

An integrated deep learning approach for 3D urban scene generation using monocular satellite imagery

Ömer Can Karadağ^{*1}, Nihal Altuntaş²

¹Department of Software Engineering, Faculty of Engineering and Architecture, İstanbul Gelişim University, İstanbul, Türkiye

²Department of Computer Engineering, Faculty of Engineering, İstanbul University-Cerrahpaşa, İstanbul, Türkiye

Cite this article as: Karadağ, Ö. C., & Altuntaş, N. (2026). An integrated deep learning approach for 3D urban scene generation using monocular satellite imagery. *J Comp Electr Electron Eng Sci*, 4(2), 56-63. <https://doi.org/10.51271/JCEES-0037>

Received: 09.12.2025

Accepted: 27.05.2026

Published: 13.10.2026

ABSTRACT

Aims: The aim of this study is to build a pipeline leveraging deep learning for 3D reconstruction of urban environments using only single high-resolution satellite images, while comparing strengths and weaknesses of the proposed system with its alternatives like multi-view or LiDAR-based methods.

Methods: We proposed a pipeline system consisting of several steps with different deep learning models. Each step works independently to complete different tasks. Firstly, the system generates digital surface model (DSM) using a Pix2Pix conditional GAN with perceptual loss, then a mask R-CNN model is used to build segmentation and finally a ResNet-50 model classifies roof type of buildings. Models were trained on SpaceNet Urban 3D, Inria aerial labeling, and GATE roof datasets, and MSE, IoU/mAP and accuracy/F1 scores are preferred as evaluation metrics for elevation accuracy, segmentation quality and classification performance, respectively.

Results: DSM generation achieved MSE of 986.70 (mean absolute deviation of 31 gray levels), with 90% of pixels maintaining errors below 50 levels. Instance segmentation attained 0.712 mAP@IoU50, declining to 0.358 at stricter IoU 75 thresholds, revealing boundary precision challenges. Roof classification reached 80% accuracy with strong performance on common types (hip roofs: F1=0.91). The system synthesizes these outputs into exportable 3D scenes combining terrain meshes, extruded building geometries, and vegetation elements.

Conclusion: This work validates that monocular satellite imagery can be sufficiently used as a sole input for a system using different deep learning models to generate functionally useful 3D urban models. The proposed system offers a practical solution for rapid urban analysis when additional data like multi-view image or laser scan is unavailable. The coordinated integration of multiple deep learning components yields promising results for urban reconstruction using conventional satellite imagery while geometric precision remains below production-grade standards.

Keywords: Digital surface model, Pix2Pix, mask R-CNN, remote sensing

INTRODUCTION

The demand for satellite imagery continues to increase each year. The information extracted from these images supports a wide range of applications including urban planning, disaster management, telecommunication infrastructure, solar power potential assessment and military strategy development. Recent advances in deep learning and computer vision have significantly expanded the types and quality of information that can be derived from static imagery. As a result, both the use of satellite data and research on deep learning methods for satellite image analysis have grown substantially.

However, the two-dimensional nature of these images imposes significant limitations. The ability to generate accurate 3D models from existing 2D data can unlock new possibilities for analysis and visualization. Until recent years, this domain was dominated by traditional methods such as LiDAR (light detection and ranging), photogrammetry and structure-from-motion (SfM). These methods are renowned for producing

highly accurate and detailed 3D models. LiDAR actively measures distances with laser pulses, directly generating dense point clouds, while photogrammetry and SfM derive 3D structure by identifying and triangulating corresponding points across multiple overlapping images. [Table 1](#) provides a comparison of traditional 3D reconstruction methods to contextualize their strengths and limitations.

Despite their proven accuracy, these conventional approaches present significant practical limitations. LiDAR surveys, whether airborne or terrestrial, require specialized, costly hardware and are often logistically complex to deploy. Similarly, multi-view photogrammetry and SfM are heavily dependent on the availability of high-quality, overlapping imagery from multiple satellite passes or aerial flight paths. This dependency not only increases data acquisition costs but also complicates and delays the reconstruction process,

*Corresponding Author: Ömer Can Karadağ, krdg.omercan@hotmail.com



Table 1. Comparison of traditional 3D reconstruction methods

Feature	Photogrammetry	LiDAR	SfM
Principle	Image analysis	Laser scanning	Multi-view geometry
Data source	RGB images	Point cloud	RGB images
Accuracy	Medium-high	Very high	Medium-high
Cost	Low-medium	High	Low-medium
Processing time	Slow-medium	Fast	Slow
Texture/color information	Rich	Limited	Rich
Key limitations	Weather conditions, textureless areas, shadows	High cost, limited texture	Computational speed, moving objects, image quality
Key applications	Visualization, mapping	High-precision modeling	Research, volume estimation, change detection

3D: Three-dimensional, LiDAR: Light detection and ranging, SfM: Structure-from-motion, RGB: Red, green, blue

making it infeasible for rapid, large-scale analyses in areas where such multi-view data is unavailable.

In response, deep learning has emerged as a powerful tool for automating complex feature extraction from satellite imagery. Significant research has demonstrated its capability to perform specific tasks such as estimating digital surface models (DSMs) (Chen et al., 2023; Liu et al., 2020), extracting instance-level building footprints (Wei et al., 2023), and classifying architectural attributes like roof types (Buyukdemircioglu et al., 2021) directly from single optical images. However, these advancements have largely progressed in isolation, solving discrete problems without converging on a unified solution for holistic 3D scene generation.

This study proposes and evaluates a novel pipeline that leverages deep learning to reconstruct 3D urban environments from a single monocular satellite image, a challenging, yet cost-effective alternative to traditional methods like stereo photogrammetry or LiDAR. The primary contribution of this work is the design and evaluation of a modular deep learning pipeline for monocular 3D urban reconstruction rather than achieving meter-level elevation accuracy. While the pipeline integrates several deep learning models to extract and assemble the necessary components for a complete 3D scene, the central goal of this work is to critically evaluate the feasibility of such fully monocular approaches. We systematically analyze the performance and crucially, the limitations and weaknesses inherent in each component of the pipeline. By providing detailed assessment of what is achievable with current methods and where the principal bottlenecks lie, this study aims to establish a clear benchmark and pave the way for more robust and accurate monocular reconstruction systems in the future.

METHODS

Ethics

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data. All procedures were carried out in accordance with the ethical rules and principles.

Study Design

We propose an integrated, monocular pipeline for 3D urban reconstruction from a single satellite image. The core challenge of this task lies in inferring a complete 3D scene encompassing terrain, buildings, and vegetation from a single 2D RGB input, an inherently ill-posed problem. To address this, our methodology decomposes the problem into a sequence of more tractable sub-tasks, each handled by a specialized deep learning model. This modular approach allows for targeted optimization of individual components such as DSM generation, building segmentation, and roof type classification before their coordinated integration. The overall pipeline is designed to sequentially extract and assemble the necessary geometric and semantic components, which are subsequently fused to construct a coherent 3D model. The following subsections detail the architecture, training, and implementation of each component.

To infer elevation from a single RGB image, we employed a conditional Generative Adversarial Network (cGAN) based on the Pix2Pix architecture (Isola et al., 2017), which is designed for image-to-image translation tasks. The model learns a mapping from an input RGB satellite image to a corresponding DSM.

The Pix2Pix framework consists of two core components: a generator and a discriminator, which are trained adversarially.

The generator is based on a U-Net architecture (Ronneberger et al., 2015). Its symmetric encoder-decoder structure compresses the input image into a latent representation via a series of convolutional layers (with batch normalization and Leaky ReLU/ReLU activations) and then reconstructs it to the target output size through transposed convolutions. A critical feature of U-Net is its use of skip connections, which concatenate feature maps from the encoder to the corresponding decoder layers. This helps preserve low-level details and enables the generation of sharper, more detailed output DSMs. To mitigate overfitting, dropout layers were incorporated in some of the upsampling blocks. The final layer uses a Tanh activation function to produce a normalized, single channel DSM output.

The discriminator utilizes a PatchGAN architecture. Instead of classifying the entire image as real or generated, it operates on local image patches, classifying each $N \times N$ patch independently. This forces the generator to produce realistic local textures and high-frequency details. The discriminator takes a concatenated channel of the input RGB image and either the real or generated DSM. Its architecture comprises convolutional layers with spectral normalization, instance normalization, and Leaky ReLU activations, progressively reducing the spatial dimension to produce a matrix of real/fake predictions for each patch.

The model was trained with a composite loss function. The adversarial loss, provided by the discriminator, uses Binary Cross-Entropy with Logits to distinguish real from generated DSMs. The total discriminator loss L_D is the average of the losses from real and generated pairs (Eq. 1)

$$L_D = 0.5 \times (L_{BCE}(D(x, y), 1) + L_{BCE}(D(x, G(x)), 0)) \quad (1)$$

The generator's loss L_G is a weighted combination of the adversarial loss L_{GAN} (Eq. 2) and a custom designed Elevation Loss L_{Elev} (Eq. 3).

$$L_{GAN} = L_{BCE}(D(x, G(x)), 1) \quad (2)$$

$$L_G = L_{GAN} + \lambda \times L_{Elev} \quad (3)$$

Elevation Loss is a key contribution to the training process. It is a composite function that integrates:

- A standard L1 loss for pixel-wise accuracy
- A weighted mechanism that assigns greater importance to a predefined critical elevation range (144-200 for our model), focusing the model's learning on structurally significant building heights.
- A VGG19-based perceptual loss to enhance structural and semantic similarity between the predicted and ground-truth DSMs.

The Adam optimizer was used with a learning rate of 1e-3 and betas of (0.5, 0.999). During training, input and target images were resized to 512x512 pixels. Data augmentation, including color jitter, was applied to the input RGB images. All images were normalized, and the data was processed in batches.

The generated DSM represents the elevation of the uppermost surfaces. To isolate the bare-earth elevation, a digital terrain model (DTM) was subsequently derived from the DSM using a cloth simulation filter (Zhang et al., 2016).

For the task of extracting building footprints, we employed mask R-CNN (He et al., 2020), a leading deep learning model for instance segmentation. This framework was selected for its ability to not only classify pixels as buildings but also to distinguish between individual building instances, which is critical for modeling separate structures in dense urban areas.

The model architecture utilized a ResNet-50 backbone fortified with a feature pyramid network (FPN). FPN enhances the model's capability to detect and segment objects at various scales by leveraging multi-scale feature maps, a crucial feature for accurately identifying buildings of different sizes in satellite imagery. The model was initialized with weights pre-trained on the COCO dataset to leverage its robust feature extraction capabilities. We then fine-tuned the entire network on our curated building dataset. Both the box predictor head (responsible for object classification) and the mask predictor head (responsible for generating pixel-wise masks) were adapted to our specific binary classification task, distinguishing only between building and background.

The model was optimized using stochastic gradient descent (SGD). As a multi-task framework, mask R-CNN is trained to minimize a combined loss function L_{total} that integrates several components (Eq. 4)

$$L_{total} = L_{rpn_cls} + L_{rpn_box} + L_{cls} + L_{box} + L_{mask} \quad (4)$$

where:

- L_{rpn_cls} and L_{rpn_box} are the classification and regression losses from the Region Proposal Network (RPN).
- L_{cls} and L_{box} are the classification and bounding-box regression losses for the detected objects.
- L_{mask} is the average binary cross-entropy loss for the mask head, which is calculated for each proposed region of interest.

During inference, the raw masks generated by mask R-CNN often exhibited irregular and noisy boundaries. To produce geometrically consistent footprints suitable for 3D modeling, a post-processing pipeline was applied to each predicted instance mask. This involved:

- **Morphological closing:** A morphological closing operation with a 5x5 rectangular kernel was used to fill small gaps and smooth the edges of the binary masks.
- **Contour extraction:** The largest connected contour was extracted from the refined mask.
- **Boundary simplification:** The contour was simplified using the Douglas-Peucker algorithm to reduce vertex count while preserving the overall shape, resulting in a regularized polygon.
- **Polygon filling:** The simplified polygon was finally filled to produce a clean, geometrically consistent building footprint.

For roof type classification, we implemented a ResNet-50 architecture (He et al., 2016) pre-trained on ImageNet. The model was adapted through targeted fine-tuning: only the final residual block (layer4) and classification head were unfrozen to preserve general features while adapting to roof-specific characteristics. The final fully-connected layer was replaced with a custom classifier (512 units, ReLU, Dropout p=0.3) matching the number of roof classes.

To address class imbalance, we employed weighted random sampling based on inverse class frequencies. The model was trained with cross-entropy loss using Adam optimizer with dual learning rates (1e-5 for backbone, 1e-4 for classifier). Training utilized standard data augmentation (random cropping, flipping, color jitter), and the best model was selected based on validation accuracy over 25 epochs (batch size=32).

Tree crown detection is performed using the DeepForest model (Weinstein et al., 2019; Weinstein et al., 2020) to identify vegetation locations and enhance the visual completeness of the reconstructed scene. Unlike the three core components, DeepForest serves a supplementary visualization role and was not quantitatively evaluated.

The final 3D scene was constructed by integrating outputs from all deep learning components using the Open3D and Trimesh libraries. The pipeline transforms 2D model predictions into a coherent 3D urban model through several systematic stages.

A base terrain mesh was created using the DTM derived from the Pix2Pix-generated DSM. Each DTM pixel was converted to a 3D vertex, with X-Y coordinates normalized and Z-values scaled by a height factor to approximate real-world elevation. Triangular faces connect these vertices to form a continuous surface, textured using the original RGB imagery and equipped with calculated surface normals for realistic lighting.

Building footprints from mask R-CNN instance segmentation masks were extruded into 3D volumes. Base heights were determined from average DTM values within each footprint, while building heights were calculated from normalized DSM values within a predefined realistic range. Walls were generated by connecting base and roof contours, with roof structures shaped according to ResNet-50 classifications.

Complex roof geometries were triangulated using Delaunay algorithms, and separate textures were mapped to walls and roofs using UV coordinates.

Tree locations identified by DeepForest were incorporated using template tree models in GLB format. Each detected tree was positioned at its bounding box center, scaled to appropriate dimensions, and grounded to the terrain using local DTM elevation data.

The complete scene, combining terrain, buildings, and vegetation, was exported in GLB format for visualization and interoperability with web-based viewers and 3D modeling software.

The data used for this research was sourced from several publicly available datasets. For the task of DSM generation, SpaceNet Urban 3D challenge dataset (Goldberg et al., 2017) was used. The dataset consists of high-resolution aerial images of three different cities in RGB and DSM format. This dataset also contains ground-truth instance labels of buildings, which were used in building segmentation task of this project. For the building segmentation task; previously mentioned Urban 3D Challenge was used alongside Inria Aerial Labeling Dataset (Maggiori et al., 2017). This dataset consists of high-resolution aerial orthorectified imagery as well as the building mask images. The dataset covers an area of 810 km² and was added to increase the image size and help the model understand different types of buildings. The roof classification model was trained using the dataset provided by GATE institute, consisting of orthophotos in 10 cm/pix resolution (Hristov et al., 2023). A small roof classification dataset from RoboFlow was also added into the mix (Surya, 2024).

The performance of each component in the pipeline was assessed quantitatively. The accuracy of the DSM generation was measured using mean squared error (MSE), root MSE and Structural Similarity Index (SSIM). The building segmentation performance of the mask R-CNN model was evaluated using the intersection over union (IoU) metric. After inference, each predicted building mask was post-processed to reduce noise and irregular boundaries. The binary masks were refined using morphological closing with 5x5 rectangular kernel to fill small gaps and smooth the edges. The largest connected contour was then extracted and simplified using the Douglas-Peucker algorithm followed by polygon filling to produce geometrically consistent footprints. For the roof type classification, standard metrics including accuracy, precision, recall, and F1-score were calculated. Because the tree detection model lacks ground-truth tree location data, no metric calculations were performed at that stage.

RESULTS

The results demonstrate the viability of the monocular approach for 3D reconstruction. The Pix2Pix-based model for DSM generation was iteratively improved over four versions. The final version, which incorporated a VGG19 based perceptual loss function, achieved the best performance with a final MSE of 986.70, RMSE of 31.4 and SSIM of 0.697 on the test dataset. The MSE was computed by comparing the absolute per-pixel intensity differences between predicted and ground-truth DSMs across all test images (8-bit scale, range 0-255). The SpaceNet Urban 3D dataset originally provides DSM values in meters. During preprocessing, these were normalized to 8-bit (0-255) for compatibility with our

Pix2Pix implementation. Based on typical elevation ranges for Jacksonville (approximately 0-50 meters), 1 gray level corresponds to roughly 0.2 meters, and our RMSE of 31.4 gray levels translates to approximately 6.2 meters. All quantitative results in this study are reported on the normalized scale for consistency with the training pipeline. Each pixel error was squared and averaged across the dataset. A quantitative comparison, shown in Figure 1, reveals that the model using VGG19 based perceptual loss produces significantly sharper and more structurally accurate elevation maps compared to initial versions, which suffered from more pronounced blurring and artifacts. Although the absolute value appears numerically high, it corresponds to an average absolute deviation of approximately 31 gray levels per pixel, or about 12% of the full dynamic range. This indicates that the majority of regions maintain moderate accuracy, while a smaller subset of pixels with large deviations, amplified by the squaring operation, contributes disproportionately to the overall error. Approximately 34% of the test pixels exhibit an absolute difference within 0-5 gray levels, suggesting locally stable performance despite the global MSE being influenced by outliers and high-contrast regions. Table 2 shows that 90% of pixels had an absolute error of less than 50 gray levels. The SSIM of 0.697 indicates that the predicted DSMs preserve structural similarity with ground-truth despite absolute pixel wise errors. To further analyze elevation accuracy across different scene components, we performed a stratified evaluation using building masks generated by our mask R-CNN model. Pixels were classified into three categories: building interiors, building boundaries (3-pixel buffer) and non-building areas. Building interiors exhibited the highest error (RMSE: 37.8, MAE: 28.2), followed by building boundaries (RMSE: 30.9, MAE: 23.1), while non-building areas showed the lowest error (RMSE: 30.4, MAE: 20.9). This confirms that elevation estimation is most challenging on building structures. To visually illustrate the spatial distribution of errors, error heatmaps from selected test samples are presented in Figure 2. The highest errors were concentrated at building edges and in areas of deep shadow, where visual cues for elevation are ambiguous.

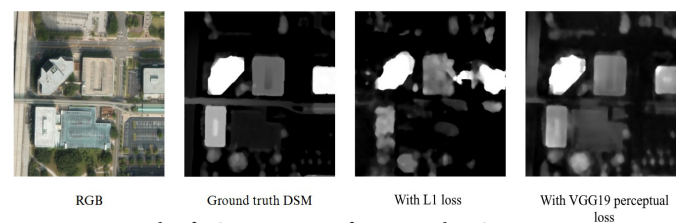


Figure 1. Example of DSM generation from a single RGB image

DSM: Digital surface model, RGB: Red, green, blue, VGG19: Visual geometry group 19-layer network

Table 2. Pix2Pix model absolute error-pixel count	
Absolute error range	Pixel count
0-5	18,559,062
5-10	6,252,400
10-20	10,001,808
20-30	6,840,507
30-50	7,375,825
50-100	4,497,386
100-150	716,163
150-255	201,466

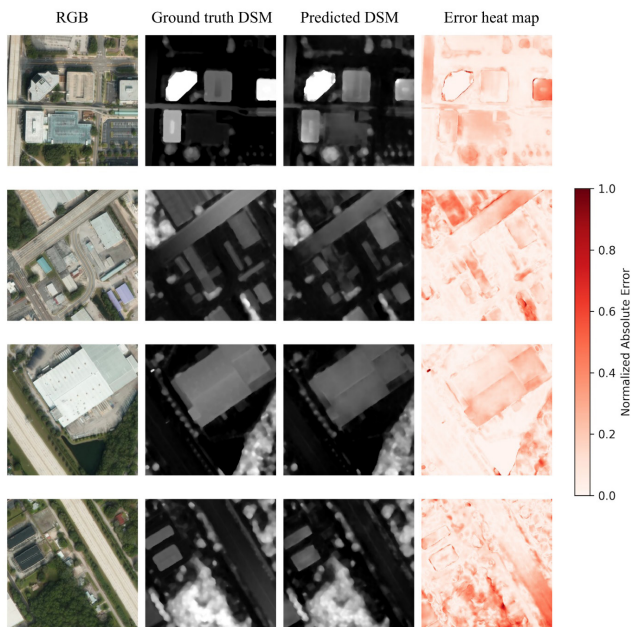


Figure 2. Error heatmaps of predicted DSM images

DSM: Digital surface model, RGB: Red, green, blue

The performance of the fine-tuned mask R-CNN model for building segmentation was evaluated using mean average precision (mAP) metric. The model achieved an mAP of 0.712 at an IoU threshold of 0.50, indicating a strong capability to correctly identify and locate building instances at a moderate overlap level. This demonstrates that the model can successfully distinguish between building and non-building areas for general detection purposes. However, the model's performance in precisely delineating boundaries was limited, as evidenced by a significant drop to 0.358 mAP at the stricter IoU threshold of 0.75. This sharp decline suggests the model struggles to define precise building contours and may face challenges in segmenting adjacent structures in dense urban environments. The overall COCO mAP (averaged over IoU thresholds from 0.50 to 0.95) was 0.383. Nevertheless, the obtained values remain acceptable for many building detection applications and were deemed sufficient for the purposes of this project, which prioritizes a complete, functional pipeline over production-grade segmentation accuracy. An example RGB image along with its ground-truth and predicted mask is shown in [Figure 3](#).



Figure 3. Comparison of ground-truth mask and predicted mask

RGB: Red, green, blue

The roof classification model achieved an overall accuracy of 80.0% across five classes. The performance was particularly strong for the most common class, hip roofs, which achieved an F1-score of 0.91. Performance varied across the remaining classes, with the model effectively identifying complex roofs (0.73 F1-score) but showing lower precision and recall for less frequent types, such as gable and flat. A pyramid class was also

included; however, it contained only three samples in the test set, making the results unreliable and not representative of the model's actual performance for this category. The weighted average F1-score was 0.80, indicating robust classification performance overall. A summary of class-wise performance is illustrated in the confusion matrix ([Figure 4](#)), which highlights both the strengths in recognizing common roof types and the limitations associated with underrepresented classes.

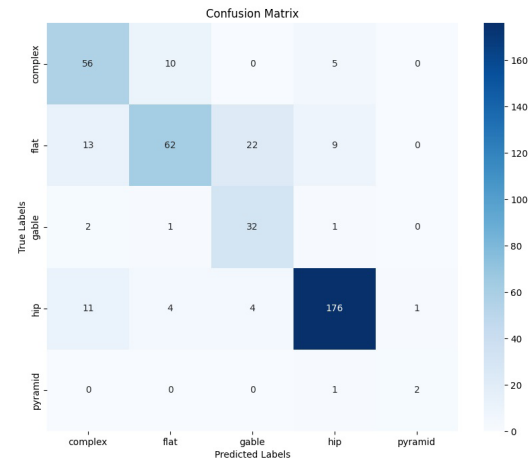


Figure 4. Confusion matrix of roof classification model

Tree crown detection was performed using the pre-trained DeepForest model to enrich the visual completeness of the reconstructed scene. The model identified vegetation locations and generated crown boundaries that were integrated into the final 3D reconstruction. As this component served a supplementary visualization role rather than a core analytical function, quantitative evaluation was not performed, as ground-truth tree annotations were unavailable.

We compared the performance of each pipeline component against recent studies employing similar methodologies. It should be noted that direct numerical comparison is challenging due to difference in datasets, evaluation metrics and output representations across studies. Nevertheless, the following comparisons provide a qualitative assessment of our method's relative performance.

For DSM generation, Ghamisi and Yokoya (2018) reported RMSE values of 3.89 m, 2.58 m and 3.66 m on Potsdam, Vaihingen and Stockholm datasets respectively, using a cGAN based approach. Our method achieves an RMSE of 31.4 on the normalized 8-bit scale, which corresponds to roughly 6.2 m. While this places our absolute accuracy below specialized urban DSM studies, several factors should be considered such as their dataset's different elevation distribution and complexity. Importantly, our SSIM of 0.697 indicates strong structural fidelity, demonstrating that perceptual loss effectively preserves elevation patterns despite the scale differences.

Table 3 compares our building segmentation results with recent studies using similar architectures. Hou and Li (2024) reported approximately 95% classification accuracy on UAV aerial imagery, while Ramadhani and Alya (2024) achieved a precision of 0.85 on UAV orthophoto data. More recently, Vincent and Varalakshmi (2024) obtained a mAP@50 of 0.79 on the WHU and Inria datasets using mask R-CNN. Although direct numerical comparison is limited by differing datasets and evaluation metrics, these results contextualize

Table 3. Comparison of recent building segmentation studies

Study	Method	Dataset	Metric	Result
Ramadhani & Alya (2024)	Mask R-CNN	UAV orthophoto	Precision	0.85
Hou & Li (2024)	Mask R-CNN (optimized)	UAV aerial	Classification accuracy	95%
Vincent & Varalakshmi (2024)	Mask R-CNN	WHU+Inria	mAP@50	0.794
This study	Mask R-CNN	Inria+SpaceNet	mAP@50	0.712

R-CNN: Region-based convolutional neural network, UAV: Unmanned aerial vehicle, WHU: Wuhan University, mAP: Mean average precision

the performance of mask R-CNN for building extraction tasks. Our lower mAP@50 of 0.71 can be attributed to the use of the SpaceNet dataset which includes more dense vegetation, presenting additional challenges compared to the datasets used in comparative studies.

Our roof classification model's performance aligns with the upper tier of results in current remote sensing literature, where architectural complexity often limits accuracy. A comparative analysis can be seen in Table 4.

Table 4. Comparison of recent roof type classification studies

Study	Method	Classes	Accuracy
Buyukdemircioglu et al. (2021)	CNN	6 types	80%-86%
Shailja et al. (2025)	Swin transformer	4 types	75%
Ölçer et al. (2023)	Siamese network	3 types (flat, gable, hip)	73%
This study	ResNet-50	5 types	80%

CNN: Convolutional neural network, ResNet-50: 50-layer residual network

The integration pipeline successfully synthesized all component outputs into a coherent 3D urban model, as illustrated in Figure 5. The final reconstruction demonstrates the practical viability of the monocular approach, producing a visually consistent scene with recognizable terrain features, building masses, and vegetation distribution.



Figure 5. Final 3D scene

3D: Three-dimensional

Qualitative assessment reveals several key outcomes; the terrain mesh accurately reflects the topographic relief of the area, building footprints maintain plausible geometric relationships, and roof types are correctly instantiated according to classification results. The scene successfully captures the overall urban layout and spatial distribution of structures.

However, systematic analysis also identified characteristic limitations. Primary failure modes include: (1) occasional building height inaccuracies in densely shadowed regions where DSM estimation proved challenging, (2) simplified roof geometries that occasionally miss intricate architectural details, and (3) minor misalignments between building footprints and terrain mesh at steep gradients. These limitations reflect the inherent constraints of monocular reconstruction

but do not fundamentally compromise the scene's utility for visualization and urban analysis applications.

The exported GLB model maintains all semantic groupings (terrain, buildings, vegetation) and proved compatible with major 3D viewing platforms, confirming the pipeline's practicality for downstream applications.

DISCUSSION

The extraction of semantically meaningful information from satellite imagery through deep learning has emerged as a prominent research direction in recent years. Within this domain, 3D reconstruction leveraging deep learning methodologies has constituted an active area of investigation. However, 3D reconstruction from satellite imagery presents distinct challenges attributable to the inherent characteristics of the data modality. Prevailing approaches in the literature predominantly depend on multi-stereo imagery (Tarabalka et al., 2022) or structure-from-motion (SfM) techniques, which utilize overlapping views to triangulate depth information through photogrammetric principles. In contrast, our monocular approach operates under significantly weaker geometric constraints, trading strict data requirements for inherent limitations in reconstruction accuracy. As Panagiotou et al. (2020) observe, single-image elevation estimation represents an ill-posed problem that inherently yields lower fidelity compared to multi-view systems. The principal advantage of our method lies in its reduced data dependency (requiring only a single satellite pass) though this practical benefit comes at the cost of diminished geometric precision. Furthermore, while off-nadir imagery has been exploited for height estimation in other studies (Li et al., 2024), our work is exclusively limited to nadir-facing orthoimagery, and consequently does not address the potential advantages of oblique viewing angles.

For DSM generation, we employ a Pix2Pix conditional generative adversarial network (cGAN) to infer elevation data from single RGB inputs. This approach aligns with established literature demonstrating that cGANs can effectively learn complex mappings from optical imagery to elevation profiles (Panagiotou et al., 2020). Our implementation initially utilized a standard L1 loss function; however, empirical evaluation revealed that a custom loss function incorporating VGG-based perceptual loss yielded superior performance in preserving structural details. Nevertheless, monocular DSM estimation remains fundamentally constrained by its inherent limitations. The absence of multiple viewpoints prevents the recovery of occluded surfaces and undercuts geometries, resulting in incomplete scene reconstruction. In practice, our Pix2Pix-generated DSMs exhibit high fidelity in planar regions with distinctive textural patterns but demonstrate reduced reliability in areas characterized by abrupt topographic transitions or shadow occlusion. Furthermore,

dataset limitations present additional challenges. Improved generalization would require more geographically diverse training data encompassing varied architectural styles and urban morphologies. Future work could explore more advanced monocular depth estimation architectures, such as DPT or MiDaS, which may offer enhanced performance for satellite-based elevation modeling.

For building segmentation, we employ mask R-CNN for instance-aware extraction, aligning with established methodologies in recent literature. This approach follows the precedent set by Zhao et al. (2018), who adopted mask R-CNN “as a base network due to its simplicity” for generating polygonal building masks from satellite imagery. The framework produces pixel-precise instance masks that enable precise rooftop delineation. This instance-level paradigm proves particularly valuable in dense urban environments, where it effectively disambiguates adjacent structures that might otherwise merge in semantic segmentation approaches. However, a recognized limitation of mask R-CNN is its tendency to generate masks with irregular, noisy boundaries. Our observations corroborate this phenomenon: while the model achieves reasonable detection accuracy (mAP 0.712 at IoU = 0.50), the resulting building contours frequently exhibit jagged or fragmented edges. Although a morphological post-processing step was implemented, it yielded only marginal improvements in IoU. This suggests that the underlying segmentation quality, rather than boundary smoothing alone, limits refinement potential. Several studies have proposed methods for regularizing building footprints (Kong et al., 2023), such post-processing refinements were not implemented in our current pipeline. Additionally, dataset limitations analogous to those affecting DSM generation persist. Improved generalization would benefit from expanded training data encompassing greater architectural diversity and geographic variation. Therefore, future work should investigate footprint regularization methods and incorporate additional training data to improve generalization across architectural styles and regions.

For roof type classification, we implemented a fine-tuned ResNet-50 convolutional neural network, a methodological choice consistent with established practices in the literature. This approach aligns with the work of Buyukdemircioglu et al. (2021), who demonstrated that a fine-tuned ResNet-50 architecture achieves competitive accuracy across six roof categories using a dataset of 10,000 roof images. Their findings confirmed the model’s efficacy on very-high-resolution aerial imagery, a result our study corroborates. We observe similar performance trends, with ResNet-50 effectively learning discriminative features for common roof types including flat, gable, and hip configurations when trained on adequate labeled exemplars. Nevertheless, classification performance remains contingent upon dataset scale and diversity, as noted in prior research. In our implementation, misclassifications predominantly occur among categories with minimal visual distinction or when roof features are partially occluded. While future enhancements could explore expanded class inventories or more sophisticated architectures, ResNet-50 provides a robust baseline, as substantiated by its consistent performance across multiple studies in the domain.

Tree crown detection was implemented using the DeepForest model, a pre-trained deep learning framework specifically designed for individual tree detection in RGB imagery. Unlike

the other pipeline components, which were trained or fine-tuned on task-specific datasets, DeepForest was deployed in its pre-trained state. While the model successfully identified tree locations and generated crown polygons for 3D scene integration, the absence of ground-truth tree annotations in our dataset precluded quantitative validation. This limitation is common in urban reconstruction pipelines where vegetation mapping is secondary to build infrastructure.

Overall, the favorable aspects of our approach are that it operates with minimal input data and integrates DSM generation, segmentation and roof classification in a single pipeline. It requires only one optical image to produce useful 3D urban models, and aligns with recent evidence that deep learning can approximate geometric reconstruction in a purely monocular setting. On the other hand, the limitations are equally clear. The DSM estimates lack the precision of multi-view methods, the segmentation masks require additional refinement to be production-ready, and the roof classifier depends on limited classes and labeled data. Our study was also restricted to high-resolution urban imagery, leaving its applicability to rural or lower resolution data untested.

CONCLUSION

As a result, the study demonstrates that a monocular deep learning-based pipeline can indeed generate useful 3D urban scenes, combining DSM prediction, building segmentation, tree crown detection and roof classification. The method reduces acquisition demands and computation compared to traditional multi-view stereo approaches, though with an expected loss in geometric accuracy. Future work should focus on integrating post-processing for building outlines, enriching the dataset and exploring additional modalities. Despite their limitations, the results show that monocular pipelines are viable for rapid urban analysis when only single view satellite data are available, echoing findings in prior literature.

ETHICAL DECLARATIONS

Ethics Committee Approval

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study.

Financial Disclosure

This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the 2209-A Research Project Support Programme for Undergraduate Students and SAYZEK-ATP.

Author Contributions

Concept: ÖCK, NA; Design: ÖCK, NA; Control: ÖCK, NA; Resources: ÖCK, NA; Materials: ÖCK, NA; Data Collection and/or Processing: ÖCK, NA; Analysis and/or Interpretation: ÖCK, NA; Literature Review: ÖCK, NA; Writing the Article: ÖCK, NA; Critical Review: ÖCK, NA.

Acknowledgments

The work was also carried out within the framework of the SAYZEK-ATP (Artificial Intelligence Talent Cluster of Defence Industry-Academic Thesis Programme). The authors would like to express their gratitude to Cevahir Çığla from ASELSAN, who served as project mentor under the SAYZEK-ATP program and provided valuable guidance throughout the study.

REFERENCES

- Buyukdemircioglu, M., Can, R., & Kocaman, S. (2021). Deep learning based roof type classification using very high resolution aerial imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B3-2021, 55-60. <https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-55-2021>
- Chen, S., Shi, Y., Xiong, Z., & Zhu, X. X. (2023). HTC-DC Net: monocular height estimation from single remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-18. <https://doi.org/10.1109/TGRS.2023.3321255>
- Ghamisi, P., & Yokoya, N. (2018). IMG2DSM: height simulation from single imagery using conditional generative adversarial net. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 794-798. <https://doi.org/10.1109/LGRS.2018.2806945>
- Goldberg, H., Brown, M., & Wang, S. (2017). A benchmark for building footprint classification using orthorectified RGB imagery and digital surface models from commercial satellites. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)* (pp. 1-7). IEEE. <https://doi.org/10.1109/AIPR.2017.8457973>
- He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2020). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 386-397. <https://doi.org/10.1109/TPAMI.2018.2844175>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). IEEE. <https://doi.org/10.1109/CVPR.2016.90>
- Hou, T., & Li, J. (2024). Application of mask R-CNN for building detection in UAV remote sensing images. *Heliyon*, 10(19), e38141. <https://doi.org/10.1016/j.heliyon.2024.e38141>
- Hristov, E., Petrova-Antonova, D., Petrov, A., Borukova, M., & Shirinyan, E. (2023). Imagery dataset for rooftop detection and classification. *Zenodo*. <https://doi.org/10.5281/zenodo.7633595>
- Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5967-5976). <https://doi.org/10.1109/CVPR.2017.632>
- Kong, L., Qian, H., Wu, Y., Niu, X., Wang, D., & Huang, Z. (2023). Simplification and regularization algorithm for right-angled polygon building outlines with jagged edges. *ISPRS International Journal of Geo-Information*, 12(12), 469. <https://doi.org/10.3390/ijgi12120469>
- Li, W., Yang, H., Hu, Z., Zheng, J., Xia, G. S., & He, C. (2024). 3D building reconstruction from monocular remote sensing images with multi-level supervisions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 27728-27737). IEEE. <https://doi.org/10.1109/CVPR.2024.02702>
- Liu, C.-J., Krylov, V. A., Kane, P., Kavanagh, G., & Dahyot, R. (2020). IM2ELEVATION: building height estimation from single-view aerial imagery. *Remote Sensing*, 12(17), 2719. <https://doi.org/10.3390/rs12172719>
- Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 3226-3229). IEEE. <https://doi.org/10.1109/IGARSS.2017.8127684>
- Ölçer, N., Ölçer, D., & Sümer, E. (2023). Roof type classification with innovative machine learning approaches. *PeerJ Computer Science*, 9, e1217. <https://doi.org/10.7717/peerj-cs.1217>
- Panagiotou, E., Chochlakis, G., Grammatikopoulos, L., & Charou, E. (2020). Generating elevation surface from a single RGB remotely sensed image using deep learning. *Remote Sensing*, 12(12), 2002. <https://doi.org/10.3390/rs12122002>
- Ramadhani, A., & Alya, N. F. (2024). Building extraction from unmanned aerial vehicle imagery using mask-RCNN (case study: Institut Teknologi Sepuluh Nopember, Surabaya). *E3S Web of Conferences*, 600, 06003. <https://doi.org/10.1051/e3sconf/202460006003>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention-MICCAI 2015 (Lecture Notes in Computer Science, Vol. 9351, pp. 234-241)*. Springer. https://doi.org/10.1007/978-3-319-24574-4_28
- Shailja, Yadava, M., Dwivedi, R., & Pardeshi, P. J. (2025). Classifying roof types from orthophotos: a Swin Transformer-based approach. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-5/W2-2025, 577-584. <https://doi.org/10.5194/isprs-annals-X-5-W2-2025-577-2025>
- Surya. (2024, July). *Roof-Type-Classifier dataset* [Data set]. Roboflow Universe. <https://universe.roboflow.com/surya-dmfwf/roof-type-classifier-bafod>
- Tarabalka, Y., Girard, N., Tripodi, S., Larrosa, C., Bauchet, J. P., Fonteix, G., ..., & Madelain, V. (2022). Reconstruction of 3D environments from satellite images by AI and computational geometry for exploitation in mixed reality. *STO-MP-MSG-197*. NATO Science and Technology Organization.
- Vincent, J. M., & Varalakshmi, P. (2024). Extraction of building footprint using mask-RCNN for high resolution aerial imagery. *Environmental Research Communications*, 6(7), 075015. <https://doi.org/10.1088/2515-7620/ad5b3d>
- Wei, S., Zhang, T., Ji, S., Luo, M., & Gong, J. (2023). BuildMapper: a fully learnable framework for vectorized building contour extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197, 87-104. <https://doi.org/10.1016/j.isprsjprs.2023.01.015>
- Weinstein, B. G., Marconi, S., Aubry-Kientz, M., Vincent, G., Senyondo, H., & White, E. P. (2020). DeepForest: a Python package for RGB deep learning tree crown delineation. *Methods in Ecology and Evolution*, 11(12), 1743-1751. <https://doi.org/10.1111/2041-210X.13472>
- Weinstein, B. G., Marconi, S., Bohlman, S. A., Zare, A., & White, E. P. (2019). Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks. *Remote Sensing*, 11(11), 1309. <https://doi.org/10.3390/rs11111309>
- Zhang, W., Qi, J., Wan, P., Wang, H., Xie, D., ..., & Yan, G. (2016). An easy-to-use airborne LiDAR data filtering method based on cloth simulation. *Remote Sensing*, 8(6), 501. <https://doi.org/10.3390/rs8060501>
- Zhao, K., Kang, J., Jung, J., & Sohn, G. (2018). Building extraction from satellite images using mask R-CNN with building boundary regularization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 242-2424). IEEE. <https://doi.org/10.1109/CVPRW.2018.00045>