# Habitation Boundary Extraction using Geospatial Artificial Intelligence on HPC

Prakhar Verma*, Sivakumar V, Vivek Singh Tomar, Soham Rangdal, Kedar N. Ghogale, Biju C, Jitendra Mhatre, Sajeevan G

Centre for Development of Advanced Computing (C-DAC), Pune, Maharashtra, India
*prakharv@cdac.in

**KEY WORDS:** GeoAI, HPC, Deep Learning, AI/ML, Semantic Segmentation, Habitation

## Abstract

Artificial Intelligence (AI), High-Performance Computing (HPC) and high-resolution satellite imagery have collectively enhanced the capability to extract meaningful geospatial information quickly and accurately. In this paper, we have proposed an automated workflow for extracting habitation boundary using Deep Learning (DL). The workflow is built on deep learning-based semantic segmentation, augmented by an adaptive contour-refinement loss, optimized to detect small and scattered habitation from satellite images. The pipeline incorporates Spatial Pyramid Dilation Convolution (SPDConv) and Effective Squeeze-and-Excitation (EffectiveSE) in the backbone, combined with a boundary-aware hybrid loss (Dice, weighted BCE, L1 boundary loss). The model was trained and validated on C-DAC PARAM Siddhi-AI HPC system using 6274 samples. The satellite image was sourced from ISRO Bhuvan and the annotations were created on QGIS. Both the satellite image and its corresponding raster annotation were tiled with resolution of 640 x 640 pixels. The training workload was distributed using PyTorch’s Distributed Data-Parallel (DDP) framework, enabling efficient scaling across multiple GPUs and optimizing experimentation workflow on large datasets. The final model attained 0.757 precision, 0.669 recall, and 0.7 Intersection-over-Union (IoU) on the validation set, indicating reliable performance in extracting habitation boundaries. The predicted segmentation masks are post-processed to create geospatial polygons of habitation boundaries. The output can be seamlessly integrated with national geospatial programmes like PMGSY National GIS.

## 1. Introduction

Artificial Intelligence (AI) and Deep Learning (DL) have reformed geospatial analysis by transforming traditional workflows into highly efficient and automated processes. GeoAI has been emerging in the field of spatial technology for rural and urban planning and development. GeoAI combines geospatial and AI technologies for modeling, prediction, and information extraction. Tasks that previously required extensive manual effort such as land use/ land cover classification and building footprint extraction can now be completed with remarkable speed and accuracy. This paradigm shift has been driven by the convergence of High-Performance Computing (HPC) and Convolutional Neural Networks (CNNs), enabling the rapid processing of high-resolution satellite imagery (Ronneberger et al., 2015; Chen et al., 2018).

Traditionally, identifying and mapping habitations involved deploying survey teams to the field, collecting GPS coordinates, and manually digitizing boundaries using GIS software. This workflow is slow, costly, and prone to inconsistency (Zhang et al., 2016). Our study extracts rural habitation (dwelling) clusters to support GIS initiatives like PMGSY National GIS by providing updated habitation map that help planning and prioritizing road connectivity.

## 2. Related Work

In the past, habitation mapping relied on manual digitization, spectral indices, and conventional machine learning models. Classical approaches using support vector machines (SVM), random forests (RF), and decision trees on spectral and texture features produced effective results but lacked generalization (Breiman, 2001; Pal and Mather, 2003). With the advent of DL, CNN-based models such as UNet, DeepLabV3+, and SegNet (Badrinarayanan et al., 2017) brought notable improvements in segmentation accuracy. UNet is widely adapted for both biomedical image and remote sensing mask generation tasks due to its encoder-decoder design and skip connections that preserve spatial information.

Real-time object detection became feasible with the advent of the YOLO (You Only Look Once) family. Initially, YOLOv3 established a practical speed, accuracy trade-off, and YOLOv4 and YOLOv5 pushed both metrics further through architectural refinements and enhanced training strategies (Jocher et al., 2021; Redmon and Farhadi, 2018; Bochkovskiy et al., 2020). Since these models produce bounding boxes only, they lack the spatial details required for precise habitation mapping. To address this limitation YOLOv5-Seg extended the YOLOv5 backbone with segmentation heads to generate pixel-level masks (Jocher et al., 2022).

Oktay et al. (2018) demonstrated Attention U-Net that integrates Attention gates into the U-Net skip connections to focus on relevant regions while suppressing background noise. Each gate uses coarse contextual features to generate a soft mask via sigmoid activation and multiplies it element-wise with incoming feature maps. The model thus learns where to attend during training and refines its focus at inference without any extra localization network or cascaded stages. Chen et al. (2022) introduced TransUNet, which combines a convolutional encoder with a Vision Transformer to capture both local detail and global context. First, CNN layers extract multi-resolution feature maps, which are tokenized and passed through Transformer blocks. A lightweight decoder then merges the Transformer’s global embeddings with high-resolution CNN features via skip connections, preserving fine spatial cues while modeling long-range dependencies in a single end-to-end framework. Xie et al. (2021) proposed SegFormer, employing a hierarchical Transformer encoder without positional embeddings alongside a simple multi-layer perceptron decoder. The encoder applies overlapped patch merging and efficient self-attention to produce multiscale features. The decoder fuses those features via MLP layers to yield final segmentation maps. SegFormer models run faster and use fewer parameters while matching or exceeding accuracy benchmarks.

Guo and Chen (2017) used an ensemble CNN to identify village buildings but did not achieve real-time performance. Anilkumar et al., (2023) applied an adaptive DeepLabv3+ model, tuned with

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-5-W2-2025-707-2025 | © Author(s) 2025. CC BY 4.0 License.

an Improved Golden Eagle optimisation algorithm, to segment building footprints in aerial imagery. Chen et al. (2019) employed a Mask R-CNN framework to delineate urban village polygons at city scale, yet this study also centered on metropolitan area. Most recently, Li et al. (2024) proposed UV-Mamba, a deformable-convolution-enhanced state-space model that achieves state-of-the-art accuracy while remaining lightweight and real-time, making it highly relevant for large-area habitation boundary mapping. However, these studies focus on urban areas, leaving rural habitation clusters under-explored.

## 3. Dataset

To construct a robust training dataset for the habitation-segmentation model, we sourced imagery from the ISRO Bhuvan. QGIS was used for annotation of habitation footprints, which were then rasterized to the native pixel grid and tiled, together with their source imagery, into 640 × 640-pixel patches with a 128-pixel overlaps. The dataset spans various terrains of plains, hilly regions, covering diverse geographic rural areas of Madhya Pradesh, Maharashtra and Assam. Dataset collected in various illumination, weather conditions which strengths the model training in diverse conditions. The resulting dataset was randomized and allocated 70% of the samples to training, 20% to validation, and 10% to testing. The augmentation techniques which we included, (i) geometric transformation such as horizontal and vertical flip, rotation (±90°), and scaling; (ii) photometric changes like brightness and contrast shift along with hue and saturation jitter; and (iii) spatial alterations including erasing, mosaic combination, and cutout augmentation.

## 4. Methodology

### 4.1 Model Architecture

We have adapted habitation extraction model that extends the standard YOLO11x-seg backbone by integrating Spatial Pyramid Dilation Convolution (SPDConv) and Effective Squeeze and Excitation (EffectiveSE) (Zhu et al., 2025) which enhance the capability of capturing irregular, multi-scale patterns of rural habitations while minimizing background noise. We have integrated additional two layers in YOLO11x-seg model. The key advantages of the adapted model are highlighted in the subsequent sub sections:

#### 4.1.1 Backbone Enhancements

Backbone network level, we retained the core YOLOv11x-seg architecture to extract multiscale features using C3K2 and SPPF modules, to allowing different resolution. To enhance contextual point of view, we inserted SPDConv blocks immediately after each 3×3 stride-2 convolution at P3, P4, and P5 levels, which adapts dilation to feature-map size, capturing both fine details (small habitations) and broad context (large clusters) without extra parameters. Additionally, we applied the EffectiveSE on the deepest 1024-channel map (P5), before the SPPF block. It uses global pooling and re-weight channels to recalibrate feature responses by this enhancing habitation prediction and reducing background noise

#### 4.1.2 Neck and Feature Fusion

The neck and feature fusion component retains the original YOLOv11x-seg structure, employing a top-down Feature Pyramid Network (FPN) followed by a bottom-up Path Aggregation Network (PAN) for effective multi-scale feature fusion. The SPDConv-enriched and EffectiveSE-recalibrated feature maps are seamlessly integrated into this FPN/PAN, enhancing semantic consistency and localization across scales without adding any additional fusion layers.

#### 4.1.3 Segmentation Head

We have maintained the segmentation head its original architecture and utilized a single multi-scale head that concatenates feature maps from P3, P4, and P5, followed by a 1×1 bottleneck convolution to generate per-class masks. It directly leverages the enhanced feature maps for more precise mask predictions, without altering head complexity.

#### 4.1.4 Boundary-aware Hybrid Loss

We used a hybrid loss function which comprises of Dice loss, a weighted background Binary Cross Entropy (BCE), and an L1 boundary loss. Each component addresses a specific challenge in the task of habitation extraction from high-resolution remote sensing imagery.

**Dice loss** focuses on optimizing the spatial overlap between predicted masks and ground truth annotations. In habitation extraction, where the target regions often occupy a small portion of the image, pixel-wise losses tend to be biased toward the background. Dice loss helps counter this imbalance by directly maximizing the intersection over union, ensuring that small and scattered habitation clusters are not overlooked.

**Weighted background BCE** addresses the tendency of the model to falsely activate on non-habitation regions. Satellite images often include background patterns that resemble built-up areas, such as bare soil, rocky terrain, or dry vegetation. By penalizing misclassified background pixels more heavily, this term improves the model’s ability to differentiate true habitation areas from confusing background textures, enhancing overall precision.

**L1 Boundary loss** is included to improve the sharpness and accuracy of predicted boundaries. Precise edge delineation is critical in geospatial applications, where small shifts in polygon boundaries can translate into significant errors in mapped footprints. The L1 term reduces pixel-wise discrepancies along object contours, producing cleaner and more reliable outlines of habitation structures.

In combination, these losses reinforce each other: Dice improves region-wise matching, weighted BCE minimizes false positives, and L1 sharpens object borders. This synergy allows the model to produce more accurate, interpretable, and geospatially consistent habitation boundaries.

### 4.2 Training Setup

The C-DAC PARAM Siddhi-AI HPC system was utilized for training the model which facilitates multi-GPU. The workload was distributed using PyTorch’s Distributed Data-Parallel (DDP) framework, which enabled efficient training over large datasets and rapid experimentation. Below is the HPC training setup:

- **Software stack:** Ultralytics segmentation framework (v8.3.133), Python 3.10, PyTorch 2.6.0+cu118
- **Hardware**: PARAM Siddhi-AI NVIDIA A100-SXM4 (4 x GPU, 40 GB VRAM each)
- **Batch size**: 8
- **Epochs**: 200
- **Optimizer**: AdamW, with a starting learning rate 0.002, weight decay 0.0005, cosine annealing schedule

- **Input Size**: Random cropping to 640×640, with on-the-fly augmentations
- **Loss Function**: Composite loss combining binary cross-entropy for pixel classification, Dice loss for mask overlap, and Distribution Focal Loss (DFL) for boundary refinement

### 4.3 Postprocessing

Automated postprocessing involved converting raw segmentation masks into clean, simplified polygons for habitation clusters that comprises the following steps:

- Mask edges are traced into polygon contours to outline each cluster accurately
- Small or noisy contours like shadows or vegetation, are filtered out based on size and length thresholds
- The polygon shapes are smoothed to reduce unnecessary detail while keeping the main structure
- The processed polygons are converted to GIS vector file format

The post processing steps are essential for taking the predicted pixel based habitation boundaries into GIS environment.

## 5. Results

Our model achieved its best validation performance with an Intersection-over-Union (IoU) of 0.7, precision of 0.757, recall of 0.669, F1-score of 0.71, and the mean Average Precision at an IoU threshold of 0.5 (mAP@0.5) was 0.671. These metrics show that the segmentation network captures the majority of true-habitation pixels while keeping false positives to a minimum, producing reliable outlines for rural planners. Table 1 shows comparative study of YOLO11x-seg baseline with improved YOLO.

| Metric | YOLO-Seg | Our Model |
|---|---|---|
| Bounding-box Precision (P) | 0.725 | 0.793 |
| Bounding-box Recall (R) | 0.649 | 0.679 |
| mAP50 (BBox) | 0.743 | 0.713 |
| mAP50–95 (BBox) | 0.531 | 0.466 |
| Mask Precision (P) | 0.593 | 0.757 |
| Mask Recall (R) | 0.597 | 0.669 |
| mAP50 (Mask) | 0.548 | 0.671 |
| mAP50–95 (Mask) | 0.318 | 0.38 |
| Peak F1 (Mask) | 0.595 | 0.71 |

Table 1: Comparative summary of YOLO-Seg and Our model

### 5.1 Quantitative Evaluation

The quantitative evaluation plots are described below:

- **Precision versus Confidence threshold**: This plot illustrates how the model's precision changes as the detection confidence threshold increases from 0.0 to 1.0, for the habitation class (thin orange curve) and for all classes combined (thick blue curve). At the lowest thresholds, precision is low because even very uncertain predictions are accepted. As the threshold rises, low-confidence detections are filtered out and precision improves steadily. The curve reaches its maximum value of 1.00 at a threshold of 0.925, indicating that beyond this point every remaining detection is a true positive.

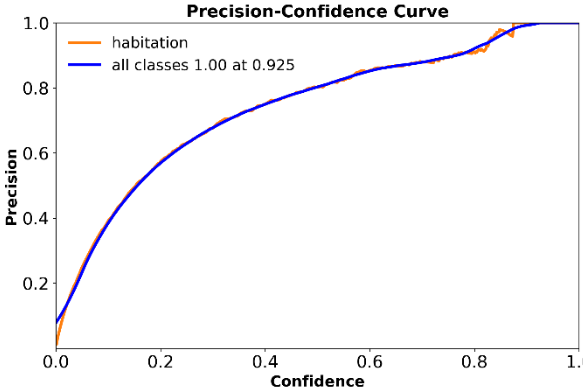


Figure 1a: Precision–confidence curve

The close alignment of the orange and blue curves shows that the habitation class precision closely tracks the overall model precision. In practice, one would rarely operate at such a high threshold because it sacrifices recall; instead, the optimal balance between precision and recall occurs at the F1-maximizing threshold of 0.406 (see Figure 1c), where precision is around 0.71. Selecting a threshold above this value can be justified in applications where false positives carry a high cost, whereas lower thresholds may be preferred when it is more important to capture every instance of habitation.

- **Recall–Confidence Curve**: This plot shows (Figure 1b) how the model's recall varies as the detection confidence threshold is increased from 0.0 to 1.0 for the habitation class (thin orange curve) and for all classes combined (thick blue curve).

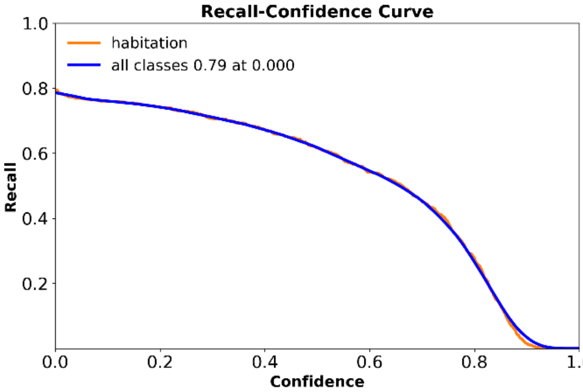


Figure 1b: Recall–confidence curve

The maximum recall of 0.79 occurs at the lowest threshold, since every predicted box is accepted regardless of confidence. As the threshold rises, the model rejects low-confidence detections and recall declines smoothly, falling below 0.40 once the threshold exceeds 0.80. The near-perfect overlap of the two curves indicates that the habitation class recall mirrors the overall model recall. In practice, choosing a higher threshold will reduce false positives but at the cost of missing true habitation areas. For applications where overlooking any habitation is critical, one might select a threshold closer to 0.0 or to the F1-optimal value of 0.406 (which yields a recall of approximately 0.71).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-5-W2-2025-707-2025 | © Author(s) 2025. CC BY 4.0 License.

- **F1–Confidence Curve**: This plot shows (Figure 1c) how the model's F1 score varies with the detection confidence threshold for the habitation class (thin orange curve) and for all classes combined (thick blue curve).

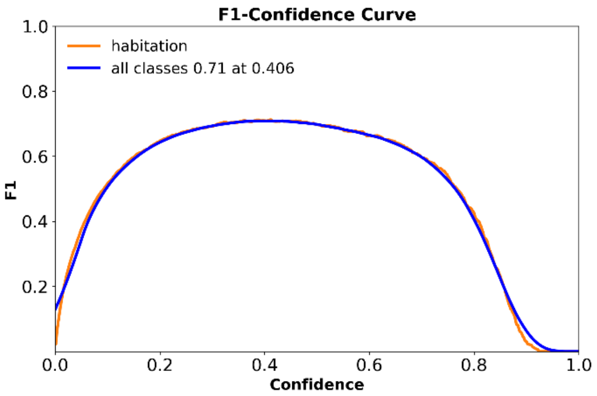


Figure 1c: F1–confidence curve

As the threshold increases from 0.0 to 1.0, the F1 score initially rises, reaching its maximum of 0.71 at a threshold of 0.406. Beyond this point, the score declines because further increases in confidence reduce recall faster than they improve precision. The close overlap of the two curves indicates that the model's overall detection performance is representative of its performance on the habitation class alone. Selecting the threshold at 0.406 therefore provides the best balance between false positives and false negatives for this application.

- **Precision–Recall Curve**: This plot (Figure 1d) shows the trade-off between precision and recall for the habitation class (thin orange curve) and for all classes combined (thick blue curve).

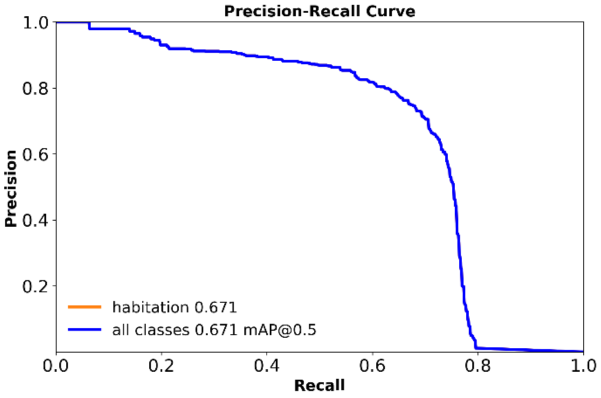


Figure 1d: Precision–recall (PR) curve

Precision is highest at low recall, since only the most confident detections are accepted, and then declines as recall increases and lower-confidence predictions are included. The area under the curve, computed at an IoU threshold of 0.5, is 0.671 (mAP@0.5), summarizing the model's overall detection quality across all recall levels. The close alignment of the two curves indicates that performance on the habitation class closely matches the aggregate performance. In practice, one would select a point along this curve that balances the cost of missed detections against that of false alarms. For example, the F1-optimal threshold of 0.406 (Figure 1c) yields a precision and recall of approximately 0.71, offering a balanced operating point.

### 5.2 Confusion Matrix Analysis

Normalized confusion matrix as shown in Figure 2, illustrates how accurately the model performed during validation across each class. Model correctly labels 76% of actual habitation pixel and 24% misclassified label of habitation.

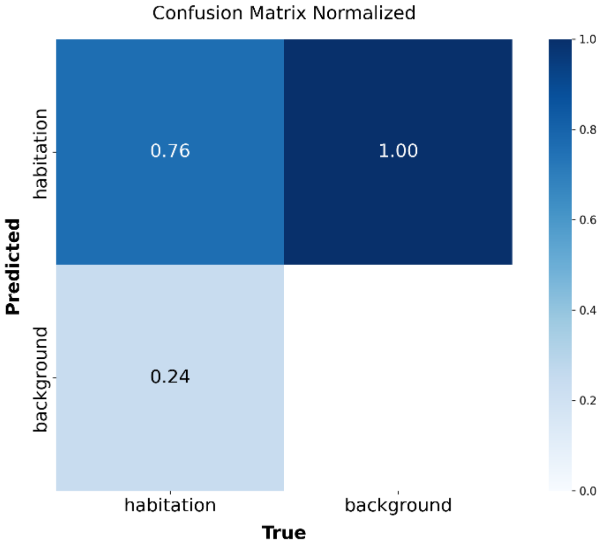
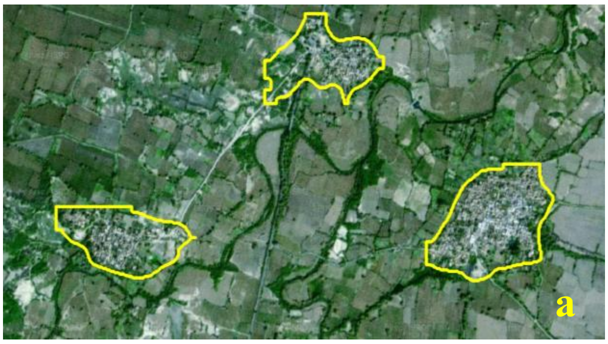


Figure 2: Normalized confusion matrix

### 5.3 Qualitative Assessment

Predicted output shown in Figure 3, we have overlaid predicted habitation clusters (yellow) on the satellite image. Each polygon closely matches the true extent of habitation areas, even under tree cover and varying light conditions. Confidence scores range from approximately 0.35 (under low illumination) to 0.75 (under high illumination), demonstrating that the model maintains reliability across different terrains and illumination conditions. Figure 3(a,b,c) indicates prediction in different illumination and geography. In summary, our segmentation model produces accurate rural habitation clusters and strong quantitative scores (IoU = 0.7, F1 = 0.71, mAP@0.5 = 0.671). These results indicate that our model is well suited for automated mapping of rural habitations.



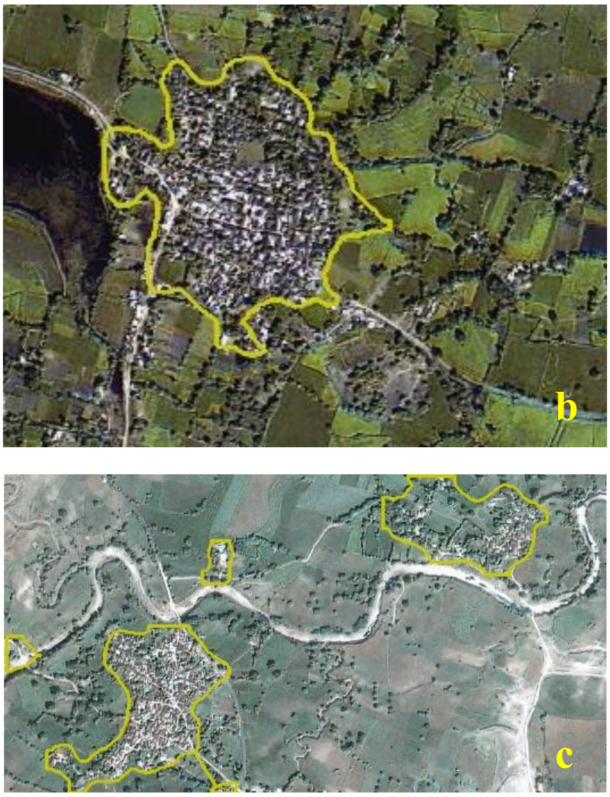


Figure 3 (a,b,c): Detected habitation boundaries highlighted in yellow color

## 6. Conclusions

We have demonstrated a semantic segmentation pipeline that reliably extracts rural habitation boundaries from high-resolution satellite imagery. Our final model achieved an IoU of 0.7, precision 0.757, recall 0.669, F1 score 0.71, and mAP@0.5 0.671. The model achieved high accuracy by balancing detection performance to minimize both missed habitations (false negatives) and incorrect detections (false positives). It also ensures scalability through HPC, using multi-GPU training on A100s to cut down model development time from weeks to hours. The proposed approach differs from prior YOLO-based segmentation methods by integrating Spatial Pyramid Dilation Convolution and Effective Squeeze-and-Excitation modules, enabling multi-scale feature representation and background suppression. The use of a boundary-aware hybrid loss further refines polygon edges and improves the detection of small, scattered habitations that are often missed in conventional models. The model is suitable for automated mapping of rural habitations which may be used in Geospatial applications such as PMGSY planning workflows.

## Acknowledgements

Authors gratefully acknowledge NRSC, ISRO for sharing Bhuvan satellite imagery. Authors also thank C-DAC for funding this R&D and providing access to PARAM Siddhi-AI HPC facilities.

## References

Anilkumar, P., Venugopal, P., Maddikunta, P.K.R., Gadekallu, T.R., Al-Rasheed, A., Abbas, M., Soufiene, B.O., 2023: An Adaptive DeepLabv3+ for Semantic Segmentation of Aerial Images Using Improved Golden Eagle Optimization Algorithm. *IEEE Access*, 11, 106688–106705. https://doi.org/10.1109/ACCESS.2023.3318867.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017: SegNet: A Deep Convolutional Encoder–Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020: YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint*. https://arxiv.org/abs/2004.10934.

Breiman, L., 2001: Random Forests. *Mach. Learn.*, 45(1), 5–32.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021: TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint*. https://doi.org/10.48550/arXiv.2102.04306.

Chen, L., Xie, T., Wang, X., Wang, C., 2019: Identifying Urban Villages from City-Wide Satellite Imagery Leveraging Mask R-CNN. In: *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. & ACM Int. Symp. Wearable Comput.*, Adjunct, pp. 29–32.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: *ECCV 2018*, pp. 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.

Guo, Z., Chen, Q., 2017: Village Building Identification Based on Ensemble Convolutional Neural Networks. *Sensors*, 17(11), 2487. https://doi.org/10.3390/s17112487.

Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., et al., 2021: YOLOv5 v6.0 – New Models, Roboflow & More. *Ultralytics Blog*, 12 Oct 2021. https://www.ultralytics.com/blog/yolov5-v6-0-is-here. Accessed on 18 July 2025.

Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., et al., 2022: Introducing Instance Segmentation in Ultralytics YOLOv5 v7.0. *Ultralytics Blog*, 23 Nov 2022. https://www.ultralytics.com/blog/introducing-instance-segmentation-in-yolov5-v7-0. Accessed on 18 July 2025.

Li, L., Chen, B., Zou, X., Xing, J., Tao, P., 2024: UV-Mamba: A DCN-Enhanced State Space Model for Urban Village Boundary Identification in High-Resolution Remote Sensing Images. *arXiv preprint*. https://doi.org/10.48550/arXiv.2409.03431.

Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018: Attention U-Net: Learning Where to Look for the Pancreas. In: *Proc. MIDL 2018*, Amsterdam, pp. 1–11.

Pal, M., Mather, P.M., 2003: An Assessment of the Effectiveness of Decision Tree Methods for Land Cover Classification. *Remote Sens. Environ.*, 86(4), 554–565.

Redmon, J., Farhadi, A., 2018: YOLOv3: An Incremental Improvement. *arXiv preprint*. https://doi.org/10.48550/arXiv.1804.02767.

Ronneberger, O., Fischer, P., Brox, T., 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *MICCAI 2015*, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.*, 34, 12077–12090.

Zhang, L., Zhang, L., Du, B., 2016: Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.*, 4(2), 22–40.

Zhu, W., Han, X., Zhang, K., Lin, S., Jin, J., 2025: Application of YOLO11 Model with Spatial Pyramid Dilation Convolution (SPD-Conv) and Effective Squeeze-Excitation (EffectiveSE) Fusion in Rail Track Defect Detection. *Sensors*, 25(8), 2371. https://doi.org/10.3390/s25082371.