3D Reconstruction from Aerial Imagery

Exploring the third dimension of aerial imagery with generative adversarial networks

Project Report

Degree programme:	CAS Practical Machine Learning
Author:	Balthasar Teuscher
Thesis advisor:	Prof. Marcus Hudritsch
Expert:	Dr. Jürgen Vogel
Date:	8. April 2019

Management Summary

Expanding into the third dimension has become a widespread trend in the course of the current digitalization. Novel key technologies that incorporating the third dimension extensively are for example augmented- and virtual reality as well as autonomous navigation. However the majority of the data respectively content is only recorded in two dimensions like most aerial and satellite imagery. Creating a three dimensional model from such two dimensional content information is often an exhaustive task considering traditional photogrammetry approaches. This work explores new techniques to extract height information from aerial imagery based on methods from machine learning.

Contents

1 Introduction 4
- 1.1 Project Proposal 4
- 1.2 Objectives 4
- 1.3 Generative Adversarial Network 4

2 Problem Statement 5

3 Other Works 6

4 Dataset 7
- 4.1 Preprocessing 7

5 Model 8
- 5.1 Results 8
- 5.2 Evaluation 9

6 Summary 10

7 Contests of the figures 11

8 Bibliography 12

9 Declaration of Authorship 13

1 Introduction

This project work is in partial fullfillment of the Certificate of Advanced Studies (CAS) in Practical Machine Learning at the Bern University of Applied Sciences. The main goal of the project work is to gain practical hands-on experience with machine learning, more specific to follow up on a machine learning task and walk through it in an iterative manner. In the present case learning a three dimensional representation from two dimensional imagery using neural networks.

1.1 Project Proposal

The proposal for learning a three dimensional representation from aerial imagery came from my interest in spatial data science and my background in geography. Given that this project is not directly tied to my company and workload, its characteristic is rather exploratory. The overall goal is to get an overview of the state of the art and experiment upon them. Though the focus lies on approaches leveraging generative adversarial networks, it should not be considered as a constraint but rather as a guideline.

1.2 Objectives

Given the exploratory characteristic the objectives of this project are represent an initial outline of the foreseen workload at the beginning. The following three steps are consider as a path forward.

1. Elaborate the state of the art of 3d reconstruction with neural networks from literature
2. Apply approaches found in the literature to aerial imagery
3. Improve and/or create own model based on the gained insight

1.3 Generative Adversarial Network

A General Adversarial Network (GAN) is a machine learning technique that became increasingly popular after the original publication by Ian Goodfellow [1]. It constitutes of two networks playing a zero sum game. The generator network generates for example from a noise vector an image, which in turn the discriminator evaluates as fake or not based on the learned experience from a real dataset.

2 Problem Statement

Generating a three dimensional representation from two dimensional imagery is a common task in computer graphics, which is heavily applied in cartography and mapping in general. Extracting the depth from the content of a two dimensional image is an ambiguous problem. To address such ambiguity, one can use more than one image from different viewpoints of the same scene or object.

A traditional approach often applied in remote sensing to generate a digital elevation model (DEM) is stereo photogrammetry[1]. It uses two oblique photographic images and involves two steps, first to figure out the camera orientation relative to the image and second the triangulation respectively stereo matching of ground points or patches. The second step can be achieved either with semi global matching or patch matching whereas the camera orientation can be derived from the image directly. [2]

Among other similar approaches that are widely used in robotics, autonomous driving and simultaneous localization and mapping (SLAM) are for example stereo, multi-view stereo, optical flow, single image depth prediction, semantic segmentation and instance segmentation.[2] In the present project single image depth prediction resembles the most reasonable method to follow up given the available data which consists of a single continuous vertical aerial image.

[1] See https://en.wikipedia.org/wiki/Digital_elevation_model#Methods_for_obtaining_elevation_data_used_to_create_DEMs
[2] For an overview see http://www.robustvision.net/index.php

3 Other Works

There exist a vast amount of research papers dealing with the broader topic of 3d reconstruction, most of them in the field of computer vision and robotics. The three papers presented hereafter are selected to highlight different aspects of the specific problem this work likes to address.

- Large-Scale Semantic 3D Reconstruction: An Adaptive Multi-resolution Model for Multi-class Volumetric Labeling [3]

The work from Blaha et al. jointly applies semantic object segmentation and 3d reasoning based on implicit volumetric surface modelling. Given a shape semantically labeled as a building wall for example indicates the shape to be rather flat and vertical.

- Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps with Accurate Object Boundaries [4]

This paper by Hu et al. acknowledges the strong performance of convolutional neural networks (CNN) in estimating the depth of a single image. Weaknesses of such methods are lossy spatial resolutions and blurry and distorted reconstruction. Hu et al. address this shortcomings by fusing spatial information extracted at the different convolutional layers of the encoder with the output of the decoder followed by several refinement layers. Further they propose a loss function that can measure step edges.

- Learning Shape Priors for Single-View 3D Completion and Reconstruction [5]

The approach of Wu et al. first extracts a sketch including a depth map from the image before generating a voxel representation from this. The voxel representation finally gets tested against previously generated “natural shapes” by a generative adversarial network in order to constrain the ambiguity of 3d shapes fitting a given 2d representation.

The first approach is similar in terms of the domain and data used, albeit it leveraged multiple oblique images. On the other hand the applied methods are not based on neural networks opposed to the other two. The seconds approach resembles the problem statement of this work, generating a depth map, the closest. The drawback is that the dataset is only composed of indoor images, similar to the objects based dataset of the last approach. The latter is the only one which leverages a generative adversarial network.

Lacking access to the code of the papers above or unsuitable data modalities in combination with lack of time to fit the input to it, the decision was made to proceed towards experimenting with a working template.

4 Dataset

The region of interest in this project is the municipality of Zürich as the city government provides a free vectorized digital elevation model (DEM) on the their open data platform[3]. It consist of several levels of detail of which the digital terrain model in the form of a triangulated irregular network (TIN) and the digital surface model of the rooftops are used in this project to generate a height map between the ground level and the rooftop (Figure 1).

Figure 1: 3D City model of Zürich. (Source: www.stadt-zuerich.ch)

The high resolution areal imagery used in the following as input is provided freely by the canton Zürich on their data portal. It has a spacial resolution of about 0.1 metres and was recorded during spring 2016 in true colour[4].

4.1 Preprocessing

The first step in preprocessing the dataset for consumption in a neural network was to spatially partitioned the region of interest into about 8500 tiles. Each tile has a size of 1024 by 1024 pixel which translates into a covered ground area of around 10'500 square meters given the spatial resolution of 0.1 meters.

In a second step the a hight map serving as ground truth was extracted from the absolute height difference in meters between the terrain model and the surface model (Figure 2). This was achieved by triangulating the position of every 8^{th} by 8^{th} pixel from an intersecting shape of both the terrain- and surface model. From the resulted heights, the absolute difference of these pixels was linearly interpolated.

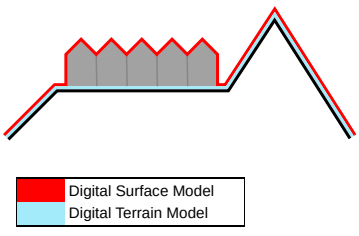

Figure 2: Digital elevation model (DEM).
(Source: www.wikipedia.com)

After initially experimenting with different input sizes, the final networks was fed with batch of 16 images each with the dimensions 128 by 128 by 3. The tiles were further split into a train and test dataset with a ration of roughly 90 to 10 percent and both the ground truth and the input were normalized by dividing by 255.

[3] See https://data.stadt-zuerich.ch/
[4] See http://maps.zh.ch/?topic=OrthoZH

5 Model

The first running model consisted of a simple autoencoder with each three convolutional layers and three deconvolutional layers, relu activation, L2 loss, adam optizer and was fed a whole tile as input. It turned out to collapse and diverge soon after a promising start. Using a densely connected layer after the encoder didn't yield any better results as the input size was to big for the parameters to fit into memory.

Following iterations included experimenting with the input size and finally applying batch normalization with the goal to help the network stabilize and converge as well as various layer configurations didn't help to make it converge. The introduction of an ensuing discriminator finally lead to a functional model (Figure 3).

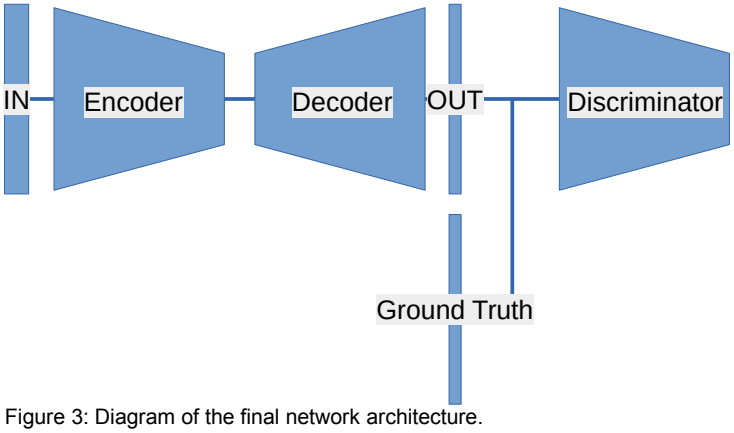

Figure 3: Diagram of the final network architecture.

The input (IN) in the final model are the tile of the aerial imagery, each split up into 16 parts to create a batch of RGB images with the dimensions of 16 by 128 by 128 by 3. The encoder was made up of four convolutional layers with a kernel size of 3 by 3 followed by leaky relu activation. The decoder consist of 4 deconvolutional layers with a kernel size of 4 by 4. All layer use relu activation except the last one using tanh to generate the output of dimension 16 by 128 by 128 by 1. The discriminator uses 6 convolutional layers alternating between feature extraction and downsampling, both with leaky relu activation and finally a fully connected layer with sigmoid activation

On all convolutional layers the stride is set to two and padding to same. In the decoder batch normalization is used, while in the decoder instance normalization is performed and spectral normalization in the discriminator at each layer to help stabilize the network. The model was trained using adam optimization.

The discriminator basically classifies the output as fake or real image base in experience from the ground truth. This further means that the model is adding the distinct loss function of generative adversarial networks [1]. This lead to the loss being calculated as L2 + L_{GAN}.

5.1 Results

The first figure illustrates the input, the ground truth and the output of some batch members (Figure 4). It shows that the model learned to distinguish the footprint of the building s quite accurate and even reproduces the unsharp edges introduced in the ground truth by interpolating during the preprocessing.

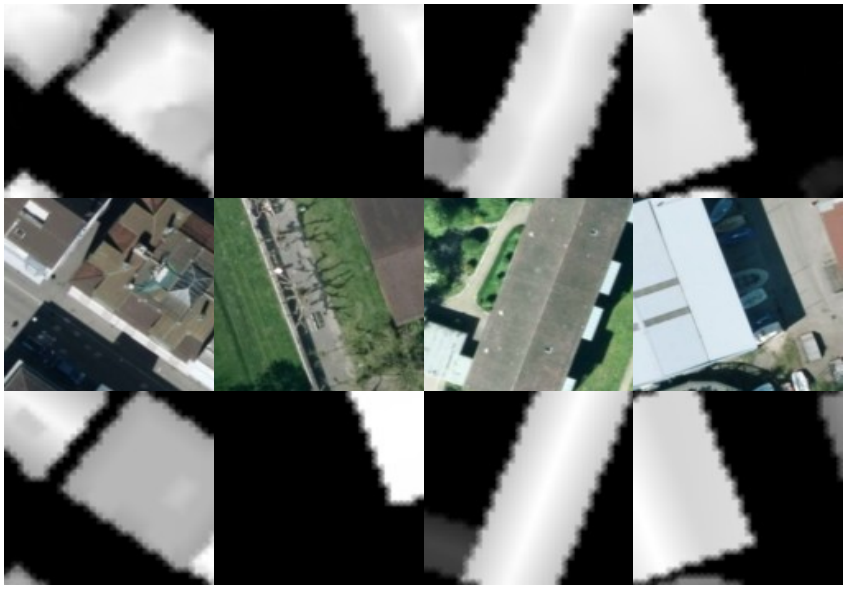

Figure 4: Input (middle), output (ip) and ground truth (down) of some batch members with size 128 by 128 pixel.

Figure 4 shows the input, output and ground truth of an image that was tested with the initial scale of 1024 by 1024 pixels. The output even shows the shape of the rooftop but also some artefacts on the street. Still the capacity to deal with shadows and streets is very good in this example as well as the ones before.

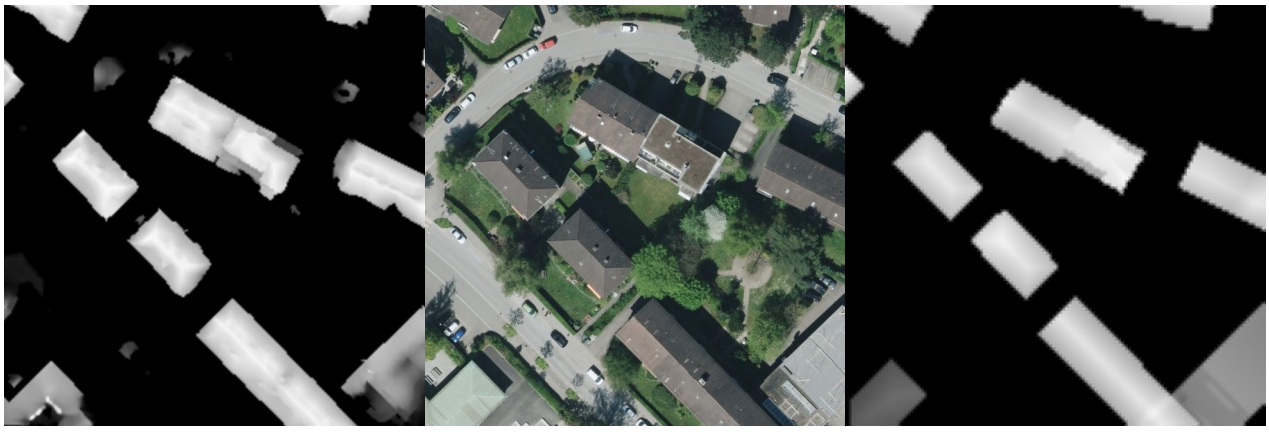

Figure 5: Input (middle), output (left) and ground truth (right) of testing a whole tile of 1024 by 1024 pixel.

That the testing with the full size tiles works means that the model can deal with different spatial resolutions. On the other hand it has difficulties with green roofs and wastewater treatment plants.

5.2 Evaluation

The mean error after training several epochs read around 4.4 meters for the training set and 5.2 meters for the test set. For the testing with the full tile of the mean error is about 8.7 meters for the test set.

It needs to be taken into account that the results shown previously are greyscale images, meaning the min and max values are scaled to 0 and 255. Further the ground truth has been interpolated in the preprocessing step leading to additional bias.

6 Summary

The most striking insight is that all the presented approaches as well as the final model implement some sort of constrain to address the ambiguity of the posed problem. In the present case this was the addition of a discriminator to the model analogous a generative adversarial network architecture. Further the results show that the model is capable of learning more complex features than simply correlating the spectral reflectance of pixel and patches, which is a common approach in image classification. Nevertheless the evaluation reveals a rather high error rate. By increasing the accuracy of the ground truth, adding multiple imagery layer, augmenting the data, training with different spatial levels, and also optimizing the hyper-parameters and architecture of the model, the overall robustness and accuracy could certainly be improved much.

7 Contests of the figures

Figure 1: 3D City model of Zürich. (Source: www.stadt-zuerich.ch) 7

Figure 2: Digital elevation model (DEM). (Source: www.wikipedia.com) 7

Figure 3: Diagram of the final network architecture. 8

Figure 4: Input (middle), output (ip) and ground truth (down) of some batch members with size 128 by 128 pixel. 9

Figure 5: Input (middle), output (left) and ground truth (right) of testing a whole tile of 1024 by 1024 pixel. 9

8 Bibliography

[1] I. J. Goodfellow *et al.*, “Generative Adversarial Networks,” Jun. 2014.

[2] X. X. Zhu *et al.*, “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources,” *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017.

[3] M. Blaha, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler, “Large-Scale Semantic 3D Reconstruction: An Adaptive Multi-resolution Model for Multi-class Volumetric Labeling,” *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3176–3184, Jun. 2016.

[4] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, “Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps with Accurate Object Boundaries,” 2018.

[5] J. Wu, C. Zhang, X. Zhang, and Z. Zhang, “Learning Shape Priors for Single-View 3D Completion and Reconstruction,” pp. 1–17, 2018.

9 Declaration of Authorship

I hereby certify that I composed this work completely unaided, and without the use of any other sources or resources other than those specified in the bibliography. All text sections not of my authorship are cited as quotations, and accompanied by an exact reference to their origin.

Place, date:

Signature: