Olson, Alexander W, and Shoshanna Saxe. 2024. "Single-Image Building Height Estimation Using EfficientNet: A Simplified, Scalable Approach." *Findings*, May. https://doi.org/10.32866/001c.116609.

URBAN FINDINGS

Single-Image Building Height Estimation Using EfficientNet: A Simplified, Scalable Approach

Alexander W Olson[1], Shoshanna Saxe[2]

[1] Centre for Analytics and Artificial Intelligence Engineering, University of Toronto, [2] Department of Civil and Mineral Engineering, University of Toronto

Keywords: Urban Technology, Building Analysis, Neural Networks, Image-Based Estimation, Urban Data Processing

https://doi.org/10.32866/001c.116609

Findings

We present a novel approach for estimating building heights using single street-level images. The method employs EfficientNet, a state-of-the-art neural network, to eliminate the need for additional data like street maps. We compare this new method with existing techniques, focusing on accuracy evaluated through metrics like Mean Absolute Error (MAE). The model is pre-trained on the Cityscapes dataset and fine-tuned on images from Toronto's 3D Massing dataset. It demonstrates strong accuracy, with an MAE of 1.21 meters, outperforming traditional methods.

1. QUESTIONS

This article explores the task of building height estimation using street-level imagery. We investigate the potential to accurately estimate building heights using an approach that relies solely on individual street-level images. This challenges the conventional need for multiple data sources or the complete visibility of a building's top, which is common in existing methods.

Our proposed method aims to streamline the process by eliminating the need for additional data such as street map information or LiDAR. Instead, it focuses on utilizing a single image combined with a state-of-the-art neural network architecture, thereby reducing both cost and time requirements while enhancing accuracy.

A significant aspect of the study is the comparison of this novel method with existing techniques, specifically in terms of accuracy. The research includes a thorough error analysis, evaluating the performance compared to existing methods using metrics like Mean Absolute Error (MAE). This comparison is critical to understanding how the new approach stands against traditional methods under varying urban conditions and settings.

In summary, the research focuses on the viability of a simplified, single-image-based method for estimating building heights, its efficiency over current methodologies, and an in-depth analysis of its accuracy compared to existing approaches.

2. METHODS

Our training method involved the use of three datasets. The first, Cityscapes, was used to pre-train the image feature extractor on images of street scenes. For the main height prediction task, we combined images of buildings from Google StreetView with corresponding ground truth height data from the Toronto 3D Massing dataset.

2.1. Cityscapes (Pre-Training)

Cityscapes (Cordts et al. 2016) is a dataset of street scenes, created with the intention of aiding research in computer vision. The Cityscapes training set used includes 2,978 images of European cities. The contents of each image are labelled, with polygons denoting the region corresponding to each label. There are 30 possible labels, including 'person', 'building', and 'car'. Fifty cities are included in the dataset, primarily in Germany, and images were taken during the day across the year.

2.2. Toronto 3D Massing (Ground Truth)

Our ground truth height measurements are acquired from the City of Toronto's Open Data Portal. The city provides 3D models of 423,710 buildings, including height data derived from LiDAR for 410,355 of them. We randomly selected 5000 buildings from this dataset for use in our approach. To ensure that the sample accurately reflects the larger dataset, we compared the distribution of building heights. The sampled building heights exhibit a slightly lower mean (6.2m) and a smaller standard deviation (4.8m) compared to the full dataset (6.9m mean, 8.8m standard deviation). This difference is primarily attributable to a few very tall outlier buildings in the full dataset. Confirming this, when excluding the top 1% of heights in both datasets, the average height converges to 6.2m.

2.3. Google StreetView (Training and Evaluation)

Google StreetView's API was utilized to gather 5000 images of Toronto buildings for this research. The 3D Massing dataset provided the geographic coordinates for these buildings. However, due to script errors during image capture, 81 buildings were excluded, resulting in 4909 images. These errors occurred when the API could not locate the given addresses or provided irrelevant user-submitted images. Despite this reduction, the quality and accuracy of the dataset remained intact. The images, each with a resolution of 640x640 pixels, form the essential input for our building height estimation model.

3. Methodology

We apply a deep neural network directly to the task of estimating building heights. Our model is derived from EfficientNet (Tan and Le 2019, 2021), a convolutional network that achieved state-of-the-art performance on several standard image prediction tasks, despite an order of magnitude fewer tunable parameters than its predecessors. This greatly speeds up the process of training,

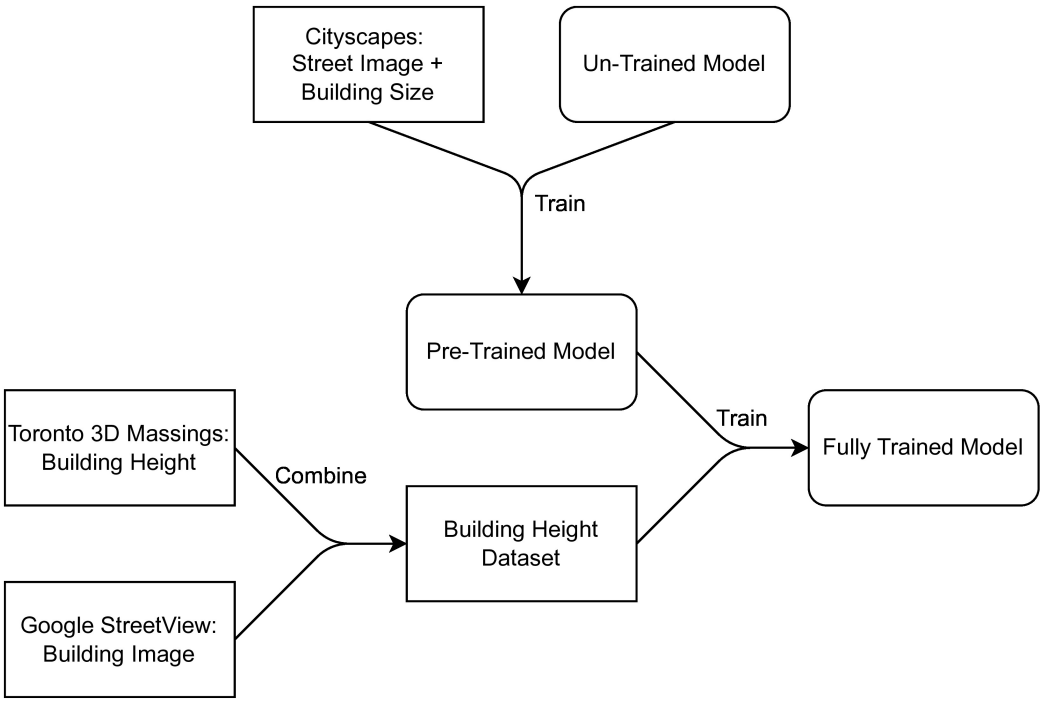

Figure 1. Diagram of the model training process.

and allows for strong performance on a comparatively smaller training set. We extend the PyTorch implementation of EfficientNetV2-S (i.e., small), with three additional fully-connected layers, which take the default 1000-dimensional output of the original network and reduce to a single prediction. Additional layers are required as the original architecture of the network is designed for a classification task, compared to our regression task.

We first pre-train our network on the Cityscapes (Cordts et al. 2016) dataset. While the labels provided with Cityscapes are intended for semantic segmentation, we want our network to ultimately predict a real value. In order to bring the Cityscapes task more in line with this, we calculate for each street-level image in the dataset the area of the image which contains a building.

For building height estimation, we use a Mean Squared Error (MSE) loss to train the model due to its sensitivity to outliers and smooth gradients. We monitor performance on the validation data using MAE for its interpretability and robustness to outliers. We split our building images into training, validation and testing sets with a ratio of 70:10:20, comprising 3439, 491, and 979 images respectively. During training, data augmentation techniques were employed to marginally alter the input images, to improve the robustness of prediction. Images could be flipped horizontally, rotated within 10 degrees of the original, and/or have their brightness slightly adjusted.

4. FINDINGS

Our model, leveraging the EfficientNet architecture, has shown strong accuracy in estimating heights from single street-level images. Using Google StreetView images as our testing set, and ground truth heights derived from

Table 1. Summary of method performance. SF stands for San Francisco.

Approach	Images	Method	City	MAE (m)	Estimates within 2m
Díaz and Arguello 2016	100	Single View Metrology	Madrid	2.48	–
Yuan and Cheriyadat 2016	400	Camera Projection	SF	–	67%
Al-Habashna 2021	20	Contour Processing	Ottawa	2.32	50%
Yan and Huang 2022	236	Deep Learning	SF	1.62	74%
Proposed	979	Deep Learning	Toronto	1.21	86%

the Toronto 3D Massing dataset, we achieved a Mean Absolute Error (MAE) of 1.21 meters. This level of precision is higher than comparable methods (Table 1), and noteworthy considering the complexity and variability inherent in urban landscapes.

A critical factor in the success of our model was the incorporation of pre-training. The impact of this pre-training is evident when comparing the model's performance with and without it. Without pre-training, the model achieved an MAE of 1.46 meters, on buildings ranging in height from 0.5 meters to 73.4 meters, demonstrating that pre-training improved accuracy by 17%.

The primary advantage of our method is its simplicity and efficiency. Unlike other methods that require multiple data sources or complex pre-processing steps, our model relies solely on a single image. This streamlined approach reduces the time and resources needed for data collection and processing, making it more accessible and scalable for various applications.

Another significant advantage is the robustness of our model against partial occlusions and varied architectural styles. The model's ability to accurately estimate building heights even when the entire structure is not visible in the image represents a substantial improvement over existing methods. This robustness is particularly beneficial in densely built urban areas, where complete views of buildings are often obstructed, and is essential for applying this method to diverse urban landscapes.

In conclusion, our study presents a novel and efficient approach to estimating building heights using deep learning and single street-level images, outperforming existing published methods for the same task. The method's accuracy, simplicity, and scalability hold significant potential for applications in urban studies, urban planning, and related fields.

ACKNOWLEDGEMENTS

This research was funded by the Low-Carbon Renewable Materials Center and the Centre for the Sustainable Built Environment (CSBE) both at the University of Toronto. CSBE in turn is funded by an NSERC Alliance Grant (ALLRP 582941 – 23), the Climate Positive Energy Initiative and the School

of Cities at the University of Toronto and 12 industry partners (Colliers; the Cement Association of Canada; Chandos Construction; Mattamy Homes; Northcrest; Pomerleau; Purpose Building, Inc.; ZGF Architects; Arup; SvN Architects + Planners; Entuitive; and KPMB Architects).

Submitted: December 07, 2023 AEST, Accepted: April 15, 2024 AEST

REFERENCES

Al-Habashna, Ala'a. 2021. "Building Height Estimation Using Street-View Images, Deep-Learning, Contour Processing, and Geospatial Data." *2021 18th Conference on Robots and Vision (CRV)*, May, 103–10. https://doi.org/10.1109/crv52889.2021.00022.

Cordts, Marius, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. "The Cityscapes Dataset for Semantic Urban Scene Understanding." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–23.

Díaz, Elkin, and Henry Arguello. 2016. "An Algorithm to Estimate Building Heights from Google Street-View Imagery Using Single View Metrology across a Representational State Transfer System." *SPIE Proceedings* 9868 (May): 98680A. https://doi.org/10.1117/12.2224312.

Tan, Mingxing, and Quoc V. Le. 2019. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." *36th International Conference on Machine Learning, ICML 2019* 2019-June (May): 10691–700. https://arxiv.org/abs/1905.11946v5.

———. 2021. "EfficientNetV2: Smaller Models and Faster Training." *Proceedings of Machine Learning Research* 139 (April): 10096–106. https://arxiv.org/abs/2104.00298v3.

Yan, Yizhen, and Bo Huang. 2022. "Estimation of Building Height Using a Single Street View Image via Deep Neural Networks." *ISPRS Journal of Photogrammetry and Remote Sensing* 192 (October): 83–98. https://doi.org/10.1016/j.isprsjprs.2022.08.006.

Yuan, Jiangye, and Anil M. Cheriyadat. 2016. "Combining Maps and Street Level Images for Building Height and Facade Estimation." *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics, UrbanGIS 2016*, October. https://doi.org/10.1145/3007540.3007548.