Semantic Segmentation of Shadow Casting Objects

Sai Mukkundan Ramamoorthy, Shrikar Nakhye, Kai Bastian Glasenapp

September 15, 2025

H-BRS DLRV Project

Hackathon Bonn 2025 - AI and Shadow

- Organized by the City of Bonn to promote AI for urban planning
- Citizens, students, and experts worked together to label trees in aerial images
- Created a public dataset with about 1700 images
- Goal: Train AI models to create a mask of objects that cast shadows across the city

Aerial Images of Bonn

- High-resolution aerial images captured every two years by the City of Bonn
- Based on CIR (Color Infrared) True Ortho TIFFs from 2024
- Captured by Aerowest using an IGI UrbanMapper-2 camera on a CESSNA 404 aircraft
- CIR includes conventional RGB channels and a near-infrared channel (700–850 nm)
- True Ortho eliminates distortions — all objects appear from a vertical perspective
- Georeferenced with `.tfw` Worldfiles, enabling precise map placement
- Resolution: **2.5 cm per pixel**, Image size: **10,000 × 10,000 px** (250 × 250 m)
- Total dataset: **2736 images**, requiring over 1.1 TB of storage

Annotation in Makesense.ai and Augmentation

- Lowered resolution to 10cm per pixel and sliced images to 500 by 500 pixels
- Makesense.ai: Web-based tool to annotate Data without account
- Single Class: tree
- Augmentation via Rotation, Flipping and Brightness/Contrast Hue/Saturation/Value Shift
- Data augmentation increases dataset to about 1700 images
- Split into training and validation subsets (1200/500)

Experimental Setup

- **Early stopping** is applied to prevent overfitting and determine the optimal number of epochs based on validation loss.
- **Identical dataset** is used for training, validation, and testing across all experiments.
- **Hardware consistency**: All models are trained on the same machine with an NVIDIA RTX 4070 Ti SUPER GPU.
- Hyperparameters such as optimizer, learning rate, and batch size are kept consistent where the model architecture allows.

Evaluation Metrics

Below metrics are used to compare model performance in terms of computational utilisation and accuracy in segmenting the objects of interest.

Training and Inference Time
To show if the model is computationally efficient to train in consumer-grade gpu and whether it can run in real time without any performance bottlenecks.

GPU Memory Utilisation
To visualise the memory utilisation by the model i.e which captures how computationally intensive to run a model with a limited VRAM.

F1 Score
F1 score is the harmonic mean of precision and recall and is commonly used in semantic segmentation to evaluate the balance (bias) between false positives and false negatives.

Mean Intersection over Union (mIoU)
To check the overlap of the regions between the ground truth and the predicted mask in percentage.

Mask R-CNN with ResNet Based backbone for Semantic Segmentation of Aerial Images

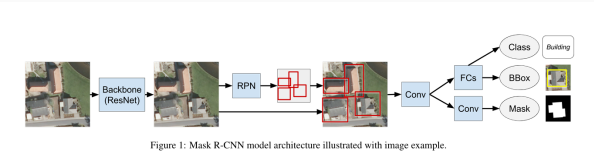


Figure 1: Mask R-CNN model architecture illustrated with image example.

- **Mask R-CNN:** Based on the Faster R-CNN architecture with the ability to add pixel masks for the objects.
- **Architecture:**
  - *Backbone:* ResNet-50 / ResNet 101 based model to generate feature maps which utilises skip connections between convolutional layers.
  - *RPN:* Proposes a regions of interests(ROI) by constructing feature pyramid.
  - *Head Units:* Three Head units (Class Head, BBox Head, Mask Head) takes ROI as input and gives labels, bounding box and binary mask for the objects.
- **Input:** RGB images resized to **512×512 px**.
- **Output:** Binary segmentation mask with bounding box and label for the class **tree**.
- **Training Setup:**
  - Augmentation: rotation, flipping, brightness/contrast and color shifts.
  - Dataset size: **1700 images** → **1200 train / 500 val**.
  - Mask Head Loss function: Cross-Entropy or Squared Hinge Loss.
- **Application:** Tree detection & segmentation in high-resolution urban aerial imagery.

YOLOv8m-seg

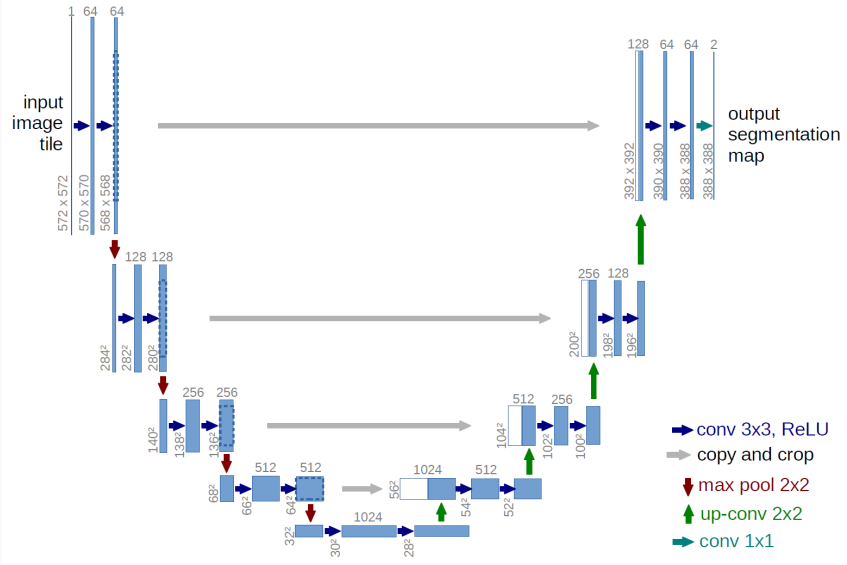
- Based on **CNN** backbone with Cross Stage Partial connections (CSP) for efficient feature reuse
- Utilizes a **decoupled head** for separate object classification, bounding box regression, and mask prediction
- Fully convolutional design — no anchor boxes required, improves generalization
- **Training**: Supervised with annotated masks, optimized via composite loss:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$$

- Hackathon Results: Training (50 Epochs) - Inference Time on Dataset and 50 Epochs: 650s and 1866s

U-Net for Semantic Segmentation of Aerial RGB Images

- **U-Net:** Fully convolutional network designed for precise pixel-level segmentation.



- **Architecture:**
  - *Encoder:* Extracts features through convolution and downsampling.
  - *Decoder:* Upsamples features to original resolution.
  - *Skip Connections:* Combine low-level and high-level features for accurate localization.

Future Work and Learn More

- Implement all Metrics and Train the other Models on the Dataset
- `https://github.com/ItsShriks/Shadow_Casting_Object_Segmentation`
- `https://github.com/MrZinken/Hackathon-Bonn`