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

January 21, 2026

Table of contents

GitHub Project Repository

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, the project aims to identify the most effective approach for reliable segmentation, supporting geospatial research where high-quality segmented imagery is essential for deeper analysis and improved results.

Project overview

The 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**, model performance is evaluated to identify the most reliable and efficient architecture.

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]

AGH

Using Adapted FCNs to Leverage Noisy Crowdsourced Data

Technology: Adapted Fully Convolutional Network (FCN)

- A variant of the FCN architecture was utilized, performing pixel-to-pixel classification and 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 Dice Multi-score for the road class).

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

AGH

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%).

Architecture Selection

After the overview, the **CSE-Unet** [3] architecture was selected for further exploration and analysis in this project due to its promising results in addressing key challenges in aerial image segmentation.

Baseline Model

Baseline Model

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

This process involved the preparation of the learning pipeline, selection of a simple architecture (**classic U-Net with Resnet50**), training the model on the dataset, and evaluation of its performance.

Input Data and Augmentation

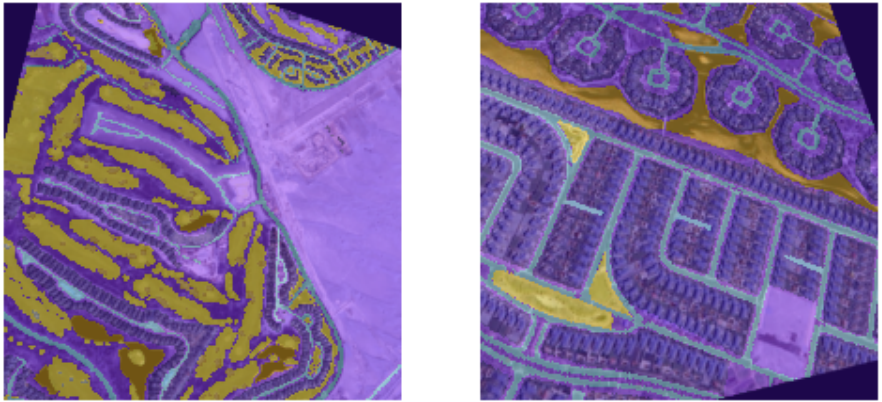

Figure 3: Sample input batch (data augmentation visible)

The model and the training process

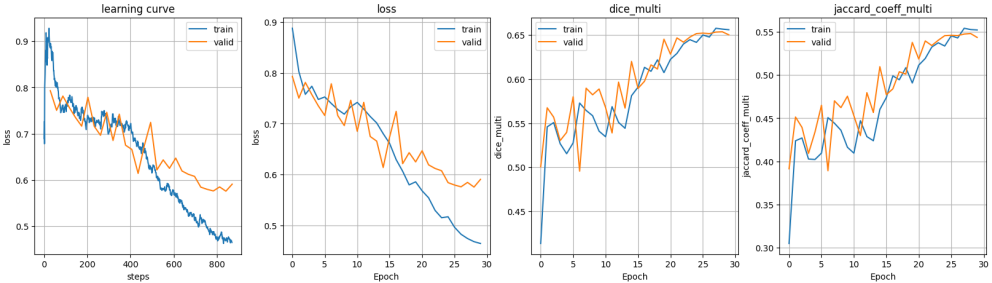

Figure 4: Learning history for classic U-Net with ResNet backbone (65% Dice score)

Evaluation

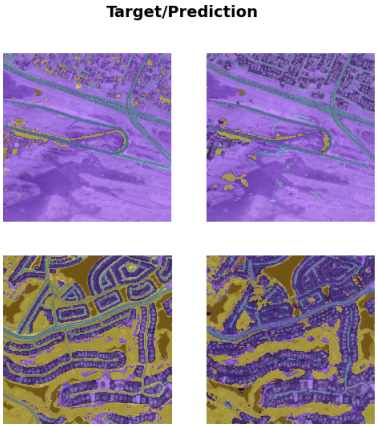

Figure 5: Target/Prediction comparison

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**.

First Implementation and Baseline Model Update

First Implementation and Baseline Model Update

For this milestone, the goal was to **implement the CSE-Unet architecture** as described in the original paper.

In addition, the baseline model established in the previous milestone was revisited and trained on the **updated dataset**.

Dataset Update

As identified in the previous milestone, the initial dataset had some issues with annotations. For this milestone, a revised version of the dataset was obtained, with improved annotations.

After a brief research, a ISPRS Potsdam dataset was selected as a replacement, which is a well-known benchmark dataset for aerial image segmentation tasks. It contains high-resolution aerial images along with their corresponding pixel-wise annotations for various land cover classes.

Class ID	Class Name	Color
0	Impervious surfaces	White
1	Buildings	Blue
2	Low vegetation	Cyan
3	Trees	Green
4	Cars	Yellow
5	Clutter	Red
6	Undefined	Black

Figure 6: Sample Image and Mask from the ISPRS Potsdam Dataset

Due to the simplicity of use, the dataset was accessed from Kaggle: ISPRS Potsdam dataset by Asad Iqbal.

> *Images are from the ISPRS Potsdam dataset. Each input image in the dataset was divided into image patches of size 300 × 300. These patches are divided into training (~2000 images) and testing (~400 images) datasets.*

Important note: To address the limitations of computational resources, the models were trained on the subset of the dataset (200 images, train/test ratio preserved) and the images were resized to 256x256 pixels. For the final evaluation, the full dataset is planned to be used.

Baseline model retraining

The baseline model from the previous milestone (classic U-Net with ResNet50 backbone) was retrained on the updated dataset twice:

1. With the pretrained weights from ImageNet (5 epochs freezed + 30 epochs unfrozen).
2. Without the pretrained weights. (38 epochs from scratch (early stopping)).

Experiment	Params	Epoch	Train Loss	Train Dice Multi	Train Jaccard Coeff Multi	Valid Loss	Valid Dice Multi	Valid Jaccard Coeff Multi	Time
Baseline	339M	38	0.869217	0.574932	0.414553	1.053347	0.399294	0.275351	00:45
Baseline*	339M	29	0.465625	0.770685	0.645282	0.493050	0.666335	0.557140	00:56

* Baseline model with pretrained weights

The results of the pretrained model are getting closer to the ones reported in the original paper, although still not matching them.

CSE-Unet Implementation

The CSE-Unet architecture was implemented as per the original paper, with the following key components:

1. The Dual-Path Encoder
2. RFB-Based Skip Pathways
3. The Decoder

The model was trained on the updated dataset for 56 epochs (early stopping) with the same training parameters as the baseline model.

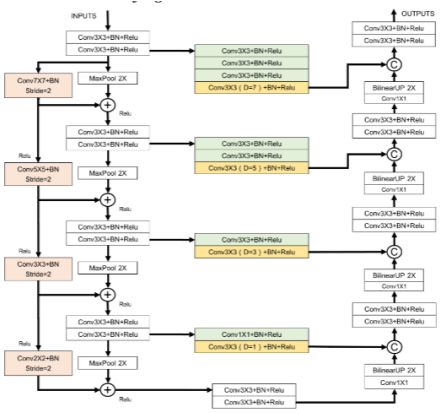

Figure 2. The architecture of CSE-UNet.

Figure 7: CSE-Unet Architecture

Experiment	Params	Epoch	Train Loss	Train Dice Multi	Train Jaccard Coeff Multi	Valid Loss	Valid Dice Multi	Valid Jaccard Coeff Multi	Time
Baseline	339M	38	0.869217	0.574932	0.414553	1.053347	0.399294	0.275351	00:45
Baseline*	339M	29	**0.465625**	**0.770685**	**0.645282**	**0.493050**	**0.666335**	0.557140	00:56
CSE-Unet	**36M**	56	0.854657	0.584462	0.434387	0.808706	0.475868	0.345522	**00:14**

* Baseline model with pretrained weights

Please note that the CSE-Unet model has significantly fewer parameters (36M vs 339M) and trains much faster due to its efficient architecture. The performance on the test set is similar to the baseline model (trained from scratch). However, on the validation set it outperforms the baseline model approx. 25%, which indicates better generalization.

Next steps

- Test baseline model with similar parameters count.
- **Train baseline model and CSE-Unet on the full dataset.**
- Tile-reconstruction for big-image segmentation.

Experiments

Preliminary experiments recap

Experiments were conducted on a subset of the data (500 out of 2500 images).

Experiment	Params	Epoch	Train Loss	Train Dice Multi	Train Jaccard Coeff Multi	Valid Loss	Valid Dice Multi	Valid Jaccard Coeff Multi	Time
Baseline	339M	38	0.869217	0.574932	0.414553	1.053347	0.399294	0.275351	00:45
Baseline*	339M	29	**0.465625**	**0.770685**	**0.645282**	**0.493050**	**0.666335**	0.557140	00:56
CSE-Unet	**36M**	56	0.854657	0.584462	0.434387	0.808706	0.475868	0.345522	**00:14**

* Baseline model with pretrained weights

Please note that the CSE-Unet model has significantly fewer parameters (36M vs 339M) and trains much faster due to its efficient architecture. The performance on the test set is similar to the baseline model (trained from scratch). However, on the validation set it outperforms the baseline model approx. 25%, which indicates better generalization.

The models

Baseline

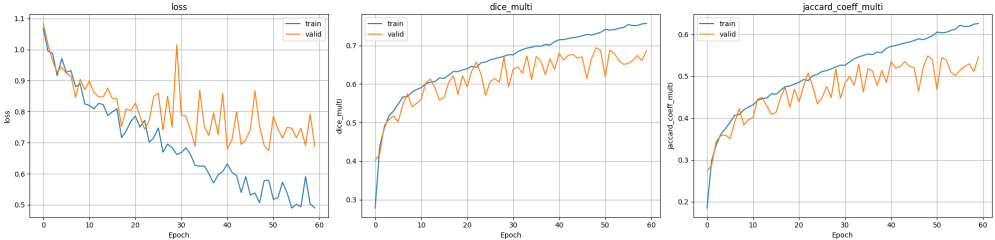

Figure 8: Baseline Training History

CSE-Unet

Base implementation

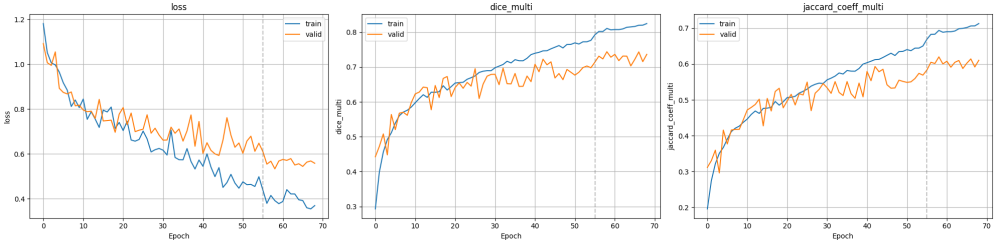

Figure 9: Base implementation History

Updated implementation

Changes

- dropout (0.2) in DoubleConv blocks
- custom loss function: combined CrossEntropyLoss + DiceLoss
- technicalities: I used some tools for the first time

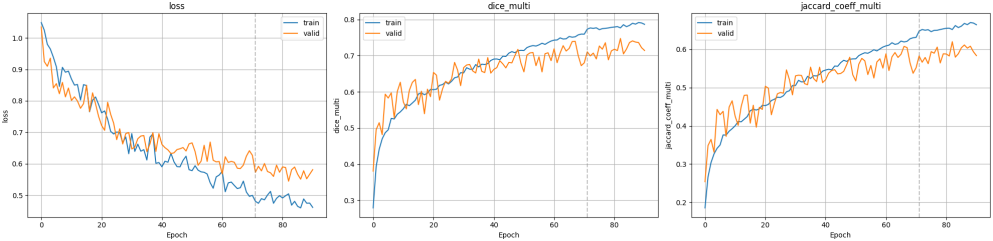

Figure 10: Updated implementation history

Results

model	t_loss	t_dice_multi	t_jaccard_coef_multi	v_loss	v_dice_multi	v_jaccard_coef_multi
Baseline (ResNet34)	0.578534	0.730393	0.591593	0.691645	0.694157	0.54865
CSE-Unet (Base)	**0.391817**	**0.810424**	**0.693776**	**0.533609**	0.743653	0.61987
CSE-Unet (UI)	0.504399	0.776692	0.652036	0.544584	**0.747400**	**0.62012**

Total params:

- Baseline (resnet34): 41,221,668
- CSE-Unet: 36,988,807

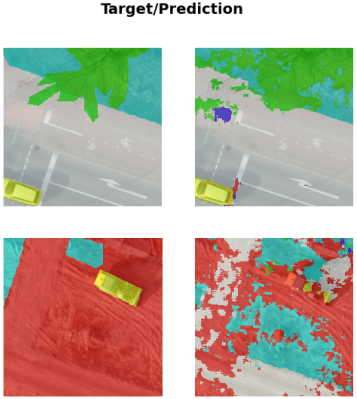

Figure 11: Baseline

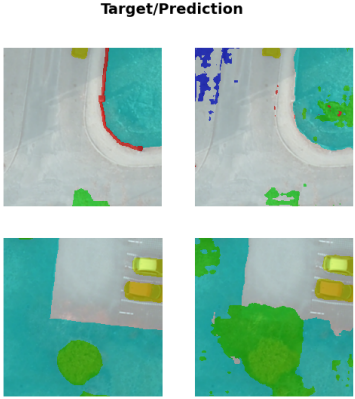

Figure 12: CSE-Unet

Conclusions

Key Findings

- **Efficiency vs. Performance:**
  - **CSE-Unet** outperforms the **ResNet34 Baseline** by ~**7 p.p.** in Jaccard Coeff Multi (0.62 vs 0.55) with a comparable parameter count (~36M vs ~41M).
  - On the data subset, it achieved competitive validation scores against the massive **ResNet50 Baseline** (339M params) while training significantly faster.

- **Architecture & Generalization:**
  - The **Updated Implementation** (Dropout 0.2 + Combo Loss) successfully mitigated overfitting.
  - While raw metrics between Base and Updated CSE are similar, the Updated model shows a healthier convergence gap between training and validation loss.

More Experiments

Previous Experiments Recap

Firstly, experiments on a subset of the data (500 out of 2500 images) were conducted to compare the baseline U-Net and CSE-Unet architectures. The baseline model, with 339M parameters, achieved a Jaccard Coefficient of 0.2754, while the CSE-Unet, with only 36M parameters, trained faster and reached a higher validation Jaccard score of 0.3455. Notably, using pretrained weights for the baseline improved its performance, yielding a score of 0.5571. Overall, the CSE-Unet demonstrated efficient training and competitive results despite its smaller size.

Further experiments were performed on the full dataset (2500 images) to evaluate the impact of dropout and a combined loss function on the CSE-Unet architecture. The base implementation of CSE-Unet showed signs of overfitting, with a significant gap between training and validation losses. By introducing a dropout rate of 0.2 in the DoubleConv blocks and employing a combined CrossEntropyLoss and DiceLoss, the updated implementation mitigated overfitting. Although raw metrics between the base and updated CSE models were similar, the updated model exhibited a healthier convergence gap, indicating improved generalization.

Migration to Full Dataset

As the dataset provided on Kaggle was incomplete and eemed mislabeled, migration to a original full dataset provided by the The International Society for Photogrammetry and Remote Sensing webpage was performed.

Such approach required some manual data preparation (i.e. cutting images into tiles, cleaning, splitting). However, the new dataset provided significantly more data ($\approx$ 3300 512x512 tiles after cleaning vs ~2400 256x256 tiles in the Kaggle dataset), which positively impacted the model performance and generalization.

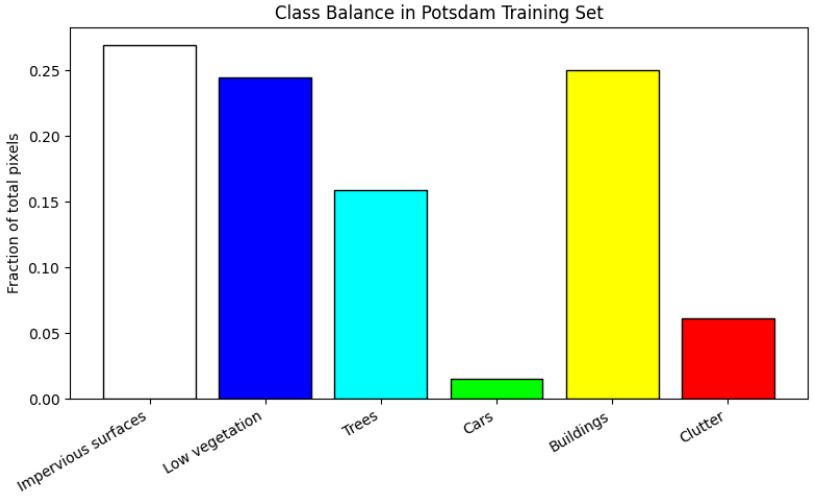

Figure 13: Class balance in the training subset of the full dataset

Initial comparison

To assess model performance on the full dataset, experiments were run with both the classic U-Net (using a ResNet34 backbone, matching the CSE-Unet in parameter count) and the CSE-Unet architecture enhanced with **dropout** and **weight decay**. Training settings largely mirrored previous experiments, with slight modifications to the learning rate and epoch count to better suit the expanded dataset.

- Learning rate: $\approx 1e-4$ (lr finder).
- Training time: ≈ 50 epochs (early stopping).
- Loss function: CrossEntropy*

The combined CrossEntropy + Dice loss was attempted, but for bigger input sizes it led to unstable training dynamics and worse preliminary results, as Dice loss is known to be unstable for the preliminary training stages.

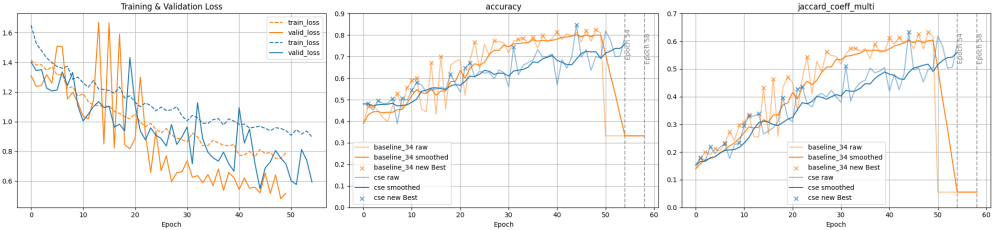

Figure 14: Training and validation histories for Baseline (Orange) and CSE-Unet (Blue) models on the full dataset.

- **Stability & Noise**: The CSE-Unet (Blue) demonstrates significantly smoother training curves across all metrics. In contrast, the Baseline (Orange) exhibits high volatility, with frequent, sharp spikes in validation loss and jagged "raw" metric lines.
- **Convergence Speed**: The Baseline reaches higher accuracy and Jaccard scores earlier in the training process (around Epoch 30–40), whereas the CSE-Unet follows a more gradual, linear improvement path.
- **Generalization Gap**: The gap between training and validation loss is narrower and more consistent for the CSE-Unet, suggesting that the architectural features and dropout are effectively regulating the model.
- **Late-Stage Collapse**: At approximately Epoch 50, the Baseline model experiences a total performance collapse. This suggests a "catastrophic forgetting" or an unstable gradient event that triggered the early stopping mechanism.
- **Convergence Illusion**: The CSE-Unet's steady improvement indicates it has not yet converged, while the Baseline's early plateau followed by collapse suggests it may have prematurely converged to a suboptimal solution.

AGH

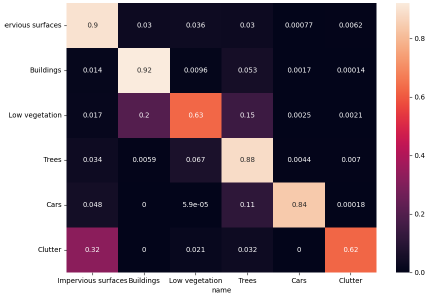

Figure 15: MC of Baseline

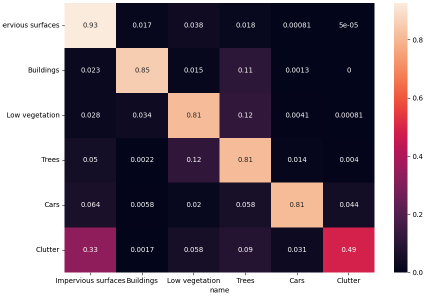

Figure 16: CM of CSE-Unet

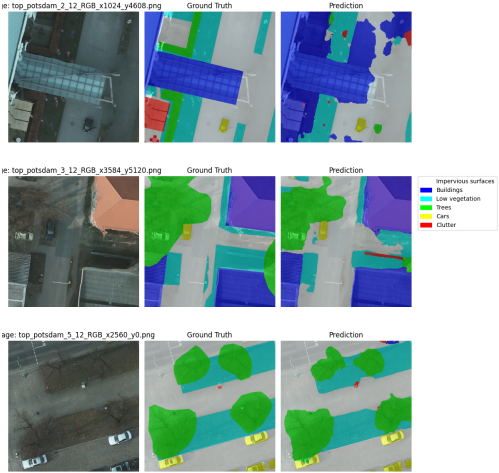

Figure 17: Predictions/Targets of Baseline

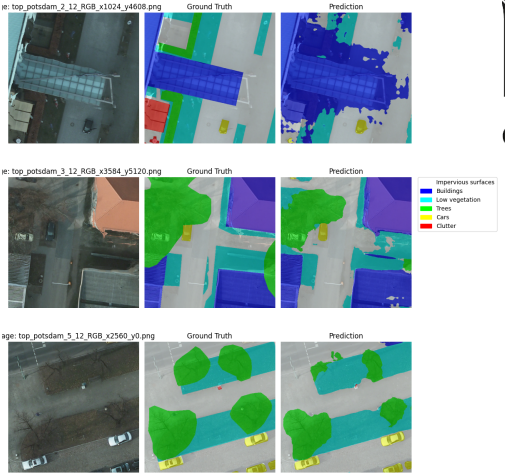

Figure 18: Predictions/Targets of CSE-Unet

Further CSE Experiments

To take the advantage of the fact that CSE-Unet trains stably on the full dataset, further experiments were run to improve its performance.

"Phase 2"

Firstly, using the lowered lr, the CSE-Unet was trained for another 30 epochs. This led to steady improvements in validation metrics without overfitting.

"Phase 3"

Then, the combined CrossEntropy + Dice loss function was utilized again with low learning rate as it is known for the ability to "understand" the "object instances" better, as it cates to both pixel-wise accuracy and overall shape similarity. Unfortunately, the try did not lead to any improvements, but did not significantly degrade the overall performance either.

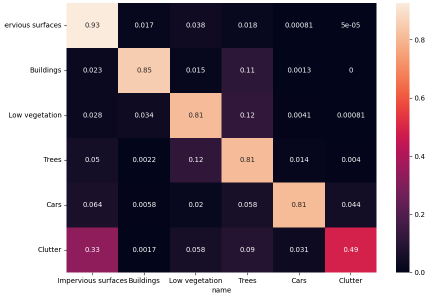

Figure 19: CM of CSE-Unet

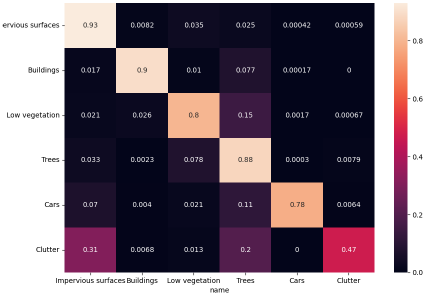

Figure 20: MC of Phase 2 CSE-Unet

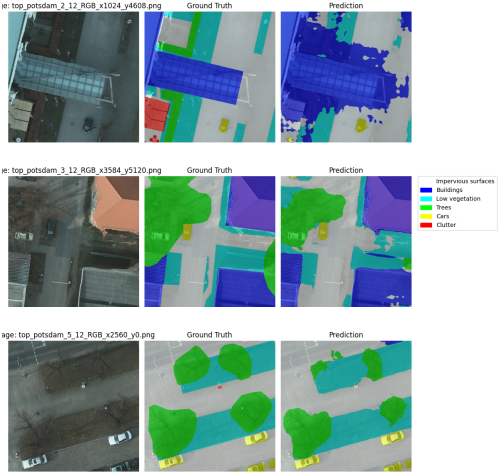

Figure 21: Predictions/Targets of CSE-Unet

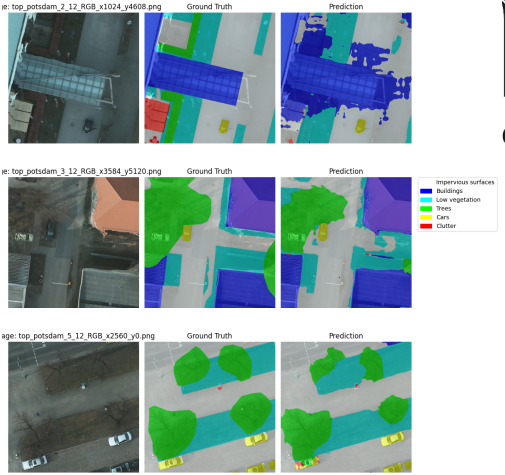

Figure 22: Predictions/Targets of Phase 2 CSE-Unet

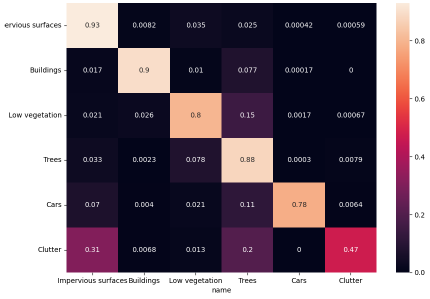

Figure 23: MC of Phase 2 CSE-Unet

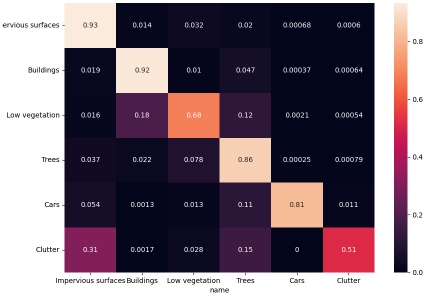

Figure 24: CM of Phase 3 CSE-Unet

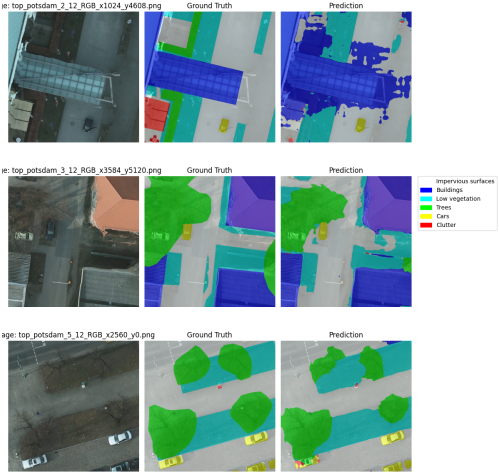

Figure 25: Predictions/Targets of Phase 2 CSE-Unet

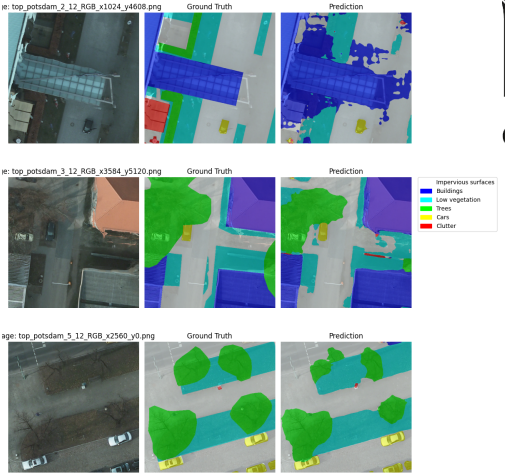

Figure 26: Predictions/Targets of Phase 3 CSE-Unet

Results summary

Model	Accuracy	Dice Multi	Jaccard Coeff Multi
baseline_34	0.83	0.75	0.63
cse_combined_loss	0.77	0.66	0.54
cse	0.85	0.75	0.63
cse_phase_2	**0.87**	**0.78**	**0.67**
cse_phase_3	0.86	0.76	0.65

Please note that there was the idea to use Test Time Augmentation (TTA) to further boost the performance, but due to performance issues on the remote server, the TTA experiments were not able to be run successfully.

IoU per class for the Phase 2 CSE-Unet model

To better understand the model performance across different classes, here are the Intersection over Union (IoU) scores for each class in the Phase 2 CSE-Unet model*:

Class ID	Name	IoU
0	Impervious surfaces	0.56
1	Buildings	0.66
2	Low vegetation	0.61
3	Trees	0.45
4	Cars	0.20
5	Clutter	0.34

* to address performance issues, the Per class IoU was calculated based a subset of validation images (10%).

Final experiments

Final experiments

To conclude the experiments on the new dataset, additional experiments were performed on the CSE-Unet model.

The current best model is cse_phase_2: best Jaccard Coef Multi 0.671; trained using lr-s `slice(1e-6, 1e-4)` and effective batch size of 16 (4 GPU batches 4 accumulation steps).

"Phase 4"

All experiments were trained with warm-start from current best model weights (unless otherwise noted).

1. **cse_phase_4_1*** – 10x smaller learning rates, effective batch size 16 – to further stabilize training.
2. **cse_phase_4_1_combined_loss*** – warm-start from cse_phase_4_1, 10x smaller learning rates, effective batch size 16, combined CrossEntropy + Dice loss – to see if combined loss helps when training is stable.
3. **cse_phase_4_2_combined_loss*** – same learning rates, batch size 8 (no Gradient Accumulation) – to see if smaller batch size helps with combined loss.
4. **cse_phase_4_3_focal_loss*** – same learning rates, batch size 8 – to evaluate focal loss with these settings.

	loss	valid_loss	accuracy	jaccard_multi
cse_phase_4_1	Cross Entropy	0.434	0.864	**0.681**
baseline_34_pretrained	Cross Entropy	0.485	N/A	0.678
cse_phase_2	Cross Entropy	0.508	**0.868**	0.671
cse_phase_4_1_combined_loss	CE + Dice	0.53	0.86	0.669
cse_phase_3	Cross Entropy	0.529	0.855	0.652
cse_phase_4_2_combined_loss	CE + Dice	0.454	0.834	0.646
cse	Cross Entropy	0.548	0.846	0.634
baseline_34	Cross Entropy	0.481	0.825	0.633
cse_phase_4_3_focal_loss	Focal	0.295	0.818	0.613
cse_combined_loss	CE + Dice	0.677	0.774	0.535

As we can see, the newly trained cse_phase_4_1 model outperforms both the previous best (cse_phase_2) and the baseline pretrained model, achieving the highest Jaccard coefficient and improved validation loss.

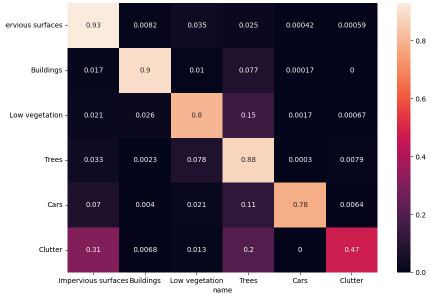

Figure 27: MC of Phase 2 CSE-Unet

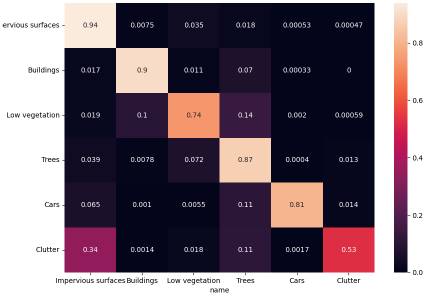

Figure 28: CM of Phase 4.1 CSE-Unet

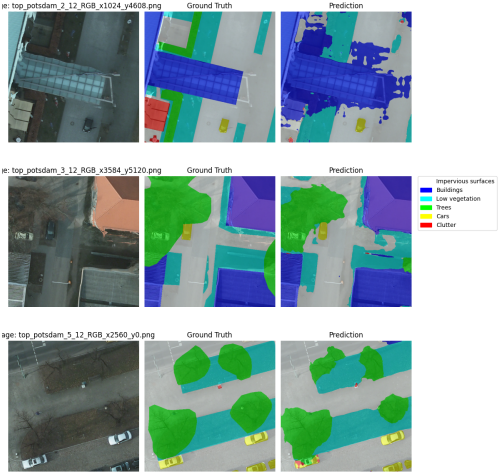

Figure 29: Predictions/Targets of Phase 2 CSE-Unet

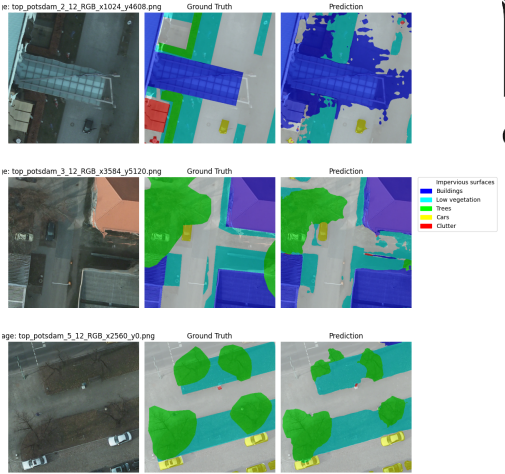

Figure 30: Predictions/Targets of Phase 4.1 CSE-Unet

Conclusions

Project Summary & Key Achievements

The **Context and Semantic Enhanced (CSE) U-Net** was successfully implemented and evaluated against both standard and pretrained ResNet-based U-Net baselines for high-resolution aerial image segmentation.

- **Surpassing the Baseline:** Through iterative optimization, the final model (**Phase 4**) achieved a **Jaccard (IoU) of 0.681**, successfully outperforming the **ResNet-34 Pretrained Baseline**.
- **Data-Centric Improvement:** Migrating from a noisy Kaggle subset to the official **ISPRS Potsdam benchmark** was a critical decision that stabilized training and enabled valid comparisons.
- **Efficiency:** The CSE architecture achieved these superior results with ~**12% fewer parameters** (36M vs 41M) than the baseline, validating the efficiency of Multi-level Receptive Field Blocks (RFB).

Architectural Analysis

The transition from Phase 2 to Phase 4 provided critical insights into the training dynamics:

- **The "Slow Burner" Confirmed:** Early experiments suggested the CSE model was a "slow burner" that hadn't fully converged. Phase 4 confirmed this: by applying a **warm restart with 10x smaller learning rates**, the model escaped its plateau and significantly reduced validation loss.
- **Volatility vs. Stability:** The ResNet Baseline exhibited high volatility and suffered "catastrophic collapse" in early epochs. In contrast, the CSE-Unet's RFB modules acted as a natural regularizer, maintaining linear stability. This stability was crucial, as it allowed for the aggressive fine-tuning in Phase 4 without destabilizing the weights.
- **Fine-Grained Feature Recovery:** The lower learning rate in Phase 4 specifically improved the model's ability to resolve difficult, small-scale classes. Confusion matrices showed distinct improvements in **Cars** and **Clutter**, proving that the final performance boost came from refining details rather than just general context.

Methodological Challenges & Limitations

Despite beating the baseline, the project target **Jaccard score of 0.75** (reported in literature) was not fully met. This gap is attributed to:

- **Computational Constraints:** High-resolution segmentation is resource-intensive. Limits on batch size and training time prevented the extensive hyperparameter grid-searches required to fine-tune a custom architecture to state-of-the-art levels. The possibilities of using gradient accumulation and mixed-precision training were explored, but still fell short of enabling exhaustive experimentation.
- **Loss Function Complexity:** Experiments with **Combined Loss (CE + Dice)** and **Focal Loss**—even in the stable Phase 4—consistently failed to outperform standard CrossEntropy. This confirms that for this specific architecture/dataset combination, auxiliary losses introduced gradient instability rather than refinement.
- **The Generalization Gap:** While Dropout and Weight Decay solved early overfitting, closing the final gap to the literature benchmarks likely requires more aggressive augmentation strategies (e.g., MixUp or Mosaic) which were outside the scope of this computational budget.

Final Takeaways

- **Custom vs. Pretrained:** It was demonstrated that a **custom-designed architecture (CSE-Unet) trained from scratch** can outperform a **heavy, pretrained ImageNet baseline**.
- **The Importance of Schedules:** The success of Phase 4 proves that for complex aerial data, a **multi-stage learning rate schedule** is just as critical as the architecture itself.
- **Viability:** The model successfully addressed "intra-class heterogeneity" and demonstrated that with proper regularization and patience, deep learning can reliably segment complex urban environments.

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.