Cities 158 (2025) 105703

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

Do You Know Your Neighborhood? Integrating Street View Images and Multi-task Learning for Fine-Grained Multi-Class Neighborhood Wealthiness Perception Prediction

Yang Qiu [a], Meiliu Wu [b], Qunying Huang [a,*], Yuhao Kang [c]

[a] *Spatial Computing and Data Mining Lab, University of Wisconsin-Madison, United States of America*
[b] *School of Geographical & Earth Sciences, University of Glasgow, Glasgow, United Kingdom*
[c] *Department of Geography and the Environment, The University of Texas at Austin, United States of America*

ARTICLE INFO

Keywords:
Wealthiness prediction
Social sensing
Urban sensing
Deep learning
Transformer
Urban perception
Human perception

ABSTRACT

The assessment of urban wealthiness is fundamental to effective urban planning and development. However, conventional methodologies often rely on aggregated datasets, such as census data, with a coarse-grained resolution at the census tract level, impeding accurate evaluation of wealthiness in individual neighborhoods and failing to capture spatial heterogeneity. This study proposes a novel approach to predict urban wealthiness at a point-scale spatial resolution by utilizing geo-tagged street view images as input for deep learning models, thereby simulating human perception of urban built environments. Using the Place Pulse 2.0 dataset, which contains over 1.2 million pairwise comparisons of 110,988 street view images from 56 cities worldwide for different urban environment perception factors (e.g., safety and wealthiness), we developed deep learning models based on the Swin Transformer and Multi-gate Mixture-of-Experts (MMOE), a multi-task learning architecture. These models extract and integrate visual features of surrounding elements, including buildings, parks, and vehicles, to classify the wealthiness of specific geo-locations into three categories: Impoverished, Middle, and Affluent. To enhance model training and ground truth data, we modified and enhanced the TrueSkill Rating System, used for scoring neighborhoods via pairwise street view image comparisons, by considering temporal decay and spatial autocorrelation factors. These modifications improved the normality of wealthiness score distribution, reducing the standard deviation from 5.385 to 4.302 and skewness from −0.055 to −0.024. Consequently, model performance improved consistently, with accuracy increases observed in Swin Transformer (63 % to 68 %), ViT (54 % to 58 %), and ResNet50 (51 % to 56 %). In addition, proposed MMOE model demonstrates a significant improvement in the differentiation and classification of wealth categories within a three-class classification system (Impoverished, Middle, Affluent). It achieves an overall accuracy of 82 %, outperforming baseline models, Swin Transformer, ViT, and ResNet50, by 14 %, 24 %, and 26 % respectively. Additionally, we compared our model's predictions with average household income data at the census block group level to elucidate its strengths and limitations. Experimental results demonstrated the efficacy of using geo-tagged street view images for predicting urban wealthiness across diverse geographic and environmental contexts. Our findings also highlight the importance of integrating both quantitative and qualitative evaluations in the prediction of urban environmental factors. By synthesizing human perceptions with advanced deep learning techniques, our approach offers a nuanced understanding of urban wealthiness, providing valuable insights for urban planning and development strategies.

1. Introduction

Wealthiness, often correlated with other socioeconomic factors, is an important indicator in determining the quality of life in urban areas. It also reflects the impact of the urban area on the broader environments, including education (Fang, 2018), public health (Foundation, R.W.J, 2018), and social cohesion (Kawachi & Kennedy, 1997). As such, understanding the wealthiness of an urban area can provide valuable

* Corresponding author.
E-mail addresses: qiu67@wisc.edu (Y. Qiu), Meiliu.Wu@glasgow.ac.uk (M. Wu), qhuang46@wisc.edu (Q. Huang), yuhao.kang@austin.utexas.edu (Y. Kang).

https://doi.org/10.1016/j.cities.2025.105703
Received 6 September 2023; Received in revised form 3 December 2024; Accepted 31 December 2024
Available online 17 January 2025

insights for urban planning and development. Recent studies have shown that wealth inequality has become increasingly concentrated in specific regions, with spatial wealth disparities growing nearly twice as much as income disparities since 1970 (Suss et al., 2024).

Recent advancements in urban studies, especially in computing and machine learning, have transformed our understanding of urban dynamics in several ways. These technologies, integrating data from remote sensing images with spatial big data from geotagged search engines and social media platforms, have greatly enhanced the analysis and interpretation of urban environments, offering a new perspective for sustainable urban planning and development (Yu & Fang, 2023).

However, measuring the wealthiness of urban areas is a complex task requiring a solid understanding of multiple factors (e.g., household income and house price) that can contribute to the livability and sustainability of cities. The existing approaches for measuring wealthiness often encounter various constraints. First, most studies used spatially aggregated datasets, such as census data (MacDonald, 2014), to provide a general indication of the wealthiness (e.g., median household income) of a land parcel. Such a method is limited by a coarse-grained resolution (e.g., at a census block level) due to spatial aggregation, and thus is not able to accurately reflect the wealthiness of individual neighborhoods or communities within the same census unit, in which the spatial heterogeneity cannot be represented. This limitation has become more pronounced following the COVID-19 pandemic, which has exacerbated existing wealth inequalities and created new patterns of spatial economic distribution (Ferreira, 2021; Raphael & Schneider, 2023). While models based on aggregated data offer valuable insights, they often lack the resolution to capture nuanced variations within urban landscapes effectively. Enhancing these models is crucial for more precise and localized urban analysis. Additionally, the use of aggregated spatial units inherently triggers the well-known modifiable area unit problem (MAUP) (Nelson & Brewer, 2015).

Recent methods have started incorporating diverse data sources such as social media and remote sensing images for wealth prediction to overcome these limitations Taubenböck et al., 2018). Second, alternative methods have been employed using novel big data datasets, such as social media data (Blumenstock et al., 2015; Fatehkia et al., 2020; Indaco, 2020), restaurant data (Dong et al., 2019) or remote sensing images (Lin et al., 2021; Yeh et al., 2020), to predict wealthiness. These methods, while innovative, often lack the direct visual context of urban environments, crucial for a holistic understanding of urban wealthiness. Correspondingly, these approaches may not fully incorporate the geographic and environmental factors or visual features that are essential for accurately representing and learning wealthiness, leading to lower performance in their prediction models. This shortfall presents a compelling case for developing methodologies that more accurately mirror the complex reality of urban areas.

Factors such as the visual appearance of properties or the built environmental perceptions from residents and workers in the area, can significantly contribute to the prediction of the wealthiness of a neighborhood. Research indicates that leveraging visual cues from urban imagery, such as street-level photographs, can yield insights that complement traditional socioeconomic data (Suel et al., 2023b). The importance of such factors is underscored by their potential application in urban planning, real estate market analysis, and social policy formulation. In fact, the visual appearance of properties serves as a tangible representation of their quality, aesthetics, and overall value. This visual aspect is not only crucial for understanding current urban landscapes but also for predicting future trends and developments in urban areas. High-end neighborhoods often exhibit visually appealing properties with well-maintained landscapes, architectural elegance, and upscale amenities (Salesses et al., 2013). Conversely, lower-income areas may have properties with visible signs of neglect, limited maintenance, and fewer aesthetic enhancements (Sampson & Raudenbush, 2004). Thus, enhancing existing models to integrate these visual cues becomes crucial for a more nuanced and actionable understanding of urban wealthiness. The visual appearance of properties and perceptions of the built environment also play a key role in revealing the social dynamics and neighborhood characteristics that influence wealthiness patterns. The influence of upscale retail establishments, well-designed public spaces, or exclusive amenities on perceptions of affluence and property values is substantiated by research (Qin et al., 2019). By incorporating these aspects, wealthiness prediction models can more accurately reflect the broader socioeconomic context and its impact on urban wealthiness distribution.

Despite recent advancements (Blumenstock et al., 2015; Fatehkia et al., 2020; Indaco, 2020), a gap remains in capturing the granular, human-scale perspective essential for a comprehensive understanding of urban wealthiness. In response, this paper proposes a novel approach to predict the wealthiness of urban environments at the most fine-grained level (i.e., the specific locations of images), by using geo-tagged street view images as the input for deep learning models. Under this context, street view imagery offers unique advantages by providing a more direct and tangible representation of the urban environment that captures details like property aesthetics, maintenance, and public amenities that are pivotal in assessing wealthiness (Biljecki & Ito, 2021a). In particular, this modeling process can be considered as the simulation of human perception within urban built environments (Piga & Morello, 2015), aiming to offer a more nuanced and sophisticated understanding of the street-view scene by incorporating human perception into the measurement process. To construct model training samples and ground truth data, we utilized the Place Pulse 2.0 dataset (Salesses & Hidalgo, 2020), which provides geo-tagged street view images and comparative wealthiness votes. Traditionally, the TrueSkill Rating System is used to rank players in competitive games based on win-loss records, where each player's skill score is updated after every match. While TrueSkill effectively handles uncertainty in rankings through a Bayesian framework, it has notable limitations when applied to wealthiness prediction in neighborhoods. Specifically, it does not account for temporal changes in urban environments, where wealthiness dynamics can shift over time, nor does it consider spatial relationships between nearby neighborhoods, which can significantly influence wealthiness distribution. To enhance model performance, we modified the TrueSkill Rating System by incorporating temporal decay, which reduces the weight of older comparison pairs to account for rapid urban development; and spatial autocorrelation, which lessens the influence of comparisons between nearby locations to prevent inaccuracies when voters may fail to distinguish minor differences in wealthiness. This modification improved the normality of wealthiness score distribution and led to consistent performance improvements across different models.

Furthermore, we incorporated a Multi-gate Mixture-of-Experts (MMOE) model into our classification framework that categorizes and predicts the wealthiness of specific geo-locations into three categories: Impoverished, Middle, and Affluent. It is a machine learning architecture used primarily in multi-task learning by dynamically selecting expert networks for each task. The proposed MMOE model utilizes five expert networks and three task-specific gate networks to dynamically select relevant features for each wealthiness category, thereby enhancing classification performance and achieving more balanced and improved accuracy across all three wealthiness categories. The proposed model was compared with widely-used vision models, including Swin Transformer (Liu et al., 2021), ViT (Dosovitskiy et al., 2020) and ResNet52 (He et al., 2016). Using the modified TrueSkill scores as input data, MMOE significantly outperformed other models, achieving substantial improvements in overall accuracy. To further evaluate the generalizability of the proposed model, it was then trained using images from New York City (NYC), and subsequently tested on images from Boston and Los Angeles (LA). Furthermore, we explored the correlation between our point-based prediction results and the aggregated median household income by census block group to evaluate and analyze their matching level. Finally, we performed a qualitative evaluation of our prediction results based on visual interpretation of the images as well as a quantitative examination by computing the average predicted wealthiness values within a certain radius for different land use types. This is to verify whether our model truly understands the built environment of images in a human-like manner and predicts accurate wealthiness scores.

To sum up, our main contributions include:

- First, we introduce a novel approach to assess neighborhood wealthiness in urban environments by leveraging deep learning techniques that simulate human perception from street view imagery. The proposed deep learning-based model can effectively incorporate visual features derived from street view imagery with strong generalizability, ultimately contributing to the improved measurement and understanding of wealthiness distribution in neighborhoods.
- Second, we innovatively modified the TrueSkill Rating System, adapting it to consider and incorporate temporal decay and spatial autocorrelation factors. These modifications allow the model to generate more accurate wealthiness scores by considering both time-related changes and geographic variations. This innovation is crucial for ensuring that the wealthiness predictions better reflect the dynamic nature of urban environments, capturing how wealthiness distribution evolves over time and across different locations. Our evaluation demonstrates that the modified TrueSkill system improved the normality of wealthiness score distribution, reducing the standard deviation from 5.385 to 4.302 and the skewness from −0.055 to −0.024. This enhancement led to a 5 % increase in overall accuracy for the Swin Transformer model, from 63 % to 68 %. ViT and ResNet50 also benefited, with increases of 4 and 5 percentage points respectively.
- Third, we implemented and adapted the MMOE model, which enhances the model's ability to differentiate between wealthiness categories in a three-class (i.e., Impoverished, Middle, and Affluent) wealthiness classification system. This innovation improves classification performance by allowing the model to dynamically allocate experts to different wealthiness categories based on the complexity of the input data. When compared with other models using the Modified TrueSkill system, the MMOE model demonstrated significant improvements. It achieved an overall accuracy of 82 %, representing a 14 % increase over the Swin Transformer (68 %), a 24 % increase over ViT (58 %), and a 26 % increase over ResNet50 (56 %).
- Fourth, our methodology achieves the most fine-grained level of wealthiness measurement, specifically at the point-scale spatial resolution. This approach contrasts with existing works and methods, such as census block analyses, enabling us to capture the spatial dependency and heterogeneity of wealthiness distribution within neighborhoods more accurately. This fine granularity helps overcome the limitations of the MAUP and reveals the nuanced spatial patterns of wealthiness that larger-scale analyses might miss.
- Finally, source codes are publicly available at https://github.com/scdmlab/Wealthiness_Prediction_Cities, contributing to the open source community.

2. Related work

This literature review focuses on two aspects: (1) the application of human perception and computer vision technology to understanding urban environments; and (2) state-of-the-art methods for measuring urban wealthiness and other indicators, along with their limitations.

2.1. Human perception and computer vision in urban environments

Human perception is critical in understanding and interpreting urban environments. It involves the processing of sensory information, such as the visual appearance of a location, which is influenced by various factors including an individual's cognitive and emotional state, past experiences, and cultural background. In recent years, computer vision has been applied to mimic human perception in capturing and analyzing urban environments, particularly in the context of understanding urban characteristics at a large scale.

Deep learning models have been trained on street view images to estimate perceived urban attributes, such as safety, aesthetics, walkability, and wealthiness (Biljecki & Ito, 2021b). These models can reveal hidden patterns in the visual characteristics of locations that may be challenging for human observers to detect, such as subtle variations in architectural style or the distribution of specific visual elements within a scene (Zhou et al., 2018). Computer vision techniques have also been used to assess the age-friendliness of urban environments using Google Street View images, achieving an accuracy of 85% in predictions (Moradi et al., 2023).

Recent studies have significantly expanded the application of computer vision in urban studies. (Wu et al., 2023) leveraged street view imagery to explore the connection between human perceptions and urban vitality in Shenzhen, China, demonstrating how visual data can be quantitatively analyzed to assess urban vitality. Street view imagery has been utilized to analyze the spatio-temporal evolution of urban visual environments in Singapore, providing insights into changes in urban aesthetics and safety perceptions over time (Liang et al., 2023). Additionally, human perceptions were extracted from street view images to better assess urban improvement opportunities, illustrating the application of computer vision in gauging urban renewal potentials (He et al., 2023). Furthermore, differences in safety perceptions across neighborhoods in Stockholm, Sweden, were assessed using a GeoAI approach and survey data, highlighting the effectiveness of combining GeoAI with traditional survey methods for a comprehensive understanding of urban safety perceptions (Kang et al., 2023). Collectively, these studies highlight the important role of street view imagery, analyzed through advanced deep learning models, in enhancing our understanding of various aspects of urban environments. By leveraging the power of deep learning, researchers can obtain a more nuanced and sophisticated prediction of the factors that shape human perception in urban environments, which in turn can be valuable for informing urban planning and development decisions (Dubey et al., 2016).

2.2. Urban wealthiness measurement

Traditional methods of measuring urban wealthiness rely on spatially aggregated socioeconomic statistics, such as median income (Gyourko et al., 2013) or poverty rates (Chakravorty, 1996a). However, these approaches may not fully capture the complexity and diversity of urban environments, especially when applied to urban regions with different built environments and landscapes (Batty, 2021). For instance, Abitbol and Karsai (Abitbol & Karsai, 2020) highlighted that accurately tracking urbanization and the corresponding socioeconomic changes has been challenging for traditional data collection methods. Despite even using the neural networks fed with satellite images to recover the socioeconomic information of an area, these models still lacked the ability to explain how visual features within a sample triggered specific predictions. Consequently, they often fail to accurately capture the wealthiness of individual neighborhoods or communities within the same census unit, where substantial spatial heterogeneity exists. Gao et al. (2020) emphasize that traditional models inadequately capture the diverse and non-uniform nature of urban dynamics, showing the importance of spatial heterogeneity in urban simulations. Spatial heterogeneity refers to the variations in wealthiness distribution that occur within localized areas, where distinct economic characteristics can exist even between neighboring streets.

Traditional models, which often rely on aggregated data such as median income at the census block group (CBG) level, fail to capture these finer differences in wealthiness distribution (Chakravorty, 1996b). This aggregation can mask important local variations in socioeconomic conditions, leading to inaccurate predictions or a lack of detail in wealthiness assessments. This lack of granularity not only affects the precision of wealthiness estimates but also limits our understanding of how economic and social resources are distributed across urban areas. Recognizing these patterns is crucial, as they often result in disparities in access to infrastructure, public services, and opportunities between wealthy and less affluent neighborhoods (Roy et al., 2023; Yang & Hu, 2022). Additionally, the use of aggregated spatial units introduces the well-documented MAUP challenge (Nelson & Brewer, 2015). By incorporating spatial heterogeneity, our approach captures wealthiness at a more detailed, point-scale level and highlights the spatial dynamics contributing to inequality. This fine-grained model, leveraging street view imagery and human perception, offers a more nuanced and sophisticated understanding of wealthiness distribution. It also provides urban planners and policymakers with the insights needed to develop tailored interventions that effectively address localized socioeconomic challenges and promote more equitable urban development (Nicoletti et al., 2022).

In fact, recent research has shown the potential of using street view imagery as a data source to complement traditional socioeconomic indicators for predicting urban environmental factors (Glaeser et al., 2016). For instance, Naik et al. (2014) developed a model called Streetscore based on Google Street View images to predict the perceived safety of a location. The model was trained on a dataset of images collected from New York City and Boston, and its generalizability was tested on 27 cities across the United States. In particular, the utilization of deep learning techniques enables the models to uncover the hidden patterns of the imagery data that are beyond human perception (Fintz et al., 2022; Kononenko & Kukar, 2007). Such learning outcomes can add more value to urban planning and development, bringing new opportunities to improve quantitative research in the measurement of urban environments. For example, a data-driven deep learning model is proposed to estimate six human perceptual indicators (i.e., safe, lively, beautiful, wealthy, depressing, and boring) in urban regions (Zhang et al., 2018). The proposed model was trained by millions of human ratings on street-level images, and can predict these indicators with a high accuracy rate of about 80%. The authors suggested that this model could be used to analyze the distribution of city-wide human perception for various urban regions. Additionally, the authors conducted a series of statistical analyses to identify visual elements that are positively or negatively correlated with each of the six perceptual indicators, shedding light on how different visual features contribute to people's perceptions of a place.

While substantial progress has been made using street view imagery to examine urban factors, key limitations persist. Prior models, such as convolutional neural network (CNN) and ResNet architectures (Zhang et al., 2018), can not effectively integrate human perception, limiting their ability to understand and predict wealthiness in a way that aligns with human perspectives. Furthermore, these works often estimate a wide range of perceptual indicators (e.g., safety, livability) with wealthiness being only one of the indicators to examine urban environments, offering a broader but less focused discussion and insights for urban wealthiness. This work, by contrast, concentrates on the wealthiness aspect using the MMOE architecture, with Swin Transformer serving as the feature extractor, providing a deeper, more nuanced understanding and interpretation of urban wealthiness, thus filling a significant gap in the specialized analysis of this particular dimension. Finally, prior approaches could only classify the perceptual indicator of an image into a binary class, e.g., wealthy or non-wealthy. However, the work provides a solution for multi-class categorization of urban wealthiness.

3. Methodology

This section elaborates on the workflow for classifying street view images for wealthiness prediction (Fig. 1). First, the workflow begins with data collection, gathering images and scores from the Place Pulse 2.0 dataset (Salesses & Hidalgo, 2020). Next, data pre-processing techniques, including the Contrast Limited Adaptive Histogram Equalization (CLAHE) (Pizer et al., 1990), are applied to the street view images to enhance image visibility, before feeding them into the proposed prediction model. In particular, both the original TrueSkill Rating System (Microsoft, 2005) for assigning wealthiness scores and our modified version of TrueSkill Rating System will be introduced, alongside the MMOE model design.

3.1. Data collection and transformation

Approximately 110, 000 street images were collected from the public dataset Place Pulse 2.0, each geo-tagged with precise coordinates. Volunteers provided wealthiness comparison results for pairs of images, allowing us to collect ordinal data reflecting public perceptions of urban wealthiness. These comparisons, though subjective and qualitative, were transformed into quantifiable wealthiness scores using a modified version of the TrueSkill Rating System.

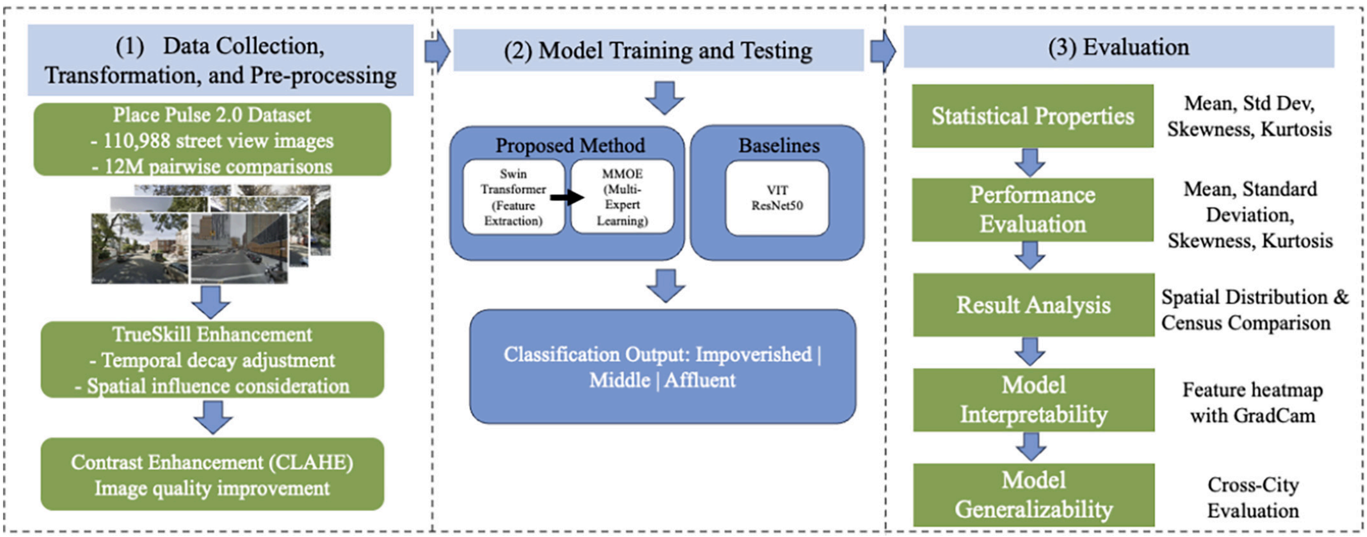

(1) Data Collection, Transformation, and Pre-processing
Place Pulse 2.0 Dataset
- 110,988 street view images
- 12M pairwise comparisons
TrueSkill Enhancement
- Temporal decay adjustment
- Spatial influence consideration
Contrast Enhancement (CLAHE)
Image quality improvement
(2) Model Training and Testing
Proposed Method
Swin Transformer (Feature Extraction)
MMOE (Multi-Expert Learning)
Baselines
VIT
ResNet50
Classification Output: Impoverished | Middle | Affluent
(3) Evaluation
Statistical Properties
Mean, Std Dev, Skewness, Kurtosis
Performance Evaluation
Mean, Standard Deviation, Skewness, Kurtosis
Result Analysis
Spatial Distribution & Census Comparison
Model Interpretability
Feature heatmap with GradCam
Model Generalizability
Cross-City Evaluation

Fig. 1. The workflow of classifying and evaluating street view images for wealthiness prediction.

3.1.1. *Spatial-temporally enhanced TrueSkill system for image wealthiness rating*

The original TrueSkill system, developed for competitive gaming, computes rankings based on pairwise comparisons without considering the time of the comparisons made by voters, as well as the location of images being compared. However, the dynamic and geographically dependent nature of urban environments makes the standard TrueSkill unsuitable for this study. Over time, neighborhoods evolve due to socioeconomic changes, and nearby areas tend to share similar characteristics due to spatial autocorrelation. Ignoring these factors can lead to outdated or geographically inconsistent wealthiness scores. To address the limitations of the standard TrueSkill algorithm, we consider both temporal decay and spatial autocorrelation into the algorithm enhancement.

3.1.1.1. Spatial influence. In urban environments, nearby locations often share similar socioeconomic characteristics, and the perceived wealthiness differences between neighboring areas should not be substantial if the images represent points within close proximity (e.g., within 1 km). Large discrepancies in wealthiness scores for such close locations might result from subjective judgments by voters or from non-representative photos, such as poorly captured images. The standard TrueSkill system does not account for geographic proximity and only updates scores based on the voting results. If two images represent geographically close locations but have significantly different scores, this discrepancy is likely unreasonable and may reflect inconsistencies in the voting process rather than actual differences in wealthiness.

Thus, introducing a spatial influence factor is necessary to prevent wealthiness scores for nearby locations from diverging excessively. To address this, we reduce the impact of votes between images of geographically close locations by weakening the influence of their comparisons, which prevents overfitting based on subtle, often non-meaningful differences. Conversely, for locations that are farther apart, where real socioeconomic differences are more likely to exist, the system retains and trusts the voting results more.

The goal here is to leverage street-level images to make more precise predictions about the wealthiness of a specific location, without solely relying on traditional Census Block Group (CBG)-based data. At the same time, we aim to consider the socioeconomic homogeneity often expected in urban environments. This approach allows us to capture more detailed and dynamic variations in urban areas, while still accounting for the general consistency of nearby locations.

The choice of a 1 km distance threshold is based on several findings in spatial autocorrelation studies. In urban settings, spatial autocorrelation—the tendency for nearby locations to have similar characteristics—often stabilizes around 1 km. For instance, Cai and Wang (2006) found that spatial autocorrelation in topographic indices tends to stabilize beyond 1 km, indicating that areas within this distance are more likely to share socioeconomic similarities. Additionally, Lin et al. (2020) used a 1 km grid in their study on urban land use changes, and Hansen et al. (2000) demonstrated that a 1 km spatial resolution is effective for global land cover classification, which highlights the utility of this distance in capturing local variations.

We define the spatial influence factor as $f(d) = 1 - e^{-3d}$. This function increases the spatial influence factor as the distance d (the distance between two points) increases, ensuring that comparisons between distant locations (e.g., beyond 1 km) have a stronger influence on updating their scores. For example, when $d = 1, f(1) = 0.95$, meaning that spatial influence is substantial, and the system allows for more variation in scores between these points. However, when $d < 1$ km, $f(d)$ decreases rapidly. This rapid decrease ensures that nearby locations, which are more likely to share similar socioeconomic characteristics, have less variation in their scores, as the differences might not be genuine or substantial. This approach helps maintain spatial consistency by minimizing unrealistic variations among proximate locations while trusting and incorporating more reliable differences from distant comparisons.

The choice of $k = 3$ balances the rapid increase in spatial influence when d is small and allows for sufficient divergence when $d > 1$. A smaller k would lead to slower growth, reducing the distinction between proximate and distant locations, while a larger k would cause too abrupt of a change, potentially neglecting genuine differences at slightly greater distances. In this setting, $k = 3$ was experimentally chosen to ensure a balance between nearby convergence and allowing distant comparisons to meaningfully update scores. This ensures that locations closer than 1 km have quickly diminishing differences in their scores, which aligns with the expected socioeconomic homogeneity in urban settings. Meanwhile, distances beyond 1 km see more substantial differences, reflecting real-world spatial-economic variations.

3.1.1.2. Temporal decay. Urban environments change over time, and a wealthiness comparison made several years ago might no longer be valid. Based on the dataset, the time span between the earliest and latest votes is approximately eight years. To address this, we apply a temporal decay factor, $\lambda(t) = e^{-ln(2)\frac{t}{T_{\text{half}}}}$, where t is the time difference in years, and $T_{\text{half}} = 4$ years. This ensures that older comparisons have diminishing influence on the current wealthiness score.

The updated TrueSkill equations incorporate both temporal and spatial factors to better reflect the fluidity of urban environments. The following equations are used, with modifications highlighted in red bold:

$$\mu_{\text{winner}} \leftarrow \mu_{\text{winner}} + \boldsymbol{\lambda}(t) \cdot \frac{\sigma^2_{\text{winner}}}{c} \cdot v\left(\frac{\mu_{\text{winner}} - \mu_{\text{loser}}}{c}, \frac{\epsilon}{c}\right) \cdot f(d)$$

$$\mu_{\text{loser}} \leftarrow \mu_{\text{loser}} - \boldsymbol{\lambda}(t) \cdot \frac{\sigma^2_{\text{loser}}}{c} \cdot v\left(\frac{\mu_{\text{winner}} - \mu_{\text{loser}}}{c}, \frac{\epsilon}{c}\right) \cdot f(d)$$

$$\sigma^2_{\text{winner}} \leftarrow \sigma^2_{\text{winner}} \cdot \left[1 - \boldsymbol{\lambda}(t) \cdot \frac{\sigma^2_{\text{winner}}}{c^2} \cdot w\left(\frac{\mu_{\text{winner}} - \mu_{\text{loser}}}{c}, \frac{\epsilon}{c}\right) \cdot f(d)\right]$$

$$\sigma^2_{\text{loser}} \leftarrow \sigma^2_{\text{loser}} \cdot \left[1 - \boldsymbol{\lambda}(t) \cdot \frac{\sigma^2_{\text{loser}}}{c^2} \cdot w\left(\frac{\mu_{\text{winner}} - \mu_{\text{loser}}}{c}, \frac{\epsilon}{c}\right) \cdot f(d)\right]$$

$$c^2 = 2\beta^2 + \sigma^2_{\text{winner}} + \sigma^2_{\text{loser}}$$

Here, β^2 captures the inherent randomness in perceptions, and ϵ is a tunable parameter representing the draw margin, accounting for the possibility of ties in the comparisons. Additionally, the Temporal Decay Factor, $\boldsymbol{\lambda}(t) = e^{-ln(2)\frac{t}{T_{\text{half}}}}$, where t is the time difference in years and $T_{\text{half}} = 4$ years, reduces the influence of past performances as time progresses. The Partial Influence Factor, $f(d) = 1 - e^{-3d}$, increases as the distance d (the distance between two points) increases, measuring the impact of spatial distance on the score updates.

3.1.2. *Wealthiness classification*

Once the wealthiness scores are computed, we classify the images into three categories—impoverished, middle, and affluent—based on the mean wealthiness score μ and variance σ. The classification function is defined as follows:

$$f\left(x\right) = \begin{cases} \text{impoverished} & \text{if score} \leq \mu - s \cdot \sigma, \\ \text{affluent} & \text{if score} \geq \mu + s \cdot \sigma, \\ \text{middle} & \text{otherwise.} \end{cases}$$

Here, $s \in [0, 1]$ is an adjustable factor that controls the interval between the categories. For example, setting $s = 0$ simplifies the classification into two categories—impoverished and affluent—while higher values of s create a larger middle category. The sensitivity of s will be evaluated in Section 5.1 Performance Evaluation.

3.2. *Data pre-processing*

To extract useful information from street view images, the importance of contrast between depicted objects cannot be overstated. For example, the contrast between the bright sky and the dark shadows cast by buildings can vary significantly (Menzel & Reese, 2021). To enhance the visibility of such details, contrast-adjusting algorithms prove to be effective. In addition, images captured by vehicle-mounted cameras are often susceptible to issues like overexposure and underexposure. Moreover, the inherently varied and dynamic composition of street view images, encompassing a broad spectrum of elements like fluctuating lighting conditions, prominent shadows, and unpredictable occlusions, often culminates in a pronounced inconsistency of contrast. This irregularity is not merely a superficial concern; it fundamentally hampers the interpretability of these images. These disparities pose significant challenges in maintaining consistent contrast across the entire image, which is crucial for accurate analysis and application in urban wealthiness prediction.

To rectify these issues, contrast-adjusting algorithms can be leveraged to equalize the contrast across the entire image. Such a uniform enhancement of contrast contributes to an overall quality improvement of the image, enabling more precise predictions from the model. An optimized contrast ensures that vital elements within an image, such as street signs, pedestrians, or vehicle details, are accurately captured and discernible, thereby equipping our model with comprehensive and precise information for reliable predictions (Shokrollahi et al., 2017).

To effectively adjust the contrast of the street view images, this study used the CLAHE method (Pizer et al., 1990), a variation of Adaptive Histogram Equalization (AHE) (Wikipedia contributors, 2022) that was first introduced by Karel Zuiderveld in 1994 (More et al., 2015; Zuiderveld, 1994). The main idea behind CLAHE is to enhance the local contrast of an image by dividing it into small, non-overlapping regions called tiles, and applying AHE to each tile independently. This approach overcomes the limitations of global histogram equalization, which may amplify noise and create false boundaries. As a result, it has showcased robust and advantageous functionality across diverse applications (Kim et al., 2016; Laksmi et al., 2016; Sundaram et al., 2011).

The CLAHE algorithm is defined by two parameters: the block size (BS) and the clip limit (CL). The block size determines the size of the tiles, while the clip limit controls contrast enhancement by restricting histogram amplification. A larger block size results in a higher dynamic range and increased image contrast, while a larger clip limit makes the image brighter. This study used a block size of 8 × 8 and a clip limit of 2, which produced subjectively good image quality based on image entropy (Fig. 2).

3.3. *Model architecture - from single-task models to multi-task learning*

3.3.1. *Single-task attention-based models for wealthiness prediction*

With a focus on wealthiness categorization, our task is framed as an image classification problem. Previous studies (e.g., (Naik et al., 2014)) mostly utilized traditional computer vision techniques, such as Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT), followed by classification with a machine learning model (e.g., Support Vector Machine (SVM)). HOG (Tomasi, 2012) is a feature descriptor capturing the local shape of an image, while SIFT (Lowe, 2004) detects and describes local features within an image. Although these techniques have found widespread use in computer vision applications, they struggle to handle large, complex datasets due to the lack of a self-attention mechanism.

Self-attention mechanisms empower a model to comprehend the relationships between all input elements simultaneously (Niu et al., 2021a), which stands in contrast to the sequential or hierarchical processing typically employed by CNNs. This capability provides two primary advantages over conventional convolutional approaches: the ability to capture long-range dependencies and the adaptability to varying spatial resolutions. Long-range dependencies allow a model to recognize patterns and relationships across distant elements in the input by enabling models to determine the importance of different parts of the input relative to each other (Vaswani et al., 2017). As a result, models with self-attention mechanisms like the Swin Transformer can effectively capture and process these long-range dependencies, which are often a challenge for traditional computer vision methods like CNNs (Niu et al., 2021b). In addition, self-attention mechanisms can adapt to different spatial resolutions, thereby eliminating the need for specific

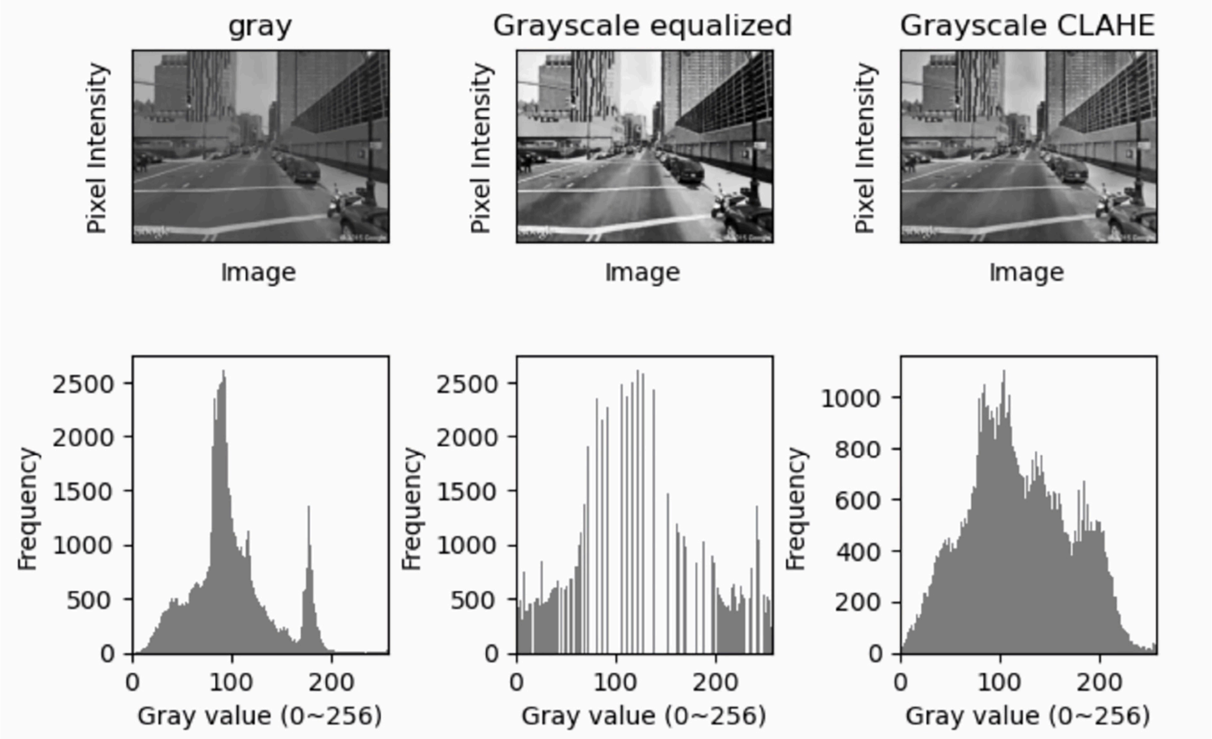


Fig. 2. Comparing CLAHE and histogram equalization - Mastering OpenCV with Python.

image preprocessing or alterations to the model's structure. This feature makes them more scalable and adaptable compared to CNN-based models, which often demand specialized architectures or adjustments to handle varying input resolutions (Ramachandran et al., 2019). Consequently, self-attention-based models can be deployed for a wide array of tasks and datasets, enabling their application in a diverse range of fields (Guo et al., 2022).

To enhance the capabilities of our image classification model for wealthiness prediction, we capitalized on the advancements in deep learning by selecting a model based on the Transformer architecture. This technology, initially developed for time-series data, has been adeptly adapted for vision tasks through careful design modifications in preprocessing and structuring of the input images, as described by (Dosovitskiy et al., 2020). However, the pioneering Vision Transformer (ViT) model, while innovative in its application of self-attention mechanisms traditionally used in natural language processing, encountered challenges, notably in processing images of varying scales which often leads to suboptimal performance with high-resolution images, and maintaining computational efficiency, as detailed by (Touvron et al., 2012). Additionally, due to its design of applying self-attention over the entire image, the model demanded high computational resources, especially for larger images.

In response to the challenges posed by the ViT in processing images of varying scales and its high computational demands, we opted for the **Swin Transformer** (large-sized model with a large number of hyperparameters) (Liu et al., 2021) for our classification task. This model, pre-trained on the extensive ImageNet-21 k dataset containing approximately 14 million images across 21,841 categories and optimized for a resolution of 384 × 384, introduces a ground-breaking approach to image classification. By employing shifted windows for localized computation, the Swin Transformer facilitates a hierarchical representation of data. This innovation significantly enhances the model's processing efficiency, making it exceptionally capable of handling the complex and diverse visual information essential for accurate urban wealthiness prediction. Moreover, the Swin Transformer retains the core advantages of the standard Transformer architecture, such as capturing long-range dependencies and adaptability to various spatial resolutions. Its hierarchical design also specifically addresses the limitations of the ViT, enabling more efficient processing of different image scales and reducing computational demands. This makes the Swin Transformer particularly effective for high-resolution images, overcoming some of the key challenges encountered with the original ViT model.

3.3.2. *Multi-task learning for enhanced multi-class classification*

We anticipated that the "middle" class would be challenging to predict due to overlapping features with both the "impoverished" and "affluent" categories. This overlapping nature often results in misclassifications and ambiguous decision boundaries, reducing the model's overall effectiveness. To address this issue, we applied the Multi-gate Mixture-of-Experts (MMOE) model, originally proposed by Ma et al. (2018), into our classification framework. MMOE is a multi-task learning architecture that incorporates several expert networks

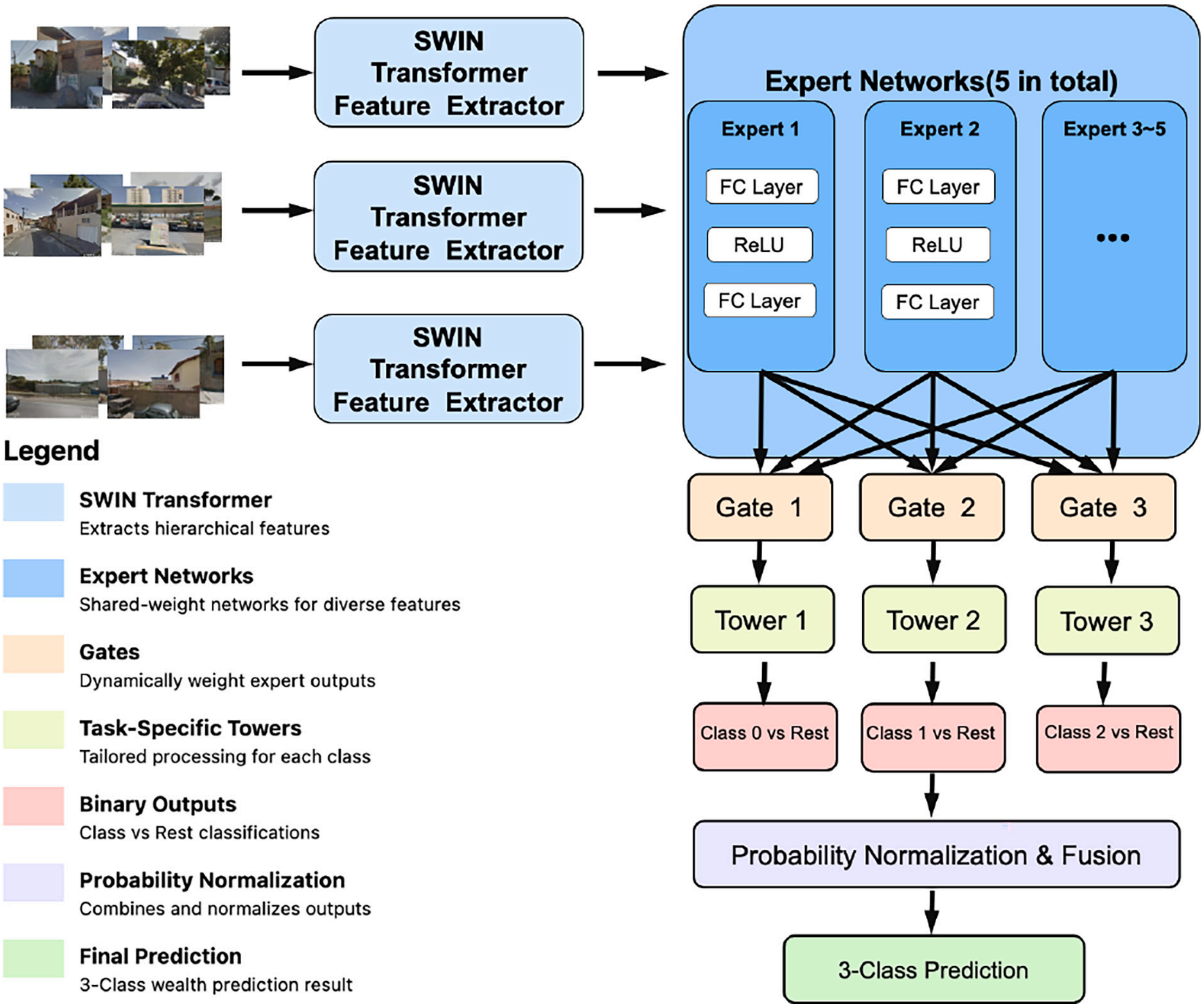

SWIN Transformer Feature Extractor
SWIN Transformer Feature Extractor
SWIN Transformer Feature Extractor
Expert Networks(5 in total)
Expert 1
FC Layer
ReLU
FC Layer
Expert 2
FC Layer
ReLU
FC Layer
Expert 3~5
•••
Gate 1
Gate 2
Gate 3
Tower 1
Tower 2
Tower 3
Class 0 vs Rest
Class 1 vs Rest
Class 2 vs Rest
Probability Normalization & Fusion
3-Class Prediction
Legend
SWIN Transformer
Extracts hierarchical features
Expert Networks
Shared-weight networks for diverse features
Gates
Dynamically weight expert outputs
Task-Specific Towers
Tailored processing for each class
Binary Outputs
Class vs Rest classifications
Probability Normalization
Combines and normalizes outputs
Final Prediction
3-Class wealth prediction result

Fig. 3. The architecture of the MMOE model, showing expert networks and task-specific gate networks.

and task-specific gate networks, each designed to dynamically select the relevant experts based on the specific task. In our case, the three tasks involve classifying images into 'impoverished,' 'middle,' and 'affluent' categories. This architecture allows the model to benefit from shared knowledge across tasks while enabling each task to specialize and learn distinct features relevant to its classification problem.

The MMOE model in our implementation consists of five expert networks, though for simplicity, our diagram (Fig. 3) shows three with an ellipsis indicating the full set. The expert networks in MMOE are a set of shared neural networks that process the same input but learn to specialize in different aspects of the data, capturing diverse patterns. Each expert network receives the input features extracted from the SWIN Transformer and processes them through two fully connected layers with a ReLU activation in between. These experts provide a rich set of features that are dynamically weighted by the gate networks. Each of the three task-specific gate networks selects relevant outputs from the experts by assigning different weights to each expert based on the input features. This dynamic selection mechanism allows the model to focus on task-specific information, reducing the overlap and confusion between the "middle" class and the other categories.

Following the gate networks, we implement task-specific towers. These towers, one for each task, are implemented as feed-forward neural networks (FFNs). Each tower consists of a fully connected layer that outputs two values, representing the binary classification for each task (e.g., 'impoverished vs. rest', 'middle vs. rest', 'affluent vs. rest'). These FFNs further process the gated expert outputs to produce task-specific binary representations. This additional layer of task-specific processing enhances the model's ability to capture nuanced features relevant to each classification task, allowing for fine-tuned discrimination between the "impoverished," "middle," and "affluent" categories.

In particular, the MMOE model addresses the challenge of class overlap by enabling more granular feature extraction through the expert networks and selecting the most relevant features for each task using the gate networks. The "middle" class often shares characteristics with both the "impoverished" and "affluent" classes, making it difficult for a traditional model to separate these classes clearly. By leveraging multiple experts and dynamically adjusting their contributions, MMOE allows the model to better differentiate between overlapping features, thus improving the classification of the "middle" category.

The fusion of the task outputs in MMOE is critical to ensure its success. Each task produces a binary output indicating the likelihood that the input belongs to its respective class. These outputs are then combined and normalized to form a final three-class prediction. We first apply a sigmoid function to each task's binary output to obtain task-specific probabilities. Then, we employ another softmax normalization over the positive class probabilities of the three tasks to generate a final probability distribution over the three classes. This two-step process ensures that the model produces a valid probability distribution, which is then used to select the final predicted class based on the highest probability. This fusion process allows the model to combine the evidence from each task and weigh the likelihood of each class appropriately, leading to more accurate final predictions.

To ensure consistency during training, we use Binary Cross-Entropy with Logits Loss for each task's binary output. These individual task losses are then combined to form the overall loss used to optimize the model performance. During evaluation, the normalized probabilities were used to compute the accuracy and F1-score based on the final predicted class. This approach allowed us to handle the overlapping features between classes while maintaining accurate and stable classification performance.

The use of MMOE in this experiment provides several key advantages. First, it enables the model to separate overlapping classes more effectively by dynamically selecting relevant features for each task. Second, the expert networks allow for diverse feature extraction, while the gate networks specialize the outputs for each task. Finally, the fusion of outputs from the binary classification tasks into a final multi-class prediction ensures that the model can handle the ambiguity and overlap between the "middle" class and other classes. This architecture proved to be a valuable addition to the classification pipeline, improving the model's ability to distinguish between the impoverished, middle, and affluent categories.

The MMOE model was trained under the same conditions as other models, using the AdamW optimizer with the same learning rate and incorporating data augmentation techniques such as random-perspective transformation. By integrating MMOE into our model, we were able to significantly improve the classification accuracy and address the challenges posed by the overlapping features in our dataset.

3.4. *Loss function*

Addressing the complexity of classifying images into three categories, namely "impoverished", "middle", and "affluent", we interpret the task as a multi-class classification problem. For such problems, the cross-entropy loss function serves as an ideal objective function for optimization.

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} log\left(p_{ic}\right)$$

where N represents the number of images, C denotes the number of classes, $y_{ic} \in 0, 1$ are the one-hot encoded labels, and p_{ic} signifies the probability that the i-th image is predicted to be in class c.

However, to counteract the issue of class imbalance, we incorporated class weights into the loss function, enhancing the training process's effectiveness. Class imbalance can often lead to a biased model that overlooks minority classes. The weighted cross-entropy loss function takes the form:

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} w_c y_{ic} log\left(p_{ic}\right)$$

where w_c denotes the weight assigned to class c, inversely proportional to the class frequency. By adjusting these weights, we can ensure a balanced contribution of each class to the total loss, leading to a more robust and reliable model.

4. Experiment design

This study first performed a classification experiment involving three categories, including impoverished, middle, and affluent. However, the three-class experiment results revealed substantial challenges in differentiating the "middle" class from both the "impoverished" and "affluent" categories due to overlapping features present in the images. To further evaluate the effectiveness of the proposed model, we then excluded the "middle" class, transforming the task into a binary classification problem by predicting each image as an "impoverished" or "affluent" category only.

4.1. *Experiment setup*

For our study, we utilized gradient accumulation (Hermans et al., 2017) to train our model. This technique enables the accumulation of gradients from multiple mini-batches before applying the weight update. This approach facilitates the use of larger batch-size updates, contributing to improved model convergence and accelerated training. We set the gradient accumulation steps to 3 in our study, which means that gradients from three mini-batches were accumulated before updating the weights. The adoption of gradient accumulation improved our model's performance and reduced memory usage during the training phase.

For model evaluation, we divided our dataset into three sections: 80% for training, and 10% each for both validation and testing. To assess the model's generalizability, we trained it on images from New York, and tested it on datasets from Boston and Los Angeles.

Table 1 provides a detailed split of the dataset, showing the total count of the data, the distribution across impoverished, middle, and affluent categories, and the division of the data into training, validation, and testing sets. It is worth noting that while the New York dataset was used for training, datasets from Boston and LA were exclusively used for testing the model's performance.

To address the issue of class imbalance in the dataset, we implemented the Weighted Random Sampler (WRS) from the PyTorch library (Paszke et al., 2019). The sampler generates indices based on the inverse class frequencies, which are used to draw samples during the training process. This ensures that all classes have the same chance of being selected for a mini-batch, thereby reducing the bias towards the majority class. By using WRS, we could provide a balanced view of our classes despite the initial imbalance in our dataset.

In the context of data augmentation, we conducted a comprehensive analysis of various techniques designed to encompass the heterogeneous visual characteristics observed in urban environments. Among these methods, the random-perspective transformation was found to be particularly effective. This data augmentation technique modifies the perspective from which an image is viewed, thereby introducing variance and simulating how an image might appear when captured from different angles or elevations. Such transformational manipulation effectively enriches our dataset by replicating the complexity and variability inherent in urban landscapes. When applied to our image classification task, the random-perspective transformation demonstrated remarkable performance. It adeptly handled the diverse range of visual features characteristic of urban environments, from towering skyscrapers to labyrinthine roadways. By providing a more nuanced and comprehensive representation of these features, it greatly improved our model's capacity for accurate image recognition and categorization.

We trained for 100 epochs using the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of $5e^{-5}$, using a warm-up mechanism (Goyal et al., 2017) to adjust the learning rate dynamically. We also utilized Grad-CAM (Selvaraju et al., 2016) to visualize which areas of the images were most important for the model's predictions, generating heatmaps from the CNN gradient information.

4.2. Evaluation

The effectiveness of the enhanced TrueSkill system was evaluated through multiple approaches. First, we compared the statistical properties of the wealthiness score distributions before and after the modification, focusing on measures of normality, standard deviation, and skewness. Second, we analyzed Q-Q plots to visually examine the normality of the original and modified TrueSkill scores. Finally, we conducted performance comparisons across different models using both the original and modified TrueSkill scores to assess the impact on model performance. In order to evaluate the effectiveness of our proposed MMOE model, we compared it with Swin Transformer, and two additional benchmark models, including ResNet50 and ViT, for comparative assessment. These models were selected due to their significant standing and widespread application in computer vision tasks. ResNet50, with its use of residual connections to address vanishing gradients, has been a cornerstone in computer vision. This model offers a solid reference for the performance of CNN architectures. Subsequently, ViT has emerged as a notable contender in the domain. Recognized as the first transformer model to show exemplary performance on image classification tasks, ViT establishes a rigorous standard for applying transformer architectures to visual data processing.

Our evaluation uses three key metrics: precision, recall, and F1-score. Precision measures the proportion of correct positive predictions, while recall captures the model's ability to identify all relevant positive cases. The F1-score, as the harmonic mean of precision and recall, provides a balanced evaluation, especially for imbalanced class distributions."

The MMOE model, which we applied to our classification framework, was evaluated using the same metrics. MMOE is expected to provide enhanced performance by dynamically selecting relevant features for each classification task, addressing the overlap between the "impoverished," "middle," and "affluent" categories. This dynamic expert selection ensures that the model can learn distinct features for each class, significantly reducing confusion between classes. For the three-class task, we performed an evaluation at $s = 1.0$ to capture the model's full potential in separating these categories.

5. Results

5.1. Comparison of original and modified TrueSkill scores

Our modified TrueSkill algorithm demonstrates significant improvements over the original version, as evidenced by key statistical measures and graphical analysis:

As shown in Table 2, while the mean remains constant, our modified algorithm reduces the standard deviation, indicating a more concentrated distribution. The skewness and kurtosis values of the modified scores are closer to zero, suggesting a distribution that more closely approximates a normal distribution.

Fig. 4 visually demonstrates the difference in score distributions between the original TrueSkill algorithm and our modified version. The modified distribution exhibits a more symmetrical and concentrated shape, aligning with the improved statistical measures.

Furthermore, the Q-Q plots (Fig. 5) visually confirm the improved normality of our modified TrueSkill scores. The modified algorithm's Q-Q plot shows points adhering more closely to the theoretical normal distribution line, particularly at the tails, indicating a better fit to normality compared to the original TrueSkill scores.

These improvements in statistical properties and distribution normality suggest that our modified TrueSkill algorithm provides a more reliable and statistically sound basis for wealthiness estimation from street view images. The reduction in standard deviation and the closer adherence to normality in both the distribution and Q-Q plots indicate that our modifications have successfully addressed some of the

Table 2
Comparison of statistical measures between original and modified TrueSkill scores.

Metric	Original TrueSkill	Modified TrueSkill
Mean	25.785	25.785
Standard Deviation	5.385	4.302
Skewness	−0.055	−0.024
Kurtosis	−0.520	−0.476

Table 1
Dataset Distribution and Split for Model Training and Testing (s = 1).

Datasets		Count	Impoverished	Middle	Affluent	Training	Validation	Test
Overall Performance	Global	111,390	18,890	73,676	18,824	89,112	11,139	11,139
Model Generalizability	New York	3396	617	2174	605	2716	340	340
	LA	482	47	290	145	0	0	482
	Boston	1333	233	881	219	0	0	1333

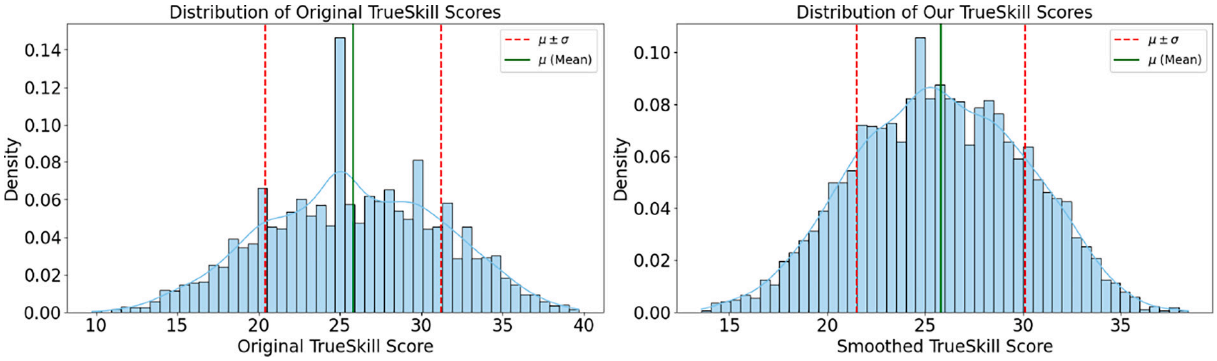


Fig. 4. Distribution comparison of original TrueSkill scores (left) and our modified TrueSkill scores (right).

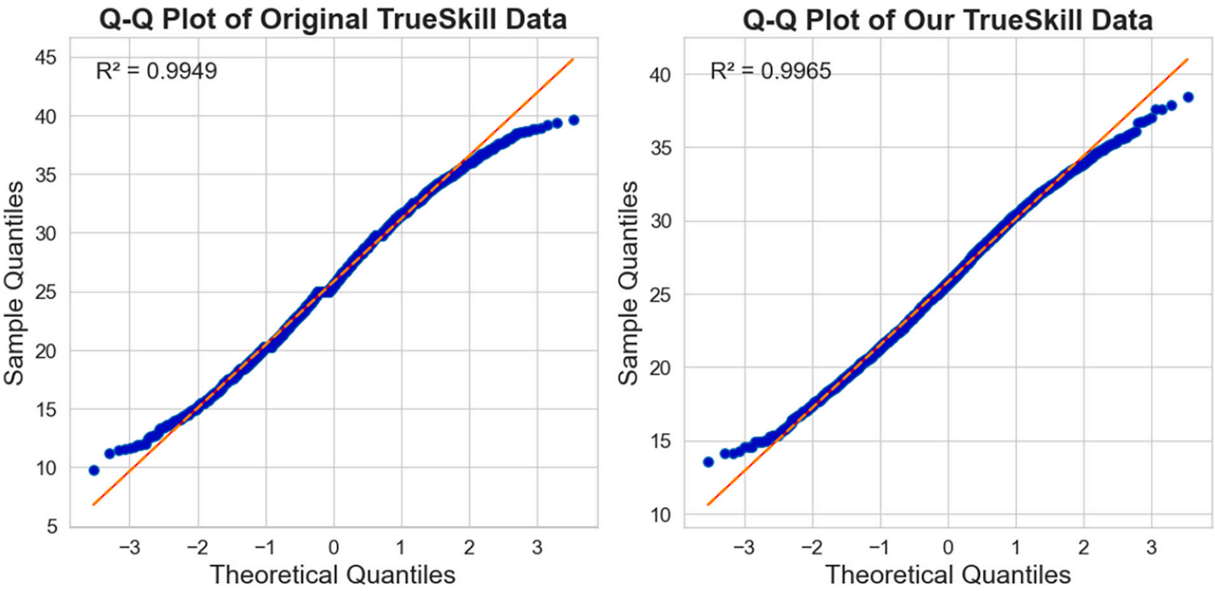


Fig. 5. Q-Q plots comparing original (left) and modified (right) TrueSkill score distributions.

limitations of the original TrueSkill algorithm in the context of urban wealthiness assessment.

5.2. Performance evaluation

Table 3 illustrates a performance comparison between three-class classification models: the proposed MMOE and the benchmark models (Swin Transformer, ResNet50 and ViT). The comparison focused on the threshold $s = 1.0$ to evaluate the model's ability to separate the "impoverished," "middle," and "affluent" categories in a clear-cut scenario. The results indicated that all models faced challenges in differentiating the "middle" category from the other two, a result of overlapping visual features commonly present in urban landscapes. However, MMOE demonstrated a clear performance advantage due to its dynamic expert network, achieving significantly better precision, recall, and F1-score.

The MMOE model clearly outperforms the other models across all metrics. The dynamic expert selection provided by MMOE allows it to adaptively select features most relevant to each task, improving its ability to handle overlapping classes such as the "middle" category. As a result, MMOE achieves significantly higher precision, recall, and F1-scores, especially at $s = 1.0$, where it records an overall accuracy (OA) of 0.82.

5.3. Binary classification performance

Despite the strong performance of MMOE in the three-class classification task, we also explored a binary classification focused solely on the impoverished and affluent categories. This simplification allowed for deeper analysis of class separation using techniques like Grad-CAM for interpretability. By removing the challenging "middle" category, the task became more defined, and the need for MMOE diminished. The expert-selection complexity was reduced, causing MMOE to effectively collapse into a single model, represented by the Swin Transformer. Therefore, we excluded MMOE from further comparisons, focusing instead on the binary performance of Swin, ViT, and ResNet50.

For binary classification, where the "middle" class was excluded, all models showed significantly improved performance. Table 4 presents the performance comparison, demonstrating how focusing on the binary task allowed for better precision, recall, and F1-scores across the models. Swin continued to outperform both ResNet50 and ViT, achieving the highest overall accuracy (OA).

The results show that Swin outperformed both ViT and ResNet50 in the binary classification task, achieving higher precision, recall, and F1-scores across all categories. This performance improvement illustrates that by focusing on the two extreme categories, the models can better differentiate between impoverished and affluent urban landscapes, leading to more accurate classification outcomes.

In conclusion, MOE demonstrated its strength in handling multi-class

Table 3
Performance comparison of the proposed MMOE, Swin, ViT, and ResNet50 for 3-class classification at $s = 1.0$ after performance improvement.

Model	Threshold	Metrics	Impoverished	Affluent	Middle	Mean
Swin + Original TrueSkill	s = 1.0	Precision	0.39	0.32	**0.68**	0.46
		Recall	0.33	0.27	**0.78**	0.46
		F1-score	0.35	0.29	**0.73**	0.46
		OA			**0.63**	
Swin + Modified TrueSkill	s = 1.0	Precision	0.45	0.38	**0.72**	0.52
		Recall	0.41	0.35	**0.84**	0.53
		F1-score	0.43	0.36	**0.77**	0.52
		OA			**0.68**	
MMOE + Modified TrueSkill	s = 1.0	Precision	0.72	0.67	**0.84**	0.74
		Recall	0.67	0.63	**0.91**	0.73
		F1-score	0.70	0.65	**0.88**	0.74
		OA			**0.82**	
MMOE + Original TrueSkill	s = 1.0	Precision	0.62	0.57	**0.75**	0.65
		Recall	0.59	0.52	**0.78**	0.63
		F1-score	0.60	0.54	**0.76**	0.63
		OA			**0.70**	
ViT + Original TrueSkill	s = 1.0	Precision	0.29	0.28	**0.65**	0.40
		Recall	0.29	0.21	**0.69**	0.39
		F1-score	0.29	0.24	**0.67**	0.40
		OA			**0.54**	
ViT + Modified TrueSkill	s = 1.0	Precision	0.32	0.31	**0.70**	0.44
		Recall	0.33	0.24	**0.72**	0.43
		F1-score	0.32	0.27	**0.71**	0.43
		OA			**0.58**	
ResNet50 + Original TrueSkill	s = 1.0	Precision	0.37	0.48	**0.61**	0.49
		Recall	0.56	0.22	**0.62**	0.47
		F1-score	0.45	0.30	**0.62**	0.48
		OA			**0.51**	
ResNet50 + Modified TrueSkill	s = 1.0	Precision	0.40	0.52	**0.65**	0.52
		Recall	0.59	0.26	**0.67**	0.51
		F1-score	0.48	0.35	**0.66**	0.50
		OA			**0.56**	

Table 4
Performance comparison of Swin, ViT, and ResNet50 for binary classification after improvement.

Model	Threshold	Metrics	Impoverished	Affluent	Mean
Swin	s = 1.0	Precision	**0.87**	0.81	0.84
		Recall	0.79	**0.88**	0.84
		F1-score	0.83	**0.84**	0.84
		OA		**0.82**	
ViT	s = 1.0	Precision	0.75	**0.79**	0.77
		Recall	**0.87**	0.63	0.75
		F1-score	**0.81**	0.70	0.76
		OA		**0.76**	
ResNet50	s = 1.0	Precision	0.65	**0.91**	0.78
		Recall	**0.95**	0.52	0.73
		F1-score	0.78	0.67	0.73
		OA		**0.73**	

tasks with overlapping features, achieving the best results in three-class classification. However, in the binary classification task, where the challenge was simplified, Swin emerged as the most robust model, delivering superior performance compared to ViT and ResNet50.

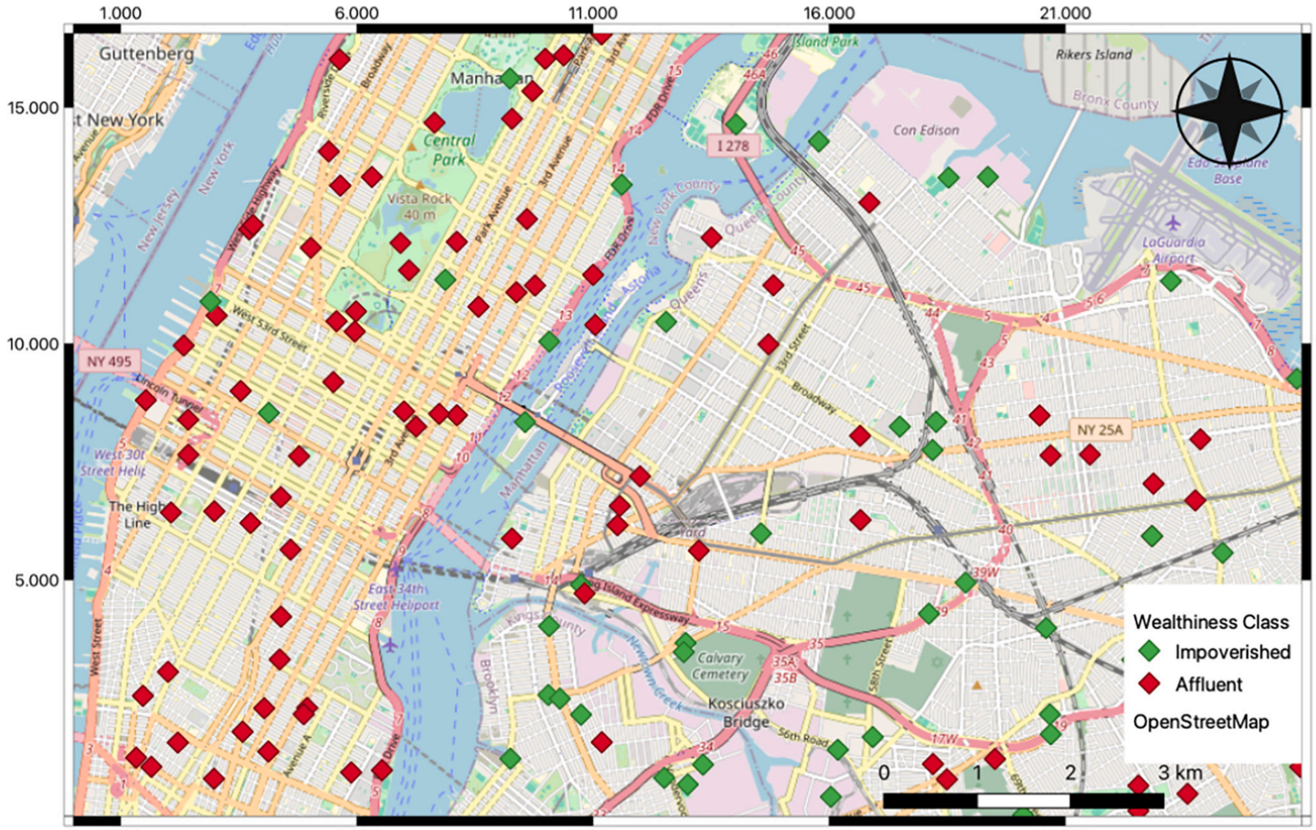
5.4. Result analysis

For a more streamlined and intuitive visualization, comparison and analysis of the model performance and interpretability, we prioritized the examination of the binary classification model over the ternary model. Binary classification models, due to their simplicity and robustness, often provide more straightforward and reliable insights than their multiclass counterparts. Focusing on the binary model simplified the analysis, highlighting key differences between the two classes and the factors driving them. However, while binary classification models can provide valuable insights, they may not capture the full spectrum of wealthiness in urban environments. Future work could consider more complex models to provide a more nuanced view of wealthiness distribution.

Fig. 6 displays the spatial distribution of the predicted images by the binary-class Swin model in the New York City area, with images classified as affluent class shown in red and impoverished ones in green. The results revealed that affluent-class locations were primarily clustered in the Manhattan and Brooklyn neighborhoods, with a few scattered throughout the Bronx, Queens, and Staten Island. These areas are characterized by features such as higher building density and greater urban activity, which are typical indicators of affluence in urban environments. On the contrary, impoverished areas were more dispersed throughout all five boroughs, with a higher concentration in the suburban areas of the city. This clear visual representation of the spatial distribution of wealthiness in NYC is useful for urban planners and policy-makers to understand spatial dependence and heterogeneity of wealthiness patterns.

A notable finding of our study is that most street scenes near cemeteries are consistently categorized as 'impoverished' in terms of wealthiness. The bottom center of Fig. 6 highlights a specific cemetery as an example. The presence of a cemetery within a community can substantially influence its perceived affluence due to its potential impacts on local property values, real estate market dynamics, the surrounding environment, and public health concerns (Jonker & Olivier, 2012; Tang, 2019). These intertwined factors collectively contribute to the lower wealthiness classification of communities near cemeteries.

To further investigate this, we performed a sensitivity analysis by varying the buffer distance around cemeteries and compared the average wealthiness scores of street scenes within each buffer distance to the average score of all street scenes. Our analysis revealed a trend where the wealthiness of street scenes is more likely to be affected by the presence of a cemetery within shorter distances, and the impact diminishes as the distance increases. This pattern is clearly illustrated in plot Fig. 7), which shows the relationship between the buffer distance and the average wealthiness score of street scenes near a cemetery. As the distance from the cemetery increases, the average wealthiness score



Map generated using QGIS with OpenStreetMap as the basemap.Data overlays highlight wealthiness prediction results from the study.

Fig. 6. Mapping of Google Street View images with binary predicted wealthiness classes with s = 1 in the study area of NYC.

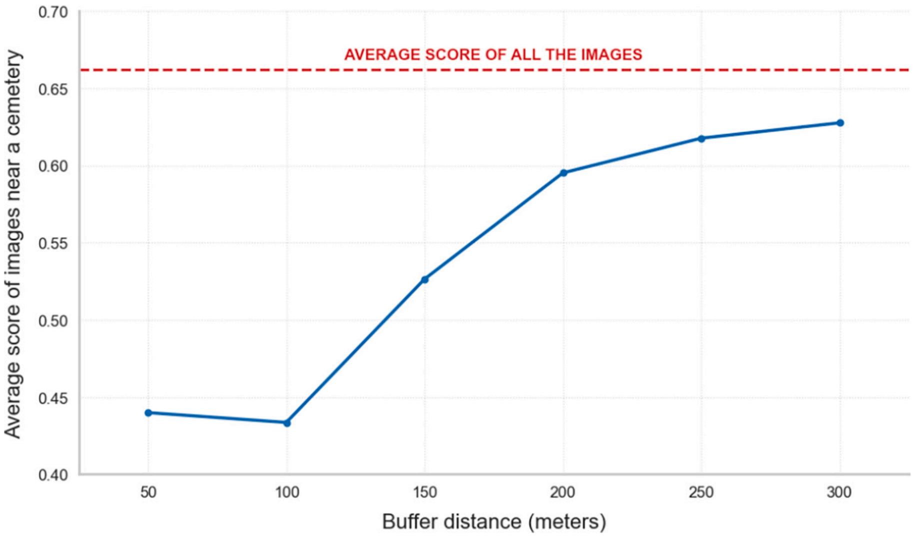


Fig. 7. Effect of buffer distance on binary scores of images near a cemetery.

approaches the overall average score for all street scenes.

5.5. Model interpretability

To decipher the decision-making process of our model, we employed Grad-CAM (Selvaraju et al., 2016) to highlight the most influential areas of the images in the model's predictions. Grad-CAM is a powerful technique for visualizing what a deep learning model is focusing on. It generates saliency maps by using the partial derivatives of the output class score concerning the feature maps of the last convolutional layer in a CNN (Selvaraju et al., 2017). This provides a measure of the importance of each feature map in making the model's decision. Due to its ability to shed light on the complex inner workings of CNNs—often viewed as black boxes—Grad-CAM proves particularly insightful. It finds extensive use in various domains such as image classification, object detection, and semantic segmentation.

In our study, we found that the Grad-CAM visualizations provided valuable insights into the factors that the model used to predict the wealthiness of a location. For example, in many cases, the Grad-CAM heatmaps showed that the model was using visual cues such as the presence of high buildings, large houses, villas, or well-maintained landscapes to predict that a location was "affluent" (Fig. 11). On the other hand, for locations that were predicted to be "impoverished", the Grad-CAM heatmaps often showed that the model focused on visual characteristics such as the presence of low buildings, old cars, or overgrown vegetation (Fig. 10).

Based on our analysis, we found that the Pearson correlation between the predictions made by our model and the household median income by census blocks is relatively low, with a correlation coefficient of 0.093. However, based on a further manual inspection examination of the images (Fig. 8 and Fig. 9), our model's predictions align better with the human perception of urban environments. This finding suggests that our model can effectively capture the nuances and subtleties of the urban environments and reflect the spatial heterogeneity in the distribution of wealthiness that cannot be learned by traditional spatial aggregation methods.

These results suggest that our model is able to accurately use the visual appearance of the images to predict the wealthiness of the locations depicted in the images, demonstrating the great potential of using human perception and deep learning to improve the prediction of the wealthiness of urban environments.

5.6. Model generalizability

To evaluate the generalizability of our proposed models, we conducted experiments using the MMOE model for three-class classification and the Swin Transformer for binary classification (Table 5). Both models were trained on street view images from New York City (NYC) along with their corresponding modified TrueSkill scores. We then tested these models on images from Boston and Los Angeles (LA) to assess their performance in distinct geographical regions.

5.6.1. East coast generalization – Boston

When generalizing to Boston, both our models demonstrated strong performance. The MMOE model achieved an average accuracy of 75 % across the three classes, with scores of 0.72, 0.78, and 0.75 for the impoverished, middle, and affluent classes respectively (Table 5). This robust performance suggests that the model effectively captured common urban features shared between NYC and Boston, despite their unique characteristics. The binary Swin model also performed well, achieving an average accuracy of 73 %, further confirming the model's ability to transfer learning between these two East Coast cities.

5.6.2. West coast generalization - Los Angeles

In the LA scenario, our models showed slightly different patterns of generalization. The MMOE model maintained strong performance with an average accuracy of 68 % (0.65, 0.70, and 0.69 for impoverished, middle, and affluent classes respectively; Table 5). While there was a slight decrease compared to the Boston results, this performance is still impressive given the significant differences between NYC and LA in terms of architectural style, urban layout, and climate. Interestingly, our binary Swin model showed even stronger generalization to LA, with an average accuracy of 81 %. This suggests that while the nuanced distinctions required for three-class classification become more challenging across disparate urban environments, the broader binary classification task remains robust.

These results underscore the effectiveness and adaptability of both our MMOE and Swin models. The MMOE model's ability to maintain high accuracy in three-class classification across diverse urban landscapes demonstrates its potential for nuanced urban analysis tasks. Meanwhile, the Swin model's exceptional performance in binary classification, particularly its strong generalization to the West Coast, highlights its potential for widespread application in urban landscape classification tasks.

The slight performance variations between East and West Coast generalizations provide valuable insights into the challenges and opportunities in cross-regional urban analysis. While both models perform well, the differences suggest that future work could explore region-specific fine-tuning or the incorporation of additional regional context to further enhance model generalizability across diverse urban environments.

5.7. Comparison with census data

This study observed a relatively low correlation between the wealthiness predicted values with median income level by census

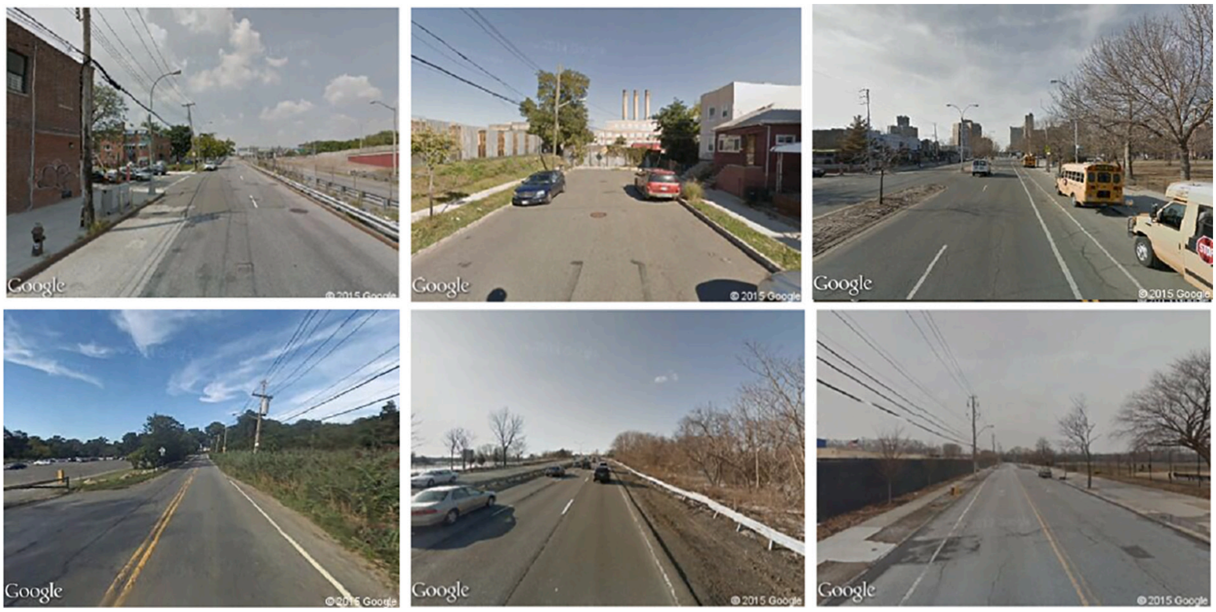


Fig. 8. Example images where the score class is impoverished while the mean household income class is affluent.

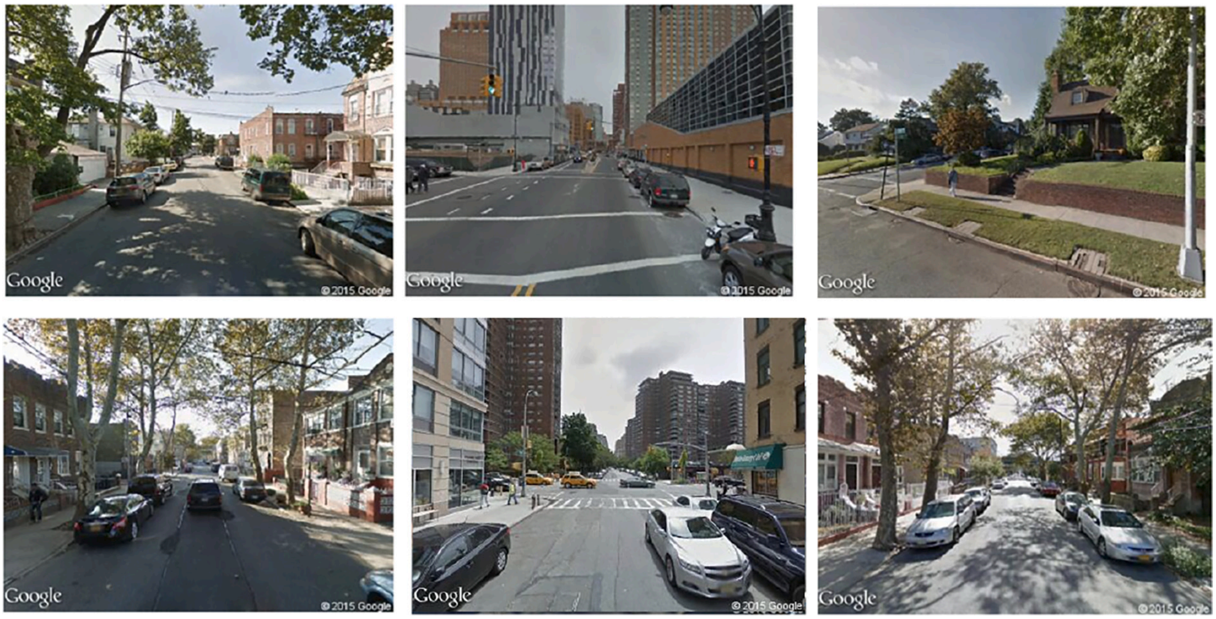


Fig. 9. Example images where the score class is affluent while the mean household income class is impoverished.

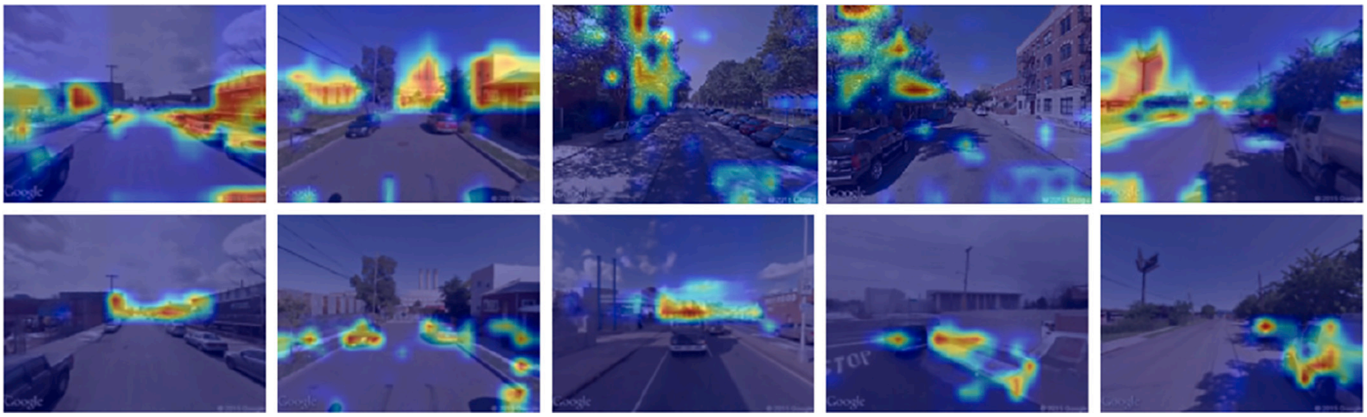


Fig. 10. The **GradCam** image of the model on the impoverished street scene (Selvaraju et al., 2016), and the highlighted part of the image shows the key point areas in the image that affect the classification of the model. For impoverished places, which are usually desolate, vehicles, plants and houses in the images are the key factors affecting the classification of the model.

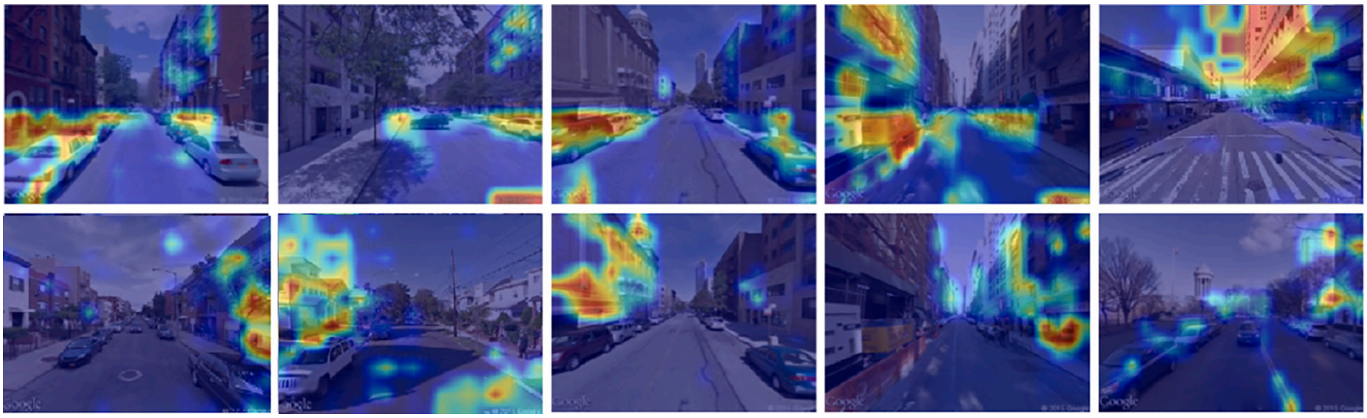


Fig. 11. The **GradCam** (Selvaraju et al., 2016) image of the model about the affluent street scene, and the highlighted part of the image presents the key point areas that affect the model classification. Compared with the impoverished images, the affluent places tend to have more vehicles, the vehicles are also more technologically advanced, and the density of buildings is also higher.

Table 5
Model Generalizability Results.

Scenario	Three-classification (MMOE)			Binary classification (Swin)	
	Impoverished	Middle	Affluent	Impoverished	Affluent
NY → Boston	0.72	0.78	0.75	**0.79**	0.67
NY → LA	0.65	0.70	0.69	**0.88**	0.74

blocks. One potential explanation for the low correlation is the differences in the granularity (or resolution) of the data. The census block group data provides a broad, aggregate view of income levels in a given area, while the Streetview class data capture more detailed and fine-grained information about the specific characteristics and features of individual neighborhoods. Hence, the two datasets may measure different aspects of wealthiness and the low correlation does not necessarily imply a lack of association. Another potential explanation is the influence of other factors that are not captured by the census block group data. For example, factors such as the quality of public services, the availability of amenities, and the level of social cohesion in a neighborhood may also play a role in shaping the human perception of wealthiness. Integrating these factors into the analysis can provide a more comprehensive understanding of the relationship between income and wealthiness.

6. Discussion

This study introduces a novel approach to predicting neighborhood wealth by integrating human perception with deep learning models trained on geo-tagged street view images. This section discuss the practical applications, advantages, and limitations of this work.

6.1. Practical applications

Beyond the technical performance of our models, it is crucial to explore the broader implications, and practical applications in urban planning, real estate development, and socioeconomic research.

6.1.1. Urban planning

Accurate, fine-grained predictions of neighborhood wealth can significantly inform urban planning and policy-making. By identifying socioeconomically disadvantaged areas, city planners can prioritize interventions such as enhancing public infrastructure, improving access to education and healthcare, and creating employment opportunities. For instance, our model can reveal neighborhoods lacking essential amenities, guiding targeted investments to promote equitable urban development. This aligns with the principles of equitable urban development, as highlighted in prior research on urban inequalities (Suel et al., 2023a).

Moreover, the dynamic capabilities of our model enable continuous monitoring of urban changes. By tracking shifts in wealth distribution, policymakers can detect early signs of gentrification or urban decline, facilitating proactive decision-making. For example, urban revitalization programs can be informed by identifying neighborhoods undergoing economic stress or rapid socioeconomic changes. This data-driven approach helps ensure that urban policies remain responsive to evolving challenges (Huang et al., 2022).

6.1.2. Real estate development

In the real estate sector, wealth predictions offer valuable insights into market dynamics. Developers and investors can use our model to identify areas with high growth potential or emerging markets. For example, predicting wealth trends in rapidly developing neighborhoods can guide decisions on where to build new residential or commercial properties. The growing use of AI in real estate demonstrates its transformative potential, particularly in improving investment decisions and property development (Viriato, 2019).

Additionally, our approach enhances traditional property valuation methods by incorporating visual features such as building aesthetics, green spaces, and street-level amenities. These factors are often overlooked in census-based models but are critical for understanding property values. By leveraging visual data, real estate stakeholders can make more informed decisions that account for both tangible and intangible neighborhood characteristics (Law et al., 2019).

6.2. Advantages over traditional methods

Compared to traditional wealth prediction approaches, which typically rely on aggregated census data, our method offers several distinct advantages. First, the use of geo-tagged street view images allows for real-time and fine-grained assessments of wealth, capturing recent urban changes that census data often misses (Naik et al., 2014). This is particularly valuable in rapidly evolving urban environments, where static data sources may no longer reflect current conditions.

Second, our incorporation of human perception into wealth prediction captures nuanced socioeconomic indicators that are difficult to quantify. For example, visual cues such as well-maintained facades, green spaces, and road quality may reflect neighborhood wealth but are often excluded from conventional datasets. By extracting these features, our model provides a more comprehensive understanding of urban environments.

Third, the adoption of advanced machine learning techniques, such as the Multi-gate Mixture-of-Experts (MMOE) model, allows for specialized feature extraction tailored to different wealth classes. This improves the model's performance in distinguishing subtle socioeconomic differences, such as those within the middle class, a challenge often overlooked in binary classification models.

6.3. Limitations

One limitation of our study is the bias present in the Place Pulse 2.0 dataset. The dataset was crowdsourced from online volunteers, and it is possible that these volunteers may not be fully representative of the broader population. This can introduce biases in the dataset, such as disproportionate representation of certain demographics or skewed perception of what might be considered an "affluent" or "impoverished" location. These biases have the potential to impact the accuracy of our model's predictions. Additionally, our study is limited by the binary and ternary classification models used to predict the level of wealthiness. While these simplified approaches allowed us to evaluate the models' performance using standard classification metrics, they may not fully capture the spectrum of wealthiness in urban environments. Looking forward, instead of employing binary or ternary classification schemes, a more nuanced approach would involve transitioning to regression-based models. This approach would aim to predict a continuous wealthiness score rather than restrict the outcome to specific classes. This change can provide a more detailed, granular, and comprehensive understanding of urban wealthiness distribution, thereby significantly enhancing the model's predictive capability and its potential real-world applications.

While binary and ternary classification models offer methodological simplicity, we acknowledge their limitations in fully representing the nuanced socioeconomic landscape of urban areas. Recent research has highlighted the multifaceted nature of urban inequality and the changing dynamics of the middle class. For example, Pew Research Center studies have shown that the American middle class has been shrinking over time, with movement into both upper and lower income tiers (Kochhar, 2018). Additionally, research on urban inequalities in the 21st century economy has revealed growing disparities not just between high and low income groups, but also within suburbs and between cities (Nijman & Wei, 2020). The emergence of the new information-based economy has created more complex patterns of inequality that go beyond simple classifications (Muniz & Bailey, 2022). We recognize that our binary/ternary models may oversimplify these intricate socioeconomic variations. However, they serve as a starting point for analysis while maintaining methodological clarity. In future work, we aim to explore more sophisticated approaches that can better capture the spectrum of urban wealthiness distribution, potentially incorporating methods like the multiple correspondence analysis used in some household classification studies (Were et al., 2022). This could provide a more nuanced understanding of socioeconomic stratification in urban settings without sacrificing the interpretability of our current approach.

Moreover, while our study provides a technical foundation for assessing urban wealthiness, its practical application in urban planning and policy-making requires further exploration. For instance, the insights gained from our model could be leveraged to inform targeted urban development projects, helping to identify areas in need of investment or redevelopment. Policymakers could use these findings to devise strategies aimed at reducing wealthiness disparities and improving overall urban livability. This practical application aspect is an area that needs more attention in future studies to bridge the gap between technical capability and real-world impact.

Additionally, the low resolution image (400 × 300 pixels) in our dataset poses a challenge. This limitation is significant as finer details could be useful in assessing wealthiness levels. However, we advocate for enhancing visual data analysis within the same resolution constraints, rather than diverging to non-visual data sources. We believe that advancements in image processing and machine learning techniques can extract more meaningful information from existing visual data. This approach aligns with our study's core objective of leveraging visual perception for wealthiness prediction and maintains the methodological consistency of relying primarily on visual data.

7. Conclusion

This study proposed a novel approach for predicting neighborhood wealthiness based on human perception and deep learning at the most fine-grained level (i.e., point-scale spatial resolution), by leveraging geotagged street view images as the input for model training. This approach is free from the cumbersomeness of MAUP and is able to predict wealthiness at a finer-grained level, better reflecting the spatial dependence and heterogeneity of the distribution of wealthiness. Our results have demonstrated more insights into using human perception and deep learning to predict neighborhood wealthiness, compared with using the income statistics aggregated by census block group.

Overall, our study underscores the importance of considering human perception in predicting neighborhood wealthiness. This has significant implications for urban planning and policy-making, where our findings can inform targeted interventions and strategic development aimed at reducing wealthiness disparities and enhancing urban livability. The practical applications of our technology in urban planning, policy formation, and even in real estate and socio-economic research, open up new avenues for understanding and addressing the complexities of urban wealthiness distribution with more social equality.

A key innovation in our study was the modification of the TrueSkill Rating System to incorporate temporal decay and spatial autocorrelation factors. This allowed for more accurate wealthiness predictions by considering the dynamic changes in urban environments over time and space. Additionally, we introduced the Multi-gate Mixture-of-Experts (MMOE) model, which significantly improved the classification performance. These advancements enhance the precision and applicability of our model in diverse urban settings.

Future research could delve into the use of higher-resolution street view images, capturing more detailed and nuanced visual cues about urban wealthiness. In addition, much more efforts are needed to apply this visual-based technology to real-world applications in urban planning and policy-making. For example, city planners can capitalize on these insights to identify socio-economic disparities, focusing on targeted interventions in underprivileged neighborhoods. Crucially, future studies should explore how to establish effective collaborations between technologists, urban planners, and policymakers. Such partnerships are essential to translate visual data insights into actionable urban development strategies, ultimately applying the model's findings to foster more equitable and socially conscious urban environments.

A particularly innovative and crucial direction for future research lies in the integration of temporal data. Urban environments are not static; they evolve and transform over time. This dynamism is often a reflection of underlying socioeconomic changes, including shifts in wealthiness distribution. By analyzing changes in street view imagery over time, we can track the evolution of neighborhoods and predict future trends in urban wealthiness. This approach goes beyond a snapshot view of urban wealthiness, offering a movie-like, dynamic perspective that captures the trajectory of neighborhoods over time. The incorporation of temporal data can lead to a paradigm shift in how we understand and analyze urban wealthiness. It allows us to capture the temporal nuances that static models might miss, providing a more accurate, nuanced, and holistic understanding of urban wealthiness. It also enables us to forecast future trends, offering valuable insights for urban planning and policy-making.

These advancements, particularly the integration of temporal data, will significantly enhance our understanding of the complexities of wealthiness distribution in urban environments. They will contribute to the development of more sophisticated, accurate, and dynamic approaches for analyzing urban wealthiness, paving the way for more informed and effective urban planning and policy decisions.

CRediT authorship contribution statement

Yang Qiu: Visualization, Validation, Methodology, Formal analysis, Writing – review & editing, Writing – original draft. **Meiliu Wu:** Methodology, Data curation, Conceptualization, Writing – review & editing. **Qunying Huang:** Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization, Writing – review & editing. **Yuhao Kang:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation.

Data availability

Data and source code are available at https://github.com/scdmlab/Wealthiness_Prediction_Cities.

References

Abitbol, J. L., & Karsai, M. (2020). Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence, 2*, 684–692.

Batty, M., 2021. Defining urban science, in: Urban informatics. Springer Singapore, pp. 15–28. URL: Doi:https://doi.org/10.1007/978-981-15-8983-6_3, doi: https://doi.org/10.1007/978-981-15-8983-6_3.

Biljecki, F., & Ito, K. (2021a). Street view imagery in urban analytics and gis: A review. *Landscape and Urban Planning, 215*, Article 104217.

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

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science, 350*, 1073–1076.

Cai, X., & Wang, D. (2006). Spatial autocorrelation of topographic index in catchments. *Journal of Hydrology, 328*, 581–591. https://doi.org/10.1016/J.JHYDROL.2006.01.009

Chakravorty, S., 1996a. Urban inequality revisited. Urban Affairs Review 31, 759–777. URL: doi:https://doi.org/10.1177/107808749603100604, doi: https://doi.org/10.1177/107808749603100604.

Chakravorty, S. (1996b). Urban inequality revisited. *Urban Affairs Review, 31*, 759–777. https://doi.org/10.1177/107808749603100604

Dong, L., Ratti, C., & Zheng, S. (2019). Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proceedings of the National Academy of Sciences, 116*, 15447–15452. https://doi.org/10.1073/pnas.1903064116

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. *European conference on computer vision, Springer.*, 196–212.

Fang, C. (2018). Growing wealth gaps in education. URL https://www.src.isr.umich.edu/growing-wealth-gaps-in-education/.

Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M., & Weber, I. (2020). Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science, 9*. https://doi.org/10.1140/epjds/s13688-020-00235-w

Ferreira, F. H. (2021). Inequality and covid-19. *Finance & Development URL*. https://www.imf.org/external/pubs/ft/fandd/2021/06/inequality-and-covid-19-ferreira.htm.

Fintz, M., Osadchy, M., & Hertz, U. (2022). Using deep learning to predict human decisions and using cognitive models to explain deep learning models. *Scientific Reports, 12*, 4736.

Foundation, R.W.J. (2018). Wealth matters for health equity. URL https://www.rwjf.org/en/insights/our-research/2018/09/wealth-matters-for-health-equity.html.

Gao, C., Feng, Y., Tong, X., Lei, Z., Chen, S., & Zhai, S. (2020). Modeling urban growth using spatially heterogeneous cellular automata models: Comparison of spatial lag, spatial error and gwr. *Computers, Environment and Urban Systems, 81*, Article 101459. https://doi.org/10.1016/j.compenvurbsys.2020.101459

Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. (2016). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry, 56*, 114–137. https://doi.org/10.1111/ecin.12364

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K., 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677.

Guo, J., Jia, N., & Bai, J. (2022). Transformer based on channel-spatial attention for accurate classification of scenes in remote sensing image. URL https://www.nature.com/articles/s41598-022-19831-z.

Gyourko, J., Mayer, C., & Sinai, T. (2013). Superstar cities. American economic journal. *Economic Policy, 5*, 167–199. https://doi.org/10.1257/pol.5.4.167

Hansen, M., DeFries, R., Townshend, J., & Sohlberg, R. (2000). Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing, 21*, 1331–1364. https://doi.org/10.1080/014311600210209

He, J., Zhang, J., Yao, Y., & Li, X. (2023). Extracting human perceptions from street view images for better assessing urban renewal potential. *Cities, 134*, Article 104189.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition. 2016 IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.90

Hermans, J. R., Spanakis, G., & Möckel, R. (2017). Accumulated gradient normalization. *Asian Conference on Machine Learning, PMLR.*, 439–454.

Huang, T., Dai, T., Wang, Z., Yoon, H., Sheng, H., Ng, A. Y., Rajagopal, R., & Hwang, J. (2022). *Detecting neighborhood gentrification at scale via street-level visual data. 2022 IEEE International Conference on Big Data (Big Data)* (pp. 1632–1640). https://api.semanticscholar.org/CorpusID:255440542.

Indaco, A. (2020). From twitter to gdp: Estimating economic activity from social media. *Regional Science and Urban Economics, 85*, Article 103591. https://doi.org/10.1016/j.regsciurbeco.2020.103591

Jonker, C., & Olivier, J. (2012). Mineral contamination from cemetery soils: Case study of zandfontein cemetery, South Africa. *International Journal of Environmental Research and Public Health, 9*, 511–520. https://doi.org/10.3390/ijerph9020511

Kang, Y., Abraham, J., Ceccato, V., Duarte, F., Gao, S., Ljungqvist, L., … Ratti, C. (2023). Assessing differences in safety perceptions using geoai and survey across neighbourhoods in Stockholm, Sweden. *Landscape and Urban Planning, 236*, Article 104768.

Kawachi, I., & Kennedy, B. P. (1997). Socioeconomic determinants of health: Health and social cohesion: Why care about income inequality? *BMJ, 314*, 1037. https://doi.org/10.1136/bmj.314.7086.1037

Kim, S. E., Jeon, J. J., & Eom, I. K. (2016). Image contrast enhancement using entropy scaling in wavelet domain. *Signal Processing, 127*, 1–11.

Kochhar, R. (2018). The american middle class is stable in size, but losing ground financially to upper-income families. URL https://www.pewresearch.org/short-reads/2018/09/06/the-american-middle-class-is-stable-in-size-but-losing-ground-financially-to-upper-income-families/.

Kononenko, I., & Kukar, M. (2007). *Machine learning and data mining.* Horwood Publishing.

Laksmi, T., Madhu, T., Kavya, K., & Basha, S. E. (2016). Novel image enhancement technique using clahe and wavelet transforms. *International Journal of Scientific Engineering and Technology, 5*, 507–511.

Law, S., Paige, B., Russell, C., 2019. Take a look around: Using street view and satellite images to estimate house prices. ACM Transactions on Intelligent Systems and Technology 10. URL: doi:https://doi.org/10.1145/3342240, doi: https://doi.org/10.1145/3342240.

Liang, X., Zhao, T., & Biljecki, F. (2023). Revealing spatio-temporal evolution of urban visual environments with street view imagery. *Landscape and Urban Planning, 237*, Article 104802.

Lin, L., Di, L., Zhang, C., Guo, L., & Di, Y. (2021). Remote sensing of urban poverty and gentrification. *Remote Sensing, 13*, 4022. https://doi.org/10.3390/rs13204022

Lin, Y., Zhao, J., Zhang, M., & Chen, G. (2020). Analysis of spatial-temporal differentiation and influence factors of construction land expansion of the urban agglomeration in Central Yunnan. *IOP Conference Series: Materials Science and Engineering, 780*. https://doi.org/10.1088/1757-899X/780/7/072042

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lowe, G. (2004). Sift-the scale invariant feature transform. *International Journal, 2*, 2.

Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., & Chi, E. H. (2018). Modeling task relationships in multi-task learning with multi-gate mixture-of-experts, in: Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining. *ACM. pp., 1930–1939*.

MacDonald, L. (2014). Median income as a better measure of development progress-nancy birdsall and christian meyer. URL https://www.cgdev.org/blog/median-income-better-measure-development-progress-nancy-birdsall-and-christian-meyer. (Accessed 21 May 2023).

Menzel, C., & Reese, G. (2021). Implicit associations with nature and urban environments: Effects of lower-level processed image properties. *Frontiers in Psychology, 12*, Article 591403.

Microsoft, T. (2005). Trueskill™ ranking system. URL https://www.microsoft.com/en-us/research/project/trueskill-ranking-system/. (Accessed 19 March 2023).

Moradi, F., Biloria, N., & Prasad, M. (2023). Analyzing the age-friendliness of the urban environment using computer vision methods. *Environment and Planning B: Urban Analytics and City Science, 50*, 2294–2308. https://doi.org/10.1177/23998083231153862

More, L. G., Brizuela, M. A., Ayala, H. L., Pinto-Roa, D. P., & Noguera, J. L. (2015). Parameter tuning of clahe based on multi-objective optimization to achieve different contrast levels in medical images. In *2015 IEEE international conference on image processing (ICIP)*. https://doi.org/10.1109/icip.2015.7351687

Muniz, J., & Bailey, S. R. (2022). Does race response shift impact racial inequality? *Demographic Research, 47*, 935–966. https://doi.org/10.4054/demres.2022.47.30

Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). *Streetscore – Predicting the perceived safety of one million streetscapes, in: 2014 IEEE conference on computer vision and pattern recognition workshops* (pp. 793–799). https://doi.org/10.1109/CVPRW.2014.121

Nelson, J. K., & Brewer, C. A. (2015). Evaluating data stability in aggregation structures across spatial scales: Revisiting the modifiable areal unit problem. *Cartography and Geographic Information Science, 44*, 35–50. https://doi.org/10.1080/15230406.2015.1093431

Nicoletti, L., Sirenko, M., & Verma, T. (2022). Disadvantaged communities have lower access to urban infrastructure. *Environment and Planning B: Urban Analytics and City Science, 50*, 831–849. https://doi.org/10.1177/23998083221131044

Nijman, J., & Wei, Y. D. (2020). Urban inequalities in the 21st century economy. *Applied Geography, 117*, Article 102188. https://doi.org/10.1016/j.apgeog.2020.102188

Niu, Z., Zhong, G., & Yu, H. (2021a). A review on the attention mechanism of deep learning. *Neurocomputing, 452*, 48–62.

Niu, Z., Zhong, G., & Yu, H. (2021b). A review on the attention mechanism of deep learning. *Neurocomputing, 452*, 48–62. https://doi.org/10.1016/j.neucom.2021.03.091

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. *Pytorch: An imperative style, high-performance deep learning library*. (2019). URL https://arxiv.org/abs/1912.01703.

Piga, B., & Morello, E. (2015). Environmental design studies on perception and simulation: An urban design approach. *Ambiances*. https://doi.org/10.4000/ambiances.647

Pizer, S., Johnston, R., Ericksen, J., Yankaskas, B., & Muller, K. (1990). Contrast-limited adaptive histogram equalization: Speed and effectiveness. In *[1990] proceedings of the first conference on visualization in biomedical computing*. https://doi.org/10.1109/vbc.1990.109340

Qin, Z., Yu, Y., & Liu, D. (2019). The effect of hopsca on residential property values: Exploratory findings from Wuhan, China. *Sustainability, 11*, 471. https://doi.org/10.3390/su11020471

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J., 2019. Stand-alone self-attention in vision models. arXiv:arXiv:1906.05909.

Raphael, S., Schneider, D., 2023. Introduction: The socioeconomic impacts of covid-19. URL: https://www.rsfjournal.org/content/9/3/1.

Roy, S., Majumder, S., Bose, A., & Chowdhury, I. R. (2023). Does geographical heterogeneity influence urban quality of life? A case of a densely populated indian city. *Papers in Applied Geography, 9*, 395–424. https://doi.org/10.1080/23754931.2023.2225541

Salesses, P., & Hidalgo, C. A. (). *Place Pulse*. https://figshare.com/articles/dataset/Place_Pulse/11859993. https://doi.org/10.6084/m9.figshare.11859993.v1

Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PLoS One, 8*. https://doi.org/10.1371/journal.pone.0068400

Sampson, R. J., & Raudenbush, S. W. (2004). Seeing disorder: Neighborhood stigma and the social construction of "broken windows". *Social Psychology Quarterly, 67*, 319–342. https://doi.org/10.1177/019027250406700401

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE international conference on computer vision (ICCV)* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D., 2016. Grad-cam: Why did you say that? Visual explanations from deep networks via gradient-based localization. CoRR abs/1610.02391. URL: http://arxiv.org/abs/1610.02391, arXiv:1610.02391.

Shokrollahi, A., Mahmoudi-Aznaveh, A., & Mazloom-Nezhad Maybodi, B. (2017). Image quality assessment for contrast enhancement evaluation. *AEU - International Journal of Electronics and Communications, 77*, 61–66. https://doi.org/10.1016/j.aeue.2017.04.026

Suel, E., Muller, E., Bennett, J. E., Blakely, T., Doyle, Y., Lynch, J., , … Nathvani, R., et al. (2023a). Do poverty and wealth look the same the world over? A comparative study of 12 cities from five high-income countries using street images. *EPJ Data Science, 12*. https://doi.org/10.1140/epjds/s13688-023-00394-6

Suel, E., Muller, E., Bennett, J. E., Blakely, T., Doyle, Y., Lynch, J., , … Nathvani, R., et al. (2023b). Do poverty and wealth look the same the world over? A comparative study of 12 cities from five high-income countries using street images. EPJ data. *Science, 12, Article 33*. https://doi.org/10.1140/epjds/s13688-023-00394-6

Sundaram, M., Ramar, K., Arumugam, N., & Prabin, G. (2011). Histogram modified local contrast enhancement for mammogram images. *Applied Soft Computing, 11*, 5809–5816.

Suss, J., Kemeny, T., & Connor, D. S. (2024). Geowealth-us: Spatial wealth inequality data for the United States, 1960–2020. *Scientific Data, 11*. https://doi.org/10.1038/s41597-024-03059-9

Tang, J. T. (2019). Cemeteries use a lot of space and are terrible for the environment. Is there a better way? URL. https://ggwash.org/view/70300/burial-culture-and-the-issues-with-using-so-much-space-for-cemeteries. (Accessed 18 April 2023).

Taubenböck, H., Staab, J., Zhu, X.X., Geiß, C., Dech, S., Wurm, M., 2018. Are the poor digitally left behind? Indications of urban divides based on remote sensing and twitter data. ISPRS International Journal of Geo-Information 7. URL: https://www.mdpi.com/2220-9964/7/8/304, doi: https://doi.org/10.3390/ijgi7080304.

Tomasi, C. (2012). Histograms of oriented gradients. *Computer Vision Sampler*, 1–6.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2012. Training data-efficient image transformers and distillation through attention (2020). Doi: 10.48550. Arxiv.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Viriato, J. C. (2019). Ai and machine learning in real estate investment. *The Journal of Portfolio Management, 45*, 43–54. https://doi.org/10.3905/jpm.2019.45.7.043

Were, V., Foley, L., Turner-Moss, E., Mogo, E., Wadende, P., Musuva, R., & Obonyo, C. (2022). Comparison of household socioeconomic status classification methods and effects on risk estimation: Lessons from a natural experimental study, Kisumu, western Kenya. *International Journal for Equity in Health, 21*. https://doi.org/10.1186/s12939-022-01652-1

Wikipedia contributors, 2022. Adaptive histogram equalization. URL: https://en.wikipedia.org/w/index.php?title=Adaptive_histogram_equalization. [Online; accessed 19-April-2023].

Wu, C., Ye, Y., Gao, F., & Ye, X. (2023). Using street view images to examine the association between human perceptions of locale and urban vitality in Shenzhen, China. *Sustainable Cities and Society, 88*, Article 104291.

Yang, M., & Hu, B. (2022). The impact of city cluster development on the inter-city disparity: Evidence from China. *Chinese Journal of Urban and Environmental Studies*. https://doi.org/10.1142/s2345748122500063

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications, 11*. https://doi.org/10.1038/s41467-020-16185-w

Yu, D., & Fang, C. (2023). Urban remote sensing with spatial big data: A review and renewed perspective of urban studies in recent decades. *Remote Sensing, 15*, 1307.

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

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*, 1452–1464. https://doi.org/10.1109/tpami.2017.2723009

Zuiderveld, K., 1994. Contrast limited adaptive histogram equalization. Academic Press Professional, Inc., USA. p. 474–485.