18TH WORLD CONFERENCE ON EARTHQUAKE ENGINEERING
WCEE2024
MILAN

MILAN, ITALY
30TH JUNE - 5TH JULY 2024
www.wcee2024.it

AUTOMATIC BUILDING HEIGHT ESTIMATION: MACHINE LEARNING MODELS FOR URBAN IMAGE ANALYSIS

M. Ureña-Pliego[1], R. Martínez-Marín[1], B. González-Rodrigo[2], B. Benito-Oterino[3] & M. Marchamalo-Sacristán[1]

[1] School of Civil Engineering. Universidad Politécnica de Madrid, Madrid, Spain, miguel.urena@upm.es

[2] School of Forest Engineering and Natural Resources. Universidad Politécnica de Madrid, Madrid, Spain

[3] School of Surveying, Geodesy and Cartography Engineering. Universidad Politécnica de Madrid, Madrid, Spain

Abstract*: Artificial intelligence (AI) is triggering major advances in construction engineering in the era of the generalization of Building Information Modelling (BIM). One of the main areas of AI development is the determination of building attributes from street view images. Building height is an essential factor for the assessment of structural vulnerability that is often missing, although sometimes available in cadastral databases. The goal of this research is to improve the automation of the collection of building attributes for seismic exposure assessment, for future applications in Central American cities. This work presents the framework and the development of machine learning techniques for urban image analysis and an application to the estimation of building heights. In this study, building heights were calculated from street-view imagery based on a semantic segmentation machine learning model. The model has a fully convolutional architecture and is based on the HRNet encoder and ResNexts depth separable convolutions, achieving fast runtime and state-of-the-art results on standard semantic segmentation tasks. The model was built from scratch and trained using the Julia programming language, the Flux library and the Cesvima Magerit supercomputer. It was possible to estimate the average height of the buildings in pilot images with a maximum error of 3 meters, under conditions that are versatile, and which allow for easy matching without the need for point clouds or any kind of image depth estimation. The image segmentation results are promising, achieving a high performance in the Citiscapes fine val set. The proposed method works without the need for extra parameters, allowing for the use of images from the Internet or images captured by mobile phones. However, the model still needs improvements in order to function adequately for non-European cities due to the lack of suitable datasets. Further research alternatives are commented, as well as the difficulties of obtaining valuable training data to apply these models in countries with no training datasets and different urbanism conditions. This line of research contributes to the characterization of buildings and the estimation of attributes essential for the assessment of seismic risk using automatically processed street view imagery.*

1. Introduction

The construction industry's ongoing digital transformation holds the promise of introducing groundbreaking technologies and tools that can usher in novel business models, materials, and solutions, ultimately benefiting the entire industry value chain (European Construction Sector Observatory, 2021). Within the realm of digital technologies in construction, artificial intelligence finds its place in the category of data information and analysis (Baldini, G. et al., 2019). The integration of artificial intelligence into construction practices has thus far been constrained primarily to pilot projects, with a primary focus on enhancing structural analysis, design, and optimization. These efforts have yielded promising results, particularly in the deployment of artificial neural networks for tasks like structural damage assessment (e.g., post-earthquake damage detection) and structural health monitoring (e.g., identifying damage and nonlinearities in wind turbine blades through pattern recognition techniques).

Over recent decades, both industry managers and researchers have grappled with the formidable challenge of devising tools for evaluating the seismic vulnerability of existing building inventories, with the overarching goal of formulating dependable risk mitigation strategies. Various large-scale methodologies have been put forth to identify the most vulnerable segments within building stocks, following established protocols (Baggio et al., 2007; Brzev et al., 2013). Notably, urban image analysis has emerged as a pivotal approach for estimating the seismic exposure of buildings—a critical step in the calculation of their seismic vulnerability. Consequently, there is a clear imperative for the development of advanced automated tools capable of rapidly assessing the vulnerability of existing structures based on visual data, exemplified by the innovative VULMA tool designed for assessing the vulnerability of individual buildings and conducting comprehensive seismic risk assessments (Ruggieri et al., 2021). This avenue of research underscores the importance of creating dedicated training datasets, as proposed by Cardellicchio et al. (Cardellicchio et al., 2022).

The assessment of seismic exposure hinges on the meticulous characterization of structural attributes at varying levels of granularity to categorize each building's exposure class (Esquivel-Salas et al., 2022), a framework established by the Global Earthquake Model (GEM) foundation (Brzev et al., 2013). Previous investigations have elucidated that one of the most influential attributes for seismic exposure classification is a building's height or the number of floors (Esquivel-Salas et al., 2022; Rodríguez-Saiz et al., 2022). Recognizing the indispensability of this attribute, an estimation of building height with a maximum absolute error of one floor suffices. To fulfill this need, automated machine learning models are indispensable, especially in cities where pertinent data is scarce, as is the case in the capitals of Central American nations, known for their susceptibility to significant seismic hazards and high levels of physical and social vulnerability (Benito et al., 2012).

In many developing cities, street view images are abundantly available and can be obtained cost-effectively. Consequently, there exists a pressing demand for the development of models adept at extracting building heights from single-view images. Such models would enable researchers to obtain this pivotal parameter in developing urban centres using readily accessible internet imagery or smartphone-captured photos, facilitating the creation of indispensable databases for the evaluation of seismic exposure in urban environments.

The principal objective of this research endeavour is to craft cost-effective methodologies for collecting building height data, with a specific focus on seismic risk assessment in Central American cities situated in developing nations. This study makes a substantial contribution to the field by harnessing street view images processed through a diverse array of convolutional neural networks to deliver precise estimates of key seismic exposure attributes, capitalizing on freely available visual data. This approach can be seamlessly applied to monitor extensive urban regions in developing countries.

Furthermore, this research innovatively repurposes technology initially designed for autonomous vehicle navigation to the realm of building data collection and automated building surveys. It introduces a novel model that combines RestNext and HRNet, accompanied by a method for estimating building heights through semantic segmentation and a single street view image.

2. Building height in the evaluation of seismic exposure

Research into the seismic vulnerability of buildings, involving the use of imagery to extract fundamental parameters like height and facade material, is experiencing global growth (Esquivel-Salas et al., 2022). Traditional surveys have conventionally relied on manual visual inspections, a process known for its time-consuming and costly nature. There's a current trend shifting towards automating these procedures, exemplified by the recent VULMA model, which employs a machine learning model for categorical classification to automate surveys (Cardellicchio et al., 2023). Previous experiences involved capturing images of individual building facades and applying the necessary labels through a classifier model. However, in real-world scenarios using street view images, multiple buildings and additional elements such as streets are present, rendering the classifier model ineffective, as it cannot isolate the specific building of interest.

The INSPIRE scheme (INSPIRE, 2013) defines a building as a permanent structure, both above and below ground, intended for accommodating people, animals, goods, services, or production. Within this definition, one or multiple BuildingParts are identified, each denoting a distinct area within a cadastral parcel with consistent volume, either above or below ground. BuildingPart elements encompass attributes related to height, including the number of above-ground floors, height below ground level in meters, and the count of below-ground floors (INSPIRE, 2013).

While building heights are typically accessible through INSPIRE in Europe, new methodologies have been developed to automatically estimate building heights, especially to facilitate data collection for urban cadastres in cities lacking such information. Leveraging ESA's Sentinel 1 and 2 data series, particularly Synthetic Aperture Radar (SAR) data (Drusch et al., 2012; Frantz et al., 2021), enabled the automated extraction of building height data on a large scale (Frantz et al., 2021). This approach yielded an error rate of under 5 meters for average-height buildings but exhibited larger errors for taller structures. Further enhancements in satellite techniques involved multi-view satellite imagery (Cao and Huang, 2021), achieving an error rate of approximately 2 meters. Aerial imagery (Xiao, Gerke, and Vosselman, 2012) and Aerial Laser Imaging Detection and Ranging (LIDAR) have been employed to measure building heights (Bonczak and Kontokosta, 2019), but these methods are costly and require manual labour, raising questions about their suitability for large-scale data collection. More recently, methods utilizing street-view imagery, captured from a driver's perspective, have emerged. These approaches employ machine learning techniques to segment images and identify the top corners of buildings. Horizontal distances from the camera to the building are then calculated using camera positions and databases like OpenStreetMap, which contain x and y coordinates of building corners (Ala, 2020) and rooflines (Zhao, Qi, and Zhang, 2019). By considering parameters such as camera height, estimation of three vanishing points, and detecting vertical lines of buildings through semantic segmentation, height can be computed based on image perspective (Yan and Huang, 2022), yielding a relative error of 5%. Alternatively, height can be determined using trigonometric calculations. Spherical images allow for more complex geometric methods that do not require 2D external corner coordinates (Díaz and Arguello, 2016).

In the realm of computer vision, significant strides have been made in recent years in image classification tasks through the development of new machine learning models such as the vision transformer (Dosovitskiy et al., 2020) or its semantic segmentation version (Strudel et al., 2021). These models, characterized by their extensive parameterization, are typically developed by large companies and pose challenges in terms of training. Conversely, simpler models have emerged for image classification, such as RestNexts (Xie et al., 2016; Hitawala, 2018). Remarkable results in semantic segmentation have been achieved with HRNets (Wang et al., 2019), a model capable of processing images at multiple resolutions. Although computer vision is primarily employed in urban contexts for tasks related to autonomous vehicle navigation, especially using publicly available datasets like Cityscapes or Vistas (Cordts et al., 2016; Neuhold et al., 2017), its application in other domains such as building inspection or data collection has remained relatively underdeveloped.

3. Methods: Artificial Intelligence Applied to Building Height Estimation

3.1 Facade Detection: Semantic Segmentation Model

The initial phase of any image-based data collection tool involves the identification of spatial and categorical attributes pertaining to specific objects within the input images. In our context, these input images consist of street view perspectives from a car driver's viewpoint, from which we extract spatial details regarding buildings, vehicles, road surfaces, and other objects. This task is executed using a machine learning model for semantic segmentation, which assigns a class to each pixel within the input image.

We have developed a fully convolutional neural network that is constructed based on the HRNet encoder framework proposed by Wang et al. in 2019. This HRNet-based model is designed to preserve a high-resolution representation of the input image throughout the entirety of the model's processing. To optimize runtime efficiency, the output labelled image possesses half the resolution of the input.

In the HRNet encoder (Figure 1), the input image is split into two branches: one maintains half of the original resolution, while the other retains just one sixteenth of the original resolution. The second branch serves as a bottleneck, while the first branch acts as a high-resolution representation of the image. The process of reducing image resolution occurs progressively through the encoder, with the second branch's resolution halved in each encoder layer. Both branches are interconnected through cross-convolutions (Figure 2).

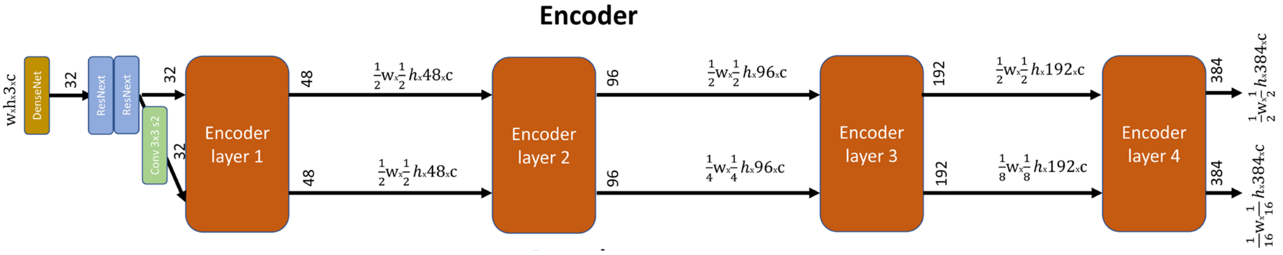


Figure 1. Model encoder. (Ureña-Pliego et al., 2023)

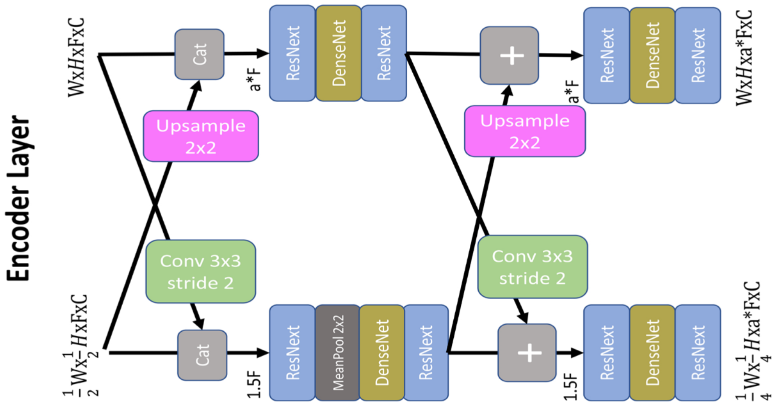


*Figure 2. Encoder layer. The first cross-convolution is not applied, and mean pool is applied on branch 1 in the encoder layer 1. × denotes array dimension and * denotes multiplication. (Ureña-Pliego et al., 2023)*

Within the encoder architecture, we employ ResNext convolutions (Figure 2), following which an atrous spatial pyramid convolution (Figure 4) is implemented. ResNext represents a recently introduced model that utilizes depth separable convolutions (Chollet, 2016) with an inverse bottleneck design. Remarkably, it achieves top-tier results while requiring fewer training parameters. DenseNet layers, on the other hand, are layers that augment the feature dimension by repeatedly concatenating convolution outputs to the input, a process that is iterated multiple times (Huang et al., 2016).

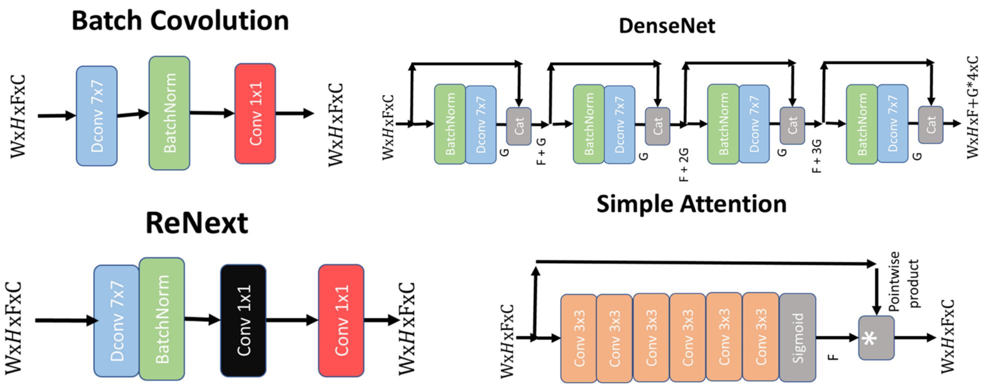


*Figure 3. Basic layers of the model. × denotes array dimension and * denotes multiplication. (Ureña-Pliego et al., 2023)*

Within the decoder architecture (Figure 5), we employ straightforward depthwise convolutions, without the inclusion of a bottleneck structure. Following this, an attention mechanism is applied (Figure 3), where skip connections originating from layers 5, 4, and 3 are integrated with decoder layers 1, 2, and 3, respectively. This integration is achieved by concatenating along the feature dimension of each branch.

In each decoder layer, the second branch undergoes upsampling until it matches the size of the first branch in the final decoder layer. Just before both branches are fused together, an attention mechanism is utilized. This attention mechanism assigns a level of attention to each pixel, determining its importance and serving as a criterion for merging two images into a unified output.

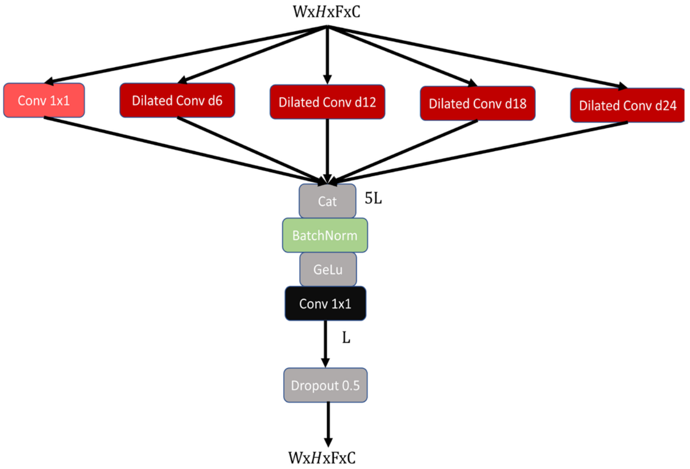


Figure 4. Atrous spatial pyramid layer. (Ureña-Pliego et al., 2023)

The model is programmed in the Julia programming language (Bezanson *et al.*, 2014) using Zygote (Innes *et al.*, 2019) and Flux (Innes *et al.*, 2018) libraries and cross-entropy (Zhang and Sabuncu, 2018) as the loss function for classification tasks. Optimisation is done with supervised learning using the ADAM optimizer (Kingma and Ba, 2014). Training is performed with a A100 Nvidia GPU on a Magerit supercomputer at Cesvima UPM. For model parameter initialization, a glorot uniform distribution (Glorot and Bengio, 2010) is used.

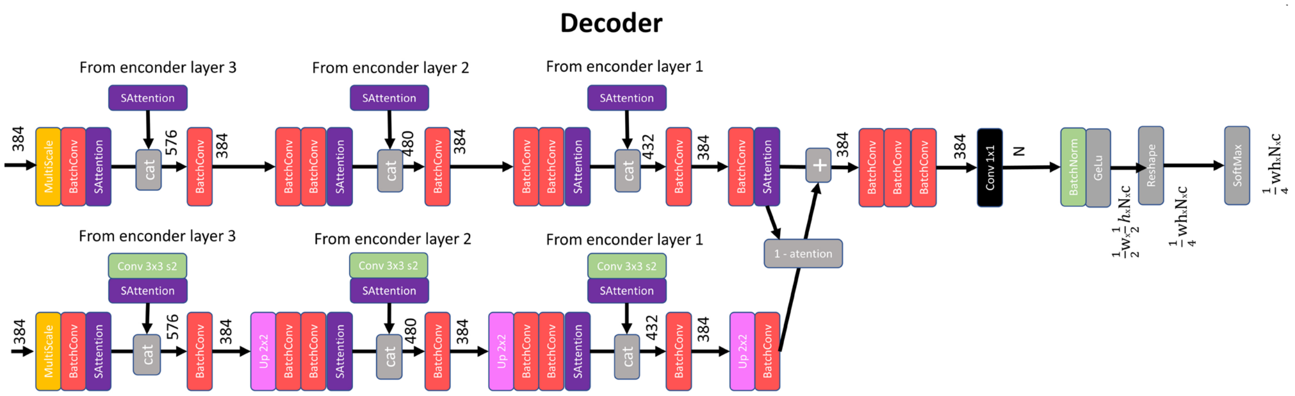


*Figure 5. Model decoder. × denotes array dimension and * denotes multiplication. (Ureña-Pliego et al., 2023)*

3.2 Extracting Spatial Information

This model offers 2D relative spatial information derived from the image. However, our goal is to acquire 3D real-world spatial information, enabling direct measurement of building height. This necessitates the generation of a point cloud from the input image. There exist machine learning models designed to provide pixelwise depth estimation from single-view images. With access to the depthmap of the image and camera data such as the focal length, we could assign distance and direction to each pixel in the image, thus forming the point cloud. Regrettably, the precision of single-view depth estimation models developed thus far (Li et al., 2022) falls short of the requirements for this task. In benchmarking tests like KITTI, even the best-performing models exhibit an error rate of approximately 8%. To put this into perspective, it translates to an 8-meter error in depth estimation for a building situated 100 meters away from the camera.

We conducted experiments using depth models trained on KITTI data, but the resulting point cloud proved to be inaccurate. Since generating a point cloud from single-view images proved unfeasible, we have pursued two alternative approaches. Firstly, a point cloud can be constructed through photogrammetry using multiple images or a video source (Müller et al., 2022). However, this approach comes with a significant computational burden and reduces the versatility of our method, as obtaining multiple images with the necessary characteristics from a video can be challenging. The second option is to devise a method that doesn't rely on a point cloud, which we have found to be the most practical approach.

3.3 Height Comparison Method

In this section, we derive building height estimates using 2D information extracted from the image. Within a given street scene, certain objects possess known heights, such as cars, people, or elements like windows and doors on buildings. By comparing the size of a building with the size of a reference class, such as the aforementioned objects, we can deduce the building's height. Ideally, windows and doors would serve as our preferred reference class; however, regrettably, these labels are absent from the training dataset. To utilize windows and doors as a reference class, the creation of our own semantic segmentation dataset would be necessary, a task that exceeds the scope of this research. Consequently, we opt to use cars and people as the reference classes, although cars only serve as a reliable reference if they are parked in close proximity to the building under measurement.

To compute the local building height, we calculate the quotient for each column of pixels belonging to a building in the segmented image, dividing it by the pixels associated with one of the reference classes. This result is then multiplied by the height of the chosen reference class. The outcome of this computation is a vector with a length equivalent to the horizontal dimension of the image.

It's crucial to note that the accuracy of this method hinges on the proximity of the reference object to the building. When the reference object is not adjacent to the building, it will appear larger in the image, rendering the measurement invalid. Such instances may arise in non-standard street geometries, like intersections, or when there are moving vehicles or pedestrians crossing the street. Additionally, reference objects with atypical sizes can also lead to inaccurate measurements.

To mitigate the impact of invalid data, we calculate the mean and standard deviation of the local height vector. Data points that deviate significantly from the mean value are then removed from the vector. Invalid measurements often result in heights that are underestimated, as inappropriately chosen reference objects tend to appear larger in most cases. Consequently, positive deviations from the mean of the height vector are given less tolerance than negative deviations.

In scenarios where multiple images and reference classes are employed, further statistical computations can be performed to enhance the accuracy of the height estimation.

3.4 Model Training Datasets

In the realm of semantic segmentation, the majority of street-view image datasets are primarily tailored for tasks related to autonomous driving. These datasets typically encompass classes like buildings, sky, and roads, all of which are relevant to our research. However, valuable classes such as windows, doors, building materials, and others are often absent from these datasets, as they hold no significance in the context of self-driving vehicles. To initiate a broad but low-precision training phase, we leveraged three datasets: Vistas (Neuhold et al., 2017), GTA5 (Richter et al., 2016), and Cityscapes coarse (Cordts et al., 2016).

The Vistas dataset boasts approximately 25,000 images from various countries around the world, making it unique in its inclusion of images from Central America. GTA5, on the other hand, comprises around 25,000 images extracted from the GTA5 video game. Labels for this dataset were automatically generated from in-game images depicting the city of Los Angeles, resulting in varying lighting conditions, including night scenes. Lastly, Cityscapes is a dataset composed of images from German cities, featuring approximately 18,000 coarsely labelled images and 5,000 finely labelled images. We utilized the coarse dataset from Cityscapes for our initial broad training, reserving the fine dataset for a subsequent refinement phase. This training approach allowed us to expose the model to a more diverse range of data, helping to mitigate issues often associated with real-world data.

The model underwent an initial training phase with the 18,000 images from the Cityscapes coarse dataset, followed by the incorporation of 25,000 images from Vistas (with label adaptation to match the Cityscapes scheme), 25,000 images from GTA5, and finally, 5,000 images from the Cityscapes fine dataset. To augment the dataset, we applied techniques such as 10-degree rotations and horizontal flipping.

4. Case Study: Extraction of Building Height in Cityscapes

4.1 Results

The semantic segmentation model achieved a notable IoU metric (intersection-over-union) of 86.13% on the Cityscapes fine validation set (Cordts et al., 2016). To elucidate these results, we present and discuss a case study utilizing a Cityscapes image. An image from the Cityscapes validation set underwent processing by the model, representing it as a (width × height × 3) matrix with values ranging between 0 and 1. This transformation yielded a (classes × width × height × 1) matrix with values between 0 and 1, signifying the model's likelihood assessment for each pixel belonging to a specific class. In this context, three classes are considered: sky, building, and car.

To facilitate analysis, the vector is reshaped to dimensions (width × height × 1), effectively assigning an integer to represent the class with the highest probability for each pixel. Subsequently, the height vector is extracted (Figure 6) by comparing pixels within the building and car classes for each row, after excluding columns devoid of pixels belonging to the sky class. In this illustrative example, after refining the height vector, the building heights span from 3.5 to 6.5, with an average of around 5. This signifies that, on this particular street, the buildings are approximately 5 floors high.

The car's height used for this purpose would be calibrated based on measurements from the city, encompassing buildings of known heights. It's important to note that this approach requires a consistent street alignment and an adequate portion of visible sky within the image. As street view images are both cost-effective and easily accessible, any erroneous measurements stemming from images with moving cars or non-uniform street alignments could be statistically filtered out.

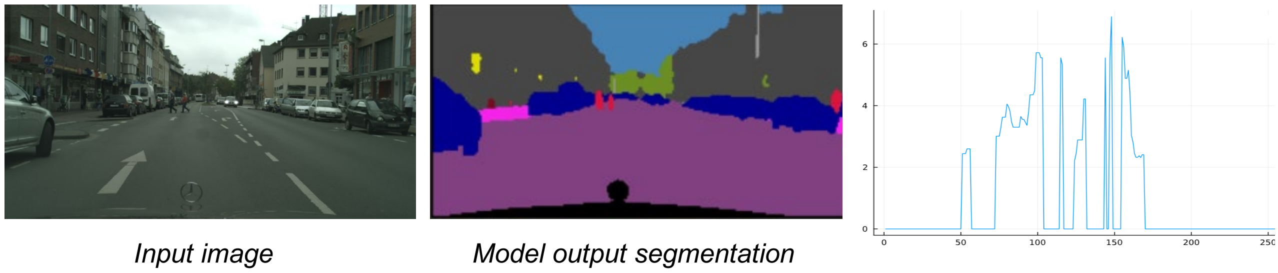
In practice, this method has the potential to yield average building heights on a street with a maximum error of 3 meters, equivalent to approximately one floor, and a relative error of 9% in cities characterized by uniform streets.

4.2 Discussion

The results of the image segmentation exhibit promise, demonstrating exceptional performance in the Cityscapes fine validation set (Cordts et al., 2016). This paves the way for further advancements in automating the detection of buildings and their constituent parts and features.

In the context of the case study, the outcomes are gratifying for estimating the "height" attribute associated with building exposure, denoting the elevation above ground in terms of storeys (e.g., a building's height equivalent to three storeys). Typically, height classes are defined within a range of storeys or floors, such as 1–3, 4–7, and > 8 (FEMA, 2001; Brzev et al., 2013). Consequently, the determination of seismic exposure categories can be carried out with building height estimates precise to a single floor. The proposed method effectively meets this requirement, exhibiting a maximum error of approximately one floor.

These results are comparable with those of other methods, such as those using satellite imagery with absolute errors of around 2–5 m, which usually require SAR data (Frantz *et al.*, 2021) or multi-view imagery (Xiao, Gerke and Vosselman, 2012; Cao and Huang, 2021) and are therefore more complicated.



Input image *Model output segmentation* *Building height vector (Number of floors)*

Figure 6. Example of the method with an image from the Cityscapes test set. (Cordts et al., 2016)

These results align favourably with other methodologies, including those utilizing satellite imagery, which often yield absolute errors in the range of 2–5 meters, typically necessitating SAR data (Frantz et al., 2021) or multi-view imagery (Xiao, Gerke, and Vosselman, 2012; Cao and Huang, 2021), thereby introducing increased complexity.

Street view methods incorporating machine learning can achieve a lower relative error of 5% but necessitate additional data such as building footprints (Zhao, Qi, and Zhang, 2019; Ala, 2020) or camera height information (Yan and Huang, 2022).

Notably, the proposed method operates without the need for supplementary parameters, enabling the utilization of images from the internet or images captured via mobile phones for various purposes, provided that potential challenges associated with dataset diversity can be addressed effectively.

5. Conclusions

Based on the findings of this study, it is evident that artificial intelligence holds significant promise for streamlining the extraction of pertinent building parameters. However, it is clear that there remains substantial progress to be made within the realm of computer vision.

A critical challenge lies in the scarcity of high-quality datasets featuring urban street imagery, especially when considering images from developing countries. Furthermore, the existing datasets have primarily been tailored for the specific demands of self-driving car technology, largely overlooking the requirements of the construction sector.

Encouragingly, the research demonstrates the ability to estimate average building heights within pilot images with only minor errors, operating within a versatile framework that simplifies matching without necessitating complex elements like point clouds or image depth estimation. The semantic segmentation model emerges as a standout performer, yielding impressive results while offering enhanced flexibility and speed compared to alternative models. It effectively tackles the issues of global information extraction deficiency and image detail loss. However, it is apparent that the model requires refinement to operate effectively in non-European urban environments, primarily due to the dearth of suitable datasets.

Future research endeavours could focus on mitigating the scarcity of training data for the segmentation model through domain adaptation, potentially employing a model capable of transforming street images to align with the visual style of a German city. However, it is imperative that comprehensive datasets be generated for various regions across the globe.

A promising avenue for future exploration involves the development of classifier models for estimating building heights. Classifier machine learning models represent a well-established field within artificial intelligence, and these models would assign categorical classes to denote the number of floors in each building. This information carries utmost importance for evaluating seismic exposure and computing building vulnerability, particularly in regions facing pressing challenges in this regard.

6. Acknowledgements

This research received public funds from a grant for the completion of industrial doctorates (IND2020/TIC-17528) funded by Comunidad de Madrid.

The authors gratefully acknowledge the KUK_AHPAN project (RTI2018-094827-B-C21/C22) for the conceptual support about seismic risk and the Universidad Politécnica de Madrid (www.upm.es) for providing computing resources on Magerit Supercomputer.

7. References

Ala, (2020) An Open-source System for Building-height Estimation using Street-view Images, Deep Learning, and Building Footprints Reports on Special Business Projects. Available at: www.statcan.gc.ca.

Baggio, C. et al. (2007) Field Manual for post-earthquake damage and safety assessment and short term countermeasures (AeDES) Translation from Italian: Maria ROTA and Agostino GORETTI. Available at: http://ipsc.jrc.ec.europa.eu.

Baldini, G. et al. (2019), ‘Digital Transformation in Transport, Construction, Energy, Government and Public Administration’, EUR 29782 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-08614-7, doi:10.2760/058696, JRC116179.

Benito, M.B. et al. (2012) ‘A new evaluation of seismic hazard for the Central America Region’, Bulletin of the Seismological Society of America, 102(2), pp. 504–523. Available at: https://doi.org/10.1785/0120110015.

Bezanson, J. et al. (2014) ‘Julia: A Fresh Approach to Numerical Computing’. Available at: http://arxiv.org/abs/1411.1607.

Bonczak, B. and Kontokosta, C.E. (2019) ‘Large-scale parameterization of 3D building morphology in complex urban landscapes using aerial LiDAR and city administrative data’, Computers, Environment and Urban Systems, 73, pp. 126–142. Available at: https://doi.org/10.1016/j.compenvurbsys.2018.09.004.

Brzev, S. et al. (2013) GEM global earthquake model GEM Building Taxonomy Version 2.0 exposure modelling.

Cao, Y. and Huang, X. (2021) ‘A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities’, Remote Sensing of Environment, 264. Available at: https://doi.org/10.1016/j.rse.2021.112590.

Cardellicchio, A. et al. (2022) ‘View VULMA: Data Set for Training a Machine-Learning Tool for a Fast Vulnerability Analysis of Existing Buildings’, Data, 7(1). Available at: https://doi.org/10.3390/data7010004.

Cardellicchio, A. et al. (2023) ‘A machine learning framework to estimate a simple seismic vulnerability index from a photograph: the VULMA project’, Procedia Structural Integrity, 44, pp. 1956–1963. Available at: https://doi.org/10.1016/j.prostr.2023.01.250.

Chollet, F. (2016) ‘Xception: Deep Learning with Depthwise Separable Convolutions’. Available at: http://arxiv.org/abs/1610.02357.

Cordts, M. et al. (2016) ‘The Cityscapes Dataset for Semantic Urban Scene Understanding’. Available at: http://arxiv.org/abs/1604.01685.

Díaz, E. and Arguello, H. (2016) ‘An algorithm to estimate building heights from Google street-view imagery using single view metrology across a representational state transfer system’, in Dimensional optical metrology and inspection for practical applications v. SPIE, p. 98680A. Available at: https://doi.org/10.1117/12.2224312.

Dosovitskiy, A. et al. (2020) ‘An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale’. Available at: http://arxiv.org/abs/2010.11929.

Drusch, M. et al. (2012) ‘Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services’, Remote Sensing of Environment, 120, pp. 25–36. Available at: https://doi.org/10.1016/j.rse.2011.11.026.

Esquivel-Salas, L.C. et al. (2022) ‘Remote structural characterization of thousands of buildings from San Jose, Costa Rica’, Frontiers in Built Environment, 8. Available at: https://doi.org/10.3389/fbuil.2022.947329.

European Construction Sector Observatory (2021) Digitalisation in the construction sector.

FEMA (2001) Hazus -MH 2.1. Technical Manual. Washington, DC: (FEMA), Federal Emergency Management Agency, pp. 1–139.

Frantz, D. et al. (2021) 'National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series', Remote Sensing of Environment, 252. Available at: https://doi.org/10.1016/j.rse.2020.112128.

Glorot, X. and Bengio, Y. (2010) Understanding the difficulty of training deep feedforward neural networks. Available at: http://www.iro.umontreal.

Hitawala, S. (2018) 'Evaluating ResNeXt Model Architecture for Image Classification'. Available at: http://arxiv.org/abs/1805.08700.

Huang, G. et al. (2016) 'Densely Connected Convolutional Networks'. Available at: http://arxiv.org/abs/1608.06993.

Innes, M. et al. (2018) 'Fashionable Modelling with Flux'. Available at: http://arxiv.org/abs/1811.01457.

Innes, M. et al. (2019) 'A Differentiable Programming System to Bridge Machine Learning and Scientific Computing'. Available at: http://arxiv.org/abs/1907.07587.

INSPIRE (2013) Infrastructure for Spatial Information in Europe D2.8.III.2, 'Data Specification on Buildings-Technical Guidelines'. Technical Report; European Commission Joint Research Centre: Luxembourg .Available at: http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2.

Kingma, D.P. and Ba, J. (2014) 'Adam: A Method for Stochastic Optimization'. Available at: http://arxiv.org/abs/1412.6980.

Li, Z. et al. (2022) BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation. Available at: https://github.com/zhyever/Monocular-Depth-Estimation-Toolbox.

Müller, T. et al. (2022) 'Instant neural graphics primitives with a multiresolution hash encoding', ACM Transactions on Graphics, 41(4), pp. 1–15. Available at: https://doi.org/10.1145/3528223.3530127.

Neuhold, G. et al. (2017) The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. Available at: www.mapillary.com.

Richter, S.R. et al. (2016) 'Playing for Data: Ground Truth from Computer Games'. Available at: http://arxiv.org/abs/1608.02192.

Rodríguez-Saiz, J. et al. (2022) 'Exposición sísmica de los edificios por métodos geoespaciales', in XIV congreso geológico de américa central &. VII congreso geológico nacional. San José, Costa Rica: Colegio de Geólogos de Costa Rica.

Ruggieri, S. et al. (2021) 'Machine-learning based vulnerability analysis of existing buildings', Automation in Construction, 132. Available at: https://doi.org/10.1016/j.autcon.2021.103936.

Strudel, R. et al. (2021) 'Segmenter: Transformer for Semantic Segmentation'. Available at: http://arxiv.org/abs/2105.05633.

Ureña-Pliego, M. et al. (2023) 'Automatic Building Height Estimation: Machine Learning Models for Urban Image Analysis', *Applied Sciences* 13, no. 8: 5037. https://doi.org/10.3390/app13085037

Wang, J. et al. (2019) 'Deep High-Resolution Representation Learning for Visual Recognition'. Available at: http://arxiv.org/abs/1908.07919.

Xiao, J., Gerke, M. and Vosselman, G. (2012) 'Building extraction from oblique airborne imagery based on robust façade detection', ISPRS Journal of Photogrammetry and Remote Sensing, 68(1), pp. 56–68. Available at: https://doi.org/10.1016/j.isprsjprs.2011.12.006.

Xie, S. et al. (2016) 'Aggregated Residual Transformations for Deep Neural Networks'. Available at: http://arxiv.org/abs/1611.05431.

Yan, Y. and Huang, B. (2022) 'Estimation of building height using a single street view image via deep neural networks', ISPRS Journal of Photogrammetry and Remote Sensing, 192, pp. 83–98. Available at: https://doi.org/10.1016/j.isprsjprs.2022.08.006.

Zhang, Z. and Sabuncu, M.R. (2018) 'Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels'. Available at: http://arxiv.org/abs/1805.07836.

Zhao, Y., Qi, J. and Zhang, R. (2019) 'CBHE: Corner-based building height estimation for complex street scene images', Association for Computing Machinery, Inc, pp. 2436–2447. Available at: https://doi.org/10.1145/3308558.3313394.