Comparative Analysis of Deep Learning Architectures for Aerial Image Segmentation

Evaluating CNN- and U-Net-based architectures
for accurate semantic segmentation of aerial imagery.

Mateusz Mazur

Faculty of Electrical Engineering, Automation, Computer Science and Biomedical Engineering, AGH
Field of study: Computer Science and Intelligent Systems
Specialization: Artificial Intelligence and Data Analysis

November 19, 2025

Table of contents

Introduction

Motivation

Accurate aerial image segmentation is vital for up-to-date mapping and urban analysis, yet manual methods are slow and costly. Deep learning offers scalable solutions, but architectures differ in handling challenges like shadows, small objects, and class variability. By comparing key models, this project seeks the most effective approach for reliable segmentation—directly supporting my geospatial research, where high-quality segmented imagery is essential for deeper analysis and better results.

Project overview

This project focuses on **aerial image segmentation**, a key task in remote sensing with applications in mapping, urban planning, and environmental monitoring. The objective is to **compare the performance of different deep learning architectures**, ranging from classic convolutional neural networks (CNNs) to advanced models such as **U-Net** and U-Net–derived architectures (e.g, MultiRes-UNet, and CSE-UNet). Using publicly available datasets like **iSAID** or the **Dubai Aerial Imagery dataset**, the project evaluates model performance to identify the most reliable and efficient one.

Literature overview

- **Learning Aerial Image Segmentation From Online Maps** [2]
  Shows that **CNNs** can learn to segment aerial images using noisy labels from online maps like OpenStreetMap. The study demonstrates that large, imperfect datasets can reduce manual annotation needs while maintaining strong performance.
- **A Context and Semantic Enhanced UNet for Semantic Segmentation of High-Resolution Aerial Imagery** [3]
  Introduces **CSE-UNet**, which enhances segmentation in high-resolution aerial imagery using multi-level receptive field blocks and a dual-path encoder. The model effectively addresses intra-class heterogeneity and inter-class homogeneity, outperforming UNet and other baselines.
- **Integrating Semantic Edges and Segmentation Information for Building Extraction from Aerial Images Using UNet** [1]
  Proposes **MultiRes-UNet**, an improved model for building extraction that integrates multi-scale feature learning and semantic edge information. Results show superior boundary accuracy compared to UNet, DeeplabV3, and ResNet.

Related datasets

To evaluate and compare different segmentation architectures, two benchmark datasets are considered:

1. Semantic segmentation dataset – The Humans in the Loop (Kaggle)

   The Humans in the Loop dataset contains aerial images of Dubai annotated with pixel-wise semantic segmentation across six classes. It is publicly available under a CC0 1.0 license, making it free for use in research and analysis.

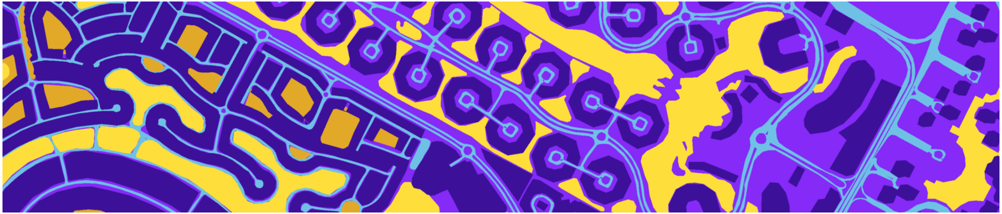

Figure 1: Sample mask for the dataset

1. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images [4] This paper introduces iSAID, the first large-scale benchmark dataset for instance segmentation in aerial imagery, featuring 655,451 objects across 15 categories. It combines object detection and pixel-level segmentation, addressing challenges like dense scenes and tiny objects. Results show that standard models like Mask R-CNN and PANet perform suboptimally, highlighting the need for specialized methods for aerial images.

Figure 2: Sample images from the iSAID dataset

Theory Overview

Learning Aerial Image Segmentation From Online Maps [2]

Using Addapted FCNs to Leverage Noisy Crowdsourced Data

Technology: Adapted Fully Convolutional Network (FCN)

- The study utilized a variant of the FCN architecture, which performs pixel-to-pixel classification, returning a structured spatially explicit label image.
- **Adaptation:** The FCN variant introduced a **third skip connection** (in addition to the original two) to preserve even finer image details and deliver sufficiently sharp edges in the segmentation results.
- **Label Generation:** OSM coordinates for buildings (polygons) and roads (centerlines with estimated widths based on highway tags) were automatically transformed into pixel-wise label maps.

Key Insights

- **Improved Generalization:** Training on a large variety of data spanning multiple different cities (e.g., Chicago, Paris, Zurich, Berlin) improves the classifier's ability to generalize to new, unseen locations (e.g., Tokyo).
- **Complete Substitution:** Semantic segmentation can be learned without any manual labeling by relying solely on large-scale noisy OSM labels, achieving acceptable results. The sheer volume of training data can largely compensate for lower accuracy.
- **Augmentation:** Even when a comfortable amount of accurate training data is available (the "gold standard"), pretraining with massive OSM data from other sites further improves results (e.g., boosting F1-score for the road class).

A context and semantic enhanced UNet for semantic segmentation of high-resolution aerial imagery [3]

Core Challenges in High-Resolution Aerial Imagery:

1. **Intra-class heterogeneity:** Objects of the same category (e.g., cars) have wide-ranging visual appearances (colors, characteristics), leading to difficulty in categorization. This stems mainly from insufficient contextual information.
2. **Inter-class homogeneity:** Objects of different categories (e.g., buildings and impervious surfaces) have similar appearances, leading to semantic ambiguity. This stems from poor semantic information.

Proposed solution: CSE-UNet: Context and Semantic Enhanced UNet

Key Insights – Addressing Heterogeneity via Context

Technology 1: Multi-level RFB-based Skip Pathways (Context Enhancement)

- **Purpose:** To strengthen the representational capacity for **multi-scale contextual features** and mitigate **intra-class heterogeneity**.
- **Mechanism:** Inspired by the concept of receptive fields in human visual systems, Receptive Field Blocks (RFB) are utilized in the skip pathways.
- **Implementation:** These pathways exploit varying convolution kernels and dilated convolutions to control the sizes and eccentricities of receptive fields, effectively highlighting informative regions.

Key Insights – Addressing Homogeneity via Semantic Enhancement and Performance

Technology 2: Multi-kernel Dual-path Encoder (Semantic Enhancement)

- **Purpose:** To extract and fuse multi-level features with **rich semantic information** and tackle **inter-class homogeneity** by enlarging the inter-class differences.
- **Mechanism:** The dual-path encoder contains an auxiliary multi-kernel based feature encoding path that provides additional semantics during the downsampling process.
- **Feature Fusion:** Feature outputs from the original UNet encoding path and the auxiliary path are fused via element-wise addition at each level to generate rich semantic representations.

Integrating semantic edges and segmentation information for building extraction from aerial images using UNet [1]

MultiRes-UNet Architecture and Multi-Scale Feature Learning

- **Goal:** To achieve **accurate mapping of building objects** from aerial imagery, overcoming challenges posed by vegetation and shadows which exhibit similar spectral values to buildings.
- **Technology: MultiRes-UNet** The MultiRes-UNet is an improved version of the original UNet network, designed to enhance feature assimilation and address inconsistencies between encoder/decoder features.

1. **MultiRes Block:** This block replaces the traditional series of two convolutions in the original UNet structure.
   - **Function:** Assimilates features learned from the data at various scales to comprise more spatial details.
   - **Mechanism:** It mimics inception-like blocks by approximating larger convolutions (like 5x5 and 7x7) using a sequence of lightweight and smaller 3x3 convolutions to extract spatial features from various scales while attempting to manage memory requirements.
2. **Res Path:** New shortcut path replaces the common skip connections used in UNet.
   - **Function:** Mitigates the **semantic gap** between the low-level features computed in the encoder and the notable higher-level features computed in the decoder.
   - **Mechanism:** Uses a **chain of convolutional operations** and residual connections instead of straightforwardly merging feature maps. Extra non-linear operations are expected to decrease semantic gaps.

Key Insights

- **Enhanced Boundaries:** Semantic edges are specifically used to enhance the boundary of semantic polygons.
- **Irregular Polygon Correction:** Edges help solve the issue of irregular semantic polygons and make them more appropriate for actual building forms.
- **Distinction:** Edges realize the distinction between adjacent buildings.
- **Performance Gain:** Integrating semantic edges enhanced the average quantitative results for Intersection Over Union (IOU) by **0.78%** (from 93.35% to 94.13%).
- **Overall Competency:** MultiRes-UNet achieved 93.14% IOU accuracy (with data augmentation), proving its success in building object extraction compared to state-of-the-art models like UNet (92.40%), DeeplabV3 (89.48%), and ResNet (88.84%).

Baseline Model

Baseline Model

For this milestone, the goal was to establish a baseline model for the project.

This involved preparing the learning pipeline, selecting a simple architecture (**classic U-Net with Resnet50**), training it on the dataset, and evaluating its performance.

Input Data and Augmentation

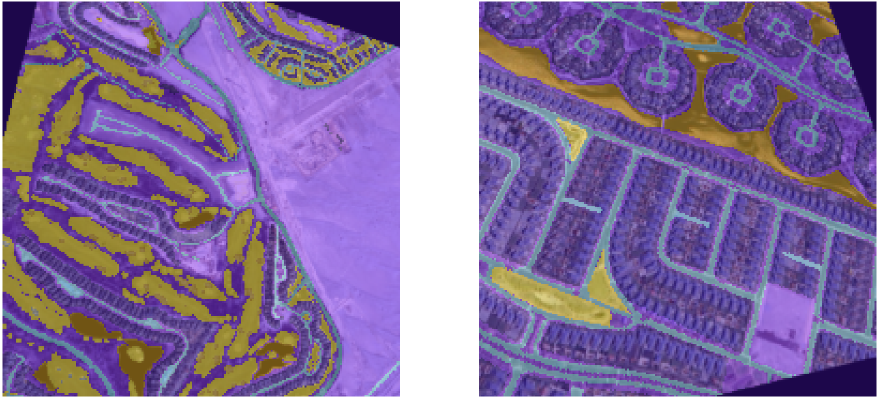

Figure 3: Sample input batch (data augmentation visible)

The model and the training process

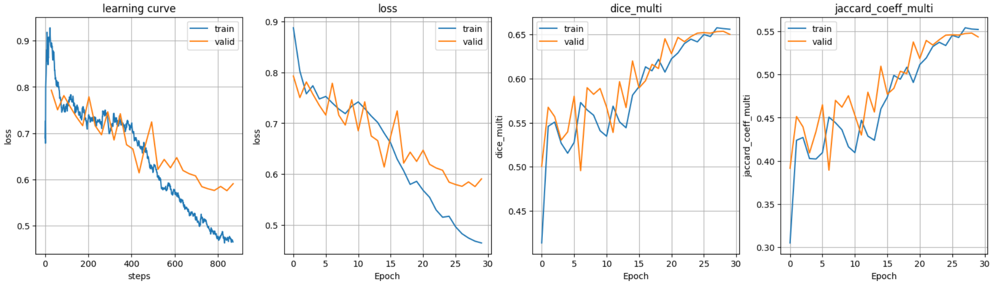

Figure 4: Learning history for classic U-Net with ResNet backbone (65% Dice score)

Evaluation

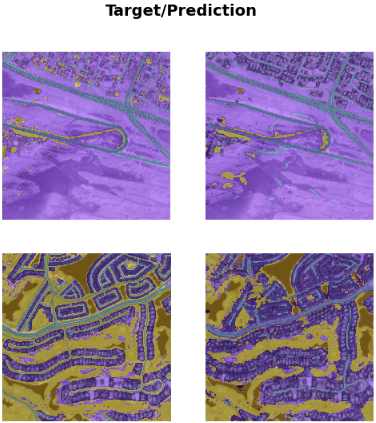

Figure 5: Target/Prediction comparison

Key notes, observations and findings

- Classic U-Net is not the best performing architecture for this task, but it serves as a good baseline (practical finding also confirmed with research papers).
  The result of ~65% Dice score is reasonable for a baseline attempt.
- Other architectures:
  - There are U-Net variants with attention mechanisms that could be explored in future milestones.
  - Advanced U-Net variants achieve better performance on this dataset (~80% Dice score).
- Data augmentation plays a crucial role in improving model generalization, especially with limited data.
  - In aerial imagery, augmentations like rotations, flips, and color adjustments are particularly effective.
  - Moreover, in contrast to natural images, aerial imagery allows for bigger zoom levels without losing context, which can be leveraged during training.

Problems identified:

- The dataset seems to be not entirely manually annotated, which may affect the model's performance – further investigation is needed and **potentially changing the dataset**.

Thank you for your attention

Questions

References

References

[1] Abdollahi, A. and Pradhan, B. 2021. Integrating semantic edges and segmentation information for building extraction from aerial images using UNet. *Machine Learning with Applications*. 6, (2021), 100194.

[2] Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T. and Schindler, K. 2017. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing*. 55, 11 (2017), 6054–6068. DOI:https://doi.org/10.1109/TGRS.2017.2719738.

[3] Wang, F. and Xie, J. 2020. A context and semantic enhanced UNet for semantic segmentation of high-resolution aerial imagery. *Journal of physics: Conference series* (2020), 012083.

[4] Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S. and Bai, X. 2019. Isaid: A large-scale dataset for instance segmentation in aerial images. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (2019), 28–37.

References