Eyes on the Street: Assessing Window-to-Wall Ratios in Google Street Views using Machine Learning

Han Tu

Massachusetts Institute of Technology, Boston, MA, United States
hantu@mit.edu

Abstract. Windows play an important role in 'Eyes on the Street' in Jane Jacobs' theory. However, vital street-level parameters in her theory, most notably windows, are rarely assessed at the urban scale due to imprecise existing datasets. To resolve this challenge, this study proposes an automated computer vision-based methodology to extract the window-to-wall ratios (WWRs) of buildings in the Bronx, New York, using semantic segmentation machine learning. This study brings together machine learning and Google Street View (GSV) to accurately assess WWRs at the urban scale. The WWR distribution results show that street-level WWRs help to analyze with other urban data, with controlled parameters, such as land use and building age. Our WWR assessment can be universally applied to other cities using geotagged street view imagery of GSV. This study can help provide a reference for precise future urban design and management assessments.

Keywords: Machine Learning, Data Analytics, Google Street View (GSV), Visual quality, Window-to-wall ratio

1 Introduction

Street-level window-to-wall ratios (WWRs) have long played a critical role in assessing the visual quality – the visual stimuli that shape our perceptions, attitudes, and general views (Sheikh & Bovik, 2006) – of urban street façades (Li et al., 2020). The measurement of WWRs at an urban scale has been conducted only for building or urban energy assessments (Shi et al., 2020; Szcześniak et al., 2022).

Information on assessing WWRs for urban street-level visual quality is rare. Planners' and managers' ability to plan and manage urban façades effectively and efficiently is, therefore, limited. However, Google Street View (GSV) provides street-level profile views of urban façades (Charreire et al., 2014; X. Li et al., 2015; Rundle et al., 2011; Taylor et al., 2011) and WWRs. Yet, little research on GSV with WWRs for urban planning exists. With the GSV information, assessed street-level WWR data can be incorporated into urban façade planning and management. Thus, we explore GSV as a street-level, urban WWR assessment tool to understand the visual qualities of the street, such as safety (Raskar et al., 2015; Zhang et al., 2018). We conduct a case study assessment of street façades using GSV images in the Bronx, New York City.

Our methodology involves four steps. First, the GSV Images (400x600 pixels) are collected in two directions vertical to the streets (Tsai & Chang, 2012) using the GSV Image API within the Bronx. Second, based on these large-scale street view image data, the pixel segmentation machine learning model – the Mask R-CNN model (He et al., 2018) trained on our labeled dataset – is used to measure the WWRs in the GSV images at an urban scale automatically. Third, the data predicted by the trained machine learning model are summarized, visualized, and analyzed.

The results reveal that GSV to be suited for assessing street-level WWRs. We suggest that WWRs extracted from GSV can be a relatively objective measurement of street-level visual quality and may be valuable in guiding urban façade planning and management.

2 Related work

In urban design, windows can increase the interaction between people in buildings and on streets. Moreover, Jane Jacobs concludes three necessary qualities that a city street needs to maintain safety as eyes on the street – a clear demarcation, eyes upon the street and sufficient buildings facing streets, and continuous eyes on the street to guarantee effective surveillance (Jacobs, 1961). In her theory, windows are one of the elements that play an essential role in the safety of eyes on the street. She further claims that "Lack[ing] the checks and inhibitions exerted by eye-policed city streets," they can become flashpoints for destructive and malicious behavior. Numerous studies have demonstrated the relationship between eyes on the street and crime rates from theoretical and criminological perspectives (Cozens & Hillier, 2012). However, none of the urban designers has ever demonstrated her theory on how much windows work as eyes on the street to build a safer community using urban data.

Visual quality (i.e., the impacts of visual features on people's feelings, aesthetics, health, etc.) stimulates our perceptions, attitudes, and general views of urban life (Forkenbrock & Sheeley, 2004). Assessing the urban visual quality accurately is important in the design and management of urban environments to demonstrate the theory of Jane Jacobs (1961). Quantitative assessments of urban visual quality aim to automatically measure the quality of images in a perceptually consistent method (Sheikh & Bovik, 2006). However, existing studies on automated quantitative assessments of street facades are not accurate enough to guide urban design. For instance, many studies that measure urban visual qualities and analyze the possible related factors (Dubey et al., 2016; Zhang et al., 2018) do not accurately measure how much each visual element impacts urban environments.

Window-to-wall ratios (WWRs) of street facades are one of the most important evaluation indicators of visual quality. WWRs are typically estimated based on either the use of 3D laser scanning from street-view or aerial images (Früh & Zakhor, 2003). Although image processing techniques based on machine learning and computer vision have been used to extract façade geometries, they are largely insufficient for precise WWR calculations (C.-K. Li et al., 2020). Moreover, their fundamental weakness is that they depend on rigorous façade styles and quality requirements (Koutsourakis et al., 2009; Riemenschneider et al., 2012; Teboul et al., 2010, 2011) or photographs collected by human labor (Früh & Zakhor, 2003, 2004). These methods are "semi-automatic" and rely on significant human input. Additionally, existing urban image segmentation datasets are noisy with shadows from urban objects such as trees, or some windows are unlabeled. These urban image datasets, such as ADE20K (Zhao et al., 2017; Zhou et al., 2018) and Mapillary (Neuhold et al., 2017), PASCAL VOC (Everingham et al., 2010), and CityScapes (Cordts et al., 2016), cannot segment windows and walls precisely for accurate WWR calculations. Thus, we need a dataset to automatically assess WWRs on the urban scale to improve urban design and management.

To overcome these limitations, our research uses Google Street View (GSV) images (which have continuous street images from the perspective of pedestrians) to assess street-level building façades. GSV, first introduced in 2007, is a library of urban panoramic images captured by cameras on cars (Rundle et al., 2011). The visual quality of GSV is similar to people's experiences on the street by car, bike, or foot.

By stitching the images together, GSV images can create continuous building façade experiences of a streetscape. Thus, the GSV image library has been proposed as an effective potential data source for urban assessments (Rundle et al., 2011; Rzotkiewicz et al., 2018), such as public open space audits (Edwards et al., 2013; Taylor et al., 2011), neighborhood environmental audits (Charreire et al., 2014; Rundle et al., 2011; Odgers et al., 2012; Griew et al., 2013), and building energy audits (Shi et al., 2020; Szcześniak et al., 2022).

While we rarely find any study in the literature using GSV images and machine learning for urban assessment of WWRs, we decide to examine whether or not the application of GSV images with computer vision machine learning for assessing street-level WWRs can work effectively and efficiently. Such an automated assessment can provide a potential method to evaluate the urban environment accurately and to guide urban design.

3 Methodology

3.1 Google street view (GSV) data collection

Our training set contains 500 Google Street View (GSV) images scraped randomly in the Bronx, New York. Our prediction set contains 1874 images in Soundview, a neighborhood in the south of the Bronx, New York City (Fig.1). The street map of the study area was processed and generated based on OpenStreetMap, an editable geographic database.

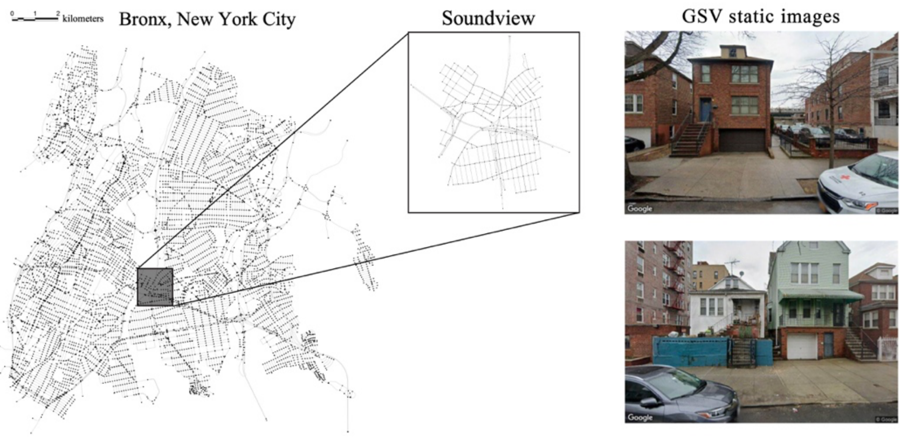

Figure 1. The location, road map, and GSV of a site in Soundview, Bronx, New York. Source: Author, 2022.

Figure 1 shows the GSV of a site in Soundview, Bronx, New York City. The street view is the same as a user sees with GSV panorama (Tsai & Chang, 2012). We collect the GSV images of 400 X 600 pixels in two directions vertical to the streets at intervals of 20m (Tsai & Chang, 2012) using the GSV Image API within the Bronx. To represent the urban WWRs of the study area, 500 sample sites are generated randomly along the road map as the training set from the scraped images.

3.2 Annotation of the training set

We annotate 500 images on the makesense.ai platform. Figure 2 shows an annotated example of our training set. In terms of the selection of labeled facades, we only label the walls and windows facing the camera because the facades with perspective angles cause the miscalculation of the window-to-wall ratio. In terms of the labeled content, the edges of the annotation boxes are the edge of the windows and walls. If a huge tree trunk is encountered, the annotation boxes should not contain the large trunk area; If a small tree branch is encountered, the annotation boxes should contain the small branch area. The area of the walls includes the area of the windows, but not the area of roofs. Finally, the labeling result is a dataset in Coco format containing two label classes - window and wall.

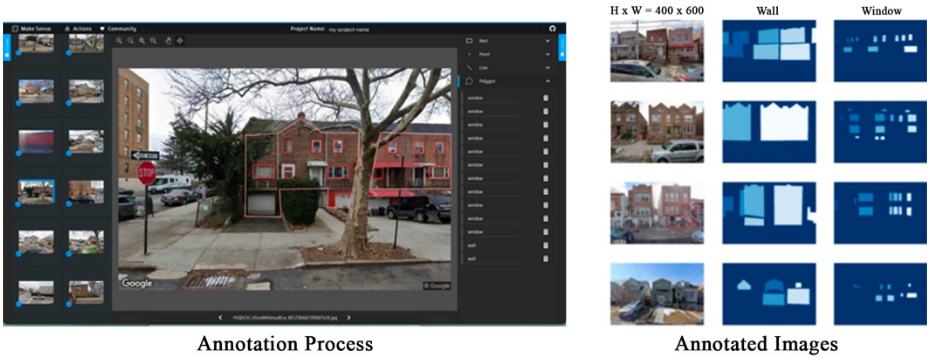

Figure 2. Annotated training set of GSV in the Bronx, New York. Source: Author, 2022.

3.3 Window and wall area extract from GSV images

The existing trained segmentation models cannot accurately detect the windows from the walls. After testing on the same model of DeepLab (Chen et al., 2017), we found that the reasons include the following 4 biased aspects: noise of the training set, different sources of the datasets, the detection of only glass, and other biases.

Some urban image datasets, such as ADE20K (Zhao et al., 2017; Zhou et al., 2018) and Mapillary (Neuhold et al., 2017), CityScapes (Cordts et al., 2016), have the classes of windows and walls. However, their datasets do not annotate every window of the facades on the street. Thus, the mAP of the window detection is lower than we need for WWR calculation. These noisy dataset annotations result in an imprecise output of specific class detection. Figure 3. shows an output example of the DeepLab model trained on the dataset of ADE20K. From the image, on the right is the GSV image from Soundview; on the left, we can see that the buildings, sky, trees, and cars are well segmented. However, the windows cannot be detected due to the noise of the dataset.

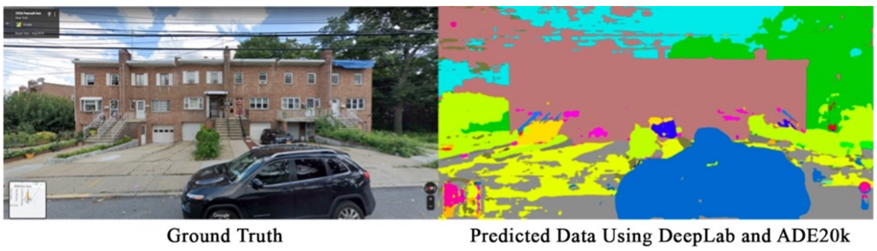

Figure 3. An example of the GSV segmentation results using DeepLab and ADE20k. Source: Author, 2022.

Some datasets are scraped from other media platforms, such as PASCAL VOC (Everingham et al., 2010) from Flickr. The image of these datasets contains different scenes from Google Street View images. For example, Flickr images include more interior, and people view images than street-level images. The training sets of these datasets do not have the same domain as the Google Street View images in the Bronx. Therefore, different databases of image sources lead to different databases suitable for use in different contexts.

Window detection is a key component in many graphics and vision applications related to 3D city modeling and scene visualization. However, some façade parsing datasets can detect the glass in the buildings, such as research by Li (2020). Moreover, these datasets cannot detect all types of windows, and some of them focus on glass detection. Thus, the existing window detection dataset cannot parse the windows in the Bronx effectively. Therefore, we train the pixel segmentation machine learning model – the Mask R-CNN model (Chen et al., 2017) on our dataset for WWR calculations. Then, the output of the trained model can automatically measure the WWRs in the GSV images on an urban scale. Finally, the mean average precision (mAP) is used to evaluate the output of the trained model.

3.4 WWR index calculation

We extract the number of masks of different colors from the image segmentation result. The output results include

- the number of windows,
- the number of walls,
- the number of pixels of windows,
- the number of pixels of walls, and
- the number of pixels in the images.

WWR measures the window area on the total amount of exterior wall area, which is the result of the total number of pixels of windows over the number of pixels of walls. The measured WWR data by machine learning prediction is calculated, mapped. Last, the data predicted by the trained machine learning model are summarized, visualized, and analyzed.

4 Results

4.1 Data augmentation in machine learning model

We use 455 training set and 55 validation set to train the Mask R-CNN model (Chen et al., 2017). In the training process, we augment the images using the augmentation function of imgaug (https://github.com/aleju/imgaug). We utilize the `Gauss. Noise + Contrast + Sharpen, Affine, Crop + Pad` to augment the street view images to reduce the large number requirement of the training set.

In our case, cropping the image is necessary because the distance of the building facade from the camera varies very much in our dataset. The impact of irrelevant patterns like tree trunk and car can be reduced. Gaussian noise is added to prevent overfitting when our model tries to learn high frequency features are not useful.

4.2 Model hyperparameters and training

Our implementation follows the Mask RCNN paper for the most part, but we adjust several hyperparameters in favor of simplicity and generalization. To support training multiple images per batch we resize all the images to the same size: 384*384. The bounding box are generated based on the smallest box that encapsulate all the pixels of the mask, which is useful for a quick assessment of the result correctness. This makes the model easy to apply images augmentations that would otherwise be harder to apply to bounding boxes.

After several iterations of adjusting hyperparameters, we end up with the 0.001 learning rate, 143 epochs, and 384*384 image size. The mAP of the training is 0.655 in testing the model on 55 validation images, more than 0.5 of the ADE20k datasets. Figure 4. shows the examples of the generated masks. The result shows that we can successfully extract the area of windows and walls to analyze the WWRs further.

Ground Truth with Annotation　　Predicted Data Using Our Model

Figure 4. The generated masks of google street view using our model. Source: Author, 2022.

Table 1. Sample of window-to-wall ratio output of the predicted images of Google Street View in the Bronx, New York

Date	lat	lon	ang	Window _num	Wall_ num	Windo w_pix	Wall_p ix	ratio
2022-02	40.835949	-73.87157	2.454806	12	5	2529	39237	0.064

2022-02	40.834112	-73.874904	7.799306	12	1	9621	39957	0.241
2022-02	40.842859	-73.881411	48.034176	6	1	5955	26715	0.223
2019-11	40.839317	-73.869859	20.287989	2	3	2694	62937	0.043

Source: Author, 2022.

4.3 Data analysis and visualization

Using the WWR index function mentioned above, we are able to calculate the WWRs from the generated mask and the number of windows and walls. The generated data from images are stored in a pandas data frame and then merged with geo-tagged data frame based on the panoID column. Among the 1,874 test images, 97.9% are distributed between 0 to 0.2 (Figure 5. right chart), suggesting a relatively precise result.

The result is mapped for a clear view of the geographical distribution. From Figure 5, we find an interesting phenomenon that main streets have relatively small WWR compared with secondar streets. One possible explanation is that these main streets have higher traffic flow, and the residents prefer smaller WWR to maintain privacy.

Figure 5. Data visualization and distribution of WWR in the Bronx, New York.
Source: Author, 2022.

5 Discussion

Urban visual data has been generated at an unprecedent speed with the rise of scene segmentation, which opens new opportunities to quantitively investigate urban physical environment from multiple perspectives. This study provides a method of evaluating WWR by integrating big geo-data from GSV images. We show how GSV along with Mask RCNN can be used efficiently to measure the features of building façade at a large scale.

Our work is limited in several aspects. First, the size of validation set is not quite large, thus may lead to overfitting in the testing set. In future work, we can add more annotated to obtain a larger dataset and achieve a maximum performance of the algorithm. Second, our self-annotated dataset contains only two annotation categories: windows and walls. Third, the segmentation of street corners remains an unsolved problem, since a specific image view does not fit all buildings with different orientations. Inevitably, a high score in the algorithm does not always mean the best state.

This paper serves as the first step in profiling the urban building facade with GSV and scene segmentation. We propose that this line of research can be extended in several ways. In the field of architecture and urban design, WWR is an interesting topic with a wide range of applications. For example, 1) Differences in WWR between cities in different geographic locations, i.e., seaside and inland 2) Differences in WWR between cultures, i.e., Islamic architecture has relatively small WWR compared with catholic buildings. And by calculating WWR on a large scale and quantitatively using google street view and computer vision, we can provide a reference for region or culture based architectural design.

Future works include training a more accurate and generalized model with annotated images from other cities, such as Paris and San Francisco. Our current dataset contains only images in Bronx, NY, where the building facades are quite simple. While considering the variety of building facades in different cities, the current model may not perform well in other cities. With a generalized model, it is possible to compare WWR among different districts of a city and even among different cities and reveal the correlation between WWR and other urban index, like the impact of culture or religion on building design.

6 Conclusion

Our aim is to develop and test an automated computer vision methodology to extract the WWR of building facades in a large scale, using state-of-the-art machine learning techniques and Google Street View (GSV) images. The methodology can be further extended to assess other features of the built environment that shape the visual experience, such as urban architectural style and the relationship between adjacent buildings.

By training the Mask RCNN model, we could achieve a satisfying performance on the annotated validation set, and the mean Average Precision (mAP) reaches 0.66. The result reveals that our model achieves a medium-togood performance of calculating WWR, which can shed light on future architectural design and urban development.

The main contributions of our paper are as follows: First, a machine learning model on GSV images for the large-scale and quantitative measurement of WWR, which is one of the key parameters involved with a meaningful understanding of the urban environment.

Second, an annotated GSV dataset that contains 500 GSV images labelled with categories of window and wall. The existing scene-centric datasets include a very limited number of images annotated with the window category and their perspectives are too oriented from the facades, which cause low accuracy of window segmentation.

Third, windows play an important role in 'Eyes on the Street' in Jane Jacobs' theory. However, vital street-level parameters in her theory, most notably windows, are rarely assessed at the urban scale due to imprecise existing datasets. To resolve this challenge, this study proposes an automated computer vision-based methodology to extract the window-to-wall ratios (WWRs) of buildings in the Bronx, New York, using semantic segmentation machine learning. This consistent measured maps on the WWR qualities in Bronx, which can be used for further studies on urban visual environment.

Acknowledgements. This research is based on courses in Senseable City Lab (SCL) and computer vision. We acknowledge Rui Wang, Nikita Klimenko, Fan Zhang, and Maoran Sun for offering help during the courses.

References

Charreire, H., Mackenbach, J. D., Ouasti, M., Lakerveld, J., Compernolle, S., Ben-Rebah, M., McKee, M., Brug, J., Rutter, H., & Oppert, J.-M. (2014). Using remote sensing to define environmental characteristics related to physical activity and dietary behaviours: A systematic review (the SPOTLIGHT project). Health & Place, 25, 1–9. https://doi.org/10.1016/j.healthplace.2013.09.017

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. ArXiv:1606.00915 [Cs]. http://arxiv.org/abs/1606.00915

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. ArXiv:1604.01685 [Cs]. http://arxiv.org/abs/1604.01685

Cozens, P., & Hillier, D. (2012). Revisiting Jane Jacobs's "Eyes on the Street" for the Twenty-First Century: Evidence from Environmental Criminology. The Urban Wisdom of Jane Jacobs, 313–344.

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep Learning the City: Quantifying Urban Perception at a Global Scale. In B. Leibe, J. Matas, N. Sebe,

& M. Welling (Eds.), Computer Vision – ECCV 2016 (pp. 196–212). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_12

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Forkenbrock, D. J., & Sheeley, J. (2004). Effective Methods for Environmental Justice Assessment. The National Academies Press. https://doi.org/10.17226/13694

Früh, C., & Zakhor, A. (2003). Constructing 3D city models by merging aerial and ground views. IEEE Computer Graphics and Applications, 23(6), 52–61. https://doi.org/10.1109/MCG.2003.1242382

Früh, C., & Zakhor, A. (2004). An Automated Method for Large-Scale, Ground-based City. . . International Journal of Computer Vision, 60(1), 5–24.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018). Mask R-CNN. ArXiv:1703.06870 [Cs]. http://arxiv.org/abs/1703.06870

Jane, J. (1961). The Death and Life of Great American... Book by Jane Jacobs. ThriftBooks. https://www.thriftbooks.com/w/the-death-and-life-of-great-american-cities_jane-jacobs/249572/

Koutsourakis, P., Simon, L., Teboul, O., Tziritas, G., & Paragios, N. (2009). Single view reconstruction using shape grammars for urban environments. 2009 IEEE 12th International Conference on Computer Vision, 1795–1802. https://doi.org/10.1109/ICCV.2009.5459400

Li, C.-K., Zhang, H.-X., Liu, J.-X., Zhang, Y.-Q., Zuo, S.-C., & Fang, Y.-T. (2020). Window Detection in Facades Using Heatmap Fusion. Journal of Computer Science Technology, 35(4), 900–912. https://doi.org/10.1007/s11390-020-0253-4

Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. Urban Forestry & Urban Greening, 14(3), 675–685. https://doi.org/10.1016/j.ufug.2015.06.006

Neuhold, G., Ollmann, T., Bulo, S. R., & Kontschieder, P. (2017). The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. 2017 IEEE International Conference on Computer Vision (ICCV), 5000–5009. https://doi.org/10.1109/ICCV.2017.534

Raskar, R., Naik, N., Philipoom, J., & Hidalgo, C. (2015). Streetscore—Predicting the Perceived Safety of One Million Streetscapes. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. https://doi.org/10.1109/CVPRW.2014.121

Riemenschneider, H., Krispel, U., Thaller, W., Donoser, M., Havemann, S., Fellner, D., & Bischof, H. (2012). Irregular lattices for complex shape grammar facade parsing. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 1640–1647. https://doi.org/10.1109/CVPR.2012.6247857

Rundle, A. G., Bader, M. D. M., Richards, C. A., Neckerman, K. M., & Teitler, J. O. (2011). Using Google Street View to Audit Neighborhood Environments. American

Journal of Preventive Medicine, 40(1), 94–100. https://doi.org/10.1016/j.amepre.2010.09.034

Rzotkiewicz, A., Pearson, A. L., Dougherty, B. V., Shortridge, A., & Wilson, N. (2018). Systematic review of the use of Google Street View in health research: Major themes, strengths, weaknesses and possibilities for future research. Health & Place, 52, 240–246. https://doi.org/10.1016/j.healthplace.2018.07.001

Sheikh, H. R., & Bovik, A. C. (2006). Image information and visual quality. IEEE Transactions on Image Processing, 15(2), 430–444. https://doi.org/10.1109/TIP.2005.859378

Shi, X., Wang, C., Wang, M., & Tang, P. (2020). An Innovative Approach to Determine Building Window- To-Wall Ratios for Urban Energy Simulation. Symposium on Simulation in Architecture + Urban Design, 57–60.

Szcześniak, J. T., Ang, Y. Q., Letellier-Duchesne, S., & Reinhart, C. F. (2022). A method for using street view imagery to auto-extract window-to-wall ratios and its relevance for urban-level daylighting and energy simulations. Building and Environment, 207, 108. https://doi.org/10.1016/j.buildenv.2021.108108

Taylor, B. T., Fernando, P., Bauman, A. E., Williamson, A., Craig, J. C., & Redman, S. (2011). Measuring the Quality of Public Open Space Using Google Earth. American Journal of Preventive Medicine, 40(2), 105–112. https://doi.org/10.1016/j.amepre.2010.10.024

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P., & Paragios, N. (2011). Shape grammar parsing via Reinforcement Learning. CVPR 2011, 2273–2280. https://doi.org/10.1109/CVPR.2011.5995319

Teboul, O., Simon, L., Koutsourakis, P., & Paragios, N. (2010). Segmentation of building facades using procedural shape priors. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 3105–3112. https://doi.org/10.1109/CVPR.2010.5540068

Tsai, V., & Chang, C.-T. (2012). FEATURE POSITIONING ON GOOGLE STREET VIEW PANORAMAS. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, I–4, 305–309. https://doi.org/10.5194/isprsannals-I-4-305-2012

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. Landscape and Urban Planning, 180, 148–160. https://doi.org/10.1016/j.landurbplan.2018.08.020

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid Scene Parsing Network. ArXiv:1612.01105 [Cs]. http://arxiv.org/abs/1612.01105

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2018). Semantic Understanding of Scenes through the ADE20K Dataset. ArXiv:1608.05442 [Cs]. http://arxiv.org/abs/1608.05442