Three Dimensional Street Scene Representation Learning for Street Frontage Classification

Conference Paper · June 2025

CITATIONS
0

READS
176

3 authors, including:

Zicheng Fan
National University of Singapore
11 PUBLICATIONS **131** CITATIONS
SEE PROFILE

Filip Biljecki
National University of Singapore
234 PUBLICATIONS **8,997** CITATIONS
SEE PROFILE

This paper is peer reviewed and accepted at the 19th International Conference on Computational UrbanPlanning and Urban Management.
https://www.ucl.ac.uk/bartlett/casa/about/cupum-2025 Doi:10.17605/OSF.IO/ABYQH

Three Dimensional Street Scene Representation Learning for Street Frontage Classification

Zicheng Fan*[1], Stephen Law†[2] and Filip Biljecki‡[1,3]

[1]Department of Architecture, National University of Singapore, Singapore
[2]Department of Geography, University College London, United Kingdom
[3]Department of Real Estate, National University of Singapore, Singapore

August 9, 2025

Summary

Street View Imagery (SVI) serves as a valuable medium for sensing and analysing street spaces, however, it inadequately represents the 3D morphological structure of streets. The study presents a novel approach that transforms a single omni-directional SVI into a coloured point cloud, explicitly enhancing its ability to represent the 3D morphological structure of street space. Building on these point clouds, experiments are conducted applying 3D deep learning methods to infer street frontage attributes, such as the activeness. Street frontage classification based on 3D point cloud structures performs comparably to traditional image-based methods and surpasses them with the addition of both 3D and colour information. The work presents great potential to understand and explain street attributes from street morphology, contributing to more effective urban design and renewal strategies.
KEYWORDS: Urban Perception, Street View Imagery, Point Clouds, Representation Learning, 3D Deep Learning.

1 Introduction

Human perceptions and experiences of street spaces are rich and varied, holding significant importance in our daily lives. A well-designed street environment enhances commercial activity (Kim and Woo 2022), safety (Cui et al. 2023), mental health (Wang et al. 2019), and property values (Law, Paige, and Russell 2019). Conversely, poor street environments often lead to stress and anxiety (Chen et al. 2023), and can be associated with poverty (Yuan et al. 2023) and even criminal activities (Z. He et al. 2022). Perception of street spaces largely depends on visual information, including the colour, material, and texture of street facades, street traffic and commercial activities, and street furniture and vegetation. Street view imagery (SVI) has been widely used in urban analytics research as a valuable medium to understand human perceptions of street spaces (Biljecki

*zicheng.f@u.nus.edu
†stephen.law@ucl.ac.uk
‡filip@u.nus.edu

and Ito 2021). The use cases include auditing street quality (Smith, Kaufman, and Mooney 2021; Li et al. 2022), detecting changes in street environments (Han et al. 2023; Stalder et al. 2024), and extracting specific street elements such as buildings, trees (Zhang et al. 2024) and urban lighting (Fan and Biljecki 2024). The advancements of deep learning techniques, especially in computer vision, have provided powerful tools for these tasks (Ibrahim, Haworth, and Cheng 2020).

However, urban imagery may not be the only approach to represent street spaces. One key limitation is that individuals experience their environment dynamically in 3D, building knowledge of the 3D morphological structure through movement. Navigating urban spaces using 3D city models was reported "more intuitive" than relying on 2D maps (Biljecki, Stoter, et al. 2015). Nevertheless, 3D information such as spatial depth, distance, enclosure and height are arguably not fully represented with SVI. The 3D morphological structure is implicitly inferred from the visual information present in the street imagery, resulting in a lack of targeted consideration of the spatial form and structure corresponding to the street scenes. Consequently, it remains unclear to what extent an individual's perception of street environments derived from SVI is based purely on visual semantics, such as building materials or street trees, as opposed to the influence of morphological characteristics inherent to the streetscape.

Given these limitations, this study aims to develop a workflow that translates SVI into point clouds, applies 3D deep learning methods to learn a representation of a street scene and evaluates the approach through a down-stream street frontage classification task. Through a series of experiments, we aim to address the following questions:

1. Can street frontages characteristics such as activeness be effectively classified solely using three-dimensional morphological structure?
2. Can the combination of three-dimensional morphological characteristics and visual features enhance the performance of classifying street frontage characteristics?

2 Methods

An overall research framework of the study is shown in Figure 1 which consists of two parts; translating SVI into point clouds, and applying 3D deep learning techniques for classifying street frontage according to activeness.

2.1 SVI to Point Cloud Workflow

We first explore a workflow to generate coloured point clouds based on panoramic SVI and their corresponding depth information. Specifically, following the method proposed by Cavallo (2015), panoramic images under the equirectangular projection were reprojected into a spherical coordinate system. The horizontal and vertical coordinates of pixels were mapped to azimuth and inclination angles and the depth information from depthmap served as the radius. Using the sphere's centre as the origin, the spherical coordinates were then transformed into Cartesian coordinates, enabling the calculation of each pixel's position in 3D space. In this way, the 2D SVIs can be converted

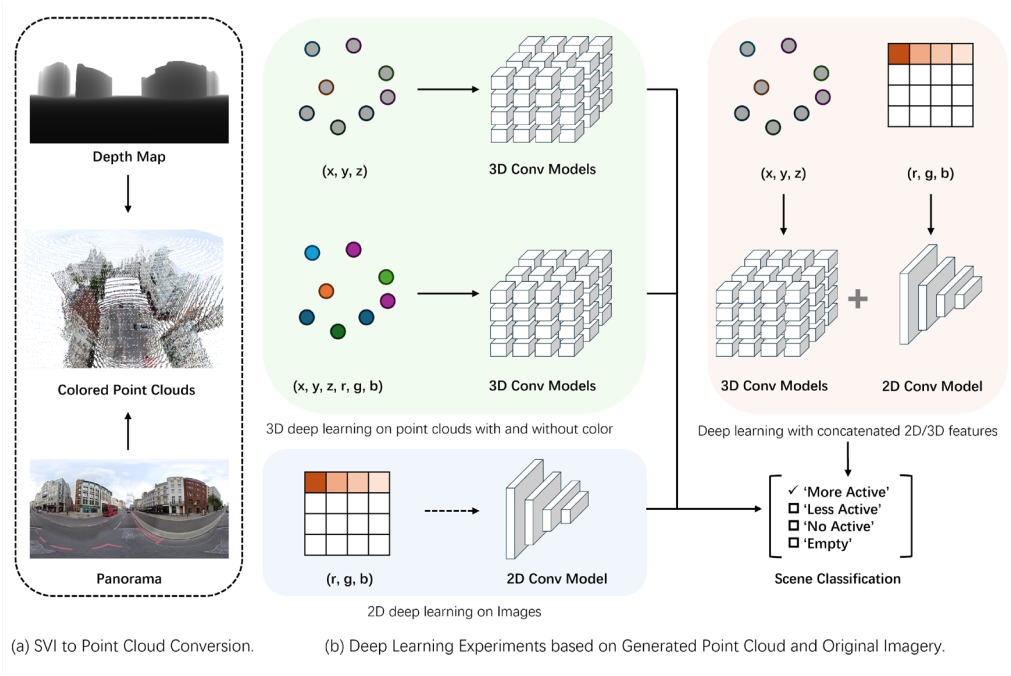

Figure 1: A Research Framework of the Study.

to coloured 3D point clouds, and the morphological structure characteristics of street scenes are restored and represented.

2.2 3D Deep Learning

The second part of this study focuses on applying 3D deep learning to the generated street view point clouds, and evaluating if classification performance can be improved compared to traditional methods. 3D deep learning primarily focuses on learning local or global structural features from point clouds or voxels. Similar to applying deep learning on images, 3D deep learning can also accomplish classification and segmentation tasks, such as distinguishing vehicles like cars and motorcycles on the street. In this study, we hypothesize that the morphological structure of a point-cloud-based street scene is linked to its frontage characteristics and can be effectively learned and classified. Based on this hypothesis, we applied PointNet (Qi, Su, et al. 2017), PointNet++ (Qi, Yi, et al. 2017), and Point Transformer (Zhao et al. 2021) architectures to train deep learning models using the generated point clouds, aiming to infer the frontage characteristics of street scenes.

2.3 Case Study and Experiments

Using Manhattan, New York City, as the study area, we downloaded 2,169 omni-directional panoramic images from Google Street View (GSV). As shown in Figure 3, these images were manually filtered and labelled into four categories based on different street frontage types: 'More-Active', 'Less-Active', 'Non-Active', and 'Empty' following a similar procedure as Law,

Seresinhe, et al. (2018) in labelling the activeness of a street frontage inspired from the seminal work of Jacobs (1961).

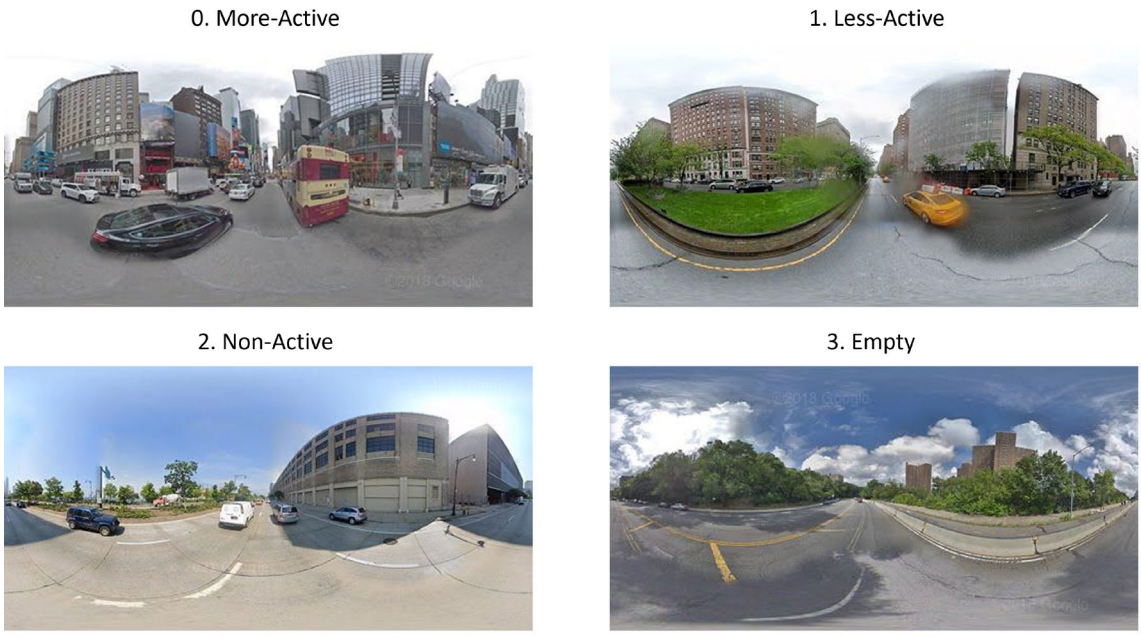

Figure 2: Sample images for the four street frontage classes.

Using the `streetlevel` library[1], we downloaded the depthmap of SVI from GSV and projected SVI and depthmap pairs into coloured point clouds. By randomly splitting the point clouds and SVI pairs into training and testing sets with a ratio of 80:20, we conducted the following experiments:

- **Point Cloud Learning**: Models (PointNet, PointNet++, Point Transformer) were trained on point clouds with spatial coordinates (x, y, z) and with colour information (x, y, z, r, g, b) for street scene classification. The aim is to evaluate the effectiveness of pure morphological features in distinguishing street scenes and to assess the impact of incorporating additional colour and texture information.
- **Image Learning**: ResNet50 architecture was used to classify street scenes from the original images, serving as a baseline.
- **Multi-modal Learning**: Features from Point Transformer (point clouds) and ResNet50 (images) were concatenated and trained with a multilayer perception (MLP) to explore integrating spatial and visual information.

[1] https://streetlevel.readthedocs.io/en/master/

3 Results

3.1 Point Cloud Learning Experiments

Table 1 compares the best performance of different model architectures on the classification task using point clouds with and without colour information as input. All experiments were optimised using the negative log-likelihood loss function with the ADAM optimiser. Each experiment was trained for 25 epochs, the initial learning rate was set at 0.0001 and halved every 10 epochs, the batch size was 16, and the number of sampled points was 1024. Each input-model configuration was tested twice, and the best performance in terms of accuracy and F1 was recorded.

Table 1: Performance metrics for different 3D deep learning architectures.

	Without Colour		**With Colour**	
Models	Accuracy	F1	Accuracy	F1
PointNet	0.657	0.618	0.671	0.667
PointNet++	0.645	0.610	0.710	0.700
PointTransformer	0.659	0.624	0.682	0.667

It is found that PointNet++ achieved the best performance when colour information was included, with an accuracy of 71.0% and F1 score of 70.0%. Figure 3 illustrates the loss and accuracy variation during the training process. When colour information was excluded, Point Transformer performed best with an accuracy score of 65.9% and F1 score of 62.4%. For all tested model architectures, incorporating colour information improved both model performance and training stability.

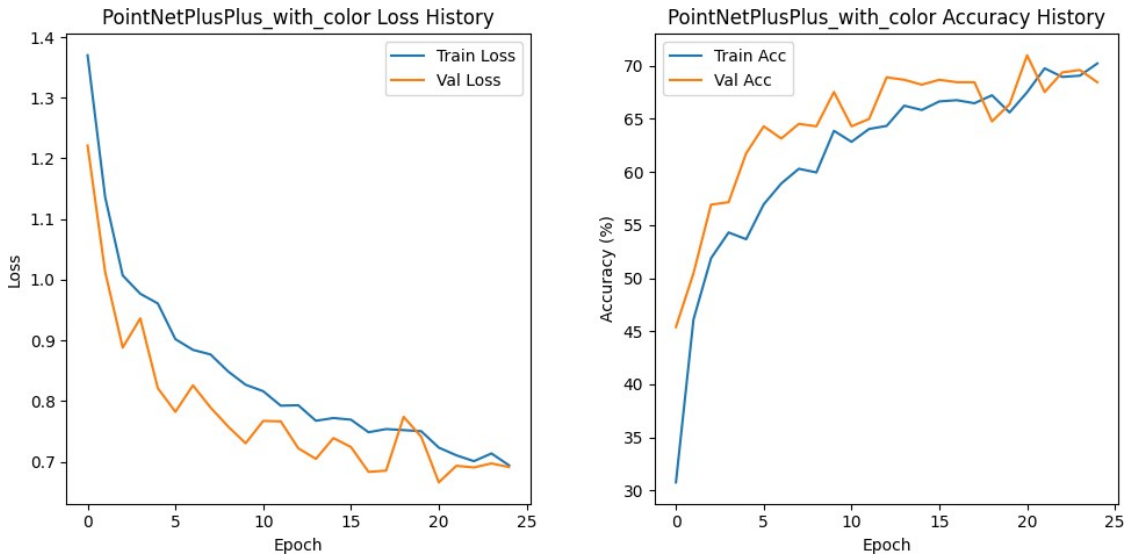

Figure 3: Loss and accuracy variation during training process.

3.2 Image Learning and Multi-modal Learning Experiments

Based on the same train-and-test split, we trained baseline image models based on a ResNet50 architecture (K. He et al. 2016). Specifically, we experiment models with one linear layer, or with 1 or 2 dense layers in a MLP. The model performance is shown in Table 2. Model with two dense layers achieves the best performance, with accuracy score of 0.703 and F1 score of 0.704.

Table 2: Performance metrics for different ResNet50 models.

Models	Accuracy	F1 Score
ResNet50 + Linear	0.698	0.703
ResNet50 + 1 Dense	0.671	0.668
ResNet50 + 2 Dense	0.703	0.704

With Point Transformer and ResNet50 as feature extractors respectively, we conduct further experiments for multi-modal learning. Point cloud and image features are concatenated and passed through either a linear layer or one or two dense layers in a MLP for comparison. Similarly, each input-model configuration was tested twice. As shown in Table 3, model with one additional dense layer achieves the best performance with an accuracy score of 0.733 and F1 score of 0.737.

Table 3: Performance metrics for different multi-modal models.

Models	Accuracy	F1 Score
Point Transformer + ResNet50 + Linear	0.719	0.707
Point Transformer + ResNet50 + 1 Dense	0.733	0.737
Point Transformer + ResNet50 + 2 Dense	0.724	0.713

4 Discussion and Conclusion

The analysis reveals that morphological features of street spaces, represented by coordinates of point cloud data, can effectively classify street frontage, although their performance is slightly weaker than traditional image-based classification models. Integrating morphological structure with colour information significantly improves classification accuracy. Furthermore, concatenating 2D and 3D features via multi-modal classification achieved the best street frontage classification results. This study highlights the importance of 3D spatial structural information in understanding street attributes, a factor often overlooked in traditional SVI-based urban studies. Through a series of experiments, the learnability of street morphological structures and their effectiveness in enhancing deep learning performance are demonstrated. Further research is expected to examine the geographical generalizability of the pattern and to explore various methods for integrating 2D and 3D information in urban image analytics.

Acknowledgements

The first author is supported by the National University of Singapore under the President's Graduate Fellowship. This research is part of the project Large-scale 3D Geospatial Data for Urban Analytics, which is supported by the National University of Singapore under the Start Up Grant R-295-000-171-133.

References

Biljecki, Filip and Koichi Ito (2021). "Street view imagery in urban analytics and GIS: A review". In: *Landscape and Urban Planning* 215, p. 104217. doi: 10.1016/j.landurbplan.2021.104217.

Biljecki, Filip, Jantien Stoter, et al. (2015). "Applications of 3D City Models: State of the Art Review". In: *ISPRS International Journal of Geo-Information* 4.4, pp. 2842–2889. doi: 10.3390/ijgi4042842.

Cavallo, Marco (2015). "3D City Reconstruction From Google Street View". In: url: https://www.semanticscholar.org/paper/3D-City-Reconstruction-From-Google-Street-View-Cavallo/83e78578851af19540a930ff4a44f3215857d1d6.

Chen, Jingjia et al. (2023). "Measuring Physical Disorder in Urban Street Spaces: A Large-Scale Analysis Using Street View Images and Deep Learning". In: *Annals of the American Association of Geographers* 113, pp. 469–487. doi: 10.1080/24694452.2022.2114417.

Cui, Qinyu et al. (2023). "Analysing gender differences in the perceived safety from street view imagery". In: *International Journal of Applied Earth Observation and Geoinformation* 124, p. 103537. doi: 10.1016/j.jag.2023.103537.

Fan, Zicheng and Filip Biljecki (2024). "Nighttime Street View Imagery: A new perspective for sensing urban lighting landscape". In: *Sustainable Cities and Society* 116, p. 105862. doi: 10.1016/j.scs.2024.105862.

Han, Yuqi et al. (2023). "Mapping seasonal changes of street greenery using multi-temporal street-view images". In: *Sustainable Cities and Society* 92, p. 104498. doi: 10.1016/j.scs.2023.104498.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, Zhanjun et al. (2022). "Multiscale analysis of the influence of street built environment on crime occurrence using street-view images". In: *Computers, Environment and Urban Systems* 97, p. 101865. doi: 10.1016/j.compenvurbsys.2022.101865.

Ibrahim, Mohamed R., James Haworth, and Tao Cheng (2020). "Understanding cities with machine eyes: A review of deep computer vision in urban analytics". In: *Cities* 96, p. 102481. doi: 10.1016/j.cities.2019.102481.

Jacobs, Jane (1961). *The Death and Life of Great American Cities*. Random House Inc.

Kim, Sohee and Ayoung Woo (2022). "Streetscape and business survival: Examining the impact of walkable environments on the survival of restaurant businesses in commercial areas based on street view images". In: *Journal of Transport Geography* 105, p. 103480. doi: 10.1016/j.jtrangeo.2022.103480.

Law, Stephen, Brooks Paige, and Chris Russell (2019). “Take a Look Around: Using Street View and Satellite Images to Estimate House Prices”. In: *ACM Trans. Intell. Syst. Technol.* 10, 54:1–54:19. doi: 10.1145/3342240.

Law, Stephen, Chanuki Illushka Seresinhe, et al. (2018). “Street-Frontage-Net: urban image classification using deep convolutional neural networks”. In: *International Journal of Geographical Information Science* 34.4, pp. 681–707. doi: 10.1080/13658816.2018.1555832.

Li, Xiao et al. (2022). “Urban infrastructure audit: an effective protocol to digitize signalized intersections by mining street view images”. In: *Cartography and Geographic Information Science* 49, pp. 32–49. doi: 10.1080/15230406.2021.1992299.

Qi, Charles R., Hao Su, et al. (2017). *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*. doi: 10.48550/arXiv.1612.00593.

Qi, Charles R., Li Yi, et al. (2017). *PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space*. doi: 10.48550/arXiv.1706.02413.

Smith, Cara M., Joel D. Kaufman, and Stephen J. Mooney (2021). “Google street view image availability in the Bronx and San Diego, 2007–2020: Understanding potential biases in virtual audits of urban built environments”. In: *Health & Place* 72, p. 102701. doi: 10.1016/j.healthplace.2021.102701.

Stalder, Steven et al. (2024). “Self-supervised learning unveils urban change from street-level images”. In: *Computers, Environment and Urban Systems* 112, p. 102156. doi: 10.1016/j.compenvurbsys.2024.102156.

Wang, Ruoyu et al. (2019). “Using street view data and machine learning to assess how perception of neighborhood safety influences urban residents’ mental health”. In: *Health & Place* 59, p. 102186. doi: 10.1016/j.healthplace.2019.102186.

Yuan, Yuan et al. (2023). “Using street view images and a geographical detector to understand how street-level built environment is associated with urban poverty: A case study in Guangzhou”. In: *Applied Geography* 156, p. 102980. doi: 10.1016/j.apgeog.2023.102980.

Zhang, Fan et al. (2024). “Urban Visual Intelligence: Studying Cities with Artificial Intelligence and Street-Level Imagery”. In: *Annals of the American Association of Geographers* 0, pp. 1–22. doi: 10.1080/24694452.2024.2313515.

Zhao, Hengshuang et al. (2021). *Point Transformer*. doi: 10.48550/arXiv.2012.09164.

Biographies

Zicheng Fan is a PhD candidate in the Department of Architecture, NUS. His research centres on spatial data science, especially the effectiveness and uncertainty of applying street view imagery in urban analytics.

Stephen Law is an associate professor in UCL Geography. His research primarily centres on applying data science in urban design analytics.

Filip Biljecki is an assistant professor in the Department of Architecture and the Department of Real Estate, NUS. His research and teaching are converging Geomatic engineering, geospatial technologies, and urban data science to support digital twins and data-driven urban planning.