DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Control of Autonomous Multi-Agent Systems Part I

Report

Convolutional Neural Networks for Automated Built Infrastructure Detection in the Arctic Using Sub-Meter Spatial Resolution Satellite Imagery

Weihao Wang
wang.1988339@studenti.uniroma1.it

Supervised by
Professor Alessandro Giuseppi
PhD Emanuele De Santis

Master in Control Engineering

Department of Computer, Control and Management Engineering
"Antonio Ruberti" (DIAG)
University of Rome "La Sapienza"
AY 2022/2023

Contents

Chapter 1

Introduction

In this report, I present a comparative study of three state-of-the-art semantic segmentation models: Unet++, DeepLabv3, and PSPNet. My goal is to investigate the performance of these models on a particular dataset which is appeared in paper Convolutional Neural Networks for Automated Built Infrastructure Detection in the Arctic Using Sub-Meter Spatial Resolution Satellite Imagery and to identify their strengths and weaknesses.

Deep learning models have shown promising results in addressing this challenge, and several models have been proposed in recent years. In this report, I focus on three popular models: Unet++, DeepLabv3, and PSPNet. Unet++ is an extension of the original Unet model that incorporates skip connections and multi-scale feature fusion to improve segmentation accuracy. DeepLabv3 is a convolutional neural network that uses atrous convolution to capture multi-scale contextual information. PSPNet is another network that uses pyramid pooling modules to capture global contextual information.

In detail, I compare the performance of these three models on a particular dataset I mentioned using a horizontal comparison experiment. I also investigate the performance of Unet++ on a sub-datasets of the same dataset using a vertical comparison experiment. My study aims to provide insights into the strengths and weaknesses of these models and to identify the factors that affect their performance in this particular task.

Chapter 2

Materials

In order to ensure the consistency with the original paper as much as possible, I re selected the dataset, that is, the satellite map range of a smaller area as shown in Figure2.1. The selection of this dataset eliminated many unlabeled noise areas that can help prevent the model from biasing towards the dominant classes and improve the segmentation accuracy for the less frequent classes. Additionally, the model can focus on learning the correct segmentation of the labeled regions, which can lead to better segmentation performance.

(a) Original Area Selection (b) New Area Selection (yellow area)

Figure 2.1: The difference between the two selected dataset areas

2.1 Dataset of Horizontal Comparison Experiment

This dataset is selected of the area in Figure 2.1, and splits the training set, test set, and validation set at a ratio of 80:10:10. My training dataset consisted of 1420 tile pairs, and both our validation and testing datasets consisted of 176 tile pairs.

For training CNNs, aerial images and corresponding annotated raster layers were split into smaller tiles sized at 256 pixels by 256 pixels. The vector of polygon features of buildings and roads are rasterized, since CNNs require their input to be in the form of an image.This time we mainly test two data sets, the first data set is used for horizontal comparison experiment, and the second data set is used for vertical comparison experiment. After splitting the data into sub-datasets, we can measure the new size of each target class as the number of pixels in the labeled masks belonging to each class, as seen in Table 2.1 and Table2.2.

	Background	Road	Residential	Public
Training	76659546	6959240	6175042	3267292
Validation	9419140	886804	698628	529764
Testing	8726324	743228	1025684	548325

Table 2.1: Class sizes in Horizontal comparison experiment dataset measured as number of pixels in the labeled masks.

2.2 Dataset of Vertical Comparison Experiment

In order to investigate the performance and generalization ability of Unet++ for this specific task, I also did a Vertical Comparison Experiment. I utilized the dataset obtained from the Horizontal Comparison Experiment, and a subset of this data was generated by manually deleting the images where the labeled pixels accounted for less than 15% of the total pixels. Figure2.2 are showed an example of deleted images. And in the end, I also did a test on original dataset.

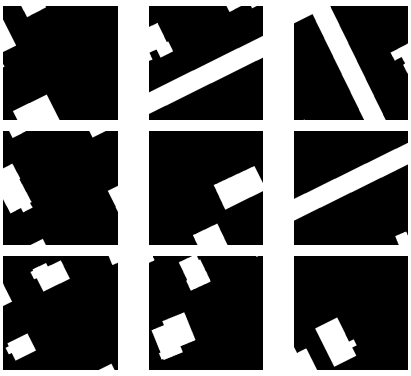


Figure 2.2: Example of deleted images.

Same as before, I split the training set, test set, and validation set at a ratio of 80:10:10. My training dataset consisted of 1420 tile pairs, and both our validation and testing datasets consisted of 176 tile pairs.

	Background	Road	Residential	Public
Training	31447209	4903509	5318238	3419812
Validation	4553245	452368	594178	187453
Testing	4007665	721372	672388	234671

Table 2.2: Class sizes in vertical comparison experiment data set measured as number of pixels in the labeled masks.

Chapter 3

Method

3.1 Experimental Design

Three state-of-the-art semantic segmentation models were selected for this study: Unet++, DeepLabv3, and PSPNet. And the study was divided into two parts: a horizontal comparison experiment and a vertical comparison experiment.

In the horizontal comparison experiment, the three models were trained and evaluated on the same dataset to compare their performance.

Models was constructed and trained using PyTorch-cuda 11.7 and the Segmentation Models for PyTorch library, with a hardware configuration of an AMD Ryzen 7 5800h 16 cores Processor and NVIDIA GeForce GTX 1650 with 4 GB of dedicated VRAM. Hyperparameters for model training are listed in Table 3.1.

Hyperparameter	Value/Type
Input size	$256 \cdot 256$ pixels
Batch size	8
Epochs	100
Loss function	CrossEntropyLoss
Optimizer	Adam
Learning rate	0.001

Table 3.1: Hyperparameters for model training

In the vertical comparison experiment, Unet++ was trained and evaluated on three different datasets, one of them is the dataset I used in last report, other two as second chapter 2.2 said. In order to investigate the effect of dataset size and class balance on the model's performance. The models were trained and evaluated using the same settings as in the horizontal comparison experiment.

3.2 Image Augmentation

Due to insufficient data, we employed image augmentation, which generates additional samples through transformations applied in the dataspace, they are: Transposition, Random 90°, Random horizontal flipping, Random vertical flipping. So my experiment consisted of three trials, in which we trained the model under different conditions: one trial applied only Transposition to the training dataset, one

trial for all of the transformations applied together, and one trial for none image augmentation.

3.3 Evaluation Metrics

The performance of the models was evaluated using precision, recall, and F1 score for each class. Precision measures the fraction of true positives among the predicted positive samples, recall measures the fraction of true positives among the actual positive samples, and F1 score is the harmonic mean of precision and recall.F1 score was averaged across all classes for an overall assessment of model performance. Furthermore, accuracy assessment was conducted for each trial of the augmentation experiment to determine the optimal augmentation method(s).

Chapter 4

Results

4.1 Horizontal Comparison Experiment

In the horizontal comparison experiment, the Unet++, DeepLabv3, and PSPNet models were trained with image augmentations and evaluated on the specified task dataset using the same training, validation, and testing sets. The performance of the models was evaluated using precision, recall, and F1 score metrics for each class, as well as the average metrics across all classes.

The results of the horizontal comparison experiment are presented in Table 4.1. It can be observed that Unet++ outperformed the other two models in terms of the average precision, recall, and F1 score across all classes. Specifically, Unet++ achieved an highest F1 score of 0.78, which is 5.41% and 6.85% higher than that of DeepLabv3 and PSPNet, respectively. Unet++ also achieved the highest F1 scores for most of the individual classes, including background, road, residents, public.

Model	Average Precision	Average Recall	Average F1-Score (No augmentation)	Average F1-Score (Transposition)	Average F1-Score (All augmentation)
Unet++	0.77	0.79	**0.78**	**0.78**	0.71
DeepLabv3	0.69	0.81	**0.74**	0.71	0.73
PSPNet	0.66	0.83	0.72	**0.73**	0.69

Table 4.1: Results of the horizontal comparison experiment

Confusion matrices showing the number of correctly and incorrectly classified pixels for each infrastructure class and augmentation trial are available in Figures 4.1, 4.2, 4.3, 4.4, and corroborate Table 4.1. These are a useful tool for visualizing the true and false positives and negatives used to calculate the reported accuracy metrics, as well as understanding how the model confuses classes during detection. It can be seen that there are varying sources of false positives and negatives. As we can see from confusion matrices, infrastructure classes are confused for each other at significantly higher rates. Background, road, Residents, public classes are largely confused for road, background, public, residents classes.

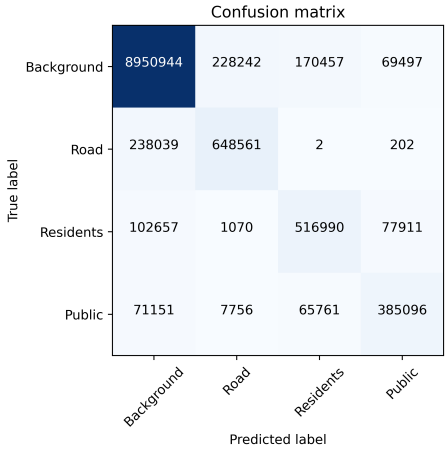


Figure 4.1: Unet++(None)

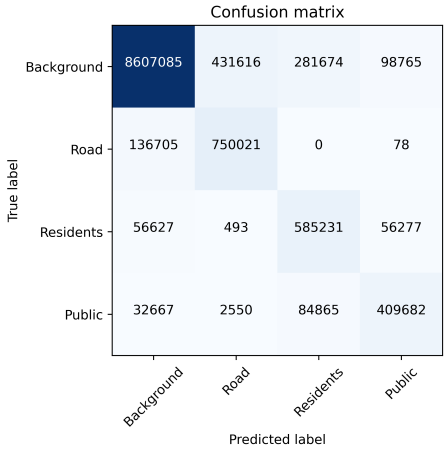


Figure 4.2: Unet++(Transposition)

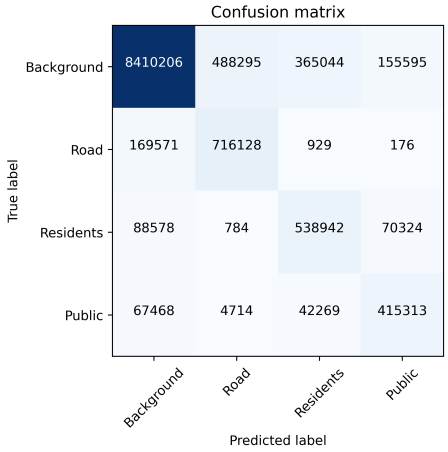


Figure 4.3: DeepLabv3(None)

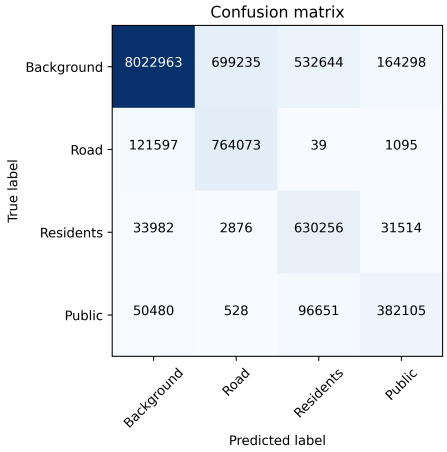


Figure 4.4: PSPNet(None)

4.2 Vertical Comparison Experiment

In the vertical comparison experiment, the Unet++ model was trained and evaluated on vertical comparison experiment of the Utqiagvik Satellite dataset to investigate the effect of dataset size and class balance on the model's performance. The performance of the model was evaluated using the same metrics as in the horizontal comparison experiment.

Dataset	Number of Images	Class Balance	Average F1-Score
Original Area	4386	Extremely Unbalanced	0.68
Reselect Area	1772	Unbalanced	0.78
Reselect Area and Noise Reduction	866	Balanced	0.87

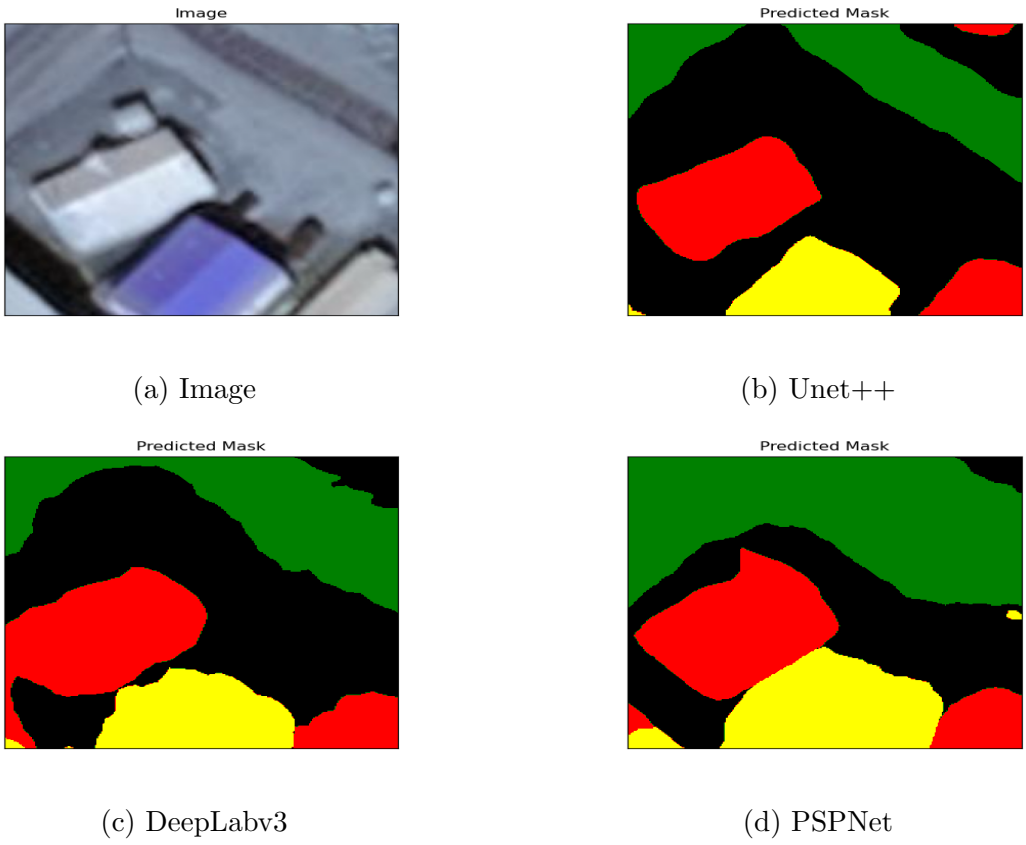
Table 4.2: Results of the horizontal comparison experiment

The results of the vertical comparison experiment are presented in Table 4.2. It can be observed that the performance of Unet++ improved as the class balance increased. Specifically, the model achieved an highest average F1 score of 0.68, 0.78, and 0.87 on the Original area dataset, reselect area sub-dataset, and reselect area

and noise reduction sub-dataset, respectively. This indicates that improving the class balance can improve the performance of semantic segmentation models.

4.3 Qualitative Results

In addition to the quantitative results, qualitative results were also obtained to visually compare the segmentation results of the models. Examples of the segmentation results for the Unet++, DeepLabv3, and PSPNet models are shown in Figure 4.5. It can be observed that the Unet++ model produced more accurate and detailed segmentation results compared to the other models.



(a) Image (b) Unet++

(c) DeepLabv3 (d) PSPNet

Figure 4.5: Examples of segmentation results for the Unet++, DeepLabv3, and PSPNet models on the Utqiagvik dataset

Overall, the results of this study indicate that the Unet++ model outperforms DeepLabv3 and PSPNet in terms of semantic segmentation accuracy on the Utqiagvik dataset . Additionally, increasing the dataset size and improving the class balance can further improve the performance of the model.