Our MapAI approach: focusing on data pipeline and loss functions

Tord Kriznik Sørensen1, Martijn Vermeer1, Jacob Alexander Hay1, Daniele Fantin1, and David Völgyes1, 2

1. Science and Technology AS, Oslo, Norway
2. E-mail any correspondence to: volgyes@stcorp.no

Abstract
Precision building detection is a difficult challenge because resolution, lighting conditions, and image quality greatly influence the performance of machine learning models. Additionally, the building types, settlement structure, road structure, soil color and texture, vegetation, and car types can also affect image segmentation, making the solutions local or regional. In this paper, we describe the solution for the MapAI challenge submitted by the ATELIER team. We focused on two primary parts: data processing and loss functions. Our main insights were that the data can be effectively resampled by exploiting its structure and that the boundary intersection over union metric is very forgiving, not giving the network enough incentive to refine the borders. The segmentation was performed with a standard 5-level deep U-Net with an additional conditional random fields (CRF) denoiser.

Keywords: artificial intelligence; machine learning; deep learning; segmentation, building detection

Introduction
MapAI challenge [1] is a precision building detection challenge with two tasks: (1) detecting buildings from aerial imagery (RGB) only. (2) detecting buildings from aerial and LiDAR combined data sources. These tasks are regional, as city structures, roads, vegetation, etc., change regionally, therefore, generalization is a hard task. In our view, the RGB-only task was the most challenging. Therefore, we focused our efforts on this task, and the second task used the same architecture as task 1, with minimal trivial modifications.

Materials and methods
Datasets and processing
We used the same preprocessing for tasks 1 and 2, and we did not apply augmentation to the LiDAR data. The training and validation data were presented as 500x500px image chips. Our data processing steps were motivated by the following observations about the data:

- The RGB, LiDAR and mask data had significant mismatches.
- The dataset is small.
- The data chips originated from slicing larger tiles into chips.

Our first processing step was to identify the chip structure and stitch the chips back to larger, at least 10x10 chip large mosaics. These larger mosaics allowed us to resample the data into new 900x900px chips with 400-pixel strides. This allowed us to apply data augmentation on these larger chips, and clip the results to 500x500px after the augmentations. The input RGB and LiDAR data were rescaled using subtraction of 115 and 2 and multiplying channels by 0.022 and 0.25, respectively, before the following augmentations.

- Affine transformations (rotation (±30°), zoom (0.8x-1.2x))
- Gaussian blur (σ = 0.4 px)
- Color jiggle (brightness change (±15%), contrast (±40%), saturation (±40%), hue (±5%))
- Additive Gaussian noise (σ = 0.05)
- CowMix blending of two samples
- Rolling the image around randomly

Augmentations were applied independently with 20% probability. The implementations were provided by the Kornia [2] augmentations, except for the CowMix [3] and rolling augmentations implemented by the authors.

We considered using alternative remote sensing datasets, e.g., SpaceNet [4], but it didn’t seem beneficial. However, we did use a secondary source: Google Maps images (zoom level 19) combined with LiDAR
data from hoydedata.no and Open Street Map for RGB, LiDAR, and label masks, respectively. Chips are generated from major Norwegian cities: Oslo, Nesodden, Drammen, Fredrikstad, Hamar, Bergen, Stavanger, Kristiansand, and Trondheim. (See the effect in the discussion section.) As a final pre-processing step, we found that there are missing buildings from the labels, and there are some non-existing buildings in the labels. We applied the following pre-processing: if the average elevation of a building footprint is below 2m, or if an early model predicted a building with at least 300 pixels but had no corresponding label, then we masked out these parts of the training and validation data by using 25-pixel dilation of the mask or detected region.

Network architecture
We used a simple U-Net architecture [4], followed by a conditional random fields denoiser [5]. The losses were applied to the U-Net raw output and the CRF output. The final predictions were made based on the CRF output. To improve performance, we used an ensemble of the 3 best models for task 1 (RGB only) and task 2 (RGB+LiDAR). The only difference between the networks was the different number of input channels.

Since the U-Net data network doesn’t like 500-pixel wide images and we wanted to reduce edge artifacts, all input images were mirror-padded up to 512x512 pixels, and the network output was clipped back to 500x500 pixels.

The network was implemented in Pytorch [6] using the Pytorch Lightning framework [7], and the U-Net network was implemented by [8]. The final ensemble models were exported into ONNX format [9] and submitted to the competition.

The ensemble model also applied test time augmentation [10]: the inputs were left-right mirrored, and the prediction of the mirrored image was transformed back. The two predictions of all three models were averaged to form the final prediction.

Evaluation metrics and loss functions
The MapAI [1] competition published their evaluation metrics, which was an intersection over union measure for the whole building footprint and another for the boundary region. The complete building metric was approximated with a Dice loss in our training. However, we found that the boundary region is very wide, and covers the majority of the typical buildings. We estimated that the border region is 14 pixels wide. We applied a Dice loss [11] on the boundaries with 4 and 2-pixel wide boundaries instead of the 14-pixel original boundary. Dice loss is known to have challenges in giving feedback on individual pixels, which is addressed by the Combo loss [12]. We used class-balanced focal loss [13] ($\gamma = 1.0$) as an auxiliary loss. With using $D_4$, $D_2$ notation for the boundary losses, $D_\infty$, the final combined loss function is as follows:

$$\text{Loss} = 4 \cdot D_2 + 2 \cdot D_4 + D_\infty + 4 \cdot \text{Focal loss}$$

Hyper-parameter search
The hyper-parameter search mostly focused on finding the optimal augmentation parameters using random search and manual tuning. The final best model was retrained three times with different seeds to provide base models for ensembling.

Post-processing
As a post-processing step, we removed all objects which are smaller than 50px, and we filled up all holes in buildings that are smaller than 150 pixels, with the help of Scikit image toolkit [14].

Results
The submission reached a score of 0.6963 for task 1, 0.7916 for task 2, and an overall score of 0.7440 on the independent test set.

Discussion
We worried that the training dataset would have a significantly different distribution from the test dataset. Even our worst-performing validation models achieved a 0.85 final score on RGB data, while on the test set, no team achieved above 0.71. This implies that a significant distribution shift can be observed between training and test data. The additional Google Satellite data didn’t improve our performance on the validation set, but we hoped to improve the generalization for the test set. It seems this additional data didn’t alleviate the problem.

The training data was extremely noisy, both in terms of imaging artifacts and label mismatch. We hypothesize that the largest source of error was the distribution shift, but averaging local metrics instead of global evaluation foreground/background predictions and not applying sliding window inference emphasized the effect of small objects at image borders. This gives a huge influence to the edge and corner pixels which was likely not the intention of the organizers.

Future work
Additional data sources, LiDAR augmentation, an ensemble of a larger variety of network architectures, and larger scale hyper-parameter search are planned as future work. Shape-aware loss functions and post-processing steps could be considered.

Conflict of interest declaration
The authors are employees of the Science and Technology AS (S&T), Oslo, Norway, and participating in the MapAI competition was a work assignment, and S&T offers deep learning-based segmentation of remote sensing imagery. However, the integrity of the results was not affected in any way by S&T, and the test data was evaluated independently by the organizers.