data sets, including high-resolution optical and multispectral imagery (Zhu et al., 2017; Bittner et al., 2018; Ji et al., 2019; Rastogi et al., 2020; Alsabhan and Alotaiby, 2022).

Although recent studies have introduced advanced architectures such as U-Net++ variants (Zuo et al., 2025) and transformer-based segmentation models (Wang et al., 2022; Gibril et al.,

* Corresponding author

2024) for building footprint extraction, most existing work emphasizes architectural innovation using optical imagery alone or small, region-specific data sets. Fewer studies systematically investigate how nationally available lidar products, such as U.S. Geological Survey (USGS) 3D Elevation Program (3DEP), can be operationally integrated with multispectral imagery to improve building extraction at scale. Moreover, there is limited emphasis on reproducible workflows tailored for repeated data updates and national mapping programs. This study addresses this gap by focusing on scalable multimodal data fusion using National Agriculture Imagery Program (NAIP) and 3DEP products within a consistent deep learning framework, with emphasis on operational applicability for large-area building footprint updating rather than solely on model architecture novelty.

Combining spectral and elevation information has been shown to improve segmentation accuracy and model generalization. For example, Yu et al. (2024) integrated DSM features to enhance building boundary definition, whereas Ji et al. (2018) demonstrated that lidar-derived heights help distinguish buildings from other impervious surfaces. A notable example is the AWS Open Data tutorial “Automatic building footprint extraction using satellite RGB and lidar elevation” (Amazon Web Services, 2021), which used RGB imagery with a lidar-derived digital surface model (DSM) to improve building segmentation in urban areas.

Building on this foundation, the present study develops a 6-band data stack that includes Red, Green, Blue, and Near-Infrared (IR) bands from the U.S. Department of Agriculture (USDA) National Agricultural Imagery Program (NAIP), along with lidar-derived feature height and intensity rasters from 3DEP. The IR band aids vegetation–building separation, height data isolates vertical structures, and intensity provides information on surface reflectivity differences between roofs, vegetation, and other materials. By aligning all layers to 1.0 m resolution, this multisource feature stack captures complementary spatial, spectral, and structural characteristics.

We propose a U-Net–based deep learning framework for building footprint extraction using this 6-band data set. NAIP provides high-resolution spectral information, whereas 3DEP lidar adds precise height and intensity features to enhance detection, particularly in vegetated or shadowed areas. The method is evaluated using Microsoft’s Global Building Footprints (GBF) (Che et al., 2024; Microsoft, 2025) as reference data with performance assessed using precision, recall, F1 score, and Intersection-over-Union (IoU). Results indicate that integrating NAIP and 3DEP data improves footprint extraction performance, with notable gains in challenging environments.

This study contributes a reproducible and scalable workflow for integrating NAIP multispectral imagery with lidar-derived structural features to improve building footprint extraction. A key strength of the framework is its adaptability: once established, the pipeline can be re-deployed with minimal model re-tuning when new NAIP or 3DEP collections become available. This characteristic makes the approach highly relevant for agencies such as the USGS, where repeat data acquisitions are frequent and efficient updating of building inventories is critical.

Specifically, this work addresses the following gaps in existing building footprint extraction studies:

- Lack of operationally scalable fusion workflows: Many prior studies demonstrate optical–lidar fusion at local or experimental scales but do not emphasize reproducibility or reusability with recurring national data sets.
- Underutilization of lidar intensity information: Although lidar-derived height or DSM products are commonly used, the complementary role of lidar intensity for distinguishing roofs from vegetation and other surfaces remains insufficiently explored.
- Limited benchmarking across data modalities: Few studies systematically compare optical-only, lidar-only, and fused inputs within a consistent deep learning framework to quantify their relative contributions.
- Insufficient discussion of label quality impacts: The influence of large, automatically generated reference data sets, such as GBF, on model training and evaluation is often overlooked.

By addressing these gaps, the proposed framework advances building footprint extraction toward repeatable, large-area deployment using nationally available multisource geospatial data.

2. DATA

2.1 NAIP imagery

The NAIP is a federally funded initiative managed by the USDA’s Farm Service Agency (USDA-FSA) that provides high-resolution aerial imagery across the continental United States. NAIP imagery is collected during the agricultural growing season and is primarily intended to support the agency’s agricultural monitoring and compliance programs. However, the high spatial resolution (typically 1-meter or finer) and four-band coverage - red, green, blue, and near-infrared (NIR) - make NAIP a substantial resource for a broad range of geospatial applications beyond agriculture, including urban planning, environmental monitoring, and land cover classification (USDA-FSA, 2023). NAIP imagery is acquired using digital sensors onboard aircraft and made publicly available shortly after collection. The consistent nationwide coverage, multi-year availability, and open access have made NAIP a widely adopted data set in remote sensing and GIS research (Subedi and Portillo-Quintero, 2025). NAIP imagery used in this study were downloaded from USGS EarthExplorer (https://earthexplorer.usgs.gov).

2.2 USGS 3DEP

The USGS 3DEP was launched to meet the increasing demand for high-quality, three-dimensional (3D) elevation data across the United States. It is a collaborative effort among federal, state, local, tribal, and private organizations to produce a national 3D elevation data set at a uniform, high-resolution quality (FGDC, 2023). Lidar is the primary data collection method for the lower 48 states, Hawaii, and U.S. territories. In Alaska, where conditions can be challenging for airborne lidar collection, airborne Interferometric Synthetic Aperture Radar (IfSAR) technology is used instead. The key features of the 3DEP include nationwide coverage, public accessibility, and regular updates. The program aims to provide high-resolution elevation data across the entire United States, and these datasets are freely available through the USGS for a wide range of geospatial applications.

Lidar point clouds are provided in the LAS or LAZ file format. The data have a nominal point density of 2 pts/square meter and

a vertical height accuracy RMSEz of 10cm, per the USGS Base Lidar Specifications (Heidemann, 2012). Each point contains detailed x, y, z coordinates as well as other attributes, including intensity, return number, and point classification. The 3DEP classification is optimized for the development of bare earth digital elevation data and not optimal for extracting above ground features (Gruen et al., 2019). To increase the accuracy of extracted build height information, a deep learning (DL) model (Liu et al., 2024) is applied to re-classify 3DEP lidar data from the test sites. The DL model is trained using the open annotated lidar training data from the Data Fusion Contest (DFC) 2019 data set (Bosch et al., 2019; Le Saux et al., 2019). The classes in the DFC 2019 training data include ground (bare earth), vegetation, building, water, and bridge deck.

2.3 Global Building Footprints (GBF)

The GBF data set is a large-scale, open-access geospatial resource that provides detailed building outlines derived from high-resolution satellite imagery using deep learning techniques. First released in 2018, the data set has since expanded to include more than 1.4 billion building footprints across over 160 countries as of 2024 (Microsoft, 2025). Footprints are generated through a two-stage pipeline involving semantic segmentation via CNNs to detect building pixels, followed by polygonization to convert segmented outputs into vector geometries (Li et al., 2022). The data set is distributed in GeoJSON format, supporting applications in urban planning, population mapping, disaster response, and AI/ML training.

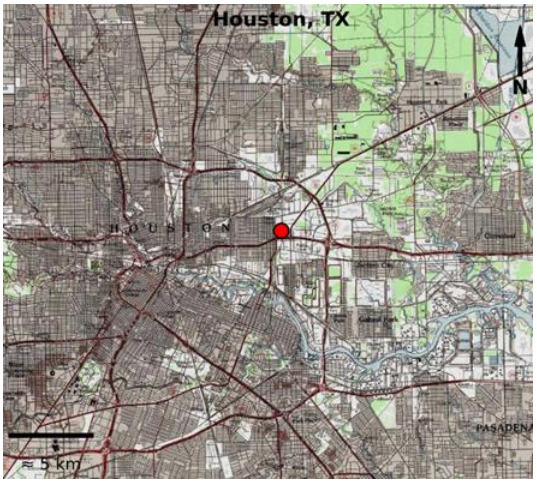
Recent advancements have extended the GBF data set into three-dimensional form, resulting in the release of 3D-GloBFP, the first global 3D building footprint data set (Che et al., 2024). In this version, building height estimates are included, derived through stereo image matching using overlapping satellite scenes. This enhancement enables volumetric urban modeling at global scale and addresses one of the key limitations of earlier GBF releases, which were strictly 2D. Although height accuracy varies by region and image quality, the 3D-GloBFP data set represents a substantial step forward in providing scalable, remotely sensed building height data, especially in areas lacking lidar or city-level elevation data sets (Che et al., 2024). In this study, GBF is used as labelling data to create a mask for training as well as validation.

3. METHODS

3.1 Study area

The Houston, Texas study site (29°49′30.0″ N, 95°19′30.0″ W; Figure 1) encompasses a dense and diverse urban landscape characterized by mixed residential, commercial, and industrial zones. The site has an area of 671.32 square kilometers. The selected study site lies within the greater Houston metropolitan area, one of the largest and most dynamic cities in the United States, notable for its complex urban morphology, extensive transportation infrastructure, and varied building types. This region presents both challenges and opportunities for building footprint extraction due to high-rise structures in the downtown core, low-rise suburban developments, and widespread tree cover that can partially obscure rooftops in aerial imagery. The test site includes numerous impervious surfaces, waterways, and vegetated zones, providing a representative mix of spectral and structural conditions for evaluating the performance of multi-source remote sensing–based deep learning models. Its geographic extent and building density make it a relevant and demanding case study for validating footprint extraction approaches that integrate NAIP aerial imagery and lidar-derived elevation products.

(a)



(b)

Figure 1. Location illustration of study area in Houston, TX. (a) OpenStreetMap Contributors (2025). (b) Basemap courtesy of U.S. Geological Survey Topographic maps.

3.2 3DEP Lidar Point Cloud Classification

Point cloud classification methods are used to assign meaningful labels to each point in a point cloud data set. In recent years, DL approaches tailored for point clouds have gained traction. A recent systematic survey and outlook on DL-based point cloud classification (Diab et al., 2022; Farshian et al., 2023; Zhang et al., 2023) indicate that Point Transformer (Zhao et al., 2021; Wu et al., 2022) has one of the best-performing approaches. Based on the statistics reported in the literature, this method was selected to classify 3DEP lidar data for this study (Liu et al., 2024).

The Transformer has been widely used in natural language processing (Diab et al., 2019) and computer vision (Zhao et al., 2021) because of its powerful self-attention mechanism (Vaswani et al., 2017). It is beneficial for unstructured point clouds because it does not require a regular data structure and is only based on pointwise operations (Zhao et al., 2021). Point Transformer adopts aggregated transformer blocks that define self-attention across point features through linear operations. It introduces a trainable, parameterized position encoding that can be trained end-to-end. This differs from existing works that apply global attention to the whole point cloud set (Xie et al., 2018; Lee et al., 2019). Conversely, Point Transformer applies self-attention locally, enabling scalability to large scenes with millions of points (Zhao et al., 2021). According to Liu et al. (2024), the Point Transformer achieved an overall classification accuracy of 96.7% from three validation sites.

3.3 U-Net

U-Net is a convolutional neural network (CNN) architecture originally developed for biomedical image segmentation

(Ronneberger et al., 2015) but has since been widely adopted in remote sensing for tasks such as building footprint extraction, land cover mapping, and road detection. Its popularity stems from a symmetric encoder–decoder design connected by skip connections, which allow the model to capture both high-level semantic context and fine-grained spatial detail. The encoder progressively downsamples the input image to learn abstract, high-level features, whereas the decoder upsamples these representations to recover spatial resolution. Skip connections link corresponding encoder and decoder layers, ensuring that localization accuracy is preserved by reintroducing low-level spatial information lost during downsampling.

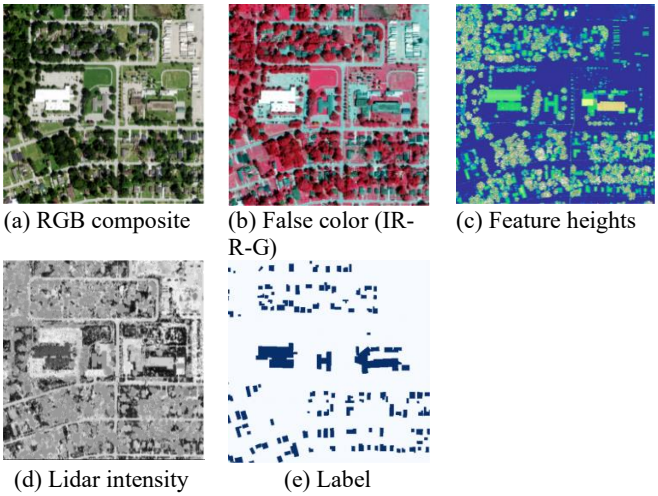
Moreover, U-Net is well-suited for applications where labeled training data is limited, thanks to its efficient architecture and compatibility with data augmentation strategies. Variants such as U-Net++ and attention U-Net have been proposed to further enhance boundary refinement and focus on relevant spatial features, but the original U-Net remains a strong baseline for semantic segmentation in remote sensing (Luo et al., 2021; Alsabhan and Alotaiby, 2022). In building extraction, the combination of U-Net’s localization precision and its ability to integrate multi-source data makes it particularly effective for generating accurate and topologically consistent building footprints, which are critical for urban planning, disaster assessment, and geospatial analysis.

3.4 Data preparation pipeline

To prepare the input data for building footprint extraction, we pre-processed both aerial imagery and lidar-derived features to ensure consistency in spatial resolution and coordinate reference systems. The National Agricultural Imagery Program (NAIP) imagery, consisting of 98 tiles covering the study area, was bilinearly resampled from its native 0.6 m spatial resolution to 1.0 m resolution. This step was performed to align the imagery with the lidar-derived data and to reduce data volume and computational requirements (Ghanea et al., 2016; Zhu et al., 2017). The lidar-derived data sets included feature height and intensity, each provided as 289 tiles. These rasters were mosaicked, reprojected into the Universal Transverse Mercator (UTM) projection (EPSG:26915), and spatially aligned with the NAIP imagery. This ensured that both spectral and elevation features could be directly integrated for model training (Park and Guldmann, 2019; Rottmann et al., 2022).

Following alignment, a six-band image stack was created for each tile. The stack combined the four spectral bands from NAIP (red, green, blue, and near-infrared) with two lidar-derived layers: feature height and intensity. This multi-source integration provided both spectral and structural information, enhancing the ability to differentiate buildings from spectrally similar features such as roads or vegetation (Gilani et al., 2015; Bramhe et al., 2018). As shown in Figure 2, the stacked images were divided into fixed-size tiles of 512 × 512 pixels to facilitate efficient training and evaluation, resulting in a total of 6,786 tiles across the study area. The study area contains 6,217 building footprint polygons. For each image tile, the corresponding building footprint polygons were spatially clipped and rasterized to generate binary mask tiles**;** 625 tiles contained no buildings, representing background-only samples in the dataset. These masks provided the ground truth for supervised model training, consistent with common practices in remote sensing deep learning workflows (Adam et al., 2023).

Finally, all input features were normalized to a range between 0 and 1 using 1st–99th percentile clipping. This normalization reduced the influence of extreme outliers and preserving the majority of data variation across spectral and lidar-derived bands (EO Research, 2020). The resulting data set consisted of paired .npz files, each containing one six-band image tile and its corresponding building footprint mask. This standardized data set served as the foundation for model training, validation, and evaluation.



(a) RGB composite (b) False color (IR-R-G) (c) Feature heights

(d) Lidar intensity (e) Label

Figure 2. An example of a stacked 6 band training tile. National Agriculture Imagery Program imagery accessed via the U.S. Geological Survey EarthExplorer portal.

3.5 Evaluation metrics

To quantitatively assess the performance of the building footprint extraction model, several commonly used metrics in semantic segmentation were employed. Precision (Equation 1), Recall (Equation 2), Intersection-over-Union (IoU, Equation 3), and F1 (Equation 4) score were calculated on a per-class basis, whereas overall Accuracy (Equation 5) was computed across all classes. These metrics provide complementary perspectives on model performance by balancing correctness of predictions (precision), completeness of detection (recall), and overall agreement with ground truth (IoU, F1, accuracy). Precision measures the proportion of predicted building pixels that are correctly classified, whereas Recall evaluates the proportion of actual building pixels that are successfully detected. IoU quantifies the overlap between the predicted and reference building masks, and the F1 score represents the harmonic mean of Precision and Recall, balancing their trade-off. Accuracy, in contrast, accounts for both building and non-building pixels across the entire image domain, providing an overall measure of classification correctness. The metrics are defined as follows:

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

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

$$IoU = \frac{TP}{TP + FP + FN} \quad (3)$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

where TP refers to true positives, FP refers to false positives, TN refers to true negatives, and FN refers to false negatives.

4. RESULTS AND DISCUSSIONS

The building footprint extraction model was trained using six-band image patches with dimensions of (6, 512, 512), where the input channels included four NAIP spectral bands (Red, Green, Blue, Near-Infrared) and two lidar-derived features (feature height and intensity). To optimize segmentation performance, the Binary Cross Entropy (BCE) loss function was employed. This formulation focuses on pixel-wise classification accuracy, ensuring that building and non-building pixels are effectively distinguished. Model optimization was performed using the Adam optimizer with a learning rate of 0.0001, which has been shown to provide stable convergence in deep learning segmentation tasks. Training was conducted on an NVIDIA CUDA-enabled GPU, enabling efficient processing of the large data set and reducing training time. The model was trained for 50 epochs, with an 80/20 split between training and validation samples to ensure robust evaluation and to prevent overfitting.

4.1 Stacked 4 bands versus 6 bands

To know if the proposed 6-band stack outperforms those of 4-band stack. This study first compared building footprint extraction performance using two different image stacks: (1) a 4-band stack comprising red, green, blue (RGB) bands from NAIP imagery and elevation from 3DEP lidar; and (2) a 6-band stack comprising RGB and infrared (IR) bands from NAIP imagery, along with feature height and intensity derived from classified 3DEP lidar. Table 1 shows that the 6-band stack consistently outperformed the 4-band stack across all metrics. Precision improved from 0.84 to 0.88, recall from 0.82 to 0.86, F1 score from 0.83 to 0.87, and IoU from 0.72 to 0.77 in training data set. The validation data set in Table 1 further reinforces this trend as the 6-band model achieved higher precision (0.86 vs. 0.82) and a modest increase in F1 score (0.87 vs. 0.85) over the 4-band model, whereas both models had identical recall (0.88). The increased precision with similar recall indicates that the 6-band model was more effective in suppressing false positives without sacrificing detection sensitivity. These gains also suggest that the additional bands, particularly the feature height and intensity, contributed to better delineation and discrimination of building structures during training.

Evaluation metrics	Training		Validation	
	4 bands	6 bands	4 bands	6 bands
Precision	0.84	0.88	0.82	0.86
Recall	0.82	0.86	0.88	0.88
F1 score	0.83	0.87	0.85	0.87
IoU	0.72	0.77	0.74	0.76

Table 1: Performance metrics for the training and validation data set using the 4-band (RGB + elevation) and 6-band (RGB, IR, feature height, and intensity) input stacks.

Figure 3 furthermore supports our quantitative findings. In the prediction results, the 4-band model occasionally misclassified non-building man-made features such as vehicles as buildings (as shown in the blue circled area in Figure 3. (c)). These misclassifications were notably reduced in the 6-band results. The inclusion of lidar-derived feature height in the 6-band stack helped distinguish true building footprints from lower-lying structures, thereby reducing commission errors and improving spatial accuracy. Overall, the results suggest that enriching the input feature space with both spectral (IR) and structural (height and intensity) information substantially enhances the performance of deep learning models in building footprint extraction tasks. This is particularly valuable when aiming to minimize false positives without sacrificing recall, a common challenge in dense urban mapping.

(a) Natural color image (b) Label

(c) Prediction on a 4-band stack (d) Prediction on a 6-band stack

Figure 3. An example of building footprint extraction results on the data. Columns (from left to right) show the RGB image (a), ground truth building mask (b), and prediction overlay on the RGB image for 4-band (c) and 6-band (d) respectively. National Agriculture Imagery Program imagery accessed via the U.S. Geological Survey EarthExplorer portal.

4.2 Prediction on unseen data

To further validate the robustness of the proposed model, we applied the trained network to 15 randomly selected tiles within the study area that were intentionally excluded from both training and validation, which offers the most meaningful evaluation of model generalization. This evaluation provides an unbiased test of generalization performance within the same geographic region. As shown in Table 2, the model achieved high predictive accuracy, with a precision of 0.78, recall of 0.81, F1 score of 0.79, and IoU of 0.66 in Houston, TX. These results are consistent with the validation outcomes, suggesting that the model is not overfitted to the training data and can reliably detect building footprints on unseen tiles from the same region with similar urban environments.

Evaluation metrics	Precision	Recall	F1 score	IoU
	0.78	0.81	0.79	0.66

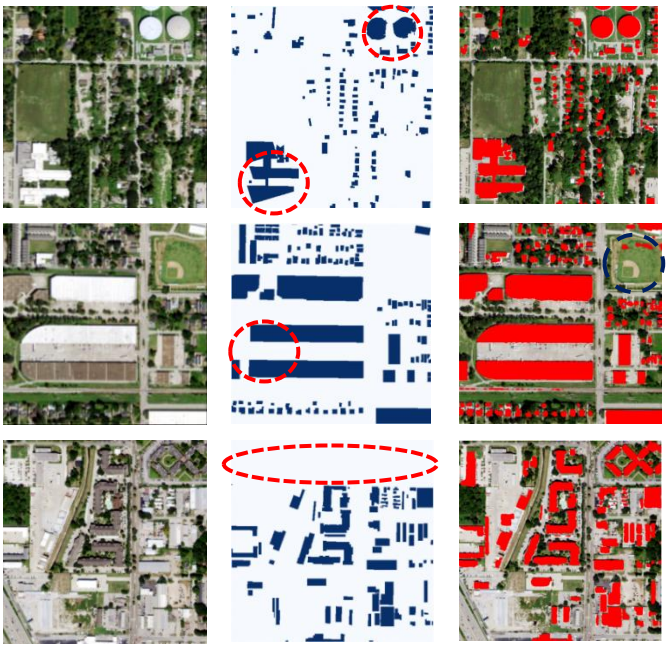
Table 2: Performance metrics for prediction/test data set using the 6-band (RGB, IR, feature height, and intensity) input stacks.

Visual examples of these predictions are shown in Figure 4, which illustrates building extraction results on the Houston test tiles. The comparison between ground truth masks and predicted footprints demonstrates that most building shapes and boundaries were accurately captured. Large structures with simple roof geometries were consistently delineated, and the model showed resilience to local variations in reflectance and elevation. Minor errors were primarily associated with smaller buildings or structures adjacent to tree canopy, where lidar height and intensity values were less distinct. These observations align with

https://doi.org/10.5194/isprs-annals-XI-4-2026-103-2026

the high F1 score and IoU values reported in Table 2, reinforcing the reliability of the approach within the training domain.

Closer inspection of the Houston results in Figure 4 also reveals two important sources of error. In the middle column, some mismatches originate from the ground truth data itself (red circled areas in Figure 4.(b)), where the GBF-derived masks include footprint inaccuracies that propagate into evaluation metrics. These examples highlight that performance assessment can be partially constrained by the quality of available reference data sets. In the right column, the model occasionally misclassifies tall tree crowns as buildings (blue circled area in the first raw of Figure 4.(c)), reflecting the spectral and structural similarities between vegetation and built surfaces in the six-band imagery. Although these errors are relatively limited, they underscore the importance of combining high-quality training labels with enhanced strategies for vegetation–building separation in future work.



(a) Image (b) Ground truth (c) Prediction

Figure 4. Prediction results for unseen test tiles in Houston, TX. Left: six-band NAIP and lidar-derived inputs; Middle: ground truth building masks; Right: predicted building footprints.

5. CONCLUSIONS AND FUTURE WORKS

This study presents a scalable deep learning framework for building footprint extraction that integrates multispectral NAIP imagery with lidar-derived feature height and intensity from the USGS 3DEP program. Methodologically, the primary contribution lies in the development of a unified six-band data stack and training pipeline that jointly exploits spectral, structural, and reflectance intensity information and remains compatible with routinely updated national data sets. The proposed workflow is designed to support repeatable deployment with minimal reconfiguration as new imagery and lidar acquisitions become available. Compared with global building mapping approaches that rely primarily on optical or stereo imagery such as GBF in this study, the inclusion of lidar-derived structural features provides more robust geometric information and reduces sensitivity to shadowing, roof texture variability, and vegetation occlusion.

Empirically, the results demonstrate that incorporating lidar-derived height and intensity features consistently improves building footprint delineation compared to optical-only or elevation-only inputs. Evaluations across training, validation, and independent Houston test tiles show gains in recall and overall segmentation performance, particularly in complex urban and vegetated environments. These findings confirm the added value of lidar-enhanced inputs for reducing omission errors and highlight the sensitivity of model performance to label quality and geographic variability.

Additional work focuses on four main directions. First, improving reference data quality is essential because reliance on GBF introduces labeling biases that propagate into both training and evaluation. This effort includes incorporating higher-quality reference data sets, such as newly released global building products, and targeted manual validation to better quantify and mitigate reference errors. Second, enhanced vegetation–building separation strategies, including spatial attention mechanisms and transformer-based architectures, can be explored to reduce tree-related false positives. Third, domain adaptation and fine-tuning approaches can be investigated to improve cross-city generalization and enable robust transfer across regions with differing urban morphologies. Finally, additional baseline architectures, including attention-enhanced U-Net variants and transformer-based segmentation models, can be incorporated to further contextualize the performance of the proposed framework. Because the framework emphasizes the integration of spectral and structural information rather than dependence on a single model architecture, these extensions can strengthen building footprint extraction performance and support the development of scalable, nationwide systems for urban analytics, hazard assessment, and infrastructure management.

DISCLAIMER

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

Adam, J.M., Liu, W., Zang, Y., et al., 2023. Deep learning-based semantic segmentation of urban-scale 3D meshes in remote sensing: A survey. *International Journal of Applied Earth Observation and Geoinformation*, 121, 103365, https://doi.org/10.1016/j.jag.2023.103365.

Amazon Web Services, 2021. Automatic building footprint extraction using satellite RGB and LiDAR elevation. AWS Open Data Tutorials. https://github.com/aws-samples/aws-open-data-satellite-lidar-tutorial (8 September 2025).

Alsabhan, W., Alotaiby, T., 2022. Automatic building extraction on satellite images using Unet and ResNet50. *Computational Intelligence and Neuroscience*, https://doi.org/10.1155/2022/5008854.

Bittner, K., Adam, F., Cui, S., et al., 2018. Building footprint extraction from VHR remote sensing images combined with normalized DSM using fused fully convolutional networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 2472–2485, https://doi.org/10.1109/JSTARS.2018.2849363.

Bosch, M., Foster, G., Christie, G., Wang, S., Hager, G. D., Brown. M., 2019. Semantic stereo for incidental satellite images.

Proc. of Winter Conf. on Applications of Computer Vision, https://doi.org/10.48550/arXiv.1811.08739.

Bramhe, V. S., Ghosh, S. K., Garg, P. K., 2018. Extraction of built-up area by combining textural features and spectral indices from Landsat-8 multispectral image. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLII-5, 727–733, https://doi.org/10.5194/isprs-archives-XLII-5-727-2018.

Cao, Y., Huang, X., 2021. A deep learning method for building height estimation using high-resolution multi-view imagery over urban area: a case study of 42 Chinese cities. *Remote Sensing of Environment*, 264, 112590, https://doi.org/10.1016/j.rse.2021.112590.

Che, Y., Li, X., Liu, X., Wang, Y., Liao, W., et al., 2024. 3D-GloBFP: the first global three-dimensional building footprint dataset. *Earth System Science Data*, 16, 5357–5374, https://doi.org/10.5194/essd-16-5357-2024.

Diab, A., Kashef, R., Shaker, A., 2022. Deep learning for Lidar point cloud classification in remote sensing. *Sensors*, 22, 7868. https://doi.org/10.3390/s22207868.

EO Research, 2020. How to normalize satellite images for deep learning. https://medium.com/sentinel-hub/how-to-normalize-satellite-images-for-deep-learning-d5b668c885af (10 September 2025).

Farshian, A., Gotz, M., Cavallaro, G., Debus, C., et al., 2023. Deep-learning-based 3-D surface reconstruction - A survey. *Proceedings of the IEEE*, 111(11), pp. 1464-1501, https://doi.org/10.1109/JPROC.2023.3321433.

FGDC (Federal Geographic Data Committee), 2023. Assessment of the 3D elevation program. https://www.fgdc.gov/ngac/meetings/june-2023/ngac-assessment-of-the-3d-elevation-program-june.pdf (8 September 2025).

Ghanea, M., Moallem, P., Momeni, M., 2016. Building extraction from high-resolution satellite images in urban areas: recent methods and strategies against significant challenges. International Journal of Remote Sensing, 37, 21, pp. 5234–5248, https://doi.org/10.1080/01431161.2016.1230287.

Gibril, M. B. A., Al-Ruzouq, R., Shanableh, A., Jena, R., et al., 2024. Transformer-based semantic segmentation for large-scale building footprint extraction from very-high resolution satellite images. *Advances in Space Research*, 73, 10, 4937-4954, https://doi.org/10.1016/j.asr.2024.03.002.

Gilani, S. A. N., Awrangjeb, M., Lu, G., 2015. Fusion of lidar data and multispectral imagery for effective building detection based on graph and connected component analysis. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XL-3/W2, 65–72, https://doi.org/10.5194/isprsarchives-XL-3-W2-65-2015..

Gruen, A., Schubiger, S., Qin, R., Schrotter, G., et al., 2019. Semantically enriched high-resolution LOD 3 building model generation. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLII-4/W15, https://doi.org/ 10.5194/isprs-archives-XLII-4-W15-11-2019.

He, H., Xu, L., Chapman, M. A., et al., 2025. Cost-effective high-definition building mapping: box-supervised rooftop delineation using high- resolution remote sensing imagery. *Photogrammetric Engineering & Remote Sensing*, 91, 4, pp. 225-239, https://doi.org/10.14358/PERS.24-00115R3.

Heidemann, H.K., 2012. Lidar base specification (No. 11-B4). U.S. Geological Survey. https://www.usgs.gov/ngp-standards-and-specifications/lidar-base-specification-online (11 September 2025).

Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE TGRS*, 57, 574-586, https://doi.org/10.1109/TGRS.2018.2858817.

Kaplan, G., Comert, R., Kaplan, O., et al., 2022. Using machine learning to extract building inventory information based on lidar data. *ISPRS Int. J. Geo-Inf, 11*, 517, https://doi.org/10.3390/ijgi11100517.

Karsli, B., Yilmazturk, F., Bahadir, M., Karsli, F., Ozdemir, E., 2024. Automatic building footprint extraction from photogrammetric and lidar point clouds using a novel improved-Octree approach. *Journal of Building Engineering*, 82, https://doi.org/10.1016/j.jobe.2023.108281.

Le Saux, B., Yokoya, N., Hänsch, R., et al., 2019. 2019 Data Fusion Contest. *IEEE Geoscience and Remote Sensing Magazine*, 7(1), https://doi.org/10.1109/MGRS.2019.2893783.

Lee, J., Lee, Y., Kim, J., et al. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. *Proceedings of the 36th International Conference on Machine Learning,* https://doi.org/10.48550/arXiv.1810.00825.

Li, J., Huang, X., Tu, L., Zhang, T., Wang, L., 2022. A review of building detection from very high resolution optical remote sensing images. *GIScience & Remote Sensing*, 59(1), 1199–1225, https://doi.org/10.1080/15481603.2022.2101727.

Liasis, G., Stavrou, S., 2016. Satellite image analysis for shadow detection and building height estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119, 437–450, https://doi.org/10.1016/j.isprsjprs.2016.07.006.

Liu, J. L., Qin, R., Song, S., 2024. Automated deep learning-based point cloud classification on USGS 3DEP lidar data using a transformer. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Athens, Greece, doi.org/10.1109/IGARSS53475.2024.10641055.

Luo, L., Li, P., Yan, X., 2021. Deep learning-based building extraction from remote sensing images: A comprehensive review. *Energies*, 14(23):7982, doi.org/10.3390/en14237982.

Microsoft, 2025. Global Building Footprints. Retrieved from https://github.com/microsoft/GlobalMLBuildingFootprints (29 July 2025).

OpenStreetMap Contributors, 2025. About OpenStreetMap. Retrieved from https://www.openstreetmap.org/about (31 July 2025).

Park, Y., Guldmann, J.-M., 2019. Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Computers, Environment and Urban Systems*, 75, 76–89, https://doi.org/10.1016/j.compenvurbsys.2019.01.004.

Rastogi, K., Bodani, P., Sharma, S. A., 2020. Automatic building footprint extraction from very high-resolution imagery using deep learning techniques. *Geocarto International*, 37, 5, 1501–1513. https://doi.org/10.1080/10106049.2020.1778100.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention,* 234-241, https://doi.org/10.1007/978-3-319-24574-4_28.

Rottmann, P., Haunert, J.-H., Dehbi, Y., 2022. Automatic building footprint extraction from 3D laserscans. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, X-4/W2-2022, 233–240, https://doi.org/10.5194/isprs-annals-X-4-W2-2022-233-2022.

Schlosser, A. D., Szabó, G., Bertalan, L., Varga, Z., Enyedi, P., Szabó, S., 2020. Building extraction using orthophotos and dense point cloud derived from visual band aerial imagery based on machine learning and segmentation. *Remote Sensing,* 12(15):2397, https://doi.org/10.3390/rs12152397.

Subedi, M., Portillo-Quintero, C., 2025. Quantifying change in urban tree cover in the city of Lubbock, Texas, using LiDAR and NAIP imagery fusion. *Science of Remote Sensing*, 11, https://doi.org/10.1016/j.srs.2025.100240.

USDA-FSA, 2023. National Agriculture Imagery Program (NAIP). U.S. Department of Agriculture, Farm Service Agency. https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/ (8 September 2025).

Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS)*, https://doi.org/10.48550/arXiv.1706.03762.

Vincent, J. M., Varalakshmi, P., 2024. Extraction of building footprint using MASK-RCNN for high resolution aerial imagery. *Environmental Research Communications*, 6, 075015, https://doi.org/10.1088/2515-7620/ad5b3d.

Wang, L., Fang, S., Meng, X., Li, R., 2022. Building extraction with vision transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-11, https://doi: 10.1109/TGRS.2022.318663.

Wu, X., Liu, X., Wang, J., et al. 2022. Point cloud classification based on transformer. *Computers and Electrical Engineering* 104(A), https://doi.org/10.1016/j.compeleceng.2022.108413.

Xie, S., Liu, S., Chen, Z., Tu, Z. 2018. Attentional shape context net for point cloud recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* https://doi.org/10.1109/CVPR.2018.00484.

Ye, X., Bai, W., Huang, X., 2025. Enhancing residential building identification in a coastal Texas city: an integrated framework leveraging remote sensing, GIS, and transfer learning techniques. *Photogrammetric Engineering & Remote Senisng*, 91, 7, 455-462, https://doi.org/10.14358/PERS.24-00048R3.

Yu, D., Ji, S., Wei, S., Khoshelham, K. 2024. 3-D building instance extraction from high-resolution remote sensing images and DSM with an end-to-end deep neural network. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-19, https://doi: 10.1109/TGRS.2024.3383432.

Zhang, H., Wang, C., Tian, S., et al. 2023. Deep learning-based 3D point cloud classification: A systematic survey and outlook. *Displays* 79, https://doi.org/10.48550/arXiv.2311.02608.

Zhao, H., Jiang, L., Jia, J., et al., 2021. Point Transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259-16268, https://doi.org/10.1109/ICCV48922.2021.01595.

Zhu, X. X., Tuia, D., Mou, L., et al., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5, 4, 8-36, https://doi.org/10.1109/MGRS.2017.2762307.

Zuo, X., Shao, Z., Wang, J., Huang, X., Wang, Y., 2025. A cross-stage features fusion network for building extraction from remote sensing images. *Geo-Spatial Information Science*, 28, 2, 387–401, https://doi.org/10.1080/10095020.2024.2307922.