Evaluating Elderly Perceptions of Green space in Ottawa Using Deep Learning

Zhewen Luo

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Master of Science in Geography

Department of Earth and Environmental Sciences
Faculty of Arts
University of Ottawa

Abstract

This thesis investigates the degree to which advanced GeoAI techniques, particularly deep learning models, and how these models can reproduce older adults perceptions of urban green spaces and to map the quality of green spaces across the city using street-level imagery. The study compares three popular deep-learning feature extractors to predict and map seniors' perceptual responses to green spaces in Ottawa, utilizing data derived from Mapillary street-level images. In these images, senior citizen volunteers were seated in front of a computer and selected between images based on perceptual dimensions such as safety and aesthetics. A few AI ranking models are constructed for the purpose of identifying how well street-level photographs are valued by seniors within Ottawa's green spaces. Our focus is mainly on finding the best-performing model and backbone training dataset via experimentation with various generations of deep learning architectures and using statistics to determine how such differences affect human perceptual modelling. Experimentation included the determination of the best ranking loss function for our data, comparing different loss functions, effects of hidden layers and how different pre-trained weights from different domains affect model performance. The RankNet algorithm demonstrated superior performance compared to RankSVM in a range of architectural configurations, including ViT-B16, VGG16, and ResNet50, with an average difference of 8.45% in accuracy for each perceptual attribute. The model deemed most effective overall adopted ViT-B16 as its foundation, providing the highest absolute accuracy despite the accuracy not being statistically different from VGG19 or ResNet50. Contrary to previous research we did not find the addition of Softmax to be of any significance in model preformance. Finally, we found that there is no statistically significant differences in using different-object-domain training weights in the transfer learning. Detailed prediction maps were predicted on park photographs from Mapillary that indicated how seniors perceive public spaces across the city. By combining deep learning, image segmentation, and spatial analysis, the research offers an empirical contribution of how AI can contribute insights into how urban planning can be enhanced to

meet the needs of an ageing population and creates a foundation for further investigation into age-friendly environments.

Acknowledgements

This work is a contribution to the Ottawa Neighborhood Study (ONS), as a part of a longitudinal study called *Using Artificial Intelligence Wisely: Age-Friendliness, 'Snow Moles' and Deep Mapping* funded by the Faculty of Social Sciences, University of Ottawa.

I extend my deepest gratitude to Dr. Michael Sawada for his invaluable guidance and supervision throughout the completion of this thesis. I also wish to express my heartfelt appreciation to Dr. Luke Copland and Dr. Anders Knudby for their insightful suggestions, which greatly enhanced the quality of this work. We thank Elizabeth Kristjansson, Daniel Baxter, Anita Acheson and all of the participants that produced data and/or contributed to focus groups for their time and effort.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Canada is experiencing rapid population aging, with seniors (65 years and older) making up an increasing percentage of the population. The 2016 census marked the first time that Canadians aged 65 and over outnumbered children aged 14 and under; in 2022, 18.53% of the overall Canadian population is aged 65 and over, compared to 12.55% at the beginning of the 21st century; it is estimated that by 2040, one-fourth of the total population will be seniors (Statistics Canada, 2022). This trend has important implications for urban planning, particularly with respect to making cities age-friendly. The World Health Organization (WHO) (2007) has identified eight domains crucial for age-friendly cities, which include outdoor spaces, social participation, and health services, in response to the situation of rapidly aging populations since it strikes not only in Canada but globally, and provided systematic and scientific advice to promote the well-being of those living in urban areas, especially for elderly individuals. Across Canada, communities have actively engaged in initiatives to enhance age-friendliness in line with these WHO recommendations, with approximately 1,000 communities actively engaged in initiatives aimed at enhancing age-friendliness in line with the WHO's eight domains (D. Baxter et al., 2017). Thus, it becomes imperative for these communities to develop well-informed plans related to the eight domains based on a thorough understanding of the community's requirements, which is essential for evaluating the actual status of age-friendliness in those communities.

Providing accessible and welcoming green spaces in urban areas is crucial to supporting seniors' physical, social, and mental well-being. The growing body of evidence on the benefits of public green spaces for the physical and mental health of older adults is encouraging. These spaces offer a promising means of both preventing health decline and providing intervention for those suffering from cognitive and physical disabilities (Akpinar, 2016; D. E. Baxter & Pelletier,

2019). It is more important to assess the perceived quality of green spaces than to determine proximity to a minimum quantity of green space. Research has shown that people prefer future investments in public green spaces to focus on improving the quality of green spaces rather than increasing the quantity (Madureira et al., 2018). Moreover, participant-derived evaluation of perceived environmental quality can be more important than that of expert-determined quality because individual perceptions of green space can influence the benefits derived from it (Y et al., 2017).

Street-level imagery is a source of geographic data which contains a large amount of information about place. The imagery refers to photos collected by map service providers like Mapillary, etc., that captured entire 360-degree panoramas using specialized vehicles (e.g., Google) or those that are taken and uploaded by users (crowdsourced) according to certain standards, along the roadways, paths or trails in cities and suburban areas. One of its applications in urban studies, thanks to the support of artificial intelligence technology, is visual perception assessment (Fan & Yu, 2021). Images capture semantic elements related to perceptions by visually conveying features of the urban environment that people subconsciously or consciously associate with certain qualities, such as safety, wealth, beauty, or liveliness. Studies of perception, such as those conducted in the Place Pulse project (Dubey et al., 2016; Salesses et al., 2013), require participants to evaluate an image based on its semantic elements, comparing these with their mental models of what constitutes a safe, wealthy, or beautiful space. These judgments are informed by prior experiences, societal norms, and expectations, making semantic image analysis a powerful tool for understanding urban perceptions.

This research represents one of the first attempts to apply deep learning specifically to seniors' perceptions of green space in a city, providing useful empirical insights into age-friendly urban design. The objective of this study is to investigate the most effective methodology for training a robust deep learning model using a limited set of street-level imagery that was assessed

according to a range of semantic categories that align with the concerns typically faced by older adults. To be explicit, the following research question is addressed: Can deep learning models be used to reproduce seniors' perception of green space using limited training data of approximately 4,000 pairs of images per perceptual dimension? In order to answer this question, a number of experimental questions will be undertaken:

1. Which model architecture produces the most accurate predictions?
2. How do different pre-trained (from a more semantically similar domain(s) or semantically more distant) model weights influence performance?
3. Do we need a more complex architecture with more layers or a simpler network, with regards to ranking function?

Chapter 2

Literature Review

This section will present an overview of current initiatives aimed at identifying key dimensions for age-friendliness, methods for collecting data to quantify perceptions of urban areas and explore deep mapping techniques to analyze these perceived attributes.

2.1 Applications of Street-level Imagery in Human Perception

Street-level imagery is a source of geographic data which contains a large amount of information about places. The imagery refers to photos collected by map service providers like Mapillary, Google Street View (GSV), Microsoft Street Side (US only) etc., that captured entire 360-degree panoramas using specialized vehicles (e.g., Google mapping automobiles) or those that are taken and uploaded by users (crowdsourced) according to certain standards (all services including Google allow this), along the roadways, paths or trails in cities and suburban areas. One street-level imagery application in urban studies, thanks to the support of artificial intelligence, is the application of street-level views in human visual perception assessment (Fan & Yu, 2021). Deep learning models can learn how a particular group of people perceive an urban scene represented by a street-level image under a particular semantic category; these models can then predict the people's perception of all relevant images in a short period over entire geographic regions like a large city of over 1 million people. The modelled perception levels or ranks can be assigned to the locations of those images, which can then be visualized through modern GIS technologies for mapping and applied to the predictions integrated spatial analyses.

The literature has examined the accessibility of urban amenities (Paez et al., 2010), spatial imbalance of urban facilities (Huang X. et al., 2022), the influence of street components on age diversity (An et al., 2023), the impact of neighbourhood land use on work participation and well-being of older people (Cheam & Bozovic-Stamenovic, 2023) and friendliness of urban

facilities concerning elderly adults (Yang et al., 2023). However, to the best of the authors' knowledge, few studies on the perceived attributes of urban public spaces have been discussed in the context of research on age-friendliness. Existing datasets that are available for analyzing perceptual attributes within urban areas are most frequently composed of manually labelled data collected through crowdsourcing. Place Pulse 1.0 is a dataset created through a crowdsourcing effort aimed at mapping urban perceptions and contains 4,136 geo-tagged images from two cities in the United States and two cities in Austria (Salesses et al., 2013). This dataset was created through a crowdsourced website where participants on the internet are shown pairs of random street-view images. They are then asked to express their preference for one image over the other in response to questions like “Which place looks more X?”, where X is a perceptual word like “safe”, “lively”, “beauty”, and “activity”. These were determined via two focus groups as the most important dimensions. To generate a truly global dataset of urban appearance, Dubey et al. (2016) developed an updated version called Place Pulse 2.0, containing 1.17 million pairwise comparisons of 110,988 images from 56 cities in 28 countries on 6 continents. A total of 81,630 online volunteers evaluated the dataset based on six perceptual dimensions: safety, liveliness, boredom, affluence, depression, and beauty. Our study data were also generated based on a Place-Pulse collection methodology, which will be elaborated on later.

2.2 Urban Perception Using Machine Learning

Several published studies have demonstrated the feasibility of quantifying urban perception through the application of machine learning methodologies. Naik et al. (2014) proposed a Streetscore algorithm based on support vector regression (SVR) and generic street-level imagery features to predict the perceived safety of street-level images within the United States using the data from Place Pulse 1.0. Porzi et al. (2015) identified the intermediate-level visual elements that contribute to the perception of safety in the Place Pulse 1.0 dataset and found that using a convolutional neural network approach had significant improvement over

conventional machine learning techniques like SVR. Dubey et al. (2016) proposed a Streetscore-CNN (SS-CNN) and Ranking Streetscore-CNN (RSS-CNN) that were built using a Siamese deep convolution architecture where weights were optimized by minimizing a joint classification loss and ranking loss function and training using the Place Pulse 2.0 dataset to predict the perceptual attributes of street-view images in terms of safety, beauty, affluence, boredom, depression, and liveliness. Across all the perceptual attributes, they achieved the best accuracy of ~70% using the global dataset. Fu et al. (2018) proposed a CNN-based StreetNet for crime type inference and crime rankings from street-view images. For an assessment of visual walkability, Blecic et al. (2018) and Zhou et al. (2019) predicted perceived ranked scores of walkability, employing the VGG16 backbone for image classification and SegNet backbone for image segmentation respectively. Blecic et al. (2018) however trained their model using human-ranked images on a Likert scale of 1 to 5 rather instead of using binary image pairwise comparisons. Wang et al. (2021), compared the performance of an end-to-end CNN-based model and a second CNN plus a random forest model applied to scenes of varying complexity in multiple regions of a Chinese city and suggested that model architecture and performance depend on the intricacy of different urban areas. Huang et al. (2023) found the correlation between residents' activities and human perception, by using a ViT-based model trained on MIT Place Pulse 2.0 dataset to extract semantic features images and using an informatics-based diversity indicator to quantify human activities.

In this study, we present the experimental results of a comparative analysis of the use of different deep learning architectures for feature extraction on the training dataset. Additionally, we examine the efficacy of various ranking methods and mappings to identify potential avenues for further research into the needs of elderly residents with respect to urban public green spaces. Finally, we use AI to help identify those environmental objects that are commonly found at the

highest, medium and lowest end of the ranking scales. These results may have practical applications in planning age-friendly cities.

Chapter 3

Data and Methodology

3.1 Elderly Perception of Green space in Ottawa

We identified key features of green space quality by asking two focus groups that comprised of seniors aged over 60. They had 7 and 3 participants respectively in a mix of male and female. Both groups were asked the questions related to green space (Table 1). Their responses were analysed and coded to determine four dimensions of green space that were used for preference capture based on street-level imagery (Table 2). The definition of the green space used in this paper concerns areas of vegetation, such as parks, community gardens, and grassy areas along roadsides. Waterfronts such as riverside or lakeside are not included.

Elderly perceptions of the four dimensions of green space quality were done through another two sessions. Each session involved 12 participants who were all aged 60 plus of male and female genders and various levels of mobility. During the sessions, the participants were being shown two side-by-side images and asked to choose either left or right image according to the qualitative question being asked. Their selections were recorded into a table to form a dataset for each dimension containing a left image identifier, a right image identifier and which the winner. The number of image pairs for each dimension is shown in Table 3.

Table 1. Green space-related Questions for the Focus Groups.

Green space-related Questions
• When you think about spending time in urban parks and trails in Ottawa, what comes to your mind? • What features of the parks and trails in Ottawa do you enjoy when you are there, and what makes you want to return to the park or trail? • When you have been at parks and trails in Ottawa, what can you remember made your experience easy and comfortable?

In total, 10,000 images downloaded from Maplillary were used in collecting elderly perceptions. The images were captured along accessible bike trails or pedestrian pathways within the Ottawa urban area and were taken during summertime under various environmental condititions, e.g., sunny and cloudy weather conditions, with no images taken at night or in the rain. Figure 1 illustrates the distribution of those Mapillary images.

Table 2. Four Dimensions of Green space Quality.

Dimension	Question Being Formed
Connection to Nature	In which environment do you think you would feel a stronger connection to nature?
Aesthetics	Which environment do you find more beautiful or appealing?
Safety	In which environment would you feel safest?
Activities	In which environment do you feel you would be best able to do the activities you enjoy?

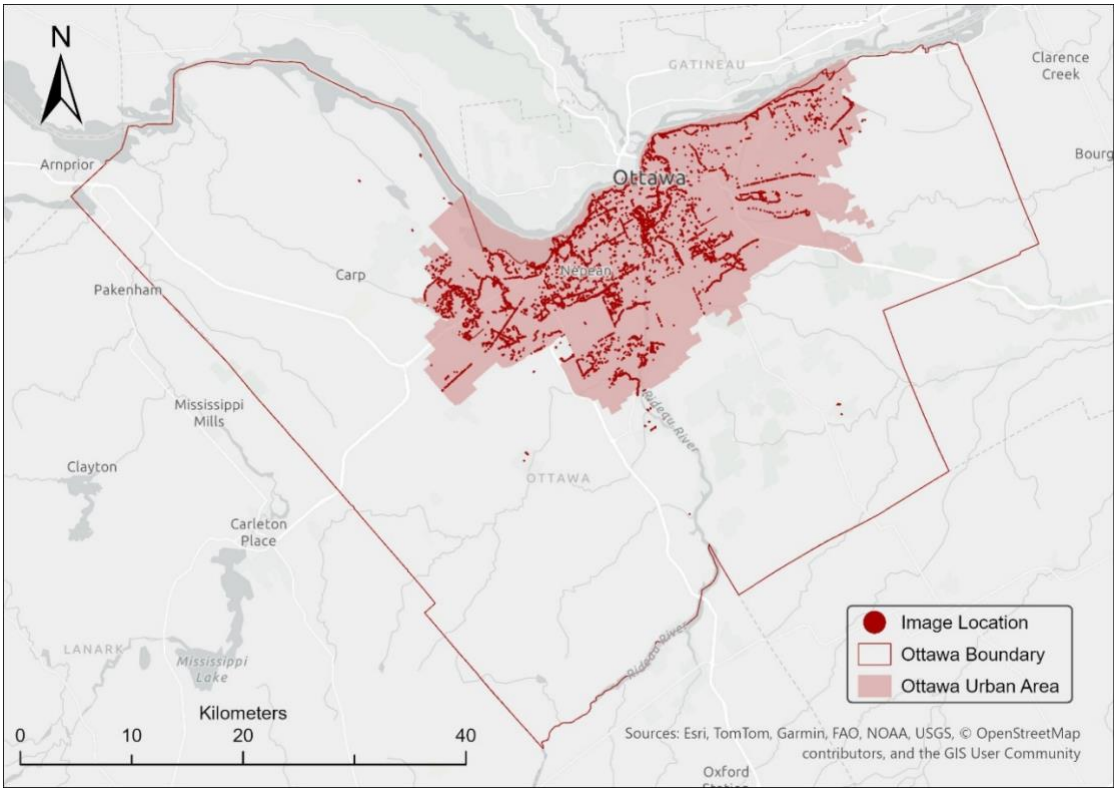

Figure 1. Distribution of Mapillary Image Samples.

Table 3. Amount of Selection Results for Each Perceptual Dimension.

Dimension	# Total	# Left-or-Right	# No-preference
Connection to Nature	5,780	4,563	1,217
Aesthetics	5,533	4,468	1,065
Safety	5,286	3,474	1,812
Activities	5,972	4,452	1,520

3.2 Siamese Deep Neural Network

To model elderly perception, we employed a Siamese deep neural network similar to the the Ranking Streetscore-CNN (RSS-CNN) (Dubey et al., 2016) that was modified to fit our specific questions and data. The RSS-CNN model consists of two identical backbone branches with shared weights for feature extraction, with a classification subnetwork connected to their concatenated output layers and two identical ranking subnetworks connected to each backbone branch. The classification subnetwork consists of a set of convolutional layers culminating in a fully connected layer (dense layer). The ranking subnetworks consist of dense layers with bounded weights. Figure 2 illustrates the model architecture of RSS-CNN.

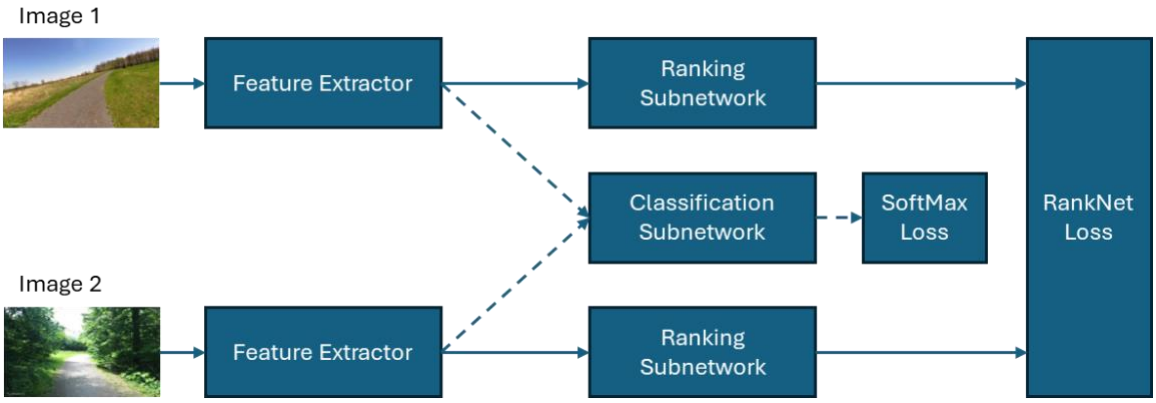

Figure 2. Architecture of Ranking Streetscore-CNN.

3.3 Image Feature Extractors with Transfer-learning

Different architectures of convolutional neural networks (CNN) (Elharrouss et al., 2022; LeCun et al., 2015) which are dominant architectures in computer vision, as well as new Vision Transformers (ViT) (Dosovitskiy et al., 2021) with self-attention-based mechanisms that have achieved scaling success for natural language processing, were adopted for experiments to extract deep features of images that are required for predicting perceptual responses.

VGG networks (Simonyan & Zisserman, 2015), developed by the Visual Geometry Group at the University of Oxford in 2014 and designed for large-scale image recognition tasks, is one of the early-stage fundamental convolutional neural networks (ConvNet) to emerge after the AlexNet breakthrough in 2012. In contrast to earlier convolutional networks that used larger and varied-size filters, it uses very small (3×3) convolutional filters that are applied uniformly throughout the network, resulting in greater depth. The simplicity and uniformity of the VGGNet has contributed to its popularity as a feature extractor for numerous other models, which are based on the VGG16 or VGG19 versions with 16 and 19 layers, respectively. These include those used for object detection, such as R-CNN and SSD, etc. (Elharrouss et al., 2022).

ResNet (He et al., 2015), short for Residual Networks, is another ground-breaking ConvNet architecture that was created to address the vanishing gradient problem (VGP) by training very deep convolutional networks. To cirumvent the VGP, the authors introduced residual connections, which create shortcuts between one or more layers that pass the input of the earlier layers directly to their output, to scale the network to 152 layers which is eight times deeper than VGG networks but has lower complexity. This result won 1st place in the ImageNet Challenge 2015. Different variants of ResNet have already been used widely for object detection and semantic segmentation, as well as leveraged for other deep learning architectures (Elharrouss et al., 2022).

Vision Transformer (ViT) (Dosovitskiy et al., 2021) represents a significant shift from convolutional neural networks by adapting the transformer architecture, originally designed for sequence-to-sequence tasks like natural language processing, to image classification. Specifically, the ViT divides an input image into fixed-size patches, typically 16x16 or 32x32 pixels, with each patch then flattened into a vector and linearly embedded into a higher-dimensional space. Positional encodings are added to the patch embeddings to retain spatial information. The core of ViT consists of multiple transformer encoder layers which includes multi-head self-attention mechanisms and feed-forward neural networks. These layers process the sequence of patch embeddings, allowing the model to learn complex dependencies and relationships between different parts of abstractions of the image.

The training data (Table 3) for any given question that seniors responded to was quite small from a 'deep learning' standard. Convolutional Neural Networks (ConvNets) are most often trained on millions of images. For example, most common deep learning models have weights from being trained on the ImageNet dataset which contains 1000 categories across millions of photographic images. Once a given architecture is trained on a large dataset of photograps, the model contains domain-specific weights that can be simply loaded and then fine-tuned on a dataset in a similar domain (street-level photographic images) that are much smaller in number. For example, a ConvNet trained on millions of photographs, in the first few layers, learns how to identify shapes of things like horizontal, vertical, and angular aspects of the real-world as seen in a photograph. That is because usually the ConvNet has millions of examples and so these geometric and abstract geometric shapes are well defined and defined so much better than if you used a ConvNet and trained it on 10,000 samples or even 50,000 samples. When the weights from a ConvNet with its perfectly abstracted geometry are available then you can use those in the same model but tune them for your task at hand. This is transfer-learning and is the standard in applied AI today because you get the perfect geometric representations (basically like a perfect edge

detection filter and other similar filters) and then keep them immutable except for the most abstract layers at the end. At that point, you can update the final abstractions from the already-trained ConvNet with your domain-specific data. Doing so generally provides a major boost in accuracy and prediction, far more so than if you only had a few thousand images or data points to model using a deep neural network. But one does not finish there, as you still need a lot of examples to tune those last more abstract layers for your domain-specific problem even when you have very small datasets.

There are various ways to make a small dataset seem larger for a ConvNet, such as augmenting the image by providing different affine transformation parameters so that the label remains the same, but the left and right images can be reused repeatedly, making the small dataset seem infinite. Thus, transfer learning is a requirement if one wants to take advantage of most of an existing model's weights. However, transfer learning begs the question as to whether models trained on different base datasets provide accuracy when applied to a dataset like ours which is street-level human perception. This is one reason why we test different weights using equivalent models and statistically compare them based on the input data they have used and that we have retrained on our data.

3.4 Ranking Algorithms

We formed the problem as described as following. The first component comprises a set of M images denoted by I represented as $\{x_i\}$ where i ranges from 1 to m, and these images exist in an n-dimensional pixel space (i.e., $I \in R^n$). The dataset's second component comprises N sets of triplets of perception results. These sets are labelled as P and denoted by the notation, $\{(i_k, j_k, y_k)\}$, for k ranging from 1 to N sets of triplet results, where $y \in \{1, -1\}$ indicates which image on the left or right in a pair is being chosen by a user. To reduce the difficulty of model training, all records labelled as "No-preference" (e.g., two images were perceived as equally safe to a participant) were excluded from the process of model training. Therefore, the perceived level

of images can be calculated as continuous scores using the pairwise Learning-to-Rank (LTR) approach. Thus, the perceptual level of an image can be modelled by methods concerning pairwise learning ranking (LTR), which is one of the applications of machine learning.

We converted these binary results (each pair's selection) into a relative ranking system that has a continuous scale, allowing a ranking score as an evaluation level (a so-called regression output) that was assigned to the winning input image and its corresponding geotagged location. After that, we rescaled the ranking scores into [0,1], making the ranking results easier to understand. Such a ranking system also facilitates detailed analysis by capturing the nuances of different street-level images under a specific perceptual dimension.

Originally, the RSS-CNN adopted a weighted sum of two losses which are a Softmax loss (cross-entropy) and a loss identical to the RankSVM formulation (Dubey et al., 2016; Joachims, 2002) for weights update. Nevertheless, newer and faster computational losses have been proposed, for instance, with RankNet (Burges et al., 2005) becoming a more popular and widely used option. We conducted a series of comparative experiments to determine the most appropriate approach for training the most reasonable model in the context of urban perception prediction. The experiments compared the performance differences between RankSVM and RankNet losses, along with matching different feature extractors.

3.5 Approaches of Spatial Characteristics

We used a variety of statistical and spatial techniques to analyze the relationships, and spatial patterns present in the prediction of ranked scores across Ottawa's parks from the AI model.

Pearson's R (Pearson, 1895) was used to assess the global linear correlation between variables, providing a general sense of how the variables were related across the study area. Because spatial autocorrelation was present between the variables used in Pearson's R, inflating the Type 1 error rates, we also used Lee's L (Lee, 2001) to control for spatial autocorrelation and

thereby provide unbiased correlations and ascertain the degree to which spatial autocorrelation affects the correlation between ranked output and thus understand how biased the Pearson's measure is.

Moran's I (Moran, 1950) was employed to detect spatial autocorrelation and identify potential outliers in the ranked score predictions. The global measure of spatial autocorrelation facilitated the identification of how similar a variable's value is to its neighbours. When spatial autocorrelation is present, Pearson's R is biased because not all data is statistically independent. Local Moran's I, alternatively, decompose the global value of Moran's I across all spatial locations with predicted ranks and can identify clusters of high or low values and areas where spatial patterns diverge from the overall trend (high surrounded by low or vice versa), thereby indicating the presence of spatial outliers or hotspots. Moran's I is a particularly useful tool in spatial analysis, as it enables the determination of whether a given spatial distribution has values whereby neighbouriung ranks are similar, dispersed, or random. This approach facilitated the identification of areas exhibiting significant spatial correlations between the two variables, thereby highlighting local patterns and anomalies.

Kriging interpolation (Chilès & Desassis, 2018) was used to estimate values in unsampled areas of the green spaces/parks based on the spatial structure of the sampled data. By creating a continuous surface, kriging allowed us to make predictions about unsampled regions with a measure of uncertainty, providing more complete spatial coverage. The technique exploits spatial autocorrelation and provides an optimal linear unbiased estimate, which is critical for accurate interpolation in geographic studies.

By combining these techniques, we were able to comprehensively analyze spatial relationships, detect anomalies, and predict values in unsampled areas, providing deeper insights into the spatial dynamics of the study area.

3.6 Predicted Images from Mapillary and Google Street View

Two distinct sources were utilized to obtain the street-level imagery utilized in this investigation. One set of data was obtained from Mapillary, the same source used for model training. A separate collection of images uploaded by any user in 2017 was used. The rationale for employing this dataset is that it comprises a sufficient number, over one hundred thousand, of images available for download in comparison to other years. The second data set was obtained from Google Street View (GSV). The objective of this assessment is to evaluate the model's capability to predict images with analogous semantics to the street environment even though they were gathered from disparate sources. The prediction dataset comprises over 20,000 images.

Chapter 4

Experiment and Results

To implement the model, it was first necessary to extract bottleneck features using pre-trained models sourced from available data sets. These pre-trained models, renowned for their robust feature extraction capabilities, were loaded and employed to process the input data, thereby obtaining bottleneck features, which are intermediate representations that capture the essential characteristics of the input data in an abstract yet high-level form. Subsequently, the bottleneck features were stored for future utilization in model training. Next, the top layers, which are fully connected layers, were added and trained on the extracted features followed by an output layer with an activation function that was appropriate for the task requirements.

To evaluate the model, a five-fold cross-validation was conducted to assess the model's robustness and generalizability across different subsets of the data. Subsequently, the model was retrained on the entire training dataset to leverage all available data, and its performance was evaluated on a separate test dataset that was withheld during the initial training phase. This comprehensive approach ensures that the model is both well-tuned and capable of generalizing effectively to new, unseen data if the model provides sufficient accuracy.

Subsequently, the dataset was partitioned into a training set and a hold-out validation set, with 80% of the data allocated for training and 20% reserved for validation (Table 4). The model was compiled and trained on the training data while continuously validating its performance on the hold-out set, thereby monitoring and preventing overfitting. This step is particularly conducted to find out the best hyperparameters of top layers, explicitly including the number of layers, the units of each layer, and the dropout rates at the output. We adopted the Bayesian Optimization approach in this hyperparameter tuning process which can be easily implemented through Keras Tuner (Chollet et al., 2015).

Table 4. Statistics of Dataset for Each Perceptual Dimension.

Dataset of Dimension	# Training (80%)	# Testing (20%)	# Total
Connection to nature	3,629	907	4,536
Aesthetics	3,516	879	4,395
Safety	2,754	688	3,442
Activities	3,520	880	4,400

4.1 Experiment 1: RankNet vs. RankSVM

The RankNet loss and the RankSVM loss were initially employed as the cost function for the model. Thus, we undertook cross-validation for the three tuned models (ViT-B16, VGG16 and ResNet50). The results are presented in Table 5. It can be concluded that the use of RankSVM or RankNet as a loss function under a variety of paired model comparisons yields statistically significant differences in all cases. The nature of the differncdes for all model architectures is that RankNet provides higher accuracies than RankSVM. These notable differences in accuracy between RankNet and RankSVM substantiate the assertion that RankNet is a superior cost function for a ranking model when confronted with tasks similar to ours. Other researchers undertaking similar domain rankings should assess both to determine the degree to which our findings can be generalized.

Table 5. Cross-validation Results RankSVM Loss and RankNet with values and t-tests.

Base Model	ViT-B16		VGG16		ResNet50	
Loss Type	RankSVM	RankNet	RankSVM	RankNet	RankSVM	RankNet
Accuracy Per Fold	77.27%	85.14%	72.59%	82.56%	71.70%	83.18%
	77.01%	83.39%	76.68%	84.26%	77.33%	85.54%
	77.48%	85.09%	72.81%	83.33%	75.93%	84.75%
	76.26%	82.50%	73.59%	82.07%	77.75%	84.39%
	76.27%	83.95%	73.42%	84.84%	77.43%	85.38%
Avg.	76.86%	84.01%	73.82%	83.41%	76.03%	84.65%

SD	0.57%	1.13%	1.65%	1.15%	2.52%	0.95%
Paired T-test p-value	1.68e-05		7.91e-05		0.00021	

4.2 Experiment 2: Does adding a Softmax loss for model training provide significantly different results?

The objective of this experiment was to ascertain whether incorporating a Softmax loss into the model would enhance its performance. As evidenced in Table 6, the application of a Softmax loss did not result in an enhancement of ranking prediction accuracy. This finding is at odds with the conclusion reached by Dubey et al. (2016), who asserted that an end-to-end learning model that incorporates both classification and ranking losses outperforms a model trained with only classification losses. This is because the classification and ranking functions are two distinct sub-networks, with their respective weights optimized through the application of a Softmax loss and a ranking loss, respectively. The transformer model is particularly vulnerable to Softmax loss and offers a promising avenue for future research on how ranking loss functions are constructed, as the lower *p*-value, although not significant, found between Softmax vs combined for the vision transformer is relatively low pointing to a stronger likelihood of finding significant differences as a function of loss function construction.

Table 6. Cross-validation results of two training configurations, A and B, which indicate with RankNet loss only and the sum of the RankNet loss and Softmax loss. Within each paired t-test, none of the average accuracies differ significantly ($p<0.05$).

Base Model	ViT-B16		VGG16		ResNet50	
Loss Type	A	B	A	B	A	B
	85.14%	83.99%	82.56%	83.39%	83.18%	82.84%
Accuracy Per	83.39%	82.98%	84.26%	83.52%	85.54%	85.75%
Fold	85.09%	85.03%	83.33%	84.20%	84.75%	84.82%
	82.50%	83.04%	82.07%	82.47%	84.39%	84.93%

	83.95%	82.66%	84.84%	83.15%	85.38%	85.05%
Avg.	84.01%	83.54%	83.41%	83.35%	84.65%	84.68%
SD	1.13%	0.97%	1.15%	0.63%	0.95%	1.09%
Paired T-test p-value	0.12		0.45		0.43	

4.3 Experiment 3: What effect does adding a hidden layer have on model training

The efficacy of incorporating a concealed layer into the model training process was investigated. Table 7 shows that the utilization of VGG16 and ResNet50 as the base model showed no statistical significant differences between models. However, a notable discrepancy was observed in the case of ViT-B16 where a statistically significant differnce between models is observed. However, as can be seen from the training curves shown in Figure 3, adding hidden layers can reduce learning efficiency and reduce the consistency of each training round. Therefore, we conclude that for such datasets as ours, using a single output neuron is sufficient.

Table 7. Cross-validation results of the experiment two model configurations: A indicates the model was trained using one output layer only, while B indicates adding one fully connected layer. Within each paired t-test, only the average accuracies of the model with ViT-B16 differ significantly ($p<0.05$ and Cohen's $d = -0.76$).

Base Model	ViT-B16		VGG16		ResNet50	
Loss Type	A	B	A	B	A	B
Accuracy Per Fold	85.14%	84.81%	82.56%	83.24%	83.18%	84.33%
	83.39%	83.86%	84.26%	82.44%	85.54%	84.34%
	85.09%	86.05%	83.33%	83.79%	84.75%	84.47%
	82.50%	84.33%	82.07%	82.34%	84.39%	84.66%
	83.95%	84.77%	84.84%	83.95%	85.38%	84.98%
Avg.	84.01%	84.76%	83.41%	83.15%	84.65%	84.55%
SD	1.13%	0.82%	1.15%	0.75%	0.95%	0.27%
Paired T-test p-value	0.04998		0.31		0.41	

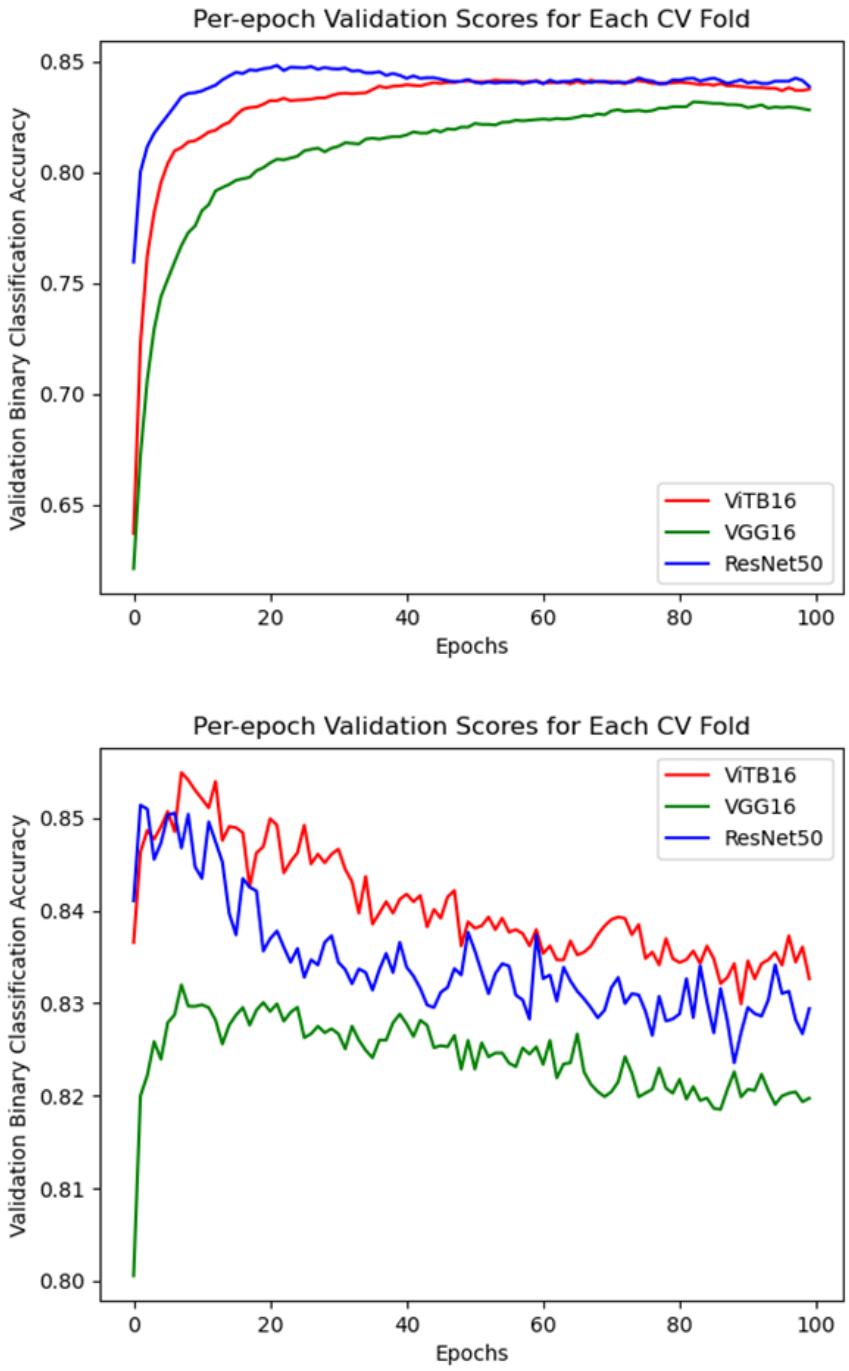

Figure 3. Curves of per-epoch validation scores for each cross-validation fold of experiment 3. The upper and lower panels show the case with no hidden layer added and with a hidden layer added, respectively.

4.4 Experiment 4: How do different pre-trained weights like Places365, ImageNet-1k and Hybrid (mix of Places365 and ImageNet) weights affect model accuracy

We conducted an experiment utilizing the VGG16 base model with pre-trained weights derived from the Places365, ImageNet-1k, and Hybrid datasets. Places365 (B. Zhou et al., 2018) is a scene recognition dataset, which is composed of 10 million images comprising 434 scene classes. ImageNet-1k, as aforementioned, is dataset which contains 1,000 categories across millions of images of generic objects. The findings shown in Table 8 along with Figure 4 show that utilizing the VGG16 model with pre-trained weights from Places365 resulted in enhanced accuracy compared to employing weights from ImageNet or Hybrid. However, Places365 shows statistically significantly higher accuracy than using the other pre-trained weights except in the case of hybrid weights, although Places 365 shows higher absolute accuracy for the majority of folds.

Table 8. Cross-validation results of the experiment of using different pre-trained weights on VGG16 feature extractor.

Pretrained Weights	In-1k	Places365	Hybrid
Accuracy Per Fold	82.23%	83.52%	83.57%
	82.23%	85.61%	84.81%
	83.93%	85.83%	84.28%
	82.88%	83.63%	82.55%
	82.26%	85.38%	84.30%
Avg.	82.70%	84.80%	83.90%
SD	0.74%	1.12%	0.87%
In-group T-test	In-1k vs. Places365	In-1k vs. Hybrid	Places365 vs. Hybrid
p-value	0.015	0.089	0.028

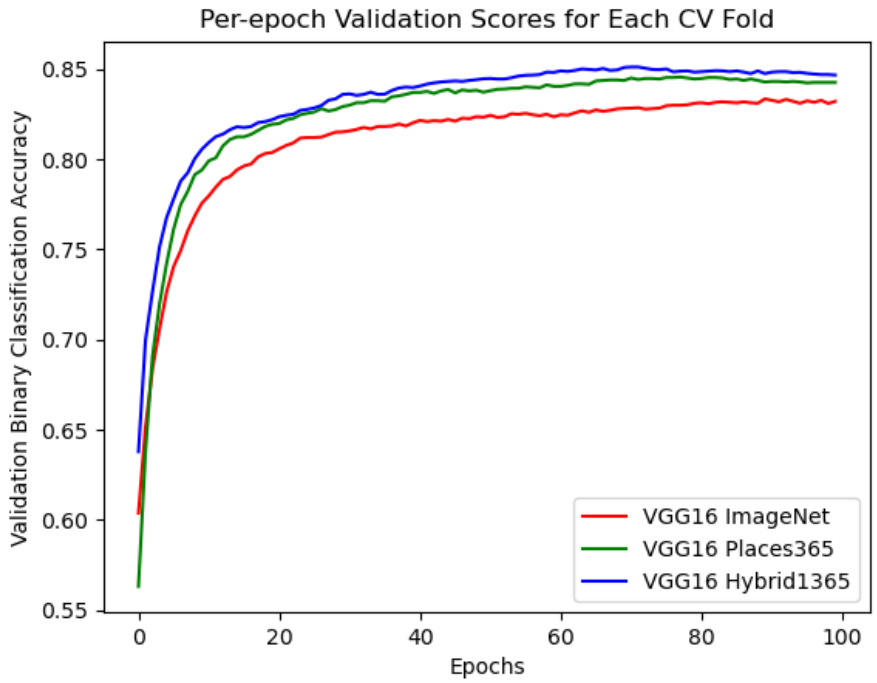

Figure 4. Curves of Per-epoch Validation Scores for Each Cross-validation Fold of Experiment 4

4.5 Predicting and Mapping Different Perceptual Dimensions

The base models used for prediction in the other dimensions were ViT-B16 (In-1k) and VGG16 (Places365). The results of the t-test shown in Table 9 and the final evaluation using testing datasets shown in Table 10 indicate that there is no statistically significant difference in the predictions made by these two models. It appears that the two models treat "connection with nature" in a similar manner, yet VGG16 (Places365) is observed to perform more effectively in several dimensions.

Moreover, the findings indicate that perceptual accuracy is predominantly affected by the perceptual category, rather than the quantity of data, as the training data sets for each dimension encompass a comparable number of image pairs. It is evident that the concept of "connection with nature" is subject to a relatively consistent standard of judgment among older individuals, whereas the selection of an image perceived as more aesthetically pleasing may be influenced by

a more diverse range of standards. It can thus be surmised that there may be a greater degree of bias present in some of the more ambiguous dimensions, which presents a significant challenge for deep learning models.

Table 9. Cross-validation results for predictions of different perceptual dimensions. "A" indicates the model using ViT-B16 pre-trained on In-1k as the base plus one hidden layer with RankNet loss; "B" indicates the model using VGG16 pre-trained on Places365 as the base plus only the output layer with RankNet loss.

Accuracy	Connection to nature		Aesthetics		Safety		Activity Enjoy	
Model	A	B	A	B	A	B	A	B
Accuracy Per Fold	84.81%	83.52%	73.15%	75.57%	73.69%	75.35%	65.91%	66.76%
	83.86%	85.61%	71.12%	71.00%	76.64%	75.87%	69.60%	68.47%
	86.05%	85.83%	71.70%	74.27%	73.31%	75.67%	69.60%	70.17%
	84.33%	83.63%	71.98%	71.00%	71.95%	72.54%	68.61%	66.62%
	84.77%	85.38%	75.40%	74.97%	74.42%	74.94%	68.47%	71.88%
Avg.	84.76%	84.80%	72.67%	73.36%	74.00%	74.87%	68.44%	68.78%
SD	0.82%	1.12%	1.70%	2.21%	1.73%	1.35%	1.51%	2.26%
Paired T-test	0.95		0.41		0.18		0.73	

Figure 5 illustrates the results generated by a model utilizing ViT-In1k as a feature extractor. Our findings indicate that images pertaining to natural settings are accorded a higher ranking than those featuring more concrete objects like buildings. These images suggest a greater sense of openness and greenness, within the parks the built-up areas are perceived to be safer than narrow, dark scenes. The activity dimension may seem contradictory, with a rather drab-looking image ranked highly and green spaces with sunny trails and a lot of vegetation very near the trailing edge. In deciding to compare imagery according to the activities that they do as aged individuals, it could make sense that wide open green space (weather seems irrelevant to the models as it should) is where perhaps fewer mobile individuals can undertake their activity, as

opposed to the more rugged vegetated trails preferred by mountain bikers and young hikers. This is consistent with the safety dimension where highly treed trails are ranked low and as vegetation decreases to open green space with a paved pathway, the scores go much higher. Treed-managed landscapes are very important to aesthetics and the connection to nature, these indicate a preference for more parkland vistas. These are physical objects within space that can be managed and contribute to the planning or redevelopment of urban space for a more aged population.

Table 10. Final Evaluation Results of the Two Models

Perceptual Dimension	Model 1: RankNet (with one hidden layer) with ViTB16-ImageNet	Model 2: RankNet (with output layer only) with VGG16-Places365
Connection to nature	85.13%	85.03%
Aesthetics	72.60%	73.74%
Safety	76.56%	77.41%
Activity to Enjoy	69.08%	67.75%

Figure 6 illustrates the predicted dimensions for each park. It is evident that the outcomes vary significantly with alterations in the dimensions. Because the ranking scales from each model are different because they are arbitrary (e.g., what is a 2 in “Beauty” is not necessarily equal to a 2 in “Activity”), thus the data was standardized for easy comparison visually so that high values have the larges colour value and low values have the smallest colour value. One of the unfortunate results of using data in Parks is the linearly clustered data, this is because most managed green spaces like Ottawa’s parks are traversed on shared pathways or trails by foot and bicycle. The question of what this means in terms of ranking an entire park or deciding which park is best and which is worst may need supplementary data, for example from satellite imagery or census variables. However, the most effective data would be a regular distribution or

randomized stratified distribution of photos. Unfortunately, these just do not exist. The following section will address the issue of spatial distribution anomalies in the prediction results.

Figure 5. Example of the results of the prediction generated by the model using ViT-In1k as the feature extractor of Mapillary images collected within a 20-meter buffer around Ottawa's parks and green spaces; the images are selected from the quantile sample in descending order from highest to lowest.

Figure 6. Examples of mapping results for each dimension: Connection to Nature (upper left), Aesthetics (upper right), Safety (lower left), and Activities (lower right).

4.6 Linear Correlation Analysis

The correlation analysis between four variables—connection to nature, aesthetics, safety, and activities—yields substantial insights into their interrelationships. The results of the Pearson correlation, Spearman correlation and Lee's L (Figure 7) demonstrate a strong positive correlation between Connection to Nature and Aesthetics, as well as between Connection to Nature and Activities. This indicates that environments with a high degree of connection to nature tend to be more aesthetically pleasing and offer a greater range of activities. In contrast, the correlation between Safety and both Connection to Nature and Aesthetics is moderately negative. This suggests that areas perceived as more connected to nature or aesthetically appealing may be considered less safe. The relationship between safety and activities is weaker and negative, indicating a slight trade-off between safety and the availability of activities. Overall, the analysis suggests a potential balance between natural, aesthetic environments and perceived safety, whereby areas rich in natural elements and activities may not always be associated with high safety ratings.

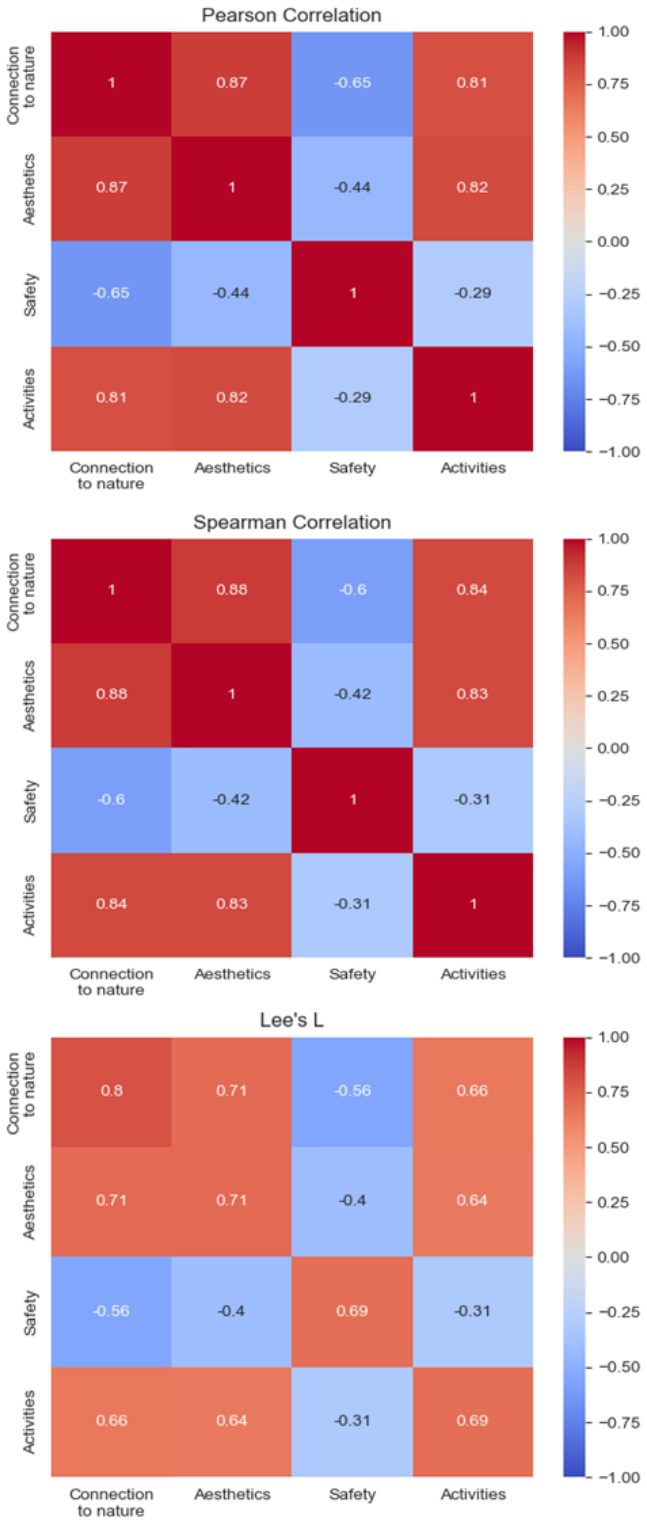

Figure 7. Matrices of correlation analysis across all perceptual dimensions.

Chapter 5

Discussion

5.1 Significances of the Experiments

Throughout the four experiments, we tested several hypotheses related to model performance, with an emphasis on improving ranking accuracy and examining the impact of architectural choices.

The comparison between RankNet and RankSVM revealed that RankNet consistently outperformed RankSVM across various architectures, including ViT-B16, VGG16, and ResNet50. This finding is significant because it suggests that RankNet is better suited for perceptual ranking tasks, likely due to its ability to handle continuous ranking scores more effectively than RankSVM, which was originally designed for binary classification problems. These results mirror the findings of RankNet (Burges et al., 2005), who demonstrated that the gradient-based optimization used in RankNet provides superior model generalization for ranking tasks. Min et al. (2020) conducted a comparable analysis of the utilization of RankNet and RankSVM for RSS-CNN training on the Place Pulse 2.0 dataset. The results indicated that the influence of these two losses is minimal regarding the performance of RSS-CNN, with a mere 0.5% difference in the accuracy of each perceptual attribute in Place Pulse 2.0. However, our findings contradict those of the previous study, with an average difference of 8.45% in accuracy for each perceptual attribute, which may be due to the use of a very small dataset for training or may be relevant to the perceptual attributes of the dataset for training. It is possible that connection to nature may contain less variance since older people generally tend to choose the same scenes as a better connection to nature. As there is a paucity of literature examining analogous perceptual dimensions within the authors' knowledge base, it is hoped that future studies will examine and elaborate upon this matter further.

The second experiment assessed whether incorporating a Softmax loss would improve ranking accuracy, given that the Softmax function can help with classification tasks by separating classes more distinctly. However, the findings indicated that Softmax loss did not improve model performance. This contradicts previous findings by Dubey et al. (2016), who proposed that combining ranking and classification losses could enhance model generalization. The discrepancy may be attributed to the nature of the data, where ranking preferences are more complex and subjective, and classification-driven learning may not fully capture the nuances of human perception. Similarly, many previous studies that have employed the RSS-CNN framework have all included Softmax for classification output. However, we present a more streamlined version of the modelling approach that can be used to address this ranking task from a human perception perspective. This may prove beneficial in situations where the available hardware is less sophisticated, as it could facilitate more efficient training of models.

In the third experiment, the inclusion of an additional hidden layer was explored. Results demonstrated that while adding a hidden layer did not significantly enhance model performance in VGG16 and ResNet50, there was a small improvement when using ViT-B16. This suggests that for certain architectures, such as transformer-based models, deeper networks can capture more complex patterns, though they may also introduce inefficiencies, as seen in the training curves. These results align with He et al. (2015), who emphasized that residual networks perform better without unnecessarily deep layers unless the additional depth directly aids in capturing more significant hierarchical representations. Furthermore, due to the absence of comparable datasets in size and the perceptual attributes in other research, we have presented this as a mere empirical observation rather than a definitive conclusion. Generally, it is widely accepted that models with a reduced number of layers are more advantageous when training with limited data.

Finally, we investigated the impact of different pre-trained weights—Places365, ImageNet-1k, and Hybrid—on model accuracy. The VGG16 model pre-trained on Places365

consistently outperformed those pre-trained on ImageNet-1k and Hybrid. This finding is crucial, as it underscores the importance of using domain-specific pre-trained models, especially when dealing with perceptual tasks such as landscape assessment. The superiority of Places365-pretrained models is consistent with the work of Zhou et al. (2018), who demonstrated that models pre-trained on scene-centric datasets provide better contextual feature extraction for environmental assessments. Here we show that to be true but also that object-based dataset weights like ImageNet-1k and the Hybrid still produce viable models and that the difference in domain alignment of the feature extractor for a well-tuned model may not be a critical factor if the domain is similar such as photographs of objects and backgrounds or places.

Table 11. Performance comparison of experiment results of our and other research in safety and aesthetics.

Model	Safety	Aesthetics	Dataset
Ours (ViT-B16)	76.56%	72.60%	Perceptions of Ottawa elder adults
RSS-CNN (Dubey et al.)	73.50%	70.20%	Place Pulse 2.0
MTDRALN-TC (Min et al.)	65.82%	72.32%	
Xu et al.	64.87%	69.20%	
DRAMA (Guan et al.)	64.81%	68.74%	

Table 11 demonstrates the efficacy of our optimal model when benchmarked against several leading models from other research studies. Only semantically similar dimensions are used for this comparison. The analysis focuses on two key perceptual attributes: safety and aesthetic perception. These attributes are particularly relevant for Place Pulse 2.0 and our dataset. Our model outperforms many existing urban perception models that employ Place Pulse 2.0 as the training dataset, particularly in terms of related perceptual attributes such as safety and aesthetics. This is because, as shown in the table, using a narrower dataset can improve the

accuracy of model projections by limiting the sampled population to a specific age group and area, which is rational, as there is greater consistency in the subjects' preferences when selecting images during data collection, resulting in a more focused representation of the intrinsic perceptual characteristics in the data.

5.2 Detection of Outliers of Ranking Results

To identify outliers in the prediction maps, we employ Local Moran's I statistic and the squared inverse distance as the definition of adjacency. Figure 8 illustrates an instance of outliers in a park area. From this example, it can be observed that the presence of outliers may be attributed to two primary factors: inaccurate model predictions and the presence of noise in the image. The latter may take the form of irrelevant shadows or the presence of people in the image, for instance. This finding indicates that a more powerful computer vision model should be employed that can resist the effects of noise and that performs preprocessing of noise cancellation before training.

Figure 8. A sample of detecting outliers of ranking scores.

5.3 Attempts of Kriging Interpolation for Mapping Spatial Trends

The map of ranking scores, created using ordinary kriging, is presented in Figure 9. The use of interpolated sample points is not a viable option for the existing Mapillary dataset. This is because the sample points available for interpolation are not randomly distributed within the defined area; rather, they are concentrated along the trails and roads within the park. Consequently, the results of kriging interpolation for a specific park are highly inaccurate and devoid of meaningful interpretation. In this regard, it is evident that a superior data source is required for interpolation mapping purposes.

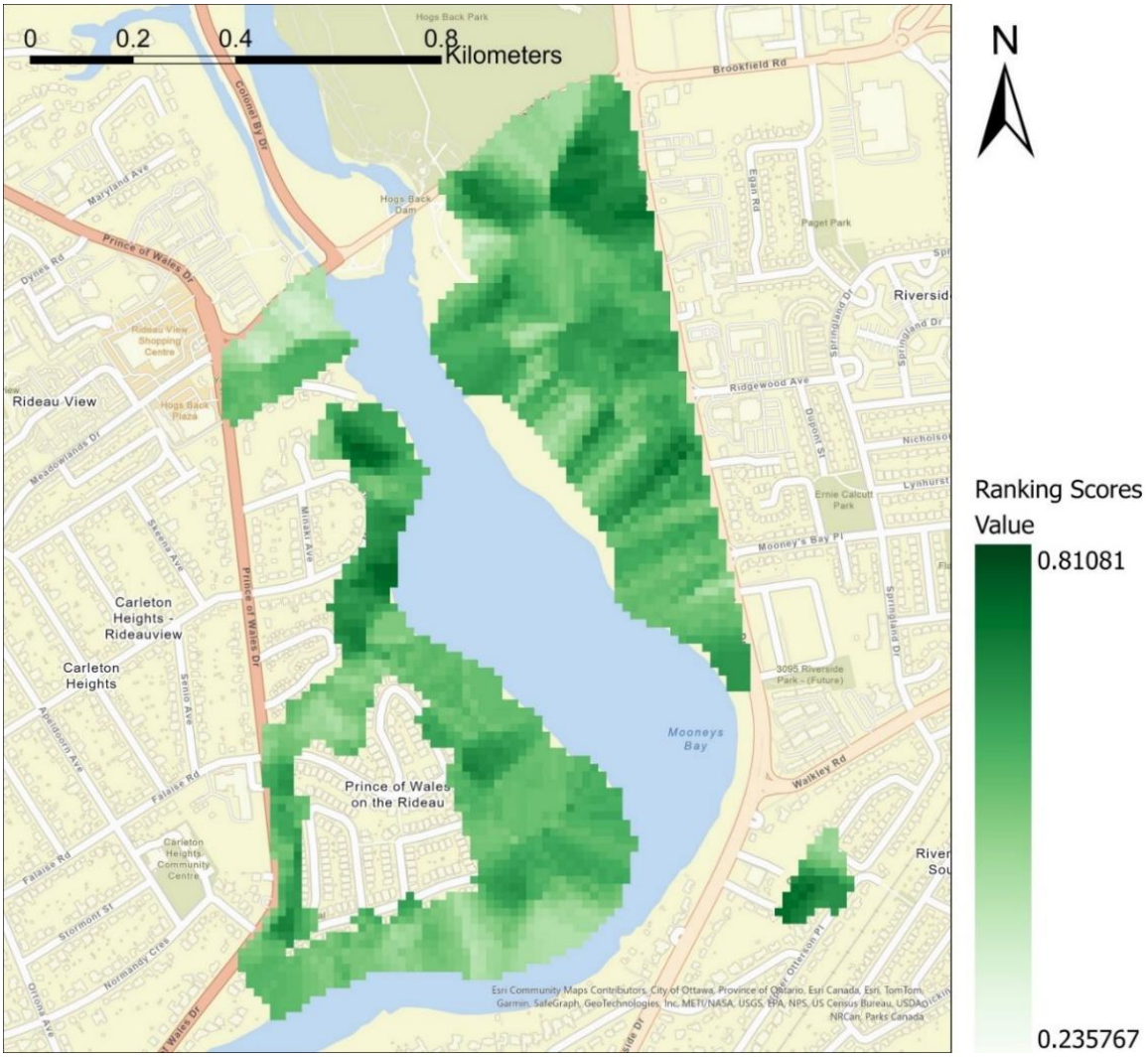

Figure 9. An example of Kriging interpolation using ranking scores.

5.4 Predicting “Connection to Nature” Using Google Street-view Images

To evaluate the model's capacity for generalization, a set of GSV images was obtained for the purpose of assessment. This enabled us to determine whether the model was able to make accurate predictions. In the absence of ground truth data, an attempt was made to visually compare the prediction results with satellite maps. An illustrative example is provided in Figure 10. As the dimension of "connect to nature" is more readily quantifiable, we observed that if the image ranking is low in urban blocks and higher in areas with high green coverage, it indicates that the model can be used to predict semantically similar datasets to some extent. Nevertheless, more rigorous and precise quantification, along with a more appropriate theoretical framework, are required for confirmation.

GSV imagery is about views on streets or motor-accessed roads, while the Mapillary images used in the training dataset are more focused on green spaces, such as parks, walking trails, and cycling paths. The use of GSV imagery was to demonstrate that the model along its weights trained using Mapillary images can be transferred to predict on GSV images. But we only exhibited the prediction results here with satellite background so that we can make visual analysis by comparing a greener area would have higher perception ranking scores where older adults could have stronger connection to nature. We will try to apply some more quantitative methods here to validate this on a firmer basis. Moreover, if we could collect the older adults’ perception on GSV we can make further comparative analysis on the differences of using the two images’ sources.

Figure 10. An example of making prediction on GSV images using the model trained with Mapillary images for "connection to nature". Higher values indicate higher ranking.

5.5 Correlations between Ranking of Perceptions and Semantic Elements for Street-level Images

To ascertain the correlation between ranking and image segments, we employed the U-Net algorithm to extract the percentage of pixels in each image. U-Net (Ronneberger et al., 2015) is a convolutional neural network (CNN) architecture specifically designed for image/object segmentation tasks. The model is widely popular in the medical imaging field due to its effectiveness in accurately segmenting images into different regions (such as detecting cancerous tissues). Figure 11 illustrates how the U-Net model works for the prediction of segmentation, whereby pixels are classified into a predefined value within a colour map that assigns a specific RGB value to a class of objects.

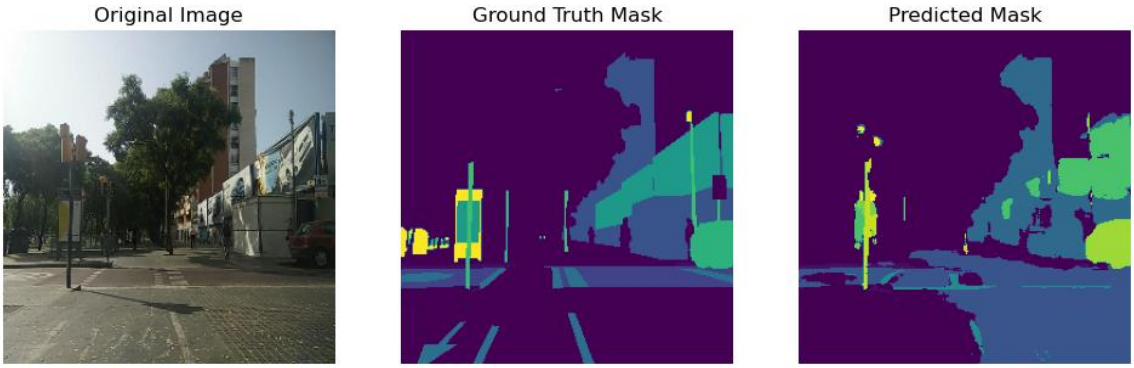

Figure 11. Example of the Usage of U-Net to Segment an Image.

We trained a U-Net model on 25,000 images in Mapillary Vistas Dataset v1.2 (MVD-v1.2) that has 66 categories of objects pre-defined in the ground truth masks. These images are captured at various conditions regarding weather, season and daytime by different experienced photographers using various imaging devices. Since the image size of the dataset used to train the ranking model is 224 by 224, we resized the images and corresponding annotation labels in the MVD-v1.2 into the same size. The model accuracy achieves ~88% in pixel-wise multi-classification while it has a very low Mean Intersection over Union (mIoU). We still used this model to predict the segmentation results on the Mapillary images of Ottawa green spaces captured in 2017 as used in the previous analysis since we want to know in general how the

perceptual ranking of those images is related to real-world semantic elements. After we calculated the portions of 66 different categories of objects, we tested correlation using Pearson's R and spatially weighted correlation using Lee's L (Appendix A, Appendix B).

The observed proportion of an image that contained phone booths, vehicles, pedestrians, bike lanes, guard rails, car mounts, motorcycles, potholes, riders and manholes were found to exhibit a markedly robust positive linear correlation with the ranking of connections to nature, aesthetics and activities. This is corroborated by the observation that each of these variables has a Pearson's R ranking at the top of the list. Moreover, the ranking scores for connections to nature, aesthetics and activities were found to be significantly and strongly correlated with one another, with a Pearson's R exceeding 0.8. No elements demonstrated a moderate negative correlation with the three dimensions. Regarding the perception of safety, the ranking exhibits a significantly strong negative correlation with the rankings of the other three dimensions, in the order of connection to nature, aesthetics and activity. Moreover, the elements identified as beneficial for the other three dimensions are inversely correlated with perceptions of safety.

The results obtained by Lee's L calculation were analogous. It was found that, for the perceptions of connection to nature, aesthetics and activities, the objects exhibiting a high positive correlation by Pearson's R also demonstrated a high positive correlation when considering spatial autocorrelation, but the unbiased coefficients of Lee's L were overall lower in value. Additionally, while most objects had a positive effect on the ranking of safety, this was not the case in reverse.

In a similar study, Min et al. (2020) investigated the relationship between visual elements and perception scores using Place Pulse 2.0 datasets. The researchers discovered that pedestrian crosswalks and road fences are associated with perceptions of safety, while greenery, such as trees and public gardens, are linked to perceptions of beauty. Our findings are generally consistent with theirs, but there are also some differences. We believe there are two reasons for

this. Firstly, the perception of older adults from a specific region may differ from a general recognition by large populations globally. Secondly, since the scene parsing model used by Min et al. is different from ours, this can result in categories of objects being segmented differently and the accuracy of the model being different. The significant drawback of U-Net is that it cannot detect the border of an object very accurately which has little effect on a simple scene but results in a very low mIoU, with increasing image complexity. The latest studies in urban perception have employed newer segmentation models, which may assist in improving the understanding of the semantic factor of perception mechanisms (J. Huang et al., 2023; Li et al., 2024; Min et al., 2020).

5.6 Shortcoming of Dataset

It should be noted that a limitation of the data set is that each image was utilized in a single pair of comparisons. Consequently, the implementation of algorithms such as TrueSkill is unfeasible, as these algorithms necessitate that an object be compared with others on approximately 30 occasions to generate a stable ranking score. Consequently, we opted to utilize the model itself to directly generate a ranking score, as opposed to employing ranking algorithms that rely on external comparisons. The model maps binary selection outcomes into a linear distributed line as floating values, enabling scores to be ordered ascending or descending. This approach was previously explored by Dubey et al. (2016), who demonstrated its effectiveness, although their training data is considerably larger than the one available to us. The latter consists of over a million perception selections from Place Pulse 2.0. However, our model's prediction accuracy, particularly in the context of safety as a common dimension in both datasets, exhibits a ~74% accuracy rate, which closely resembles the ~72% accuracy rate attained by the model. This suggests that the amount of data in our possession is sufficient. Nevertheless, we contend that, in the absence of extensive individual engagement in the training process and a paucity of shared urban living experiences in Ottawa, the model's capacity to discern the most salient features in

images that influence selection judgments is enhanced. This predicament gives rise to the utilization of a model that has been trained with our data, which can result in predictions that are not representative of the aged population in Ottawa. This underscores the necessity for the collection of more comprehensive data.

5.7 Influnence of Prior Experiences of Places on Perception Selection

A potential source of bias arises from participants' prior knowledge of the locations depicted in the image pair. For instance, if an image or both images are recognized by the participant, their selection may be influenced by prior experiences at that location. For instance, the recollection of pleasant memories at that location can significantly influence their perception of the image, even if it is not objectively appealing. To mitigate such potential influences, we can incorporate a protocol during data collection wherein participants are prompted to indicate whether they have previously visited the location depicted in an image. Alternatively, we can employ images captured in disparate locations; however, this approach does not guarantee the exclusion of all bias.

5.8 Influence of Downsampling Images

In the context of training a deep learning model, the process of downsampling is a mandatory step, as it significantly impacts the efficiency of the training procedure. The utilization of original, high-resolution images as training input necessitates the implementation of a more substantial model and deeper convolutional layers to attain optimal performance in terms of objectives. This assertion is supported by the available evidence. A further rationale for this approach stems from the necessity to employ a set of pretrained weights for the model's backbone, which remains constant during the training process and is solely updated through the process of inference. This is a common and critical technique called transfer learning, which reinforces the ability to represent spatial features of a backbone at both simple and abstract levels during training on a large image dataset in the early layers of the network. It would not be

possible to produce such weights that are generally good for any visual learning process from a small dataset like ours. Consequently, leveraging domain-specific weights and refining abstract representations to align with our specific dataset enables the backbone to adapt to our specific inquiries. Given that most readily available pretrained weights were trained on images of 224 by 224, we resized images in our dataset accordingly.

Chapter 6

Conclusion

This thesis presents the findings of empirical studies on the performance of deep learning methods to see the degree to which an AI can perceive the environment along four human dimensions of perception. The experiments underscored the significance of meticulous consideration when selecting the model architecture, loss function, and pre-trained weights, which are intimately connected to the nature of the task. Our findings build upon earlier research on human perception, which also explored how people perceive urban environments based on street-level photographs/images. In comparison to the Place Pulse dataset, which is widely used in this field, our study narrows the focus to the elderly and green spaces, examining the unique preferences of this population in greater detail. This study focuses exclusively on parks, whereas other studies have concentrated on streets. This marks the first instance of AI being used to map human perception of parks from the perspective of the elderly.

The results of our experiments demonstrate that the RankNet loss function consistently outperforms RankSVM, particularly in models such as ViT-B16, VGG16, and ResNet50. This indicates that the continuous optimization of RankNet is more effective in capturing the complexity of perceptual ranking. In contrast with the findings of previous research, the incorporation of a Softmax loss did not result in enhanced model performance. This highlights the distinctive challenges inherent to the perceptual ranking task. The impact of additional hidden layers on the results was inconclusive. While some architectures demonstrated enhanced performance, others exhibited a decline in learning efficiency. This emphasizes the importance of a thorough approach to adjusting model depth. The use of domain-specific pre-training weights derived from Places365 also demonstrated a clear advantage over generic object-based weights, such as those obtained from ImageNet. This further supports the idea that perceptual models are

highly context-sensitive and benefit from task-relevant pre-training. Noted that all conclusions were drawn by using our dataset, thus while the results were encouraging, several constraints were also identified. The relatively limited size of the dataset (comprising 4,000 image pairs per dimension) limits the model's capacity to generalize to the broader green space. While the results are promising, they are limited to the Ottawa context. To validate these findings, it would be beneficial to extend the research to other cities or countries. However, many of our results were similar to other studies of similar alignment. Furthermore, as is the case with any study of perception, there is an inherent subjectivity to the training data that may introduce bias. Despite the measures taken to mitigate this, future research should investigate the potential impact of such biases on model predictions, particularly given the involvement of a representative group of older adults. Furthermore, the model exhibited suboptimal performance in capturing specific perceptual dimensions, such as safety and activity. The weak correlations between some of these dimensions (e.g., activity and safety etc.) suggest that the subjectivity of these concepts may present a challenge for the interpretation of AI models. Further research should investigate alternative model architectures to more effectively capture the spatial relationships and contextual cues that influence perceptions of safety and activity.

Furthermore, we addressed the issue of spatial autocorrelation, which is a common challenge in studies that rely on photos taken in proximity. To this end, we employed Lee's L value to integrate spatial correlations, thereby providing more reliable comparisons. The results show that Lee's L value produces lower correlation values than the conventional Pearson's R value. This indicates that spatial correlations have a significant impact on the model's predictions. This impact highlights the importance of using spatially sensitive correlation measures when working with geographically structured data. For example, some perceptual dimensions (e.g. aesthetics and safety) demonstrated spatial clustering, indicating that older adults' perceptions are shaped not only by visual elements within individual images but also by their spatial context.

In conclusion, this study presents a framework for elucidating older adults' perceptions of green spaces and has the potential to inform urban planning strategies aimed at creating more age-friendly environments. Further work should extend these findings to optimize AI models for perception tasks and explore their application in a wider range of urban settings.

References

Akpinar, A. (2016). How is quality of urban green spaces associated with physical activity and health? *Urban Forestry & Urban Greening*, *16*, 76–83. https://doi.org/10.1016/j.ufug.2016.01.011

An D., Liu Y., & Huang Y. (2023). The Influence of Street Components on Age Diversity: A Case Study on a Living Street in Shanghai. *Sustainability (Switzerland)*, *15*(13). Scopus. https://doi.org/10.3390/su151310493

Baxter, D. E., & Pelletier, L. G. (2019). Is nature relatedness a basic human psychological need? A critical examination of the extant literature. *Canadian Psychology / Psychologie Canadienne*, *60*(1), 21–34. https://doi.org/10.1037/cap0000145

Baxter, D., Enns, A., & Kristjansson, E. (2017). *A Tale of Eight Cities*.

Blečić, I., Cecchini, A., & Trunfio, G. A. (2018). Towards Automatic Assessment of Perceived Walkability. In O. Gervasi, B. Murgante, S. Misra, E. Stankova, C. M. Torre, A. M. A. C. Rocha, D. Taniar, B. O. Apduhan, E. Tarantino, & Y. Ryu (Eds.), *Computational Science and Its Applications – ICCSA 2018* (14 citation(s); pp. 351–365). Springer International Publishing. https://doi.org/10.1007/978-3-319-95168-3_24

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, 89–96. https://doi.org/10.1145/1102351.1102363

Cheam M., & Bozovic-Stamenovic R. (2023). Neighborhood Land Use, Work Participation, and Quality of Life Among Workers in Mid and Late Life: Exploratory Analysis of Singapore's Public Housing Neighborhoods. *Journal of Aging and Environment*, *37*(3), 314–340. Scopus. https://doi.org/10.1080/26892618.2022.2092928

Chilès, J.-P., & Desassis, N. (2018). Fifty Years of Kriging. In B. S. Daya Sagar, Q. Cheng, & F. Agterberg (Eds.), *Handbook of Mathematical Geosciences: Fifty Years of IAMG* (pp. 589–612). Springer International Publishing. https://doi.org/10.1007/978-3-319-78999-6_29

Chollet F., asd, & asd. (2015). *Keras*. Keras. https://keras.io/

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., & Houlsby N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (No. arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). *Deep Learning the City: Quantifying Urban Perception At A Global Scale* (221 citation(s); No. arXiv:1608.01769). arXiv. http://arxiv.org/abs/1608.01769

Elharrouss, O., Akbari, Y., Almaadeed, N., & Al-Maadeed, S. (2022). *Backbones-Review: Feature Extraction Networks for Deep Learning and Deep Reinforcement Learning Approaches* (No. arXiv:2206.08016). arXiv. http://arxiv.org/abs/2206.08016

Fan Z., & Yu L. (2021). Street view imagery: Methods and applications based on artificial intelligence. *National Remote Sensing Bulletin*, *25*(5), 1043–1054. https://doi.org/10.11834/jrs.20219341

Fu, K., Chen, Z., & Lu, C.-T. (2018). StreetNet: Preference learning with convolutional neural network on urban crime perception. *In Proceedings of 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, November 6–9, 2018 (SIGSPATIAL '18)*, 269–278. https://doi.org/10.1145/3274895.3274975

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (No. arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Huang, J., Qing, L., Han, L., Liao, J., Guo, L., & Peng, Y. (2023). A collaborative perception method of human-urban environment based on machine learning and its application to the case area. *Engineering Applications of Artificial Intelligence*, *119*, 105746. https://doi.org/10.1016/j.engappai.2022.105746

Huang X., Gong P., & White M. (2022). Study on Spatial Distribution Equilibrium of Elderly Care Facilities in Downtown Shanghai. *International Journal of Environmental Research and Public Health*, *19*(13). Scopus. https://doi.org/10.3390/ijerph19137929

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142. https://doi.org/10.1145/775047.775067

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, S.-I. (2001). Developing a bivariate spatial association measure: An integration of Pearson's r and Moran's I. *Journal of Geographical Systems*, *3*(4), 369–385. https://doi.org/10.1007/s101090100064

Li, X., Beaucamp, B., Tourre, V., Leduc, T., & Servières, M. (2024). *Evaluation of Urban Perception Using Only Image Segmentation Features*. 200–207. https://www.scitepress.org/Link.aspx?doi=10.5220/0011969700003473

Madureira, H., Nunes, F., Oliveira, J. V., & Madureira, T. (2018). Preferences for Urban Green Space Characteristics: A Comparative Study in Three Portuguese Cities. *Environments*, *5*(2), 23. https://doi.org/10.3390/environments5020023

Min W., Mei S., Liu L., Wang Y., & Jiang S. (2020). Multi-Task Deep Relative Attribute Learning for Visual Urban Perception. *IEEE Transactions on Image Processing*, *29*, 657–669. https://doi.org/10.1109/TIP.2019.2932502

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, *37*(1/2), 17–23. https://doi.org/10.2307/2332142

Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore – Predicting the Perceived Safety of One Million Streetscapes. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 793–799. https://doi.org/10.1109/CVPRW.2014.121

Paez A., Mercado R. G., Farber S., Morency C., & Roorda M. (2010). Accessibility to health care facilities in Montreal Island: An application of relative accessibility indicators from the perspective of senior and non-senior residents. *International Journal of Health Geographics*, *9*. Scopus. https://doi.org/10.1186/1476-072X-9-52

Pearson, K. (1895). Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

Porzi L., Rota Bulò S., Lepri B., & Ricci E. (2015). Predicting and Understanding Urban Perception with Convolutional Neural Networks. *Proceedings of the 23rd ACM international conference on Multimedia*, 139–148. https://doi.org/10.1145/2733373.2806273

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation* (No. arXiv:1505.04597). arXiv. https://doi.org/10.48550/arXiv.1505.04597

Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The Collaborative Image of The City: Mapping the Inequality of Urban Perception. *PLoS ONE*, *8*(7), e68400. https://doi.org/10.1371/journal.pone.0068400

Simonyan, K., & Zisserman, A. (2015, April 10). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. https://doi.org/10.48550/arXiv.1409.1556

Statistics Canada. (2022). *Population estimates on July 1st, by age and sex*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501

Wang, R., Ren, S., Zhang, J., Yao, Y., Wang, Y., & Guan, Q. (2021). *A comparison of two deep-learning-based urban perception models: Which one is better?* (2 citation(s)). *1*, 3. https://doi.org/10.1007/s43762-021-00003-0

World Health Organization. (2007). *Global age-friendly cities: A guide*. https://iris.who.int/handle/10665/43755

Y, Z., Ae, V. den B., T, V. D., & G, W. (2017). Quality over Quantity: Contribution of Urban Green Space to Neighborhood Satisfaction. *International Journal of Environmental Research and Public Health*, *14*(5). https://doi.org/10.3390/ijerph14050535

Yang L., Chang H.-T., Li J., Xu X., Qiu Z., & Jiang X. (2023). A Comprehensive Evaluation of the Friendliness of Urban Facilities for the Elderly in Taipei City and New Taipei City. *Sustainability (Switzerland)*, *15*(18). Scopus. https://doi.org/10.3390/su151813821

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Zhou, H., He, S., Cai, Y., Wang, M., & Su, S. (2019). 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. https://doi.org/10.1016/j.scs.2019.101605

Appendix A

Pearson's correlation coefficient between each object category and each perceptual ranking.

Category	Connection to nature	p-value	Aesthetics	p-value	Safety	p-value	Activities	p-value
Barrier	0.0707	0.0000	0.0715	0.0000	-0.0036	0.1819	0.0522	0.0000
Bench	0.0287	0.0000	0.0287	0.0000	-0.0022	0.4043	0.0289	0.0000
Bicycle	0.1010	3.9602e-311	0.1094	0.0000	-0.0147	0.0000	0.0769	0.0000
Bicyclist	0.2051	0.0000	0.1614	0.0000	-0.1622	0.0000	0.1368	0.0000
Bike Lane	0.4836	0.0000	0.4640	0.0000	-0.3091	0.0000	0.4241	0.0000
Bike Rack	0.0768	0.0000	0.0801	0.0000	0.0076	0.0048	0.0637	0.0000
Billboard	0.0731	0.0000	0.0697	0.0000	0.0013	0.6158	0.0586	0.0000
Bird	0.0572	0.0000	0.0562	0.0000	0.0027	0.3187	0.0825	0.0000
Boat	0.0492	0.0000	0.0490	0.0000	-0.0048	0.0767	0.0343	0.0000
Bridge	-0.0040	0.1407	-0.0014	0.6009	0.0648	0.0000	0.0329	0.0000
Building	0.0988	0.0000	0.1067	0.0000	-0.0147	0.0000	0.0739	0.0000
Bus	0.2093	0.0000	0.1690	0.0000	-0.1637	0.0000	0.1436	0.0000
Car	0.0253	0.0000	0.0259	0.0000	0.0013	0.6375	0.0242	0.0000
Car Mount	0.4627	0.0000	0.4494	0.0000	-0.2827	0.0000	0.4071	0.0000
Caravan	0.1080	0.0000	0.1081	0.0000	-0.0071	0.0085	0.1240	0.0000
Catch Basin	0.0722	0.0000	0.0800	0.0000	0.0085	0.0016	0.0740	0.0000
CCTV Camera	0.0717	0.0000	0.0717	0.0000	-0.0027	0.3168	0.0517	0.0000
Crosswalk - Plain	0.1101	0.0000	0.1201	0.0000	-0.0102	0.0001	0.0934	0.0000
Curb	0.0487	0.0000	0.0474	0.0000	-0.0032	0.2285	0.0334	0.0000
Curb Cut	0.2094	0.0000	0.1607	0.0000	-0.1738	0.0000	0.1348	0.0000
Ego Vehicle	0.0751	0.0000	0.0755	0.0000	-0.0012	0.6641	0.0577	0.0000
Fence	0.0222	0.0000	0.0249	0.0000	0.0101	0.0002	0.0225	0.0000
Fire Hydrant	0.0780	0.0000	0.0752	0.0000	0.0051	0.0568	0.0620	0.0000
Ground Animal	0.0072	0.0071	0.0177	0.0000	0.0655	0.0000	0.0483	0.0000
Guard Rail	0.4765	0.0000	0.4604	0.0000	-0.3004	0.0000	0.4174	0.0000
Junction Box	0.1450	0.0000	0.1449	0.0000	-0.0996	0.0000	0.1265	0.0000
Lane Marking Crosswalk	0.2732	0.0000	0.2933	0.0000	-0.0992	0.0000	0.2669	0.0000

Category	Connection to nature	p-value	Aesthetics	p-value	Safety	p-value	Activities	p-value
Lane Marking - General	0.2107	0.0000	0.1668	0.0000	-0.1628	0.0000	0.1412	0.0000
Mailbox	0.0496	0.0000	0.0499	0.0000	-0.0043	0.1071	0.0344	0.0000
Manhole	0.3884	0.0000	0.3494	0.0000	-0.3017	0.0000	0.2898	0.0000
Motorcycle	0.4315	0.0000	0.3740	0.0000	-0.3102	0.0000	0.3845	0.0000
Motorcyclist	0.3098	0.0000	0.2562	0.0000	-0.2852	0.0000	0.2133	0.0000
Mountain	-0.0134	0.0000	-0.0908	0.0000	-0.1165	0.0000	-0.2367	0.0000
On Rails	0.0595	0.0000	0.0634	0.0000	0.0111	0.0000	0.0466	0.0000
Other Rider	0.3998	0.0000	0.3413	0.0000	-0.3463	0.0000	0.2473	0.0000
Other Vehicle	0.4919	0.0000	0.4401	0.0000	-0.3754	0.0000	0.4005	0.0000
Parking	0.1789	0.0000	0.1141	0.0000	-0.2584	0.0000	0.0117	0.0000
Pedestrian Area	0.4887	0.0000	0.4367	0.0000	-0.3727	0.0000	0.4001	0.0000
Person	0.1643	0.0000	0.1012	3.4542e-312	-0.2150	0.0000	-0.0022	0.4200
Phone Booth	0.5545	0.0000	0.5006	0.0000	-0.5010	0.0000	0.4275	0.0000
Pole	0.0568	0.0000	0.0433	0.0000	-0.0264	0.0000	0.0178	0.0000
Pothole	0.4121	0.0000	0.3694	0.0000	-0.2916	0.0000	0.3577	0.0000
Rail Track	0.0681	0.0000	0.0429	0.0000	-0.0908	0.0000	0.0432	0.0000
Road	0.0108	0.0001	-0.0512	0.0000	-0.1696	0.0000	-0.1814	0.0000
Sand	0.0468	0.0000	0.0499	0.0000	-0.0044	0.0980	0.0421	0.0000
Service Lane	0.0476	0.0000	0.0253	0.0000	-0.0397	0.0000	0.0114	0.0000
Sidewalk	0.2001	0.0000	0.2207	0.0000	-0.1346	0.0000	0.1713	0.0000
Sky	0.1224	0.0000	0.1183	0.0000	-0.0206	0.0000	0.1214	0.0000
Snow	0.0076	0.0049	0.0307	0.0000	0.0274	0.0000	0.0133	0.0000
Street Light	0.0709	0.0000	0.0415	0.0000	-0.0892	0.0000	0.0348	0.0000
Terrain	0.1949	0.0000	0.1964	0.0000	-0.2430	0.0000	0.1272	0.0000
Traffic Light	0.1356	0.0000	0.1190	0.0000	-0.0527	0.0000	0.1050	0.0000
Traffic Sign (Back)	0.1134	0.0000	0.1163	0.0000	-0.0548	0.0000	0.0815	0.0000
Traffic Sign (Front)	0.0488	0.0000	0.0531	0.0000	0.0214	0.0000	0.0538	0.0000
Traffic Sign Frame	0.0526	0.0000	0.0682	0.0000	0.0389	0.0000	0.0637	0.0000
Trailer	0.0336	0.0000	0.0121	0.0000	-0.0057	0.0335	0.0220	0.0000
Trash Can	0.0669	0.0000	0.0601	0.0000	-0.0236	0.0000	0.0426	0.0000
Truck	0.0721	0.0000	0.0396	0.0000	-0.0847	0.0000	0.0496	0.0000
Tunnel	0.2960	0.0000	0.3100	0.0000	-0.1161	0.0000	0.2696	0.0000

Category	Connection to nature	p-value	Aesthetics	p-value	Safety	p-value	Activities	p-value
Unlabeled	0.0295	0.0000	-0.0340	0.0000	-0.1865	0.0000	-0.1690	0.0000
Utility Pole	0.0749	0.0000	0.0819	0.0000	-0.0490	0.0000	0.0715	0.0000
Vegetation	0.0042	0.1153	0.0251	0.0000	0.1133	0.0000	0.0849	0.0000
Wall	-0.0330	0.0000	-0.0217	0.0000	0.1161	0.0000	0.0299	0.0000
Water	0.0244	0.0000	0.0279	0.0000	0.0593	0.0000	0.0468	0.0000
Wheeled Slow	0.0271	0.0000	-0.0038	0.1540	-0.0524	0.0000	-0.0253	0.0000

Appendix B

Lee's L correlation coefficient between each object category and each perceptual ranking.

Category	Connection to nature	p-value	Aesthetics	p-value	Safety	p-value	Activities	p-value
Banner	0.1817	0.8558	0.1758	0.8604	-0.0549	0.9562	0.1982	0.8428
Barrier	0.0587	0.9532	0.0560	0.9554	-0.0016	0.9987	0.0473	0.9622
Bench	0.0241	0.9808	0.0213	0.9830	-0.0096	0.9924	0.0244	0.9805
Bicycle	0.0833	0.9336	0.0869	0.9307	-0.0092	0.9927	0.0680	0.9458
Bicyclist	0.1744	0.8615	0.1443	0.8852	-0.1372	0.8908	0.1219	0.9029
Bike Lane	0.4334	0.6646	0.4155	0.6777	-0.2733	0.7846	0.3842	0.7007
Bike Rack	0.0647	0.9484	0.0643	0.9487	-0.0046	0.9963	0.0574	0.9542
Billboard	0.0593	0.9527	0.0547	0.9564	-0.0008	0.9994	0.0508	0.9595
Bird	0.0621	0.9505	0.0557	0.9556	-0.0304	0.9758	0.0538	0.9571
Boat	0.0405	0.9677	0.0374	0.9702	0.0006	0.9995	0.0318	0.9746
Bridge	0.0107	0.9915	0.0110	0.9912	0.0211	0.9831	0.0188	0.9850
Building	0.0811	0.9353	0.0852	0.9321	-0.0090	0.9928	0.0665	0.9470
Bus	0.1771	0.8594	0.1488	0.8817	-0.1378	0.8903	0.1263	0.8995
Car	0.0204	0.9837	0.0188	0.9850	-0.0080	0.9936	0.0206	0.9836
Car Mount	0.4127	0.6797	0.3998	0.6892	-0.2500	0.8025	0.3671	0.7135
Caravan	0.0878	0.9300	0.0901	0.9282	-0.0024	0.9981	0.1036	0.9175
Catch Basin	0.0590	0.9530	0.0666	0.9469	0.0141	0.9887	0.0679	0.9458
CCTV Camera	0.0577	0.9540	0.0544	0.9566	-0.0008	0.9993	0.0468	0.9626
Crosswalk	0.0964	0.9232	0.1037	0.9173	-0.0190	0.9849	0.0858	0.9316
Curb	0.0407	0.9675	0.0373	0.9702	0.0010	0.9992	0.0310	0.9753
Curb Cut	0.1777	0.8589	0.1440	0.8854	-0.1458	0.8840	0.1198	0.9046
Ego Vehicle	0.0620	0.9505	0.0592	0.9528	-0.0021	0.9983	0.0510	0.9593
Fence	0.0164	0.9869	0.0158	0.9874	-0.0028	0.9977	0.0168	0.9866
Fire Hydra	0.0624	0.9502	0.0577	0.9540	0.0007	0.9995	0.0527	0.9580

Category	Connection to nature	p-value	Aesthetics	p-value	Safety	p-value	Activities	p-value
Ground Animal	0.0220	0.9825	0.0276	0.9780	0.0158	0.9874	0.0211	0.9832
Guardrail	0.4266	0.6696	0.4117	0.6804	-0.2649	0.7910	0.3786	0.7049
Junction Box	0.1242	0.9012	0.1184	0.9057	-0.0852	0.9321	0.1073	0.9146
Lane Marking Crosswalk	0.1789	0.8580	0.1486	0.8818	-0.1381	0.8902	0.1257	0.9000
Lane Marking - General	0.2294	0.8185	0.2468	0.8050	-0.0816	0.9349	0.2229	0.8236
Mailbox	0.0409	0.9674	0.0384	0.9693	-0.0002	0.9999	0.0324	0.9741
Manhole	0.3533	0.7238	0.3221	0.7473	-0.2567	0.7974	0.2814	0.7783
Motorcycle	0.3849	0.7003	0.3387	0.7348	-0.2714	0.7860	0.3481	0.7277
Motorcyclist	0.2962	0.7670	0.2483	0.8039	-0.2553	0.7984	0.2180	0.8274
Mountain	-0.0215	0.9828	-0.0786	0.9374	-0.0711	0.9433	-0.1940	0.8461
On Rails	0.0534	0.9574	0.0528	0.9579	0.0038	0.9970	0.0417	0.9667
Other Ride	0.3537	0.7235	0.3072	0.7586	-0.2893	0.7723	0.2317	0.8167
Other Vehicle	0.4530	0.6504	0.4057	0.6848	-0.3363	0.7366	0.3810	0.7031
Parking	0.1555	0.8764	0.1025	0.9183	-0.2076	0.8355	0.0217	0.9827
Pedestrian	0.4517	0.6514	0.4038	0.6863	-0.3356	0.7371	0.3804	0.7036
Person	0.1393	0.8891	0.0906	0.9278	-0.1655	0.8685	0.0101	0.9920
Phonebooth	0.5140	0.6071	0.4553	0.6488	-0.4525	0.6508	0.4030	0.6869
Pole	0.0494	0.9606	0.0377	0.9699	-0.0109	0.9913	0.0236	0.9812
Pothole	0.3756	0.7071	0.3388	0.7347	-0.2594	0.7952	0.3339	0.7384
Rail Track	0.0612	0.9512	0.0398	0.9683	-0.0797	0.9364	0.0379	0.9698
Road	0.0092	0.9927	-0.0320	0.9745	-0.1184	0.9057	-0.1376	0.8905
Sand	0.0444	0.9646	0.0409	0.9674	-0.0140	0.9889	0.0382	0.9695
Service Lane	0.0454	0.9638	0.0250	0.9801	-0.0382	0.9695	0.0139	0.9889
Sidewalk	0.1686	0.8660	0.1719	0.8635	-0.1020	0.9187	0.1454	0.8844
Sky	0.1015	0.9191	0.0987	0.9213	-0.0137	0.9891	0.1035	0.9175

Category	Connection to nature	p-value	Aesthetics	p-value	Safety	p-value	Activities	p-value
Snow	0.0108	0.9914	0.0299	0.9761	0.0175	0.9861	0.0113	0.9910
Streetlight	0.0490	0.9609	0.0281	0.9775	-0.0494	0.9606	0.0270	0.9785
Terrain	0.1725	0.8630	0.1684	0.8662	-0.2037	0.8385	0.1148	0.9086
Traffic Sign (Back)	0.0518	0.9587	0.0489	0.9610	-0.0113	0.9910	0.0479	0.9618
Traffic Sign (Front)	0.0513	0.9591	0.0643	0.9487	0.0142	0.9887	0.0498	0.9603
Traffic Light	0.1060	0.9156	0.0988	0.9212	-0.0415	0.9669	0.0846	0.9325
Traffic Sign	0.0946	0.9246	0.0922	0.9265	-0.0328	0.9738	0.0733	0.9416
Trailer	0.0453	0.9639	0.0178	0.9858	-0.0632	0.9496	-0.0193	0.9846
Trash Can	0.0601	0.9521	0.0501	0.9600	-0.0220	0.9824	0.0411	0.9672
Truck	0.0668	0.9467	0.0375	0.9701	-0.0827	0.9340	0.0434	0.9653
Tunnel	0.2467	0.8051	0.2591	0.7955	-0.0871	0.9306	0.2278	0.8197
Unlabeled	0.0279	0.9778	-0.0148	0.9882	-0.1336	0.8937	-0.1236	0.9016
Utility Pole	0.0693	0.9447	0.0736	0.9413	-0.0456	0.9636	0.0598	0.9523
Vegetation	0.0266	0.9787	0.0318	0.9746	0.0285	0.9772	0.0447	0.9643
Wall	-0.0122	0.9903	-0.0083	0.9933	0.0524	0.9582	0.0052	0.9958
Water	0.0262	0.9791	0.0289	0.9769	0.0170	0.9864	0.0369	0.9706
Wheeled Slow	0.0233	0.9814	-0.0030	0.9976	-0.0342	0.9727	-0.0169	0.9865