LEVERAGING 3D MODEL IMAGERY TO AUTOMATICALLY ESTIMATE A NEW WINDOW VIEW INDEX

Stephen Law *
Department of Geography
University College London
London, UK
stephen.law@ucl.ac.uk

Esra Suel
Centre for Advance Spatial Analysis
University College London
London, UK
e.suel@ucl.ac.uk

Steven Stalder
Swiss Data Science Center
ETH Zurich and EPFL
Zurich, Switzerland
steven.stalder@sdsc.ethz.ch

Atsushi Takizawa
Department of Living Environment Design
Osaka Metropolitan University
Osaka, Japan
takizawa@omu.ac.jp

ABSTRACT

We yearn for a connection with nature and a sense of refuge from our living spaces, seeking solace in the views from our windows at home. Despite these inclinations, existing window view indices are often time-consuming to collect and over-simplified in their construction. The limited research can be attributed to the lack of data and computational methods for analyzing vistas from properties. To address this gap, this study proposed a novel data collection and resampling procedure that leverages the newly accessible photo-realistic 3D modelled imagery from Google Maps, along with existing machine learning techniques, to improve housing price/rent prediction and to establish a novel window view index for automatically assessing the appeal of views at home.

1 INTRODUCTION

Extensive research has been conducted in environmental psychology that points to the benefits of high-quality views which are mentally restorative and provide opportunities for outlook, refuge, and solace (Kaplan, 1995; Appleton, 1984). This is reflected in differences of premiums paid for a flat with a better view (Baranzini & Schaerer, 2011) and the effects urban view quality has on depression symptoms (Helbich et al., 2019). Empirical evidence also demonstrates that more scenic views are associated with individual happiness (Seresinhe et al., 2019). An abundance of urban view indices have been proposed including ones using remotely sensed data such as the Normalised Difference Vegetation Index or more recently street view-derived visual desirability indicators (Law et al., 2019), visual walkability indicators (Zhou et al., 2019), and skyview factors (Liu et al., 2017). The majority of these indicators are either from overhead aerial views or at the street level using simple semantic attributes and often do not consider the complex 3D environment from a residential window view. While existing window view indices are either time-consuming to collect or over-simplified in their construction, the lack of enriched window view indices can be attributed to the shortage of both data and computational methods for analyzing the vistas from residential properties. To bridge this research gap, this study will introduce a novel data collection pipeline in making use of the newly available photo-realistic 3D-modelled imagery from Google Maps, coupled with existing machine learning and vision techniques to tackle the following aims;

- First, to introduce a novel data collection and resampling procedure in retrieving window views in Tokyo, Japan.

*Corresponding Author.

- Second, to explore how the views from homes can enhance the accuracy of housing rents prediction beyond simple semantic and depth features, testing a series of pretrained and newly trained vision backbones.
- Third, to propose an innovative window view metric extending existing research, for assessing the desirability of views from properties in Tokyo, Japan.

2 Related Works

2.1 Window view quality

Window view quality is commonly assessed by both subjective and objective methods. The former often uses stated preference methods where participants would provide a subjective judgment on views (assess or rank) which often tend to be smaller in scale (Lottrup et al., 2015). The latter utilises objective measurements including 3D visibility analysis and image analysis using machine learning methods. As an illustration, Li et al. (2021) used a transfer learning approach to categorise a window photograph view as either nature or construction. In a subsequent study, Li et al. (2022) employed a 3D photo-realistic City Informational Model (CIM) alongside a fine-tuned semantic segmentation model to quantify a window view index, encompassing factors such as greenery, sky, construction, and water bodies. However, previous research; i. was only able to retrieve a small number of window views from custom-designed CIM models, and ii. used semantic attributes which might not capture more holistic image features from a window scene such as the overall composition and the skyline.

2.2 Visual desirability from prices

Previous research has demonstrated the association between visual attributes derived from urban images, socioeconomic profiles (Gebru et al., 2017), perceived safety (Naik et al., 2014) and individual happiness (Seresinhe et al., 2019). In particular, we have seen the increasing use of urban imagery to estimate house price/rent in urban economics. For instance, Ahmed & Moustafa (2016) supplemented traditional housing features with visual features extracted from property photos to estimate house prices. However, these studies primarily focused on predictive accuracy rather than interpretability. In contrast, Law et al. (2019) proposed a semi-interpretable hedonic price model aimed at assessing the visual appeal of London's street views based on housing prices. This study will adopt a similar approach to forecast housing rental values in Tokyo and propose a novel window-view desirability index that can be automatically mapped using 3D modeling imagery data from Google Maps.

3 Method and Materials

3.1 Architecture

Adopting the architecture introduced by Law et al. (2019), we present a two-stage pipeline illustrated in Figure 1 for forecasting Tokyo property rental rates. In the first stage, we employ a conventional hedonic regression model, denoted as $F(X;\theta)$ parameterised by θ, which maps housing characteristics X to rental prices Y. These typical housing attributes encompass the property's size, proximity to the nearest rail station, passenger traffic at the nearest rail station, the property's construction year, and the building's height. We minimise the mean squared error loss function $L(\theta)$ between the predicted and the observed rent with added l2 regularisation on the regression model's weights. More formally, $L(\theta) = \frac{1}{n}\sum_{i=1}^{N}\left(y_i - F(x_i;\theta)\right)\Big)^2 + \lambda\left\|\theta\right\|^2$.

In the second stage, we begin by extracting deep image features S from the modelled imagery I using different pretrained/trained vision backbone models $V(\cdot)$. We then learn a second regression model, denoted as $G(S;\theta^{[s]})$ parameterised by $\theta^{[s]}$ that takes the deep image features S to predict the difference $W = Y - \hat{Y}$ from the first stage to infer a window-view index $\hat{W}$. Similar to the first stage, we minimise the mean squared error loss function $L(\theta^{[s]})$ with added l2 regularisation on the regression model's weights. The losses are optimized in the learning process using an ADAM optimiser with a learning rate of 0.01 for 5000 epochs.

3.2 VISION BACKBONES

A number of vision backbones $V(\cdot)$ have been tested. This includes; i. Vision Transformer (Dosovitskiy et al., 2021), a state-of-the-art self-attention vision transformer encoder pretrained on ImageNet, ii. Clip (Radford et al., 2021), a Contrastive Vision-Language foundation model pretrained on CoCo dataset, iii. Street2Vec (Stalder et al., 2023), a self-supervised vision model that was trained on pairs of time series street view images in London, following the non-contrastive Barlow Twins learning objective (Zbontar et al., 2021), iv. SatlasPretrain Bastani et al. (2023), an aerial image foundational model pretrained on high-resolution NAIP. As an additional baseline, v. we combine the frequencies of urban semantic classes extracted from Segformer (Xie et al., 2021) pretrained from the ADE20k dataset and statistics from the depth maps. Finally, vi. we also train two simple Encoder-Decoder models (five $Conv - BatchNorm - Maxpool$ encoder blocks and five $ConvT - BatchNorm - UnMaxpool$ decoder blocks) on the synthetic modelled imagery. The first model contains only the coloured imagery (RGB) and the second contains both coloured and depth information (RGBD).

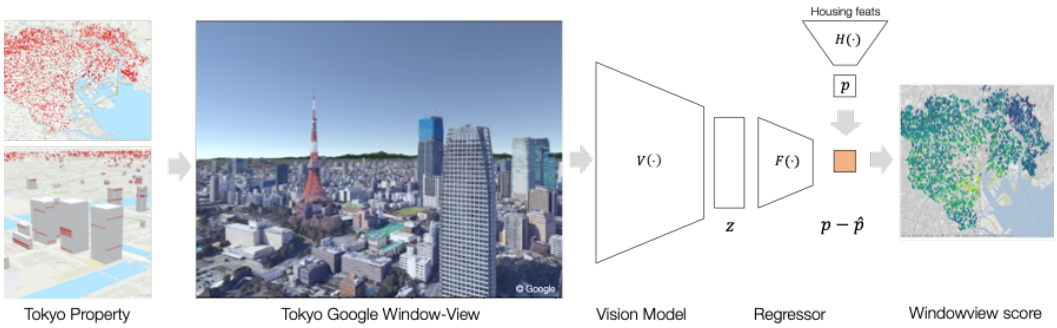


Figure 1: Pipeline for predicting Tokyo's window view scores.

3.3 MATERIALS AND DATA COLLECTION PIPELINE

Below is a concise overview and description of the data collection pipeline and the three sources of data used, namely; i. the Lifull property dataset, ii. the Plateau building dataset and iii. the Google 3D model imagery dataset. Figure 2a shows the data collection pipeline. We first retrieve Tokyo's (23 wards) rental data from the Lifull property dataset between 07-2015 and 06-2017 [1]. The dataset contains property rents, locations, and the attributes used for the regression model (size, distance to railway station, station passenger volume, station accessibility and the height of building). Next, we merged the Lifull property dataset with the Plateau building dataset [2] which contains 3D information on the floor and orientation of each flat. When processing the building data, we generated points along the building's perimeter at 5-meter intervals. These points were then extended vertically for each floor to capture views of all properties. Subsequently, we filtered these views based on matching floor levels and unit orientations, resulting in a viewpoint dataset that contains the view parameters $c_i = [lon_i, lat_i, height_i, heading_i]$ specific to each property. This dataset was utilized as input for Google Maps photorealistic 3D tiles application programming interface [3] via the Cesium plugin within Unity to generate high-resolution coloured and depth imagery based on the window views of each property. The camera angle of view is configured with a focal length of 24mm, and the light source is set to a default of 55 degrees with shadows turned off. The processed dataset has been cleaned and comprises 124,476 3Bedroom rental transactions[4]. The output variable, representing property rents, has been log-transformed for the analysis.

[1] https://www.nii.ac.jp/dsc/idr/en/lifull/

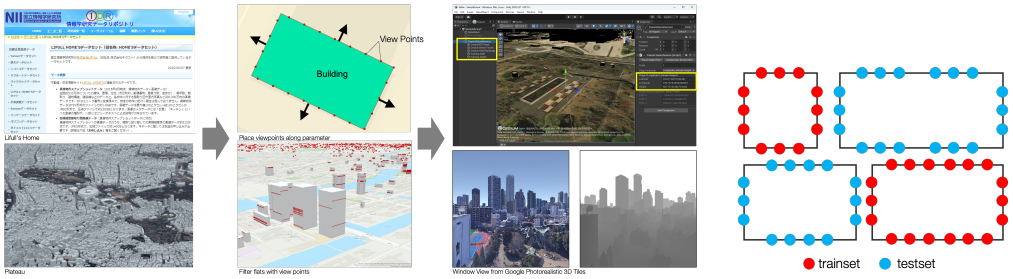
[2] https://www.geospatial.jp/ckan/dataset/plateau-tokyo23ku

[3] https://developers.google.com/maps/documentation/tile/3d-tiles

[4] 3BR market has been selected due to computation efficiency and being a popular housing type for family.

3.4 RESAMPLING AND TRAINING PROCEDURE

In our experiment, we avoided random data splitting because views from the same building could be very similar, which could lead to data leakage between the train and test sets. To address this issue, we developed a novel building resampling approach where we divided the buildings into a 80-20 train and test set. To achieve this split, we needed to identify a group of buildings with a sample size that met this requirement. We treated the resampling as a linear programming problem to ensure an even division of buildings as shown in figure 2b is achieved.



(a) A novel data collection pipeline that leverages on the Lifull property dataset, plateau building dataset and google 3D model imagery to retrieve both the window view and depth imagery for Tokyo rental properties.

(b) To avoid potential data linkage, we proposed a building-level train-test set split.

Figure 2: Novel data collection pipeline and resampling procedures.

4 RESULTS

We estimated a simple linear hedonic model as a baseline, incorporating standard housing attributes ($R2$=62.0% and $rmse$=0.202) and then nine other regression models building on top of the baseline one (refer to Table 1). The $base + SegDepth$ model, which adds urban semantic classes and depth information from the window views, showed improvements compared to the baseline model ($R2$ =65.0% and $rmse$= 0.194). Similarly, the $base + vit$ model, which uses the Vision Transformer encoder, also showed improvements over the baseline model ($R2$=65.1% and $rmse$=0.194). As ablations, we ran two additional vit models ($vit128$ and $vit256$), both trained with a reduced dimension principal components model. The out-of-sample $R2$ and $rmse$ exhibit only minor differences, with an $R2$ of 66.3% and $rmse$ of 0.191. We then tested $BarlowSV$, a Barlow Twins-based non-contrastive self-supervised learning model, pretrained on time series of street images. Its performance was slightly worse than that of the Vision Transformer but still surpassed the baseline model, yielding an out-of-sample $R2$ of 64.1% and $rmse$ of 0.197. Following this, we tested a *CLIP* contrastive learning model, pretrained on CoCo dataset, achieving an out-of-sample $R2$ of 67.2% and $rmse$ of 0.189. We also tested a *SatlasPretrain* foundation model with reduced dimensions $Satlas128$ achieving an out-of-sample $R2$ of 66.9% and a $rmse$ of 0.191. Finally, we tested two simple Encoder-Decoder models, one trained on the modelled coloured imagery (RGB) which boosted the out-of-sample $R2$ and $rmse$ to 70% and 0.180, and the other trained on both modelled coloured and depth imagery (RGBD) which further boosted the $R2$ and $rmse$ to 70.7% and 0.178. These results show the use of simple segmentation features and pretrained vision models on natural images are not as performant for predicting rents involving synthetic data. As a result, we will use our trained $base + EncDecRGBD$ for interpretations and inferring the new window view index.

To interpret the results, we visualised the Tokyo Window View Index WVI geographically. The new WVI map of Tokyo shows that the neighbourhoods in the southwest with views towards Tokyo Bay have a higher window view desirability, while the neighbourhoods to the east of Arakawa River have lower window view desirability. Furthermore, we visualised a sample of images with higher WVI to the left of the map and lower WVI to the right of the map. The findings indicate that views with higher WVI scores offer more expansive vistas of the skyline from higher floors, whereas views with lower WVI scores present more confined perspectives, such as those overlooking buildings and car parks.

Table 1: building out of sample testset results.

model	r2	mse
$base$	0.620	0.202
$base + SegDepth$	0.650	0.194
$base + vit$	0.651	0.194
$base + vit128$	0.663	0.191
$base + vit256$	0.663	0.190
$base + Clip$	0.672	0.189
$base + BarlowSV$	0.641	0.197
$base + Satlas128$	0.669	0.189
$base + EncDecRGB$	<u>0.700</u>	<u>0.180</u>
$base + EncDecRGBD$	**0.707**	**0.178**

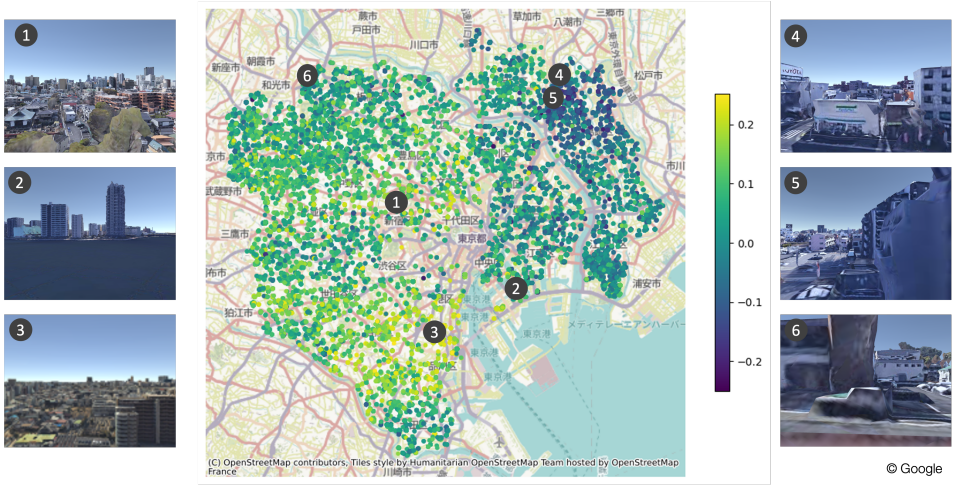


Figure 3: Tokyo new Window View Index

5 Discussion

In summary, this research introduces a novel data collection pipeline that uses newly accessible 3D-modelled imagery data from Google Maps and applied existing machine vision models to enhance the predictive accuracy of housing rent forecasts and in deriving a window view index WVI. This novel holistic index goes beyond the mere semantics of the window scene, providing insights into the quality of city views automatically. The results indicate that people generally prefer higher floors with expansive views overlooking the Tokyo Bay area, than lower floors with shallower views overlooking buildings. These findings underscore the need to improve the quality of window views for properties located on lower ground floors in Tokyo possibly through the incorporation of on-street greenery and art. Several limitations remain. Most importantly, the WVI needs to be validated and examined carefully through human subject surveys. This validation process can help determine the reliability of the machine-generated rating and its alignment with subjective assessments. Secondly, the image quality reduces significantly with close-up views. Verification with on-site window view data might be necessary to alleviate this concern. Thirdly, further research is required to better understand the composition of the scenes. For example, would people prefer architecturally complex or simpler views? Given the opaque nature of deep image representations, the application of explainability methods can help reveal patterns of what constitutes a favourable or less favourable view (Law et al., 2023). Lastly, the inclusion of location information can further improve the predictive accuracy of the model (Mai et al., 2020; Rußwurm et al., 2023). Despite these limitations, this research offers a new way to understand the city from the sky, demonstrating the usefulness of 3D-modelled imagery data and machine learning. Such research affords valuable insights for housing and urban design policymakers, helping them grasp the importance of window views in design and identify geographic areas that would need enhancement.

ACKNOWLEDGMENTS

We thank the support of the Japan Society for the Promotion of Science (JSPS) through the JSPS Summer Fellowship Program and the Nikken Sekkei Research Institute which made this research and publication possible.

REFERENCES

Eman Ahmed and Mohamed Moustafa. House price estimation from visual and textual features, 2016.

Jay Appleton. Prospects and refuges re-visited. *Landscape journal*, 3(2):91–103, 1984.

Andrea Baranzini and Caroline Schaerer. A sight for sore eyes: Assessing the value of view and land use in the housing market. *Journal of Housing Economics*, 20(3):191–199, 2011.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16772–16782, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114 (50):13108–13113, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1700035114. URL https://www.pnas.org/content/114/50/13108.

Marco Helbich, Yao Yao, Ye Liu, Jinbao Zhang, Penghua Liu, and Ruoyu Wang. Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in beijing, china. *Environment international*, 126:107–117, 2019.

Stephen Kaplan. The restorative benefits of nature: Toward an integrative framework. *Journal of environmental psychology*, 15(3):169–182, 1995.

Stephen Law, Brooks Paige, and Chris Russell. Take a look around: Using street view and satellite images to estimate house prices. *ACM Transaction Intelligent, Systems and Technology*, 10(5), 2019. ISSN 2157-6904.

Stephen Law, Rikuo Hasegawa, Brooks Paige, Chris Russell, and Andrew Elliott. Explaining holistic image regressors and classifiers in urban analytics with plausible counterfactuals. *International Journal of Geographical Information Science*, pp. 1–22, 2023.

Maosu Li, Fan Xue, Anthony GO Yeh, and Weisheng Lu. Classification of photo-realistic 3d window views in a high-density city: The case of hong kong. In *Proceedings of the 25th International Symposium on Advancement of Construction Management and Real Estate*, pp. 1339–1350. Springer, 2021.

Maosu Li, Fan Xue, Yijie Wu, and Anthony GO Yeh. A room with a view: Automatic assessment of window views for high-rise high-density areas using city information models and deep transfer learning. *Landscape and Urban Planning*, 226:104505, 2022.

Lun Liu, Elisabete A. Silva, Chunyang Wu, and Hui Wang. A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Computers, Environment and Urban Systems*, 65:113 – 125, 2017. ISSN 0198-9715. doi: https://doi.org/10.1016/j.compenvurbsys.2017.06.003. URL http://www.sciencedirect.com/science/article/pii/S0198971516301831.

Lene Lottrup, Ulrika K Stigsdotter, Henrik Meilby, and Anne Grete Claudi. The workplace window view: a determinant of office workers' work ability and job satisfaction. *Landscape Research*, 40 (1):57–75, 2015.

Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. *arXiv preprint arXiv:2003.00824*, 2020.

Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. Streetscore – predicting the perceived safety of one million streetscapes. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '14, pp. 793–799, Washington, DC, USA, 2014. IEEE Computer Society. ISBN 978-1-4799-4308-1. doi: 10.1109/CVPRW.2014.121. URL http://dx.doi.org/10.1109/CVPRW.2014.121.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. *arXiv preprint arXiv:2310.06743*, 2023.

Chanuki Illushka Seresinhe, Tobias Preis, George MacKerron, and Helen Susannah Moat. Happiness is greater in more scenic locations. In *Nature Scientific Reports*, 2019.

Steven Stalder, Michele Volpi, Nicolas Büttner, Stephen Law, Kenneth Harttgen, and Esra Suel. Self-supervised learning unveils change in urban housing from street-level images. *arXiv preprint arXiv:2309.11354*, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Hao Zhou, Shenjing He, Yuyang Cai, Miao Wang, and Shiliang Su. Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning. *Sustainable cities and society*, 50:101605, 2019.