Abstract
Building segmentation from aerial images and 3D laser scanning (LiDAR) is a challenging task due to the diversity of backgrounds, building textures, and image quality. While current research using different types of convolutional and transformer networks has considerably improved the performance on this task, more precise and accurate segmentation methods for buildings are desirable for applications such as automatic mapping. In this study, we propose a general framework termed BuildSeg employing a generic approach that can be quickly applied to segment buildings. Different data sources were combined to increase generalization performance.

The approach yields good results for different data sources as shown by experiments on high-resolution multi-spectral and LiDAR imagery of cities in Norway, Denmark, and France. We applied ConvNeXt and SegFormer-based models on the high-resolution aerial image dataset from the MapAI-competition. The methods achieved an IoU of 0.7902 and a boundary IoU of 0.6185 on the test set. We used post-processing to account for the rectangular shape of the objects. This increased the boundary IOU from 0.6185 to 0.6189. The code is publicly available.¹

Keywords: Image Segmentation; Remote Sensing

Introduction
Detecting buildings from remote sensing imagery has been extensively studied [1, 2, 3] as it is of great importance for many fields, such as urban planning, population estimation, and topographic map production. Since the amount of data cannot be processed manually, data-driven machine-learning methods are needed to reduce the manual work required to obtain reliable urban development mappings. Segmenting buildings on a large scale is a challenging task because satellite or aerial images can be very diverse, for example, because of different styles of architecture, building materials, and topography. There are multiple benchmarks for the segmentation of buildings to use when testing building segmentation methods [4, 5, 6, 7, 8, 9, 10, 11, 12]. Since the silhouettes of buildings can be very different, combining several datasets with different characteristics can lead to more generally applicable building segmentation models.

We propose a framework for building segmentation referred to as BuildSeg. We consider the Inria Aerial Image Labeling Benchmark [4] combined with the MapAI-competition dataset [6] to improve the segmentation performance of the latter. When designing the BuildSeg framework, our goal was to design a segmentation pipeline that is generally applicable. Therefore, several benchmarks and corresponding models are available within the framework [4, 5, 6, 7, 8, 9, 10, 11, 12].

Neural networks and in particular convolutional neural networks (CNNs) have become the go-to methods for image segmentation, see [13] for a recent review. The U-Net [14] is one of the fundamental segmentation architectures, and we have been successfully applying

¹Code: https://github.com/llei66/BuildSeg

Figure 1: Examples from the MapAI-competition [6]. The left image (a) shows the ground-truth data, and the right image (b) the prediction created by the ConvNeXt model.
it to remote sensing and natural imagery (e.g., [15, 16, 17, 18, 19, 20, 21]). It uses an encoder-decoder structure that processes the input image at different scales and allows it to detect high-frequency patterns while being computationally feasible. The original U-Net architecture can be generalized by replacing the encoder and decoder with tailored networks. This makes the U-Net very versatile and allows us to utilize state-of-the-art encoders. In our framework, we consider two different U-Net variants, SegFormer [22] and ConvNeXt U-Net [23], where the decoder of the ConvNeXt U-Net are backward-strided convolutions [24].

The main contributions of this study can be summarized as follows: (1) We propose a general framework called BuildSeg based on [25] for segmenting buildings in aerial images of different resolutions; (2) We explore how 3D information from LiDAR affects the performance of deep CNN models; (3) We combine different datasets and apply rectangle-aware post-processing to create rectangular boundaries that match the labels more accurately. The proposed approach achieved the IOU of 0.7902 for the segmentation of images in the test set from the MapAI: Precision in Building Segmentation [6] benchmark.

**Method**

We developed our framework for the MapAI challenge [6], which provides both aerial images and LiDAR data. The challenge formulates two tasks. The first is the segmentation of buildings only using aerial imagery. In the second task, the LiDAR data must be segmented with or without aerial images. We additionally used the data from [4] to improve the performance. A subset of 5000 images was considered additional training data. To align the image sizes, we cropped the input images to 500×500.

To further increase diversity in the data, we applied augmentations, including random cropping, random vertical and horizontal axis flipping, and random changes to brightness (δ = 32), contrast (range: 0.5 − 1.5), saturation (range: 0.5 − 1.5), and hue (δ = 18).

We tried different models, such as the standard U-Net and variants of it, namely ConvNeXt and SegFormer [22]. For the ConvNeXt model, ConvNeXt [23] is used as the encoder, and backward-stride convolution [24] as the decoder. We also tried EfficientNet [26] as an encoder, but the results were not as good. All encoders were pre-trained on ImageNet. Two metrics were considered to measure the performance: intersection over union (IoU) and boundary intersection over union (BIOU) [27]. LiDAR height data were added directly as an additional channel to the multi-spectral data when available. For post-processing, we applied a sequence of morphological opening and closing operations to detect lines and removed points not matching the hypothesis of a rectangular structure.

**Experimental Results**

The results are summarized in Table 1. The model SegFormer-B5 performed best in terms of IoU and BIOU for the images. Note that SegFormer-B5 and SegFormer-B4 have more layers than SegFormer-B0.

Figure 2 illustrates the results of the SegFormer-B5 model, and it can be seen that the buildings were nicely captured. The averaged score, computed as the mean of IoU and BIOU, was 0.7044 without post-processing and slightly increased to 0.7045 after post-processing, so the latter should be preferred if IoU and BIOU weigh the same. When combined with LiDAR, the method reached an IOU of 0.8506 and a BIOU of 0.7461.

**Conclusion**

For the MapAI-competition, we proposed a solution that utilizes additional building datasets and current state-of-the-art deep learning architectures. The method achieved an IoU of 0.7902 and a BIOU of 0.6185 for the task of segmenting buildings in aerial images. Using additional information from LiDAR further improved the results, increasing the IoU to 0.8506 and the BIOU to 0.7461.

<table>
<thead>
<tr>
<th>Model</th>
<th>IOU</th>
<th>BIOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>0.7611</td>
<td>0.5823</td>
</tr>
<tr>
<td>ConvNext</td>
<td>0.7841</td>
<td>0.6105</td>
</tr>
<tr>
<td>SegFormer-B0</td>
<td>0.7632</td>
<td>0.5901</td>
</tr>
<tr>
<td>SegFormer-B4</td>
<td>0.7844</td>
<td>0.6116</td>
</tr>
<tr>
<td>SegFormer-B5</td>
<td>0.7902</td>
<td>0.6185</td>
</tr>
</tbody>
</table>
References

2. Li Y and Wu H. Adaptive building edge detection by combining LiDAR data and aerial images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2008; 37:197–202
