Mapping Human Activity Volumes Through Remote Sensing Imagery

Xiaoyue Xing, Zhou Huang, Ximeng Cheng, Di Zhu, Chaogui Kang, Fan Zhang, and Yu Liu

***Abstract*—The spatial concentration of the human activity is a crucial indication of socioeconomic vitality. Accurately mapping activity volumes is fundamental to support the regional sustainable development. Current approaches rely on mobile positioning data, which record information about human daily activity but are inaccessible in most cities due to privacy and data sharing concerns. Alternative methods are needed to provide more generalized predictions on extensive areas while maintaining low cost. This study demonstrates how remote sensing imagery can be used through an end-to-end deep learning framework for reliable estimates of human activity volumes. The neighbor effect, representing the inherent nature of spatial autocorrelation in the volumes, is incorporated to improve the network. The proposed model exhibits strong predictive power and demonstrates great explainability of physical environment on variations of activity volumes. Landscape interpretations based on hierarchical features provide both object-based and region-based insights into the coevolvement of landscape and human activity. Our findings indicate the possibility of extensively predicting activity volumes, especially in areas with limited access to mobile data, and provide support for the promising framework to better comprehend broad aspects of the human society from observable physical environments.**

***Index Terms*—Deep convolutional neural network (DCNN), human activity, neighbor effect, physical environment.**

I. Introduction

GLOBAL sustainable development goals (SDGs) require essential knowledge of where and how crowded people are to "make cities and human settlements inclusive, safe, resilient, and sustainable" (UN SDG N.11). Accurate estimations of fine-grained population distribution have remarkably promoted sustainable land-use planning, resource management, and risk reduction [1]–[3]. Most existing population products count the census-based residential population, regardless of where people are located during the day [3]. Compared with the simply residential concept, people's daily attendances in all residence-, work-, and leisure-related places jointly reflect averaged population concentrations on various functional zones over a day, which are more critical for regional infrastructure allocations [4], [5]. Therefore, it is appropriate to use population counts averaged over 24 h in certain zones as one of the fundamental socioeconomic characteristics. We define such population as the human activity volume V, since it considers collective daytime activities into the measurement.

A *social sensing* framework [6] has been recently advocated to depict multiple facets of the human society from big geo-data, such as mobile phone records [7], social media data [8], street view images [9], [10], and so on. Widely generated individual mobile data have shown the potential to record the activity volumes and to further assist policymaking related to regional vitality [5], [11], [12]. In particular, as direct indications of human presence, mobile positioning data enable disaggregations of census data with socioeconomic weights [13], [14] and provide ancillary statistics of population dynamics [15]. Owing to the high penetration of location-based services (LBS) into a variety of daily activities, including instant communication, navigation, online business, and entertainment [16]–[18], geographical coordinates of most activity locations can be automatically recorded every time subscribers send location requests and authorize LBS-based mobile applications. This information stands well for the spatial distribution of individuals' daily activities, and yet, the access is impeded by privacy issues and enterprise data sharing concerns [19]. It is an open and active challenge to find an appropriate alternative for the positioning records and broadly predict human activity volumes. Fortunately, the interplay between the physical environment and human activity provides insightful evidences for tackling this problem [20]–[22].

Physical characteristics in built-up structures [23], greening covers [24], surface temperatures [25], [26], and atmosphere emissions [27] are greatly shaped by and inversely shapes how people communicate, produce, and live. This unravels a possibility of estimating human activity volumes based on a glimpse of the physical environment, which is traceable through *remote sensing* (RS) techniques. The RS observations comprehensively capture the physical environment properties of the Earth's surface and have advantages in terms of low acquisition cost and

Manuscript received July 12, 2020; revised August 21, 2020; accepted September 4, 2020. Date of publication September 14, 2020; date of current version September 30, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0196100 and in part by the National Natural Science Foundation of China under Grant 41771425, Grant 41830645, and Grant 41625003. *(Corresponding author: Zhou Huang.)*

Xiaoyue Xing, Zhou Huang, Ximeng Cheng, Di Zhu, and Yu Liu are with the Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing 100871, China (e-mail: xyxing@pku.edu.cn; huangzhou@pku.edu.cn; chengximeng@pku.edu.cn; patrick.zhu@pku.edu.cn; liuyu@urban.pku.edu.cn).

Chaogui Kang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the Center for Urban Science and Progress, New York University, NY 10012 USA (e-mail: cgkang@whu.edu.cn).

Fan Zhang is with Senseable City Laboratory, Massachusetts Institute of Technology, MA 02139 USA (e-mail: zhangfan@mit.edu).

Digital Object Identifier 10.1109/JSTARS.2020.3023730

broadly scanned coverages, including areas with low-frequency census or sparse network base stations. However, the complex landscape scenes and their associations with human activities urge feasible solutions to exploit distinguishable clues from RS imagery to estimate activity volumes with limited positioning data as guidance.

Deep convolutional neural networks (DCNNs) [28] are appropriate algorithms to capture hidden hierarchies of geographical patterns. They have shown great ability to detect deep knowledge for land cover classification [29]–[31], image semantic segmentation [32], [33], object localization [34], [35], and spatial interpolation [36]. The traditional DCNNs utilize local connections of features inside a target image. Nonetheless, different from object-centric photographs in natural image sets such as ImageNet, precisely estimating the regional characteristics from RS scenes requires understanding plenty of ground details and more complex spatial relationships [37]. For most geographical phenomena, neighbor characteristics are critical for estimating the centric targets due to their spatial associations, as a manifestation of *the first law of geography* [38]. The concentration of human activity is typically influenced by neighbor environments, which have shown close associations with the centric land cover [37], welfare [39], attractiveness [40], dynamics [41], [42], and further human–environment interactions [43], [44]. Therefore, RS-oriented strategies and specific network adjustments considering neighbor associations should be taken into account.

In this research, we develop Neighbor-ResNet, an end-to-end deep learning model with spatially neighbor augmentation, to achieve accurate estimations of the human activity volumes. Using RS images as inputs and LBS data as labels in 18 cities in China, we find that the end-to-end model shows strong feasibility and generalizability, and introducing the neighbor effect greatly enhances the model performance. Based on the landscape interpretation and deviation analysis, our model suggests the heterogeneity of RS observations reflecting the activity volumes and deepens our understanding of the interactions between the human activity and physical environments. In a summary, the main contributions of this article are as follows.

1) We propose a general data-driven framework to estimate socioeconomic variables by exploring informative landscape knowledge. As the model shows great performances on measuring the human activity, it exhibits the potential for estimating the broader socioeconomic indicators at a fine-grained level.
2) We provide a reliable estimator for an extensive mapping of the human activity volumes, which is especially meaningful in low income regions with sparse network infrastructures or limited access to mobile positioning data. Since the spatial scale is consistent with current census-based population products, our mappings can be useful complements to reflect averaged activity-based concentrations. This provides fundamental support for the regional sustainable development.
3) We investigate the heterogeneity of landscape traces of the human activity based on hierarchical features, and we reveal how the neighbor integration improves the estimations. These shed lights on the mechanism advantages of the proposed model, and provide a new perspective to understand the interaction between human activities and physical environments.

II. Materials and Methods

A. Data Sources and Preprocessing

We retrieve human activity volumes in 18 cities in China from Tencent LBS records. The data cover a period of five weekdays from 18 January 2016 to 22 January 2016. As a major social network platform in China, Tencent had monthly active users exceeding 877 million for QQ and 762 million for Wechat by the first quarter of 2016 [45]. Rapid-developed Tencent LBS platforms can consistently provide services covering various aspects of daily activities [18] when people send location requests and authorize related Tencent applications, which have reached 98% of Chinese internet users in 2017 [46]. Geographical coordinates of a large proportion of populations can be recorded owing to the huge subscriber bases. Therefore, Tencent LBS positioning record is a reliable proxy for the human activity volumes.

In this study, Tencent data are aggregated at a grid level of 0.01° in latitude by 0.01° in longitude with an area of around 1 km^2. Given human mobility, aggregated activity volumes in a walkable extent employed as a restriction of "neighborhood of opportunity"[47] are more appropriate to summarize local vitality than the raw coordinate points. The value of a unit grid is the total amount of coordinate records located in it during a certain period. We use the spatial density of activity volumes to eliminate ground area fluctuations of $0.01^\circ \times 0.01^\circ$ in different latitudes and longitudes due to map projection. Given that the data distribution is heavy tailed, with a skewness of 6.76 and a kurtosis of 70.3, we balance the dataset by taking logarithms of the activity densities as training labels, calculated as.

$$V_i = \log\left(\frac{1}{A_i}\frac{1}{D}\sum_{d=1}^{D}\bar{C_{id}}\right), \bar{C_{id}} = \frac{1}{H}\sum_{h=1}^{H}C_{idh} \quad (1)$$

where V_i is the label of the unit grid i, A_i is the ground area of the grid i (km^2), $\bar{C_{id}}$ is the daily record count of the grid i averaged over H hours ($H = 24$) of the day d, and it is then averaged again over the studied five weekdays ($D = 5$). C_{idh} is the raw data representing the total amount of positioning records located in the grid i within a 1-h period $(h-1, h]$ at the day d. Repeated requests of the same user within a period are counted only once.

RS images are obtained from the open-sourced Google Maps datasets with three multispectral bands (red, green, and blue; RGB) at a 19.1-m spatial resolution. True color images with RGB channels are suitable to reflect the detailed physical living environment, and they are consistent with human vision cognition about the physical space. Intuitively, different magnitudes of activity volumes correspond to distinct RS scenes (see Fig. 1), confirming it is possible to figure out their associations for the accurate estimations. Consistent with human activity data, RS

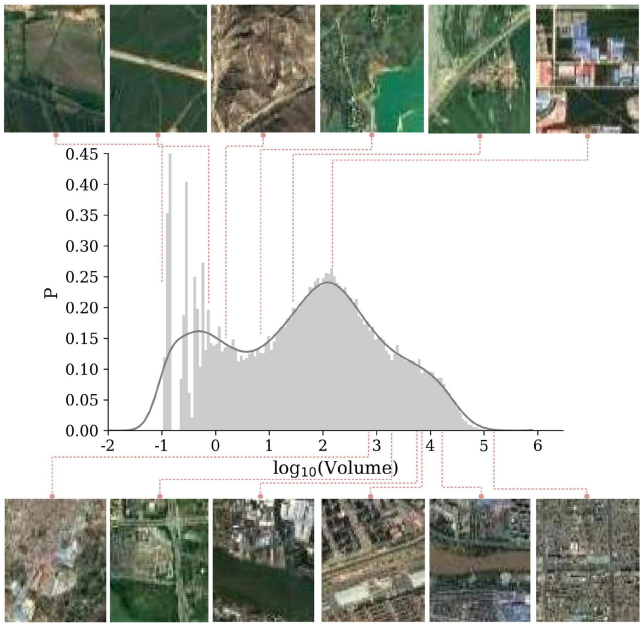


Fig. 1. Probability density histograms of the human activity volumes in dataset A, and examples of RS images associated with different volumes, linked by red dash line.

images are cropped into $0.01^\circ \times 0.01^\circ$ tiles with the same spatial extents and coordinate systems (WGS84 Web Mercator in this study).

B. Neighbor-ResNet Architecture

The physical characteristics and neighbor effects on the human activity are nonlinear and complex. We propose an end-to-end architecture called Neighbor-ResNet (see Fig. 2) by embedding the neighbor knowledge into ResNet-50 [48]. Since the volume data are in raster formats, we fix neighbor patches located at eight nearest units of a target sample. Surrounding information is included by searching for those neighbor RS tiles, and then, they are spatially concatenated from $0.01^\circ \times 0.01^\circ$ target areas to $0.03^\circ \times 0.03^\circ$ with the nearest neighbors. Different from deepening the input channels or concatenating feature vectors of parallel image patches, the input extension explicitly utilizes spatial relationships between center and neighbor tiles. In addition, the optimal scales to evaluate the local socioeconomic characteristics have been estimated to be 600–1000 m [49]. Therefore, we choose $0.01^\circ \times 0.01^\circ$ as a fixed neighbor extend, approximate to the upper bound of the optimal scales, to provide enough spatially associated information regardless of city diversity as well as to simplify the neighbor extension processes.

The model summarizes individual knowledge of the target and neighbors and assembles their features via layer-wise convolutional operations. The convolution filters, sliding on feature maps of each layer, can extract interior characteristics when covering network cells with only target or neighbor information, and conversely, it can integrate them at adjacent parts, as shown in the amplified part in Fig. 2. As layers go deeper, the proportions of integrated features increase as shown in Table I.

TABLE I
NUMERICAL PROPORTIONS OF NETWORK CELLS WITH INFORMATION OF THE INDIVIDUAL TARGET, NEIGHBORS, AND INTEGRATED PARTS

	Conv1	Conv2	Conv3	Conv4	Conv5
Target	0.104	0.085	0.005	0.002	0.000
Neighbour	0.874	0.843	0.443	0.234	0.000
Integrated	0.021	0.072	0.479	0.764	1.000

To maintain the extended spatial knowledge and avoid overfitting, we replace the final fully connected layers in ResNet-50 with convolutions generating outputs after the average pooling $3 \times 3 \times 2048$ layer. The output scalar is then an estimated volume of the $0.01^\circ \times 0.01^\circ$ area integrating both individual and associated features of the target RS tile and its neighbors. All outputs compose the whole regional distributions of human activity volumes.

C. Network Training

Dataset A in this study contains eight Chinese cities (Harbin, Beijing, Shanghai, Wuhan, Guangzhou, Kunming, Lanzhou, and Lhasa) for model training (60%, 45 998 images), validating (20%, 15 333 images), and testing (20%, 15 333 images). The locations and RS images are shown in Fig. 3. These cities vary in their geological landforms, urban landscapes, populations, and economic development levels. Therefore, they form a diverse and balanced dataset. In addition, ten randomly chosen cities apart from those used in model training compose another test set (dataset B) totally including 110 808 images to evaluate the model generalizability to new regions.

We use the L_1 loss function for back-propagation learning and weight updating. Hyperparameters are tuned empirically according to the model performance on validating sets. The learning rate and batch size are 10^{-4} and 32, respectively. After 15 000 epochs, the loss of the model convergences to a basically stable value.

We use Spearman's rank correlation coefficient (r_s), the mean absolute percentage error (MAPE), and the coefficient of determination (R^2) to assess the rank-fitting performance, absolute errors, and explained variances of the proposed model, respectively. Among these indices, the MAPE fluctuates greatly and introduces numerical noises when the denominators (real values) are small. Additionally, given the heavy-tailed characteristics of the data, large activity volumes exhibit low frequency but contain more valuable information about activity concentrations than those with high frequency [50]. Considering the different significances of those volume magnitudes, we adopt a weighted MAPE, as (2) shows, and set the weight as the reciprocal of the proportion of real volume data in different numerical intervals.

$$\begin{aligned} \text{MAPE} &= \frac{1}{\sum_{j=1}^{N} \frac{1}{p_j}} \sum_{i=1}^{N} \frac{1}{p_i} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \\ p_i &= \frac{N_c}{N},\ y_i \in [\delta \cdot c, \delta \cdot (c+1)) \end{aligned} \tag{2}$$

where y_i and $\hat{y}_i$ are the label V_i and the estimated value of the ith image, respectively; N is the testing size; and N_c is the amount

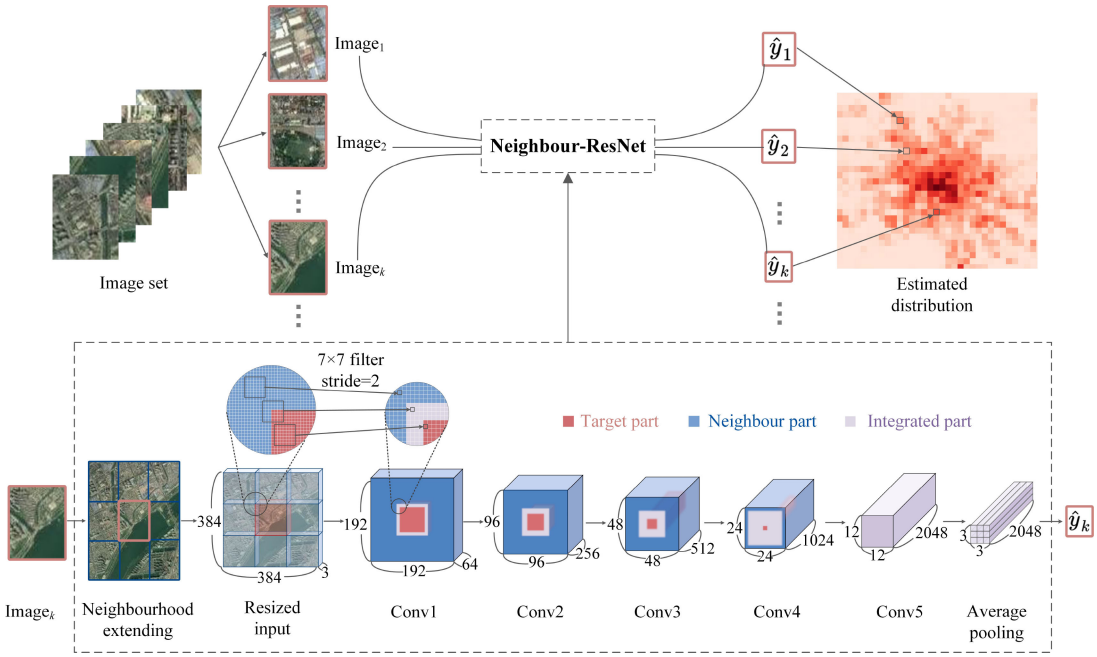


Fig. 2. End-to-end framework of Neighbor-ResNet using RS images (denoted as $\text{Image}_1, \text{Image}_2, \ldots, \text{Image}_k$) to estimate human activity volumes (denoted as $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k$). RS image tiles are resized to $(3 \times 128) \times (3 \times 128)$, including 128×128 pixels in every target (in red) and neighbor (in blue) unit. Blocks named "Conv" contain a group of convolutional layers and shortcut structures consistent with ResNet-50 in [48]. We amplify a part of the convolution processes on pixels in the top circles to show integrated information (in purple) generated by algorithmic convolutions at adjacent parts of the target and neighbor tiles.

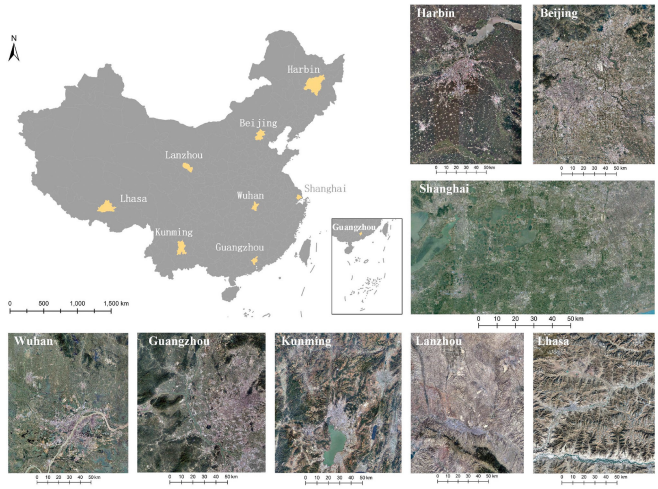


Fig. 3. Locations and RS images of the study cities in dataset A used for network training. The sampled areas, including city centers and suburbs, are set to be $1.0^\circ \times 1.0^\circ$, with 10 000 image tiles except for Shanghai due to the specialty of the city morphology.

TABLE II
ACCURACY ASSESSMENT OF TEST CITIES IN DATASET B

	ResNet			Neighbour-ResNet		
City	r_s	MAPE	R^2	r_s	MAPE	R^2
Hefei	0.525	69.5	0.484	**0.609**	**53.2**	**0.681**
Jinan	0.710	60.6	0.563	**0.790**	**49.2**	**0.691**
Luoyang	0.711	57.2	0.557	**0.784**	**46.1**	**0.677**
Shenzhen	0.886	77.4	0.011	**0.929**	**74.2**	**0.068**
Tianjin	0.682	56.4	0.610	**0.799**	**47.9**	**0.645**
Shijiazhuang	0.757	64.6	0.474	**0.860**	**48.1**	**0.678**
Shenyang	0.574	58.6	0.607	**0.690**	**56.8**	**0.612**
Nanchang	0.641	60.1	0.639	**0.765**	**49.4**	**0.707**
Changsha	0.698	67.4	0.471	**0.800**	**52.8**	**0.645**
Dalian	0.563	64.6	0.533	**0.650**	**57.8**	**0.605**
avg	0.675	63.6	0.495	**0.768**	**53.6**	**0.601**
std	0.100	6.2	0.171	0.091	7.8	0.180

The results in bold indicate better estimation performances in the comparison between ResNet and Neighbour-ResNet.

of the real volumes in $[\delta \cdot c, \delta \cdot (c+1))$. We choose the interval (δ) as 400, the maximum segment length under which the data do not significantly heavy-tailed distributed, to adjust the MAPE measurement.

III. RESULTS

A. Accuracy

The proposed model provides an advantage of utilizing limited volume labels and widely available RS images for activity volume estimation, with high feasibility and generalizability. For testing sets in dataset A, Neighbor-ResNet ($r_s = 0.942, \text{MAPE} = 37.7\%$) exhibits higher accuracy than ResNet ($r_s = 0.894, \text{MAPE} = 51.4\%$). The improved R^2 (0.803 for Neighbor-ResNet and 0.675 for ResNet) indicates that Neighbor-ResNet can explain a greater proportion of the volume variance from RS observation clues than ResNet can. For generalization, we apply the well-trained model directly to cities in dataset B (see Table II). The average (avg) and standard deviation (std) of R^2 (0.601 ± 0.180) imply that, through the layer-wise feature assembling and evolving, we generally explore a potential 60.1% explainability of remotely sensed physical environments on the activity volumes. Spatial distributions of estimated outputs of Neighbor-ResNet are in

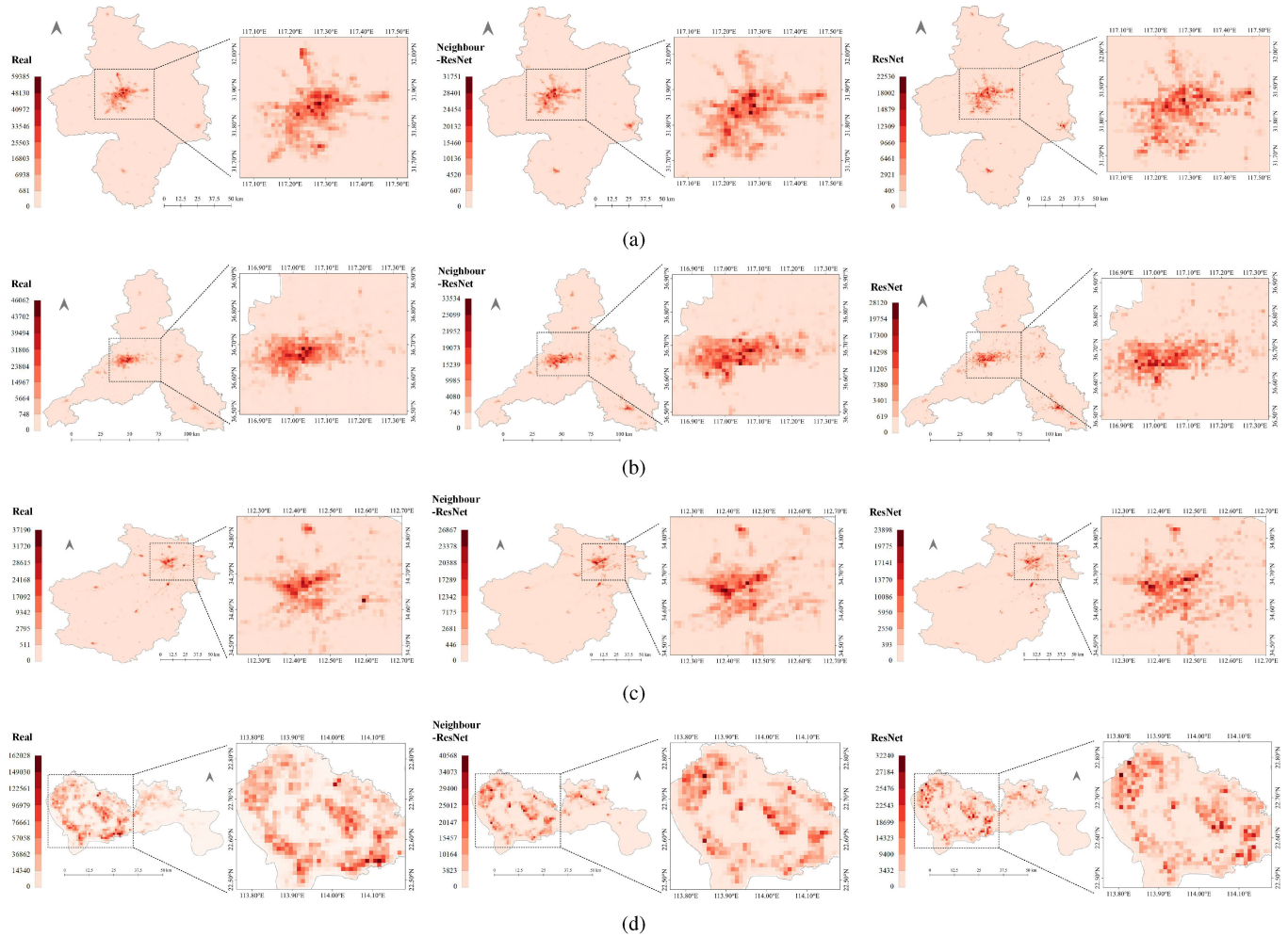


Fig. 4. (Left) Spatial distributions of real values, (middle) estimated outputs of Neighbor-ResNet, and (right) ResNet in four test cities: (a) Hefei, (b) Jinan, (c) Luoyang, and (d) Shenzhen. Distribution comparisons of all test cities are shown in Appendix D, Fig. 12. Breakpoints of the color ramp are determined by the head/tail breaks [50].

better agreement with the real volumes than those of ResNet (see Fig. 4). Densely populated urban centers and city-wide spatial patterns can be clearly recognized in the estimated mappings. We also find that Shenzhen is an exception, with a high rank correlation ($r_s = 0.929$) but low absolute accuracy (MAPE $= 74.2\%$, $R^2 = 0.068$) with the misrecognition of the south downtown. The specific analysis is provided in Appendix C.

B. Landscape Interpretation

Exploring the landscape details underpinning the network recognition gives us some insights into how human activity interplays with physical environments. We decompose the network to see how layer-wise features evolve through the deep architecture (in Appendix B). The finding shows how the end-to-end network captures hidden hierarchies of RS images, from gathering fine-grained information of ground objects to extracting characteristics of land parcels, and finally, assembling abstract features as high-level regional representations. Thus, we conduct the interpretation from both object-based and region-based perspectives.

In shallow layers, distinctive ground objects provide physical clues about volume variations. We use gradient-weighted class activation mapping (Grad-CAM) [51] to figure out ground indicators of the human activity. The obtained heat maps quantify relative contributions of input pixels to the estimates and highlight distinctive ground objects. As summarized in Fig. 5, we find some informative objects that are commonly shared, regardless of city specialties.

Through assembling image characteristics and generating high-level features, our model distinguishes subtle differences of regional layouts and reveals heterogeneous landscape traces of the human activity. We cluster the RS scenes based on the learnt features to analyze such heterogeneity. The images are grouped by the minibatch k-means method [52] using feature vectors of the last network layer as representations of the inputs. Beijing in dataset A and Shenzhen in dataset B are selected as

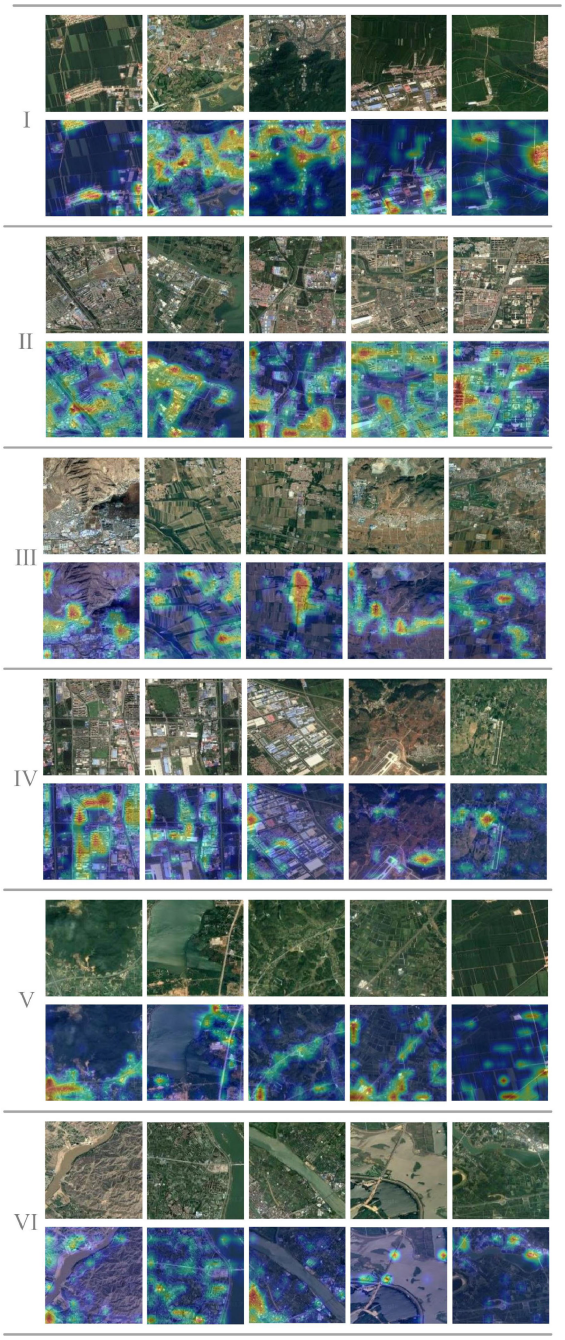


Fig. 5. Grad-CAMs for object interpretations. Labels I–VI represent six typical recognition scenes in $0.03° \times 0.03°$, each with five examples. RS images are in the top rows, and corresponding heat maps are in the bottom rows. When surrounding lands are easily distinguished by color as in Scene I or built-up areas cover most of the input regions as in Scene II, compact buildings are highlighted; Scene III shows the situation when constructions border on farmland or on hillside. Although they have similar hues, the network focuses on building areas accurately. One explanation is that the coarse texture makes building areas distinctive from natural objects; Scene IV proves the ability of our model to recognize different artificial objects. For buildings like industrial workshops, agricultural greenhouses or airport pavement, large regular shapes, and highly saturated roof colors may be informative indicators to distinguish them from crowded downtown buildings; Scene V illustrates that roads and their intersections are strongly related to human activities. Whether a road exists indicates accessibility, especially in rural areas; in Scene VI, the model avoids locating highlighted places on rivers, which plays a negative role on volume increasing. When bridges appear, the adjacent points linking the bridge and the shore are highlighted.

analysis cases, since they are metropolises in the north and south of China, respectively, and cover a wide range of per-unit volume magnitudes. The activity volumes increase with the growth of built-up areas, the decrease of green cover rates and transitions of architectural appearances in two cities [see Fig. 6(a) and (b)], while the building density shows little variation especially in Shenzhen. It indicates that increased human activity leads to built-up area expansion but building density may remains at a consistent level for daily activity demands. For nearby classes in high indexes, such as Classes #9 and #10, it is difficult to visually identify their differences. Nonetheless, they are distinguished clearly by our model, with real volumes displaying individual distributions [see Fig. 6(c) and (d)] and their units spatially clustering in distinct urban functional contexts [see Fig. 6(e) and (f)]. Detailed analysis is provided in Appendix C-A. These results demonstrate that high-level distinctions of human activity traces on landscape layouts can be effectively captured by our model.

The explainability of the physical environment on human activity volumes varies in these feature-based classes [see Fig. 6(c) and (d)]. We see that medium classes are more accurate. For classes in lower indexes, the accuracy is influenced by small population bases and noises. For those in higher indexes, when construction layouts have been largely covered and fixed, the increase of activity volumes probably follows or leads to changes that are less recognizable through RS observations, such as more effective utilization of interior architectural spaces, larger transportation capacity, stronger infrastructure support, or more attractive markets.

C. Neighbor Effect

The neighbor landscape provides knowledge about geographical contexts of the target area and enhances the estimations, but its importance is inconsistent under different types of spatial associations including high–high clustering with the centric unit as a hot spot (HH), low–low clustering with the center as a cold spot (LL), low–high clustering with the center as a low outlier (LH), high–low clustering with the center as a high outlier (HL), and nonsignificant association with a 95% confidence interval (NS). Local Moran's I [53] describes such associations, that is, the degrees to which observations at certain areas are spatially autocorrelated to those nearby. It recognizes autocorrelation types in real volumes based on statistical tests (in Appendix A). We separately analyze the performance improvement of Neighbor-ResNet on units with different association types (see Fig. 7). The result shows incorporating neighbor knowledge is more beneficial for estimating the human activity in hot spots than cold spots and nonsignificant units, while for outliers, the effect fluctuates.

We find that on the detailed numerical distributions of activity volumes neighbor effects are differentiated by the spatial associations, as shown in Fig. 8 and Table III. For hot spots mostly located in urban districts such as units in Classes #9 and #10 in Shenzhen, neighbor landscape amplifies the difference of activity concentrations in target areas, leading to larger variances and larger averages in estimates. This shows the aggregation

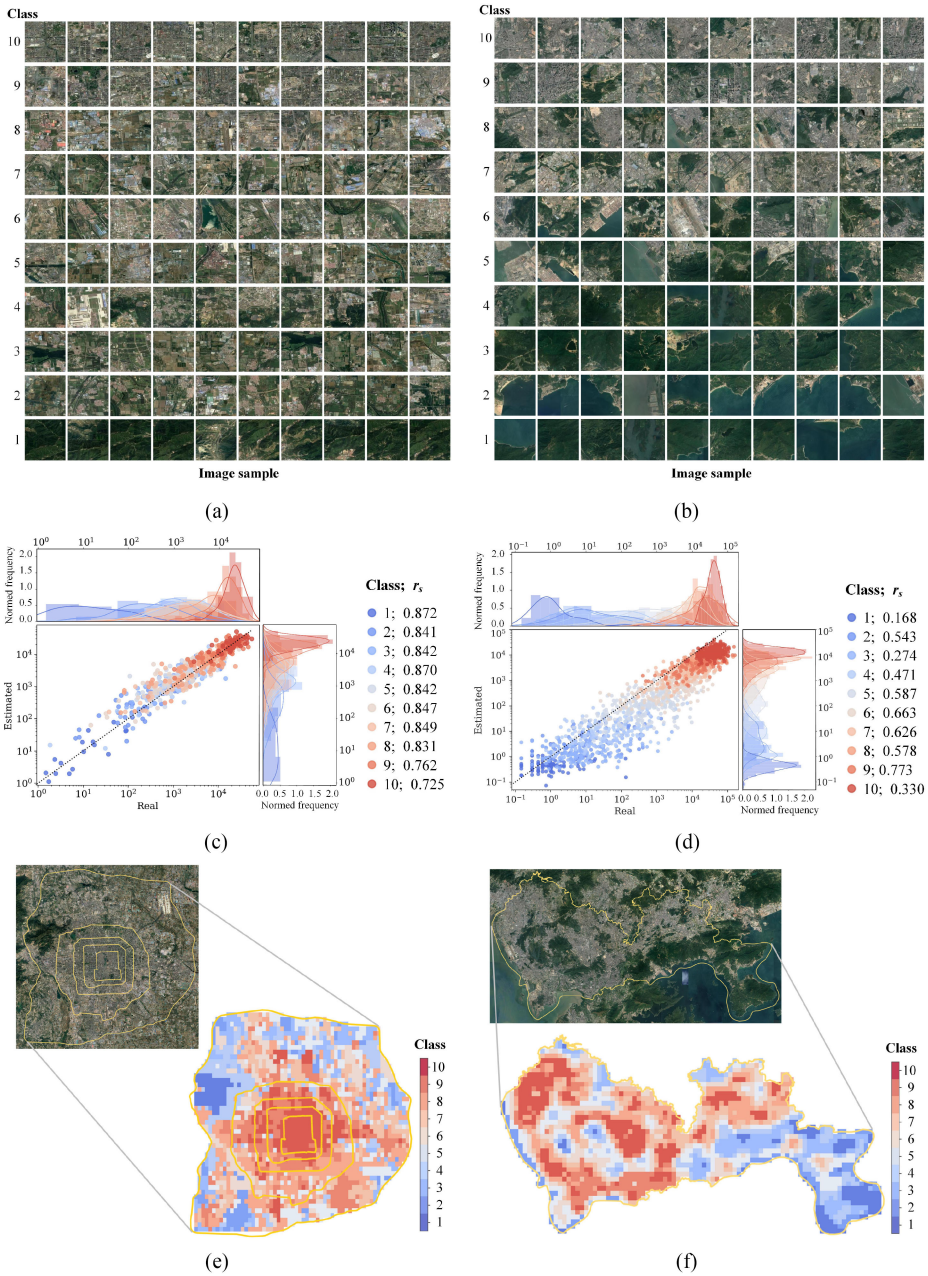


Fig. 6. (a) and (b) RS image samples, (c) and (d) test accuracies, and (e) and (f) spatial distributions of classified landscapes in (left) center Beijing inside the Sixth Ring Road and (right) Shenzhen. The landscapes are divided into ten classes labeled by numerical ranks of averaged real volumes. The optimal number of classes is determined by maximizing interclass distances and minimizing intraclass distances. In (c) and (d), correlations between real and estimated volumes of test data are listed. Black dotted lines in the scatter plot are $y = x$ lines. The top and right histograms show numerical distributions of real and estimated values, respectively. The classes, sorted from the lowest to the highest correlation r_s, are Classes #1, #10, #3, #9, #8, #2, #5, #4, #6, and #7 for Beijing and Classes #1, #3, #10, #4, #2, #8, #5, #7, #6, and #9 for Shenzhen. The five yellow loop lines in (e) represent the Second Ring Road (R2) to the Sixth Ring Road (R6) in Beijing. The yellow line in (f) represents the city boundary of Shenzhen.

TABLE III
INDICES DESCRIBING NUMERICAL DISTRIBUTIONS OF REAL VOLUMES, OUTPUTS OF NEIGHBOR-RESNET AND RESNET IN HOT SPOTS (HH), COLD SPOTS (LL), AND NONSIGNIFICANT AREAS (NS), INCLUDING AVERAGES (AVG), STANDARD DEVIATIONS (STD), SKEWNESS (SKEW), AND KURTOSIS (k)

	HH				LL				NS			
	avg	std	$skew$	k	avg	std	$skew$	k	avg	std	$skew$	k
Real	12949	14322.9	2.91	13.6	40	284.0	26.56	879.4	456	2348.1	17.22	474.2
Neighbour-ResNet	9070	6710.1	1.07	1.2	33	143.0	52.47	4026.8	394	1177.5	8.47	117.5
ResNet	6285	5206.7	1.23	2.0	40	166.7	34.18	2003.2	337	1064.3	7.85	92.31

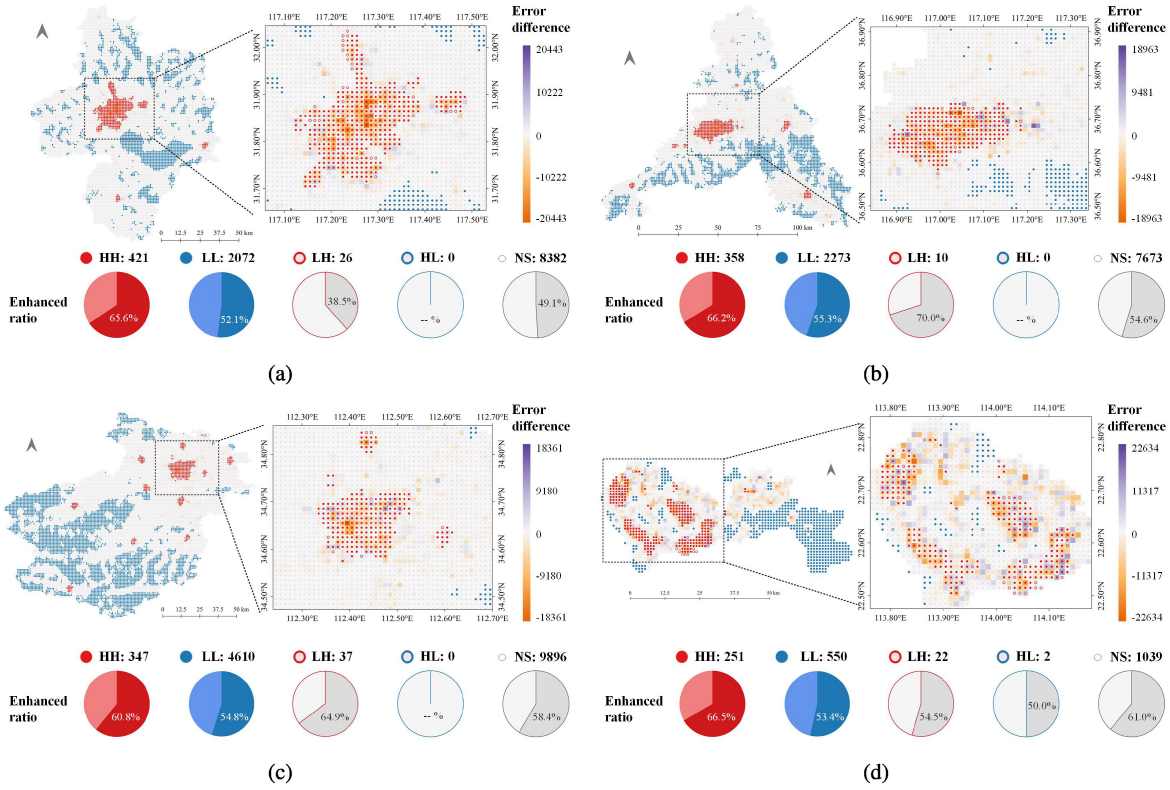


Fig. 7. Neighbor enhancement under different spatial associations in four cities: (a) Hefei, (b) Jinan, (c) Luoyang, and (d) Shenzhen. Results of all test cities are shown in Appendix D, Fig. 13. Unit values are differences of absolute estimation errors of Neighbor-ResNet and ResNet ($\text{Error}_{\text{Neighbor-ResNet}} - \text{Error}_{\text{ResNet}}$). Negative values mean accuracy improvements using Neighbor-ResNet. Five dot symbols indicate five local autocorrelation types. Pie plots annotate the ratios of units with error decrease using Neighbor-ResNet. From the results for test cities, Neighbor-ResNet generally performs better on 62.7% of hot spots, 56.9% of cold spots, and 56.5% of nonsignificant units. The enhancement on outlier units is unstable.

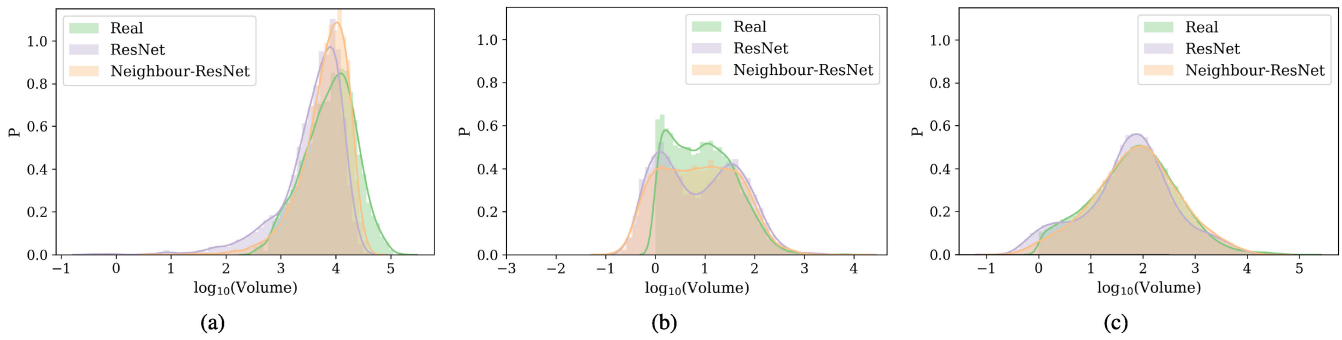


Fig. 8. Probability density histograms of real volumes and outputs of Neighbor-ResNet and ResNet in (a) hot spots, (b) cold spots, and (c) nonsignificant areas.

effect, partly explained by spatially assembled agglomeration economics in urban growth [54]. For cold spots covering major suburban areas, such as Classes #1–5 in Shenzhen, characteristics of neighbor environments are different from those of center areas. They smooth the volume variations and rectify extreme values. This shows dispersion effect contrary to the downtown. In nonsignificant areas, neighbor knowledge is also informative, producing outputs with slightly larger variances and better agreement with real volumes than those of ResNet. These results inspire future model designs by separately training groups of urban districts and suburbs to reinforce the knowledge of each single effect of neighbors.

D. Constraining Factors

We identify factors limiting the model performance according to RS techniques, regional socioeconomic functions, and specific ground appearances. In addition, we pinpoint units with the largest deviations in test cities as complementary instances for the analysis (see Fig. 9).

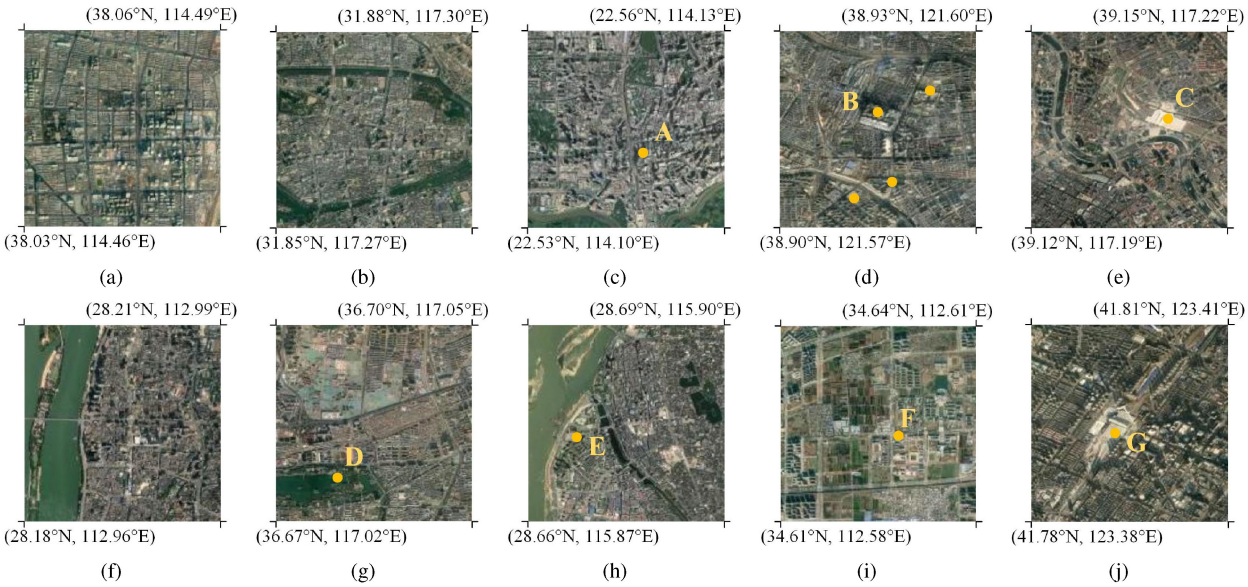


Fig. 9. Input RS images of the units with the largest estimate deviations in test cities: (a) center area in Shijiazhuang, covering high-rise buildings; (b) area in Hefei, with the circular waterway park attracting tourists for its emerald-necklace-like appearance; (c) downtown district in Shenzhen, covering international financial center (point A) along the south boundary; (d) center area in Dalian, covering high-rise building ranges (points B); (e) area in Tianjin, covering sightseeing and center economic belts along rivers and the Tianjin railway station (point C); (f) area in Changsha, covering river scenic belt and financial center; (g) area in Jinan, covering an attractive historic Daming lake (point D); (h) area in Nanchang, covering famous historic pavilion (point E) and developed river tunnels; (i) suburb in Luoyang, covering a college town (point F); and (j) area in Shenyang, covering the Shenyang railway station (point G). Deviations in these areas are all underestimated.

1) *Building height:* Building height is one of the key variables reflecting the human activity, but it presents limited information in 2-D scanned RS images. Although we can use building shadows in 2-D images for height extractions, it is greatly influenced by photographing orientations and the relative positions of the satellite, the Sun, and the buildings [55]. Instead, 3-D RS techniques, such as laser altimeter [56], multiangular observation [57], and airborne light detection and ranging (LiDAR) [58], present better measurement capability. The dimension limitation reduces the distinguishability of architecture heights. This partly explains local deviations in high-rise building ranges [see Fig. 9(a)–(d)]. In particular, this influence is greater in Shenzhen because this city has more high rises and larger population densities after the rapid urbanization.
2) *Socioeconomic backgrounds:* Particular land functions and demographic backgrounds are less traceable from RS images. First, while natural scenes negatively affect activity increasing in general, functional natural zones are exceptions, such as sightseeing and economic belts [Fig. 9(e) and (f)] and tourist attractions [see Fig. 9(b), (h), and (g)]. Given their low frequency in training data and the preservation principle of natural scene development, increases of activity volumes in these functional scenes are hard to detect. Second, significantly unbalanced demographic structures cause deviations. Since young adults have greater presences in mobile networks [59], [60], recorded activity magnitudes tend to increase in regions with greater proportions of young people [see Fig. 9(i)] or rapidly growing cities such as Shenzhen. The infrequent large volumes are hard to accurately predict (see details of Shenzhen in Appendix C-B).
3) *Specific appearances:* Layouts and buildings with special appearances have unique features that beyond the general knowledge learnt by our model. Layouts covering demarcation lines show discontinuous transformation from downtown constructions to natural scenes. This causes underestimations, such as those along city boundary of Shenzhen and Hong Kong [see Fig. 4(d)] or along coastal lines in Dalian [see Fig. 4(j)]. Specific constructions may also be misrecognized by the model. Diverse transport stations are typical examples. They can be designed similar to factories with regular bright roofs [see Fig. 9(f)] or with a unique appearance as a city symbol [see Fig. 9(j)]. High volumes but low recognizability make those buildings the sources of deviations.

IV. Discussion

In this work, deviations indicate the explainability limits of physical environment, which inspire further utilizations and enhancements. On the one hand, they highlight key areas that need additional attention for regional management. Deviating from the general knowledge learned by the model, these regions reveal mismatches between local activity volumes and the environmental carrying capacity. Overestimation may occur when local decision-makers have not realized and developed the potential of the regions, whereas underestimation emerges

TABLE IV
INDICES DESCRIBING NUMERICAL DISTRIBUTIONS OF REAL VOLUMES IN TEN LANDSCAPE CLASSES (C) IN SHENZHEN AND THE TEST SET OF BEIJING

	Shenzhen				Beijing			
C	avg	std	skew	k	avg	std	skew	k
1	5	19	6.5	41.3	39	57	2.0	3.4
2	76	167	4.8	29.4	732	1702	5.1	27.9
3	86	349	7.3	57.6	1692	2565	3.4	12.6
4	255	724	6.7	53.9	2686	2504	1.5	3.2
5	1567	2740	3.8	19.3	3640	4996	2.5	6.3
6	8380	9609	2.2	6.9	3947	5373	2.5	7.0
7	13537	11843	2.3	10.3	3483	4657	2.4	6.9
8	24216	18419	2.1	8.1	6643	5517	0.9	0.1
9	30881	30054	1.8	3.8	15589	9084	0.7	0.5
10	42336	20155	0.9	0.7	23344	11208	0.7	0.2

The class labels are ranks of averages of all real volumes; thus, the averages of only test set in Beijing are not strictly increasing.

when physical settings in those regions do not fit well with large populations. On the other hand, the limits inspire further model designs considering region diversity. This can be achieved by preclassifying regions for separate training, such as areas in downtowns and suburbs or in metropolises and small cities. Moreover, prior knowledge about city specialties can be added through input feature enrichment.

Temporal and spatial scales also influence the estimation. The changing frequency of the human activity is higher than that of the physical environment. Correspondingly, updating periods of RS imagery and positioning data are inconsistent. Based on this concern, our work is conducted with a relatively low temporal frequency, showing general interplays between daily averaged human activity and the physical environment. Spatially, owing to the invariance of convolutional operations, the architecture is universally fitted for diverse scales regarding different estimated socioeconomic factors and potential applications. The model is also promising to generate stratified population mappings across multiple spatial scales. In addition, the definition of "neighbors" influences the effectiveness of incorporated information. Although the selected $0.01^\circ \times 0.01^\circ$ neighbor extend has proved to be efficient in most cities, it still lacks the flexibility to fit regional specialties. The model can be further improved by adaptively adjusting optimal neighbor extensions based on ancillary data, such as road structures, land use types, and urban morphologies.

V. CONCLUSION

In this study, we develop a new deep-learning-based framework for estimations of human activity volumes from widely available RS imagery. The model needs only limited cover of mobile positioning data as the training guidance, and it learns the generalizable knowledge of physical environments to achieve extensive predictions. The spatial distributions of human activity resulting from our model are in agreement with the real data, and integrating neighbor knowledge enhances the estimation. Our findings and further interpretations suggest that our model can capture heterogeneous interactions governing human activity on different landscapes and neighbor associations.

Through the end-to-end model, we directly build a bridge between socioeconomic and physical environments and extract informative landscape traces of the human activity, one of the

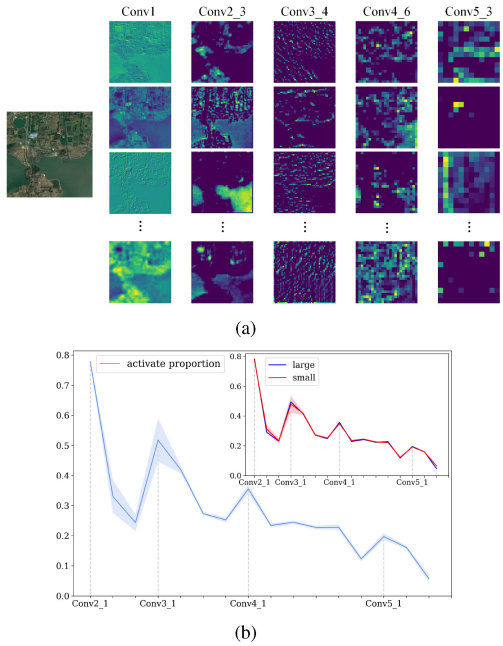


Fig. 10. Layer-wise analysis. (a) Examples of feature maps in five convolutional layers. The detailed layer names are consistent with those of ResNet-50 in [48]. (b) Proportions of activated units after the rectified linear unit (ReLU) activation functions in all test images. The value decreases in deep layers except for slight increases after max-pooling. Input areas with higher than 3000 activity volumes (large) or under ten records (small) are compared in the subfigure. Values in shallow layers have large variances and are more affected by raw input data since they extract most details of the images, while those in deep layers are more stable.

critical socioeconomic factors. This framework shows some advantages and applications: Theoretically, it provides support to track the coevolvement of the human activity with physical landscape through hierarchical RS features; the model enhancement validates the feasibility of integrating the geographical laws into networks; and the great predictability reminds us to consider the collinearity of activity-related indices and environmental factors when they are both used as explanatory variables in related tasks. Practically, it provides a universal architecture supporting a wide range of socioeconomic measurements, such as gross domestic product (GDP), crime rates, and housing prices; it can be a reliable human activity estimator with great generalizing performances on extensive unsampled areas, especially in low-income countries and regions; and the estimated values provide a basic magnitude reference to adjust diverse activity data from different mobile sources, and thus, make them comparable. With the advantages of the generalizability of the DCNN and scale invariance of convolutions, future model improvement can be achieved by enriching input features about region specialties and utilizing adaptive strategies for scale selections and neighbor extensions.

APPENDIX A
RECOGNIZING TYPES OF LOCAL SPATIAL AUTOCORRELATIONS

Local Moran's I [53] measures local spatial autocorrelations. It can be classified into five types based on a collection of

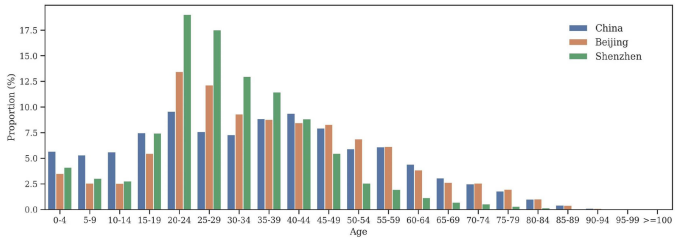


Fig. 11. Demographic structures of the whole of China, Beijing, and Shenzhen. The proportions of population aged from 20 to 40 in Shenzhen is larger than the averaged level around China. Although Beijing and Shenzhen are similar metropolises, the latter has a greater proportion of young adults than the former has.

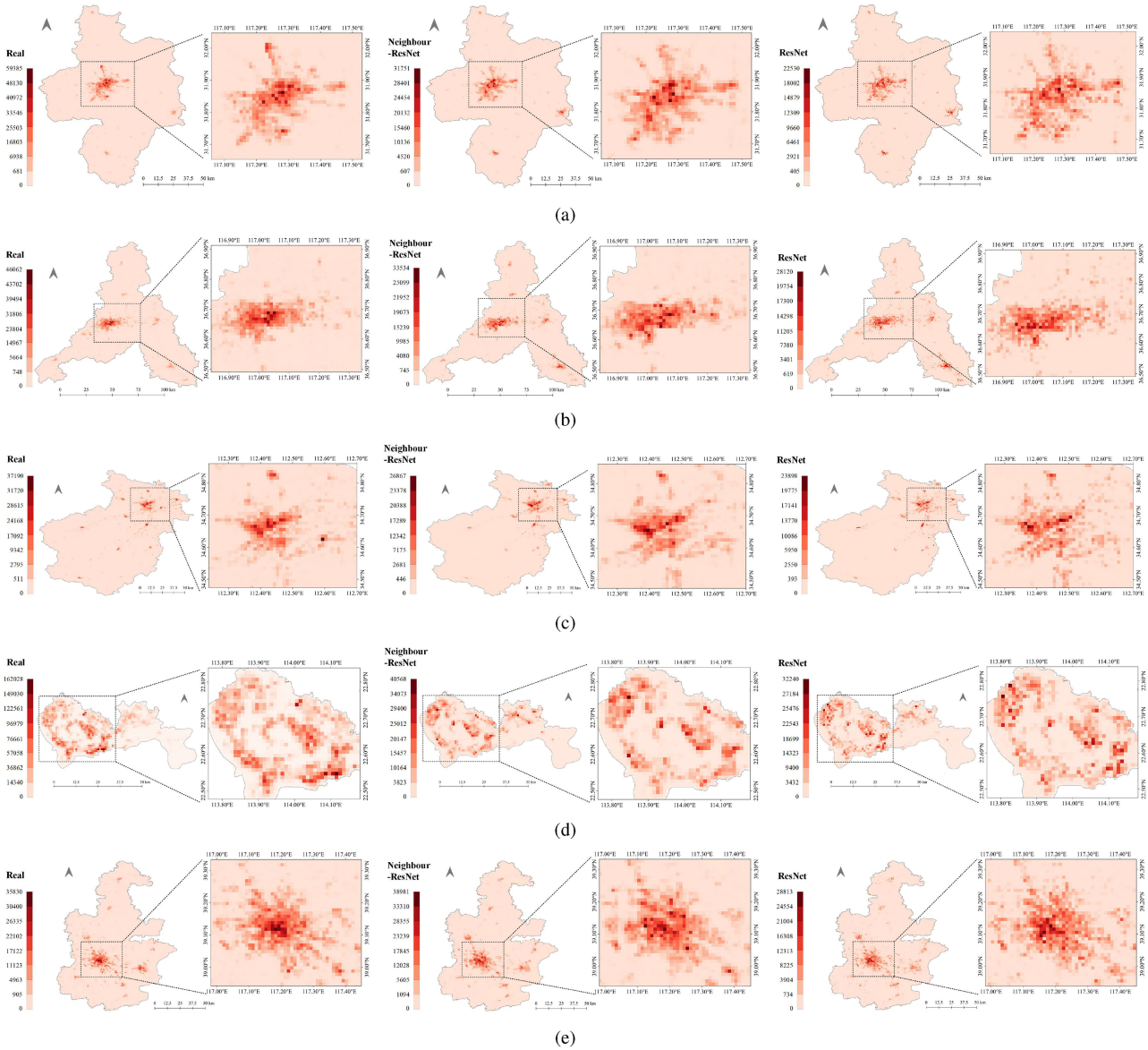


Fig. 12. (Left) Spatial distributions of real values, (middle) estimated outputs of Neighbor-ResNet, and (right) ResNet in all test cities: (a) Hefei, (b) Jinan, (c) Luoyang, (d) Shenzhen, (e) Tianjin, (f) Shijiazhuang, (g) Shenyang, (h) Nanchang, (i) Changsha, and (j) Dalian. Breakpoints of the color ramp are determined by the head/tail breaks [50].

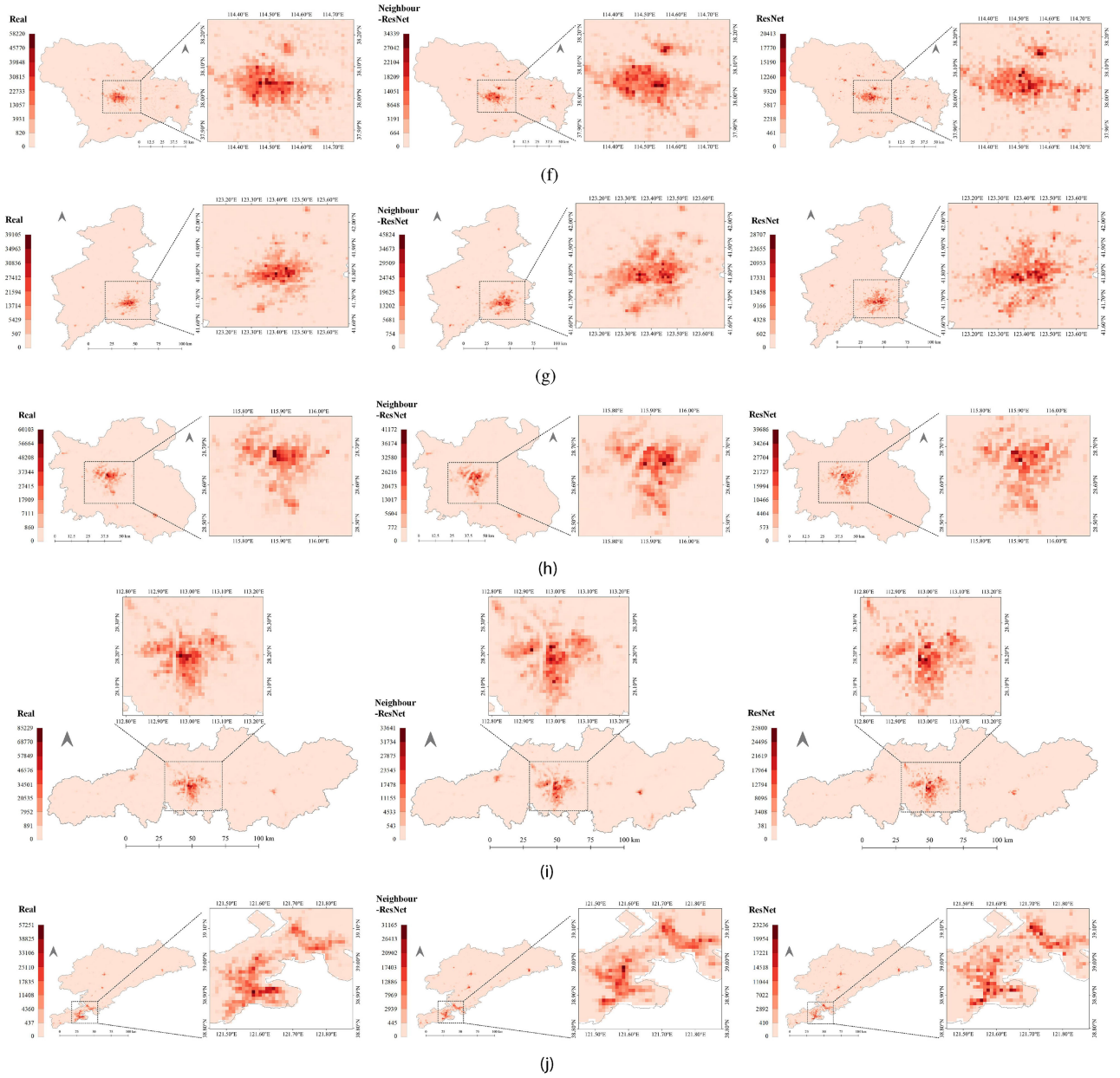


Fig. 12. (Continued.) (Left) Spatial distributions of real values, (middle) estimated outputs of Neighbor-ResNet, and (right) ResNet in all test cities: (a) Hefei, (b) Jinan, (c) Luoyang, (d) Shenzhen, (e) Tianjin, (f) Shijiazhuang, (g) Shenyang, (h) Nanchang, (i) Changsha, and (j) Dalian. Breakpoints of the color ramp are determined by the head/tail breaks [50].

statistical indicators explicitly describing the spatial association of a certain observation with those nearby.

The local Moran's I for an observed location i is defined as

$$I_i = z_i \sum_j \omega_{ij} z_j$$

where z_i and z_j are standardized observations at locations i and j, while w_{ij} is the spatial weight element measuring the ith and jth relationships. We set the weight of the ith observation with its eight neighbor units to be 1, while others are set as 0. Therefore, only neighbor values are included in the weighted summations over j.

We use the normally standard Z-score of the local Moran's I and its statistical significance (p-value) to detect local spatial clusters and outliers. The Z-score can be further translated into q-values as quadrant numbers of Moran scatter plots [61]. The five recognition rules are as follows.

1) $q = 1, p < 0.05$: High–high clustering with the center unit as a hot spot (HH).
2) $q = 2, p < 0.05$: Low–high clustering with the center unit as a low outlier (LH).
3) $q = 3, p < 0.05$: Low–low clustering with the center unit as a cold spot (LL).
4) $q = 4, p < 0.05$: High–low clustering with the center unit as a high outlier (HL).

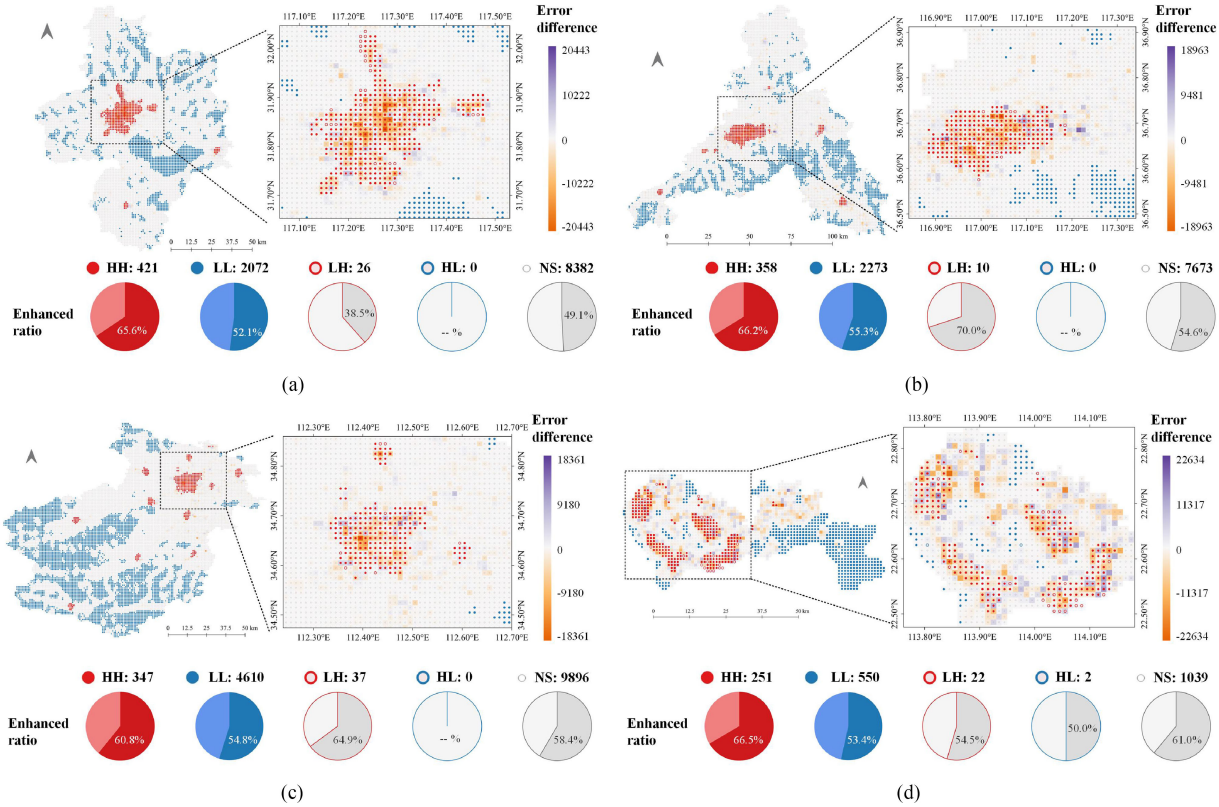


Fig. 13. Neighbor enhancements in all test cities: (a) Hefei, (b) Jinan, (c) Luoyang, (d) Shenzhen, (e) Tianjin, (f) Shijiazhuang, (g) Shenyang, (h) Nanchang, (i) Changsha, and (j) Dalian. Unit values are differences of absolute estimation errors of Neighbor-ResNet and ResNet ($\text{Error}_{\text{Neighbor-ResNet}} - \text{Error}_{\text{ResNet}}$). Negative values mean accuracy improvements using Neighbor-ResNet. Five dot symbols indicate five local autocorrelation types: high–high clustering with the center unit as a hot spot (HH), low–low clustering with the center as a cold spot (LL), low–high clustering with the center as a low outlier (LH), high–low clustering with the center as a high outlier (HL), and nonsignificant association with a 95% confidence interval (NS). Pie plots annotate the ratios of units with error decrease using Neighbor-ResNet. The hot spots show larger proportions of the enhancement using Neighbor-ResNet than the cold spots and nonsignificant units do, while the ratios in the outliers fluctuate.

5) $p \geq 0.05$: Nonsignificant association with a 95% confidence interval (NS).

APPENDIX B
NETWORK INTERPRETATION

Based on diverse training samples, our network explores general knowledge linking human activities with satellite observations. To understand how the end-to-end network learns, we decompose and interpret the deep architecture. The network explores hidden hierarchies [28] of RS images as shown in layer-wise feature maps [see Fig. 10(a)], from detailed geometric information to land parcel characteristics, and finally, to high-level regional representations. In Conv1, details are detected, such as edges and shapes. In Conv2_3, land parcels are highlighted, such as water areas, farmlands, and paths. Textural information is clearly detected in Conv3_4. For the deeper layers at Conv4_6 and Conv5_3, the features are more abstract. Different channels in one layer highlight diverse parts and generate different feature maps at the same abstract level. In addition, valid information becomes sparse when traveling through the layers [see Fig. 10(b)]. These results demonstrate how the inputs evolve to the output scalars, from gathering abundant fine-grained information to abstracting hierarchical features and assembling them as high-level representations.

APPENDIX C
SHENZHEN ANALYSIS

The estimation result for Shenzhen shows a high rank-fitting performance ($r_s = 0.929$) but a low absolute accuracy (MAPE $= 74.2\%$, $R^2 = 0.068$). The spatial distribution comparison reveals that the south center of Shenzhen is not consistently recognized. We conduct the analysis by combining landscape interpretations and deviation assessments.

A. *Model Effectiveness*

We validate that the model works in Shenzhen, since subtle differences of RS images are recognized with individual numerical volume distributions and distinct urban functional contexts. As shown in the histograms [see Fig. 6(c) and (d)], both real volumes in Beijing and Shenzhen in different classes follow approximately log-normal distributions with different averages, standard deviations, skewness, and kurtosis outcomes (see Table IV). Classes #9 and #10 in Shenzhen in Fig. 6(f)

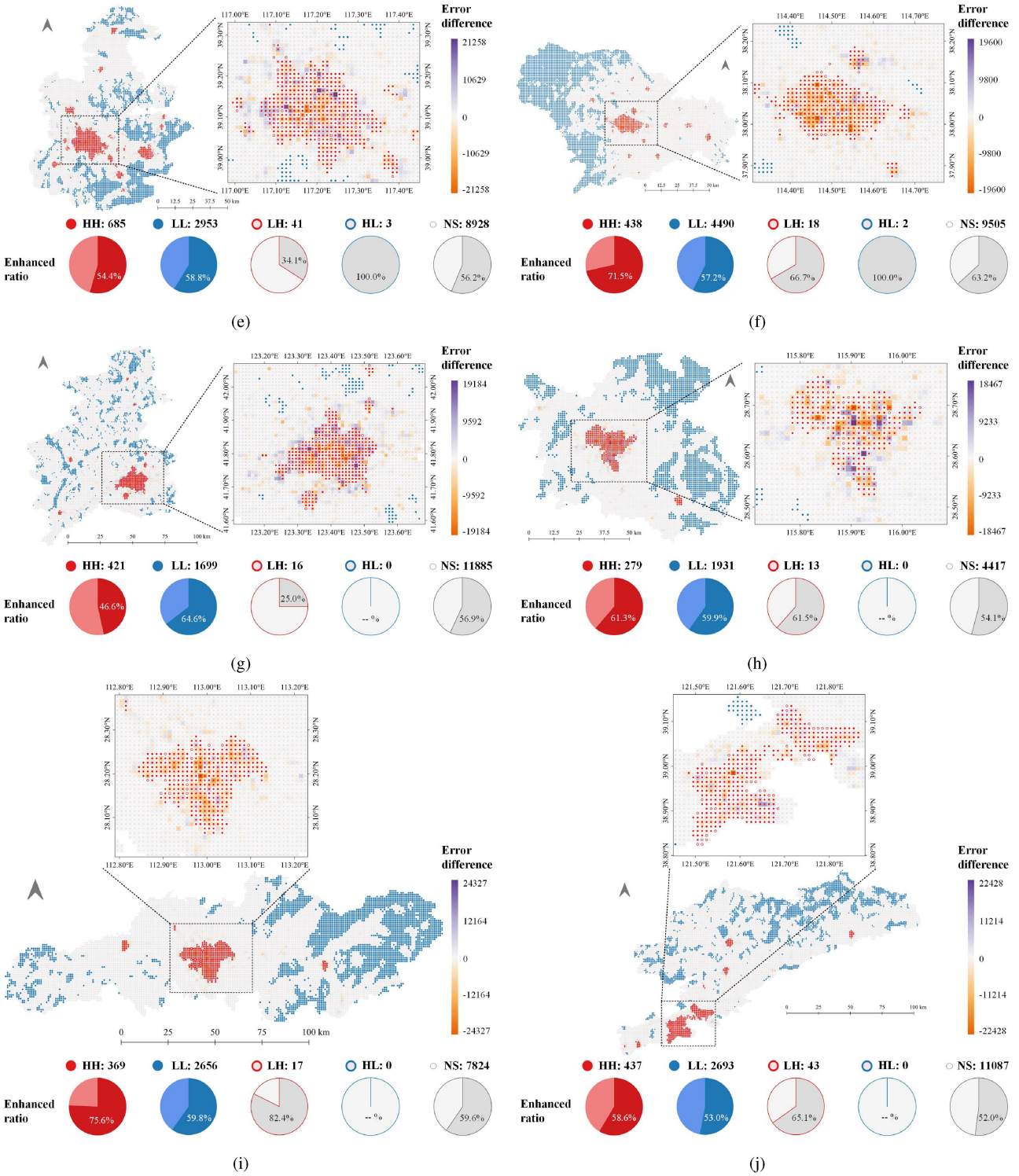


Fig. 13. (Continued.) Neighbor enhancements in all test cities: (a) Hefei, (b) Jinan, (c) Luoyang, (d) Shenzhen, (e) Tianjin, (f) Shijiazhuang, (g) Shenyang, (h) Nanchang, (i) Changsha, and (j) Dalian. Unit values are differences of absolute estimation errors of Neighbor-ResNet and ResNet ($\text{Error}_{\text{Neighbor-ResNet}} - \text{Error}_{\text{ResNet}}$). Negative values mean accuracy improvements using Neighbor-ResNet. Five dot symbols indicate five local autocorrelation types: high–high clustering with the center unit as a hot spot (HH), low–low clustering with the center as a cold spot (LL), low–high clustering with the center as a low outlier (LH), high–low clustering with the center as a high outlier (HL), and nonsignificant association with a 95% confidence interval (NS). Pie plots annotate the ratios of units with error decrease using Neighbor-ResNet. The hot spots show larger proportions of the enhancement using Neighbor-ResNet than the cold spots and nonsignificant units do, while the ratios in the outliers fluctuate.

cluster separately in the south-west and north-west zones, although they have similar construction layouts. This is highly consistent with the urban spatial arrangement: developing areas with factories and industries in the north-west and synthesis high-technological zones in the south-west [62]. Analogously in Beijing [see Fig. 6(e)], Class #10 continuously locates around the urban core and has larger proportions in the north, while its periphery is mostly covered by Class #9. It corresponds to the concentric circle development and the north-south differentiation in Beijing [63], [64]. The results demonstrate that our model is effective in distinguishing diverse landscape characteristics in Shenzhen, with the high Spearman's rank correlation coefficient (r_s) of 0.929.

B. Sources of Overall Magnitude Deviations

The demographic background of Shenzhen explains the magnitude deviations between estimated and real volumes shown in Fig. 6(d). Since the positioning data are collected from mobile networks where young adults have greater presence [59], the demographic structures influence the recorded activity magnitudes. The comparison of demographic structures in Shenzhen, Beijing, and the whole of China in Fig. 11 demonstrates the uniqueness of Shenzhen. Distinct from Beijing, as a northern metropolis that naturally grew via a long history, Shenzhen experienced a rapid urbanization in the past 35 years and formed a more youthful demographic structure. It was voted as China's Most Dynamic City and the City Most Favored by Migrant Workers in 2014 [65] and boasted a population of over 10 million people in 2016. For cities with such positively skewed age distributions, the recorded activity intensities tend to have over two times greater values than those of most cities. The infrequent large volumes of active population are hard to predict based on the learnt knowledge. This partly explains why Shenzhen shows general underestimations.

C. Sources of Local Deviations in the South Center

The geographical location of city center partly explains the local deviations. Shenzhen is a southern coastal city, extending from east to west. The downtown districts (Luohu, Futian, and Nanshan) are located along the south boundary adjacent to Shenzhen Bay and the north of Hong Kong [see Fig. 6(f)]. This leads to discontinuous transformation from downtown areas to natural scenes in the north-south direction. The RS observations, including neighbor water areas of Shenzhen Bay and green covers in the north of Hong Kong, show negative indicators for volume increasing based on the knowledge of learned general associations. This further illustrates that, while integrating neighbor knowledge enhances the estimate performance [see Fig. 7(d)], prior knowledge of directional allocations of neighbor weights may be beneficial for the estimates.

Architecture layouts also presumably influence the estimates. Shenzhen has developed simpler architectures comprising many high rises after rapid growth, without various large and regular-shaped workshops or greenhouses as in Beijing [compared in Fig. 6(a) and (b)]. As analyzed in Section III-D, building heights in the 2-D RS images are harder to recognize than building the footprint areas. Therefore, large population concentrations in high buildings tend to be underestimated. Enriching input features or adding training samples of special cities that have experienced rapid urbanization analogous to that of Shenzhen may be promising ways to improve the accuracy and enhance the model in the future.

Appendix D
Supplementary Figures

Spatial distributions of estimation results and neighbor enhancements in all test cities are shown in Figs. 12 and 13, respectively.

Acknowledgment

The authors would like to thank G. Xiu for his constructive suggestions on this article, and would also like to appreciate the comments from anonymous reviewers and editors.

References

[1] E. S. Parish, E. Kodra, K. Steinhaeuser, and A. R. Ganguly, "Estimating future global per capita water availability based on changes in climate and population," *Comput. Geosci.*, vol. 42, pp. 79–86, May 2012.
[2] D. Ehrlich *et al.*, "Remote sensing derived built-up area and population density to quantify global exposure to five natural hazards over time," *Remote Sens.*, vol. 10, no. 9, Aug. 2018, Art. no. 1378.
[3] S. Leyk *et al.*, "The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use," *Earth Syst. Sci. Data*, vol. 11, no. 3, pp. 1385–1409, Sep. 2019.
[4] E. A. Bright, P. R. Coleman, and J. E. Dobson, "Landscan: A global population database for estimating populations at risk," *Photogrammetric Eng. Remote Sens.*, vol. 66, pp. 301–314, Mar. 2003.
[5] R. Li *et al.*, "Simple spatial scaling rules behind complex cities," *Nature Commun.*, vol. 8, no. 1, Nov. 2017, Art. no. 1841.
[6] Y. Liu *et al.*, "Social sensing: A new approach to understanding our socioeconomic environments," *Ann. Assoc. Amer. Geographers*, vol. 105, no. 3, pp. 512–530, Apr. 2015.
[7] Y. Lu and Y. Liu, "Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies," *Comput., Environ. Urban Syst.*, vol. 36, no. 2, pp. 105–108, Mar. 2012.
[8] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A new insight into land use classification based on aggregated mobile phone data," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 9, pp. 1988–2007, May 2014.
[9] J. Kang, M. Krner, Y. Wang, H. Taubenbck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 44–59, Nov. 2018.
[10] S. Srivastava, J. E. V. Muoz, S. Lobry, and D. Tuia, "Fine-grained landuse characterization using ground-based pictures: A deep learning solution based on globally available data," *Int. J. Geographical Inf. Sci.*, vol. 34, no. 6, pp. 1117–1136, Nov. 2018.
[11] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, Nov. 2007.
[12] R. T. Ilieva and T. Mcphearson, "Social-media data for urban sustainability," *Nature Sustainability*, vol. 1, no. 10, pp. 553–565, Oct. 2018.
[13] P. Deville *et al.*, "Dynamic population mapping using mobile phone data," *Proc. Nat. Acad. Sci.*, vol. 111, no. 45, pp. 15 888–15 893, Oct. 2014.
[14] N. A. Wardrop *et al.*, "Spatially disaggregated population estimates in the absence of national population and housing census data," *Proc. Nat. Acad. Sci.*, vol. 115, no. 14, pp. 3529–3537, Mar. 2018.
[15] C. Kang, Y. Liu, X. Ma, and L. Wu, "Towards estimating urban population distributions from mobile call data," *J. Urban Technol.*, vol. 19, no. 4, pp. 3–21, Oct. 2012.
[16] A. Deuker, "Del 11.2: Mobility and LBS," *FIDIS Deliverables*, vol. 11, no. 2, Jul. 2008.
[17] B. Guo, R. Fujimura, D. Zhang, and M. Imai, "Design-in-play: Improving the variability of indoor pervasive games," *Multimedia Tools Appl.*, vol. 59, no. 1, pp. 259–277, Jan. 2011.

[18] Research and Markets, "Location-Based Services (LBS) Market in China 2015–2019," Jul. 2015. [Online]. Available: https://www.reportbuyer.com/product/3085882/location-based-services-lbs-market-in-china-2015-2019.html

[19] Y.-A. D. Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, no. 1, Mar. 2013, Art. no. 1376.

[20] P. Gordon, A. Kumar, and H. W. Richardson, "The influence of metropolitan spatial structure on commuting time," *J. Urban Econ.*, vol. 26, no. 2, pp. 138–151, Sep. 1989.

[21] R. Ahas, S. Silm, O. Jrv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *J. Urban Technol.*, vol. 17, no. 1, pp. 3–27, Apr. 2010.

[22] F. Zhang, L. Wu, D. Zhu, and Y. Liu, "Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns," *ISPRS J. Photogrammetry Remote Sens.*, vol. 153, pp. 48–58, Jul. 2019.

[23] F. X. Zhao, F. W. Zhao, and H. Sun, "A coevolution model of population distribution and road networks," *Physica A, Statistical Mech. Appl.*, vol. 536, Dec. 2019, Art. no. 120860.

[24] R. M. Searns, "The evolution of greenways as an adaptive urban landscape form," *Landscape Urban Planning*, vol. 33, no. 1-3, pp. 65–80, Oct. 1995.

[25] G. Manoli *et al.*, "Magnitude of urban heat islands largely explained by climate and population," *Nature*, vol. 573, no. 7772, pp. 55–60, Sep. 2019.

[26] L. Feng, H. Tian, Z. Qiao, M. Zhao, and Y. Liu, "Detailed variations in urban surface temperatures exploration based on unmanned aerial vehicle thermography," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 204–216, Dec. 2019.

[27] B. Walsh *et al.*, "Pathways for balancing CO_2 emissions and sinks," *Nature Commun.*, vol. 8, no. 1, Apr. 2017, Art. no. 14856.

[28] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[29] W. Zhao, Z. Guo, J. Yue, X. Zhang, and L. Luo, "On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery," *Int. J. Remote Sens.*, vol. 36, no. 13, pp. 3368–3379, Jul. 2015.

[30] G. J. Scott, R. A. Marcum, C. H. Davis, and T. W. Nivin, "Fusion of deep convolutional neural networks for land cover classification of high-resolution imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1638–1642, Sep. 2017.

[31] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[32] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 680–688.

[33] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.

[34] Y. Cao, X. Niu, and Y. Dou, "Region-based convolutional neural networks for object detection in very high resolution remote sensing images," in *Proc. 12th Int. Conf. Natural Comput. Fuzzy Syst. Knowl. Discovery*, Aug. 2016, pp. 548–554.

[35] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[36] D. Zhu, X. Cheng, F. Zhang, X. Yao, Y. Gao, and Y. Liu, "Spatial interpolation using conditional generative adversarial neural networks," *Int. J. Geographical Inf. Sci.*, vol. 34, no. 4, pp. 735–758, Apr. 2019.

[37] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3007029.

[38] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Econ. Geography*, vol. 46, Jun. 1970, Art. no. 234.

[39] J. E. Steele *et al.*, "Mapping poverty using mobile phone and satellite data," *J. Roy. Soc. Interface*, vol. 14, no. 127, Feb. 2017, Art. no. 20160690.

[40] P. Chhetri, R. J. Stimson, and J. Western, "Modelling the factors of neighbourhood attractiveness reflected in residential location decision choices," *Chiikigaku Kenkyu (Studies in Regional Science)*, vol. 36, no. 2, pp. 393–417, 2006.

[41] T. V. T. Nguyen, H. Han, and N. Sahito, "Role of urban public space and the surrounding environment in promoting sustainable development from the lens of social media," *Sustainability*, vol. 11, no. 21, 2019, Art. no. 5967.

[42] D. Houston, "Implications of the modifiable areal unit problem for assessing built environment correlates of moderate and vigorous physical activity," *Appl. Geography*, vol. 50, pp. 40–47, Jun. 2014.

[43] E. C. Rodrigues and R. Assuno, "Bayesian spatial models with a mixture neighborhood structure," *J. Multivariate Anal.*, vol. 109, pp. 88–102, Aug. 2012.

[44] R. White, I. Uljee, and G. Engelen, "Integrated modelling of population, employment and land-use change with a multiple activity-based variable grid cellular automaton," *Int. J. Geographical Inf. Sci.*, vol. 26, no. 7, pp. 1251–1280, Jul. 2012.

[45] *Tencent Announces 2016 First Quarter Results*, Tencent, May 2016. [Online]. Available: https://www.tencent.com/en-us/investors/financial-news.html

[46] C. Smith, Tencent statistics and facts, DMR, Jan. 2020. [Online]. Available: https://expandedramblings.com/index.php/tencent-statistics/

[47] H. Millward, J. Spinney, and D. Scott, "Active-transport walking behavior: Destinations, durations, distances," *J. Transp. Geography*, vol. 28, pp. 101–110, Apr. 2013.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[49] L. Chen, Y. Gao, D. Zhu, Y. Yuan, and Y. Liu, "Quantifying the scale effect in geospatial big data using semi-variograms," *PLOS ONE*, vol. 14, no. 11, Nov. 2019, Art. no. e0225139.

[50] B. Jiang, "Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution," *Professional Geographer*, vol. 65, no. 3, pp. 482–494, Aug. 2013.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2017, pp. 618–626.

[52] D. Sculley, "Web-scale k-means clustering," in *Proc. Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 1177–1178.

[53] L. Anselin, "Local indicators of spatial association-LISA," *Geographical Anal.*, vol. 27, no. 2, pp. 93–115, Sep. 2010.

[54] M. Fujita and J.-F. Thisse, "Economics of agglomeration," *J. Japanese Int. Econ.*, vol. 10, no. 4, pp. 339–378, Dec. 1996.

[55] X. Wang, X. Yu, and F. Ling, "Building heights estimation using ZY3 data—A case study of Shanghai, China," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2014, pp. 1749–1752.

[56] P. Gong, Z. Li, H. Huang, G. Sun, and L. Wang, "ICESat GLAS data for urban environment monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1158–1172, Mar. 2011.

[57] G. A. Licciardi, A. Villa, M. D. Mura, L. Bruzzone, J. Chanussot, and J. A. Benediktsson, "Retrieval of the height of buildings from worldview-2 multi-angular imagery using attribute filters and geometric invariant moments," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 71–79, Feb. 2012.

[58] X. Li, Y. Zhou, P. Gong, K. C. Seto, and N. Clinton, "Developing a method to estimate building height from sentinel-1 data," *Remote Sens. Environ.*, vol. 240, Apr. 2020, Art. no. 111705.

[59] K. Zickuhr, "Location-based services," Internet and American Life Project, Pew Research Center, Washington, DC, USA, 2013.

[60] Y. Yuan, Y. Lu, T. E. Chow, C. Ye, A. Alyaqout, and Y. Liu, "The missing parts from social media-enabled smart cities: Who, where, when, and what?" *Ann. Amer. Assoc. Geographers*, vol. 110, no. 2, pp. 462–475, Aug. 2019.

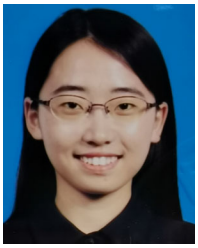
[61] L. Anselin and A. Getis, "Spatial statistical analysis and geographic information systems," in *Perspectives on Spatial Data Anal*. Berlin, Germany: Springer, Aug. 2008, pp. 35–47.

[62] W. Tu *et al.*, "Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns," *Int. J. Geographical Inf. Sci.*, vol. 31, no. 12, pp. 2331–2358, Jul. 2017.

[63] H. Ju *et al.*, "Driving forces and their interactions of built-up land expansion based on the geographical detector—A case study of Beijing, China," *Int. J. Geographical Inf. Sci.*, vol. 30, no. 11, pp. 2188–2207, Mar. 2016.

[64] J. Peng, P. Xie, Y. Liu, and J. Ma, "Urban thermal environment dynamics and associated landscape pattern factors: A case study in the Beijing metropolitan region," *Remote Sens. Environ.*, vol. 173, pp. 145–155, Feb. 2016.

[65] *Shenzhen Population*. World Population Review, (2020). Accessed: Feb. 18, 2020. [Online]. Available: https://worldpopulationreview.com/world-cities/shenzhen-population



Xiaoyue Xing received the B.S. degree from the Faculty of Geographical Science, Beijing Normal University, Beijing, China, in 2018. She is currently working toward the master's degree with the Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing.

Her current research focuses on social sensing, spatial analysis, and deep learning techniques.

Zhou Huang received the B.Sc. degree in geographical information systems (GIS) and the Ph.D. degree in cartography and GIS from Peking University, Beijing, China, in 2004 and 2009, respectively.

He is currently an Associate Professor of GIScience with the Institute of Remote Sensing and Geographical Information Systems, Peking University. In addition, he serves as the Deputy Director with the Institute of Remote Sensing and GIS, Peking University; the Beijing Key Laboratory of Spatial Information Integration and Its Applications; and the Engineering Research Center of Earth Observation and Navigation, Ministry of Education, China. He has authored and co-authored more than 50 academic papers in international journals or conferences. His current research interests include big geo-data, high-performance geocomputation, distributed geographic information processing, spatial data mining, and spatial database.

Dr. Huang was selected for the Youth Talent Innovation Plan in Remote Sensing Science and Technology, in 2015, funded by the Ministry of Science and Technology of China.

Ximeng Cheng received the B.S. and M.S. degrees from the China University of Geosciences, Beijing, China, in 2013 and 2016, respectively, the Ph.D. degree in cartology and GIS from Peking University, Beijing, China, in 2020.

His research interests include GIScience, spatiotemporal data mining, GeoAI, and urban studies.

Di Zhu received the Ph.D. degree in cartology and geographical information systems from Peking University, Beijing, China, in 2020.

His research interests include GIScience, GeoAI, geospatial analysis, and social sensing.

Chaogui Kang received the B.S. degree in geographic information science from Nanjing University, Jiangsu, China, in 2009, and the Ph.D. degree in cartography and geographic information systems from Peking University, Beijing, China, in 2015.

Since April 2015, he has been an Assistant Professor, and then, an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University, Hubei, China. His research interest include the intersection of travel behavior, built environment, and social inequality. Since July 2019, he has been a Visiting Scholar with the Center for Urban Science and Progress, New York University, New York, NY, USA.

Fan Zhang received the B.S. degree from Beijing Normal University, Zhuhai, China, in 2012. He received the M.Sc. and Ph.D. degrees from the Chinese University of Hong Kong, Hong Kong, in 2013 and 2017, respectively.

His research interests include spatiotemporal data mining, computer vision, and data-driven urban studies.

Yu Liu received the B.S., M.S., and Ph.D. degrees from Peking University, Beijing, China, in 1994, 1997, and 2003, respectively.

He is currently a Professor with the Institute of Remote Sensing and Geographical Information Systems, Peking University. His research interest mainly concentrates in humanities and social science based on big geo-data.