The Journal of Transport and Land Use jtlu.org
Vol. 19 No. 1 [2026] pp. 413-440
https://doi.org/10.5198/jtlu.2026.2873

JTLU

Urban visual AI for travel behavior prediction: A comparative analysis of satellite and street view imagery

Jinghao Cao, New Jersey Institute of Technology, jc2687@njit.edu
Yuebing Liang (corresponding author), Tsinghua University, liangyb@mail.tsinghua.edu.cn
Meiqing Li (corresponding author), University of Central Florida, meiqing@ucf.edu
Shenhao Wang, University of Florida, shenhaowang@ufl.edu

Abstract: The increasing availability of satellite and street-view images offers new opportunities for visual AI in commuting mode share analysis. However, few studies have systematically examined how different visual representations of the built environment encode census tract-level commuting mode shares. This study evaluates the predictive and interpretative performance of satellite and street-view imagery for modeling census tract-level commuting mode shares in Florida, USA. We apply semantic segmentation to extract land-use and streetscape features from both imagery types and integrate them with socio-demographic data in multiple machine learning models. We further apply SHAP (SHapley Additive exPlanations) to interpret how visual features contribute to model predictions across commuting modes. Results show that satellite imagery generally outperforms street-view imagery in terms of out-of-sample R^2, although the magnitude of improvement varies across models, travel modes, and satellite zoom levels. For the tree-based models, the relative improvement of satellite imagery over street-view imagery typically ranges from approximately 2% to 8%. Lower zoom levels with a broader spatial context better predict driving and transit use, whereas higher zoom levels with finer details improve walking and biking prediction. Despite the performance differences, SHAP analysis reveals largely consistent predictive associations: lower-density and more vegetated environments are associated with higher car dependence, while denser and pedestrian-oriented environments are associated with higher transit and active mode shares. This study demonstrates the value of interpretable GeoAI in linking visual urban form to commuting mode shares. The results highlight the importance for planners of using both satellite and street-view imagery, as the two data sources offer complementary capacities for interpreting neighborhood-level commuting mode-share patterns. The proposed approach can help practitioners compare neighborhoods and flag areas where commuting patterns or visible built-environment conditions warrant closer transportation or land-use analysis.

Keywords: Travel behavior, AI, Computer vision, Satellite images, Street view images

1 Introduction

Predicting travel mode share is essential to understanding how people interact with urban environments and to inform urban planning and travel decisions. However, mode-share data are often sparse, temporally dated, and spatially incomplete, which can limit their usefulness for timely and large-scale assessment. This creates a need for methods that can estimate commuting mode-share patterns across large areas using more widely available data sources. For practitioners, these estimates can help compare neighborhoods and identify places that may need further transportation or land-use analysis. In recent years, satellite imagery and street-view imagery have emerged as valuable resources for

Copyright 2026 Jinghao Cao, Yuebing Liang, Meiqing Li & Shenhao Wang
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0
Received March 2, 2026; Revised August 28, 2026; Accepted October 7, 2026; Available online November 9, 2026

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

uncovering commuting mode share. Prior research has demonstrated that such imagery, being globally available and publicly accessible, can be used to predict a wide range of urban indicators, including socioeconomic characteristics (Jean et al., 2016), land use (Albert et al., 2017), and travel patterns (Goel et al., 2018). Beyond enhancing predictive accuracy, satellite imagery and street-view imagery enable the fine-grained extraction of built environment features, thereby deepening our understanding of how urban form influences travel mode share (Hipp et al., 2022; Wu et al., 2023).

Recent studies have increasingly used satellite and street-view imagery to predict travel mode shares (Ganji et al., 2022; Gebru et al., 2017), without comparing the performance between these two types of images, presenting a significant research gap. Satellite images provide a broad spatial coverage and capture large-scale land use and infrastructure patterns, while street view images offer detailed, ground-level perspectives of the urban environments. Without comparing both types of imagery, studies may have missed opportunities to improve predictive accuracy and develop a better understanding of how different environmental factors are associated with aggregate travel patterns. Xie et al. (2016) and Yuan et al. (2020) show the potential of both types of imagery, yet their relative performance remains largely unexplored. The lack of comparative analysis prevents researchers and practitioners from choosing appropriate image sources to make more accurate and interpretable prediction of travel patterns in a variety of urban contexts.

Given that satellite and street view images are both important imagery inputs, it is essential to compare their performance and identify the optimal spatial resolution of urban imagery for commuting mode-share prediction. Figure 1 provides sample satellite and street-view images used in this study, illustrating the different visual perspectives captured by the two imagery sources. However, there has been minimal investigation into how different spatial resolutions, particularly satellite imagery at varying spatial resolution, affect predictive performance. Ayush et al. (2020) and Jean et al. (2016) applied satellite imagery at specific resolutions but did not systematically compare the effect of spatial resolution, leaving questions regarding the optimal spatial resolution to capture relevant urban attributes that influence travel behavior.

By directly comparing satellite and street-view imagery and examining the effects of satellite image scale, this study evaluates how different imagery sources and Google Maps satellite zoom levels affect census-tract-level commuting mode-share prediction performance, and how segmentation-derived built-environment features contribute to model interpretation. The analysis focuses on aggregate commuting mode-share patterns at the census-tract level rather than disaggregate individual- or trip-level travel decision-making. The key contributions are as follows:

- We develop an image-based modeling framework for census tract-level commuting mode-share prediction. Built-environment features extracted from satellite and street-view imagery using pre-trained semantic segmentation models are combined with sociodemographic variables and evaluated using multiple machine learning algorithms, including multiple linear regression (MLR), random forest (RF), and gradient boosting decision trees (GBDT).
- We examine how prediction performance differs across imagery sources and satellite spatial scales. Specifically, we compare street-view imagery with satellite imagery and evaluate satellite imagery across zoom levels 14–17 to assess how broader spatial context and finer local detail affect commuting mode-share prediction. Here, zoom level refers to the Google Maps image retrieval setting that controls the spatial extent and visible detail of the downloaded satellite image.

- We compare the interpretability of satellite and street-view imagery by integrating semantic segmentation with SHAP (SHapley Additive exPlanations) analysis. SHAP is an interpretable machine-learning method that quantifies how each input feature contributes to a model prediction. This comparison examines how the different visible built-environment features derived from each source contribute to tract-level commuting mode-share prediction and interpretation.

The comparative results help clarify how different visual representations encode distinct dimensions of the built environment. Satellite imagery captures macro-scale land-cover and spatial-structure features, while street-view imagery captures micro-scale streetscape conditions. By comparing these two imagery sources for the same census tracts, this study helps distinguish effects attributable to visual representation from those attributable to regional context.

The remainder of this paper is organized as follows. Section 2 reviews related studies on imagery-based urban analysis and travel mode-share prediction. Section 3 describes the data used in this study. Section 4 presents the methodological framework. Section 5 reports the empirical results, including the comparison of predictive performance across imagery sources and satellite zoom levels, two illustrative case studies, and the SHAP analysis of imagery-derived features. Section 6 concludes the paper by summarizing the main findings, discussing planning and methodological implications, and outlining limitations and future research directions.

2 Literature review

Many studies have explored the relationship between urban environment and travel behavior, which serves as essential basis for policy decisions in urban and transportation planning. Prior research on travel behavior has predominantly relied on structured, tabular data—such as socio-demographic attributes, land-use characteristics, and transportation network variables—because these inputs integrate naturally into traditional statistical and machine learning models. Crane (2000) offered an in-depth review of foundational studies examining the relationship between urban form and travel behavior. Foundational studies and meta-analyses have consistently demonstrated the impact of the built environment on travel outcomes, including trip frequency, mode choice, and travel distance. More specifically, built-up density and land-use mix are commonly associated with lower automobile dependence and greater use of public transit and active modes (Cervero & Kockelman, 1997; Ewing & Cervero, 2010). Pedestrian-oriented infrastructure, including sidewalks and connected street networks, is closely related to walking and walkability (Ewing & Handy, 2009; Saelens et al., 2003). A number of empirical studies have investigated travel behavior in greater detail, focusing on specific population groups, travel modes, and spatial settings. Zhang et al. (2008) investigated female travel behavior in terms of travel rate, consumption, trip purpose, and mode using data from the Beijing 2005 travel survey. Larsen and El-Geneidy (2011) examined cycling facility use in Montréal, Canada, and found that both route-level built-environment factors and personal cycling habits influenced facility use. Cycling facility availability, CBD destination location, accident exposure, and intersection density were associated with route choice, while respondents' bicycle-use frequency and comfort under different cycling conditions also shaped cycling facility use. Hong et al. (2014) employed Bayesian hierarchical models to examine the influence of built-environment factors measured across various geographic scales on travel behavior. A major advantage of such traditional dataset is their tabular structure and thus can be easily used as model input for travel behavior modeling. Although such datasets remain highly useful for transportation analysis, assembling built-environment indicators in tabular form can be

labor-intensive, and the available variables may not always capture visually observable urban form in a consistent and scalable way across locations.



a. Satellite Images b. Street View Images

Figure 1. Exemplary satellite and street view images

To overcome this limitation, recent studies have increasingly used satellite imagery and street-view imagery data as complementary and potentially less labor-intensive sources for characterizing urban settings. These images have proven effective in predicting a range of urban variables. For instance, Jean et al. (2016) demonstrated that a convolutional neural network (CNN) could extract features from satellite images to explain up to 75% of the variation in local economic outcomes. Similarly, Albert et al. (2017) utilized large-scale satellite imagery combined with deep CNN methods to analyze land-use patterns in urban neighborhoods, while Ayush et al. (2020) successfully predicted local poverty levels by applying object detection techniques to high-resolution satellite images. Beyond improving predictive accuracy, satellite imagery enables the detailed extraction of urban features, thereby enhancing our understanding of how the built environment impacts various target variables, including travel behavior. Prior remote-sensing studies have shown that satellite imagery can be used to extract interpretable land-cover and built-environment features (Bagwari et al., 2023; Ma et al., 2019; Neupane et al., 2021). These studies underscore the dual benefits of satellite imagery: not only does it enhance the predictive performance of models for critical urban variables, but it also allows for the extraction of interpretable built-environment features, such as building density, road coverage, vegetation and water, which can contribute to interpretable visual AI and GeoAI research. Here, GeoAI refers to the integration of geospatial data, artificial intelligence, and spatial analytical methods for modeling and interpreting spatial patterns in urban systems. Visual AI refers to computer vision and machine learning methods that convert visual data into structured and interpretable built-environment features. (Berriel et al., 2017; Ning et al., 2022)

Selecting the appropriate zoom level for satellite imagery is critical, as it directly affects both the spatial extent and the level of detail captured in the images. Existing studies have employed a range of zoom scales—from high-resolution, block-level analyses (Novack & Stilla, 2018; Tapiador et al., 2011), to neighborhood-level (Jing et al., 2020) and even city-level examinations (Kuang & Dou, 2020). High-resolution imagery provides detailed information about specific areas, albeit over smaller spatial extents, whereas lower-resolution imagery covers larger areas but with less granular

detail. This inherent trade-off continues to pose a challenge in determining the optimal spatial resolution for satellite image analysis.

In addition to satellite imagery, street view images have become increasingly valuable in urban research, offering human-perspective views of city environments. Recent studies have utilized street view images for various urban prediction tasks. For example, Naik et al. (2014) used Google Street View images to predict perceived safety at the neighborhood level, while Goel et al. (2018) demonstrated that street imagery could be used to estimate city-level travel patterns. Other works, such as those by Cheng et al. (2017) and Wang et al. (2021) extracted built environment features from street images to study associations with urban form. Fan et al. (2023) found that urban features extracted from street view images can predict a variety of city profiles, including people's travel behavior, poverty status, crime, and health behaviors. Beyond enhancing predictive performance, street-view images facilitate interpretative analyses by allowing researchers to extract semantic information about urban features. Semantic segmentation models have been widely used to convert street-view images into interpretable streetscape features (Chen et al., 2018; He et al., 2017). These extracted semantic features not only improve the accuracy of urban outcome predictions but also deepen our understanding of the relationship between urban characteristics and variables such as public health (Nguyen et al., 2019), crime (Hipp et al., 2022), and travel behavior (Wu et al., 2023).

Some studies discussed the applications and characteristics of the two types of images in urban research (W. Li et al., 2023; Nice et al., 2020), however, to the best of our knowledge, there is no former paper which directly compared the predictive performance of these two types of imagery. A key research gap lies in the lack of direct comparison between the predictive performance of satellite imagery and street view imagery for travel behavior prediction. Although both imagery types are extensively utilized in urban and transportation research, prior studies have largely focused on the individual contributions of each imagery type. Satellite imagery has been widely applied in tasks such as urban feature segmentation (Yi et al., 2019), traffic volume prediction (Ganji et al., 2022), and integration with socio-demographic data for travel behavior modeling (Wang et al., 2024). Similarly, street view imagery has been employed to measure social and environmental inequalities (H. Zhou et al., 2019), predict neighborhood demographics (Gebru et al., 2017; Nguyen et al., 2021), and enhance traffic safety (Cai et al., 2022; Hamim & Ukkusuri, 2024). However, no prior studies have systematically evaluated their comparative advantages or unique contributions to predictive modeling of travel behavior. This gap concerns not only prediction accuracy, but also whether features derived from satellite and street-view imagery contribute differently to census tract-level commuting mode-share prediction and interpretation.

3 Data

This study focuses on Florida because it serves as a useful test bed for comparing the predictive and interpretive value of satellite and street-view imagery. Florida exhibits substantial heterogeneity in built-environment conditions, meaningful variation across metropolitan areas, and diverse commuting patterns across urban contexts. Together, these characteristics provide an appropriate setting for evaluating how different imagery sources capture built-environment information relevant to travel mode share. This section describes the datasets used in the analysis, including commuting mode-share data, satellite imagery and street-view imagery. We first introduce the 2020 Census-based commuting data, including the census tract selection procedure and the distribution of mode shares. We then describe the satellite and street-view imagery, together with their data collection and preprocessing procedures.

3.1 Travel mode data

The travel mode data are derived from the 2020 Census of Florida, which reports the percentage of individuals commuting by automobile, public transit, and active modes within each census tract. A census tract is a small statistical geographic unit defined by the U.S. Census Bureau, typically containing between 1,200 and 8,000 residents, with an optimum size of about 4,000 residents. Because census tracts are primarily population-based rather than area-based units, their geographic size is not fixed: tracts in dense urban areas cover smaller areas, while tracts in lower-density areas cover larger areas. To construct the analytical sample, we first identified selected cities in Florida, including Miami, Tampa, Orlando, Jacksonville, Fort Lauderdale, Tallahassee, Gainesville, Pensacola, Sarasota, St. Petersburg, West Palm Beach, Boca Raton, Clearwater, Fort Myers, and Naples, and recorded their central latitude and longitude coordinates. Around each city center, we generated random spatial offsets within a 10-mile radius and matched the resulting points to census tract polygons. We then excluded tracts with missing commute mode-share data or unavailable imagery from either source. For each remaining tract, the tract centroid was used as the reference location for downloading satellite imagery at zoom levels 14–17 through the Google Maps API and a corresponding street-view image through the Google Street View Static API. After these filtering steps, the final analytical sample contained 1,700 census tracts. The 1,700 census tracts included in the final analytical sample contain a total population of 6,777,261 residents.

Figure 2 presents detailed histograms showing the distribution of commuting mode shares for driving, public transit, and active modes. As shown in Figure 2.a, Florida exhibits a high rate of car-based commuting. The distribution of driving ratios follows a bell-shaped curve, with most values concentrated between 0.70 and 0.96. In contrast, the shares of public transit and active commuting (Figure 2.b and Figure 2.c) are significantly lower, with many census tracts reporting values below 3%, including a substantial number with zero usage. Only a small fraction of tracts report more than 10% of commuters using public transit or active modes. These patterns indicate a strong reliance on automobiles for daily commuting, with limited uptake of alternative modes. Given the potential benefits of public transit and active travel for sustainable urban development, it is important to understand how the built environment influences travel mode choices.

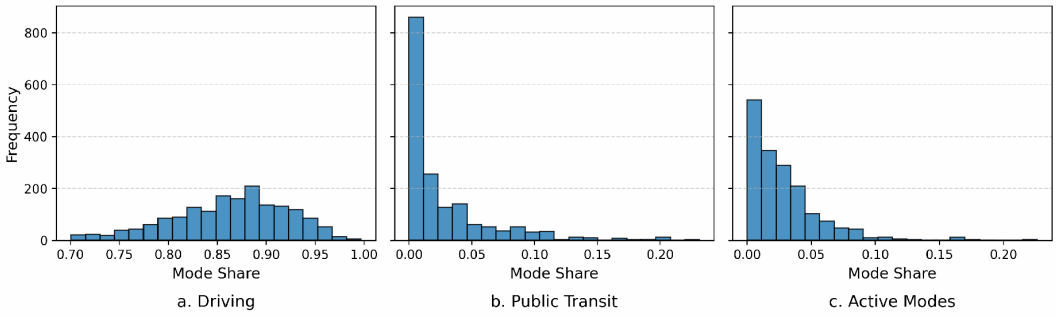


Figure 2. The distribution of travel mode shares

3.2 Urban images

The satellite and street view images used in this study are obtained via Google Maps API (https://developers.google.com/maps) and Google Street View Static API (https://developers.google.com/maps/documentation/streetview). For each census tract, we extracted its geographic centroid as the reference point for image retrieval. Satellite

imagery was downloaded at multiple Google Maps zoom levels (14–17) through the Google Maps API. Street-view imagery was retrieved through Google Street View Static API using the same census tract centroid as the query location. For the street-view query, we used a search radius of 2,700 m around the centroid to identify an available street-view scene. Each census tract was represented by one street-view image, and one satellite image at each zoom level. Street-view imagery was retrieved as a single image using a fixed street-view setting rather than multiple zoom levels. The original images from both sources were RGB images with dimensions of 1024 × 1024 pixels. The full images were resized to 224 × 224 pixels rather than split into smaller tiles. In total, the dataset included 1,700 street-view images and 6,800 satellite images. We used the census tract centroid as a standardized sampling location to ensure that satellite and street-view imagery were retrieved using a consistent spatial reference point across all census tracts.

To enhance the robustness of the training process, we apply standard data augmentation techniques, including random horizontal flipping and rotation. Examples of satellite images at different Google Maps zoom levels, together with the street-view image, are shown in Figure 3.



Figure 3. Examples of satellite images at different Google Maps zoom levels and the corresponding street-view image

4 Method

This study presents a comprehensive framework to systematically compare street view and satellite imagery across different zoom levels for travel behavior prediction. The framework evaluates these imagery types from two key aspects: predictive performance and interpretability. This section begins by presenting the overall analytical framework (Section 4.1), followed by detailed discussions on the specific methodologies for image segmentation (Section 4.2), predictive performance evaluation (Section 4.3), and interpretability analysis (Section 4.4).

4.1 Framework

Figure 4 shows the analytic framework of this study, consisting of four stages:

- Data Collection: We collect satellite imagery and Streetview data from Google Map API, as well as travel behavior data from United States Census Bureau official website (Census.gov) and spatially align them with GeoPandas package in Python. More details regarding the data processing will be introduced in Section 3.
- Feature Extraction: Pre-trained semantic segmentation tools are used to extract the proportions of various urban features from images, converting raw images into measurable feature proportions that can be used as model inputs. Here, semantic segmentation models assign a predefined category label to each pixel in an image, allowing unstructured imagery to be converted into structured built-environment features. Specifically, we use DeepLabV3+ for street-view imagery and the MIT CSAIL semantic segmentation model (ResNet-50 + PPMdeepsup) for satellite imagery. This facilitates comparisons of predictive performance and interpretability

across different image types. Details of the segmentation tools are outlined in Section 4.2.

- Travel Mode Share Modeling: We employ three machine learning models to predict travel mode shares, Multiple Linear Regression (MLR), Random Forest (RF), and Gradient Boosting Decision Trees (GBDT), using segmented urban features and socio-demographic variables as input. To assess the predictive performance of different urban image types and spatial resolutions, we compare results using the same model while varying the input imagery. Additional details about the supervised learning models and the experiment design are presented in Section 4.3.
- Model Interpretation: Beyond predicting travel behavior, understanding how the built environment shapes travel decisions is essential for informing policy implications. In this stage, we apply SHAP (SHapley Additive exPlanations), to assess the impact of urban features extracted from street view and satellite images on travel mode shares. Details of the SHAP methodology are provided in Section 4.4.

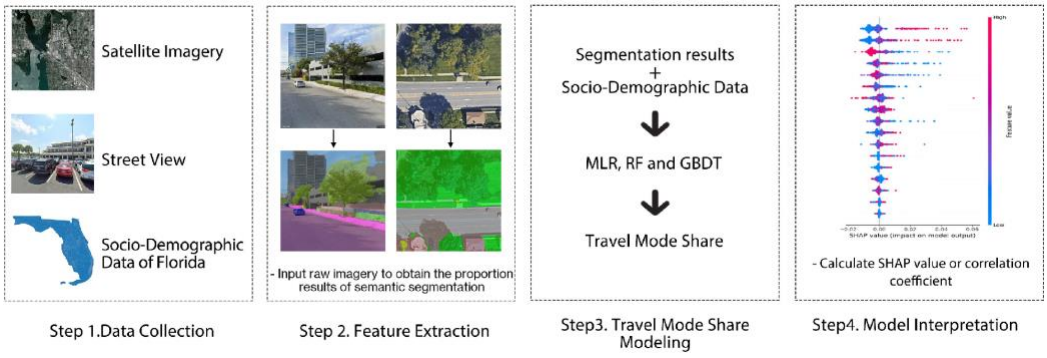


Figure 4. Framework diagram

4.2 Image segmentation

Satellite imagery and street-view imagery provide valuable information about the built environment, but they are unstructured, which means that the relevant information is embedded in pixel-level visual patterns rather than predefined tabular variables. This makes it challenging to directly use raw images in the predictive models. To address this, we use pre-trained image segmentation tools to identify and quantify features (e.g., roads, buildings, vegetation) in street view and satellite images. The extracted features provide structured inputs to compare the predictive performance and interpretability of the two image types, as will be discussed in subsequent sections. The selected segmentation classes were chosen to reflect interpretable built-environment features visible in each image source and relevant to travel behavior analysis. Because street-view and satellite imagery capture different visual perspectives, the class sets are not identical; accordingly, the comparison in this study is based on the predictive and interpretive value of the resulting feature vectors rather than on one-to-one correspondence between classes. Additional architecture details are provided in Appendix B.

For street-view images, we use DeepLabV3+ (Chen et al., 2018) for semantic segmentation. The implementation is based on the open-source repository (https://github.com/VainF/DeepLabV3Plus-Pytorch). The model takes a street-view image as input and outputs a segmentation map in which each pixel is assigned to a predefined category. Figure 5.a shows an example of a street-view image and its corresponding segmentation result. From the segmented map, we compute the relative

proportions of visible urban features and use them as image-derived predictors. The street-view feature vectors include the following categories: road, building, sky, sidewalk, plant, car, and person.

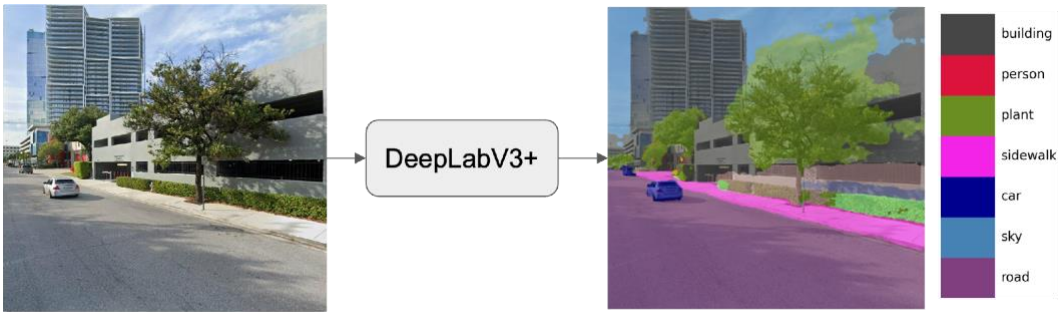
For satellite images, we employ the semantic segmentation implementation developed by MIT CSAIL (https://github.com/CSAILVision/semantic-segmentation-pytorch/tree/master). Specifically, we use a model with a ResNet-50 encoder and a PPMdeepsup decoder, trained on the ADE20K dataset (Zhou et al., 2017; B. Zhou et al. 2019). Like the DeepLabV3+ procedure for street-view images, this model takes a satellite image as input and outputs a segmentation map in which each pixel is assigned to one of the predefined land-cover categories. Figure 5.b presents an example of a satellite image together with its corresponding segmentation result. From the resulting segmentation map, we compute the relative proportions of different land-cover features and convert each satellite image into a structured feature vector. In this study, the satellite-image feature vector includes the following categories: building, road, plant, water, earth, and bridge.

The two segmentation models were selected because they provide established and reproducible tools for extracting interpretable visual features from the two image types. DeepLabV3+ was used for street-view imagery because it is a widely adopted semantic segmentation model for natural and street-level scene parsing and can identify streetscape features relevant to commuting mode-share prediction. For satellite imagery, we used the MIT CSAIL semantic segmentation model with a ResNet-50 encoder and PPMdeepsup decoder because it provides a reproducible pre-trained semantic segmentation model for extracting land-cover-related features.

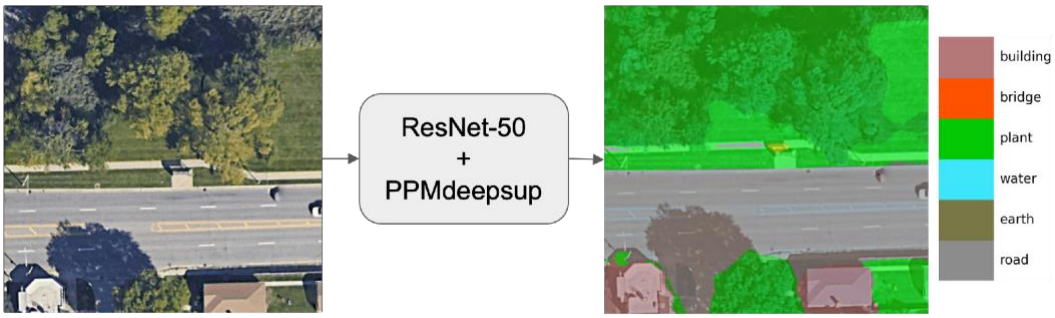
4.3 Predictive performance evaluation

In this stage, we assess the predictive power of different urban image types in modeling travel behavior. Specifically, for each location (e.g., census tract), the model takes key socio-demographic indicators (e.g., population) as control variables and urban images, either street view or satellite imagery at varying zoom levels as inputs. The output consists of the proportions of individuals using cars, public transit, or active modes for travel in that location. For each predictive method, the same outcome definition and input-variable specification are used under a given imagery setting; the differences across specifications therefore come from the imagery-derived feature vectors and, for satellite imagery, the zoom-level setting. Detailed model specifications and input-variable definitions are provided in Appendix C.

To ensure a comprehensive comparison, we employ three representative machine learning approaches: MLR, RF, and GBDT were selected because they provide complementary modeling baselines with increasing flexibility. MLR serves as a linear baseline, RF captures nonlinear relationships and feature interactions through bagged decision trees, and GBDT represents a boosting-based ensemble model commonly used for structured tabular prediction tasks. Using these three models allows us to examine whether the relative predictive value of satellite and street-view imagery remains consistent across different model families.



a . Street-view semantic segmentation (DeepLabV3+ on Cityscapes)



b. Satellite-image semantic segmentation (CSAILVision model; trained on ADE20K)

Figure 5. Examples of semantic segmentation results for the two image sources. (a) Street-view image and its segmentation result generated by DeepLabV3+. (b) Satellite image and its segmentation result generated by the MIT CSAIL semantic segmentation model with a ResNet-50 encoder and PPMdeepsup decoder trained on the ADE20K dataset.

- **Multiple Linear Regression (MLR):** A basic linear model that assumes a direct relationship between the input features and the target variable. It takes socio-demographic attributes (as will be introduced in Section 3 and extracted features from either satellite or street view images (detailed in Section 4.2) as inputs, generating a value between 0 and 1 that represents the relative share of a specific travel mode (e.g., driving).
- **Random Forest (RF):** An ensemble of numerous decision trees that can model nonlinearities and complex feature interactions. Like MLR, it incorporates socio-demographic variables and extracted image features as inputs and outputs travel mode share.
- **Gradient Boosting Decision Tree (GBDT):** A boosting technique that iteratively refines its estimates by fitting new trees to the residuals of previous ones. It utilizes the same socio-demographic and image-based features as the previous models, improving accuracy through adaptive learning.

To measure model performance, we use the coefficient of determination (R^2), which quantifies how much of the observed variation in mode share is explained by the predictions:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \left(y_i - \underline{y}\right)^2} \quad (1)$$

where y_i and $\hat{y}_i$ represents the ground-truth and predicted mode share respectively. Comparing R^2 scores across models and input types helps us determine which types of urban images yield more accurate and robust travel mode share predictions.

To ensure that model performance reflects generalization rather than goodness-of-fit on the training data, we adopted a held-out evaluation protocol. Specifically, the dataset was randomly divided into an 80% training set and a 20% held-out test set. All models were trained on the training set and evaluated only on the held-out test set. Thus, all R^2 values reported in this study are out-of-sample estimates computed on unseen test data. To quantify uncertainty, we computed bootstrap 95% confidence intervals for the held-out test-set R^2 values based on the test-set predictions.

4.4 Interpretation analysis

Beyond prediction performance, understanding how the built environment influences travel behavior is crucial for informing urban and transportation policy. In this stage, we assess the interpretability of street view and satellite images in capturing the relationship between urban form and travel behavior. Our analysis relies on Shapley Additive Explanations (SHAP), an explainable AI technique grounded in game theory (Lundberg & Lee, 2017). SHAP assigns each feature an optimal Shapley value, quantifying its contribution to the model's prediction. The Shapley value ϕ_i for a feature$i$ is calculated as:

$$\phi_i = \sum_{S \subseteq \{1,\dots,n\} \setminus \{i\}} \frac{|S|!\ (n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)] \tag{2}$$

where S is a subset of all features excluding i, n is the total number of features, and $f(S)$ is the model's prediction given the feature subset S.

In this study, we applied SHAP to GBDT because it achieved strong overall out-of-sample predictive performance across the evaluated settings and is well suited for tree-based SHAP interpretation. The model takes segmented features from urban images as inputs—representing the relative proportions of different urban features in the scene—and outputs predicted commuting mode share. SHAP has been increasingly used in interpretable machine-learning studies to explain feature contributions in travel behavior and transportation analysis (Kashifi et al., 2022; Yin et al.,2024). In this study, a SHAP value indicates the contribution of a feature to the predicted outcome relative to the model baseline: a positive SHAP value means that the feature increases the predicted mode share, whereas a negative SHAP value means that it decreases the prediction. Larger absolute SHAP values indicate stronger feature influence, and the range of SHAP values is not fixed a priori but depends on the scale of the model output and the fitted model. The resulting SHAP values reveal how various urban features are associated with travel behavior and how these effects differ between satellite and street view imagery.

5 Results

This section presents a comparative analysis of street view and satellite imagery in modeling travel mode share, focusing on both predictive ability and interpretability. Section 5.1 systematically evaluates the performance of various training models using different types of imagery data. Section 5.2 presents detailed case studies to illustrate when and why satellite and street view imagery yield differing predictive outcomes by providing numeric and visual analyses. Section 5.3 examines the interpretability of these models by comparing SHAP value analyses across imagery types.

5.1 Comparing the predictive ability of satellite and street view images

Table 1 summarizes the predictive performance of various models in estimating the mode shares of driving, public transit, and active modes, using either street view or satellite imagery at different zoom levels as input. Across the evaluated machine learning approaches, satellite imagery generally achieves higher held-out test-set R^2 values than street-view imagery, although the magnitude of the advantage varies by model, travel mode, and satellite zoom level. The bootstrap confidence intervals also indicate that some differences are modest and should be interpreted cautiously when intervals overlap. Overall, the results suggest that satellite imagery provides broader spatial context that is useful for predicting tract-level commuting mode share, whereas the single street-view image available for each tract captures more localized streetscape information.

Table 1. Held-out test-set R2 values for predicting commuting mode share. Values in brackets indicate bootstrap 95% confidence intervals. Rows represent model--travel mode combinations, and columns represent imagery source and satellite zoom level. The highest R2 value in each row is highlighted in bold.

Model	Travel Mode	Satellite-14	Satellite-15	Satellite-16	Satellite-17	Street View
MLR	Driving	0.329 [0.233, 0.411]	**0.351 [0.262, 0.430]**	0.327 [0.234, 0.405]	0.327 [0.225, 0.412]	0.321 [0.228, 0.403]
	Public Transit	0.340 [0.201, 0.434]	0.341 [0.206, 0.439]	**0.344 [0.208, 0.437]**	0.330 [0.189, 0.431]	0.343 [0.210, 0.437]
	Active modes	**0.028 [-0.079, 0.131]**	0.026 [-0.079, 0.126]	-0.001 [-0.119, 0.120]	0.005 [-0.116, 0.131]	0.005 [-0.096, 0.119]
	Mean across modes	0.232 [0.118, 0.325]	**0.240 [0.130, 0.332]**	0.223 [0.107, 0.321]	0.221 [0.100, 0.325]	0.223 [0.114, 0.319]
RF	Driving	**0.694 [0.607, 0.769]**	0.675 [0.583, 0.759]	0.684 [0.594, 0.762]	0.655 [0.561, 0.738]	0.663 [0.571, 0.742]
	Public Transit	0.811 [0.731, 0.863]	**0.815 [0.733, 0.868]**	0.809 [0.715, 0.864]	0.787 [0.704, 0.841]	0.762 [0.679, 0.820]
	Active modes	**0.554 [0.382, 0.739]**	0.548 [0.382, 0.736]	0.548 [0.379, 0.733]	0.466 [0.308, 0.646]	0.507 [0.349, 0.686]
	Mean across modes	**0.687 [0.573, 0.790]**	0.679 [0.566, 0.787]	0.680 [0.563, 0.786]	0.636 [0.524, 0.741]	0.644 [0.533, 0.749]
GBDT	Driving	**0.715 [0.616, 0.792]**	0.695 [0.586, 0.780]	0.687 [0.585, 0.771]	0.666 [0.547, 0.760]	0.655 [0.541, 0.744]
	Public Transit	**0.835 [0.737, 0.893]**	0.812 [0.703, 0.881]	0.789 [0.642, 0.875]	0.788 [0.654, 0.868]	0.774 [0.641, 0.856]
	Active modes	0.534 [0.368, 0.725]	0.548 [0.374, 0.755]	**0.559 [0.380, 0.762]**	0.502 [0.315, 0.711]	0.524 [0.368, 0.714]
	Mean across modes	**0.695 [0.573, 0.803]**	0.685 [0.555, 0.805]	0.678 [0.536, 0.803]	0.652 [0.505, 0.780]	0.651 [0.516, 0.771]

A closer inspection of the results confirms the overall advantage of satellite imagery, while showing that the magnitude of this advantage varies across models and travel modes. For MLR, the gains from satellite imagery are small to modest. For example, the highest driving R^2 increases from 0.321 for street view to 0.351 for satellite imagery, while public transit shows nearly identical performance between street view and the best satellite setting. Active modes remain difficult for MLR, with low R^2 values and wide confidence intervals. For the tree-based models, RF and GBDT achieve substantially higher predictive performance than MLR, and the advantage of satellite imagery becomes more consistent. Under GBDT, the best satellite settings increase R^2 from 0.655 to 0.715 for driving and from 0.774 to 0.835 for public transit. For active modes, the improvement is more moderate, with the best satellite setting reaching R^2=0.559 compared with R^2=0.524 for street view. Overall, these results indicate that satellite imagery generally provides stronger out-of-sample predictive performance than street-view imagery, particularly when used with nonlinear tree-based models, but the size of the improvement depends on the travel mode and satellite zoom level.

In addition to image types, Table 1 shows that satellite zoom level affects predictive performance, although the effect varies by model and travel mode. Four distinct zoom levels (14 to 17) are considered in this study, each of which captures a different spatial extent and level of detail. For instance, a zoom level of 17 typically covers an area of approximately 500m * 500m, roughly corresponding to a neighborhood scale. The detailed calculation of the cover area at different zoom levels is shown in Appendix A. Lower zoom levels capture much larger areas, potentially spanning several kilometers, but with a coarser resolution. The results suggest the following patterns:

- **Driving and Public Transit:** For driving and public transit, lower zoom levels generally provide stronger performance, especially for the tree-based models. For example, when using GBDT to predict driving mode share, zoom 14 achieves the highest R^2 of 0.715, compared with 0.666 at zoom 17. Similarly, for public transit, zoom 14 achieves the highest GBDT performance (R^2=0.835), compared with 0.788 at zoom 17. RF shows a similar pattern, with the highest driving performance at zoom 14 (R^2=0.694) and the highest public transit performance at zoom 15 (R^2=0.815). These results suggest that broader spatial context helps capture features such as major road networks, land-use structure, and regional accessibility, which are important for modeling driving and transit commuting patterns.
- **Active Modes:** In contrast, active modes show a less uniform relationship with zoom level. With GBDT, zoom 16 achieves the highest performance for active modes (R^2=0.559), slightly outperforming zoom 14 (R^2=0.534), zoom 15 (R^2=0.548), and zoom 17 (R^2=0.502). RF also performs best at zoom 14 for active modes (R^2=0.554), but the differences between zoom levels 14--16 are relatively small, and their confidence intervals overlap. These results indicate that active-mode prediction may benefit from finer local detail, but the advantage of higher zoom levels is not consistent across models. This likely reflects the fact that active modes depend on both neighborhood-scale spatial structure and highly localized pedestrian or cycling conditions, which may not be fully captured by satellite imagery alone.

These findings highlight the importance of selecting an appropriate zoom level. The choice of zoom level represents a trade-off between the breadth of contextual information and the granularity of local details. A larger spatial extent (lower zoom level) provides a comprehensive view of the urban environment that benefits predictions for long-distance travel modes, including driving and public transit. On the other hand, higher zoom levels provide finer details required for active mode prediction.

Given the same prediction model and zoom level, predictive accuracy varies by travel mode. Public transit mode shares generally yield higher R^2 values than driving or active modes. For example, using RF, public transit predictions range from 0.733 to 0.750, compared to 0.650–0.693 for driving and 0.675–0.703 for active travel. A similar pattern holds for GBDT, with public transit achieving R^2 values of 0.729–0.758, outperforming the other modes. This superior performance may be attributed to the fact that public transit infrastructure (e.g., dense transit networks and station proximities) is more readily discernible in aerial views. In contrast, driving and active modes are less predictable. This may be explained that driving mode shares tend to involve broader urban patterns that are less distinct than transit-related features, while active mode share depends on highly localized factors, such as pedestrian pathways, sidewalk conditions, and minor urban design elements, which might not be as prominent in satellite images.

In summary, these results highlight the superior predictive power of satellite imagery over street view imagery, the crucial role of appropriate zoom level selection, and the predictability of different travel modes. This multifaceted insight provides valuable guidance for future research understanding and modeling travel behavior using remotely sensed imagery.

5.2 Case study

To further investigate the strengths and limitations of satellite and street view imagery in predicting travel mode shares, we present two representative examples in Table 2. The satellite and street view images corresponding to the two examples are shown in Figure 6.

These case studies showcase the predicted demand distributions from both imagery types, alongside ground-truth travel behavior, and provide qualitative reasoning based on the visual features observed. The two case-study locations were selected to illustrate contrasting built-environment contexts and differing prediction patterns between the satellite- and street-view-based models. Case 1 is located in the Boca West area of West Boca Raton, Palm Beach County, in southeastern Florida. Case 2 is located near Page Field in Fort Myers, Lee County, in southwestern Florida. For each case-study location, the street-view image shown in Figure 6 corresponds to the single street-view image retrieved for that census tract using its geographic centroid as the reference point. By examining these contrasting contexts—a suburban residential setting in Example I and a dense suburban setting in Example II—we highlight when and why each imagery source may overestimate or underestimate certain transportation modes.

Table 2. Observed and predicted shares of commuting by driving, public transit, and active modes for two illustrative case-study locations

Example I: Suburban residential setting

Source	Driving	Public Transit	Active Modes
Observed mode share	0.715	0.164	0.046
Satellite-based prediction	0.728	0.156	0.044
Street-view-based prediction	0.758	0.134	0.037

Example II: Dense suburban setting

Source	Driving	Public Transit	Active Modes
Observed mode share	0.854	0.042	0.068
Satellite-based prediction	0.863	0.033	0.051
Street-view-based prediction	0.839	0.033	0.041



(a) Street view (Example I: Suburban residential setting)

(b) Satellite (Example I: Suburban residential setting)

(c) Street view (Example II: Dense suburban setting)

(d) Satellite (Example II: Dense suburban setting)

Figure 6. Satellite and street-view images used in case studies

In Example I, the satellite image presents a well-ordered suburban scene: clearly delineated housing parcels, extensive patches of grassland, and a meshed road network that stitches neighborhoods together. These macro-scale elements—continuous rows of roofs, evenly spaced streets, and open green buffers—immediately convey a planned but moderately dense setting where cars are expected to circulate easily yet walking and transit remain plausible. On the other hand, the street view imagery primarily illustrates immediate roadside conditions, characterized by abundant greenery, spacious roads, but notably lacking visible pedestrian paths or cycling infrastructure. This narrow field of view obscures deeper cues about public-transit accessibility or pedestrian permeability.

For the prediction results, both image types reveal the correct mode-share ranking, but accuracy differs. The satellite-based model aligns closely with ground truth, while the street-view model nudges its prediction toward a more car-oriented profile than ground truth warrants, over-estimating driving share while under-scoring transit and active modes. This suggests that macro-level cues from satellite imagery provide structural

land-use patterns to guide predictions, whereas street-view estimates are more sensitive to localized visual context.

In Example II, the case represents a dense suburban context, combining relatively dense buildings patterns with transportation and land-use features. The satellite image is dominated by an "X-shaped" pair of airport runways. There are some large warehouse-style buildings, scattered low-rise housing blocks, and several broad parking lots. Roadways form a loose rectilinear grid. Continuous green belts separate land-use clusters, further lengthening trip distances. Overall, the picture is of a coarse-grained, car-oriented environment with plenty of space for vehicle movement. At street view, there is a gravel drive lined with a pickup truck, farm or light-industrial equipment, a metal shed, and mature trees edging an informal lawn. The lane lacks sidewalks, curbs, bus stops, or marked cycle lanes. The visual emphasis on private vehicles, unpaved surfaces, and sparse pedestrian infrastructure reinforces the expectation of automobile dominance.

For this example, the satellite imagery slightly overestimates the driving mode share, while the street-view imagery underestimates it, with a larger error. This can be because the satellite imagery captures stronger macro-scale indicators of car-oriented environments—such as runways, large parking lots, and land-use separation. Meanwhile, the street-view model also underestimates active modes more than the satellite-based model. One possible reason is that the satellite image reveals medium-density development patterns—such as dispersed low-rise housing and warehouse clusters—which can support active modes. In contrast, the street view lacks visible pedestrian or cycling infrastructure, which may lead the model to further underestimate active modes.

In summary, these two cases highlight the strengths of satellite imagery in capturing macro-scale urban features—such as building density, land-use patterns, and spatial organization—that are critical for accurately predicting travel mode share. In contrast, street-view imagery tends to be more sensitive to localized sampling contexts, where subtle visual cues—like the presence or absence of sidewalks, curbs, or bike lanes—can disproportionately influence the model's perception of travel behavior.

5.3 SHAP analysis of imagery-derived features

This subsection reports the main SHAP patterns observed in the prediction models, while broader interpretation of possible mechanisms and planning implications is provided in the Discussion section. In this section, we apply SHAP to GBDT because it achieves strong overall out-of-sample predictive performance across the evaluated settings and is well suited for tree-based SHAP interpretation. The SHAP summary plots are shown in Figure 8. In these plots, each point represents one observation. The horizontal position of a point indicates the SHAP value of a feature for that observation: positive SHAP values mean that the feature increases the predicted mode share relative to the model baseline, whereas negative SHAP values mean that the feature decreases it. The color of each point represents the original feature value, with warmer colors indicating higher feature values and cooler colors indicating lower feature values. Features with larger absolute SHAP values and wider dispersion are interpreted as having stronger or more heterogeneous contributions to the model predictions. Figure 7 presents the spatial distribution of the sampled locations across Florida, together with the census tract boundaries used as the geographic units of analysis.

5.3.1 SHAP value for driving

The SHAP summary plots for satellite and street-view imagery reveal complementary insights into the imagery-derived features associated with predicted driving commuting mode share. In the satellite-based analysis, vegetation ("plant") is one of the most

important predictors and shows a context-dependent pattern across observations. Similarly, extensive bare soil or undeveloped land (classified as "earth") shows a positive SHAP contribution to predicted driving mode share. Large water bodies show a negative SHAP contribution to predicted car use.

At the street-level view, the importance of micro-scale urban design attributes emerges distinctly. Building coverage emerges as the most influential predictor, where increased visible building area at street level is negatively correlated with the share of driving. Similarly, openness to the sky also positively correlates with driving. Street-level vegetation positively contributes to increased driving predictions. However, visible cars on streets negatively influence driving predictions. Extensive visible road surface area shows a mixed pattern in SHAP analysis rather than a uniformly positive or negative association with driving predictions. In contrast, sidewalk coverage consistently shows a negative contribution to predicted driving mode share.

Taken together, these satellite and street-view analyses suggest that dispersed suburban areas are associated with higher predicted driving dependence, while dense suburban environments are associated with lower predicted private-car use.

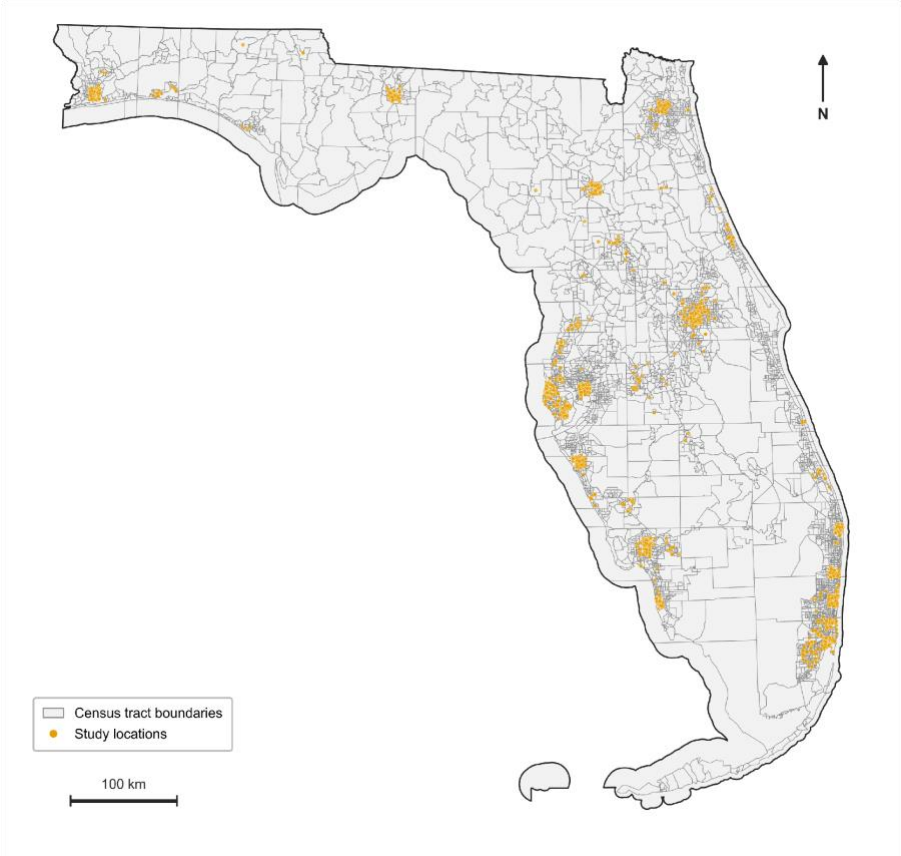


Figure 7. Spatial distribution of the sampled locations across Florida

5.3.2 SHAP value for public transit

The SHAP analysis of public transit mode share based on satellite imagery reveals that the vegetation coverage exhibits a mixed relationship with transit use, with SHAP values widely spread across both sides of 0. Building density shows a similarly mixed association with predicted public transit mode share. Notably, road, water bodies, and

bridges all show positive SHAP contributions to predicted public transit mode share. Undeveloped land has minimal systematic influence, suggesting limited predictive power for transit share.

The street-level analysis shows that sky openness has a complex yet overall positive relationship with transit use. Street-side vegetation also shows complexity but tends toward a negative influence. Building frontage coverage displays a weak but consistently positive correlation. Visible pedestrians and sidewalk coverage further reinforce transit use with consistent positive impacts. Meanwhile, visible road surfaces show more mixed influences. A higher presence of visible cars shows a positive association with public transit use.

Overall, satellite and street-level analyses consistently highlight that denser, pedestrian-oriented environments with moderate vegetation are associated with higher predicted public transit mode share, while low-density, undeveloped contexts are associated with lower predicted public transit mode share.

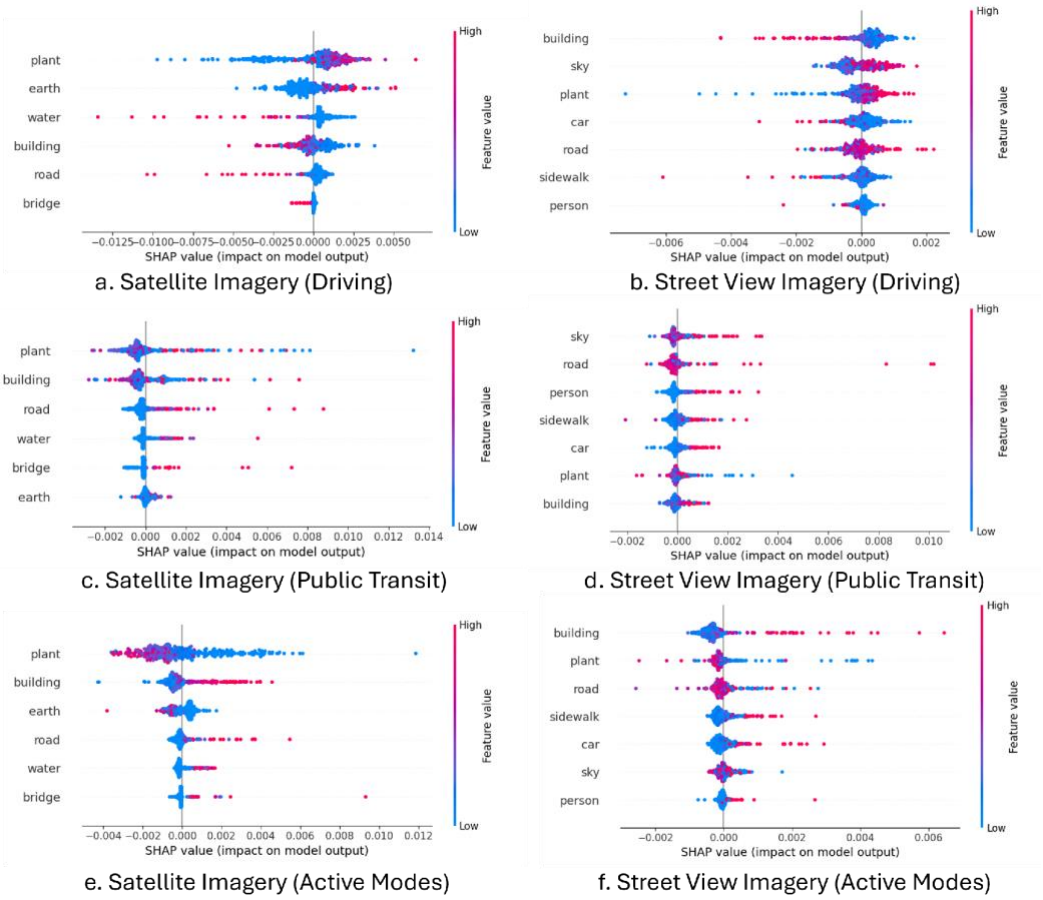


Figure 8. SHAP summary for three mode predictions: driving, public transit, and active modes

5.3.3 SHAP value for active modes

The SHAP analysis of active mode share from satellite imagery reveals several distinct associations between land-cover features and predicted active mode share. Vegetation coverage shows a clear negative SHAP contribution. In contrast, building coverage exhibits a strong and consistent positive contribution, indicating that compact, built-up environments with higher destination density are associated with higher

predicted active mode share. Undeveloped land also shows a negative relationship. On the other hand, road, water, and bridge coverage all show positive SHAP contributions to predicted active mode share.

The street-view-based SHAP analysis highlights micro-scale environmental features associated with predicted active mode share. Buildings contribute positively, indicating that visually high-density environments are associated with higher predicted walking and biking shares. The presence of sidewalks and visible pedestrians also contributes positively to the model's predictions. Visible cars are associated with a modest increase in active mode share. In contrast, vegetation and road surface coverage both exhibit negative relationships with active modes. Sky openness shows a complex pattern. The number of people visible in street view imagery has a mild but consistent positive impact on active mode share prediction.

Overall, both satellite and street-level analyses indicate that high-density, pedestrian-friendly environments—such as those with sidewalks and waterfronts—are associated with higher predicted active mode share, whereas low-density areas with pedestrian-unfriendly features, such as bare ground, are associated with lower predicted active mode share.

6 Conclusions

6.1 Summary of main findings

Previous studies have demonstrated that urban images can effectively predict perceived safety (Naik et al., 2014), poverty (Jean et al., 2016), income (Suel et al., 2019), and other socioeconomic indicators (Gebru et al., 2017) in cities. However, a significant research gap remains in how and why the two dominant categories of imagery data, street-view imagery and satellite imagery, differ in their predictive performance and interpretability for commuting mode-share modeling. This study addresses this gap by predicting tract-level commuting mode shares using either satellite or street-view imagery in Florida. By applying machine learning models to segmentation-derived image features and interpreting the results with SHAP, our empirical analysis yields the following findings.

First, models using satellite imagery generally outperform those based on street-view imagery across the evaluated travel modes. Across multiple predictive approaches, including MLR, RF, and GBDT, satellite-based models achieve higher out-of-sample R2 values than street-view-based models in most comparisons. For the tree-based models, the relative improvement is typically moderate, ranging from approximately 2% to 8% in most comparisons, with the largest observed gain around 9%. One possible explanation is that the satellite imagery captures visible built-environment patterns associated with commuting behavior, such as building density, land-use patterns, and spatial organization (Sun et al., 2017). At the same time, using only one street-view image per census tract may limit how much local street-level information is captured by the street-view models, which may partly contribute to their lower predictive performance. Commuting mode share may also be related to non-visible factors, such as socio-demographic variables and local policy conditions (Cervero, 2002; Ding et al., 2018).

Second, Google Maps satellite zoom level influences predictive performance. Across travel modes, differences in zoom level result in noticeable variation in model performance. In general, lower zoom levels, which provide broader spatial context, perform better for driving and public transit, while active modes show a more mixed relationship with zoom level. This suggests that the appropriate imagery scale depends on the travel mode being modeled.

Third, SHAP analysis highlights the complementary interpretability of the two imagery sources. Satellite imagery emphasizes macro-scale land-cover and spatial-structure features, whereas street-view imagery emphasizes localized streetscape features such as sidewalks, sky openness, visible pedestrians, visible cars, and building frontage.

6.2 Interpretation of SHAP results

The SHAP results show how segmentation-derived features contribute to model predictions, but they do not establish that these visible features independently cause changes in commuting mode share (Janzing et al., 2020; Lundberg & Lee, 2017).

For driving mode share, satellite-based SHAP results indicate that broad land-cover features are important contributors to prediction. Vegetation and bare ground generally show positive SHAP contributions for many observations, whereas water tends to show negative contributions. These patterns suggest that the satellite-based model relies on spatial-context and land-cover signals when predicting driving mode share. Street-view results, by contrast, emphasize localized streetscape features. Building coverage and sidewalk coverage generally show negative SHAP contributions to predicted driving mode share, while sky openness and street-level vegetation tend to contribute positively. Visible road surface area shows mixed SHAP values, indicating that this feature may correspond to different functional street contexts.

For public transit mode share, satellite-based SHAP results show that road, water, and bridge coverage generally contribute positively to predicted transit share, while vegetation and building coverage show more mixed associations. Street-view results indicate that the positive association of sky openness with both driving and public transit should not be interpreted as a mode-specific effect. Rather, sky openness may reflect wide and open roadway corridors that can accommodate both automobile flow and bus service along major arterial roads.

For active modes, satellite-based SHAP results show that building coverage has a consistently positive contribution, while vegetation and undeveloped land generally contribute negatively. Road, water, and bridge coverage also show positive contributions in some observations. Street-view results show vegetation and road surface coverage tend to show negative contributions. These findings are broadly consistent with the idea that active modes are associated with more compact and pedestrian-oriented environments, although the observed SHAP patterns should still be interpreted as predictive associations.

Overall, the SHAP analysis suggests that some well-established built-environment relationships, such as the association between lower-density development and higher car use or the association between pedestrian infrastructure and active modes, can be recovered from imagery-derived features. The contribution of this study is therefore not that these relationships are substantively new, but that they can be identified from image-derived feature representations and compared across satellite and street-view imagery within the same geographic context.

6.3 Planning and methodological implications

These findings have several implications for applying visual AI to commuting mode-share prediction. First, satellite imagery appears particularly useful for large-area mobility modeling and regional comparison because it captures broader spatial context, including land-use structure, development intensity, and transportation corridors. This makes satellite imagery suitable for applications where practitioners need to compare neighborhoods or identify areas that may warrant closer transportation or land-use analysis, consistent with recent studies using satellite imagery to examine urban growth-

transportation interactions and to enhance travel demand and behavior modeling (Patil et al., 2025; Wang et al., 2024). Building on these literatures, our analysis evaluates its predictive value for commuting mode share and directly compares it with street-view imagery within the same study areas. These results highlight the value of satellite imagery as a scalable source of built-environment information for aggregate commuting mode-share prediction.

Second, the sensitivity of predictive performance to Google Maps satellite zoom level suggests that image scale should be selected according to the planning question and travel mode of interest. Lower zoom levels may be more useful for capturing broader spatial patterns related to driving and public transit, while finer-scale imagery may provide additional information for active modes, although the revised results show that this relationship is not uniform across models. This finding is consistent with previous studies showing that the scale at which urban features are represented can affect the information extracted from imagery and subsequent classification or modeling performance (Pacifici et al., 2009; Zhang & Du, 2016). Our study extends this multi-scale perspective to commuting mode-share prediction by showing that the most informative Google Maps zoom level can vary across travel modes. Rather than assuming that a single image scale is universally optimal, these results suggest that image scale should be treated as an important methodological choice when using satellite imagery for transportation-related prediction.

Third, the two imagery sources should not be treated as directly equivalent. Satellite imagery captures macro-scale land-cover and spatial-structure features, while street-view imagery captures micro-scale streetscape features such as sidewalks, pedestrian presence, vehicles, and sky openness. From a policy perspective, satellite imagery may be more suitable for large-area comparison, while street-view imagery can provide complementary information on local streetscape conditions. This distinction can help practitioners select imagery sources according to the spatial scale of the planning question. This distinction is consistent with previous studies showing that aerial or satellite imagery and street-view imagery provide different but complementary perspectives of the built environment, with top-down imagery offering broader spatial coverage and street-view imagery capturing features visible from the human-scale street perspective (Larkin et al., 2021; Gaw et al., 2022;). Prior work has also shown that street-view imagery is particularly useful for measuring fine-scale, human-scale environmental characteristics that may be difficult to observe from conventional remote-sensing data alone (Biljecki & Ito, 2021). Our study extends this distinction to commuting mode-share prediction by directly comparing the predictive and interpretive value of these two imagery sources for the same census tracts. These findings therefore suggest that the choice of imagery source should depend not only on predictive performance but also on the type and spatial scale of built-environment information required for a particular planning application.

These findings suggest that planners can benefit from using satellite and street-view imagery together because the two sources provide complementary perspectives on neighborhood commuting patterns: satellite imagery captures broader land-use and spatial-structure features, while street-view imagery reveals localized streetscape conditions. Because satellite imagery captures macro-scale features, it can help planning agencies identify areas where observed commuting mode shares, predicted mode shares, or visible built-environment conditions appear inconsistent with policy goals, while street-view imagery can be used for localized diagnostic assessment.

More broadly, this study demonstrates the value of interpretable GeoAI in linking visual urban form to tract-level commuting mode shares. These results should be interpreted as predictive associations rather than isolated built-environment effects, since

commuting mode choice is also shaped by travel time, travel cost, transit accessibility, parking conditions, and other transportation-system factors not directly captured by imagery.

6.4 Limitations and future research

This study also presents several limitations that warrant consideration. First, the analysis was confined to a limited set of Google Maps satellite zoom levels, which may restrict the generalization of findings to other image scales or retrieval settings. The centroid-based sampling strategy may also underrepresent within-tract heterogeneity, especially in large or internally diverse census tracts where localized streetscape differs. Prior studies have demonstrated the value of aggregating multiple street-view observations to characterize neighborhood-level built environments (Nguyen et al., 2018). Accordingly, when data permit, future research should test multi-point sampling strategies within each census tract and evaluate whether additional street-view observations improve predictive performance and interpretability.

Second, although satellite imagery generally outperformed street-view imagery in predictive performance, this comparison should be interpreted cautiously because the two data sources produce heterogeneous feature spaces. For both image types, we used pre-trained semantic segmentation tools to extract the proportions of visible urban features from images. Thus, the feature space of each imagery source is defined by segmentation-derived urban-feature proportions. The satellite-image feature space is six-dimensional, including building, road, plant, water, earth, and bridge, whereas the street-view feature space is seven-dimensional, including road, building, sky, sidewalk, plant, car, and person. Some features, such as building, road, and plant, are identified in both image types and are partially comparable. However, other features are view-specific: sky, sidewalk, vehicles, and pedestrians are captured only in street-view imagery, whereas water, earth, and bridge are captured only in satellite imagery. Therefore, the comparison should be interpreted as a comparison between two segmentation-derived visual representations of the built environment rather than a direct comparison between identical feature sets. Future research could integrate satellite and street-view imagery within a unified multi-view representation, allowing the two image sources to contribute complementary information within the same modeling framework. This direction is consistent with recent multi-view urban sensing studies that combine satellite and street-view imagery to improve urban representation and prediction (Huang et al., 2025; Li et al., 2026; Ye et al., 2024).

Third, the potential value of hybrid models integrating both satellite and street-view imagery was not fully explored. Such models could provide a more comprehensive representation of both macro-scale spatial structure and micro-scale streetscape conditions. Another limitation is that the Census data used in this study capture commuting mode shares only. As a result, the analysis reflects work-related travel patterns rather than the full spectrum of travel behavior, such as shopping, leisure, school, or other non-work trips.

In addition, this study does not explicitly address spatial autocorrelation in model validation. Nearby census tracts may share similar built-environment and socio-demographic characteristics, so conventional random train-test splits may overestimate out-of-sample predictive performance. Future work should use spatial cross-validation or geographically blocked validation to further test model generalizability. A related data limitation concerns temporal alignment. The satellite and street-view imagery data were collected in mid-2023, whereas the commuting mode-share data were derived from the 2020 Census. Although major built-environment features such as road layout and urban

form generally evolve over longer time scales, local redevelopment, new construction, roadway changes, or changes in Google Street View imagery over time may introduce temporal mismatches for some census tracts. Herold et al. (2003) and M. Li et al. (2023) indicate that large-scale urban form generally evolves over time, while street-level infrastructure can also exhibit detectable changes across years. Therefore, the imagery-derived features should be interpreted as approximate representations of the built environment around the Census period rather than exactly at the same time as measurements.

Finally, the geographic scope of the study was limited to Florida. We would expect the main findings to remain broadly similar in states with built-environment characteristics comparable to Florida. However, the relative predictive value of satellite and street-view imagery may differ in denser states or regions with more compact urban form and stronger transit orientation, where fine-grained streetscape characteristics may play a more important role in shaping travel behavior. Expanded studies across varied locations are needed to validate and extend these findings.

Author contribution

The authors confirm their contributions to this paper as follows: study conception and design: J. Cao, S. Wang; methodology, data collection, software development, formal analysis, and visualization: J. Cao; original manuscript drafting: J. Cao, M. Li, Y. Liang; manuscript review and editing: J. Cao, M. Li, Y. Liang, S. Wang; supervision: M. Li, Y. Liang, S. Wang. All authors reviewed and approved the final manuscript.

Code and data availability

This study was conducted by academic institutions for scholarly research purposes and was not profit-driven. The commuting mode share data used in this study were derived from publicly available 2020 U.S. Census data. The code used for image feature processing, model training, and evaluation will be made available through https://github.com/JinghaoStarCao/Urban-Visual-AI-for-Travel-Behavior-Prediction upon acceptance or publication, subject to institutional and data-use restrictions. The satellite imagery was obtained through Google Maps API, and the street-view imagery was obtained through Google Street View Static API. Both imagery sources are subject to Google Maps Platform Terms of Service. As a result, the raw imagery cannot be publicly archived or redistributed. Derived features and related metadata may be shared only subject to applicable platform terms and other restrictions.

Declaration of AI-assisted writing

During the preparation of this manuscript, the authors used ChatGPT to assist with language editing and improving the clarity of the text. The authors reviewed and revised all AI-assisted outputs and took full responsibility for the content of the manuscript.

Appendix

Appendix available as a supplemental file at https://doi.org/10.5198/jtlu.2026.2873.

References

Albert, A., Kaur, J., & Gonzalez, M. C. (2017). Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1357–1366). https://doi.org/10.1145/3097983.3098070

Ayush, K., Uzkent, B., Burke, M., Lobell, D., & Ermon, S. (2020). Generating interpretable poverty maps using object detection in satellite images. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (pp. 4410–4416). IJCAI. https://doi.org/10.24963/ijcai.2020/608

Bagwari, N., Kumar, S., & Verma, V. S. (2023). A comprehensive review on segmentation techniques for satellite images. *Archives of Computational Methods in Engineering*, *30*(7), 4325–4358. https://doi.org/10.1007/s11831-023-09939-4

Berriel, R. F., Lopes, A. T., De Souza, A. F., & Oliveira-Santos, T. (2017). Deep learning-based large-scale automatic satellite crosswalk classification. *IEEE Geoscience and Remote Sensing Letters*, *14*(9), 1513–1517. https://doi.org/10.1109/LGRS.2017.2719863

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, *215*, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

Cai, Q., Abdel-Aty, M., Zheng, O., & Wu, Y. (2022). Applying machine learning and Google street view to explore effects of drivers' visual environment on traffic safety. *Transportation Research Part C: Emerging Technologies*, *135*, 103541. https://doi.org/10.1016/j.trc.2021.103541

Cervero, R. (2002). Built environments and mode choice: Toward a normative framework. *Transportation Research Part D: Transport and Environment*, *7*(4), 265–284. https://doi.org/10.1016/S1361-9209(01)00024-4

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, *2*(3), 199–219. https://doi.org/10.1016/S1361-9209(97)00009-6

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision* (pp. 833–851). Springer International Publishing. https://doi.org/10.1007/978-3-030-01234-2_49

Cheng, L., Chu, S., Zong, W., Li, S., Wu, J., & Li, M. (2017). Use of tencent street view imagery for visual perception of streets. *ISPRS International Journal of Geo-Information*, *6*(9), 265. https://doi.org/10.3390/ijgi6090265

Crane, R. (2000). The influence of urban form on travel: An interpretive review. *Journal of Planning Literature*, *15*(1), 3–23. https://doi.org/10.1177/08854120022092890

Ding, C., Cao, X., & Wang, Y. (2018). Synergistic effects of the built environment and commuting programs on commute mode choice. *Transportation Research Part A: Policy and Practice*, *118*, 104–118. https://doi.org/10.1016/j.tra.2018.08.041

Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American Planning Association*, *76*(3), 265–294. https://doi.org/10.1080/01944361003766766

Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design*, *14*(1), 65–84. https://doi.org/10.1080/13574800802451155

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

Ganji, A., Zhang, M., & Hatzopoulou, M. (2022). Traffic volume prediction using aerial imagery and sparse data from road counts. *Transportation Research Part C: Emerging Technologies, 141*, 103739. https://doi.org/10.1016/j.trc.2022.103739

Gaw, L., Chen, S., Chow, Y., Lee, K., & Biljecki, F. (2022). Comparing street view imagery and aerial perspectives in the built environment. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *10*, 49–56. https://doi.org/10.5194/isprs-annals-X-4-W3-2022-49-2022

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google street view to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, *114*(50), 13108–13113. https://doi.org/10.1073/pnas.1700035114

Goel, R., Garcia, L. M., Goodman, A., Johnson, R., Aldred, R., Murugesan, M., Brage, S., Bhalla, K., & Woodcock, J. (2018). Estimating city-level travel patterns using street imagery: A case study of using Google street view in Britain. *PloS One*, *13*(5), e0196521. https://doi.org/10.1371/journal.pone.0196521

Hamim, O. F., & Ukkusuri, S. V. (2024). Towards safer streets: A framework for unveiling pedestrians' perceived road safety using street view imagery. *Accident Analysis & Prevention*, *195*, 107400. https://doi.org/10.1016/j.aap.2023.107400

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2961–2969). https://doi.org/10.1109/ICCV.2017.322

Herold, M., Goldstein, N. C., & Clarke, K. C. (2003). The spatiotemporal form of urban growth: Measurement, analysis and modeling. *Remote Sensing of Environment, 86*(3), 286–302. https://doi.org/10.1016/S0034-4257(03)00075-0

Hipp, J. R., Lee, S., Ki, D., & Kim, J. H. (2022). Measuring the built environment with Google street view and machine learning: Consequences for crime on street segments. *Journal of Quantitative Criminology, 38*, 537–565. https://doi.org/10.1007/s10940-021-09506-9

Hong, J., Shen, Q., & Zhang, L. (2014). How do built-environment factors affect travel behavior? A spatial analysis at different geographic scales. *Transportation*, *41*(3), 419–440. https://doi.org/10.1007/s11116-013-9462-9

Huang, Y., Sanatani, R. P., Liu, C., Kang, Y., Zhang, F., Liu, Y., Duarte, F., & Ratti, C. (2025). No "true" greenery: Deciphering the bias of satellite and street view imagery in urban greenery measurement. *Building and Environment*, *269*, 112395. https://doi.org/10.1016/j.buildenv.2024.112395

Janzing, D., Minorics, L., & Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. *International Conference on Artificial Intelligence and Statistics*, 2907–2916. https://doi.org/10.48550/arXiv.1910.13413

Jean, N., Burke, M., Xie, M., Alampay Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Jing, C., Zhou, W., Qian, Y., & Yan, J. (2020). Mapping the urban population in residential neighborhoods by integrating remote sensing and crowdsourcing data. *Remote Sensing*, *12*(19), 3235. https://doi.org/10.3390/rs12193235

Kashifi, M. T., Jamal, A., Kashefi, M. S., Almoshaogeh, M., & Rahman, S. M. (2022). Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel Behaviour and Society*, *29*, 279–296. https://doi.org/10.1016/j.tbs.2022.07.003

Kuang, W., & Dou, Y. (2020). Investigating the patterns and dynamics of urban green space in China's 70 major cities using satellite remote sensing. *Remote Sensing*, *12*(12), 1929. https://doi.org/10.3390/rs12121929

Larkin, A., Gu, X., Chen, L., & Hystad, P. (2021). Predicting perceptions of the built environment using GIS, satellite and street view image approaches. *Landscape and Urban Planning*, *216*, 104257. https://doi.org/10.1016/j.landurbplan.2021.104257

Larsen, J., & El-Geneidy, A. (2011). A travel behavior analysis of urban cycling facilities in Montréal, Canada. *Transportation Research Part D: Transport and Environment*, *16*(2), 172–177. https://doi.org/10.1016/j.trd.2010.07.011

Li, M., Sheng, H., Irvin, J., Chung, H., Ying, A., Sun, T., Ng, A. Y., & Rodriguez, D. A. (2023). Marked crosswalks in US transit-oriented station areas, 2007–2020: A computer vision approach using street view imagery. *Environment and Planning B: Urban Analytics and City Science*, *50*(2), 350–369. https://doi.org/10.1177/23998083221112157

Li, W., Lai, Y., Xu, L., Xiangli, Y., Yu, J., He, C., Xia, G.-S., & Lin, D. (2023). OmniCity: Omnipotent city understanding with multi-level and multi-view images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 17,397–17,407). IEEE. https://doi.org/10.1109/CVPR52729.2023.01669

Li, Z., Su, Y., Biljecki, F., & Zhao, W. (2026). BuildingMultiView: powering multi-scale building characterization with large language models and multi-perspective imagery. *International Journal of Applied Earth Observation and Geoinformation*, *146*, 105034. https://doi.org/10.1016/j.jag.2025.105034

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*. https://doi.org/10.48550/arXiv.1705.07874

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 779–785). https://doi.org/10.1109/CVPRW.2014.121

Neupane, B., Horanont, T., & Aryal, J. (2021). Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sensing*, *13*(4), 808. https://doi.org/10.3390/rs13040808

Nguyen, Q. C., Khanna, S., Dwivedi, P., Huang, D., Huang, Y., Tasdizen, T., Brunisholz, K. D., Li, F., Gorman, W., Nguyen, T. T., & Jiang, C. (2019). Using Google street view to examine associations between built environment characteristics and US health outcomes. *Preventive Medicine Reports*, *14*, 100859. https://doi.org/10.1016/j.pmedr.2019.100859

Nguyen, Q. C., Sajjadi, M., McCullough, M., Pham, M., Nguyen, T. T., Yu, W., Meng, H.-W., Wen, M., Li, F., Smith, K. R., Brunisholz, K., & Tasdizen, T. (2018). Neighbourhood looking glass: 360º automated characterisation of the built environment for neighbourhood effects research. *Journal of Epidemiology and Community Health*, *72*(3), 260–266. https://doi.org/10.1136/jech-2017-209456

Nguyen, T. T., Nguyen, Q. C., Rubinsky, A. D., Tasdizen, T., Deligani, A. H. N., Dwivedi, P., Whitaker, R., Fields, J. D., DeRouen, M. C., Mane, H., Lyles, C. R., Brunisholz, K. D., & Bibbins-Domingo, K. (2021). Google street view-derived neighborhood characteristics in California associated with coronary heart disease,

hypertension, diabetes. *International Journal of Environmental Research and Public Health*, *18*(19), 10428. https://doi.org/10.3390/ijerph181910428

Nice, K. A., Thompson, J., Wijnands, J. S., Aschwanden, G. D., & Stevenson, M. (2020). The "Paris-end" of town? Deriving urban typologies using three imagery types. *Urban Science*, *4*(2), 27. https://doi.org/10.3390/urbansci4020027

Ning, H., Ye, X., Chen, Z., Liu, T., & Cao, T. (2022). Sidewalk extraction using aerial and street view images. *Environment and Planning B: Urban Analytics and City Science*, *49*(1), 7–22. https://doi.org/10.1177/2399808321995817

Novack, T., & Stilla, U. (2018). Classifying the built-up structure of urban blocks with probabilistic graphical models and Terrasar-x spotlight imagery. *Remote Sensing*, *10*(6), 842. https://doi.org/10.3390/rs10060842

Pacifici, F., Chini, M., & Emery, W. J. (2009). A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sensing of Environment*, *113*(6), 1276–1292. https://doi.org/10.1016/j.rse.2009.02.014

Patil, G. R., Bhatia, S. Y., Yadav, M., Buddhiraju, K. M., & Rao, K. V. (2025). Spatiotemporal urban growth dynamics and its interactions with transportation networks in an unplanned urban conglomeration using satellite imagery: A case study of Mumbai Metropolitan Region, India. *Land Use Policy*, *157*, 107617. https://doi.org/10.1016/j.landusepol.2025.107617

Saelens, B. E., Sallis, J. F., & Frank, L. D. (2003). Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures. *Annals of Behavioral Medicine*, *25*(2), 80–91. https://doi.org/10.1207/S15324796ABM2502_03

Suel, E., Polak, J. W., Bennett, J. E., & Ezzati, M. (2019). Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific Reports*, *9*(1), 6229. https://doi.org/10.1038/s41598-019-42036-w

Sun, B., Ermagun, A., & Dan, B. (2017). Built environmental impacts on commuting mode choice and distance: Evidence from Shanghai. *Transportation Research Part D: Transport and Environment*, *52*, 441–453. https://doi.org/10.1016/j.trd.2016.06.001

Tapiador, F. J., Avelar, S., Tavares-Corrêa, C., & Zah, R. (2011). Deriving fine-scale socioeconomic information of urban areas using very high-resolution satellite imagery. *International Journal of Remote Sensing*, *32*(21), 6437–6456. https://doi.org/10.1080/01431161.2010.512928

Wang, Q., Wang, S., Zheng, Y., Lin, H., Zhang, X., Zhao, J., & Walker, J. (2024). Deep hybrid model with satellite imagery: How to combine demand modeling and computer vision for travel behavior analysis? *Transportation Research Part B: Methodological*, *179*, 102869. https://doi.org/10.1016/j.trb.2023.102869

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? *Computational Urban Science*, *1*(1), 3. https://doi.org/10.1007/s43762-021-00003-0

Wu, F., Li, W., & Qiu, W. (2023). Examining non-linear relationship between streetscape features and propensity of walking to school in Hong Kong using machine learning techniques. *Journal of Transport Geography*, *113*, 103698. https://doi.org/10.1016/j.jtrangeo.2023.103698

Xie, M., Jean, N., Burke, M., Lobell, D., & Ermon, S. (2016). Transfer learning from deep features for remote sensing and poverty mapping. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1), pp. 3929–3935. https://doi.org/10.1609/aaai.v30i1.9906

Ye, J., Lv, Z., Li, W., Yu, J., Yang, H., Zhong, H., & He, C. (2024). Cross-view image geo-localization with Panorama-BEV co-retrieval network. In *European Conference*

on Computer Vision (pp. 74–90). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72913-3_5

Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W., & Zhao, T. (2019). Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network. *Remote Sensing*, *11*(15), 1774. https://doi.org/10.3390/rs11151774

Yin, G., Huang, Z., Fu, C., Ren, S., Bao, Y., & Ma, X. (2024). Examining active travel behavior through explainable machine learning: Insights from Beijing, China. *Transportation Research Part D: Transport and Environment*, *127*, 104038. https://doi.org/10.1016/j.trd.2023.104038

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., & Zhang, L. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, *241*, 111716. https://doi.org/10.1016/j.rse.2020.111716

Zhang, M., Quanxin, S. U. N., Jinchuan, C. H. E. N., & Jifu, G. U. O. (2008). Travel behavior analysis of the females in Beijing. *Journal of Transportation Systems Engineering and Information Technology*, *8*(2), 19–26. https://doi.org/10.1016/S1570-6672(08)60015-X

Zhang, X., & Du, S. (2016). Learning selfhood scales for urban land cover mapping with very-high-resolution satellite images. *Remote Sensing of Environment*, *178*, 172–190. https://doi.org/10.1016/j.rse.2016.03.015

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* (pp. 5122–5130). IEEE. https://doi.org/10.1109/CVPR.2017.544

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, *127*(3), 302–321. https://doi.org/10.1007/s11263-018-1140-0

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