MONOCULAR DEPTH ESTIMATION OF GOOGLE EARTH IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

M. Najaf[1*], H. Arefi[1], H. Amini Amirkolaee[1], B. Farajelahi[1]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Iran

Commission IV, WG IV/3

KEY WORDS: Depth Estimation, Scene Recognition, Convolutional Neural Network, Height Image, Digital Elevation Model, Google Earth Image.

ABSTRACT:

Depth estimation from images is an important task using scene understanding and reconstruction. Recently, encoder-decoder type fully convolutional architectures have gained great success in the area of depth estimation. Depth extraction from aerial and satellite images is one of the important topics in photogrammetry and remote sensing. This is usually done using image pairs, or more than two images. Solving this problem using a single image is still a challenging problem and has not been completely solved. Several convolutional neural networks have been proposed to extract depth from a single image, which act as encoders and decoders. In this article, we use one of these networks, which has performed well for depth estimation, in order to extract height from aerial and satellite images. Our main goal is to investigate the performance of Google Earth satellite data in order to produce a digital elevation model. At first, we extracted the digital model of the target area using ISPRS benchmark data, then we did the same thing using Google Earth satellite images. The paper presents a convolutional neural network for computing a high-resolution depth map given a single RGB Google Earth image. The results show the proper performance of Google Earth satellite images for height extraction. We achieved values of 2.07 m and 0.36 m for the RMS and REL metrics, respectively, which are very comparable and acceptable to the values of 2.04 m and 0.39 m obtained from the ISPRS benchmark images.

1. INTRODUCTION

Depth estimation from images is one of the important topics in computer vision, remote sensing and photogrammetry, which has many applications, such as the preparation of a digital elevation model and maps, reconstruction, change detection, robotics and autonomous vehicle control. Depth estimation from images is usually terminated using stereo images or more than two images, and these methods are still the most accurate, but with the spread of deep learning methods, recent efforts have been made to estimate the depth using a single image. Regarding the preparation of the digital elevation model, which is the most important application of depth estimation from images in photogrammetry and remote sensing. In recent years, there have been attempts at preparing a digital elevation model using a single image based on deep learning and neural networks. Amini and Arefi (2019), proposed an architecture based on a deep convolutional neural network (CNN) in order to estimate the height values of a single Areal image (Amirkolaee and Arefi 2019). Also in another work, they proposed a CNN architecture for estimating the digital surface model (DSM) from a single airborne or spaceborne image (Amirkolaee and Arefi 2019). Amini and Arefi (2019), proposed a novel approach for 3D change detection in urban areas using only a single satellite image. Therefore, a dense convolutional neural network (DCNN) is utilized so as to estimate a digital surface model (DSM) from a single image. The changed areas are detected by subtracting the estimated DSMs (Amini Amirkolaee and Arefi 2019). Amini and Arefi (2021), proposed a different approach based on convolutional neural networks (CNNs) to generate a digital surface model (DSM) from a single high-resolution satellite image (Amini Amirkolaee and Arefi 2021). Alhashim and Wonka (2018), proposed a convolutional neural network for computing a high-resolution depth map given a single RGB image with the help of transfer learning (Alhashim and Wonka 2018). Farooq Bhat et al. (2021), presented a block based on the transform architecture, which divides the depth ranges into bins whose center value is estimated according to each image. Finally, the depth values are the results of the linear combination of the centers of these bins. (Bhat, Alhashim et al. 2021). In this study, we implemented a CNN-based network to estimate elevation values from single true orthophoto images and Google Earth satellite images. We employ an encoder-decoder network (Alhashim and Wonka 2018), where the encoding part is based on DenseNet-169 (Huang, Liu et al. 2017) and is used to prepare a depth map from a close-range single image. The remainder of this paper is organized as follows: Section 2 describes the proposed methodology. Section 3 includes the experimental and evaluation on both implementation in qualitative and quantitative aspects. Section 4 presents the final conclusion of depth estimation.

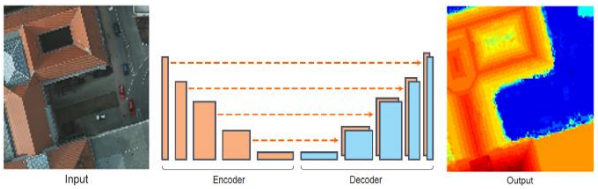

Figure 1. Overview of the proposed network architecture.

2. PROPOSED METHOD

In this section, we describe the method used to estimate the height map from a single RGB image. First, we describe an encoder-decoder architecture. In the following, we apply this architecture to our datasets for training, then we evaluate the results using the

training data set. In this study, we have used a simple encoder-decoder architecture with skip connections. The encoder part of this network is a pre-trained DenseNet-169 and has no changes. In the decoder part, we have basic blocks of convolution layers that are applied to the concatenation of 2×bilinear up sampling of the previous block, whose spatial size is similar to the size after applying up sampling in the encoder part. (Huang, Liu et al. 2017).

2.1 Methodology

The proposed methodology for estimating depth maps from DSMs is shown in this section (see Figure. 2). First, we have a data pre-processing, then we have data augmentation. After that the proposed CNN architecture which contains an encoder and decoder is used which is described in section 2.2.

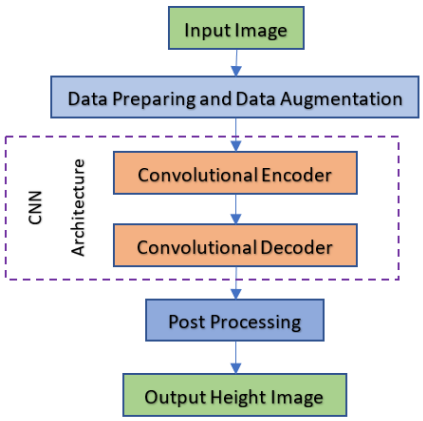

Figure 2. The flowchart of the proposed methodology.

2.2 Data Preparing and Data Augmentation

To prepare the training data, we must note that in each tile there must be various features, such as buildings and trees with specific shapes and sizes. For this purpose, all the data were resampled to a resolution of 40 cm, and the size of 128×128 was suitable to cover the complications in this resolution. For data augmentation, rotation and overlap have been used, also some data has been generated randomly, in such a way that a random point is selected in the overall image and, according to that, the training data with the size of 128×128 is produced.

2.3 The Proposed Network Architecture

The neural network contains both an encoder and decoder. In the encoder part, color images are encoded into a feature vector by using DenseNet-169 network. Then the decoder is used. The decoder includes skip-connections and up-sampling layers. The vector will be in a series of up-sampling layers so as to build the final depth map at half the input resolution. The decoder starts with a 1×1 convolutional layer with the same number of output channels as the output of the truncated encoder. Then the up-sampling blocks are used, which are composed of a 2×bilinear up-sampling followed by two 3×3 convolutional layers with output filters set to half the number of inputs filters, and where the first convolutional layer of the two is applied on the concatenation of the output of the previous layer and the pooling layer from the encoder having the same spatial dimension. The activation function is a leaky ReLU with a parameter of $\alpha = 0.2$ for each up-sampling block except for the last one. The input images are represented by their original colors in the range [0; 1]. This network doesn't consist of any Batch Normalization. Figure 1. shows the overview of the proposed network architecture (Alhashim and Wonka 2018), (Huang, Liu et al. 2017). Table. 1 illustrates the structure of our encoder-decoder with skip connections network.

Layer	Output	Function
INPUT	128×128×3	
CONV1	64×64×64	DenseNet CONV1
POOL1	32×32×64	DenseNet POOL1
POOL2	16×16×128	DenseNet POOL2
POOL3	8×8×256	DenseNet POOL3
…	…	…
CONV2	4×4×1664	Convolution 1×1 of DenseNet block4
UP1	8×8×1664	Upsample 2×2
CONCAT1	8×8×1920	Concatenate POOL3
UP1-CONVA	8×8×832	Convolution 3×3
UP1-CONVB	8×8×832	Convolution 3×3
UP2	16×16×832	Upsample 2×2
CONCAT2	16×16×960	Concatenate POOL2
UP2-CONVA	16×16×416	Convolution 3×3
UP2-CONVB	16×16×416	Convolution 3×3
UP3	32×32×416	Upsample 2×2
CONCAT2	32×32×480	Concatenate POOL1
UP2-CONVA	32×32×208	Convolution 3×3
UP2-CONVB	32×32×208	Convolution 3×3
UP4	64×64×208	Upsample 2×2
CONCAT4	64×64×272	Concatenate CONV1
UP2-CONVA	64×64×104	Convolution 3×3
UP2-CONVB	64×64×104	Convolution 3×3
CONV3	64×64×1	Convolution 3×3

Table 1. Network architecture.

2.4 Post Processing

Network outputs are 64×64 in size, and their values are local. As a post-processing, we first resample the image to 128×128 size, then make the values absolute by adding the minimum height value of that patch.

3. EXPERIMENTAL RESULTS

3.1 Dataset

In this work, we used the data of the city of Potsdam, which includes true ortho photos and DSM obtained from dense image matching with a resolution of 5 cm, and Google satellite images corresponding to this area were downloaded with a zoom level of 21, which its resolution was about 8 cm. This is a nominal resolution and cannot realistically expect such a resolution, and it is not comparable to the 5 cm resolution of the benchmark dataset. For this reason, both True Ortho Photo and Google satellite image data were resampled to 40 cm resolution. The training data has been prepared in the size of 128×128, which is comprehensive data with shapes and sizes. Since these data were collected in the winter season, the cover of tree crowns is less visible and there are almost no trees and, Correspondingly, there is no tree in the DSM data. Google Earth satellite images were prepared correspondingly in the same size. It should be noted that the Google data was related to another season and time in which the trees are known.

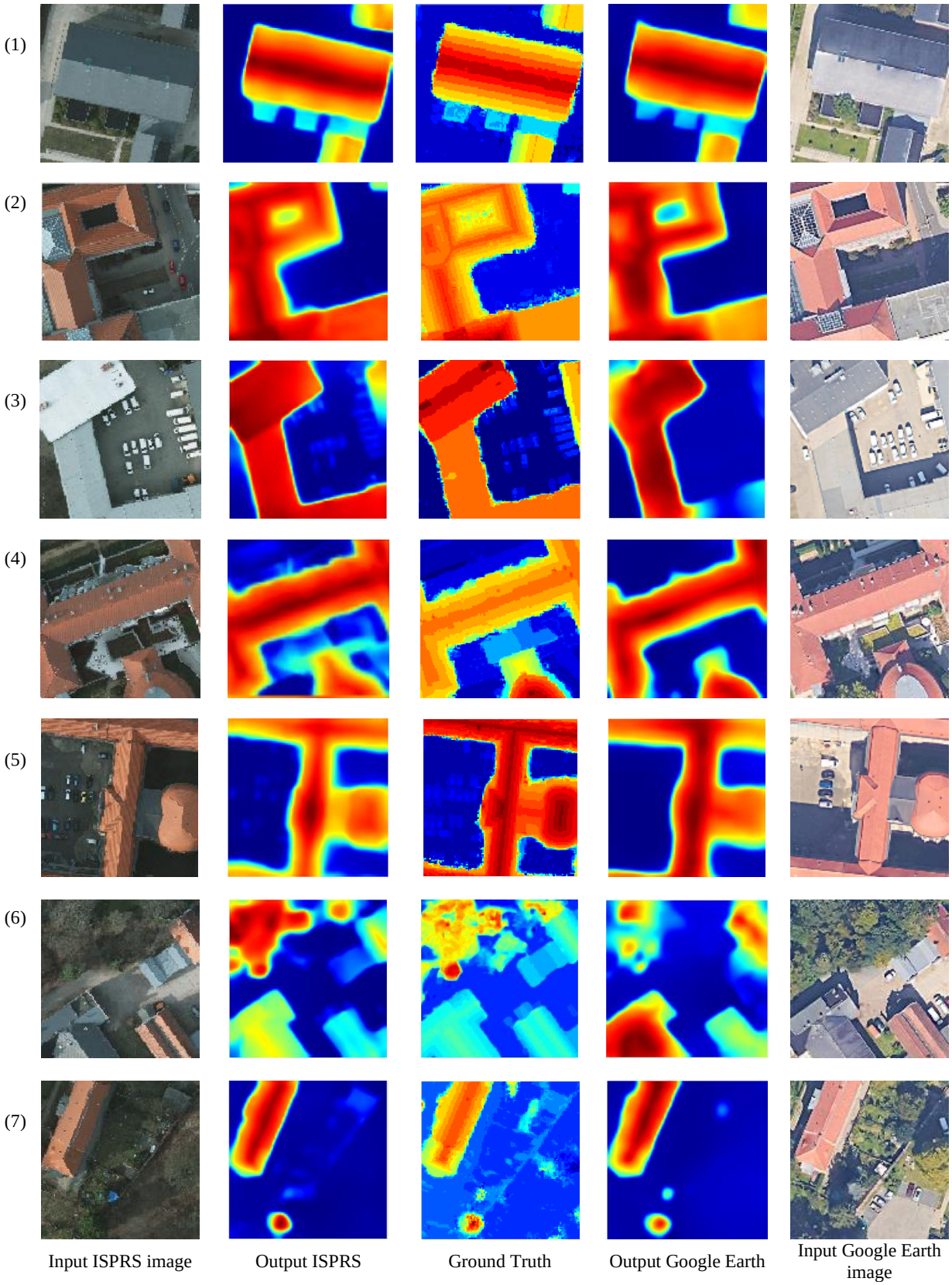

Figure 3. Qualitative comparison of the deep models on the ISPRS Potsdam dataset.

3.2 Evaluation

3.2.1 Qualitative Results

First, we observe the results of the network implementation using the Potsdam ISPRS benchmark dataset. Regarding the estimation of the height of the buildings, the results are very acceptable and in terms of shapes, sizes and borders, the buildings have been estimated with appropriate accuracy. As we mentioned in section 3.1, this dataset is related to the winter season and the absence of leaves on the trees in the true ortho photo data as well as the DSM data has caused fewer non-structural features (trees) to be extracted. Compared to the results of the proposed DenseNet (Amirkolaee and Arefi 2019), we achieved satisfactory results. But our main goal is to check the performance of Google Earth satellite images to estimate the depth of a single image. It can be seen in the figure (3) that Google Earth data has also shown a very suitable performance. The results of this data are very close to the results of the benchmark dataset on the Dense Depth (Alhashim and Wonka 2018) and the visual results seem to be better than the DCNN (Amirkolaee and Arefi 2019). The Google Earth images are taken in another season and its tree coverage is fully completed. On the other hand, the corresponding DSM images do not contain tree cover. As a result, the trees still did not appear in the height estimation results, since the trees did not exist in both trained datasets. These visual inspections indicate that the performance of the training network performed well in this study and it would have much better depth estimated results if the tree cover exists in both training datasets. The input training data contains 30K samples. The performance of the proposed method is evaluated in the benchmark dataset of Potsdam and its corresponding Google Earth satellite images .Before starting to interpret and compare the results of using the network on ISPRS benchmark images and Google Earth satellite images, we must pay attention to a few points.

1. The DSM used as training data is used for both ISPRS and Google Earth images, it is completely consistent with ISPRS data and is the result of Dense Image Matching of the images of this dataset.

2. The images of ISPRS and Google differ in terms of time, and in some details, such as the cover on the trees, due to the difference in the seasons of imaging, there are differences between the two images. But in terms of existing buildings and structures, the two images are completely identical and have the same georeferencing.

3. Google Earth images do not have any geometric corrections, and this makes us see tilt and height displacement in some parts, and in general, vertical geometry is not maintained throughout the image, and compared to ISPRS images that are true orthophotos, there is no perfect pixel correspondence.

4. The last point is that no radiometric correction has been applied to the Google Earth images, and some parts may have too much brightness or shadow in Google images due to sunlight.

In the following, we will analyze the obtained results by categorizing the visible scene. In figure 3 (1). where there is almost no other complication than the building, we see that the results are very acceptable. In this type of scene, almost all main parts, borders, shapes, and sizes are correctly extracted. In some parts where the size of the complications is small, especially at shallower depths, we can see that the boundaries are better extracted in the benchmark images. Also, in relation to the extracted height values, this style of the image has the best output results and the height difference with the ground truth in this style of image is about 1 meter.

In figure 3 (2). again, the main cover of the scene is the buildings, with the difference that these buildings are more complex. In this case, the boundary and shape of the whole building are well extracted from both the ISPRS benchmark dataset and Google Earth satellite images, and the difference is the estimated height values in these areas. The main reason for this issue is the color contrast difference between the images of these two datasets. In this case, the results of the ISPRS benchmark dataset are better because the DSM is completely corresponding to it, but this does not mean that the Google results are not suitable and these images also performed well.

In figure 3 (3). in addition to the building, there are also small objects such as cars in the scene. Also, these types of scenes have high levels of gray values and brightness, which, due to the higher contrast of Google Earth images, the brightness in the Google image is also higher. The results of the benchmark images, both in extracting buildings and extracting cars, have been very acceptable. In the Google Earth images, the cars are not extracted and there is a weakness in this field, but we should note that the brightness of the cars was very high due to direct sunlight. On the other hand, we can see that the buildings are generally extracted well from Google images, but in the part where the brightness was very high, we again have poor performance. But a very important and positive point can be seen here. In the lower part of the image, we can see that the sun is shining on the building in such a way that a relatively large shadow is created on the ground. The color and dimensions of the shadow are such that the network may make a mistake and recognize it as a complication, like a roof. But we can see that such an incident did not happen and the network was correctly recognized the shadow and even the shadow boundaries did not have any effect on the results, so it performed very well.

In figure 3 (4) and (5), scenes consisting of buildings with special shapes and cars can be seen. In this type of scene where there are a few more details, small details may not be extracted well, but large objects are extracted correctly, and both datasets performed suitably in this type of scene. But we still see that Google images did not perform acceptably in extracting cars.

And finally, in figure 3 (6) and (7), we have scenes that consist of different structures, such as buildings and trees. In this type of image, the buildings are not very large and there are also shadows in these images. In general, in this state of a scene where there are many contrasting differences throughout the image, the results are appropriate, but there are two weaknesses in the results of Google Earth images. First, it is related to small complications that are not extracted well, and secondly, due to high brightness in some parts, especially on the roofs that have neutral colors such as white and gray, despite the acceptable extraction of the borders, the height values have not been estimated correctly and the difference between the ground truth images might be up to 4 meters in these parts. But in general, in these types of scenes, the accuracy is between 2 and 3 meters.

Figure 4 and Figure 5 show the profiles of the obtained results and the ground truth image in two different examples. The red profile corresponds to the ground truth image, while the green profile represents the result of the ISPRS benchmark image and the blue profile corresponds to the result of the Google Earth satellite image. Due to the resampling of height data from 5 cm to 40 cm using the bilinear method, the amount of stepping is observed in the ground truth image.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W1-2022-589-2023 | © Author(s) 2023. CC BY 4.0 License.

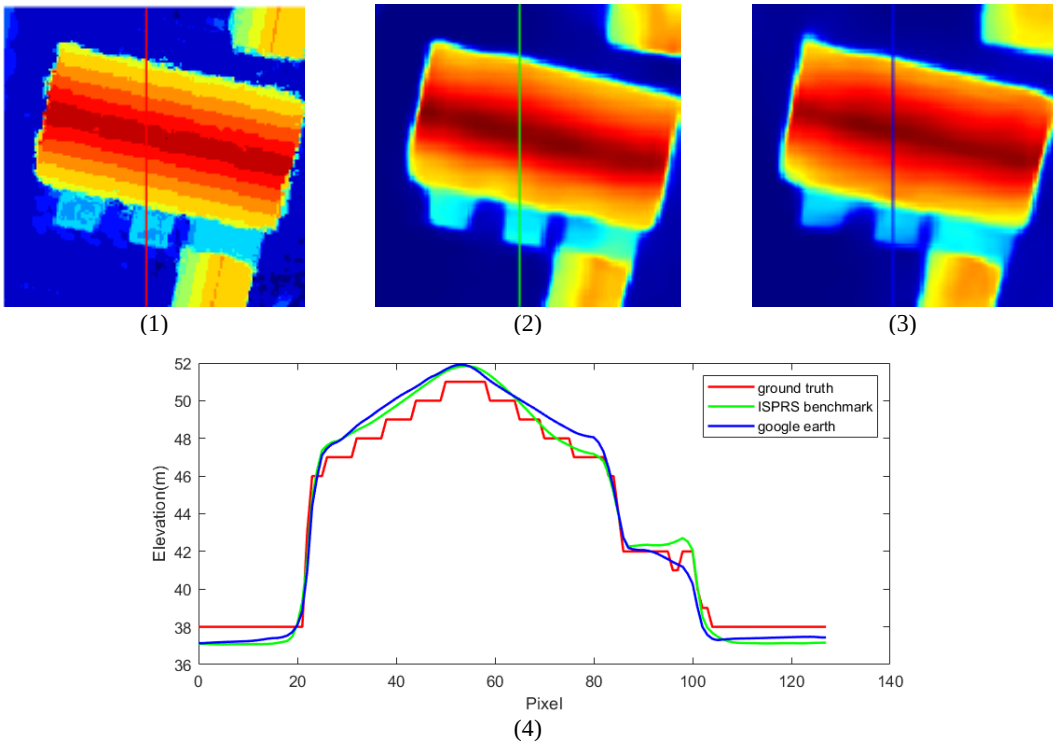

Figure 4. The comparative results in Example 1: (1) Ground-truth image, (2) ISPRS benchmark, (3) Google earth image, (4) The 3 different profiles.

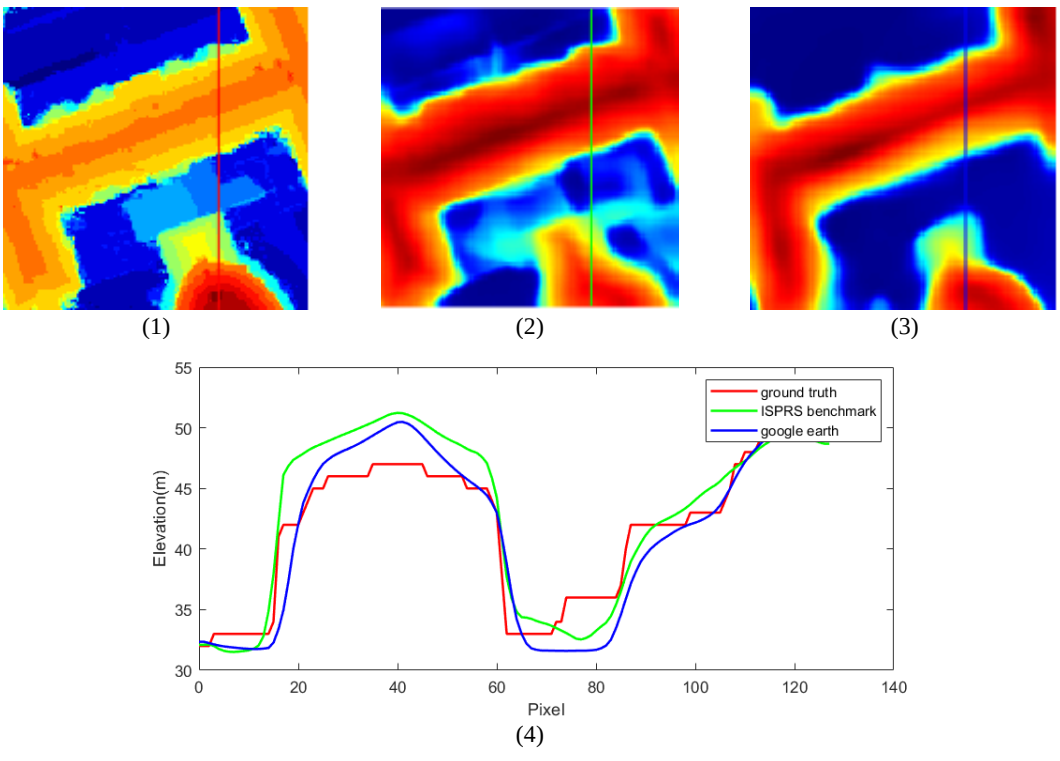

Figure 5. The comparative results in Example 2: (1) Ground-truth image, (2) ISPRS benchmark, (3) Google earth image, (4) The 3 different profiles.

However, the stepping and discontinuity are not present in the results of the proposed method, and both results are smooth and continuous and are considered suitable. In the building section, it can be seen that the profile of the proposed method is higher than the profile of the ground truth image. On the other hand, in the section related to the ground, the ground truth image profile is higher than the profile of the proposed method, which increases the error, and the height of the building itself will be affected by

these two differences. Generally, it can be seen that the results are very satisfactory and promising, and in terms of extracting the boundary and general trend of the scene, the results are completely suitable and reliable.

3.2.2 Quantitative Results

We completed the standard five metrics used in this study. In this part, some quantitative metrics are presented in order to distinguish the accuracy of the proposed height estimation approach using a single image. The DSM obtained from dense image matching with a resolution of 40 cm is employed as reference data and compared with the estimated corresponding DSM. In this regard, some various criteria are introduced, including the average relative error (Rel), RMSE.

$$Rel = \frac{1}{n}\sum \frac{|H_r - H_e|}{H_r} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum (H_r - H_e)^2} \quad (2)$$

$$\delta_i = \max\left(\frac{H_r}{H_e}, \frac{H_e}{H_r}\right) < 1.25^i, i \in \{1,2,3\} \quad (3)$$

where H_r = reference height pixel value
H_e = estimated height pixel value
n = number of pixels

Moreover, another criterion: other height values are close to the reference height values. The ratio of the estimated height value to the reference height values is calculated in three levels defined by δi (Eigen, Puhrsch et al. 2014). The criteria defined in Eqs. (1)-(2) shows that even if their values are lower, the performance is better. In contrast, Eq. (3) shows the accuracy of the proposed methodology.

Metrics	Dataset	Google Earth	DCNN
δ_1	0.5262	0.5482	0.342
δ_2	0.5887	0.6223	0.601
δ_3	0.6328	0.6605	0.782
Rel	0.3952	0.3628	0.571
RMS	2.0434	2.0768	3.468

Table 2. Evaluation metrics for depth estimation.

4. CONCLUSIONS

In this paper, we investigated the performance of high-resolution Google Earth satellite images for height extraction using a deep learning network. At first, we trained the network using ISPRS benchmark data, then we repeated this task using Google Earth satellite images to evaluate and compare the performance of these images in the application of height extraction. The obtained results indicate the proper and promising performance of these images. Although the DSM which is used for training did not exactly match the Google Earth images, the results were satisfactory, so we can expect better results if we have a more suitable DSM. It was also observed that the absence of vertical geometry in Google Earth images compared to vertical geometry in ISPRS benchmark images did not have much effect on the results. Regarding the radiometric effects, it was observed that this factor has a great influence and is very effective in the results. Therefore, by applying radiometric corrections, we can expect better results. Finally, we should announce that the results of using Google Earth images to extract the height are suitable and satisfactory. We achieved values of 2.07 m and 0.36 m for the RMS and REL metrics, respectively, which are very comparable and acceptable to the values of 2.04 m and 0.39 m obtained from the ISPRS benchmark images.

REFERENCES

Alhashim, I. and P. Wonka (2018). "High quality monocular depth estimation via transfer learning." arXiv preprint arXiv:1812.11941.

Amini Amirkolaee, H. and H. Arefi (2019). "3D CHANGE DETECTION IN URBAN AREAS BASED ON DCNN USING A SINGLE IMAGE." International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.

Amini Amirkolaee, H. and H. Arefi (2021). "Generating a highly detailed DSM from a single high-resolution satellite image and an SRTM elevation model." Remote Sensing Letters **12**(4): 335-344.

Amirkolaee, H. A. and H. Arefi (2019). "Convolutional neural network architecture for digital surface model estimation from single remote sensing image." Journal of Applied Remote Sensing **13**(1): 016522.

Amirkolaee, H. A. and H. Arefi (2019). "Height estimation from single aerial images using a deep convolutional encoder-decoder network." ISPRS journal of photogrammetry and remote sensing **149**: 50-66.

Bhat, S. F., et al. (2021). Adabins: Depth estimation using adaptive bins. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Eigen, D., et al. (2014). "Depth map prediction from a single image using a multi-scale deep network." Advances in neural information processing systems **27**.

Huang, G., et al. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition.