Street view imagery: AI-based analysis method and application

Haoyang Song

School of Geographic Sciences, East China Normal University, Shanghai, China

10213903427@stu.ecnu.edu.cn

Abstract. Street view imagery is an emerging form of geographic big data. It presents urban visual environments from the perspective of urban residents and also contains non-visual environment of cities, such as urban human activities and socio-economic development. However, traditional digital image processing has its limitations, and the continuous development of artificial intelligence, especially computer vision and deep learning, provides strong technical support for exploring the rich semantic information in street view imagery. This paper reviews the related research on street view imagery and its artificial intelligence analysis methods and applications. It outlines the acquisition, storage, and common data sources of street view imagery. Then it introduces computer vision, deep learning, and commonly used open-source datasets in street view imagery analysis. It also detailed three aspects of AI-based street view imagery applications, namely quantification of the physical space, urban perception, and spatial semantic speculation. Finally, issues like data acquisition, domain adaption and deep learning black box are discussed. The hotspots and prospects for the development of this research topic are also prospected.

Keywords: Street View Imagery, Computer Vision, Deep Learning.

1. Introduction

With the rise of big data and advancements in information technology, high-precision satellite navigation and positioning technology, as well as map services, are gradually becoming more popular in the field of geographic information science, thus giving rise to a novel type of geographic big data: street view imagery. It is based on the perspective of urban residents, enabling users to remotely explore the real natural and human environment of the city [1]. Initially, street view imagery was used for quantitative assessments of the urban visual environment. However, in recent years, more and more studies on street view imagery have proved its ability to implicitly express the non-visual environment of cities, such as urban human activities and socio-economic development.

Traditional urban research methods have certain limitations in the context of continuous urban expansion and development: research based on remote sensing images do have the advantage of large scale, but it cannot be applied to detailed analysis within the city [2]. There are also numerous studies relied on small-scale localized surveys and measurements. These methods are time-consuming, inefficient, and they cannot be extended to large-scale urban areas. Therefore, street view imagery big data has become an important data source for comprehensive, holistic, and refined quantitative analysis of cities.

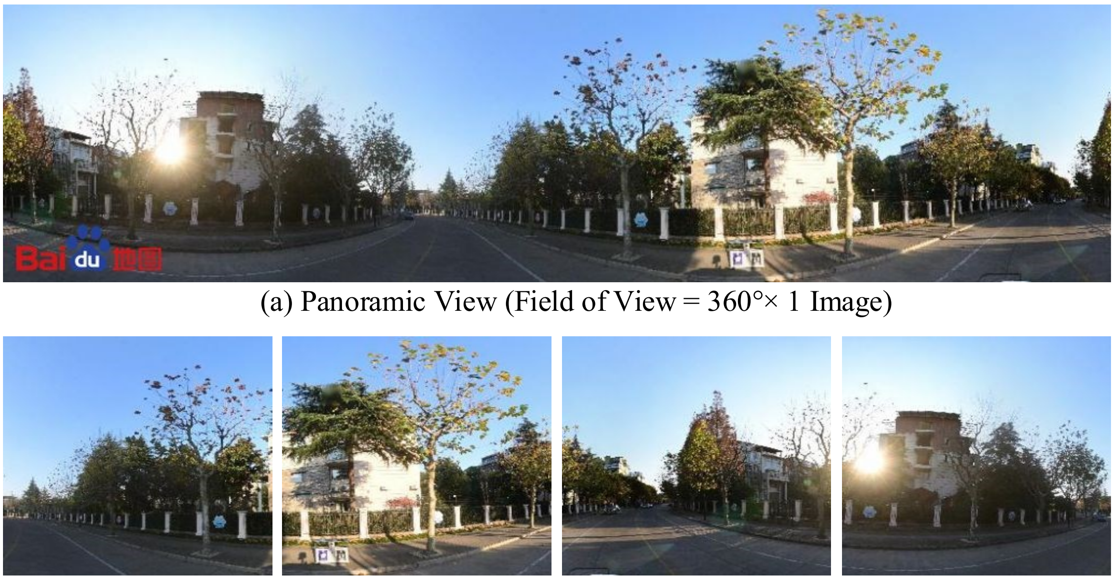
In recent years, artificial intelligence has been developing and evolving rapidly, and has made remarkable achievements in various fields as natural language processing (NLP), automatic driving, medical diagnosis, etc. The outstanding performance of AI in image understanding represented by computer vision and deep learning provides strong technical support for large-scale and intelligent analysis of street view imageries. This article focuses on artificial intelligence and street view imagery analysis. Firstly, it introduces the storage and acquisition of street view imageries and compares different data sources. Then it presents the methods of artificial intelligence in street view imagery analysis, along with commonly used open-source datasets. Then it summarizes the application directions of street view imagery based on artificial intelligence. Finally, it discusses the challenges and opportunities for the future development.

2. Street View Imagery

2.1. Street View Imagery Acquisition & Storage

Street view imagery, as a kind of widely used geo-big data, is mostly offered by map service providers worldwide. They drive street view sampling vehicles equipped with multiple cameras, sensors, including Lidar and other high-tech equipment on roads in the city in order to obtain street view imageries. In addition, areas that are inaccessible to automobiles can also be photographed by backpacks (such as the LiBackpack) equipped with cameras and Lidar.

Street view imagery is generally stored in the form of panoramas, and it contains 360° panoramic visual information around photographing center. In practical use or research, a 360° panorama is often split into several images with different horizontal views (e.g., front, back, left, and right view). Figure 1(a) and 1(b) respectively show the panoramic view and its corresponding normal view of a location in the Pudong New District of Shanghai, China, which are obtained through the Baidu Street View API.



(a) Panoramic View (Field of View = 360°× 1 Image)

(b) Normal View (Field of View = 90° × 4 Images)

Figure 1. Street view imagery of Pudong New District, Shanghai, China with different field of view (fov).

2.2. Major Sources of Street View Imagery

Currently, there are numerous companies around the globe that provide street view imagery services. The most representative ones are Google street view (GSV), Baidu street view (BSV) and Tencent street view (TSV). Table 1 shows the coverage and maximum image resolution of these three street view imagery services.

Table 1. Major Street View Imagery Services Provider.

Service Provider	Coverage	Maximum Resolution (Width × Height)
Google street view (GSV)	More than 100 countries	2048 × 2048
Baidu street view (BSV)	Only China	1024 × 512
Tencent street view (TSV)	Only China	1680 × 1200

GSV is the most popular street view service provided by Google, an American multinational corporation that operates globally. It covers more than 100 countries in the world and introduced indoor services in 2017. However, GSV is not available in countries like China. In China, there are two main street view services, namely BSV and TSV, relatively provided by two major internet companies Baidu and Tencent. Therefore, many researches focusing on street view imagery in China relies on BSV and TSV as data sources. Compared of BSV and TSV, BSV has a larger coverage and as of August 2023, BSV covers more than 80% of the cities in China, totaling 652 cities. All these street view services can be accessed through either the website or the API interface. However, it is worth noting that the API interface does not support accessing the historical street view imageries, and the resolution of the retrieved imagery is relatively low.

3. AI-based Street View Imagery Analysis Method

AI-based street view imagery analysis method applies both deep learning and computer vision into street view imagery processing and urban-oriented application practice. Most digital image processing and traditional computer vision approaches struggle to convey deep semantic information in street view imagery efficiently and completely. However, the integration of deep learning and computer vision can more effectively and precisely distinguish semantic items and scene elements in street view imageries, offering a powerful tool for extracting street view semantic information and analyzing urban settings [2].

Computer vision seeks information from images or high dimensional data by using imaging equipment and computers to recognize and measure objects instead of human eyes. In street view imagery, computer vision plays a pivotal role in analyzing and interpreting visual aspects of urban environments. However, traditional computer vision methods, although effective in many scenarios, often require the use of manually designed feature extraction algorithms. These algorithms, for example, Generalized Search Trees (GIST), Haar-like Features (HARR), Scale Invariant Feature Transform (SIFT), and Histogram of Oriented Gradient (HOG), may not fully capture the complexities and nuances of street scenes.

The introduction of the AlexNet in 2012 marked a significant breakthrough in computer vision, particularly in street view imagery analysis. AlexNet demonstrated the power of deep learning in automatically learning relevant features from high-dimensional data, such as street view imageries. It allowed the network to autonomously extract hierarchical features directly from raw pixel data, enabling better understanding and analysis of street views.

As a digital representation of urban space, street view imageries contain rich geographic, architectural, and social information. In the field of street view imagery analysis, Convolutional Neural Network (CNN) shows great potential in interpreting and understanding urban landscapes. GoogLeNet, with its multi-scale feature extraction capability, adapts well to the various scales and levels of geographic elements present in streetscape images. ResNet, with its residual connectivity, can dig deeper into the intricate structure of cityscapes, making it helpful in tasks like target detection and building identification. VGG, known for its layer-by-layer abstraction structure, facilitates the transformation of urban elements into abstract features within street view images. On the other hand, DenseNet lays emphasis on the dense transfer of features, which plays a crucial role in exploring the rich contextual relationships within street view images.

4. Major Open-Source Datasets Used in Street View Imagery Analysis

Despite the structure of deep learning, the training set has a substantial influence on both the model's generalization capacity and the number of recognizable categories. In the field of street view imagery analysis, the main datasets used for urban scene understanding are shown in Table 2.

Table 2. Major training sets used in street view imagery analysis.

Datasets	Objects	Categories
Cityscapes	25,000 images (5,000 with high quality pixel-level annotations)	30
ADE20K	Objects: 25,000 images	150
Mapillary Vistas	25,000 high resolution images	66
Places	10 million scene photographs	434
ImageNet	More than 14 million images	1000
Microsoft COCO	328,000 images	80

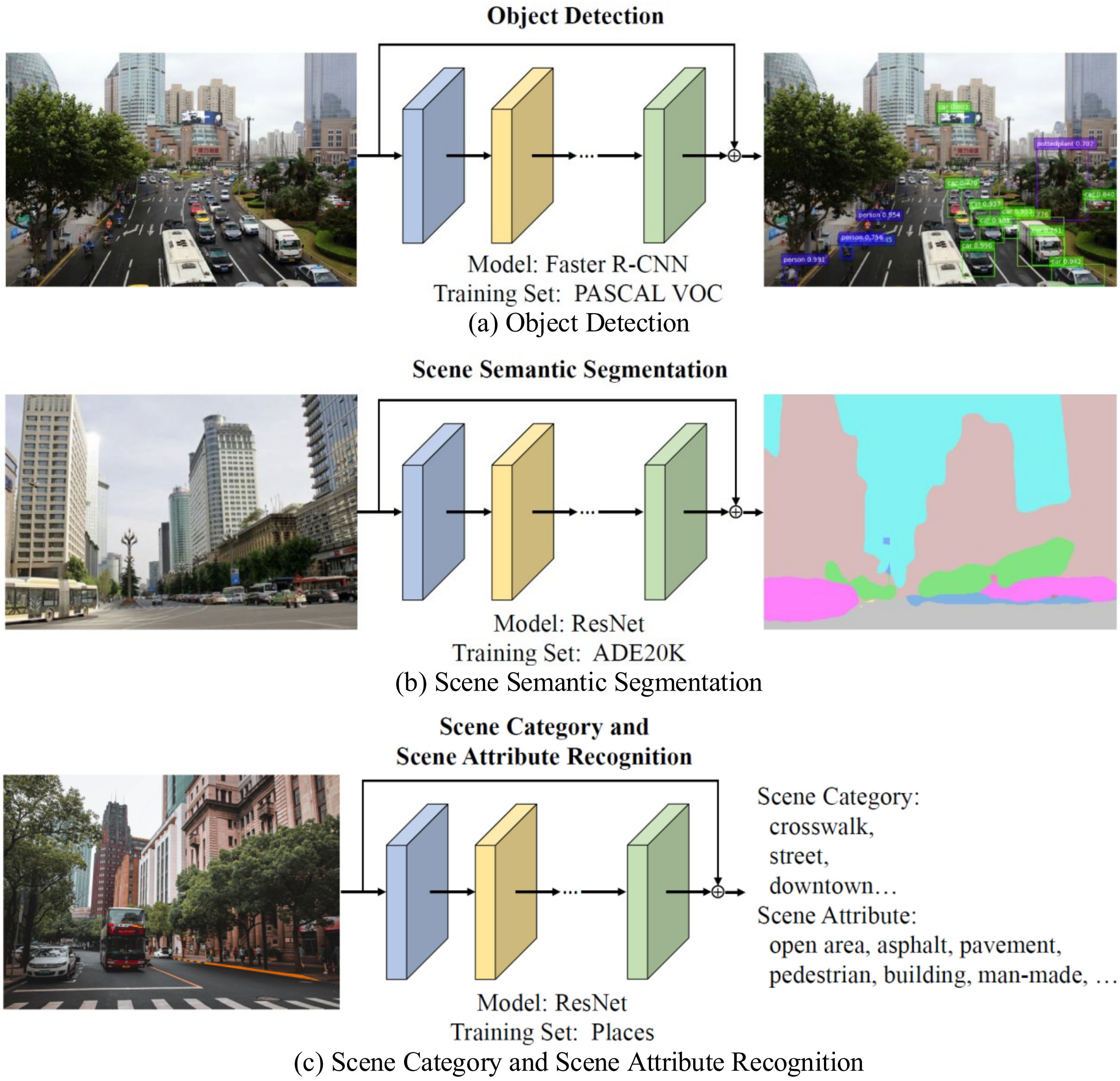
The Cityscapes dataset is a widely used open-source dataset for semantic segmentation studies of street view imagery. It contains a large number of high-resolution images covering various aspects of urban streets, including roads, buildings, vehicles, pedestrians, etc. By pro viding pixel-level annotations in the images, the Cityscapes dataset can make the deep learning model understand the semantic information in street view imageries more accurately and efficiently. The ADE20K dataset, on the other hand, is more extensive because it covers a wide range of scenes both indoors and outdoors. The ADE20K dataset covers a wider range of object categories and environmental variations, enabling the model to better understand fine-grained objects and regions in the image. Besides, it also provides pixel-level annotations which makes the result of semantic segmentation more precise and accurate.

The Mapillary Vistas dataset is a sizable collection of street-level images that covers global urban and rural environments. It can help train models for accurate semantic segmentation and help solve the problem of object recognition and understanding in street view imageries. The Places dataset contains images covering a wide range of different scenes from indoor to outdoor, natural landscape to urban street scenes. The extensive scene coverage provided by the Places dataset contributes to enhancing the model's generalization capacity. As for the ImageNet and Microsoft COCO datasets, they are both widely used generalized object recognition datasets that contain a large amount of data. Although they are not specifically designed for street scene analysis, they can also be used for pre-training and transfer learning for tasks like object detection and object segmentation in street view imagery analysis.

5. AI-based Street View Imagery Application

5.1. Quantification of the Physical Space

Quantification of the Physical Space is one of the important applications of AI-based street view imagery analysis. It involves the recognition of visual objects in street view imageries as well as scene classification. In terms of visual object recognition, there are two types of tasks: object detection and object segmentation, as shown in Figure 2(a) and 2(b) respectively. Object detection can locate and classify elements in street scene imageries, such as traffic signs, buildings, cars, pedestrians, etc., while object segmentation can accurately determine the boundaries of each element for every pixel in the street view imagery. Commonly used deep learning models in physical space quantification include Faster Mask Region-based Convolutional Neural Network (Mask R-CNN), Region-based Convolutional Neural Network (Faster R-CNN), Pyramid Scene Parsing Network (PSPNet), Single Shot MultiBox Detector (SSD), etc. In addition, scene classification is also an important task in street view imagery analysis (as shown in Figure 2(c)). The ResNet model trained on scene-oriented datasets (such as Places) is able to classify over 400 scene types and 100 scene attributes [3].



(a) Object Detection

(b) Scene Semantic Segmentation

(c) Scene Category and Scene Attribute Recognition

Figure 2. Methods for physical space quantification using street view imagery.

Therefore, with the help of artificial intelligence methods, we can quantify the elements and their attributes in large-scale street view imageries. Two of the most prominent indicators are the Tree view factor (TVF) and the Green view index (GVI). These indicators are calculated through semantic segmentation methods and can assess the visibility of forests in the city and quantify the urban greenery. For instance, researchers from Senseable City Laboratory in MIT used computer vision to extract greenspace from street view imageries in New York and Boston, USA [4]. Besides, Stubbings et al. used PSPNet to calculate tree coverage along the street network in Cardiff, Wales, and also pointed out its connection with citizens' mental health and well-being [5].

Except greenery, AI-based analysis methods can also be used to calculate the Sky view factor (SVF). Researchers like Gong et al. used PSPNet to extract and calculate SVF in a dataset containing 33,544 GSV images within densely populated urban regions in Hong Kong and validated the results through traditional field surveys [6]. Moreover, there are other studies using GSV images and CNN models to calculate the spatio-temporal coverage of solar radiation [7] and photovoltaic potential in urban streets, and predict the occurrence and locations of solar glare that affect the safety of driving [8].

5.2. Urban Perception

Street view imagery may also be used for urban perception with the aid of deep learning. The relationship between human emotions and the urban environment is a crucial part of urban perception. Early studies primarily relied on interpersonal communication to investigate human emotions towards cities. However, this method is limited in spatial scope, data volume, and susceptible to human biases. The emergence of deep learning and street view big data provides new analysis methods and data sources. Firstly, object detection is used to target human faces within massive street view imageries, then emotion recognition and analysis are performed with the help of deep learning models, and finally human emotions are linked back to the streetscape environment. In order to construct a better comprehensive urban perception model, MIT Media Lab released the Place Pulse dataset containing perception scores by engaging numerous volunteers in scoring the emotional perception of street view imageries. Zhang et al. utilized GSV images and applied a Deep Convolutional Neural Network (DCNN) model trained on Place Pulse to extract image features and predict human perceptions, including six dimensions: safety, liveliness, beauty, wealth, depression, and boredom [9].

In addition, street view imagery can be utilized for perceiving the urban environment, especially in the field of community security and public health. In terms of community safety, Amiruzzaman et al. used PSPNet to extract semantic categories from GSV images and predicted high and low crime rate areas in US cities with the accuracy of more than 95% [10]. Besides, on this basis, Zhang et al. used street view imageries of Huston to calculate the perception of safety score by using a pre-trained scene understanding model and generated the perception of safety map (as shown in Figure 3). He also matched the result with the crime rates and proposed the concept of perception bias [11]. In terms of public health, Nguyen et al. used VGG-19 in Tenserflow to extract street greenness, crosswalks and building types of GSV images in American cities and with the help of statistical model, showing a significant correlation between street elements and health issues, like obesity and diabetes [12]. Further researches include using street view imagery to perceive the psychological well-being and happiness of urban residents. Furthermore, during the pandemic, street view imageries and CNN model pre-trained on ImageNet can evaluate the COVID-19 risks in different regions of the city by detecting urban architectural features in street view imageries [13]. Apart from COVID-19, similar analytical methods are also be used for other highly contagious diseases like dengue and influenza.

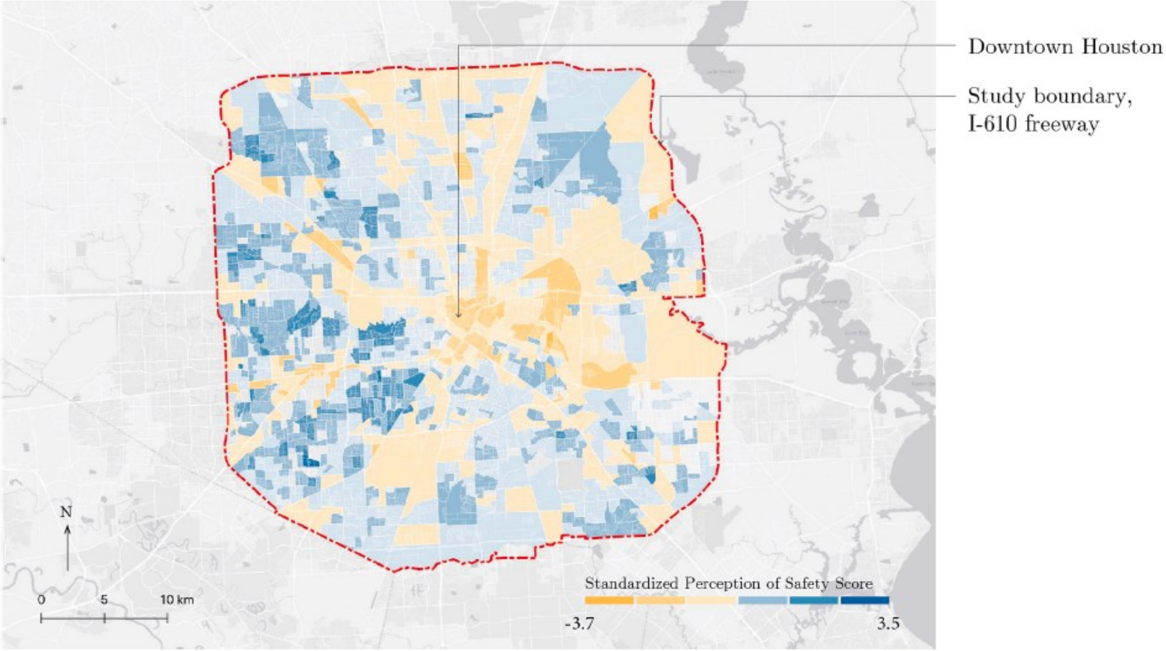


Figure 3. Perception of safety map of Houston, USA. Yellow represents places with the relatively lower safety scores, while Blue represents places with relatively higher safety scores. [11]

5.3. Spatial Semantic Speculation

Street view imageries not only depict the visual information of urban streets, but also convey the non-visual information of the city, such as the city's culture, history, function, socio-economic conditions, and human activities implicitly. For example, Ma et al. used ResNet-101 Model to detect the font types of stores in street view imageries of London. They found that font types can serve as alternative indicators for evaluating urban economy and population [14]. Gabru et al. inferred the income, ethnicity, and political inclination of neighborhoods by identifying 22 million cars detected in GSV images in America [15]. There are also other studies using street view imageries to predict house price, crime rates, etc. These applications benefit from the large amount of street view data and its rich information within, as well as the artificial intelligence like deep learning, which is a powerful tool to mine the deep semantic information inside street view imageries.

6. Challenges and Opportunities

6.1. Challenges

6.1.1. Street View Imagery Acquisition. The difficulty and cost of downloading street view imageries are continuously increasing. For example, BSV API requires applying for additional permissions to use, and the daily API call limit is only 100 times. Second is the issue of image quality. Due to the extensive workload of capturing street view imageries, some of them inevitably have problems such as poor lighting, blurriness, and unfavorable weather condition. In addition, obstacles such as passing cars and pedestrians often appear in street view imageries which affects the analysis results. Third, street view service providers are unable to guarantee the timeliness of street view imageries. For instance, on average, each location in Chinese cities has only two images taken at different times from the BSV [2].

6.1.2. Domain Adaption. In terms of today's AI-based street view imagery analysis methods, researchers mainly use computer vision and deep learning models for the segmentation and analysis of street scenes. This method can recognize the semantic information of street view imageries to a certain extent. However, there are differences between the training set of pre-trained models and the actual test imageries, which leads to domain adaptation problems and causes the decrease of model accuracy.

6.1.3. Deep Learning Black-box. Furthermore, utilizing deep learning for the analysis of street view imageries, such as ResNet trained on the Places dataset, can yield 512-dimensional feature vectors extracted from the images [3]. While the vectors can capture the essence of the entire image and perform visual similarity analysis, the deep learning black-box poses challenges in interpreting these vectors. For instance, a model might recognize elements like buildings and vehicles for classification, but the concrete interpretation of internal weights and patterns within the model remains non-intuitive. Some methods such as heatmaps can assist in visualizing the model's region of interest, yet they still provide only approximate explanations.

6.2. Opportunities

Data Sources: Crowdsourcing and Autonomous Driving Data In terms of data, apart from traditional street view service providers, there are numerous street view photos and videos contributed by users worldwide on crowdsourcing platforms like OpenStreetCam and Mapillary, which helps solve the spatiotemporal coverage challenges in street view imagery. Meanwhile, a large amount of high-precision road measurement data and visual information is needed in autonomous driving and this series of data can also be utilized to address issues like image quality, resolution, and obstacles in street view research.

6.2.1. Generative Adversarial Network (GAN) in Street View Imagery Analysis. In terms of deep learning methods, GAN is usually used for image generation, image restoration and enhancement, and style transformation. In street view imagery analysis, for example, Joglekar et al. developed FaceLift by using GAN and it is mainly used for street view imagery beautification and style transformation [16]. However, there are far fewer GAN-based studies in street view imagery analysis than CNN-based studies, which may offer more research prospects.

7. Conclusion

Before the emergence of street view imagery, there has long been a lack of a cost-effective and efficient approach to evaluate the urban visual environment and non-visual information from a human perspective. However, street view imagery services like GSV, BSV, TSV act as an important street view database and lay a solid foundation for urban analysis. Simultaneously, the continuous development artificial intelligence has provided technical support for large scale and high efficiency extraction and analysis of rich semantic information from street view big data.

In terms of specific application, this paper introduces the AI-based street view imagery application from three aspects of quantification of the physical space, urban perception, and spatial semantic speculation. By employing deep learning models such as Mask R-CNN, Faster R-CNN, PSPNet, and SSD, it becomes possible to quantitatively calculate urban elements such as greenery, sky view, solar radiation, solar glare, etc. By combining human perception with the urban environment, urban perception is enhanced, thus providing a deeper understanding in areas like community security and public health. Moreover, the potential of street view imagery in spatial semantic speculation has also been explored. It can be used to predict urban economic indicators, social factors, and urban attributes.

Despite challenges such as data access barriers, domain adaptation, and interpretability of deep learning models, opportunities continue to emerge. Crowdsourcing and autonomous driving data offer additional data sources, and GAN also provides new avenues for further studies. As technology iterates, AI-based street view imagery method and application is expected to drive deeper insights into the dynamics of urban environments, paving the way for building smarter cities.

References

[1] Li Y C, Peng L, Wu C W and Zhang J Z 2002 Street View Imagery (SVI) in the Built Environment: A Theoretical and Systematic Review. *Buildings-Basel* **12**(8) 1167

[2] Zhang F and Liu Y 2021 Street view imagery:Methods and applications based on artificial intelligence. *Nrsb* **25**(5) 1043-1054

[3] Zhou B L, Lapedriza A, Khosla A, Oliva A and Torralba A 2018 Places: A 10 Million Image Database for Scene Recognition. *Tpami* **40**(6) 1452-1464

[4] Seiferling I, Naik N, Ratti C and Proulx R 2017 Green streets - Quantifying and mapping urban trees with street-level imagery and computer vision. *Landscape urban plan* **165** 93-101

[5] Stubbing P, Peskett J, Rowe F and Arribas-Bel D 2019 A Hierarchical Urban Forest Index Using Street-Level Imagery and Deep Learning. *Remote Sens-Basel* **11**(12) 1395

[6] Gong F Y, Zeng Z C, Zhang F, Li X J, Ng E and Norford L K 2018 Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Build environ* **134** 155-167

[7] Li X J and Ratti C 2019 Mapping the spatio-temporal distribution of solar radiation within street canyons of Boston using Google Street View panoramas and building height model. *Landscape urban plan* **191** 103387

[8] Li X J, Cai B Y, Qiu W S, Zhao J H and Ratti C 2019 A novel method for predicting and mapping the occurrence of sun glare using Google Street View. *Transport res c-emer* **106** 132-144

[9] Zhang F, Zhou B L, Liu L, Liu Y, Fung H H, Lin H and Ratti C 2018 Measuring human perceptions of a large-scale urban region using machine learning. *Landscape urban plan* **180** 148-160

[10] Amiruzzaman M, Curtis A, Zhao Y, Jamonnak S and Ye X Y 2021 Classifying crime places by neighborhood visual appearance and police geonarratives: a machine learning approach. *Jcss* **4**(2) 813-837

[11] Zhang F, Fan Z Y, Kang Y H, Hu Y J and Ratti C 2021 “Perception bias”: Deciphering a mismatch between urban crime and perception of safety. *Landscape urban plan* **207** 104003

[12] Nguyen Q C, Sajjadi M, McCullough M, Pham M, Nguyen T T, Yu W J, Meng H W, Wen M, Li F F, Smith K R, Brunisholz K and Tasdizen T 2018 Neighbourhood looking glass: 360 degrees automated characterisation of the built environment for neighbourhood effects research. *Jech* **72**(3) 260-266

[13] Nguyen Q C, Huang Y R, Kumar A, Duan H S, Keralis J M, Dwivedi P, Meng H W, Brunisholz K D, Jay J, Javanmardi M and Tasdizen T 2020 Using 164 Million Google Street View Images to Derive Built Environment Predictors of COVID-19 Cases. *Int J Env Res Pub He* **17**(17) 6359

[14] Ma R X, Wang W, Zhang F, Shim K and Ratti C 2019 Typeface Reveals Spatial Economical Patterns. *Sci Rep-uk* **9** 15946

[15] Gebru T, Krause J, Wang Y L, Chen D Y, Deng J, Aiden E L and Li F F 2017 Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *P natl Acad Sci* **114**(50) 13108-13113

[16] Joglekar S, Quercia D, Redi M, Aiello L M, Kauer T and Sastry N 2020 FaceLift: a transparent deep learning framework to beautify urban scenes. *Roy Soc OpenSci* **7(1)** 190987