




Article

Quantifying the Urban Visual Perception of Chinese Traditional-Style Building with Street View Images

Liyang Zhang ^{1,2,3} , Tao Pei ^{3,4,5,*} , Xi Wang ^{3,4}, Mingbo Wu ^{3,4}, Ci Song ^{3,4} , Sihui Guo ^{3,4} and Yijin Chen ²

¹ College of Information Science and Engineering, China University of Petroleum, Beijing 102249, China; lyzhang1980@cup.edu.cn

² School of Geosciences & Surveying Engineering, China University of Mining & Technology, Beijing 100083, China; y.j.chen@263.net

³ State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; wangxi@lreis.ac.cn (X.W.); wumingbo14@mailsucas.edu.cn (M.W.); songc@lreis.ac.cn (C.S.); guosh@lreis.ac.cn (S.G.)

⁴ University of Chinese Academy of Sciences, Beijing 100049, China

⁵ Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

* Correspondence: peit@lreis.ac.cn; Tel.: +86-010-6488-8960

Received: 20 July 2020; Accepted: 21 August 2020; Published: 28 August 2020



Abstract: As a symbol of Chinese culture, Chinese traditional-style architecture defines the unique characteristics of Chinese cities. The visual qualities and spatial distribution of architecture represent the image of a city, which affects the psychological states of the residents and can induce positive or negative social outcomes. Hence, it is important to study the visual perception of Chinese traditional-style buildings in China. Previous works have been restricted by the lack of data sources and techniques, which were not quantitative and comprehensive. In this paper, we proposed a deep learning model for automatically predicting the presence of Chinese traditional-style buildings and developed two view indicators to quantify the pedestrians' visual perceptions of buildings. Using this model, Chinese traditional-style buildings were automatically segmented in streetscape images within the Fifth Ring Road of Beijing and then the perception of Chinese traditional-style buildings was quantified with two view indicators. This model can also help to automatically predict the perception of Chinese traditional-style buildings for new urban regions in China, and more importantly, the two view indicators provide a new quantitative method for measuring the urban visual perception in street level, which is of great significance for the quantitative research of tourism route and urban planning.

Keywords: urban perception; Chinese traditional-style building; street view images; view indicator; deep learning

1. Introduction

Urban perception has been receiving more and more attention because it plays an important role in urban studies [1–4]. Urban streets, as representations of urban landscapes, determine the qualities of visual perception of urban style, and buildings are the most pronounced elements in urban streets, which shape and articulate the urban style due to their dominant shapes and vivid colors, historical sites, and unique symbols [5,6]. Thus, urban perception studies have been carried out via visual perception of buildings [7,8]. Early urban perception studies were carried out at small scale through a field study of people's perceptions and documentation of urban landscapes, which involved considerable collecting effort [4]. Later, in large-scale urban perception research, urban elements could

be perceived with the help of conventional computer vision techniques; however, these methods faced scalability issues because the feature engineering relied heavily on manual extraction [9]. To address the issues, we proposed a method based on deep learning techniques to study the urban visual perception of Chinese traditional-style building with street view images.

We propose a deep learning network model based on convolutional neural networks and transfer learning that can automatically extract Chinese traditional-style architecture from thousands of street view images with pixel-level semantic segmentation. Then, based on the pixel-level image segmentation, we developed the traditional building view index to quantify pedestrians' visual perceptions of Chinese traditional-style buildings in large areas. In this research, we used Beijing as a case study. Beijing is ancient cultural capital with rich historical sites. In addition, Beijing is one of the most rapidly urbanizing cities in China, and modern buildings and historical sites have been mixed in the process of urban transformation, which allow testing of the effectiveness of the method and contribute to the planning and protection of the architecture of Beijing [10]. The main contributions of our study are as follows: (1) construction of categories of Chinese traditional-style buildings to train datasets and testing and verifying the method with Tencent street view (TSV) images; (2) suggestion of the model traditional building mask region-based convolutional neural networks (TBMask R-CNN) to achieve automatic classification and segmentation of Chinese traditional-style buildings; (3) development of the traditional building view index (TBVI) and the street of traditional building view index (STBVI) to quantify pedestrians' visual perceptions of Chinese traditional-style buildings; and (4) calculation of the TBVI and STBVI values within the Fifth Ring Road of Beijing and analysis of the spatial distribution of the TBVI and STBVI of traditional-style buildings.

2. Related Works

In this section, we review key works in urban perception about data sources and techniques, such as urban perception with street view images and urban perception with deep learning.

2.1. Urban Perception with Street View Images

Street views are interactive electronic maps that can provide 360° panoramas along urban streets; therefore, the street view images from these maps can be regarded as representations of urban landscapes. Currently, Google, Tencent and Baidu are the three mapping service providers of street view images in the world [5,11,12]. Street view images have become an important new data source in urban studies due to their high accessibility, high resolution and broad coverage [2,13,14]. In recent years, researchers have exploited these images to study urban perception with a certain urban element or synthesis of the perception results of several urban elements [15]. Generally, the perceived urban elements are mainly divided into five categories: buildings, green plants, roads, traffic signs, and pedestrians. The perception of a certain urban element or synthesis of the perception results of several urban elements can be used to study urban perception. For instance, the urban physical environment can be assessed by perceiving urban elements such as buildings, green plants, roads, etc. from Google Street View images, and the results are generally accordant with the field assessments [16–20]. By perceiving trees and grass using street view images, a series of quantitative methods for urban greenery were developed [21–23]. Through analysis of the perception of urban elements such as road suitability and green vegetation, urban safety made great progress at the street level in large areas [3,24,25]. The above studies show that street view is a very reliable data source for urban studies; however, for large-scale quantitative urban perception studies, processing large numbers of images is very challenging.

2.2. Urban Perception with Deep Learning

The traditional methodology for studying urban perception was carried out on small scales through field studies; photographs or videotapes were manually reviewed and urban residents were interviewed due to inadequate technologies, which made it difficult to analyze urban perception

in a quantitative manner and thus became a bottleneck for urban planning based on quantitative studies [5,26]. Later, in large-scale urban perception studies, approaches based on conventional computer vision techniques were developed, such as the histogram of oriented gradient (HOG) descriptor [27], scale invariant feature transform (SIFT) method [28], and edge detection methods with different filters for image processing; however, feature engineering relies on manual extraction and then manual coding based on the domain and data types, which is both difficult and expensive [9].

Deep learning has the ability to automatically learn hierarchical features and has proven its potential for processing large image datasets with near-human accuracy in image classification and pattern recognition [29,30], which provides the possibility of large-scale automatic urban perception studies. Convolutional neural networks (CNNs) are one of the deep learning techniques that have made breakthroughs in image processing and are being explored by urban researchers to study urban perception. Ref. [3] introduced CNN to predict urban safety using Google Street View images, which significantly improved prediction accuracy compared to previous methods. Ref. [2] used a CNN-based model to study the relationship between the visual attributes of cities and urban perceptions of safety and liveliness at the global scale. Ref. [31] trained the places convolutional neural network (CNN) to extract features from 0.2 million images to perceive beauty in outdoor spaces. Ref. [32] used SegNet, a semantic segmentation model-based CNN, to extract urban features, such as trees, sky, roads, buildings, from Google Street View images and then compared the differences in urban perception in four cities using street view elements. To our knowledge, the existing deep learning network models can extract buildings; however, they cannot realize the fine-grained segmentation of building subclasses, such as those of Chinese traditional-style buildings.

3. Study Area and Dataset

3.1. Study Area

Beijing is a world-famous historical and cultural city that has undergone dramatic transformation from being the capital of China's six dynasties to the administrative center and, even at present, to China's political, cultural and economic center. Beijing has many historical sites and traditional Chinese architecture, which made it a powerful case study area for our research. We focused on the area within the Fifth Ring Road, which includes the traditional Chinese landscape area with the Old City as the core, and the area from Old City to the Fifth Ring Road (Figure 1). The study region has a total area of 668.4 km² and is home to approximately 10.54 million people [33].

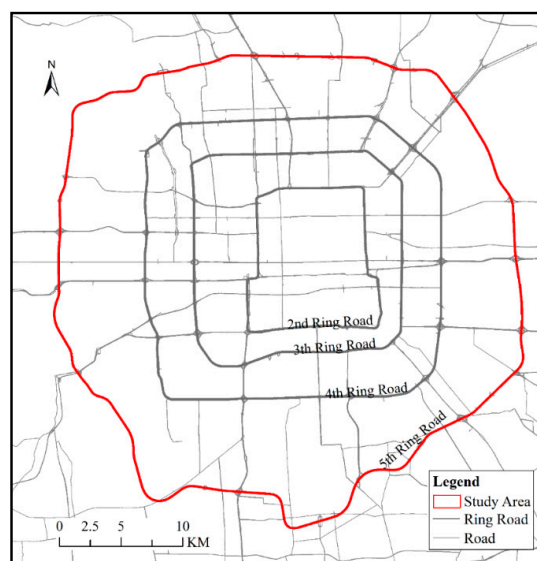


Figure 1. Study area.

3.2. Data Acquisition

We used street view images obtained from Tencent Maps, which provides an application programming interface (API) for querying and downloading street view images with parameters such as size, location/pano, heading, pitch, and key [12]. The TSV panoramas capture photos from horizontal and vertical cameras and stitch them together to produce panoramas of the original street view. The TSV image’s visual fields are approximately equal to those of the normal human visual field when the pitch angle of the camera is 0° [12]; considering that the four images were sufficient to form panoramic images that simulated the 360° visual landscape of a pedestrian at a fixed point, we downloaded a total of four images (each with a pitch angle of 0° and headings of 0°, 90°, 180°, and 270°) at a location to represent its panoramic view.

The procedure for obtaining street view images included three steps: (1) obtaining the location coordinates at an interval of 100 m along all the streets in the vector road network within the Fifth Ring Road, (2) searching the nearest TSV ID within 50 m based on the Tencent web service API and using the TSV API to download the four TSV images at each location, and (3) cleaning the data and eliminating the location coordinates without obtaining all four TSV images.

For detailed instructions on the TSV API, see reference [34]. An example of the acquisition of TSV images through the TSV API is shown in Figures 2 and 3 with a uniform resource locator (URL) example of the Tencent static image API: <https://apis.map.qq.com/ws/streetview/v1/image?size=960x640&pano=10011512120530135909300&pitch=0&heading=0&key=5MJBZ-NEIE6-QMLSW-MX4KF-5OEDE-C7B4Z>.

Response:

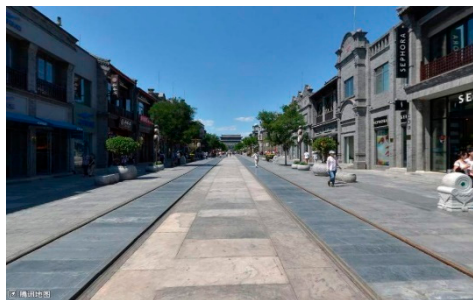


Figure 2. Tencent Street View static image application program interface.

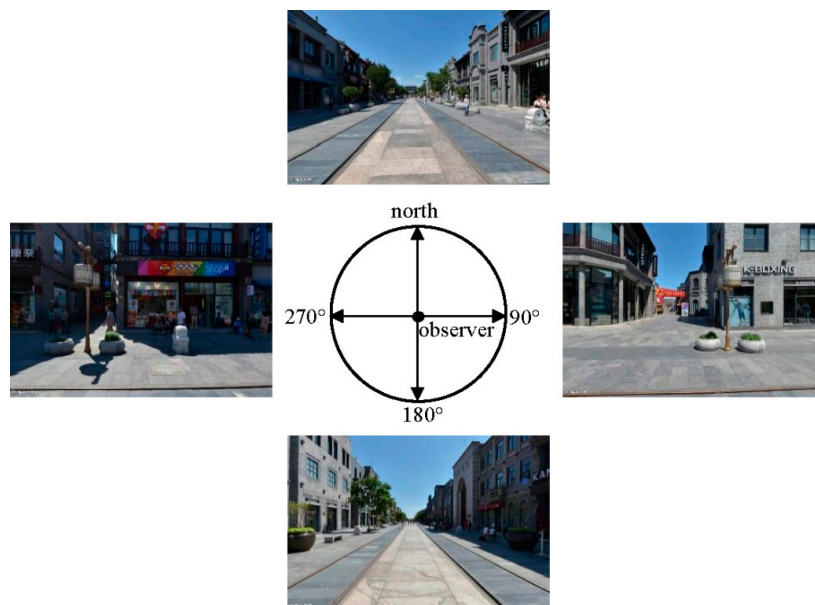


Figure 3. Tencent Street View image download parameter configuration.

In this study, the resolution of the TSV images was 960 × 640, and the format was jpg. After data cleaning, a total of 272,244 TSV images were obtained for 68,061 street points with streetscape images in four directions. The TSV images in the study area were downloaded during November 2018.

3.3. Categories of Chinese Traditional-Style Buildings

Generally, traditional Chinese buildings can be divided into official architecture and folk architecture according to their forms [35,36]. The official building style, which is also known as palace-style architecture, is characterized by high-grade buildings in ancient Chinese architecture that are rigorous, dignified, and elegant with mainly red, yellow, green, and blue colors, which reflect the royal authority and feudal hierarchy. The palace, prince palaces, temple and religious buildings are representative buildings of official style. In contrast with the official building style, there is also folk architecture, which is also known as the local building style. This style is characteristically pleasant, not large in scale, simple in shape, and mainly composed of blue-gray and gray-white colors [37]. Referring to the classification system of traditional Chinese architecture, buildings are divided into three categories: the official-style architecture, folk-style architecture, and nontraditional-style architecture. Among them, official-style and folk-style buildings belong to Chinese traditional style buildings. Table 1 shows the official-style architecture, folk-style architecture, and nontraditional-style architecture buildings.

Table 1. Description of building category.




| Category | Features | Representative Buildings |
|----------------|---|--|
| Official style | The building has a grand scale, luxurious and elegant style, with glazed tiles, red columns, and colored paintings. The color of the building is mainly red, yellow, green, and blue. |  |
| Folk style | The shape of building is simple, and the materials are blue brick, gray tile, and stone. The color of the building is mainly blue-gray and gray-white. |  |

Table 1. Cont.

| Category | Features | Representative Buildings |
|-----------------------|---|--|
| Non-traditional-style | Simple, plain, geometric forms, rectangular shapes, linear elements, and a rejection of ornament, particularly the use of glass, steel and reinforced concrete. |  |

To realize the intelligent recognition of Chinese traditional-style buildings based on deep learning in large regions, we first constructed datasets for the training and evaluation models; 4310 images were selected from the TSV images, and the training set, validation set and test set were constructed with a ratio of 8:1:1. There were 3448 images for the training set, 431 for the validation sets and 431 for the test sets. Among them, the training set included 1148 images with official-style buildings, 1150 images with folk-style buildings, and 1150 images with nontraditional-style buildings. Table 2 shows the description of the datasets. Data annotation was performed with the open source image annotation tool VGG Image Annotator (VIA) [38], which was developed by the Visual Geometry Group.

Table 2. Description of datasets.

| Categories of Building Style | Categories of Datasets | | |
|------------------------------|------------------------|------------|------|
| | Training | Validation | Test |
| Official-style | 1148 | 143 | 143 |
| Folk-style | 1150 | 144 | 144 |
| Non-traditional-style | 1150 | 144 | 144 |
| Total | 3448 | 431 | 431 |

4. Methodology

Here, we introduce the three-step workflow procedure for automatically quantifying the visual perception of Chinese traditional-style buildings on a large scale. Steps 1 and 2 are presented together in Section 4.1: Deep Learning Network for Traditional Style Building Segmentation, which describes our deep learning model that can classify and predict building categories from fine-grained images; Step 3 is introduced in Section 4.2: Traditional Building View Index, which describes the method for quantifying the pedestrians' perceptions of traditional-style buildings based on the predicted building categories generated in Step 2. In Section 4.3, we introduce Moran's I index to evaluate the degree of spatial autocorrelation of the Chinese traditional-style buildings in the study area. In addition, several important abbreviations are defined in Table 3 to make an easier reading of this paper.

Table 3. List of Abbreviations.

| Abbreviation | Full Name |
|--------------|--|
| TSV | Tencent Street View |
| TBMask R-CNN | Traditional Building Mask Region-based Convolutional Neural Networks |
| TBVI | Traditional Building View Index |
| STBVI | Street of Traditional Building View Index |
| RPN | Region proposal network |
| ResNet | Residual network |
| ROI | Region of interest |
| AP | Average precision |
| MAP | Mean average precision |
| O-TBVI | The TBVI of official-style buildings |
| F-TBVI | The TBVI of the folk-style buildings |
| O-STBVI | The STBVI of official-style buildings |
| F-STBVI | The STBVI of the folk-style buildings |

4.1. Deep Learning Network for Traditional-Style Building Segmentation

Image segmentation is a necessary step for extracting pixel-level urban features from street-level images. Recent progress in image segmentation based on deep learning enables researchers to extract objects more accurately. However, the existing deep learning network model cannot obtain the fine-grained segmentation of building subclasses, such as those of Chinese traditional-style buildings. To realize the automatic segmentation of traditional-style buildings, we proposed a deep learning network model: Traditional Building Mask Region-based Convolutional Neural Networks (TBMask R-CNN), which is based on pixel-level semantic segmentation, Mask R-CNN [39] and transfer learning. The framework of TBMask R-CNN is shown in Figure 4. The network architecture of TBMask R-CNN is composed of three modules: the backbone architecture, region proposal network (RPN) and head architecture.

The backbone architecture uses a deep convolution neural network (CNN) to better extract features from each street view images. The residual network (ResNet) is a CNN architecture that revolutionizes the CNN architectural race and devises an efficient methodology for training deep convolution neural networks. ResNet101 is a 101 layers residual network and compared with ResNet50, has a higher accuracy in this experiment. Therefore, we used the convolution layer weights of the ResNet100 pre-training model based on the Microsoft Common Objects in Context (MS COCO) dataset as the backbone architecture of TBMask R-CNN. The convolution kernels of the first and second layers of ResNet101 mainly extract the low-level features of the building, such as edges, colors, pixels, contours, gradients, etc. The third layer extracts the larger and more complex local features of the building, which is derived from the low-level features of the second layer. By analogy, layer by layer corresponds to more and more complex concepts, that is, higher-level networks recognize higher-level semantic features of building categories. Low-level features are universal features, and they are also suitable for traditional style buildings. Therefore, we used the fine-tuned approach based on transfer learning to train our model. The specific steps are as follows: first, the first and second layers' weights of the pre-training model were frozen; then, the traditional-style building training set and test set were used to train the top-level weights of the backbone network to obtain the features of the official-style buildings, folk-style buildings and nontraditional-style buildings.

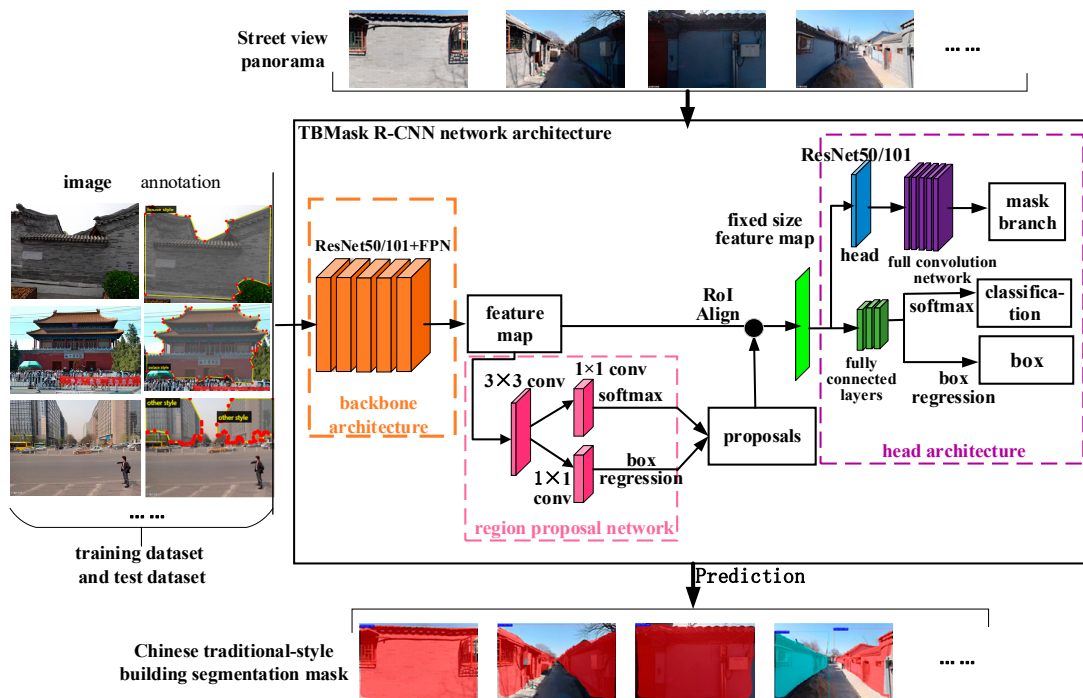


Figure 4. The traditional building mask region-based convolutional neural network (TBMask R-CNN) framework for Chinese traditional-style building segmentation.

The region proposal network was used to sample multiscale regions of interest (ROI) for the head network, which was implemented with a full convolution network to input feature map images and output a set of rectangular target proposals with object scores. The region proposal network uses a 3×3 convolution kernel network to slide on the feature map output by the last shared convolution layer of ResNet101, and maps the features of each sliding window to a low-dimensional feature, which is input to two 1×1 fully connected network layers: box regression layer and box classification layer. ROI align layer finely aligns the multi-scale features extracted by the ROI with the input content, and generates a fixed-size feature map for each ROI through the ROI align layer.

The head architecture computed the bounding box, category, and mask prediction for each ROI, which consisted of three branches: the box prediction, classification prediction, and mask prediction. For box prediction and classification prediction, classification and further box positioning are performed through the fully connected layer; for mask prediction, the head after ROI align is to expand the output dimension of ROI align as the input of the full convolutional network, which is implemented in each ROI based on the pixel-wise prediction, a more accurate segmentation mask is generated.

4.2. Traditional Building View Index

Image classification and segmentation of the buildings is a necessary step in calculating the view index of traditional-style buildings from the pedestrian perspective. Inspired by the green view index, which uses color pictures to evaluate the visibility of the surrounding urban forests as representative of the pedestrians' view of the greenery [40], we developed the traditional building view index (TBVI) to quantify pedestrians' visual perception of Chinese traditional-style buildings by interpreting the street view images. The TBVI is defined as the proportion of Chinese traditional building area in the total area of the building in the normal field of view of the pedestrians.

Based on the building detection and recognition for the panoramic imagery of each location, the TBVI is the ratio of the number of pixels of the Chinese traditional-style building to the total number of building pixels. The TBVI formula is as follows:

$$TBVI = \frac{\sum_{i=1}^m Area_{t_i}}{\sum_{i=1}^m Area_{b_i}} \times 100\% \quad (1)$$

where $Area_{t_i}$ represents the total number of pixels of the Chinese traditional-style building extracted by the TBMask R-CNN model in direction i ; $Area_{b_i}$ is the total number of building pixels in the whole image in direction i ; and the parameter m is the number of images facing different directions in the horizontal orientation of the camera lens; in this study, $m = 4$.

To measure the visibility of Chinese traditional-style buildings at street level, the street of traditional building view index (STBVI) was defined. We provide the STBVI Formula (2) as follows:

$$STBVI = \frac{\sum_{j=1}^n \sum_{i=1}^m Area_{t_ij}}{\sum_{j=1}^n \sum_{i=1}^m Area_{b_ij}} \times 100\% \quad (2)$$

where $Area_{t_ij}$ indicates the total number of traditional-style building pixels extracted from the image by the TBMask R-CNN model in the i direction of location j on the street. $Area_{b_ij}$ is the total number of building pixels extracted from the image by the TBMask R-CNN model in the i direction of location j on the same street. Parameter m is the number of images with different horizontal orientations, and n is the number of locations on the street.

4.3. Moran's I

Global spatial autocorrelation analysis mainly uses the global Moran's I index to reflect whether the pattern expressed is clustered, dispersed, or random in the whole study area. Moran's I is expressed as Formula (3) [41,42]:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_{i=1}^n \sum_{j=1}^n w_{ij}\right) \sum_{i=1}^n (x_i - \bar{x})^2} \quad (3)$$

where n is equal to the total number of regions of the study object space; x_i and x_j represent the attribute value in the i -th and the j -th region, respectively; \bar{x} is the average value of the attribute value of the studied region; w_{ij} represents the spatial weight between the attribute values in the i -th and the j -th region, and i is not equal to j . Moran's I is $[-1, 1]$, in general, a Moran's I value near 1.0 indicates clustering or a stronger spatial positive autocorrelation while a value approaching -1.0 indicates dispersion or a stronger spatial negative autocorrelation. Random distribution exists when the value is closer to zero.

In this paper, TBVI at the street spot location is used as the attribute value of the point to calculate the global Moran's I, which measures the degree of spatial autocorrelation of the Chinese traditional-style buildings in the study area.

5. Results

5.1. Accuracy of Building Segmentation

We used the training set and the validation set to build the TBMask R-CNN model, the training set is used to train the network model to adjust the weights, and the validation set is used to minimize overfitting. The test set is only for testing the trained models in order to confirm the actual predictive power of the network.

Here, two training strategies are designed based on transfer learning. Strategy 1: freeze the convolutional base, which directly uses the model pre-trained on the COCO data set to extract features of the buildings, training set and validation set are only used for the retraining of the pre-trained model's head layer to obtain the classifier of the building categories; Strategy 2: fine-tuning, which trains some layers and leaves others frozen. Here, we freeze the weights of the low-level convolutional layer of the pre-training model ResNet101, and start fine-tuning the parameters from the third layer to obtain the parameters of TBMask R-CNN model.

The test set is used to evaluate the performance of the trained models by two strategies. We use the evaluation performance indicators of the average precision (AP) of each category and the mean average precision (MAP) of all categories. Table 4 present the evaluation performance indicators of the two trained models on test dataset, which show that the second model using Strategy 2 is better than the first model using Strategy 1. With the second model, the MAP was 0.8, AP of the official-style and folk-style buildings were 0.78 and 0.81, respectively, which indicated that the MAP of the segmentation building could reach 0.8, AP of the segmentation of official-style buildings could reach 0.78, and AP segmentation of folk-style buildings could reach 0.81, which indicated that the second model can predict the architectural style categories. Finally, we applied the second model to predict 272,244 TSV images and complete the pixel-level image segmentation, and the results provide the data source for calculating the TBVI of traditional-style buildings from the pedestrian perspective.

Table 4. Average precision (AP) and mean average precision (MAP) of the TBMask R-CNN model on test dataset.

| Strategy | Building Style | AP | MAP |
|------------|-----------------------|------|------|
| Strategy 1 | Official style | 0.53 | 0.54 |
| | Folk-style | 0.55 | |
| | Non-traditional-style | 0.54 | |
| Strategy 2 | Official style | 0.78 | 0.80 |
| | Folk-style | 0.81 | |
| | Non-traditional-style | 0.82 | |

Figure 5 shows the original TSV images and the manual segmentation and TBMask R-CNN segmentation results for one random location on Dafengxiang Hutong within the Second Ring Road. The first and second columns in Figure 5 present the TSV images in the four horizontal directions and the artificial annotation for the buildings, and the last column shows the results identified by using the pixel-level segmentation model (TBMask R-CNN), which had an accuracy of 0.983. The TBVI value of the location was 1.0 according to Formula (1).

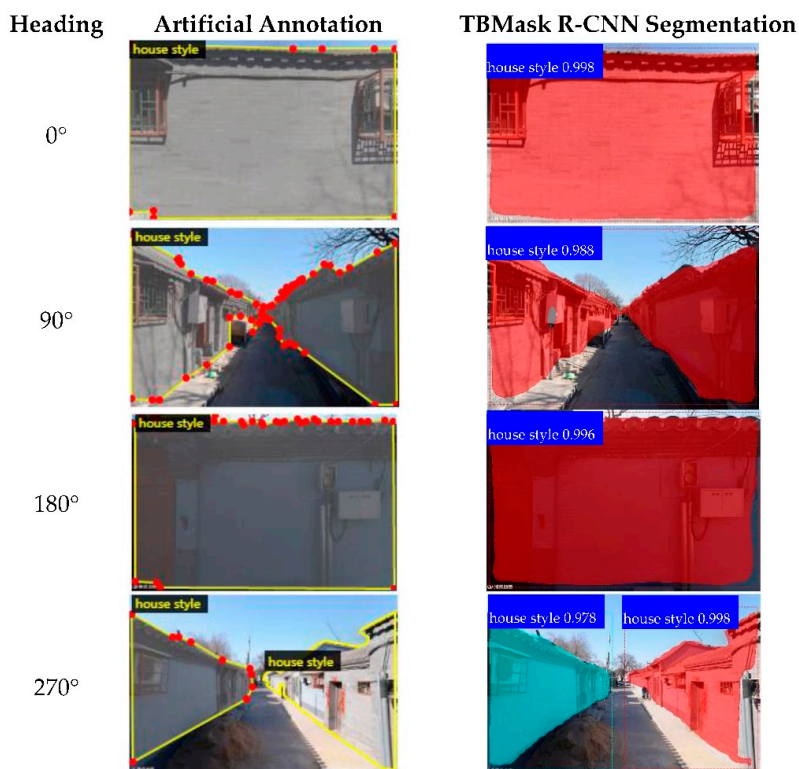


Figure 5. Comparison of the segmentation results (116.381335, 39.937614).

5.2. TBVI Results

According to Formula (1), the TBVI of the official-style buildings (O-TBVI) and the TBVI of the folk-style buildings (F-TBVI) at each location were calculated, and the TBVI of the Chinese traditional-style building (TBVI) was equal to the sum of the O-TBVI and F-TBVI. Figure 6 shows the TBVI results of the Chinese traditional-style buildings for all locations in the study area. The TBVI ranged from 0 to 1, with a mean of 0.137. The red dot indicates the largest TBVI, while the green dot indicates the smallest TBVI. Generally, the values of the TBVI were most pronounced inside the Second Ring Road, first decreasing from the Second Ring Road to the Fourth Ring Road, and then increasing toward the Fifth Ring Road to beyond the Fourth Ring Road. The spatial distribution of the Chinese traditional-style buildings within the Second Ring Road was aggregated overall, and the northern area had a higher concentration than the southern area. However, the spatial distribution was dispersed in the region outside the Second Ring Road.

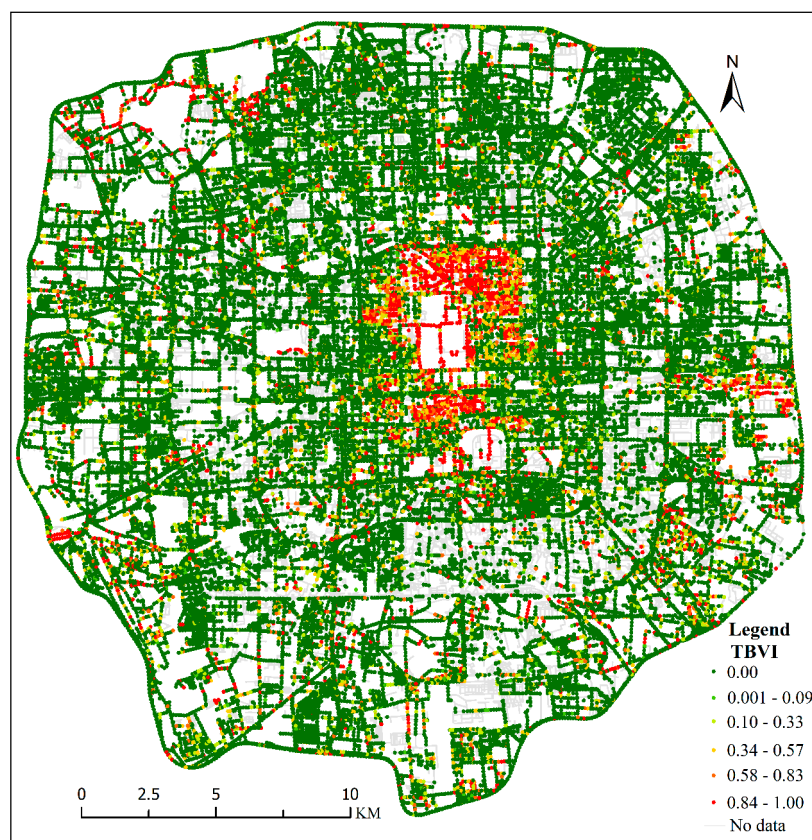


Figure 6. The overall spatial distribution of the traditional building view index (TBVI) of the Chinese traditional-style buildings.

The results of the O-TBVI and F-TBVI were shown in Figures 7 and 8. The Moran's I values of the O-TBVI, F-TBVI and TBVI in the study area are shown in Table 4. In the study area, Moran's I of the Chinese traditional-style building's TBVI, the official-style building's O-TBVI and the folk-style building's F-TBVI were 0.27, 0.09 and 0.28, respectively. The results indicated that the spatial distribution of the Chinese traditional-style buildings and its two subcategories presented aggregation and had a positive spatial correlation, in which the positive spatial correlation of the folk-style was the largest. The Moran's I values of the TBVI within the Second Ring Road, between the Second Ring Road (excluding the Second Ring Road) and Third Ring Road, between the Third Ring Road (excluding the Third Ring Road) and the Fourth Ring Road, and between the Fourth Ring Road (excluding the Fourth Ring Road) and the Fifth Ring Road were 0.47, 0.13, 0.033, and 0.145, respectively, which decreased from the Second Ring Road to the Fourth Ring Road, but the trend toward the Fifth Ring

Road was increasing. The positive spatial correlation of the TBVI was the largest within the Second Ring Road, while it had the smallest positive correlation from the Third Ring Road to the Fourth Ring Road. The spatial correlation patterns of the O-TBVI and F-TBVI were the same as those of the TBVI. In particular, Moran’s I of the O-TBVI was lower than that of the F-TBVI (Table 5), indicating that the Chinese traditional-style buildings that pedestrians perceived from the street level were mainly folk-style buildings.

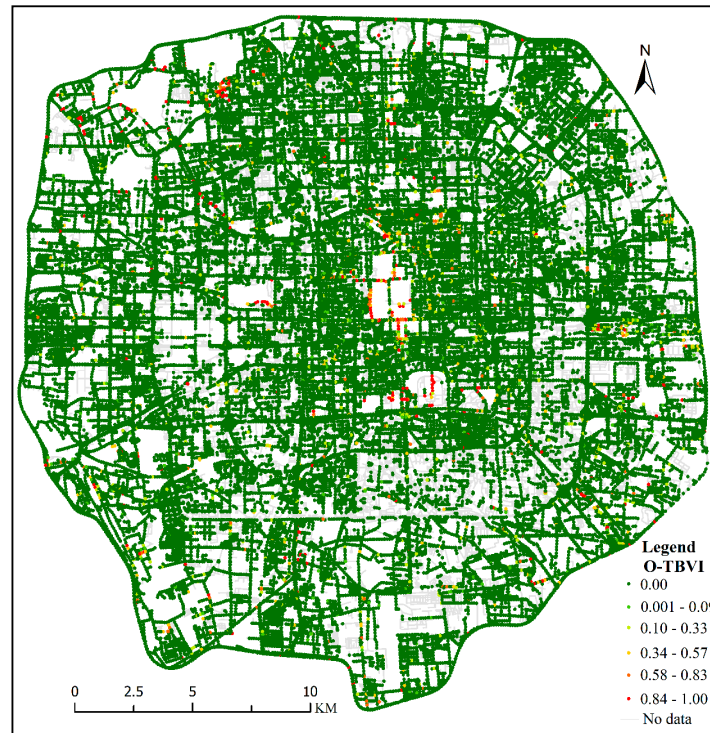


Figure 7. Spatial distribution of the TBVI of the official-style buildings.

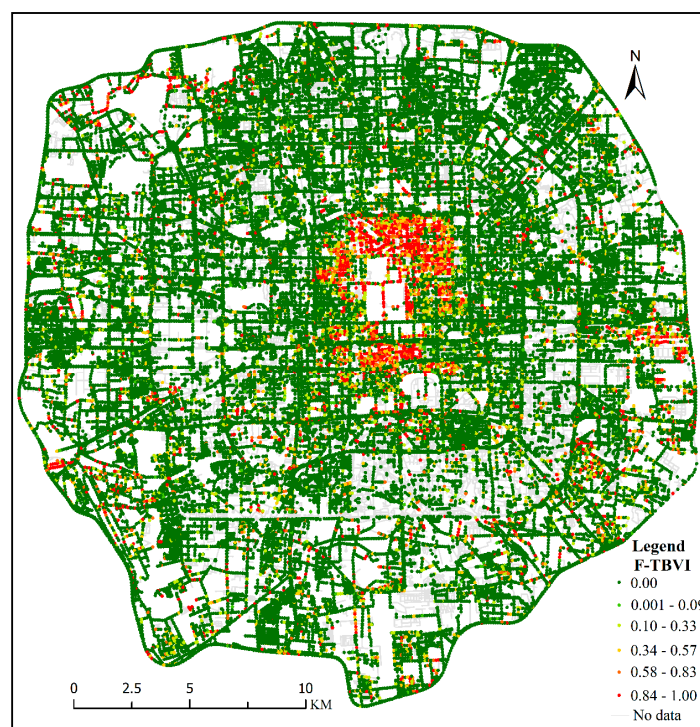


Figure 8. Spatial distribution of the TBVI of the folk-style buildings.

Table 5. Moran’s I of the Chinese traditional-style buildings.

| Area | TBVI | O-TBVI | F-TBVI |
|-------------------------------------|-------|--------|--------|
| Within Second Ring Road | 0.47 | 0.17 | 0.45 |
| Between Second and Third Ring Roads | 0.13 | 0.14 | 0.11 |
| Between Third and Fourth Ring Roads | 0.033 | 0.026 | 0.029 |
| Between Fourth and Fifth Ring Roads | 0.145 | 0.094 | 0.129 |
| Total | 0.27 | 0.09 | 0.28 |

5.3. STBVI Results

According to Formula (2), the STBVI of the official-style buildings (O-STBVI) and the STBVI of the folk-style buildings (F-STBVI) at each street were calculated, and the STBVI of the Chinese traditional-style building (as STBVI) is equal to the sum of O-STBVI and F-STBVI. Figure 9 shows the STBVI of the traditional Chinese building for all streets in the study area. Based on the natural break’s method, the STBVI, O-STBVI and F-STBVI were divided into five intervals: very low (0.00, 0.09), low (0.10, 0.33), medium (0.34, 0.57), high (0.58, 0.83) and very high (0.84, 1.0). From the city center to the Fifth Ring Road in the study area, the streets with very high STBVI values were mainly distributed to the north of the second ring area, such as outside the eastern gate of the Palace Museum, Jingshan Qian Street, Wenjin Street, Houhai Beiyan Street and Gulou West Street (see Marking T1); to the south of the Second Ring Road, such as Yongding Men West Street, Qianian Street, Dongxiao Street, Xiting Hutong Street (see Marking T2); and to the west of the second ring area, such as the roads from the north of the western fourth ring area to the eighth north of the western fourth ring area, Baochan Hutong, Qianmao Hutong and the southern side of Chang’an Street (see marks T3 and T4). In addition, the very high STBVI values were also concentrated in the southeast, northwest, and southwest corners of the fifth ring area (see marks T5, T6 and T7). The spatial distributions of O-STBVI and F-STBVI are shown in Figures 10 and 11. Compared with Figures 10 and 11, it can be seen that the buildings of official-style and folk-style were mainly concentrated in the second ring area, and the spatial distribution of the O-STBVI was obviously sparser than that of the F-STBVI. Official-style buildings were concentrated in the area near the Temple of Heaven and the Palace Museum, and the folk-style buildings were concentrated to the northeast, northwest and south of Chang’an Street in the second ring area. Table 6 shows the mean and percentage of the STBVI. From the table, it can be seen that the mean O-STBVI values from the Second Ring Road to the Fifth Ring Road were 0.018, 0.005, 0.006 and 0.009; the mean F-STBVI values were 0.347, 0.050, 0.047 and 0.067, and the mean STBVI values were 0.365, 0.055, 0.052 and 0.076, respectively. The overall spatial distribution of STBVI was similar to that of TBVI. The streets with high STBVI were densely distributed inside the Second Ring Road and scattered on the fringe around the Fifth Ring Road. Very few streets between the Second Ring Road and Fourth Ring Road had visible traditional-style buildings.

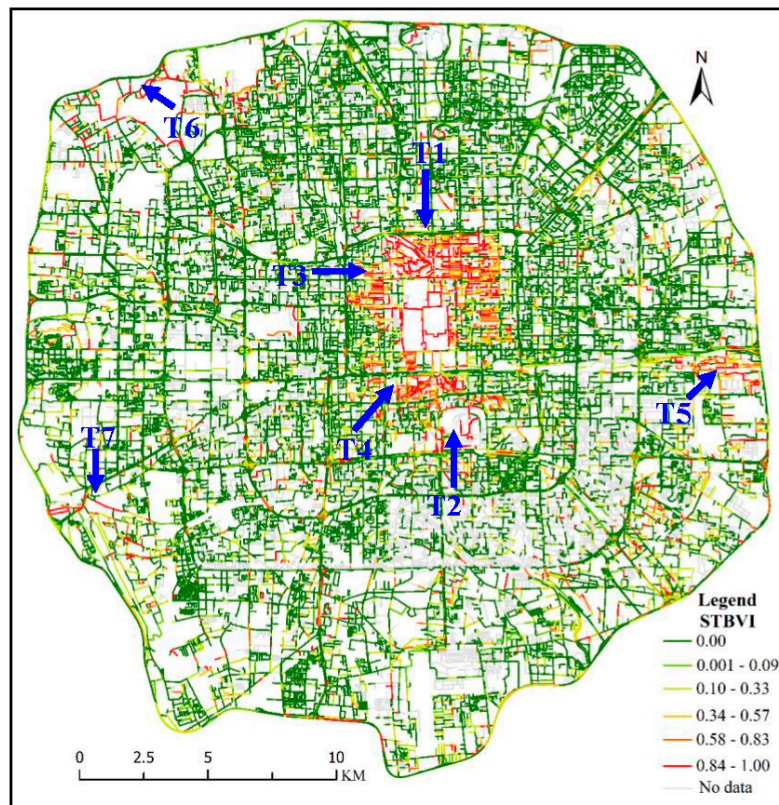


Figure 9. The overall spatial distribution of the street of traditional building view index (STBVI) of the Chinese traditional-style buildings.

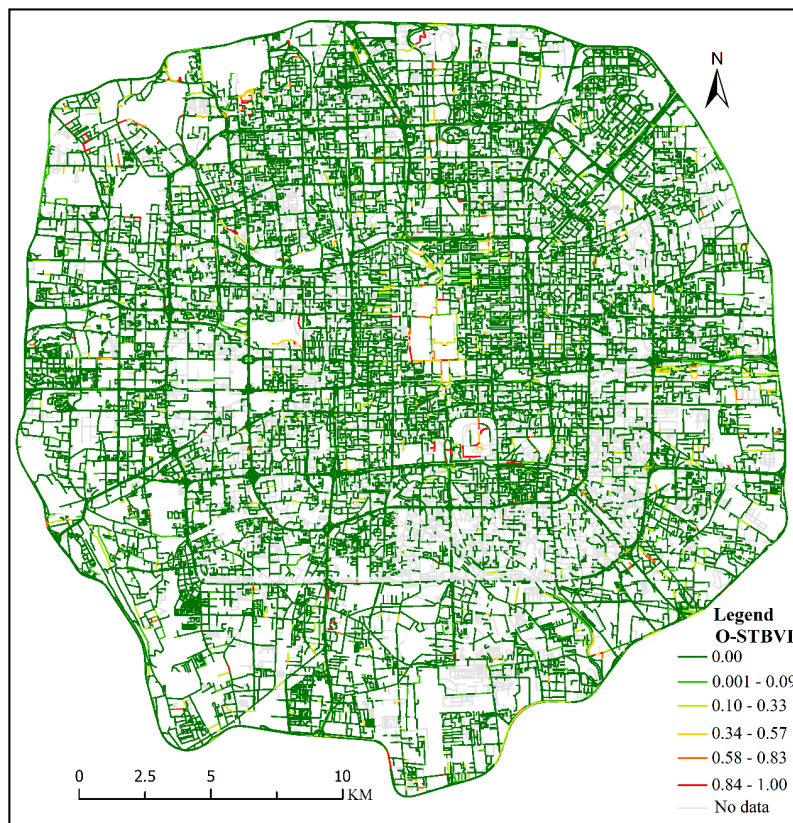


Figure 10. Spatial distribution of the STBVI of the official-style buildings.

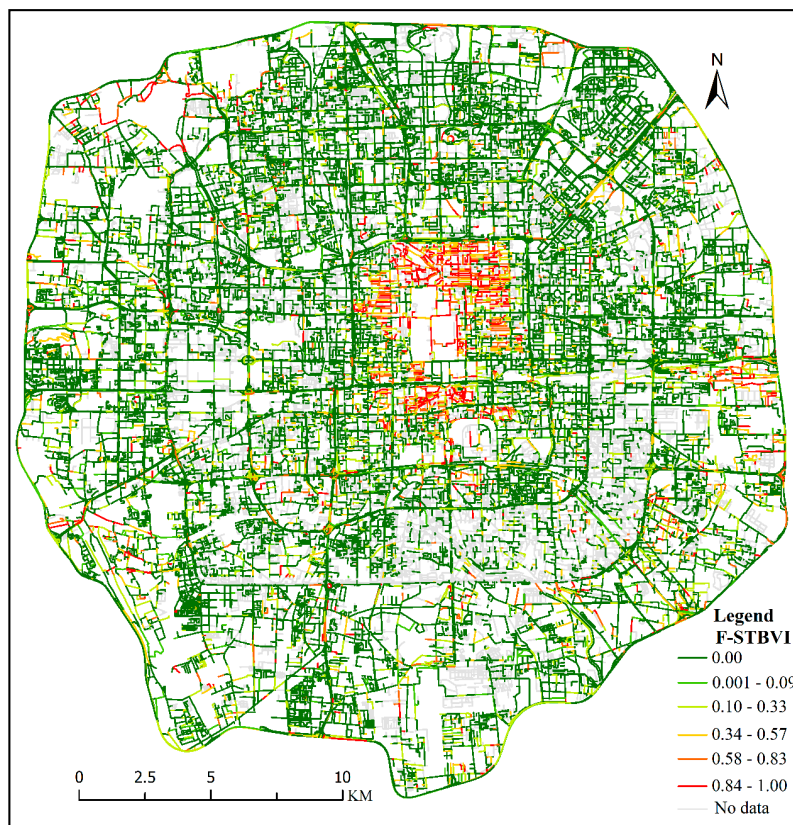


Figure 11. Spatial distribution of the STBVI of the folk-style buildings.

Table 6. STBVI statistics of the Chinese traditional-style buildings and its two subcategories.

| Area | Chinese Traditional-Style | | Subclass 1: Official-Style | | Subclass 2: Folk-Style | |
|--|---------------------------|----------------|----------------------------|----------------|------------------------|----------------|
| | STBVI Mean | Percentage (%) | STBVI Mean | Percentage (%) | STBVI Mean | Percentage (%) |
| Within Second Ring Road | 0.365 | 31.76 | 0.018 | 7.69 | 0.347 | 30.19 |
| Between Second and Third Ring Roads | 0.055 | 14.40 | 0.005 | 3.18 | 0.050 | 12.70 |
| Between Third and Fourth Ringing Roads | 0.052 | 13.81 | 0.006 | 3.77 | 0.047 | 11.79 |
| Between Fourth and Fifth Ring Roads | 0.076 | 17.11 | 0.009 | 5.16 | 0.067 | 14.78 |

6. Discussion and Conclusions

6.1. Discussion on TBVI

The spatial pattern of TBVI can be interpreted from two perspectives. First, the history of urban development determined the location and quality of Chinese traditional-style buildings, which was the basis of future preservation. The area within the Second Ring Road was the Old City, which had existed since the Ming Dynasty. Historically, the Old City was once divided into the South City and the North City. The North City has the Imperial City, while the South City has the Seoul City, where the ordinary artisans and the captured Han people lived [43]. Therefore, the quality of the buildings in the North City is higher than that of the buildings in the South City, and they more likely to be preserved as heritage. In addition, the Summer Palace and Yuanmingyuan, which are at the northwest corner of the Fifth Ring Road, are the royal courts that were built during the Qing Dynasty; to the east of the Fifth Ring Road is the village of Gaobeidian, which was a center for the distribution of the grains transported via the Grand Canal from South China to the capital during the Qing Dynasty.

Second, the development and renewal of modern Beijing based on the historical cities greatly affect the level of preservation of traditional buildings. During this process, there are different development strategies for historical sites. The choice between preservation or reconstruction is heavily influenced by the attitudes toward traditional buildings. During the early years of urban development after the founding of the People's Republic of China, the value of historical districts was not widely recognized, which resulted in massive demolition in the area of active urban construction from the Second Ring Road [43]. As the city expanded further from the Old City, the awareness of the historical value increased, resulting in greater preservation in the later-developed areas near the Fifth Ring Road.

6.2. Discussion on STBVI

In the Old City of Beijing, the STBVI value of the east-west streets was generally higher than that of the north-south streets. This was probably a result of the configuration of hutongs and courtyards, which are the traditional form of residence in Beijing. Due to the long-standing emphasis on orientation in China, each traditional courtyard housing one large family usually extended in the north-south direction, while several courtyards would line up side by side with an alley, called a hutong, connecting their northern and southern entrances from the east to west; several rows of courtyards and hutong were aligned one after another from the south to the north, forming a basic residential unit [44,45]. Because the east-west streets mainly accommodated the entrances while the north-south streets could only see the sidewalls, the traditional buildings could be perceived more easily on the east-west streets.

The streets with high STBVI values were interconnected with some districts inside the Second Ring Road and near the eastern and northwestern Fifth Ring Road, while the streets where mostly buildings of the traditional style were visible were scattered isolated in between. On the one hand, this distribution is a result of the history of Beijing, which is similar to the reason behind the distribution of the TBVI. On the other hand, this difference was partly the outcome of the different conservation strategies. There were several historic conservation districts classified by the planning institute and the city government inside the Second Ring Road, where urban development was required to pay special attention to the overall protection of the traditional style and features [46]. Meanwhile, the streets with high STBVI values scattered between the Second and Fourth Ring Roads were mostly located along historic buildings, for which the protection covers only the building and its surroundings instead of the whole district. Thus, it is implied that the overall protection strategy in the historic conservation districts was effective inside the Second Ring Road and that the interconnected streets could form multiple convenient touring paths and increase the integration of visual perception, contributing to harmony and impressiveness in the local image of the city.

The results provide a reference for the realization of the goal of "forming a landscape control area between the Second and Third Ring Road", which was proposed by the Beijing urban master plan in 2017 [47–49]. According to the requirements for coordination with the ancient capital landscape, it is necessary to strengthen the modern representation of the connotation of traditional-style architectural culture.

6.3. Conclusions

This paper used TSV images with deep learning and transfer learning to construct a deep learning model, TBMASK R-CNN, which automatically extracts Chinese traditional-style architecture from street view images and proposed TBVI and STBVI to quantify pedestrians' visual perceptions of Chinese traditional-style buildings. With the support of TBMASK R-CNN, the images of the Chinese traditional-style buildings within the Fifth Ring Road of Beijing were automatically segmented, and then the TBVI was used to quantify the spatial distribution of perception of Chinese traditional-style buildings from the pedestrian's perspective.

In general, the spatial distribution of the Chinese traditional-style buildings was characterized by aggregation within the Second Ring Road, with a higher concentration of buildings in the north than in the south, and dispersion between the Second Ring Road (excluding the Second Ring Road)

and the Fifth Ring Road. The average value of the TBVI indicated that traditional-style buildings within Beijing's Fifth Ring Road need to be protected compared with other styles of buildings, and the buildings to be planned within the Third Ring Road should consider whether their style will contribute to the Beijing urban planning goals proposed in 2017. The Moran's I of the O-TBVI was lower than that of the F-TBVI, which indicated that the Chinese traditional-style buildings that pedestrians perceived from street view were mainly folk-style buildings. Our work could be useful for urban street management and planning. Specifically, (1) we propose a method that can automatically perceive the urban landscape on a large scale; (2) the proposed indicators can quantitatively evaluate the overall spatial distribution of the urban traditional style, and provide an overview of the spatial distribution of traditional urban features, which helps urban management and planning. For example, location planning issues: when planning such as hotels, shopping malls, public stadiums and other large facilities, considered traditional landscapes as their viewing points; (3) street view images are updated quickly, and our proposed method can be directly applied to new street view images to perceive the city scene, and can help evaluate urban renewal projects with historical street view images. The project's influence on the perception of traditional-style buildings can be easily clarified quantitatively by comparison of the TBVI and STBVI values calculated through street view images before and after the project. All these analyses could contribute to the realization of the Beijing urban planning goal of "forming a landscape control area between the Second and Third Ring Road", which was proposed by the Beijing urban master plan in 2017.

In this experiment, the Chinese traditional-style building recognition model, TBMask R-CNN, had MAP of 0.8 for the verification set, but the building occlusion problem caused by trees and utility poles in the street view images affected the accuracy of the recognition and segmentation of the Chinese traditional-style building images. Therefore, strategies for dealing with occlusion and improving the accuracy of the recognition and segmentation requires further study.

Author Contributions: Conceptualization, T.P., L.Z. and C.S.; methodology, L.Z.; formal analysis, M.W.; data curation, X.W.; writing—original draft preparation, L.Z.; writing—review and editing, T.P., L.Z. and M.W.; visualization, S.G.; supervision, Y.C.; funding acquisition, T.P. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 41525004, 41421001, and 41877523, and supported by a grant from the State Key Laboratory of Resources and Environmental Information System, and the Science Foundation of China University of Petroleum, Beijing, Grant No. ZX20200100.

Acknowledgments: Many thanks to the Tencent company for authorizing us to use Tencent street view pictures in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dulinkeita, A.; Thind, H.; Affuso, O.; Baskin, M.L. The associations of perceived neighborhood disorder and physical activity with obesity among African American adolescents. *BMC Public Health* **2013**, *13*, 440. [[CrossRef](#)] [[PubMed](#)]
2. Dubey, A.; Naik, N.; Parikh, D.; Raskar, R.; Hidalgo, C.A. Deep learning the city: Quantifying urban perception at a global scale. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 196–212.
3. Porzi, L.; Rota Bulò, S.; Lepri, B.; Ricci, E. Predicting and understanding urban perception with convolutional neural networks. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 139–148.
4. Naik, N.; Philipoom, J.; Raskar, R.; Hidalgo, C.A. Streetscore—Predicting the perceived safety of one million streetscapes. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 793–799.
5. Liu, L.; Silva, E.A.; Wu, C.; Wang, H. A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Comput. Environ. Urban Syst.* **2017**, *65*, 113–125. [[CrossRef](#)]

6. He, S.; Yoshimura, Y.; Helfer, J.; Hack, G.; Ratti, C.; Nagakura, T. Quantifying memories: Mapping urban perception. *arXiv* **2018**, arXiv:1806.04054.
7. Doersch, C.; Singh, S.; Gupta, A.; Sivic, J.; Efros, A.A. What makes Paris look like Paris. *CACM* **2015**, *58*, 103–110. [[CrossRef](#)]
8. Lee, S.; Maisonneuve, N.; Crandall, D.; Efros, A.A.; Sivic, J. Linking past to present: Discovering style in two centuries of architecture. In Proceedings of the IEEE International Conference on Computational Photography, Houston, TX, USA, 24–26 April 2015; pp. 1–10.
9. Seide, F.; Li, G.; Chen, X.; Yu, D. Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription. In Proceedings of the Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011; pp. 24–29.
10. Xu, Y.; Yang, Q.; Cui, C.; Shi, C.; Song, G.; Han, X.; Yin, Y. Visual urban perception with deep semantic-aware network. In Proceedings of the Conference on Multimedia Modeling, Thessaloniki, Greece, 8–19 January 2019; pp. 28–40.
11. Anguelov, D.; Dulong, C.; Filip, D.; Frueh, C.; Lafon, S.; Lyon, R.; Ogale, A.; Vincent, L.; Weaver, J. Google street view: Capturing the world at street level. *Computer* **2010**, *43*, 32–38. [[CrossRef](#)]
12. Cheng, L.; Chu, S.; Zong, W.; Li, S.; Wu, J.; Li, M. Use of tencent street view imagery for visual perception of streets. *Int. J. Geo-Inf.* **2017**, *6*, 265. [[CrossRef](#)]
13. Hara, K.; Le, V.; Froehlich, J. Combining crowdsourcing and google street view to identify street-level accessibility problems. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 631–640.
14. Hwang, J.; Sampson, R.J. Divergent pathways of gentrification: racial inequality and the social order of renewal in Chicago neighborhoods. *Am. Sociol. Rev.* **2014**, *79*, 726–751. [[CrossRef](#)]
15. Zhang, L.; Pei, T.; Chen, Y.; Song, C.; Liu, X. A Review of urban environmental assessment based on street view images. *J. Geo-Inf. Sci.* **2019**, *21*, 46–58. [[CrossRef](#)]
16. Kelly, C.M.; Wilson, J.S.; Baker, E.A.; Miller, D.K.; Schootman, M. Using google street view to audit the built environment: Inter-rater reliability results. *Ann. Behav. Med.* **2013**, *45*, 108–112. [[CrossRef](#)]
17. Rundle, A.G.; Bader, M.D.M.; Richards, C.A.; Neckerman, K.M.; Teitler, J.O. Using google street view to audit neighborhood environments. *Am. J. Prev. Med.* **2011**, *40*, 94–100. [[CrossRef](#)]
18. Clarke, P.; Ailshire, J.; Melendez, R.; Bader, M.; Morenoff, J. Using google earth to conduct a neighborhood audit: Reliability of a virtual audit instrument. *Health Place* **2010**, *16*, 1224–1229. [[CrossRef](#)] [[PubMed](#)]
19. Badland, H.M.; Opit, S.; Witten, K.; Kearns, R.A.; Mavoa, S. Can virtual streetscape audits reliably replace physical streetscape audits? *J. Urban Health* **2010**, *87*, 1007–1016. [[CrossRef](#)] [[PubMed](#)]
20. Naik, N.; Kominers, S.D.; Raskar, R.; Glaeser, E.L.; Hidalgo, C.A. Computer vision uncovers predictors of physical urban change. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 7571–7576. [[CrossRef](#)] [[PubMed](#)]
21. Dong, R.; Zhang, Y.; Zhao, J. How green are the streets within the sixth ring road of Beijing? An analysis based on tencent street view pictures and the green view index. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1367. [[CrossRef](#)] [[PubMed](#)]
22. Li, X.; Zhang, C.; Li, W.; Kuzovkina, Y.A.; Weiner, D. Who lives in greener neighborhoods? The distribution of street greenery and its association with residents' socioeconomic conditions in Hartford, Connecticut, USA. *Urban For. Urban Green.* **2015**, *14*, 751–759. [[CrossRef](#)]
23. Li, X.; Zhang, C.; Li, W.; Ricard, R.; Meng, Q.; Zhang, W. Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban For. Urban Green.* **2015**, *14*, 675–685. [[CrossRef](#)]
24. Li, X.; Zhang, C.; Li, W. Does the visibility of greenery increase perceived safety in urban areas? Evidence from the place pulse 1.0 dataset. *ISPRS Int. Geo-Inf.* **2015**, *4*, 1166–1183. [[CrossRef](#)]
25. Li, H.; Páez, A.; Liu, D. Built environment and violent crime: An environmental audit approach using Google Street View. *Comput. Environ. Urban Syst.* **2017**, *66*, 83–95.
26. Liang, J.; Gong, J.; Sun, J.; Zhou, J.; Li, W.; Li, Y.; Liu, J.; Shen, S. Automatic sky view factor estimation from street view photographs—A big data approach. *Remote Sens.* **2017**, *9*, 411. [[CrossRef](#)]
27. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
28. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]

29. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
30. Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 647–655.
31. Seresinhe, C.I.; Preis, T.; Moat, H.S. Using deep learning to quantify the beauty of outdoor places. *R. Soc. Open Sci.* **2017**, *4*, 170170. [[CrossRef](#)]
32. Shen, Q.; Zeng, W.; Ye, Y.; Arisona, S.M.; Schubiger, S.; Burkhard, R.; Qu, H. StreetVizor: Visual exploration of human-scale urban forms based on street views. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 1004–1013. [[CrossRef](#)] [[PubMed](#)]
33. Trends and Characteristics of Population Change in Beijing in 2014. Available online: http://tj.beijing.gov.cn/tjsj/zxcdcsj/rkcydc/dcsj_4597/201601/t20160128_171191.html (accessed on 6 July 2019).
34. Tencent Street View (TSV) API. Available online: https://lbs.qq.com/panostatic_v1/guide-getImage.html (accessed on 10 August 2018).
35. Li, S. *The Arts of China*; Inner Mongolia People's Publishing House: Hohhot, China, 2006.
36. Liu, S. *Construction Civilization—Chinese Traditional Culture and Traditional Architecture*; Tsinghua University Press: Beijing, China, 2014.
37. Cao, X. The Research on Urban Color of Historic Sites in the Old City of Beijing. Master's Thesis, Beijing University of Civil Engineering and Architecture, Beijing, China, 2012.
38. Dutta, A.; Gupta, A.; Zissermann, A. Image Annotator. Available online: <http://www.robots.ox.ac.uk/~{v}gg/software/via> (accessed on 20 August 2018).
39. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
40. Yang, J.; Zhao, L.; McBride, J.; Gong, P. Can you see green? Assessing the visibility of urban forests in cities. *Landsc. Urban. Plan.* **2009**, *91*, 97–104. [[CrossRef](#)]
41. Dubin, R.A. Spatial autocorrelation: A primer. *J. Hous. Econ.* **1998**, *7*, 304–327. [[CrossRef](#)]
42. Kelejian, H.H.; Prucha, I.R. On the asymptotic distribution of the Moran I test statistic with applications. *J. Econom.* **2001**, *104*, 219–257. [[CrossRef](#)]
43. Liu, L. Study on the Protection of District Protectde Historical Sites in Beijing Old City. Master's Thesis, Beijing University of Civil Engineering and Architecture, Beijing, China, 2018.
44. Ping-Fang, X.U. The Planning and preservation of the streets in the old city of Beijing. *J. Beijing Union Univ.* **2008**, *6*, 23–27.
45. Yao, T.; Yang, X. A Comparative study on the street space form in the old city of Beijing: A case study of Shijia Hutong, the White Stupa Temple area, and dashilanr. *J. Landsc. Res.* **2018**, *10*, 22–26.
46. Whitehand, J.; Gu, K. Urban conservation in China: Historical development, current practice and morphological approach. *Town Plan. Rev.* **2007**, *78*, 643–670. [[CrossRef](#)]
47. Beijing Urban Master Plan (2016–2035). Available online: http://ghzrzyw.beijing.gov.cn/zhengwuxinxi/zxzt/bjcsztgh20162035/202001/t20200102_1554606.html (accessed on 20 May 2018).
48. Lambiotte, R.; Blondel, V.D.; Kerchove, C.D.; Huens, E.; Prieur, C.; Smoreda, Z.; Dooren, P.V. Geographical dispersal of mobile communication networks. *Phys. A Stat. Mech. Its Appl.* **2008**, *387*, 5317–5325. [[CrossRef](#)]
49. Wang, M.H.; Schrock, S.D.; Broek, N.V.; Mulinazzi, T. Estimating dynamic origin-destination data and travel demand using cell phone network data. *Int. J. ITS Res.* **2013**, *11*, 76–86. [[CrossRef](#)]

