Segment Anything Meets Urban Remote Sensing: Automated Building and Tree Mapping from UAV Imagery

Reedhi Shukla[1]; Sampath Kumar P.[2]; Kamini J.[3]; Narender B.[4]

[1,2,3,4] National Remote Sensing Centre (NRSC), ISRO, Hyderabad

Publication Date: 2026/07/18

Abstract: The Segment Anything Model (SAM) has shown very good generalisation across many image types, but when we applied it directly to drone imagery, we found it does not work well — the top-down geometry, spatial scale, and object shapes in aerial data are quite different from the natural images SAM was trained on. Based on this, we designed SAMD (Segment Anything Model for Drone data), which combines a U-Net with a ResNet-34 encoder and SAM into a single pipeline. The U-Net is fine-tuned on drone imagery and produces pixel-wise masks for buildings and trees; these masks are then used as spatial prompts for SAM, which refines the boundaries and converts everything into georeferenced vector shapefiles. Building heights are derived from DSM-DMT differencing, and the final output is a 3D city model rendered in CesiumJS. We tested SAMD on drone imagery of Kanchipuram city, Tamil Nadu, India. Building extraction reached 85% F1 and tree delineation reached 80% F1. SAMD performed better than standalone SAM on drone data and is also slightly better than U-Net alone on boundary precision.

Keywords: *SAM; U-Net; ResNet-34; Drone Imagery; UAV Data; Building Extraction; Tree Extraction, Urban Mapping.*

How to Cite: Reedhi Shukla; Sampath Kumar P.; Kamini J.; Narender B. (2026) Segment Anything Meets Urban Remote Sensing: Automated Building and Tree Mapping from UAV Imagery. *International Journal of Innovative Science and Research Technology*, 11(7), 293-298. https://doi.org/10.38124/ijisrt/26jul153

I. INTRODUCTION

Drone image data at 2–10 cm resolution provides much more detail than satellite or airborne sensors, and cities like Kanchipuram, which are growing rapidly, need this kind of data to support planning, disaster risk assessment, and environmental monitoring. From drone imagery, building footprints and tree canopy maps are generated, which feed into green-space measurement, solar analysis, and runoff modelling.

Object-based image analysis (OBIA) methods have been the usual approach for this, but they require extensive parameter tuning and do not generalise well across datasets with different illumination or sensor settings. U-Net [1] and related encoder-decoder architectures have significantly improved this by performing pixel-wise classification with much less manual effort.

The Segment Anything Model (SAM) [2] was trained on over one billion masks from 11 million images and can segment objects in images it has never seen, using point, box, or mask prompts. The architecture includes a ViT image encoder that runs once per image and a lightweight mask decoder that takes prompts and produces masks. Since we can run the encoder once and then query it many times with different prompts, it is fast at inference. The problem we found is that SAM's training data is almost entirely natural close-range images, so when we gave it drone images directly, it did not do well at all — the nadir geometry, rooftop textures, and tree crown shapes from above are not what it was trained on.

Since the combination of U-Net for domain-adapted semantic classification and SAM for boundary refinement made sense to us, we built SAMD as a sequential pipeline. The U-Net handles the scene-specific classification that SAM cannot do in zero-shot mode on drone data, and SAM handles the precise boundary work that the U-Net sometimes gets slightly wrong. The result is a complete pipeline from raw drone survey outputs to a 3D city model in the browser.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul153

We tested this on Kanchipuram city — a good test case because of its mix of large temple complexes, densely packed old residential structures, multi-storey commercial buildings, and varied tree canopy — and we report the numbers on each component along with the failure cases we observed.

II. LITERATURE REVIEW

➢ *Building and Vegetation Extraction from Aerial Imagery*

U-Net with skip connections between encoder and decoder that enable the network to localise object boundaries well while still using multi-scale context. Further, pre-trained backbones such as ResNet and EfficientNet were used as encoders, thereby improving feature learning with less training data. Chen et al. *(2022)* used a dual-branch architecture that jointly processes RGB and DSM inputs for building segmentation in dense urban scenes. More recently, transformer-based architectures have pushed building extraction accuracy further on benchmark datasets, though at higher compute cost [3]. For tree crown detection from UAV imagery, both instance- and semantic-segmentation approaches have been tested [4]. Dense South Asian urban vegetation — with many species and overlapping crowns — makes this harder than the forest canopy cases these methods were originally developed for.

➢ *SAM and its Application to Remote Sensing*

SAM was trained in a three-stage loop between human annotators and the model, producing the SA-1B dataset with over one billion masks [2]. When people tried applying SAM directly to remote sensing data, it did not work well. RSPrompter [5] trained a prompt generation network on remote sensing data to produce better SAM prompts. SAMGeo [6] added geospatial coordinate systems and tile-based processing for large satellite images. GeoSAM [7] fine-tuned SAM on geographic imagery. All of these confirm that SAM needs domain adaptation to work properly on remote sensing data, and the best results come from combining it with something that already understands the domain.

➢ *3D Urban Modelling from Drone Data*

3D urban modelling from drones has followed two main paths: photogrammetric reconstruction using Structure-from-Motion to get textured meshes or point clouds, and the simpler approach of combining 2D footprints with DSM-DTM height differencing to get LoD1 building models. The DSM-DTM approach gives polygon-with-height datasets that work directly in CesiumJS, QGIS, and ArcGIS, and can be produced from any drone survey with a standard RGB camera and GNSS [8][9]. We used this approach since it is lighter and more practical than full SfM mesh reconstruction for city-scale coverage.

III. STUDY AREA AND DATASETS

➢ *Study Area*

Kanchipuram city is in Tamil Nadu, India, about 75 km south-west of Chennai, centred at 12.8342° N, 79.7036° E. It is one of the seven sacred cities of India, known as the City of a Thousand Temples, with over 100 active temple complexes. As per the 2011 census, the population is about 165,000, and the city is growing as part of the broader Chennai Metropolitan Region.

The urban sprawl features a mix of ancient temple complexes and walled enclosures, alongside densely packed single-storey residential areas, multi-storey commercial buildings, and a variety of tree canopy — managed street trees, compound-wall vegetation, and temple garden trees. This mix made it a good test case for us, since it is harder than cities with a more uniform building stock.

➢ *Datasets*

High-resolution drone imagery is acquired over a 4.2 km² central area of Kanchipuram using a fixed-wing UAV with a 24-megapixel RGB camera, flown at 120 m above ground level, giving a ground sampling distance (GSD) of about 3.5 cm per pixel. The survey had 80% forward overlap and 70% lateral overlap. Photogrammetric processing produced an ortho-mosaic, a DSM, and a DTM at 5 cm spatial resolution.

For training data we combined local digitised annotations from the Kanchipuram ortho-mosaic with tiles from the ISPRS Benchmark datasets, OpenDroneMap community datasets, and the Inria Aerial Image Labelling Benchmark. After quality control we had about 15,000 annotated image-mask pairs. The training set was split 80% training, 10% validation, 10% test.

IV. METHODOLOGY

The SAMD pipeline has eight sequential stages. It starts with raw drone survey outputs and ends with a georeferenced, height-attribute 3D city model. The three core components are: the U-Net segmentation model, SAM in prompted mode, and the DSM-DTM height estimation module. Supporting geospatial processing uses GDAL, Rasterio, GeoPandas, and Shapely.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul153

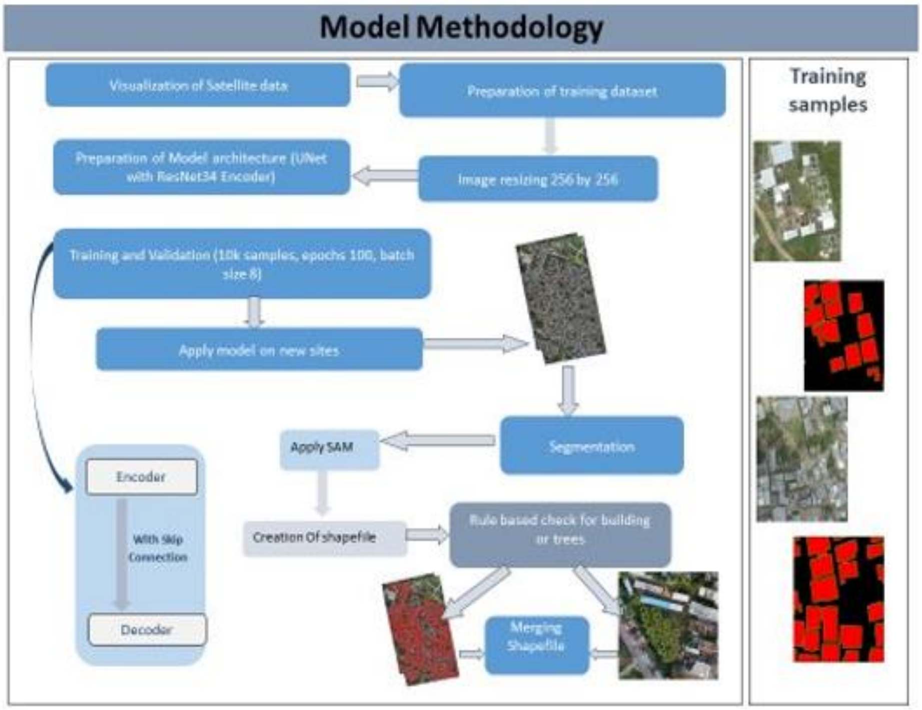

Fig 1. The SAMD Model Methodology Pipeline.

- *Data Pre-Processing*

The ortho-mosaic was tiled into 512×512 pixel patches at 5 cm resolution, with 64-pixel overlap at tile edges to reduce boundary effects during inference. Per-channel mean and standard deviation were computed from the training tiles and used to normalise all tiles to zero mean and unit variance. Data augmentation on training tiles included random flips, 90° rotations, elastic deformations, scale jitter between 0.9 and 1.1, and brightness, contrast, and saturation perturbations.

- *U-Net with ResNet-34 encoder*

We used a Dynamic U-Net with a ResNet-34 encoder pre-trained on ImageNet-1k. The encoder downsamples the 512×512 input using five convolutional blocks, batch normalisation, ReLU, and max pooling, producing feature maps at 256, 128, 64, 32, and 16 pixels. The 16×16 bottleneck captures the global context needed to tell apart large rooftops and dense tree canopy from a spectrally similar background.

We trained with the Fastai library using one-cycle learning rate scheduling with a maximum learning rate of 1×10^{-3}, a batch size of 16, and up to 40 epochs, with early stopping at a patience of 8 based on validation loss. We used a combined Binary Cross-Entropy and Dice loss to handle the class imbalance between building and background pixels.

- *SAM Integration and Vectorisation*

After U-Net inference we run SAM in automatic mask generation mode, a regular grid of point prompts is placed across the image and SAM produces candidate masks for all detected objects. We keep only the SAM masks whose centroids fall inside the U-Net foreground region. This stops SAM from producing spurious masks in areas the U-Net correctly classified as background. The U-Net handles domain-specific semantic classification that SAM cannot perform on drone data in zero-shot mode, and SAM handles boundary refinement that the U-Net sometimes misses. After this, the kept SAM masks are binarised, vectorised using *gdal_polygonize*, and simplified using the Douglas-Peucker algorithm with a tolerance of 0.5 pixels.

- *Height Estimation and 3D Model*

Building heights are estimated from the normalised DSM ($nDSM = DSM - DTM$), which gives above-ground height for every pixel by removing the terrain surface. We compute zonal statistics for each building footprint polygon, using the 95th percentile nDSM value as the building height. Each polygon is then extruded to this height and converted to GlTF format, giving LoD1 building prisms. These are loaded into CesiumJS for browser-based 3D rendering.

- *Tree Extraction*

For trees, we threshold the U-Net tree probability map at 0.5 to identify candidate crown centroids, then use these as point prompts for SAM to generate individual crown masks. Overlapping detections with an IoU above 0.5 are removed by non-maximum suppression. Each surviving crown polygon receives canopy area, an estimated height from the nDSM, and a species likelihood class (broadleaf or palm) based on the crown shape's circularity, elongation, and fractal dimension.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul153

V. RESULTS

➢ *Building Extraction*

We evaluated on the held-out test subset of 1,200 image-mask tile pairs. Table 1 shows the results for all three methods.

Table 1. Building Extraction Results on the Kanchipuram Test Set.

Method	Precision	Recall	F1	IoU
SAMD (proposed)	0.87	0.83	0.85	0.74

SAMD performed with an F1 score of 0.85 and an IoU of 0.74. The standalone U-Net without SAM boundary refinement achieved an F1 score of 0.82 and an IoU of 0.69. Standalone SAM in zero-shot automatic mode achieved an F1 score of 0.51 and an IoU of 0.34 — confirming that SAM on its own does not work on drone data without domain guidance.

Evaluating the predictions, when assessing qualitatively, SAMD identifies most building rooftops for both small and large structures. Two of the most common failure modes, with low-contrast rooftops appearing spectrally similar to road surfaces, are the under-segmentation and the over-segmentation of large temple enclosures. In this case, the partial misclassification of interior courtyards occurs considering them non-building areas. We believe that the lack of training for specific roofs types is likely causing the first problem. The second problem is likely due to the atypical design of temple complexes, which is not well represented in the consolidated training data.

➢ *Tree Extraction*

Tree extraction and mapping were assessed on 300 manually annotated crown polygons covering crown sizes from 5 to 220 m², including neem, tamarind, coconut palm, and banyan. Table 2 shows the results by crown type.

Table 2. Tree Crown Extraction Results.

Crown type	Precision	Recall	F1	Mean IoU
All tree types (overall)	0.78	0.80	0.79	0.68
Broadleaf (neem, tamarind, banyan)	0.82	0.83	0.82	0.71
Palm (coconut)	0.61	0.63	0.62	0.55

The tree module achieved an F1 score of 0.79 and a mean crown IoU of 0.68. Palm trees were detected at F1 score of 0.62, because coconut palms appear as narrow, vertically elongated crowns from above, and the small crown area makes them easy to miss or merge with adjacent crowns. Broadleaf species did much better at F1 score 0.82.

Fig 2. Input and Output of Building Extraction (Green Overlay) Results Over a Dense Mixed-Use Area

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul153

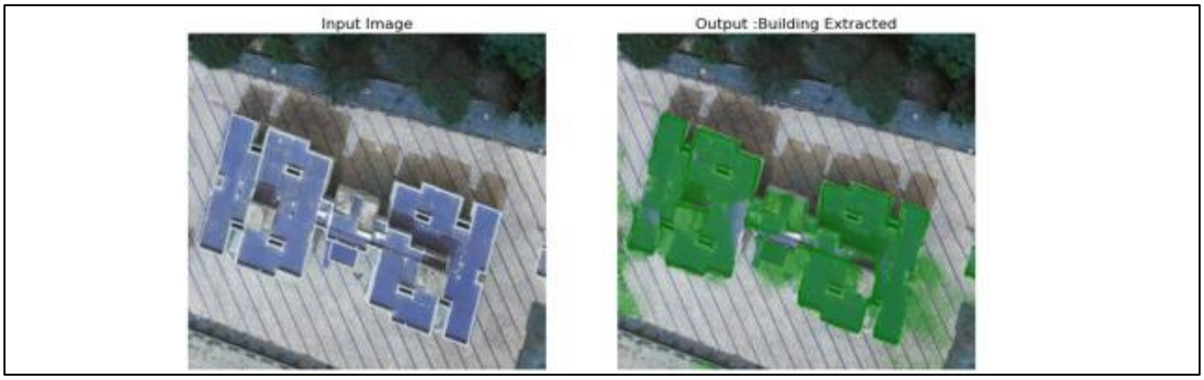

Fig 3. Input and Output of Building Extraction (Green Overlay) Results of an Institutional Complex

Fig 4. Tree Crown Detection Results in a Residential Neighbourhood

➢ *3D City Model*

The Kanchipuram study area includes 14,823 building footprints with heights ranging from 2.8 m to 34.6 m, and 6,241 tree crowns with heights ranging from 3.2 m to 22.7 m. The model correctly shows the tiered height structure of the city — temple gopuram towers are the tallest elements, and the transition from single-storey residential to multi-storey commercial is visible in the height distribution. The 3D model generated is in CesiumJS and runs in a standard browser without any GIS software. Having tree volumes alongside buildings is useful for shadow analysis and green-space measurement from a single dataset.

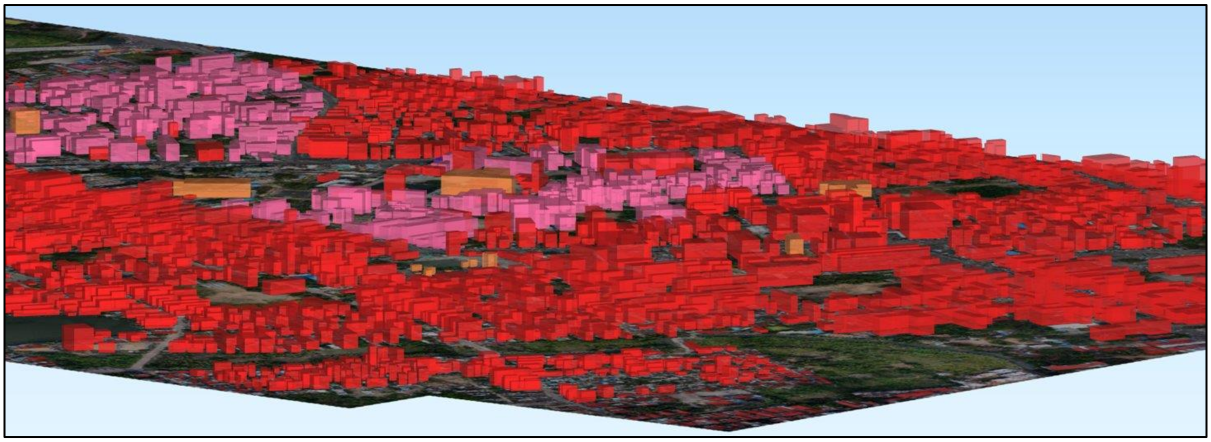

Fig 5. Three-Dimensional LoD1 City Model of Central Kanchipuram Rendered in the CesiumJS Web Platform. Building Footprints are Extruded to their DSM-Derived Heights; Colour Coding Distinguishes

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul153

VI. DISCUSSION

The results show that SAMD performs better than standalone SAM on drone data and is also slightly better than the domain-adapted U-Net on its own. The improvement in F1 from 0.82 to 0.85 when we add SAM boundary refinement is consistent with what similar ensemble approaches have reported in the literature and reflects SAM sharpening boundaries when it receives reliable spatial prompts from the U-Net rather than having to work from scratch.

Since SAM handles boundary refinement, we did not need to train entirely from scratch. We achieved good results with about 12,000 annotated image-mask pairs, which is much fewer than fully supervised approaches usually require. This matters for urban mapping projects in developing countries where annotation time is limited.

Some limitations are worth noting. The U-Net needs fine-tuning on annotated data from each new city; if the building typologies or roof materials are very different from Kanchipuram, performance will drop. The LoD1 model does not capture roof geometry detail, which limits its use for analyses such as solar suitability. Tree species classification from RGB alone is limited, and adding near-infrared or multispectral bands would help substantially. On the compute side, SAM's ViT image encoder requires about 16 GB of GPU memory for 512×512 patches; the smaller ViT-B and ViT-L variants use less memory at some boundary-precision cost, and we have not yet systematically compared these within SAMD.

VII. CONCLUSION

We have built SAMD, a pipeline that combines a domain-adapted U-Net with SAM to extract building footprints and tree canopies from high-resolution drone imagery, and extends this to a 3D city model using DSM-DTM height differencing and CesiumJS rendering. On Kanchipuram drone imagery the system reaches 85% F1 for buildings and 80% F1 for trees, and is clearly better than zero-shot SAM on drone data.

The three things we contributed here are: first, the SAMD architecture that uses U-Net masks as prompts for SAM, which gives us domain-specific accuracy from the U-Net and foundation-model boundary quality from SAM; second, an end-to-end pipeline from drone data to a 3D model in the browser using only open-source tools; and third, results on a morphologically complex Indian city covering a range of building types and vegetation structures.

Further work will focus on multispectral drone data for improved vegetation analysis, adding road and water body extraction, upgrading from LoD1 to LoD2 building models using the SfM point cloud, and testing on drone surveys of other South Asian cities to assess how well the system transfers.

REFERENCES

[1]. Ronneberger, O., Fischer, P. and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. MICCAI 2015, LNCS 9351, 234–241.

[2]. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026)..

[3]. Song, J., Zhu, A.-X., & Zhu, Y. (2023). Transformer-Based Semantic Segmentation for Extraction of Building Footprints from Very-High-Resolution Images. Sensors, 23(11), 5166. https://doi.org/10.3390/s23115166.

[4]. Weinstein, B. G., Marconi, S., Bohlman, S., Zare, A., & White, E. (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.

[5]. Keyan Chen , Chenyang Liu , Hao Chen , Haotian Zhang , Wenyuan Li , Zhengxia Zou ,Zhenwei Shi (2024), "RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model," in IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-17, 2024, doi: 10.1109/TGRS.2024.3356074.

[6]. Wu, Q., & Osco, L. (2023). samgeo: A Python package for segmenting geospatial data with the Segment Anything Model (SAM). Journal of Open Source Software, 8(89), 5663. https://doi.org/10.21105/joss.05663.

[7]. Zhao, X., et al. (2023). GeoSAM: Fine-tuning SAM for geographic remote sensing image segmentation. arXiv:2308.09634.

[8]. Gerke, M. and Nex, F. (2020). 3D building reconstruction from UAV images. In: Handbook of UAV Applications (Springer).

[9]. Bittner, K., et al. (2018). Building footprint extraction from VHR remote sensing images combined with normalized DSMs. IEEE JSTARS, 11(8), 2615–2629.