3D Map Reconstruction From Single Satellite Image Using a Deep Monocular Depth Network

2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN) | 978-1-6654-8550-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICUFN55119.2022.9829688

Changmin Son
School of Electronic and Electrical Engineering
Kyungpook National University
Daegu, South Korea
tptkddmldkdl@gmail.com

Soon-Yong Park
School of Electronic Engineering
Kyungpook National University
Daegu, South Korea
sypark@knu.ac.kr

Abstract**—** In this paper, we propose a 3D reconstruction scheme from single image with deep monocular depth estimation network, BTS (From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation) [1]. Furthermore, we expand it to height estimation focused building from remote sensing images. To address these issues, we substitute depth estimation loss function with height estimation loss function. Moreover, considering improving the quality of the building height map and looking as similar as possible to the ground-truth view, we apply building adaptive loss function.

Keywords—3D map; Monocular height estimation; Deep learning; Remote sensing image;

I. Introduction

Recently, with the development of remote sensing, height estimation is of great importance in understanding geometric relations with a scene. However, to project 3D points with publicized satellite 2D images, we must know RPCs (Rational Polynomial Coefficients). Nevertheless, in the stance of common users, it is not always possible to obtain these coefficients, but also it takes a relatively long time to get the 3D points. So for considering these things, we approach views from computer vision matching methods.

Firstly, stereo matching method is that the disparity(depth) is estimated using pixel matching in two or more input images. Since the disparity in all pixels of the input image is estimated, it can perform dense 3D reconstruction, but unfortunately, stereo matching method require focal length or camera pose.

Secondly, monocular depth and height estimation methods are estimated from a single-view image. Only reconstructing single-view image contain a lack of geometrics information, which is inherently ambiguous, and a technically ill-posed problem, with a large source of uncertainty coming from the overall scale to estimate depth and height. Despite this common ground, monocular depth estimation advantage from the only single-view image is a popular part with computer vision has more information and application methods than height estimation because of real-life utilization.

To overcome these issues, we focus on the way to estimate height from the former monocular depth estimation network. Moreover, we propose building adaptive loss function to denote a 3D map focused on building. To evaluate this, the US3D dataset (Urban Semantic 3D Dataset)[2] and Google Earth remote sensing images from the Google Earth application[3] demonstrate the effectiveness of the proposed method for the monocular height estimation. The proposed model achieves remarkable performance on both datasets.

II. Proposed method

A. Adaptive Loss function

In our experiment, we purpose to reconstruct a 3D map from a single-view remote sensing image. To do this, we extract height map from a deep learning architectured based on BTS network. The BTS network is an existing outperforming monodepth estimation network focusing on indoor and outdoor datasets. Additionally, we apply an adaptive loss function that substitutes the height estimation loss function[4] for the scale-invariant error loss function widely used for estimating monocular depth [5].

$$L = l_{height} + \delta\, l_{grad} + \mu\, l_{normal} \tag{1}$$

Equation (1) is an entire height estimation loss function, the parameters δ and μ are weight coefficients. Experimentally, these things are set to 1 fitted by the training process.

Also, each component is as follows.

$$l_{height} = \frac{1}{n}\sum_{n=1}^{i}(e_i), \qquad e_i = \left|\hat{H}_i - H_i\right| \tag{2}$$

Equation (2) is the absolute value of the difference between the predicted height and the ground truth height. L1 norm is used to increase the loss according to the height compared to the former method while being robust to outliers.

$$l_{grad} = \frac{1}{n}\sum_{n=1}^{i}\left(ln\left(\frac{\partial e_i}{\partial x}+1\right) + ln\left(\frac{\partial e_i}{\partial y}+1\right)\right) \tag{3}$$

Equation (3) is gradient information called spatial gradient. By using the Sobel filter operation, we calculate gradient spatial coordinate in (x,y). We use it to reduce the sudden change in the slope of the roof from same building and also help to recognize the edge information.

$$l_{normal} = \frac{1}{n}\sum_{n=1}^{i}\left(1 - \frac{\langle n_i^d, n_i^g \rangle}{\sqrt{\langle n_i^d, n_i^d \rangle}\sqrt{\langle n_i^g, n_i^g \rangle}}\right) \tag{4}$$

Equation (4) is a normal vector of information. n_i^d, n_i^g are respectively the normal vectors of the predicted height and the ground truth height. To use cosine similarity for measuring the similarity between the two vectors. ($\langle n_i^d, n_i^g \rangle$ is dot product)

B. Blurring

As the training process goes on, high elevation building causes some problems which do not look as similar as ground-truth view because it influenced by noise or and so on. To address these issues, we employ a blur filter like equation (6) to show flat rooftop images by reducing sudden changes in the slope of the roof influenced by noise.

$$l_{height} = \frac{1}{n}\sum_{n=1}^{i}(H_i), \quad H_i = \left|\hat{F}(P_i) - F(P_i)\right| \tag{5}$$

Equation (5) is the absolute value of the difference between the blurred predicted height and the blurred ground truth height which is similar to equation(2). $F(P_i)$ means blur filter function like equation (6).

$$F(P_i) = \frac{1}{k^2}\left(\sum_{j=1}^{k^2-1}(p_j) + p_i\right) \tag{6}$$

Equation (6) is a blur function. Each thing is meaning for k= kernel, p_j is the neighbor pixel of p_i. These things have effects that do look similar to ground-truth view in case of high elevation buildings.

In the point of reconstruction view, unlike the former depth map reconstruction method, the height map reconstruction is restored by raising the z-axis from the ground to the height, rather than using disparity. Furthermore, RGB images and AGL (Above Ground Level, Lidar data) are extracted and used in 512×512 (pixels) in Urban Semantic 3D Dataset (US3D, 2019).

III. Experiments

In our experiment, NVIDIA TITAN V 12GB GPU was used for model training US3D datasets which are 141,819 extracted images. Note that the US3D dataset provides three areas geographic tiles from Jacksonville, Florida Omaha, Atlanta Georgia in the United States. To verify the performance of the proposed method, we use cropped google earth image, Jacksonville area in the United States, which had similar domain as trained datasets, Fig. 1. Moreover, we also verify the performance of the proposed method without considering domain differences in Fig. 2, Phuket, and Fig. 3, Barcelona.

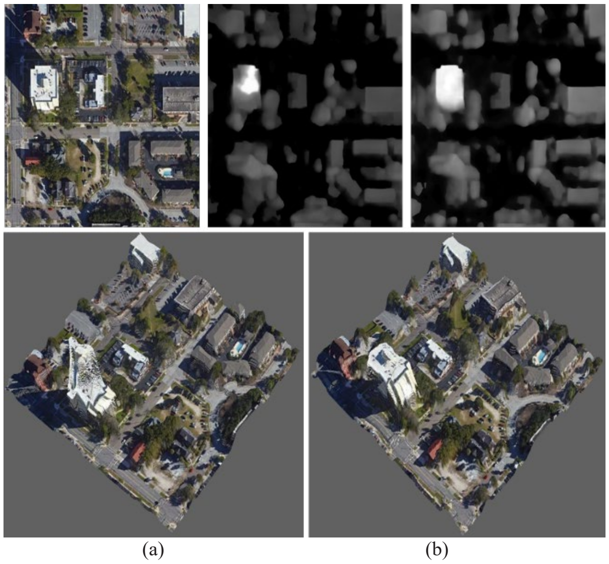

Fig. 1. Top: From left to right: remote sensing image, height map from height loss function and adaptive loss function.
Bottom: (a): 3D map reconstruction from height loss function, (b): 3D map reconstruction from adaptive loss function.

Fig.1 presents an example of comparing height loss function[4] and adaptive height loss function results by reconstructing 3D map from google earth cropped image, Jacksonville area in the United States. We observe that our reconstruction result presents less noise and succeeds in extracting the outlines of the high elevation buildings.

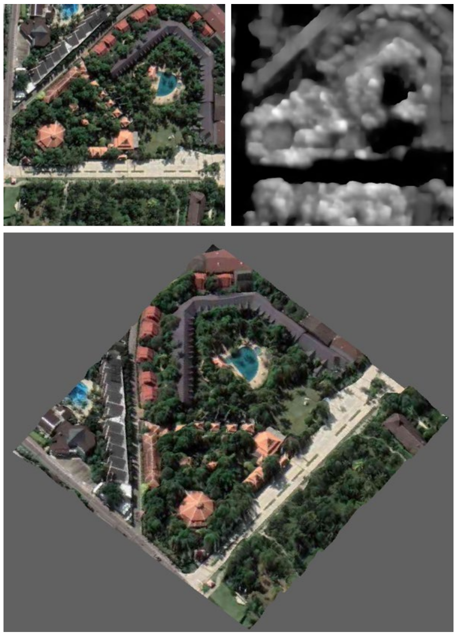

Fig. 2. Top: From left to right: remote sensing image, height map from height adaptive loss function.
Bottom: 3D map reconstruction from adaptive loss function

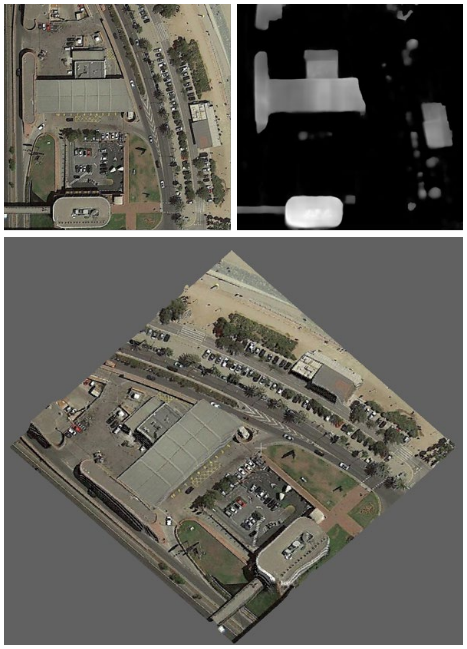

Fig. 3. Top: From left to right: remote sensing image, height map from height adaptive loss function.
Bottom: 3D map reconstruction from adaptive loss function

Fig.2, Fig.3 present examples of 3D map from google earth cropped image, Phuket area in Thailand and Barcelona area in Spain. We observe that it shows less accuracy than what we trained domains like Jacksonville. Nevertheless, it also looks similar to ground-truth view.

IV. Conclusions

In this experiment, we attempt to perform 3D map reconstruction on a different dataset from cropped google earth images. However, we observe that it is worse to work in a different view from what we trained view. In the future, we envisage looking into a way less influenced by domain differences and using refinement, for a higher level of accuracy wherein all domains. Furthermore, by applying computer vision methods, likes segmentation, 6DoF pose estimation, we will not only improve look similar to the ground-truth view but estimate accuracy likes as δ scores.

Acknowledgment

This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (No. 2021R1A6A1A03043144)

This work has supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2021R1A2C2009722)

References

[1] J. Lee, M. Han, D. Ko, and I. Suh. “From big to small: Multi-scale local planar guidance for monocular depth estimation.” arXiv preprint arXiv:1907.10326. 2019.

[2] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic Stereo for Incidental Satellite Images. In WACV, 2019.

[3] Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. 2017.

[4] C. Liu, V. A.Krylov, P. Kane, Geraldine Kavanagh and Rozenn Dahyot. “IM2ELEVATION: Building Height Estimation from Single-View Aerial Imagery”. (remote sensing). vol.12, no.17, pp. 2719,2020.

[5] D. Eigen, C. Puhrsch and R. Fergus. “Depth map prediction from a single image using a multi-scale deep network”. (Advances in neural information processing systems). vol.27, pp. 2366-2374,2014.