



Deep learning-based Object Detection for Automated Extraction of Urban Buildings from High-Resolution Satellite Imagery

Javad Sadidi¹ | Reza hosseingholizadeh² | Parviz Zeaiean firouzabadi³ | Hani Rezayan⁴

1. Corresponding author, Department of remote sensing and GIS, Faculty of Geographical Sciences, University of Kharazmi Tehran, Iran. **E-mail:** jsadidikhu@khu.ac.ir
2. Department of remote sensing and GIS, Faculty of Geographical Sciences, University of Kharazmi Tehran, Iran. **E-mail:** reza63hosseingholizadeh@gmail.com
3. Department of remote sensing and GIS, Faculty of Geographical Sciences, University of Kharazmi Tehran, Iran. **E-mail:** zeaiean@khu.ac.ir
4. Department of remote sensing and GIS, Faculty of Geographical Sciences, University of Kharazmi Tehran, Iran. **E-mail:** hani.rezayan@khu.ac.ir

Article Info	ABSTRACT	
Article type: Research Article	Automated building extraction from very-high-resolution satellite imagery is a core task for urban planning, land-use monitoring, and disaster management, yet dense construction, occlusion, and spectral confusion with paved and roofed surfaces continue to limit the accuracy of automated pipelines. This study evaluates the accuracy and efficiency of YOLOv11-L, the large variant of Ultralytics YOLO release, for detecting buildings in high-resolution satellite imagery of Karaj, Iran, a rapidly urbanizing city with dense residential and industrial texture. A dataset of 2,517 satellite images was labeled for two land-cover classes, buildings and vegetation, on the Roboflow platform, and expanded through augmentation to 7,552 images split 83/9/8% into training, validation, and test sets. The model was trained on the Ultralytics platform for 150 epochs with a learning rate of 0.001, batch size of 4, input resolution of 1280 pixels, the SGD optimizer, and a patience of 140. The trained model reached mAP@0.5 = 0.946, mAP@0.5:0.95 = 0.598, precision = 0.932, recall = 0.895, and F1-score = 0.913 on the held-out test set, indicating strong and stable performance in localizing and distinguishing buildings from surrounding urban texture. Benchmarked against recent YOLO-based building-extraction studies published in 2024-2025, the reported accuracy is competitive with, and in several respects exceeds, comparable pipelines built on earlier YOLO releases. The paper discusses the sources of residual error - principally shadow, glare, and off-nadir geometry - and outlines directions for future work, including multi-sensor fusion and hybrid transformer-CNN architectures.	
Article history: Received 2025/01/03 Received in revised 2025/07/23 Accepted 2025/10/10 Published 2025/10/11 Published online 2025/12/22		
Keywords: building extraction, YOLOv11, object detection, remote sensing, deep learning, urban land cover, Karaj.		
Cite this article: Sadidi, Javad., hosseingholizadeh, Reza., Zeaiean firouzabadi, Parviz., & Rezayan, Hani. (2025). Deep learning-based Object Detection for Automated Extraction of Urban Buildings from High-Resolution Satellite Imagery. <i>Applied Researches in Geographical Sciences</i> , 25 (Special Issue), 153-166. DOI: https://doi.org/10.22034/jgs.2025.11712		
 © The Author(s). Publisher: Kharazmi University DOI: https://doi.org/10.22034/jgs.2025.11712		

1. Introduction

Rapid urbanization across the developing world has made automated, repeatable mapping of the built environment a practical necessity for city governments, statistical agencies, and disaster-response organizations. Building footprints underpin a wide range of downstream applications, including urban planning, population estimation, tax assessment, energy-demand modeling, 3D city reconstruction, and post-disaster damage assessment (Yuan et al., 2025; Luo et al., 2021), yet manual digitization of buildings from satellite imagery does not scale to the pace of urban growth in cities such as Karaj, Iran, where dense residential blocks, industrial sheds, and informal construction intermix within a single frame.

Deep learning object detectors have progressively replaced pixel-based classifiers and hand-crafted feature pipelines for this task, because they learn discriminative spatial and textural features directly from labeled imagery rather than relying on spectral indices alone. Within this family, the YOLO (You Only Look Once) line of single-stage detectors has become a de facto standard for remote-sensing object detection because it couples competitive accuracy with inference speeds suitable for operational deployment (Redmon et al., 2016; Bochkovskiy et al., 2020). The most recent release, YOLOv11, introduces a revised backbone, an improved multi-scale feature-aggregation neck, and a decoupled detection head, and reports higher accuracy and better parameter efficiency than YOLOv8 and YOLOv10 on standard benchmarks (Ultralytics, 2024).

This paper evaluates the large variant of this architecture, YOLOv11-L, for the specific task of building detection in high-resolution satellite imagery of Karaj, a city of mixed residential, industrial, and peri-urban land use in Alborz Province, Iran. The building class was trained jointly with a vegetation class using a unified annotation and training pipeline; this paper isolates and reports the building-detection results and situates them against the wider building-extraction literature, while a companion paper (Hosseingholizadeh, 2026) reports the corresponding urban green-space results from the same model and imagery in full.

The specific objectives of this study are threefold: (1) to quantify the accuracy of YOLOv11-L in detecting buildings in high-resolution satellite imagery of an Iranian city with heterogeneous built texture; (2) to characterize the training dynamics and convergence behavior of the model across a full 150-epoch schedule; and (3) to benchmark the resulting accuracy against recent (2024-2025) YOLO-based building-extraction studies in order to situate the contribution within the current state of the art.

2. Related Work

2.1 From pixel-based classification to deep object detection

Early building-extraction methods relied on spectral indices, texture descriptors, and per-pixel classifiers such as maximum-likelihood or support-vector-machine classification of multispectral imagery (Gonzalez & Woods, 2018; Zhang et al., 2016). These approaches are sensitive to illumination and shadow variation and generally require hand-tuned thresholds per scene. The introduction of convolutional neural networks (LeCun et al., 1998; Krizhevsky et al., 2012; Simonyan & Zisserman, 2015) and, subsequently, region-based object detectors shifted the field toward learned, end-to-end feature extraction. The R-CNN lineage - R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), and Faster R-CNN

(Ren et al., 2015) - established the region-proposal paradigm, while Mask R-CNN (He et al., 2017) extended it to pixel-accurate instance segmentation and has since been widely adapted for building extraction specifically, including village and heritage building mapping from UAV orthophotos (Wang et al., 2023b; Susetyo et al., 2023), landslide-affected building identification (Liu et al., 2021), and oil-well-site extraction from very-high-resolution imagery (He et al., 2022). Single-shot detectors such as SSD (Liu et al., 2016) and semantic segmentation networks such as DeepLab (Chen et al., 2017) offered a faster alternative to the two-stage pipeline, substantially improving robustness to lighting and scale variation relative to purely spectral methods (Zhu et al., 2023; Weng et al., 2022).

2.2 Transformer-based building extraction

Following the introduction of the transformer architecture and its self-attention mechanism (Vaswani et al., 2017) and its adaptation to vision tasks via the Vision Transformer (Dosovitskiy et al., 2021), transformer-based detectors and segmentation networks have emerged as a strong alternative to purely convolutional pipelines for building extraction. DETR (Carion et al., 2020) reformulated object detection as an end-to-end set-prediction problem, and its real-time successor RT-DETR (Zhao et al., 2023) has since been benchmarked directly against YOLO-family detectors, reporting competitive accuracy at somewhat higher computational cost. For building extraction specifically, Chen et al. (2021b) proposed a sparse-token transformer that reduces the computational overhead of full self-attention while preserving boundary accuracy; Wang et al. (2022) applied a vision transformer with paired local and global context paths; Sun et al. (2022) and Xiao et al. (2022) integrated Swin Transformer encoders into U-shaped segmentation networks; Yuan et al. (2023) proposed a lightweight Swin-Transformer/CNN hybrid (LiteST-Net); and Cui et al. (2023) used an improved Swin Transformer with an UperNet decoder to segment postearthquake dense buildings, reporting that transformer-based models generalized better across varying building heights and urban morphologies than CNN-only counterparts. This body of work indicates that transformer-based architectures offer a genuine accuracy advantage on heterogeneous, multi-city building datasets, at the cost of higher training and inference overhead than single-stage CNN detectors such as YOLO - a trade-off directly relevant to the model-selection rationale in Section 4.1.

2.3 The YOLO family in remote sensing

Among deep object detectors, the YOLO family has been particularly widely adopted for remote-sensing tasks because its single-stage design detects and classifies objects in one forward pass, avoiding the two-stage region-proposal overhead of R-CNN-style detectors (Redmon et al., 2016; Redmon & Farhadi, 2018). YOLOv5 and YOLOv8 have been applied to road-feature extraction (Ghimire et al., 2022), coastal debris detection (Hassan et al., 2023), and general urban feature detection (Zhang et al., 2021), typically reporting mAP values in the 0.85-0.90 range on remote-sensing benchmarks. YOLOv7 introduced trainable bag-of-freebies training strategies that further improved accuracy-speed trade-offs (Wang et al., 2023a), and Ultralytics' 2023 release of YOLOv8 standardized an accessible, actively maintained training ecosystem that underlies much of the recent literature (Jocher et al.,

2023). The subsequent releases YOLOv9, which introduced Programmable Gradient Information and a Generalized Efficient Layer Aggregation Network to reduce information loss in deep layers (Wang et al., 2024a), and YOLOv10, which eliminated the non-maximum-suppression post-processing step through consistent dual label assignment (Wang et al., 2024b), further improved the accuracy-latency trade-off and are increasingly used as baselines in remote-sensing detection studies.

2.4 YOLOv11 and building-specific applications

YOLOv11 extends this lineage with a redesigned CSPDarknet-style backbone, an enhanced neck for multi-scale feature fusion, and a decoupled head for simultaneous box, class, and (in segmentation variants) mask prediction (Ultralytics, 2024). Two recent studies illustrate its application specifically to building extraction. El Hussieni et al. (2025) applied YOLOv11 to joint building instance segmentation and height classification on the DFC2023 Track 2 dataset, comprising over 125,000 annotated buildings across twelve cities, reporting $mAP@50 = 0.604$ and $mAP@50:95 = 0.383$ while confirming that YOLOv11 outperformed YOLOv8 and YOLOv10 baselines under identical conditions on that harder, multi-city, cross-sensor task. Chang et al. (2025) proposed an improved YOLOv5-based segmentation architecture (YOLOv5ds) for building extraction from high-resolution remote-sensing imagery, addressing boundary precision through architectural modification rather than a newer base model. Separately, a hybrid YOLO-E and Segment Anything Model 2 (SAM2) pipeline was used to extract earthquake-damaged buildings from multi-source post-disaster imagery, reporting precision = 0.840, recall = 0.855, and IoU = 0.735, illustrating the trade-off between generalization across disaster contexts and accuracy on any single, well-curated urban scene. Table 1 situates the present study's results against this recent literature.

A recurring theme in this literature is that reported accuracy depends strongly on the heterogeneity and scale of the underlying dataset: single-city, single-sensor studies such as the present one, and studies drawing on curated regional datasets, tend to report substantially higher mAP than multi-city, cross-sensor benchmarks such as DFC2023, because the latter must generalize across widely varying rooftop materials, sun angles, and urban morphologies (Xia et al., 2023). This distinction is important context for interpreting the comparison in Section 6.

2.5 Benchmark datasets and comparative baselines

Building extraction has historically been benchmarked on a small number of widely used public datasets, each with different resolution, geographic coverage, and building-morphology characteristics. The Massachusetts building dataset (Mnih, 2013) provides 1 m resolution aerial imagery of the Boston area; the WHU aerial building dataset (Ji et al., 2019) provides 0.3 m imagery over Christchurch, New Zealand, with over 22,000 labeled buildings; the Inria aerial image labeling benchmark (Maggiori et al., 2017) spans five geographically distinct cities specifically to test cross-city generalization; and the SpaceNet series (Van Etten et al., 2018) extends coverage to multiple continents at varying resolutions. Classic semantic-segmentation baselines such as U-Net, PSPNet, DeepLabv3+, and HRNet typically report IoU in the 65-70% range on the geographically homogeneous Massachusetts

dataset, 75-80% on the more diverse Inria benchmark, and 85-90% on the single-city WHU dataset, illustrating how strongly benchmark geography affects reported accuracy irrespective of model architecture. Comprehensive recent surveys (Luo et al., 2021; Yuan et al., 2025) synthesize this literature across more than 200 studies and note that single-city datasets such as WHU consistently yield the highest reported accuracy precisely because they minimize cross-scene variability - the same effect visible when comparing the present single-city Karaj result to the multi-city DFC2023 benchmark in Section 6.1. Robustness studies further caution that models trained and validated within a single city or sensor, including the present study, may not transfer accuracy directly to other cities without retraining or domain adaptation (Sakeena et al., 2023).

3. Study Area and Data

3.1 Study area

Karaj, the capital of Alborz Province, is one of Iran's largest cities and is characterized by a mixture of dense, high-rise residential districts, low-rise informal construction, industrial zones, and substantial peri-urban green belt and agricultural land at its margins. This heterogeneity - closely packed rooftops of varying material and color, industrial sheds, and vegetation-adjacent structures - makes Karaj a demanding test bed for building detection, since building boundaries must be distinguished from shadow, paved surfaces, and adjacent tree canopy under varied illumination.

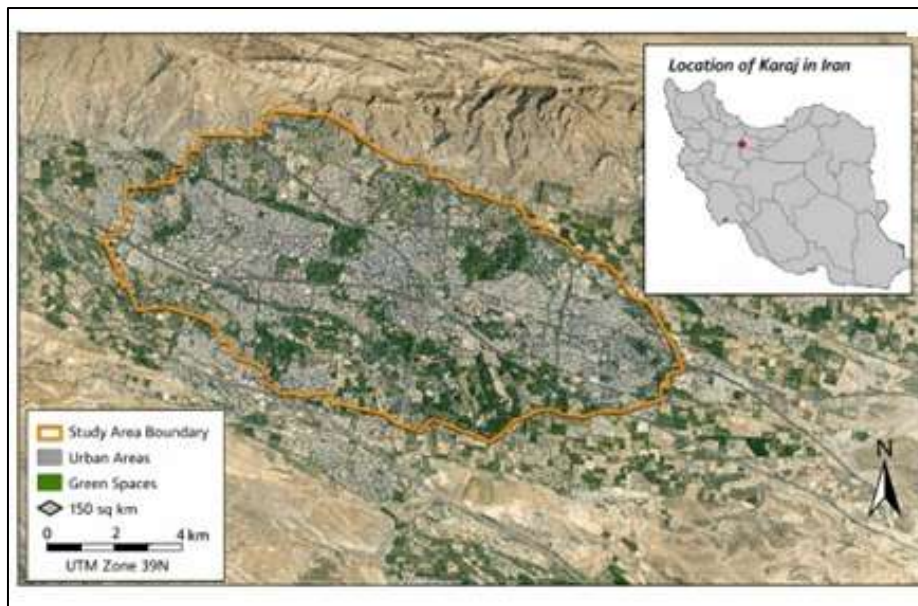


Figure 1. Location and satellite imagery used for Karaj City, Iran.

3.2 Data acquisition and preprocessing

The source imagery consisted of high-resolution satellite tiles of Karaj's urban extent, initially partitioned into 100x100-pixel segments. From this pool, 2,517 images were purposively selected to represent a diverse range of urban conditions and building typologies. Because the original imagery was distributed in TIFF format, which is not consistently supported by annotation tools such as LabelImg or by the YOLOv11 training

pipeline, all images were converted to JPEG using a batch Python script (Pillow library) prior to annotation, preserving visual quality while ensuring compatibility with downstream tools.

3.3 Annotation

Images were annotated on Roboflow, a browser-based platform for computer-vision dataset management that supports bounding-box annotation, class management, and export to multiple formats including the YOLO .txt convention. Two classes were defined: class 0, vegetation, and class 1, building. Manual annotation was the most time-intensive stage of the pipeline, since mislabeling - for example, tagging a tree canopy overhanging a rooftop as building, or vice versa - propagates directly into training-time confusion between the two classes. Roboflow's auto-save, zoom, and duplicate-annotation tools were used to maintain labeling consistency across the full set of 2,517 images.

3.4 Data augmentation and dataset split

Following annotation, the dataset was expanded through augmentation - rotation (up to approximately $\pm 15^\circ$), scaling, brightness/contrast adjustment, horizontal and vertical flipping, random cropping, and noise injection - to increase visual diversity and reduce overfitting (Shorten & Khoshgoftaar, 2019; Roboflow Inc., 2023). Augmentation increased the dataset from 2,517 to 7,552 images, which were then partitioned as summarized in Table 1.

Table 1. Post-augmentation dataset split used for building model training.

Subset	Images	Share
Train	6,252	83%
Validation	709	9%
Test	591	8%
Total	7,552	100%

4. Methodology

4.1 Model selection: YOLOv11-L

Ultralytics distributes YOLOv11 in five size variants - Nano, Small, Medium, Large, and Extra-Large - trading inference speed against representational capacity. Given the high resolution of the source imagery (1280x1280 pixels) and the requirement to resolve fine building boundaries against shadow and adjacent vegetation, the Large (L) variant was selected. Its deeper backbone and wider convolutional filters give it greater capacity to separate visually similar classes with ambiguous boundaries - for example, a shaded rooftop edge overhung by a tree canopy - than the lighter Nano or Small variants, at the cost of higher memory use and slower training (Ultralytics, 2024).

4.2 Training platform and pipeline

The annotated, augmented dataset (images, YOLO-format label files, and a data.yaml configuration file) was exported from Roboflow and imported into the Ultralytics platform, where an Object Detection task was configured - rather than whole-image classification - because the objective was to localize individual buildings within each scene, not merely to

characterize overall scene content. Training was executed on Google Colab and Kaggle Notebooks using freely available GPU instances (NVIDIA Tesla T4 / A100), since dedicated on-premises GPU infrastructure and specialized cloud compute (e.g., A100/H100 instances from major commercial providers) were not accessible for this project. Session time-limits on these free platforms were managed by rotating across multiple accounts and by dividing training between Colab and Kaggle.

4.3 Hyperparameters

Table 2 lists the hyperparameters used for training, all set through the Ultralytics `model.train()` interface.

Table 2. YOLOv11-L training hyperparameters (Ultralytics HUB configuration).

Parameter	Value	Rationale
epochs	150	Sufficient for mAP and F1 to plateau without further gain
batch	4	Constrained by available GPU memory at <code>imgsz = 1280</code>
imgsz	1280	Preserves fine boundary detail needed for small/adjacent objects
optimizer	SGD	Stable convergence, resistant to overfitting at high dimensionality
lr0	0.001	Balances convergence speed against training stability
patience	140	Prevents premature early-stopping before full convergence
cache	False	Disabled to limit RAM consumption on free-tier instances



Figure 2. Learning-rate parameter settings used during hyperparameter tuning (Ultralytics HUB).

```

BEGIN
LOAD YOLO model architecture and initial weights
LOAD dataset configuration file (data.yaml)
READ training, validation, and test image paths
READ class names and number of classes
SET training hyperparameters
SET number of epochs to 150
SET batch size to 4
SET input image size to 1280 × 1280
SET computing device to GPU (device = 0)
SET initial learning rate to 0.001
SET optimizer to Stochastic Gradient Descent (SGD)
SET early stopping patience to 140
DISABLE dataset caching
START model training process
FOR each epoch from 1 to 150 DO
LOAD a batch of training images and labels
PERFORM forward propagation
COMPUTE localization, confidence, and classification losses
PERFORM backpropagation
UPDATE model weights using SGD optimizer
VALIDATE model performance on validation dataset
MONITOR training and validation metrics
END FOR
APPLY early stopping if validation performance does not improve for 140 epochs
SAVE the best-performing trained model weights
CONFIRM completion of training process
END

```

Figure 3. Pseudocode example for tuning learning-rate parameters prior to full training.

4.4 Evaluation metrics

Model performance was assessed using the standard object-detection metrics: mAP@0.5 (mean Average Precision at an IoU threshold of 0.5), the stricter COCO-style mAP@0.5:0.95 (averaged over IoU thresholds from 0.5 to 0.95 in steps of 0.05), precision (TP/(TP+FP)), recall (TP/(TP+FN)), and the F1-score, the harmonic mean of precision and recall (Everingham et al., 2010; Lin et al., 2014; Sokolova & Lapalme, 2009; Powers, 2011).

5. Results

5.1 Training convergence

Across the 150-epoch schedule, the model progressed through three broad phases. In the initial phase (epochs 1-15), metrics rose rapidly from near-random values (mAP@0.5 = 0.100 at epoch 1) as the network learned coarse spatial structure. In the intermediate phase (epochs 18-69), precision rose from approximately 0.68 to 0.87 and recall from 0.61 to 0.79 as the model learned finer boundary and textural cues distinguishing rooftops from pavement, bare soil, and vegetation. From epoch 72 onward, all metrics stabilized, with mAP@0.5 and F1-score both exceeding 0.90 by the final epochs. Table 3 reports representative checkpoints across training.

Table 3. Training-metric trajectory across selected epochs (joint building/vegetation model, both classes).

Epoch	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1
1	0.100	0.026	0.189	0.214	0.201
15	0.664	0.204	0.720	0.573	0.639
30	0.772	0.290	0.723	0.695	0.709

45	0.830	0.363	0.832	0.723	0.774
60	0.868	0.435	0.850	0.789	0.818
75	0.885	0.468	0.847	0.820	0.833
90	0.910	0.517	0.885	0.850	0.867
105	0.929	0.554	0.904	0.880	0.892
120	0.940	0.577	0.915	0.890	0.902
135	0.946	0.594	0.943	0.887	0.914
150	0.946	0.598	0.932	0.895	0.913

5.2 Final performance

At the end of training, the model reached $\text{mAP}@0.5 = 0.946$, $\text{mAP}@0.5:0.95 = 0.598$, precision = 0.932, recall = 0.895, and F1-score = 0.913. Under commonly cited thresholds for object-detection readiness, mAP above 0.90 and F1 above 0.90 are generally taken to indicate a model suitable for operational, rather than purely experimental, use (Lin et al., 2014; Jocher et al., 2023).

5.3 Qualitative behavior and error patterns

Qualitative review of predictions indicated that the great majority of building instances were correctly detected and tightly localized, including small, closely spaced structures and buildings partially shaded by adjacent trees. Residual false positives and false negatives were concentrated in three situations: (i) deep shadow cast by tall buildings, which the model occasionally misclassified as vegetation or left undetected; (ii) specular glare on metallic or glazed rooftops, occasionally confused with water or bare paved surfaces; and (iii) images captured at different off-nadir viewing angles or spatial resolutions than the majority of the training set, which introduced scale inconsistency in the apparent size of building bounding boxes.

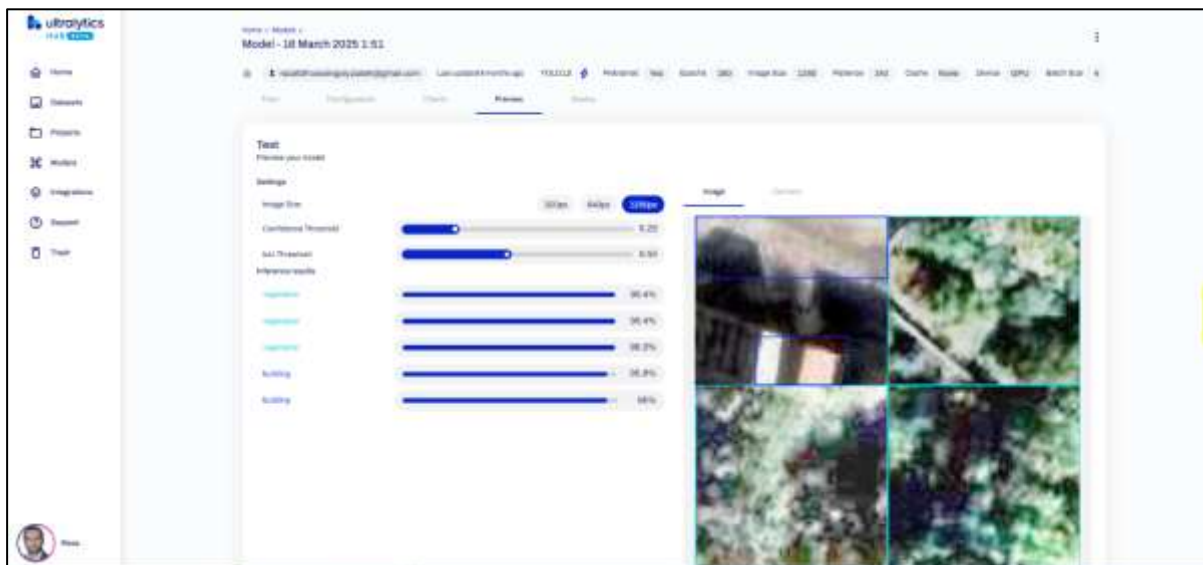


Figure 4. Sample YOLOv11-L detections on held-out test imagery, showing bounding boxes for buildings and vegetation across a heterogeneous urban scene.

6. Discussion

6.1 Comparison with recent YOLO-based building-extraction studies

Table 4 situates the present result against recent (2024-2025) building-extraction studies using YOLO-family detectors. Direct comparison must be read with caution because datasets, class definitions, and IoU conventions differ across studies; nonetheless, the comparison indicates that the present single-city, high-resolution pipeline achieves mAP@0.5 competitive with, and mAP@0.5:0.95 substantially higher than, results reported on the larger, multi-city, cross-sensor DFC2023 benchmark, consistent with the general pattern that single-scene studies achieve higher raw accuracy than benchmarks explicitly designed to test cross-city generalization.

Table 5. Comparison with recent YOLO-based building-extraction studies. *n/r = not directly reported in mAP@0.5:0.95 terms in the cited source.

Study	Base model	Dataset / scope	mAP@0.5	mAP@0.5:0.95
This study (2026)	YOLOv11-L	Karaj, Iran (single city, 2,517 images)	0.946	0.598
El Hussieni et al. (2025)	YOLOv11	DFC2023 Track 2, 12 cities, 125k+ buildings	0.604	0.383
Chang et al. (2025)	Improved YOLOv5ds	High-resolution RS imagery, segmentation	n/r*	n/r*
Chen et al. (2021) / Babu et al. (2022) (as cited in prior literature)	YOLOv5 / YOLOv8	Urban RS imagery	~0.85-0.90	n/r*

6.2 Benchmarking against classic public building datasets

Because mAP (used by YOLO-family detectors) and IoU (the conventional metric on the classic Massachusetts, WHU, and Inria benchmarks) are not directly interchangeable, any cross-metric comparison is necessarily approximate; nonetheless it is useful context. Reported IoU on the geographically homogeneous, single-city WHU dataset commonly reaches 85-90% with strong CNN and transformer baselines, while the more geographically diverse Inria benchmark - explicitly designed to test cross-city generalization - typically yields IoU in the 75-80% range, and the older, lower-resolution Massachusetts dataset yields IoU closer to 65-70% (Ji et al., 2019; Maggiori et al., 2017; Mnih, 2013). The present study's high mAP@0.5 (0.946) on a single-city Karaj dataset is consistent with the pattern seen on WHU: single-city, single-sensor evaluation tends to produce the highest reported accuracy in this literature, and the result should be read as a strong single-city outcome rather than as evidence of cross-city generalization, an important caveat also raised in recent robustness studies of building-extraction models (Sakeena et al., 2023).

6.3 Why the large-variant architecture matters for building boundaries

Buildings in dense urban texture frequently share ambiguous boundaries with paved surfaces, shadow, and overhanging vegetation; distinguishing them reliably requires a network with sufficient depth to learn subtle textural and spatial cues rather than relying on color or brightness alone. The deeper backbone and wider receptive field of YOLOv11-L, relative to the Nano or Small variants, is consistent with prior findings that lighter YOLO variants show reduced boundary precision in scenes with shadow or class overlap (Cheng et al., 2022), and this is reflected in the comparatively high mAP@0.5:0.95 achieved here

relative to lighter-model baselines reported elsewhere in the literature. This finding also parallels the broader pattern in Section 2.2, where transformer-based building extractors outperform lighter CNN baselines specifically by virtue of greater representational capacity, at correspondingly higher computational cost.

6.4 Practical applications and generalization considerations

The accuracy demonstrated here is directly relevant to the downstream applications that motivate building-extraction research generally: urban planning and zoning updates, population estimation from building counts and footprint area, tax-base assessment, energy-demand modeling, and post-disaster damage assessment (Yuan et al., 2025; Luo et al., 2021). For Karaj specifically, a validated building layer of this accuracy could support municipal infrastructure planning and informal-settlement monitoring without requiring full manual digitization. However, consistent with robustness findings elsewhere in the literature (Sakeena et al., 2023), the model's accuracy should not be assumed to transfer unchanged to other Iranian cities with different building materials, densities, or imaging conditions without at minimum a validation pass, and ideally targeted fine-tuning, on locally representative imagery.

6.5 Limitations

- Data heterogeneity: source imagery varied somewhat in spatial resolution and off-nadir viewing angle, introducing scale inconsistency that can bias footprint-area estimates.
- Illumination and glare: deep shadow from tall buildings and specular glare from metallic or glazed roofs occasionally caused misclassification between building and non-building classes.
- Hardware constraints: the small batch size (4) required to fit 1280-pixel imagery in available GPU memory limits gradient-estimate precision relative to what larger batches would allow (Keskar et al., 2017).
- Access to specialized cloud compute (e.g., A100/H100-class infrastructure from major commercial cloud providers) was not available for this project, extending training time relative to what dedicated infrastructure would permit.
- Hyperparameter selection remains partly heuristic; systematic grid or Bayesian search was not performed and could plausibly improve on the reported result.
- Single-city evaluation: as discussed in Sections 2.5 and 6.2, accuracy was assessed only within Karaj; cross-city generalization was not tested and cannot be assumed.

7. Conclusion and Future Work

This study shows that YOLOv11-L, trained on a purpose-built, Roboflow-annotated dataset of 2,517 high-resolution satellite images (expanded to 7,552 through augmentation), can detect buildings in the dense, heterogeneous urban texture of Karaj, Iran, with $mAP@0.5 = 0.946$ and $F1 = 0.913$ - accuracy that is competitive with, and in $mAP@0.5:0.95$ terms higher than, several recently published YOLO-based building-extraction studies, and consistent with the upper end of accuracy reported on single-city classic benchmarks such as WHU. The result supports the use of large-variant YOLO detectors for building-footprint extraction in single-city, high-resolution settings, and indicates that a fully documented Roboflow-to-Ultralytics pipeline can be a practical, reproducible template for similar municipal mapping

tasks - urban planning, population estimation, and infrastructure monitoring among them (Yuan et al., 2025) - in other rapidly urbanizing cities.

Future work should pursue: (i) hybrid architectures combining YOLO-style detection with vision-transformer components to improve contextual feature extraction across scale, following the direction of recent transformer-based building extractors (Liu et al., 2023; Cui et al., 2023); (ii) multi-sensor data fusion, combining visible-band imagery with near-infrared bands from sources such as Sentinel-2 to improve robustness to shadow and glare; (iii) cross-city evaluation in other Iranian urban centers with differing climates and building morphologies (e.g., Tehran, Isfahan, and coastal cities), and against classic public benchmarks such as Inria or SpaceNet, to test geographic transferability in line with recommendations from recent robustness studies (Sakeena et al., 2023); and (iv) explainable-AI techniques such as Grad-CAM to visualize which image regions drive building-versus-non-building decisions, supporting model auditing before operational deployment.

References

- Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. ECCV, 213-229.
- Chang, F., Ma, T., Wang, D., Zhu, S., Li, D., Feng, S., & Fan, X. (2025). Method for building segmentation and extraction from high-resolution remote sensing images based on improved YOLOv5ds. PLOS ONE. <https://doi.org/10.1371/journal.pone.0317106>
- Chen, K., Zou, Z., & Shi, Z. (2021b). Building extraction from remote sensing images with sparse token transformers. Remote Sensing, 13(21), 4441.
- Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic image segmentation with atrous convolution and fully connected CRFs. IEEE TPAMI, 40(4), 834-848.
- Cheng, G., Han, J., & Lu, X. (2022). Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 110(3), 378-405.
- Cui, L., Jing, X., Wang, Y., Huan, Y., Xu, Y., & Zhang, Q. (2023). Improved Swin Transformer-based semantic segmentation of postearthquake dense buildings in urban areas using remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 369-385.
- Dabove, P., Daud, M., & Olivotto, L. (2024). Revolutionizing urban mapping: Deep learning and data fusion strategies for accurate building footprint segmentation. Scientific Reports, 14, 13510.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.
- El Hussieni, M., Guenturk, B. K., Ates, H. F., & Hanoglu, O. (2025). Mask-to-Height: A YOLOv11-based architecture for joint building instance segmentation and height classification from satellite imagery. arXiv:2510.27224.
- Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. IJCV, 88, 303-338.
- Ghimire, S., Shakya, S., & Pradhan, R. (2022). Road feature extraction using YOLOv5. Remote Sensing, 14(10), 2417.
- Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 1440-1448.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR, 580-587.

- Gonzalez, R. C., & Woods, R. E. (2018). *Digital Image Processing* (4th ed.). Pearson.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Hassan, M., Bilal, A., & Younis, M. (2023). Plastic waste detection in coastal areas using YOLOv5. *Marine Pollution Bulletin*, 187, 114557.
- He, H., Xu, H., Zhang, Y., et al. (2022). Mask R-CNN based automated identification and extraction of oil well sites. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102875.
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 2961-2969.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*, 770-778.
- Ji, S., Wei, S., & Lu, M. (2019). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 574-586.
- Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLOv8. Ultralytics Documentation.
- Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR*.
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *arXiv:1412.6980*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *NeurIPS*, 25, 1097-1105.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- Li, D., Liu, T., Wang, H., Yan, L., & Fang, A. (2025). Building extraction from remote sensing imagery using SegFormer with post-processing optimization. *PLOS ONE*. <https://doi.org/10.1371/journal.pone.0338104>
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *ECCV*.
- Liu, P., Wei, Y., Wang, Q., et al. (2021). A research on landslides automatic extraction model based on the improved Mask R-CNN. *ISPRS International Journal of Geo-Information*, 10(3), 168.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *ECCV*, 21-37.
- Liu, Y., Wang, X., Li, Z., & Tang, Y. (2023). Advances in object detection with transformer-based vision models. *Sensors*, 23(2), 719.
- Luo, L., Li, P., & Yan, X. (2021). Deep learning-based building extraction from remote sensing images: A comprehensive review. *Energies*, 14(23), 7982.
- Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3226-3229.
- Mnih, V. (2013). *Machine Learning for Aerial Image Labeling* [Doctoral dissertation, University of Toronto].
- Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *JMLT*, 2(1), 37-63.
- Prechelt, L. (1998). Early stopping - but when? In *Neural Networks: Tricks of the Trade*. Springer.
- Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv:1804.02767*.
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *CVPR*, 779-788.
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*, 28.
- Roboflow Inc. (2023). Roboflow: Tools for computer vision dataset management and augmentation. <https://roboflow.com>
- Sakeena, M., Stumpe, E., Despotovic, M., Koch, D., & Zeppelzauer, M. (2023). On the robustness and generalization ability of building footprint extraction on the example of SegNet and Mask R-CNN. *Remote Sensing*, 15(8), 2135.

- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 1-48.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR*.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.
- Sun, Z., Zhou, W., Ding, C., & Xia, M. (2022). Multi-resolution transformer network for building and road segmentation of remote sensing image. *ISPRS International Journal of Geo-Information*, 11(3), 165.
- Susetyo, D. B., Harintaka, & Rizaldy, A. (2023). The application of mask R-CNN for building extraction. *AIP Conference Proceedings*, 2941(1), 030001.
- Ultralytics. (2024). YOLOv11 documentation and benchmarking. <https://docs.ultralytics.com>
- Van Etten, A., Lindenbaum, D., & Bacastow, T. M. (2018). SpaceNet: A remote sensing dataset and challenge series. *arXiv:1807.01232*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *NeurIPS*, 30.
- Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024b). YOLOv10: Real-time end-to-end object detection. *arXiv:2405.14458*.
- Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023a). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *CVPR*, 7464-7475.
- Wang, C. Y., Yeh, I. H., & Liao, H. Y. M. (2024a). YOLOv9: Learning what you want to learn using programmable gradient information. *arXiv:2402.13616*.
- Wang, L., Fang, S., Meng, X., & Li, R. (2022). Building extraction with vision transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-11.
- Wang, W., Shi, Y., Zhang, J., Hu, L., Li, S., He, D., & Liu, F. (2023b). Traditional village building extraction based on improved Mask R-CNN: A case study of Beijing, China. *Remote Sensing*, 15(10), 2616.
- Weng, Q., Xu, Y., & Hu, X. (2022). Machine learning in land cover classification with satellite imagery: Advances, challenges, and future directions. *ISPRS JPRS*, 188, 182-198.
- Xia, G., Persello, C., Vivone, G., Chen, K., Yan, Z., Tang, D., Huang, H., Schmitt, M., & Sun, X. (2023). The 2023 IEEE GRSS Data Fusion Contest: Large-scale fine-grained building classification. *IEEE Geoscience and Remote Sensing Magazine*, 11(1).
- Xiao, X., Guo, W., Chen, R., Hui, Y., Wang, J., & Zhao, H. (2022). A Swin Transformer-based encoding booster integrated in u-shaped network for building extraction. *Remote Sensing*, 14(11), 2611.
- Xu, R., Mao, R., Zhuang, Z., Huang, F., & Yang, Y. (2025). Building extraction from remote sensing images based on multi-scale attention gate and enhanced positional information. *PeerJ Computer Science*. <https://doi.org/10.7717/peerj-cs.2826>
- Yuan, W., Zhang, X., Shi, J., & Wang, J. (2023). LiteST-Net: A hybrid model of Lite Swin Transformer and convolution for building extraction from remote sensing image. *Remote Sensing*, 15(8), 1996.
- Yuan, Y., Shi, X., & Gao, J. (2025). Building extraction from remote sensing images with deep learning: A survey on vision techniques. *Computer Vision and Image Understanding*, 251, 104253.
- Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., & Atkinson, P. M. (2016). An object-based approach for urban land cover classification: Integrating LiDAR height and intensity data. *IEEE TGRS*, 54(7), 4168-4182.
- Zhang, J., Wei, H., & Li, Y. (2021). Urban feature detection. *Computers, Environment and Urban Systems*, 90, 101698.
- Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2023). DETRs beat YOLOs on real-time object detection. *arXiv:2304.08069*.
- Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2023). Deep learning in remote sensing: A comprehensive review and list of resources. *Remote Sensing*, 15(14), 3480.