ANALYSIS OF DIFFERENCES IN STREET VISUAL WALKABILITY PERCEPTION BETWEEN DCNN AND VIT MODEL BASED ON PANORAMIC STREET VIEW IMAGES

YUCHEN XIE[1], YUNQIN LI[2], JIAXIN ZHANG[3], JIAHAO ZHANG[4], ZHEYUAN KUANG[5]
[1,2,3,5]*Architecture and Design College, Nanchang University*
[2,3,4]*Graduate School of Engineering, Osaka University*
[1]*xieyuchen@ncu.edu.cn, 0009-0007-2938-5003*
[2]*liyunqin@ncu.edu.cn, 0000-0002-1886-0477*
[3]*jiaxin.arch@ncu.edu.cn, 0000-0002-6330-6723*
[4]*jiahao@is.ids.osaka-u.ac.jp, 0000-0001-5522-9759*
[5]*6008119039@email.ncu.edu.cn, 0009-0009-0184-6159*

Abstract. In measuring Urban Street Visual Walkability Perception (VWP) using Street View Images (SVIs), the VWP classification deep multitask learning (VWPCL) model based on the Deep Convolutional Neural Network (DCNN) shows notable deficiencies in recognizing local features within panoramic images. Addressing this, the study introduces a Vision Transformer (ViT)-based VWPCL model and employs various methods comparing its performance with DCNN. Initially, we assess the basic accuracy and validity performance using traditional metrics such as recall rates, and precision. Furthermore, we use the SHAP model for interpretable machine learning to analyse the significance and contribution of streetscape elements. Finally, the results of panoramic SVIs classification and feature display from different angles at the same location are compared by the Grad-CAM model to further visualise and explain the differences in feature elements that affect the classification of the computer vision model. Findings show the ViT-based VWPCL model, as compared to the traditional DCNN framework, mitigates image distortions in panoramic SVIs while demonstrating higher accuracy that aligns more closely with human visual cognition. The primary contribution of this study lies in qualitatively and quantitatively comparing the performance disparities between ViT and DCNN in the realm of street VWP.

Keywords. Visual Walkability Perception, Panoramic Street View Images, Deep Convolutional Neural Network, Vision Transformer, Grad-CAM.

1. Introduction

Urban studies have long explored how cities' appearance can be understood based on Street View Images (SVIs). Computer vision models based on 360-degree panoramic

ACCELERATED DESIGN, Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2024, Volume 2, 29-38.

SVIs have not only enhanced the efficiency of large-scale urban street perception measurements but also mitigated perceptual biases arising from the 90-degree conventional SVIs viewed from different angles. Prior research has measured street visual walkability perception using the VWP classification deep multitask learning (VWPCL)model based on the DCNN architecture and panoramic SVIs from VR audits. However, traditional computer vision models based on the Deep Convolutional Neural Network (DCNN) architecture exhibit certain limitations in recognizing panoramic SVIs, particularly regarding issues such as angular distortions in panoramas, limited capture of image-specific local features, and insufficient interpretability. In recent years, the rapid development of the Vision Transformer (ViT) model architecture has showcased outstanding performance in image classification and object detection, surpassing the performance ceiling set by DCNN. Nevertheless, the extent of trustworthiness regarding street visual walkability measurements provided by computer vision models employing different existing architectures remains unclear. Thus, the introduction of ViT in the street visual walkability research for comparison with the VWPCL model based on the DCNN architecture, concerning accuracy, validity, and reliability, holds significant importance. Within the deep feature interpretation of extensive panoramic SVIs, semantic segmentation algorithms can systematically analyse the proportion of physical elements in these SVIs. Feature interpretability models—the SHAP model offers explanations for predictive results by computing the impact of each physical element on predictions. Visual Interpretation Model Grad-CAM visualises the image areas that are in focus during the prediction process, providing an intuitive interpretation of the model predictions.

The objectives of this study are twofold: (1) to compare the performance of models based on DCNN and ViT architectures in street VWP, uncovering disparities in recognizing panoramic SVIs and understanding the importance of physical elements; (2) to comprehensively assess the performance in terms of accuracy, validity, and reliability, deeply understanding and validating the potential and advantages of ViT model in perceiving street visual walkability. Contributions of this research include:

(1) The proposal of a better performance VWPCL model, comparing the performance disparities in street VWP between DCNN and ViT architectures.

(2) Unveiling differences in understanding the importance and contribution of physical elements between DCNN and ViT in street VWP through analyses using two interpretable machine learning models.

(3) Revealing disparities in panoramic SVIs recognition capability and visual interpretability between DCNN and ViT.

2. Related Work

In the realm of image feature extraction, Convolutional Neural Network (CNN) have long dominated almost all image-processing tasks. Their more advanced DCNN, such as the DenseNet, have been widely adopted in the evolving field of image classification and have exhibited significant performance enhancements (G. Huang et al., 2017). However, the emerging deep learning model, ViT, is now becoming an alternative to CNN (Dosovitskiy et al., 2020). Lu et al. (2022) showed that ViT outperforms CNN after training on large, medium, and small datasets. While some studies have applied

ViT to image classification tasks in fields such as remote sensing and medical imaging (Bazi et al., 2021), research evaluating and recognizing panoramic SVIs using the ViT architecture in the street VWP domain remains limited. The mainstream approach still revolves around using models based on the CNN architecture (Y. Huang et al., 2023; Li et al., 2020). Additionally, most studies segment panoramic SVIs into 90-degree cubic patches to mitigate local angular distortions in panoramic images (Fan et al., 2023). However, this segmentation of cubic patches leads to the omission of physical elements like the sky, high-rise buildings, and trees, among others. Recently, there has been research comparing the performance of representative CNN and ViT in estimating land prices based on SVIs, yet there is a lack of in-depth evaluation of model differences through feature interpretation and visualized explanations (Zhao et al., 2023). Therefore, there is a pressing need for a more comprehensive and visual approach to examine and compare the performance of street VWP models based on DCNN and ViT architectures to enhance the accuracy, validity, and reliability of street visual walkability perception measurement models.

3. Methods and Datasets

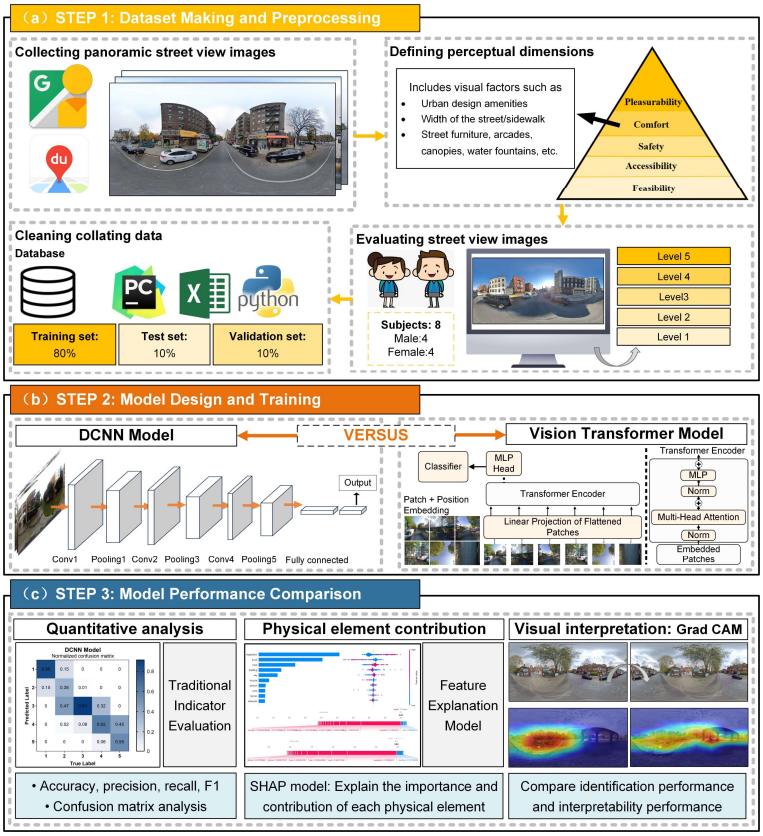

Figure 1. Research framework

This study has devised a three-phase research framework, as illustrated in Figure 1. The framework encompasses (1) Dataset making and preprocessing; (2) Model design and training; and (3) Comparative evaluation of model performance.

This framework not only enables a straightforward comparison of deep learning model performances based on different architectures but also facilitates a multifaceted explanation and validation of these models. In this study, we selected the sub-models of the comfort dimension under multiple dimensions of the VWPCL model for further study. The influences of this dimension include visual streetscape elements as well as psychological perceptual factors, thus allowing for a more comprehensive exploration of perceptual differences between models (Li et al., 2022).

3.1. DATASET MAKING AND PREPROCESSING

Figure 1(a) illustrates the process of dataset creation. The perceptual rating dataset comprises 4009 panoramic SVIs from 8 cities around the world, with different image styles from different streets highlighting the diversity of the dataset. Initially, panoramic SVIs were collected from Google and Baidu Street View, all sized at 1024*512 pixels. Following this, eight subjects with backgrounds in architecture and urban planning were invited to perform a five-category rating (1-5, from low to high) on all panoramic SVIs. To mitigate subjective bias, participants received training on the visual factors of the perception dimension of the pedestrian experience. Subsequently, the consistency of ratings from the subjects was assessed using Cronbach's Alpha, as per Equation (1). We observed high reliability and substantial consistency in the data (α =0.819>0.70).

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum S_i^2}{S_x^2}\right) \quad (1)$$

Where α is the reliability coefficient, K is the number of test questions, S_i^2 represents the variance of the scores of all subjects on question i, and S_x^2 is the variance of the total score obtained by all subjects.

3.2. DCNN, VIT MODEL DESIGN AND TRAINING

In the second phase, perceptual model designs based on DCNN and ViT architectures were developed. DCNN represents an evolved form of CNN. Its distinguishing features involve the convolutional layers extracting features from input data, pooling layers reducing the dimensionality of feature maps, and non-linear activation functions enhancing the model's non-linear expressive capacity. With an increase in network depth, DCNN can learn more abstract and intricate feature representations. On the other hand, ViT represents a neural network architecture based on attention mechanisms. ViT relies instead on self-attention mechanisms, allowing the network to capture global information from input images. This innovative architecture empowers ViT to exhibit exceptional performance in image classification and other visual tasks. Considering the performance disparities between these two model types, representative advanced architectures from DCNN and ViT, namely Desnet-169 and Swin V2, were selected for comparison. The architectures of Desnet-169 and Swin V2 are depicted in Figure 1(b). We partitioned the 4000 pairs of data in the Comfort perception dimension into training (80%), validation (10%), and testing (10%) datasets.

3.3. COMPARATIVE EVALUATION OF MODEL PERFORMANCE

3.3.1. Evaluation of Precision and Validity of Traditional Indicators

In the third phase, we initially used traditional performance evaluation metrics—accuracy, recall, precision, F1 score, and confusion matrix—to assess and compare the two models, thereby initially revealing the performance characteristics of the models concerning the accuracy and validity aspects of street visual walkability perception.

3.3.2. Credibility Evaluation via Semantic Segmentation and SHAP Feature Interpretation Model

Subsequently, employing the DeepLabv3+ model, we conducted semantic segmentation to compute the percentage of physical elements within the built environment of the streetscape, obtaining area ratios for 19 physical elements. Concurrently, utilising the SHapley Additive exPlanation (SHAP) model based on the concept of Shapley values from cooperative game theory, we ranked and assessed the importance and contributions of each physical element within all semantic segmentation maps. Notably, we provided explanations for typical individual samples, as depicted in Figure 1(c). The SHAP model's strength lies in its ability to handle intricate nonlinear relationships and facilitate both global and local explanations. Through these methodologies, we gained insights into the models' comprehension of various physical elements, thereby enabling further assessment of model performance.

3.3.3. Credibility Evaluation Through Visual Explanation Model Grad-Cam

Lastly, we aim to compare qualitatively and quantitatively the models' abilities in, +120°, and +180°, as illustrated in Figure 1(c). Subsequently, the two models predicted classifications for the shifted panoramic SVIs. We assessed the performance of the models in recognising the panned images compared to the original ones using two metrics: prediction accuracy and offset standard deviation. Higher accuracy indicates better performance in recognising panned images as the original ones, while a lower offset standard deviation suggests that the model's classification of panning images is closer to that of the original ones. The offset standard deviation is computed using Equation (2).

$$\sigma_{offset} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - C)^2}{n-1}} \quad (2)$$

Where σ_{offset} is the offset standard deviation， x_i is the classification of the panned image, C is the classification value of the original image, n is the number of images.

Subsequently, regarding angle distortion and visual interpretability of panoramic SVIs, we generated visual interpretation heatmaps using the classical class activation map technique, Grad-CAM, as depicted in Figure 1(c). Subjects were then asked to choose the heatmap that aligns more with their visual perception and provide reasons, thereby qualitatively assessing the model's performance. Grad-CAM calculates the importance weight of each feature computes the weighted activation of feature maps based on these weights, and produces the gradient-weighted class activation maps.

4. Experiments and Results

4.1. RESULTS OF EVALUATION OF PRECISION AND VALIDITY OF TRADITIONAL INDICATORS

The results of the comparison between DenseNet-169 and Swin V2 are shown below, and the evaluation metrics such as Accuracy, Recall, Precision, and Confusion Matrix evaluated based on 400 test samples are listed in Table 1 and Figure 2.

As shown in Table 1, the overall accuracy of the two models stands at 0.68 and 0.79, respectively. Swin V2 significantly outperforms DenseNet-169. DenseNet-169 exhibits an unsatisfactory recall rate of only 0.36 in Class 2, indicating ambiguity in its discrimination between Class 2 and 3. Correspondingly, in the macro average of each metric, Swin V2 surpasses DenseNet-169 by a considerable margin. The confusion matrix demonstrates that Swin V2 outperforms DenseNet-169 significantly in the classification tasks of levels 2, 4, and 5, equals in level 1, and slightly lags in level 3. Moreover, Swin V2 showcases fewer errors in off-diagonal positions compared to DenseNet-169. Hence, the comparison based on traditional metrics straightforwardly indicates the superior performance of Swin V2 across the entire dataset.

Table 1. Classification accuracy, precision, recall, and F1 score of the DCNN and ViT model

Type	Approach	Category	Precision	Recall	F1 score	No. samples
DCNN	DenseNet-169	Class 5	0.65	0.55	0.59	20
		Class 4	0.72	0.62	0.67	100
		Class 3	0.65	0.91	0.76	160
		Class 2	0.90	0.36	0.51	100
		Class 1	0.53	0.85	0.65	20
		Macro avg	0.69	0.66	0.64	400
ViT	Swin V2	Class 5	0.94	0.75	0.83	20
		Class 4	0.76	0.74	0.75	100
		Class 3	0.73	0.87	0.79	160
		Class 2	0.83	0.70	0.76	100
		Class 1	0.89	0.85	0.87	20
		Macro avg	**0.83**	**0.78**	**0.80**	400

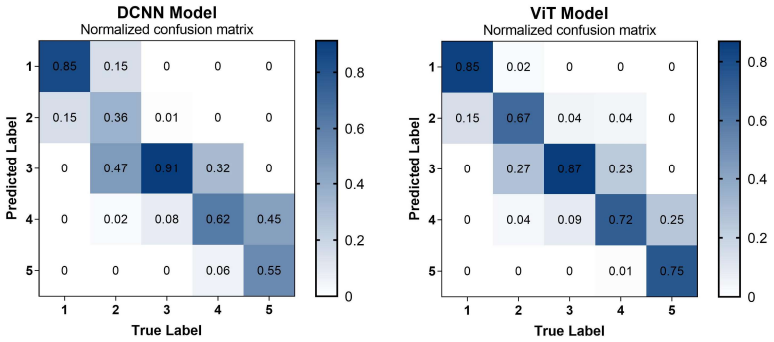

Figure 2. Confusion matrix of DCNN and ViT model

4.2. RESULTS OF CREDIBILITY EVALUATION VIA SEMANTIC SEGMENTATION AND SHAP FEATURE INTERPRETATION MODEL

However, qualitative comparisons of traditional model evaluation metrics may not fully reveal the model's performance, considering the significance of the model's understanding of physical elements. Utilizing the DeepLabv3+ model, we performed semantic segmentation on panoramic SVIs to calculate the area ratios of 19 physical elements. The top five elements by area ratio are road, sky, vegetation, building, and car, accounting for 0.30, 0.25, 0.13, 0.11, and 0.06, respectively. Then, SHAP model-based feature interpretation analysis was conducted on individual physical elements.

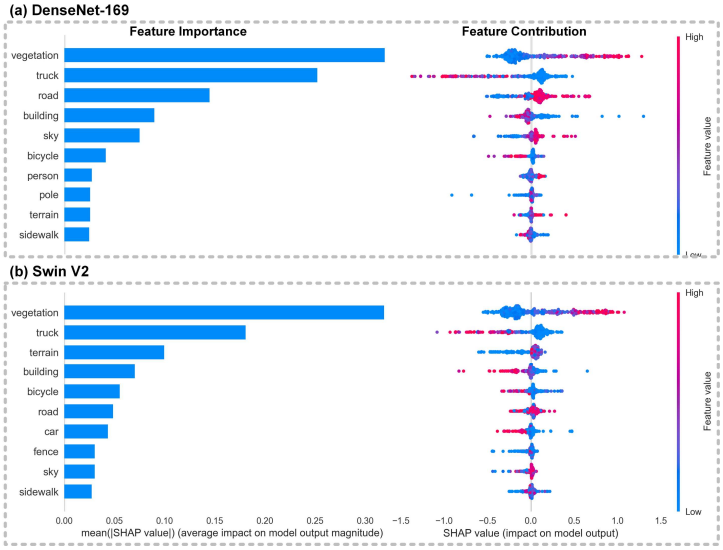

Figure 3. Feature importance and contribution of DCNN and ViT model

Figure 3 illustrates the data regarding the top 10 physically important and contributory ranked elements. It can be observed that DenseNet-169 and Swin V2 exhibit similar rankings in terms of the importance of most physical elements. However, disparities exist in certain key elements, such as the road and sky elements. DenseNet-169 ranked the road element in the top three in terms of positive importance. In contrast, Swin V2 identifies the road element as having some negative contribution. On the sky element, the two models differ in importance, although they consider the contribution to be similar. On the terrain element, which DenseNet-169 considers to have low importance and a neutral contribution, Swin V2 is contrasted.

In terms of interpreting individual samples, Swin V2 demonstrates a more accurate understanding of physical elements. Figure 4 presents two typical SVIs. In the Class 4 image on the left, both models identify vegetation elements as the primary positive factor, but they differ in the second-ranking positive factors, identified as truck and road elements, respectively. Concerning negative factors, DenseNet-169 highlights wall elements, whereas Swin V2 identifies building elements as the most significant. Observing the image, the road appears broad, unobstructed, and clean, evidently a positive factor, and no wall element is evident, with the majority being building structures. This indicates that Swin V2 can precisely and accurately identify primary influencing factors, whereas DenseNet-169 may misidentify. In the Class 1 image on

the right, both models recognise truck elements as the main negative element. In identifying the secondary negative element, DenseNet-169 identifies terrain, whereas Swin V2 identifies road and building elements. Human perception recognises congested roads filled with vehicles and older-looking building facades, contributing to low comfort. Therefore, Swin V2's identifications align more with human perception.

Figure 4. Local feature explanation of DCNN and ViT model

4.3. VISUAL INTERPRETATION OF MODEL ASSESSMENT CONFIDENCE RESULTS

4.3.1. Results of SVIs Comparison from Different Perspectives

As depicted in Table 2, both models demonstrate outstanding performance in identifying panning panoramic SVIs. They achieve accuracy rates of over 90% in recognising panning images as the original ones, with Swin V2 slightly outperforming DenseNet-169, achieving an overall accuracy of 95.50%. Regarding standard deviation, DenseNet-169 exhibits a higher value compared to Swin V2, signifying a higher level of dispersion in the model's predicted classification rating data for panning images, indicating less alignment with the original image's classification (σ=17.87＞11.31).

Table 2. Results of SVIs comparison from different angles of panning

Type	Approach	Accuracy			Overall Accuracy	Offset Standard Deviation
		Panning 60°	Panning 120°	Panning 180°		
DCNN	DenseNet-169	90.50%	96.00%	95.00%	93.83%	17.87
ViT	Swin V2	95.00%	95.50%	96.00%	**95.50%**	**11.31**

4.3.2. Results of Visual Interpretable Model Grad-Cam Comparison

We asked the subjects to choose the heat map of the panoramic SVIs generated through Grad CAM that was closer to their perception and provided reasons, enabling a qualitative and quantitative analysis. We selected representative panoramic SVIs from

Class 1, 3, and 5 (low, medium, high) for demonstration, as illustrated in Figure 5.

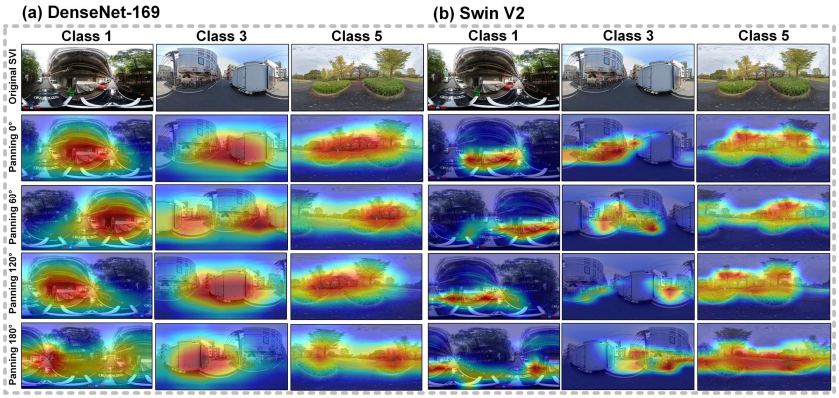

Figure 5. Results of visual interpretable model Grad-CAM comparison

As illustrated in the figure, darker colours represent a greater impact. In the original SVIs of Class 1, Swin V2 exhibits a more refined identification of the primary negative factors comprising waste bins in front of buildings, cars occupying the street, and deteriorating facades. However, DenseNet-169 considers buildings as the primary negative influence, which contradicts subject perceptions. In the panning panoramic SVIs, Swin V2 continues to identify waste bins as the main negative influence, while DenseNet-169 changes the element to the road. Within Class 3, Swin V2 demonstrates superior performance, identifying neatly arranged bicycles as the primary influencing factor, which starkly contrasts with DenseNet-169's recognition of a truck. Swin V2's predictions align entirely with subject perceptions. A travelling truck on a street does not significantly affect pedestrian perception, whereas bicycles neatly arranged on the pavement can positively influence pedestrians. Similarly, in panning images, Swin V2 still accurately identifies bicycles, while DenseNet-169 misidentifies the street as a negative factor in the 180° panning image. For Class 5, Swin V2 recognises the positive influences with semantic segmentation-like precision. Conversely, DenseNet-169's identification is coarse, and the discrepancy is more pronounced in the 120° panning image. Therefore, it can be concluded that compared to DenseNet-169, Swin V2 can more finely and accurately recognise physical elements that affect model classification and also can mitigate the effects of local distortions in local feature information in street VWP, simultaneously offering higher visual interpretability.

5. Discussion and Conclusion

This study compares qualitatively and quantitatively the street visual walkability models based on prevalent deep learning architectures. Primarily, we made a dataset comprising panoramic SVIs from multiple cities. Subsequently, two deep learning models based on distinct architectures were trained utilising this dataset. Finally, through a three-step method, we determined the performance differences between the models. Comparative analysis reveals that ViT's prevalent architecture, Swin V2, outperforms DenseNet-169 significantly in traditional metrics. Simultaneously, the evaluation results from the SHAP feature interpretation model indicate certain

discrepancies in the global feature assessment between the two models regarding the importance and contribution of elements. In local feature assessment, Swin V2 showcases a more rational and precise element recognition, whereas DenseNet-169 exhibits a certain degree of error. In the evaluation using the visual interpretability model, Grad-CAM, DenseNet-169's heat zones appear more diffuse, whereas Swin V2 not only excels in accuracy but also aligns better with human perception, improving accuracy, validity, and reliability in the model. ViT introduces a self-attention mechanism from the Transformer model that enables it to model global information about the entire image, giving ViT an advantage in interpreting model predictions and attention weights. In addition, it is better able to understand image structure and learn long-range dependencies between different regions in an image. In contrast, feature representations of DCNN are less sensitive to spatial information and are often difficult to interpret due to the loss of information during convolution and pooling operations.

The newly proposed VWPCL model in this study presents a novel method for street visual walkability perception measurement that aligns more closely with human visual subjective audits. In future work, we will incorporate models based on spherical CNN architectures for street vision walkability in the recognition of panoramic SVIs to eliminate the effect of angular distortions in panoramas to a greater extent.

References

Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021). Vision Transfo rmers for Remote Sensing Image Classification. *Remote Sensing*, *13*(3), Article 3. https:// doi.org/10.3390/rs13030516

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehgh ani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020, October 22). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXi v.Org. https://arxiv.org/abs/2010.11929v2

Fan, Z., Zhang, F., Loo, B. P. Y., & Ratti, C. (2023). Urban visual intelligence: Uncovering hi dden city profiles with street view images. *Proceedings of the National Academy of Scien ces*, *120*(27), e2220417120. https://doi.org/10.1073/pnas.2220417120

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). *Densely Connected Con volutional Networks*. 4700–4708. https://openaccess.thecvf.com/content_cvpr_2017/html/ Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html

Huang, Y., Zhang, F., Gao, Y., Tu, W., Duarte, F., Ratti, C., Guo, D., & Liu, Y. (2023). Comp rehensive urban space representation with varying numbers of street-level images. *Compu ters, Environment and Urban Systems*, *106*, 102043.

Li, Y., Yabuki, N., & Fukuda, T. (2022). Measuring visual walkability perception using panor amic street view images, virtual reality, and deep learning. *Sustainable Cities and Society*, *86*, 104140.

Li, Y., Yabuki, N., Fukuda, T., & Zhang, J. (2020). A big data evaluation of urban street walk ability using deep learning and environmental sensors-a case study around Osaka Universi ty Suita campus. *Proceedings of the 38th eCAADe Conference, TU Berlin, Berlin, Germa ny*, 319–328.

Lu, Z., Xie, H., Liu, C., & Zhang, Y. (2022). Bridging the Gap Between Vision Transformers and Convolutional Neural Networks on Small Datasets. *Advances in Neural Information Processing Systems*, *35*, 14663–14677.

Zhao, C., Ogawa, Y., Chen, S., Oki, T., & Sekimoto, Y. (2023). Quantitative land price analys is via computer vision from street view images. *Engineering Applications of Artificial Inte lligence*, *123*, 106294. https://doi.org/10.1016/j.engappai.2023.106294