UVLens: Urban Village Boundary Identification and Population Estimation Leveraging Open Government Data

LONGBIAO CHEN, CHENHUI LU, FANGXU YUAN, and ZHIHAN JIANG, Xiamen University, China
LEYE WANG, Peking University, China
DAQING ZHANG, Peking University, China and Telecom SudParis, France
RUIXIANG LUO, XIAOLIANG FAN, and CHENG WANG*, Xiamen University, China

Urban villages refer to the residential areas lagging behind the rapid urbanization process in many developing countries. These areas are usually with overcrowded buildings, high population density, and low living standards, bringing potential risks of public safety and hindering the urban development. Therefore, it is crucial for urban authorities to identify the boundaries of urban villages and estimate their resident and floating populations so as to better renovate and manage these areas. Traditional approaches, such as field surveys and demographic census, are time consuming and labor intensive, lacking a comprehensive understanding of urban villages. Against this background, we propose a two-phase framework for urban village boundary identification and population estimation. Specifically, based on heterogeneous open government data, the proposed framework can not only accurately identify the boundaries of urban villages from large-scale satellite imagery by fusing road networks guided patches with bike-sharing drop-off patterns, but also accurately estimate the resident and floating populations of urban villages with a proposed multi-view neural network model. We evaluate our method leveraging real-world datasets collected from Xiamen Island. Results show that our framework can accurately identify the urban village boundaries with an IoU of 0.827, and estimate the resident population and floating population with R^2 of 0.92 and 0.94 respectively, outperforming the baseline methods. We also deploy our system on the Xiamen Open Government Data Platform to provide services to both urban authorities and citizens.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: urban village, population estimation, heterogeneous data, urban computing

*This is the corresponding author.

Authors' addresses: Longbiao Chen, longbiaochen@xmu.edu.cn; Chenhui Lu, chenhuilu@stu.xmu.edu.cn; Fangxu Yuan, fxyuan@stu.xmu.edu.cn; Zhihan Jiang, zhihanjiang@stu.xmu.edu.cn, Xiamen University, Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China; Leye Wang, leyewang@pku.edu.cn, Peking University, Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education, Department of Computer Science and Technology, Peking University, Beijing, China; Daqing Zhang, dqzhang@sei.pku.edu.cn, Peking University, Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education, Department of Computer Science and Technology, Peking University, Beijing, China and Telecom SudParis, Institut Mines, Telecom SudParis, Evry, France; Ruixiang Luo, ruixiangluo@stu.xmu.edu.cn; Xiaoliang Fan, fanxiaoliang@xmu.edu.cn; Cheng Wang, cwang@xmu.edu.cn, Xiamen University, Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China.

2474-9567/2021/6-ART57 $15.00
https://doi.org/10.1145/3463495

ACM Reference Format:
Longbiao Chen, Chenhui Lu, Fangxu Yuan, Zhihan Jiang, Leye Wang, Daqing Zhang, Ruixiang Luo, Xiaoliang Fan, and Cheng Wang. 2021. UVLens: Urban Village Boundary Identification and Population Estimation Leveraging Open Government Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 57 (June 2021), 26 pages. https://doi.org/10.1145/3463495

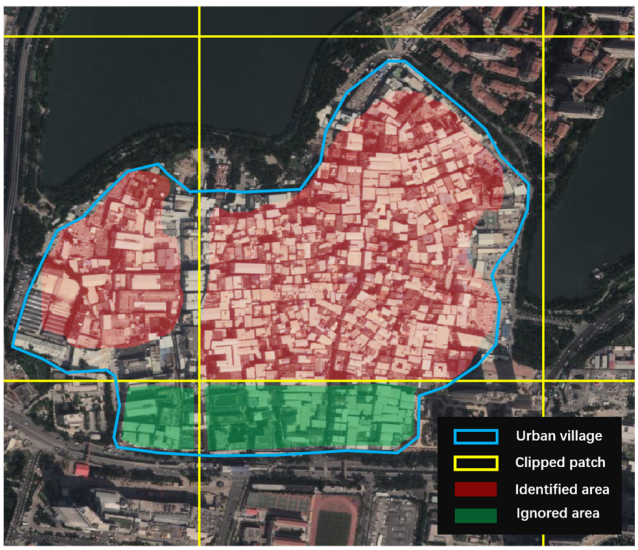
1 INTRODUCTION

Urban villages emerge with the rapid urbanization process in many developing countries. They refer to the residential areas that are lagging behind the pace of city development, lack of management of modern cities, and with low living standards [22]. In China, the key issues of urban villages are *overcrowded buildings* and *high population density* [11]. A large number of non-local graduates and many migrant workers, due to low income, often choose to live in the low-rent houses of urban villages. In order to earn more income from house rent, the landlords of these houses often illegally add floors and expand rooms in the absence of planning and management. As a result, there are an increasing number of houses and residents in urban villages. The overcrowded and poor environment of urban villages bring serious social risks to the city, including potential epidemic risks, frequent fire alarms, and high crime rates [3]. Therefore, accurately identifying the geographic boundaries of urban villages and estimating their populations are crucial for the urban authorities to renovate these urban villages, and to eliminate potential social risks.

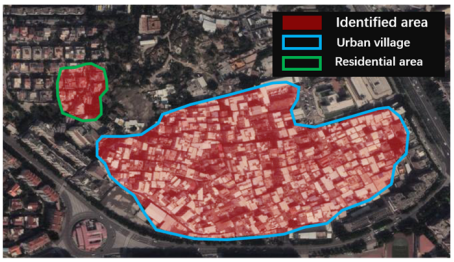
Traditionally, identifying urban village boundaries mainly relies on field surveys of city planners, which usually is time consuming and human intensive. Moreover, since urban villages are constantly being renovated, it is difficult for traditional methods to capture these changes timely and accurately. Recently, with the increasing availability of high-resolution remote sensing data, researchers have used different strategies to locate urban villages from satellite imagery. A common approach is to ask city planning experts to label the boundaries of urban villages and train a convolutional neural network to extract features for identifying the boundaries of urban villages [5]. However, two of the key issues exist in this method:

- *Incorrect boundaries.* City-widely high-resolution satellite imagery is usually very large, making it computationally intractable to directly process such a large image for identification tasks. Therefore, previous works usually clip satellite imagery into fixed-size patches for training [5]. However, such a method may clip an integrated urban village into several patches, which ruin the integrity of urban villages. Consequently, the identified urban villages with this method may be divided into small pieces by artificial boundaries between adjacent patches and the small areas in a patch may be incorrectly ignored, as illustrated in Figure 1(a). Therefore, generating patches to keep integrity of urban villages for accurate boundary identification is challenging.
- *Identification errors.* Existing methods merely rely on imagery features to characterize urban village boundaries. However, without sufficient prior knowledge, similar function areas (e.g., old residential area) are likely to be misidentified as urban villages, as shown in Figure 1(b). Therefore, reducing identification errors caused by similar imagery features is also challenging.

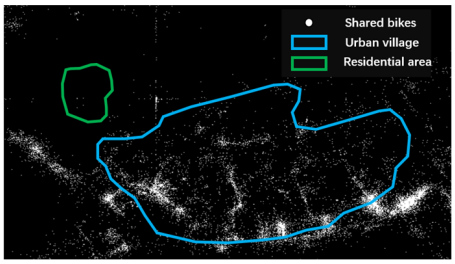
In order to solve the above-mentioned problems, we propose to incorporate heterogeneous crowdsensing data to identify urban village boundaries. Based on the prior knowledge that urban villages are usually surrounded by city road networks, we propose to segment a large city-wide satellite imagery into patches guided by city road networks. However, directly using the road topology to segment the city-wide imagery into properly-sized patches for urban village boundary identification is not realistic. Since the topology of city road networks is hierarchical, the patches segmented by urban trunk roads are usually too large (as shown in Figure 2(a)), while the patches segmented by fine-grained urban branch roads tend to be too small and may divide urban village into pieces (as shown in Figure 2(b)). In contrast, taxi trajectories not only successfully recover major road networks surrounding urban villages, but also avoid dividing urban village into pieces as taxis are usually not allow to



(a) Identified urban villages using fixed-size clipping method



(b) Urban village and residential area



(c) Bike-sharing drop-off records

Fig. 1. Figure (a) shows the identified results using fixed-size clipping method, where the blue line surrounds the ground-truth of urban village, the yellow boxes represent the clipped patches, and the red masks and the green masks are the identified areas and the ignored areas respectively. Figure (b) and Figure (c) are the comparison of urban village area and residential area in satellite imagery and bike-sharing drop-off image.

(a) Urban trunk roads (left) and taxi trajectories-based roads (right)

(b) Urban branch roads (left) and taxi trajectories-based roads (right)

Fig. 2. The comparison between road topology and taxi trajectories. The red mask represents the urban village areas and the yellow line denotes the urban roads. Figure (a) shows the comparison of area divided by urban trunk roads and trajectories-based roads, and Figure (b) shows the comparison of area divided by urban branch roads and taxi trajectories-based roads.

enter urban villages due to their narrow roads (as shown in Figure 2). Meanwhile, we note that due to the rapid urban renovation process in China, the boundaries of urban villages are constantly shrinking and even vanishing, resulting in dynamic changes of the road networks around these urban villages, as shown in Figure 13 of case studies. Regularly using taxi trajectories to update road networks is helpful to capture these changes. Therefore,

we extract road networks from large-scale crowdsensing taxi trajectories to segment satellite imagery [41]. In this way, these semantically clipped patches can effectively keep the integrity of urban villages and potentially avoid incorrect boundaries. Meanwhile, we note that commercial shared bike (e.g., Mobikes[1]) is one of the common transportation means of urban village residents and widely found in urban villages, but these bikes are usually not allowed to enter planned function areas, such as business districts and residential areas. Consequently, the bike-sharing drop-off patterns in urban villages demonstrate significant differences from those in other function areas, as shown in Figure 1(c). Therefore, we propose to fuse the bike-sharing drop-off data and the satellite imagery to distinguish similar areas, so as to reduce misidentified boundaries.

After accurately identifying urban village boundaries, our next goal is to assess the living environment and social risks. One of the key indices is the accurate population estimation in urban villages. Although governments and city administrations regularly conduct demographic censuses to collect urban population data, this data are not frequently updated due to the substantial amount of time and human labor consume. For example, in China, the demographic census happens every 10 years[2]. However, since the boundaries of urban villages are constantly changing and their residents are frequently migrating, their census-based population data tend to be outdated and not accurate. Therefore, we need a new approach to accurately estimate population in urban villages in a low-cost and near-real-time manner. In the literature, estimating population from satellite imagery has been studied in [33]. This method can extract regular residential buildings from satellite imagery, and count the number of dwelling units in planed residential areas to estimate population. However, this method does not work well in urban village population estimation due to the following reasons. First, the buildings in urban village vary in size and height. Due to lack of management, most of the buildings in urban villages are illegally extended by landlords [11]. Second, in many buildings of urban villages, a normal suite may be divided into more rooms to host more people [19]. Therefore, simply counting the number of buildings in urban villages from satellite imagery may ignore these extra tenants, leading to inaccurate population estimation.

In this work, we address the above-mentioned challenges in two folds. In order to accurately estimate the population base from the static view of living space, we conducted a series of field studies in several urban villages to identify their typical building types, and estimate population capacity of building according to the derived empirical formula. In order to estimate the floating population (e.g., migrant workers) from the dynamic view of human activities, we seek to incorporate heterogeneous urban crowdsensing data to capture the intensities of human mobility and commercial activities, so as to improve the accuracy of population estimation. Particularly, we note that human mobility intensity in urban villages can be represented by bike-sharing drop-off count, since they are common means of transportation in the narrow streets of urban villages. Similarly, we extract the distributions of points of interest (POIs) related with daily life, e.g., convenient stores, grocery shops, restaurants, to depict the hotness of commercial activities in urban villages [14]. With these insights, we design a multi-view neural network model (MvNN) to fuse these features extracted from satellite imagery, bike-sharing drop-off and geography POIs data for population estimation. Specifically, the MvNN model estimates the population base through household capacity features from the static view, and incorporate human mobility features and commercial hotness features to further fine-tune the population from the dynamic view.

In this paper, we propose a two-phase framework for urban village boundary identification and population estimation. In the first phase, we extract city road networks from taxi GPS trajectories, and segment a large city-wide satellite imagery into small patches that are surrounded by road networks. We then ask professionals to label a set of urban villages boundaries as training data and train an instance segmentation model (Mask-RCNN) to identify urban village from each patch. In particular, in order to distinguish similar area in satellite imagery, we augment each patch with bike-sharing drop-off data as an extra imagery channel to train the instance

[1]https://mobike.com/

[2]http://www.stats.gov.cn/tjzs/cjwtjd/201308/t20130829_74322.html

segmentation model simultaneously. In the second phase, we extract three features related to urban village population from the corresponding urban data sources and propose a multi-view neural network model to estimate urban village population based on both static and dynamic views. Specifically, we identify different buildings from urban village satellite imagery using another instance segmentation model, and derive an empirical formula to obtain household capacity features. Furthermore, we extract bike-sharing drop-off data as human mobility features and POIs related with daily life as commercial hotness features. Finally, we train a multi-view neural network model to estimate urban village population, which exploits the household capacity features to estimate the population base from the static view of living space, and incorporate human mobility features and commercial hotness features to further fine-tune the population from the dynamic view of human activities.

Briefly, the contributions of this paper include:

- We investigate the program on urban village boundary identification and population estimation by fusing remote sensing satellite imagery with heterogeneous crowdsensing data, which provides a low-cost decision support for urban village renovation and risk management.
- We propose a *heterogeneous data fusion* framework to identify urban village boundaries and estimate population. Our method can not only accurately identify urban village boundaries by fusing road networks guided patches with bike-sharing drop-off patterns, but also accurately estimate the resident and floating populations leveraging three types of heterogeneous features based on both static and dynamic views, i.e., household capacity features from satellite imagery, human mobility features from bike-sharing drop-off data and commercial hotness features from POIs.
- We evaluate our approach leveraging the high-resolution remote sensing satellite imagery, the large-scale vehicle (e.g., taxi and shared bike) trajectories datasets and the POIs datasets from the Xiamen Open Government Data Platform. Results show that compared to labeled urban villages boundaries from professionals, our method successfully identifies urban village boundaries with an IoU of 0.827. Compared with official population census result, our method successfully achieves accurate resident population and floating population estimation with R^2 of 0.92 and 0.94 respectively.
- We make a real-world system deployment collaborating with the Xiamen municipal government, as shown in Figure 10. Our system is working on the government portal to provide services to query relevant statistics and information of urban villages in Xiamen.

2 PRELIMINARY AND FRAMEWORK OVERVIEW

Definition 2.1. Satellite Imagery: the satellite imagery is a photograph of Earth or other planets collected by imaging satellites in space. Satellites get multi-spectral imagery by resolving different electromagnetic wave bands. In this paper, we use visible-band spectral remote sensing satellite imagery with the resolution of 1.09 meter and 0.14 meter.

Definition 2.2. Taxi Trajectory: the taxi trajectory data we collect records taxi positions every minute, which can be described by a set of GPS points denoted by 4-tuples:

$$P = \{p | p = (u, t, lat, lng)\}$$

where u, t, lat, lng are the unique taxi ID, time stamp, latitude, and longitude from GPS transmitters.

Definition 2.3. Bike-sharing Drop-off: shared bike is a kind of popular commercial public bike that can easily be picked and returned in service area. We define the bike-sharing drop-off as a set of GPS points, which are the parking positions of shared bikes after being used. The GPS point contains the 4-tuples:

$$S = \{s | s = (v, t, lat, lng)\}$$

where v, t, lat, lng are the unique bike ID, time stamp, latitude, and longitude from GPS transmitters.

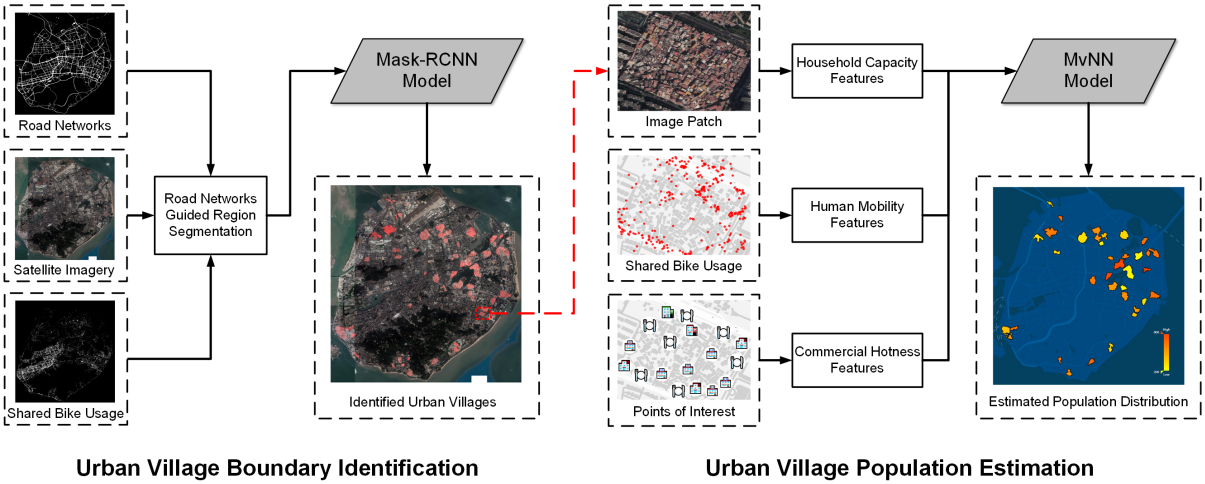


Fig. 3. Overview of the framework.

Definition 2.4. Point of Interest (POI): a point of interest is a specific point location that someone may find useful or interesting, such as hotel, hospital, restaurant, etc. This term is widely used in Geographic Information System (GIS). We can easily obtain distribution of POIs from map service.

We propose a two-phase framework to identify urban village boundaries and estimate population of each urban village, as shown in Figure 3. In the urban village identification phase, we extract city road networks from taxi trajectories, and segment the city-wide satellite imagery into small patches by the road networks. We then augment each patch with bike-sharing drop-off as an extra imagery channel to reduce probability of misidentification. Moreover, we exploit Mask-RCNN model to identify urban village boundaries in each patch, and merge all urban village masks to obtain a view of urban village distribution. In the urban village population estimation phase, we identify three types of features from the corresponding urban data sources, including household capacity features from high-resolution remote sensing satellite imagery, human mobility features from bike-sharing drop-off data, and commercial hotness features from POIs. We then feed these features to designed multi-view neural network model (MvNN) to estimate population of each urban village, and obtain a view of urban village population distribution. We elaborate the key steps of the framework in the following sections.

3 URBAN VILLAGE BOUNDARY IDENTIFICATION

In this phase, our objective is to detect and segment urban villages from city-wide remote sensing satellite imagery. In order to maintain the integrity of urban villages, we propose a road networks guided region segmentation approach to generate semantic patches. This approach is based on the prior knowledge that urban villages are located in the regions surrounded by city road networks. In order to overcome the misidentification of similar areas in satellite imagery, we incorporate bike-sharing drop-off data to generate hybrid imagery. This is because we note that commercial shared bikes are normally not allowed to enter planned housing estates, whereas they can be used in urban villages. Moreover, in order to transfer the pre-trained Mask-RCNN model to our urban village boundary identification task, we conduct two adaptions in both feature space and label space. On the one hand, we redistribute color values of original remote sensing satellite imagery to facilitate image features learning. On the other hand, we exploit crowdsourcing mechanism to label urban village labels to perform transfer learning.

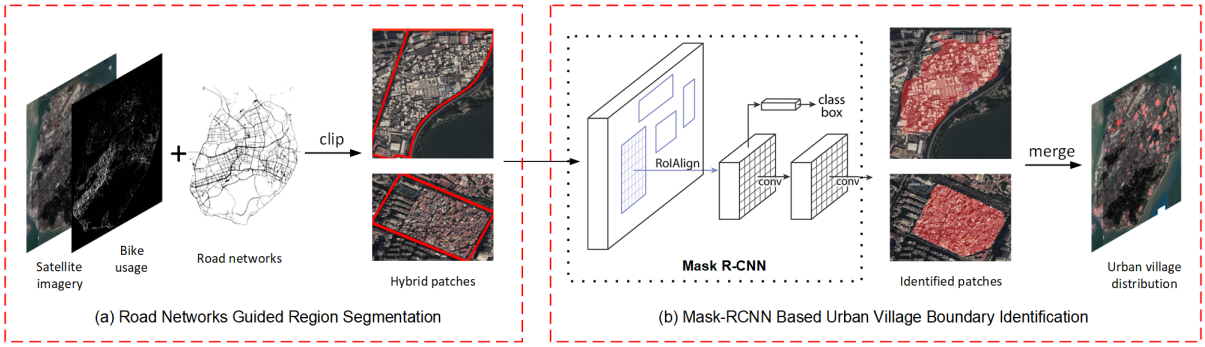


Fig. 4. The flow diagram of the urban village boundary identification approach. (a) We concatenate satellite imagery of three channels of RGB and bike-sharing drop-off gray scale image of single channel to generate the hybrid imagery, and then clip hybrid imagery to hybrid patches by road networks. The red lines in hybrid patches represent city roads. (b) We use trained Mask-RCNN model to identify urban village boundaries in each patch, and merge all urban village masks to obtain the view of urban village distribution.

The flow diagram of the identification approach is illustrated in Figure 4. We elaborate on the details of our approach as follows.

3.1 Road Networks Guided Region Segmentation

In this step, we first extract city road networks from taxi trajectories. Then, we generate the four channels hybrid imagery based on remote sensing satellite imagery and bike-sharing drop-off data. Finally, we adaptively clip the hybrid imagery into regions of different sizes based on the road networks.

More specifically, we first extract city road networks from taxi trajectories data. We grid the remote sensing satellite imagery according to the pixel ratio, and then map the GPS coordinates of each taxi into the corresponding grids, adding up to the number of taxis in each grid. We then select an appropriate threshold T to generate a binary image of road networks based on empirical studies. We note that there is no GPS positioning of taxis in urban tunnels, so we manually connect the tunnel portals and further process the image to make the urban road networks clear and obvious using morphological techniques. Second, we construct the hybrid imagery consisting of remote satellite imagery and bike-sharing drop-off data. Similarly, we use the above-mentioned method to map the bike-sharing drop-off data into grids, and count the number of bike-sharing drop-off in each grid. We then obtain a gray scale image by normalize these values into the range of 0-255. Moreover, we concatenate the three channels of the satellite imagery and the single channel of the bike-sharing drop-off gray scale image to generate the four-channel hybrid imagery. Third, we use the extracted road networks to segment the generated hybrid imagery to obtain corresponding hybrid patches. We employ the *Opencv-python* library[3] to extract and segment connected regions surrounded by road networks. the *Opencv-python* library is an effective image processing toolkit, and its *findContours* function can find closed contours in binary image based on a border following algorithm which determines the surroundness relations among the borders of a binary image by topological analysis [32]. Consequently, we clip the hybrid imagery into regions of different sizes. For example, Figure 4(a) shows the two hybrid patches by road networks guided segmentation, and the red lines in the hybrid patches represent city roads.

[3]https://pypi.org/project/opencv-python/

3.2 Mask-RCNN Based Urban Village Boundary Identification

3.2.1 Urban Village Mask Labeling. With semantic patches obtained, we then need to train a Mask-RCNN model to identify urban village boundaries. The first step is to construct an appropriate training set. A sufficient training set can help the model achieve excellent performance, while labeling urban village masks in each patch is time consuming and expertise required. Therefore, we exploit the crowdsourcing mechanism to ask a group of professionals.

First, we recruit a group of professional participants with incentives, including researchers and students in urban planning and remote sensing. We then develop a web-based crowdsourcing platform with an opensource image mask labeling tool *labelme*[4] as shown in Figure 5, which automatically assign image patches to participants. In order to ensure the quality of labeling, we design a cross-validation paradigm with field studies. Specifically, each patch is assigned to three participants to avoid labeling bias. If three participants hold the same opinion on the boundaries of urban village in the same patch, the label is considered as ground truth. Otherwise, we further conduct field studies to correct urban village boundaries. Finally, we obtain ground truth of all urban village boundaries with the help of the crowdsourcing platform.

3.2.2 Mask-RCNN Model Training and Predicting. In this step, we randomly select a set of hybrid patches to construct training set to train a Mask-RCNN model, and identify urban village boundaries in other hybrid patches using the trained model. The Mask-RCNN model is an object instance segmentation model using convolutional neural network structure, which efficiently detect objects in an image while simultaneously generating a high-quality segmentation mask for each instance [13]. Based on Faster-RCNN [26], this model proposes a RoIAlign layer to preserve pixel-to-pixel alignment between network inputs and outputs and decouple mask and class prediction to avoid competition among classes. Since the Mask-RCNN model achieves a good identification performance in instance segmentation tasks, we choose it to identify urban village boundaries.

However, traditional Mask-RCNN-based instance segmentation tasks are conducted on normal images rather than remote sensing satellite imagery. We note that the original RGB channels from remote sensing satellite imagery are generally dark because of the uneven color distribution, making it difficult for the model to effectively extract informative features. Therefore, we first exploit a histogram equalization algorithm [25] to redistribute the color values between 0-255 for the original satellite imagery, so as to facilitate Mask-RCNN to learn informative features of urban villages. Then, we randomly select the training set to train a Mask-RCNN model, and use the trained model to detect urban villages and predict their masks simultaneously in other hybrid patches, and merge all urban village masks to obtain a view of the urban village distribution.

4 URBAN VILLAGE POPULATION ESTIMATION

With the urban village boundaries identified, our next objective is to estimate the resident and floating populations of each urban village. In order to model the contextual information based on the static and dynamic views of urban village population, we extract related three types of features from heterogeneous urban data. Specifically, we first identify buildings from satellite imagery of urban villages, and derive the empirical formula to estimate the resident population base of urban village named as household capacity features. Then, we extract bike-sharing drop-off data as human mobility features and POIs related with daily life as commercial hotness features in urban villages. Furthermore, we design a multi-view neural network model (MvNN) to fuse the above-mentioned heterogeneous features. This model keeps household capacity features as the population base from the static view of living space, and incorporate human mobility features and commercial hotness features to further fine-tune the population from the dynamic view of human activities. We elaborate on the details as follows.

[4]https://pypi.org/project/labelme/

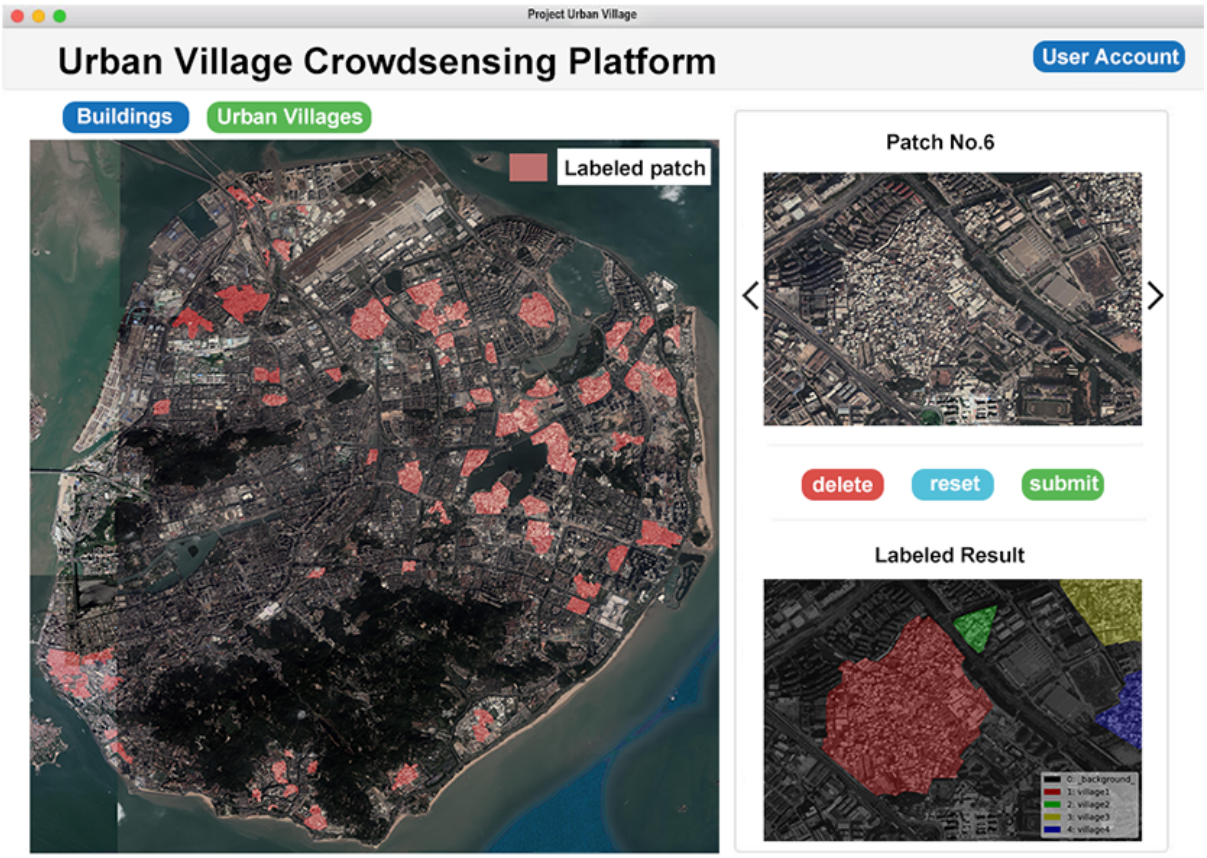


Fig. 5. The developed crowdsourcing platform for labeling tasks.

4.1 Household Capacity Features Extraction

Since most of the buildings in urban villages are built by residents themselves, and they are likely to added floors or expand rooms illegally, resulting in the buildings in urban villages with various sizes and heights. In order to accurately estimate the resident population base, we conducted a series of field studies in several urban villages to identify their typical building types. As shown in Figure 6, we divide urban village buildings in Xiamen Island into four categories:

- *Resident houses.* These buildings with red roofs are newly built or renovated in recent years. They are about 2 or 3 stories, and such building can contain 2-4 households. The residents living in these red buildings are generally large families with higher incomes, including elders, their children and children's separate families.
- *Rental apartments.* These buildings that appear bright or off-white in satellite imagery are often roofed with thermal insulation panels or cleaner cement roofs, since they may be added floor in recent years. These buildings are very commercial, and they are often converted by the owners into a large number of single rooms to rent for the greater economic benefits. The ground floors of such buildings are often converted into shops. Such buildings typically have 5-7 floors, and each floor can contain around 3 households. The

(a) Resident houses (b) Rental apartments (c) Resident apartments (d) Old houses

Fig. 6. The categories of the buildings in urban villages of Xiamen Island. The upper figures show the buildings in the satellite imagery, and the below figures are photos in field studies.

residents who rent here are usually migrant workers with low income or young people who have just started working.

- *Resident apartments*. These buildings with gray mud roofs are usually built in the last 20 to 30 years. Most of these buildings are 4-5 stories, and each floor can accommodate around 2 households. Nowadays, the residents living in these buildings are generally local residents and some migrant workers who have lived here for many years.
- *Old houses*. These low and decrepit houses with tiled roofs are typically built by local residents in the 1970s and 1980s. Such houses are usually one story and can only accommodate around 2 households. The residents who lived in these houses today tend to be elderly locals.

With the types of urban village buildings obtained, our next step is to detect and classify buildings in urban village satellite imagery. Due to the good performance of the Mask-RCNN model, we still use it to identify buildings. More specifically, we first clip identified urban village satellite imageries into fixed-size patches (i.e., 1024x1024 pixels), and randomly select N patches in all patches which include different categories of buildings. Then, we manually label the buildings in each patch, and use them to train a Mask-RCNN model. In order to improve the efficiency of labeling tasks, we also exploit the web-based crowdsourcing platform above-mentioned to outsource the label masking tasks to the professionals. Similarly, we use cross-validation paradigm with field studies to ensure the quality of ground truth. Finally, we exploit the trained model to identify buildings in each urban village satellite imagery, and obtain the number of buildings per category. Figure 7 shows the mask prediction results of each building in urban village.

Moreover, we estimate the population capacity in each building category of urban village as household capacity features. Based on the field studies and the census report from urban housing authority, we derive the empirical formula to estimate the population capacity in building category j of urban village i, defined as follows:

$$H_j^{(i)} = c_j^{(i)} h_j r \tag{1}$$

where $c_j^{(i)}$ is the number of buildings in category j of urban village i, which is derived using a Mask-RCNN model from the satellite imagery of the urban village. h_j represents the average number of households in building category j, and is compiled from our field studies in urban villages. Besides, h_j is also available from existing

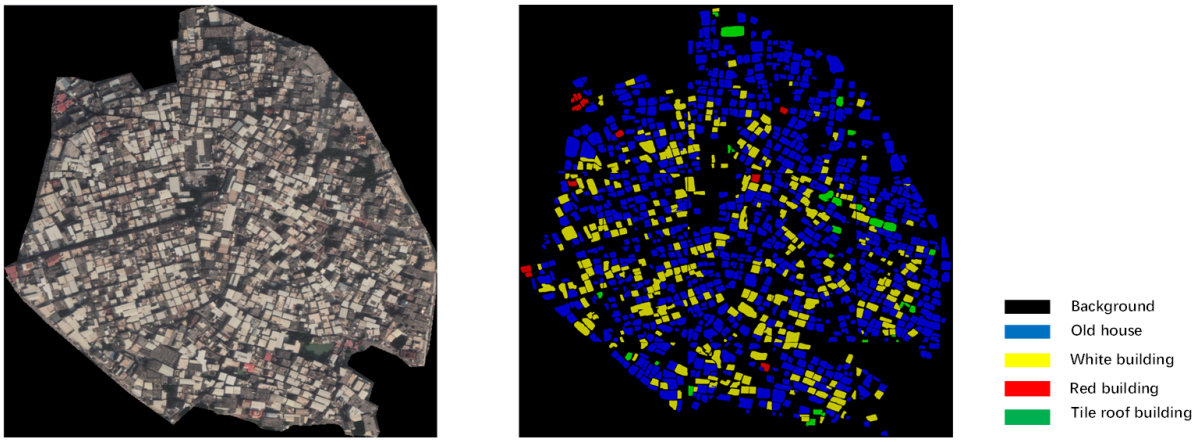


Fig. 7. The identification result of urban village buildings.

Table 1. Top-10 POI categories most relevant to urban village population with the corresponding correlation coefficient.

	Resident Population		Floating Population	
	POI categories	**correlation coefficient**	**POI categories**	**correlation coefficient**
1	Shopping	0.93	Shopping	0.84
2	Hospital	0.87	Hospital	0.74
3	Life service	0.85	Life service	0.72
4	Restaurant	0.84	Restaurant	0.72
5	Traffic facility	0.75	Traffic facility	0.67
6	Entertainment	0.66	Auto service	0.67
7	Real estate	0.66	Company	0.56
8	Auto service	0.63	Entertainment	0.56
9	Government agency	0.55	Real estate	0.54
10	Company	0.53	Government agency	0.36

literature. For example, many existing works [4, 16, 23, 39] detect and classify buildings from remote sensing satellite imagery, and there is a great deal of building information deposited in them. r denotes the average population per household in the city. For different city, r is obtained from local urban housing authority (e.g., Xiamen statistics bureau[5]). Specifically, the household capacity features of the urban village i are formulated as $H^{(i)} = (H_1^{(i)}, H_2^{(i)}, \ldots, H_m^{(i)})^T$, where i is the index of urban village and m is the number of building types in urban village.

4.2 Human Mobility Features and Commercial Hotness Features Extraction

In this step, we first extract the human mobility features from bike-sharing drop-off data. A greater frequency of shared bike usage in an urban village may represent a greater population in this urban village. Therefore, we analyze the daily shared bike usage in urban village to describe the human mobility features. Specifically, we map the GPS longitude and latitude coordinates of shared bikes to the interior of urban villages, and count the

[5]http://tjj.xm.gov.cn

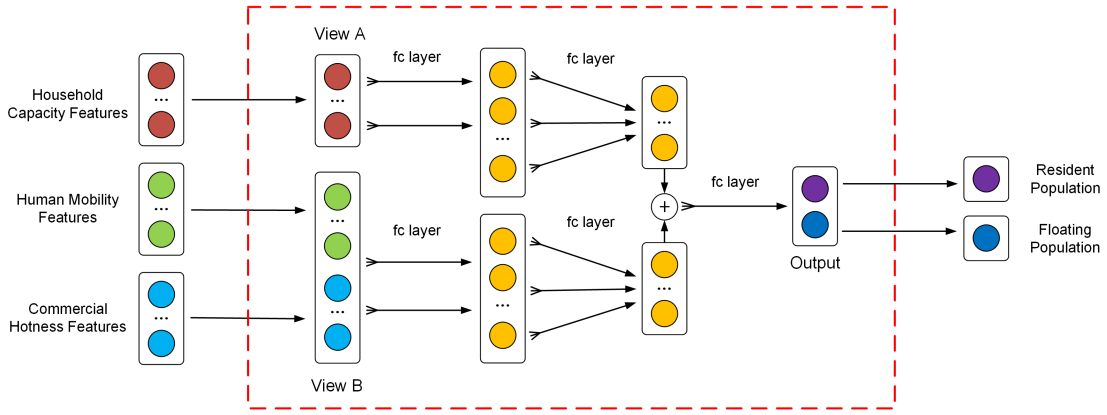


Fig. 8. The framework of the multi-view neural network model.

number of them. Based on the historical bike-sharing drop-off data, we observe that the human mobility patterns are highly dynamic in different temporal contexts. We propose to characterize the human mobility features of each urban village using a temporal-context-based profile. Specifically, given an urban village and its historical bike-sharing drop-off intensity vector measured in a week, we aggregate and average daily shared bike usage to build a typical weekly human mobility profile. The human mobility features of the urban village i are formulated as $U^{(i)} = (U_1^{(i)}, U_2^{(i)}, \ldots, U_7^{(i)})^T$, where $U_j^{(i)}$ is the average number of bike-sharing drop-off records of day j in a week in urban village i.

Second, we extract the commercial hotness features of urban villages from the categorical distribution of POIs. We note that an urban village with lots of shops and restaurants may represent more people living there. However, not all POI categories are highly correlated with the population of urban villages, so we need to select appropriate POI categories to participate in population estimation. Specifically, we retrieve all POIs within each urban village area, and group them into a set of POI categories. For each category, we measure their correlation by computing the Spearman's correlation coefficient [37] between the number of POIs and the population of urban villages. We take into account the top-n POI categories and count the number of the POIs of each category as a regional commercial hotness features. The commercial hotness features of the urban village i are formulated as $C^{(i)} = (C_1^{(i)}, C_2^{(i)}, \ldots, C_n^{(i)})^T$, where $C_j^{(i)}$ is the number of POIs of category j in urban village i and n is the number of POIs categories. In this paper, we conduct correlation analysis for resident population and floating population respectively. Table 1 shows the top-10 POIs categories that are most relevant to the resident population and floating population of urban villages. We can see that there are the same top-5 POI categories for resident population and floating population, which demonstrate these categories are essential to people's daily life. In addition, we find that the correlation coefficient in resident population is generally higher than that in floating population, which mean that the commercial hotness features have a greater impact on the resident population compared to the floating population.

4.3 Population Estimation

In this step, we use a regression model to estimate resident population and floating population in urban villages based on three types of extracted features. Intuitively, we can directly concatenate these features as a vector to train a regression neural network to estimate population. However, this method does not work well, since it ignores the different roles of these features in our population estimation task. The household capacity features

estimate the population base of urban villages to some extent, which reasonably reflect population under normal living situation from the static view of living space. Nevertheless, different urban villages with similar building distribution may have different populations. For example, urban villages in prosperous areas where more tenants choose to live have more population, while urban villages in remote areas have relative smaller population. Thus, from the dynamic view of human activities, we involve human mobility features and commercial hotness features to further fine-tune the population and improve the accuracy of estimation.

With the perspective of static living space and dynamic human activities, we design a multi-view neural network model (MvNN), as shown in Figure 8. In this model, household capacity features are extracted separately as view A = $(H_1^{(i)}, H_2^{(i)}, \ldots, H_m^{(i)})^T$, and human mobility features and commercial hotness features are concatenated as another view B = $(U_1^{(i)}, U_2^{(i)}, \ldots, U_7^{(i)}, C_1^{(i)}, C_2^{(i)}, \ldots, C_n^{(i)})^T$, where i is the index of urban village, m is the number of urban village building types, and n is the number of POI categories. For each view, we use two fully-connected layers to produce the predicted vector with the same size. Then, we add the two predicted vectors, and connect a full-connected layer to predict the final resident and floating populations. In addition, we add dropout layer after each full-connected layer, which discard part of neurons in the full-connected layer according to a certain probability to avoid over-fitting. Meanwhile, we use ReLU as the activation function in each full-connected layer, and exploit Adam optimizer to converge the model.

5 EVALUATION

In this section, we first introduce the experiment settings, and then present the evaluation results on urban village boundary identification and population estimation. We also conduct a series of case studies to demonstrate the effectiveness and deficiency of our method.

5.1 Experiment Settings

5.1.1 Datasets. We evaluated our framework based on the real-world open government data in Xiamen in 2017. We collected high-resolution remote sensing satellite imagery of Xiamen Island, large scale vehicle (e.g., taxi, shared bike) trajectories datasets, POIs datasets and population dataset from the Xiamen Open Government Data Platform, as summarized in Table 2. The dataset details are elaborated as follows.

Satellite imagery. In order to meet the different resolution requirements of urban village boundary identification and building detection, we obtained two different high-resolution remote sensing satellite imagery of Xiamen Island. The resolution of satellite imagery to identify urban village boundaries and detect buildings of urban villages are 1.09 meter and 0.14 meter, respectively.

Vehicle GPS data. The taxi trajectories dataset contains GPS trajectories of 5,486 taxis reported every 1 minute during September 2017, and the bike-sharing drop-off dataset contains usage records of 93,071 shared bikes from June 2017 to October 2017.

Point of interest (POI). We obtained all POIs distributions in Xiamen, which contains 47,439 POIs. We group all POIs into fifteen categories, including *Hotel and Hostel*, *Restaurant*, *Road and Street*, *Real estate*, *Company*, *Shopping*, *Traffic facility*, *Finance institution*, *Tourist attraction*, *Auto service*, *Business building*, *Life service*, *Entertainment*, *Hospital*, *Government agency*.

Population data. The population dataset is collected from demographic census in 2017 in Xiamen by urban authorities for facilitating the BRICS summit, including resident population and floating population for each urban village.

5.1.2 Evaluation Metric. We compared the segmented urban villages with the ground-truth dataset to evaluate the detection accuracy and the segmentation accuracy of identification method.

Table 2. Summary of Datasets

Data type	Item	Value
Satellite imagery	Resolution	1.09 meter / 0.14 meter
	Collection time	11/14/2017
Taxi trajectories data	The number of taxis	5,486
	Sampling rate	every minute
	Collection time	09/01/2017-9/30/2017
Bike-sharing drop-off data	The number of bikes	93,071
	Collection time	06/01/2017-10/31/2017
Points of interest	The number of POIs	47,439
	The number of categories	15
	Collection time	11/14/2017
Urban village population data	The number of urban villages	61
	Total resident population	584,579
	Total floating population	233,725
	Collection time	08/01/2017
Geographic coverage area	Southwest: [24.423240, 118.064736], Northeast: [24.561492, 118.198513]	

Detection accuracy. If a detected urban village has a spatial overlapping with urban villages in the ground-truth dataset, we call it a *true positive* (TP), and otherwise a *false positive* (FP). For an urban village in the ground-truth dataset that is not detected, we call it a *false negative* (FN). Based upon this, the precision and recall are calculated as follows:

$$precision = \frac{|TP|}{|TP| + |FP|}, \quad recall = \frac{|TP|}{|TP| + |FN|}, \quad F1\text{-}Score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{2}$$

Segmentation accuracy. We adopt the popular Intersection over Union (IoU) metric to evaluate the segmentation accuracy over the city-wide imagery, as follows:

$$IoU = \frac{|\{\text{ground-truth pixel}\} \cap \{\text{detected pixel}\}|}{|\{\text{ground-truth pixel}\} \cup \{\text{detected pixel}\}|} \tag{3}$$

Population estimation accuracy. We use three commonly used regression model metrics to evaluate the performance of population estimation method, including 1) Root Mean Squared Error (RMSE), 2) Mean Absolute Percentage Error (MAPE), 3) R^2 score. They are defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|, \quad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

where $\hat{y}_i$ and y_i are the predicted population and ground-truth population in urban village i respectively, n is the number of all urban villages, and $\bar{y}$ stands for the average value of all urban village population.

5.1.3 Baseline Methods. We compared our method with various baseline methods with regard to urban village boundary identification and population estimation. For urban village boundary identification, we compared our

proposed method fusing road networks and bike-sharing drop-off data based on Mask-RCNN model (RNBS-MR) with following baselines.

MiSR. This baseline method uses a multi-index scene representation model (MiSR) to detect urban villages [15]. Specifically, it divides city-wide satellite imagery to small patches (i.e., 120x120 pixels) with an overlapping (i.e., 60 pixels), then constructs characteristic histograms through extracting buildings and vegetation information by MBI [16] and NDVI respectively. Furthermore, it exploits SVM to classify each image patch.

MNC. This baseline method identifies urban village boundaries from city-wide satellite imagery using a multi-task network cascades model (MNC) [7]. Specifically, it clips city-wide satellite imagery to fixed-size patches (i.e., 512x512 pixels) and then trains a MNC model to identify urban village boundaries from each patch.

MR. This baseline method identifies urban village boundaries from city-wide satellite imagery using Mask-RCNN model [5]. Similarly, it clips city-wide satellite imagery to fixed-size patches (i.e., 512x512 pixels) and then trains a Mask-RCNN model to identify urban village boundaries from each patch.

BS-MR. This baseline method fuses satellite imagery and bike-sharing drop-off data to identify urban village boundaries using Mask-RCNN model, and it still exploits fixed-size clipping method to generate image patches.

RN-MR. This baseline method uses road networks guided segmentation method to clip satellite imagery without fusing other data, and generate semantic image patches of different sizes. Similarly, the Mask-RCNN model is also exploited to identify urban village boundaries from each patch.

For urban village population estimation, we compare our multi-view neural network method (MvNN) with the following baselines:

RF. This baseline method identifies three building types (low, medium and high elevation) from remote sensing satellite imagery of urban village, and uses them as features to train a random forest model (RF) to estimate population [10].

CNN. This baseline method estimates urban village population based on a convolutional neural network (CNN) from remote sensing satellite imagery [27]. Specifically, it first divides urban village into small patches, and classifies different patches to different population level based on a VGG-A neural network. Furthermore, for each urban village, this method sums output vector in the last layer of CNN of each patch as features, then trains a gradient boosting model to estimate population.

LRM. This baseline method exploits three types of features to train a simple linear regression model (LRM) for population estimation.

HU-ANN. This baseline method only concatenates household capacity features and human mobility features to build a population estimator using artificial neural network (ANN).

HC-ANN. Similarly, this baseline method only concatenates household capacity features and commercial hotness features to train artificial neural network (ANN) for population estimation.

HUC-ANN. This baseline method directly feeds three types of features to an artificial neural network (ANN) for population estimation without multi-view architecture.

5.2 Evaluation Results

We first present the overall results of urban village boundary identification and population estimation in Xiamen Island, and then elaborate two more experiments and results to examine the generalization ability of our models. Finally, we study the parameter selection strategies in our models and present runtime performance.

(a) The view of urban village distribution in Xiamen Island

(b) The improvements of urban village boundary identification using our method

Fig. 9. The results of the urban village boundary identification

5.2.1 Urban Village Boundary Identification Results. In Xiamen Island, we have obtained 217 patches segmented by the road network. Each patch is assigned to three participants for labeling, obtaining 381 raw boundaries labels as urban villages. After cross validation and data cleansing, we finally obtain 61 ground-truth patches.

From Table 3, we can see that our method achieves an F1-score of 0.950 (precision=0.966 and recall=0.934) and an IoU of 0.827, outperforming the other baseline methods. The MiSR method performs the worst IoU, since compared with pixel-level instance segmentation methods, this clipping patch based classification method cannot identify urban village boundaries in a fine-grained manner. The MNC model predicts segment proposals, followed by classification, which is slow and less accurate. Instead, Mask-RCNN model is based on parallel prediction of masks and class labels, which is simpler, more flexible and accurate. Results show that the IoU of the MNC method is lower than that of Mask-RCNN-based methods, validating the effectiveness of Mask-RCNN-based methods. The MR method performs the low identification accuracy, since the fixed-size clipping method ruins the integrity of urban villages. The BS-MR method achieves a relatively high precision, meaning that fusing bike-sharing drop-off data to the satellite imagery can effectively reduce misidentified areas. The RN-MR method achieves a relatively high recall and IoU, which means that using road networks guided region segmentation method detects more correct urban villages and achieves better segmentation boundaries of urban villages. Our RNBS-MR method combines road networks guided region segmentation and bike-sharing drop-off data fusion to further improve the overall identification accuracy. From Figure 9(b), we can clearly observe the improvements of urban village identification using our method. Meanwhile, Figure 9(a) shows the view of urban village distribution in Xiamen Island, where we identified 59 urban villages as denoted in red masks.

5.2.2 Urban Village Population Estimation Results. We present the estimation results of resident, floating and total populations of urban villages respectively in Table 4. It shows that the proposed MvNN method achieves the best performance among all the baselines. The RF method performs the worst estimation accuracy, since it merely consider the static buildings features and ignore dynamic information of human activities. Moreover, the coarse-grained building types do not always fit in the complicated urban village architecture style. The CNN method does not perform well, since the classification of population level for each patch is coarse-grained, which is difficult to accurately estimate urban village population. Both the RF and CNN methods only use single

Table 3. Urban village boundary identification results in Xiamen Island

Methods	Precision	Recall	F1	IoU
MiSR	0.933	0.689	0.793	0.574
MNC	0.828	0.787	0.807	0.641
MR	0.908	0.872	0.890	0.745
BS-MR	0.940	0.883	0.911	0.764
RN-MR	0.922	0.926	0.924	0.805
RNBS-MR	**0.966**	**0.934**	**0.950**	**0.827**

Table 4. Urban village population estimation results in Xiamen Island

	Resident Population			Floating Population			Total Population		
Methods	**RMSE**	**MAPE**	$\mathbf{R^2}$	**RMSE**	**MAPE**	$\mathbf{R^2}$	**RMSE**	**MAPE**	$\mathbf{R^2}$
RF	7410.76	87.18	0.57	2970.24	151.05	0.61	10218.24	81.72	0.59
CNN	6758.77	67.51	0.65	1903.73	91.95	0.79	7658.90	66.66	0.76
LRM	5753.77	73.03	0.79	2419.50	63.57	0.78	5559.62	44.23	0.81
HU-ANN	4563.73	39.02	0.84	2029.73	44.24	0.83	4915.49	36.99	0.87
HC-ANN	3657.69	36.32	0.83	1764.44	44.79	0.87	4732.27	31.02	0.88
HUC-ANN	3020.17	31.74	0.88	1615.70	38.28	0.89	4331.72	36.76	0.90
MvNN	**2517.65**	**25.05**	**0.92**	**1199.14**	**36.63**	**0.94**	**3517.41**	**27.66**	**0.93**

view (satellite imagery) to estimate population, which cannot achieve a good performance. The LRM, HU-ANN, HC-ANN and HUC-ANN methods consider two views of the static and dynamic. The LRM method is too simple to achieve a high performance. The HU-ANN and HC-ANN methods justifies that only concatenating household capacity features with single features (human mobility features or commercial hotness features) is not enough for building an effective estimation model. In contrast, the HUC-ANN method performs better by fusing three types of features. Our proposed MvNN method further improves the performance leveraging multi-view architecture to effectively fuse three types of features, achieving R^2 of 0.92, 0.94 and 0.93 in resident, floating and total population estimations respectively, outperforming the other baseline methods. In general, our method successfully predicts population of urban villages and achieves relatively high accuracy.

We developed an urban village information service system, and applied it to the Xiamen Open Government Data Platform, as shown in Figure 10. In the system, we can clearly see urban village boundaries, population distribution and other useful information for urban villages in Xiamen Island, providing a low-cost decision support for urban village renovation and risk management. We elaborate on the details in discussion.

5.2.3 Generalization Ability to Other Regions. In order to examine the generalization ability of our proposed models, we have conducted two more experiments, one in suburb area of Xiamen (Jimei District) and another in large city (Shanghai).

Specifically, we first apply the trained model in Xiamen Island to its suburb area, Jimei District. Similarly, we obtain the high-resolution remote sensing satellite imagery, the large-scale vehicle (e.g., taxi and shared bike) trajectories datasets and the POIs datasets in Jimei District from the Xiamen Open Government Data Platform. Then, we directly input these data to the trained model to identify urban village boundaries and estimate their population. Results show that our model accurately identify urban village boundaries with an F1-score of 0.916 and an IoU of 0.751, and estimate total population with an R^2 of 0.86, which demonstrate that our trained model

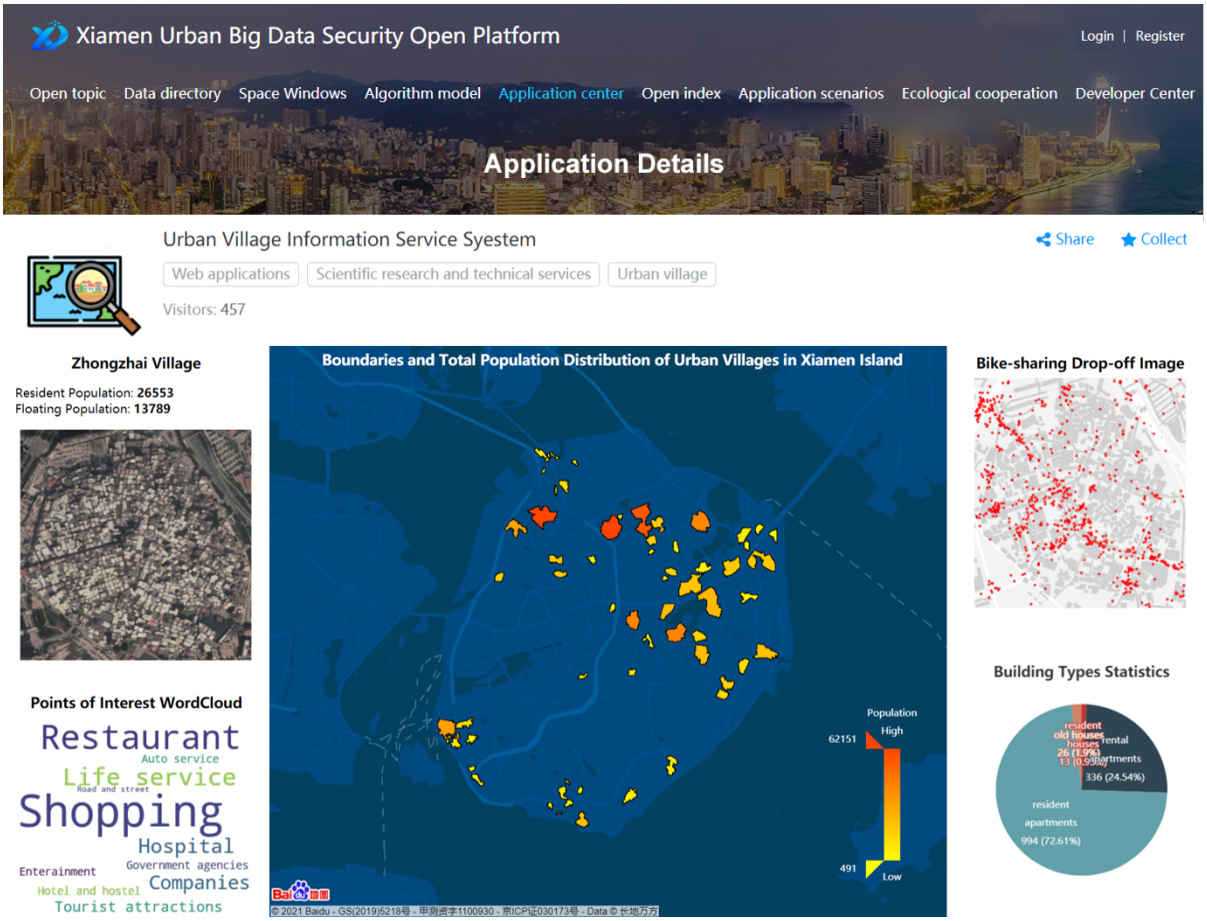


Fig. 10. The developed system of urban village information service

can be generalized to the suburb area of Xiamen. The identified urban village boundaries and estimated population distribution in Jimei District are shown in Figure 11(a).

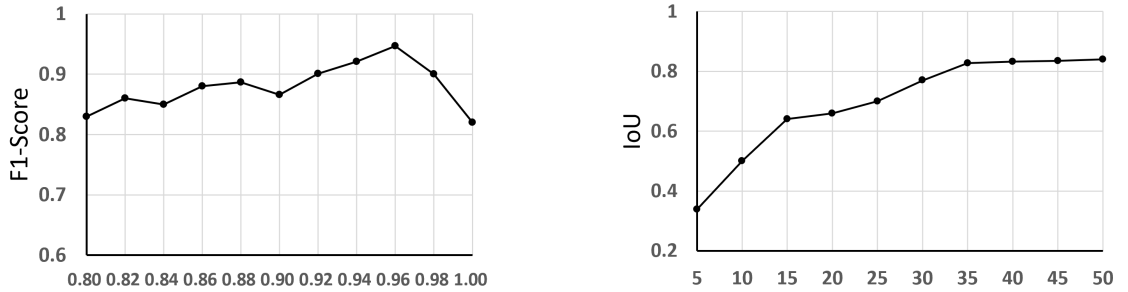
Furthermore, we extend our trained model in Xiamen Island to a large city of China, Shanghai. First, we collect the following datasets in Shanghai: 1) the high-resolution remote sensing satellite imagery from Google Earth[6], 2) the POIs datasets from Baidu Map service[7], 3) the large scale vehicle (e.g., taxi and shared bike) trajectories datasets from the Shanghai Open Data Application competition (SODA)[8] and 4) the population data from the Worldpop website[9]. Since the architecture style and population distribution of urban villages in Shanghai may be different from that of Xiamen, we exploit a transfer learning-based method to fine-tune the trained model before applying it to Shanghai. Specifically, we collect a few labels of urban village boundaries and building types in Shanghai, and then use these samples to fine-tune the boundary identification and population estimation

[6]http://earth.google.com
[7]http://map.baidu.com
[8]http://shanghai.sodachallenges.com
[9]http://worldpop.org/geodata/summary?id=5777

(a) Results in Jimei District

(b) Results in Jiading District of Shanghai

Fig. 11. Urban village boundaries and population distribution in Jimei District and Jiading District of Shanghai. The polygonal area denotes identified urban village boundaries and the color corresponds to the estimated population.



(a) The confidence threshold c in the Mask-RCNN model

(b) The urban village training set size G in the Mask-RCNN model

Fig. 12. Parameter impact analysis and optimal parameter selection.

models based on the pre-trained model. Results show that the fine-tuned models accurately identify urban village boundaries with an F1-score of 0.906 and an IoU of 0.771, and estimate total population with an R^2 of 0.78, which validate that our trained model can be generalized to other cities. Figure 11(b) shows the identified urban village boundaries and estimated population distribution in Jiading District, one of the suburb areas of Shanghai.

5.2.4 Parameter Study. We discuss the parameter selection in the RNBS-MR and MvNN methods as follows:

The instance confidence c. In the Mask-RCNN model of identifying urban village boundaries, each instance detected from each patch has a confidence. We need to carefully select the confidence threshold to filter out incorrect urban village instances and retain correct urban village instances, and to balance the precision and recall of the model. We vary c from 0.8 to 1, and present the F1-score under different c values in Figure 12(a). We can see that the confidence threshold of 0.96 can achieve an optimal result, and thus we select $c = 0.96$ in our experiments.

The urban village training set size G. In the urban village boundary identification phase, we select a series of labeled urban village patches to train Mask-RCNN model. It is well known that a large training set can help achieve high identification accuracy. However, outsourcing the labeling tasks to workers still requires great effort and time consumption. Therefore, we need to select the appropriate size of urban village training set to trade-off the model performance and labeling workload. We study the IoU values against the different urban village training set size G from 5 to 50 in Figure 12(b). We can see that the training set of size 35 can achieve an IoU higher than 80%, and the evaluation results of the model fails to improve greatly after continuing to add the G. Therefore, we select the urban village training set size $G = 35$ in our experiments.

Furthermore, for the threshold T of road network extraction, we set $T = 6$ to obtain clear road networks based on repeated experiments. With regard to the building training set size N in Mask-RCNN model, we find that selecting $N = 40$ in all 985 patches is sufficient for training a good Mask-RCNN model to identify different categories of buildings based on the repeated experiments. These patches contain a 1:3:5:1 ratio of buildings corresponding to *Resident houses*, *Rental apartments*, *Resident apartments* and *Old houses*. For the empirical formula of the household capacity features, we determine the average number of households per category in Xiamen Island based on the field studies, as follows: *the resident house* accommodates 3.45 households, *the rental apartment* accommodates 17.26 households, *the resident apartment* contains 9.63 households and *the old house* contains 2.31 households, and we obtain $r = 2.06$ from the urban housing authority in Xiamen.

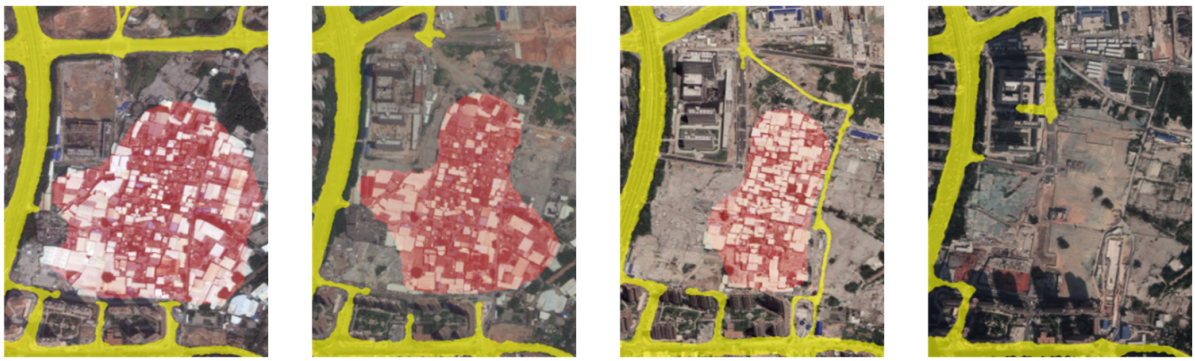
5.2.5 Runtime Performance. We implement the RNBS-MR and MvNN methods using tensorflow. We deploy our framework on a server with an nVIDIA GeForce GTX 2080 graphic card and 11GB RAM, and it takes an average of 34 minutes to identify urban village boundaries in Xiamen Island, and takes an average of 17 minutes to estimate population per urban village.

5.3 Case Studies

We conduct case studies on urban village boundary identification and population estimation in Xiamen Island. First, applying our urban village boundary identification method in different time, we observe the changes of urban village boundaries. Second, through comparing the population results of two urban villages with similar housing distribution but with large population differences, we demonstrate the effectiveness of our population estimation method. Third, in order to more comprehensively complement our models, we analyze failure cases in urban village boundary identification and population estimation.

5.3.1 The Changes of Dongzhai Village from 2015 to 2018. Since urban villages bring potential risks to society and hinder the development of cities, urban authorities often gradually demolish urban villages and transfer the residents to improve the living standards of citizens. Through urban village boundary identification, we can clearly identify these changes from remote sensing satellite imagery taken in different time. As shown in Figure 13, the Dongzhai Village was an old village whose residents lived on rent. However, since this area was far away from the city center, the rent is low and the living standard of the residents is low. Therefore, the urban authority decided to demolish this village and relocated the residents. This village occupied a large area in 2015, and then it gradually shrunk over time. The demolition of Dongzhai Village finished in 2018, and several new hospitals had been built in the upper left corner of the satellite imagery. Meanwhile, we can observe that the shrinking of the urban village boundaries also resulted in the dynamic changes of the surrounding road network. For example, there are a narrow road to surround Dongzhai Village in 2017 in Figure 13(c), and it is gone after the village vanished in 2018 in Figure 13(d).

5.3.2 The Comparison of Houkeng Village and Maohou Village. Figure 14 presents two urban villages with similar areas. The Houkeng Village is $0.15km^2$ and the Maohou Village is $0.14km^2$. Besides, they have similar building distributions. Therefore, if we estimated their populations merely based on the empirical formula, we would get



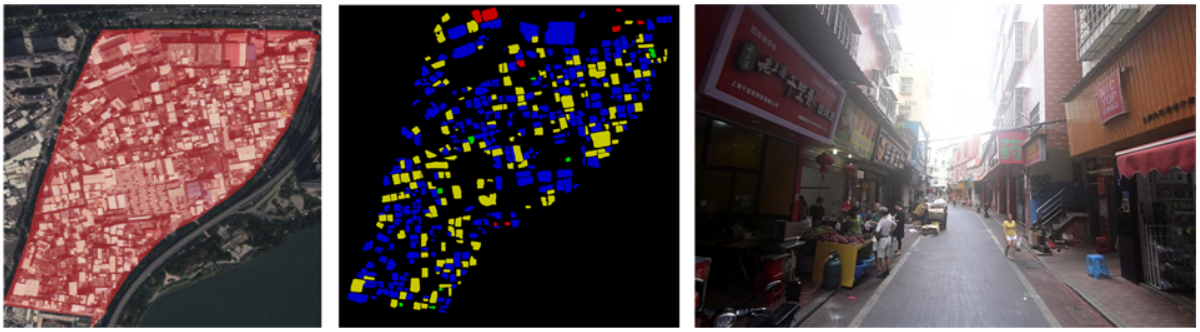
(a) The Dongzhai Village in 2015. (b) The Dongzhai Village in 2016. (c) The Dongzhai Village in 2017. (d) The Dongzhai Village in 2018.

Fig. 13. The changes of the boundary and surrounding road network of the Dongzhai Village from 2015 to 2018. The red part represents the urban village areas and the yellow part denotes the surrounding road network.

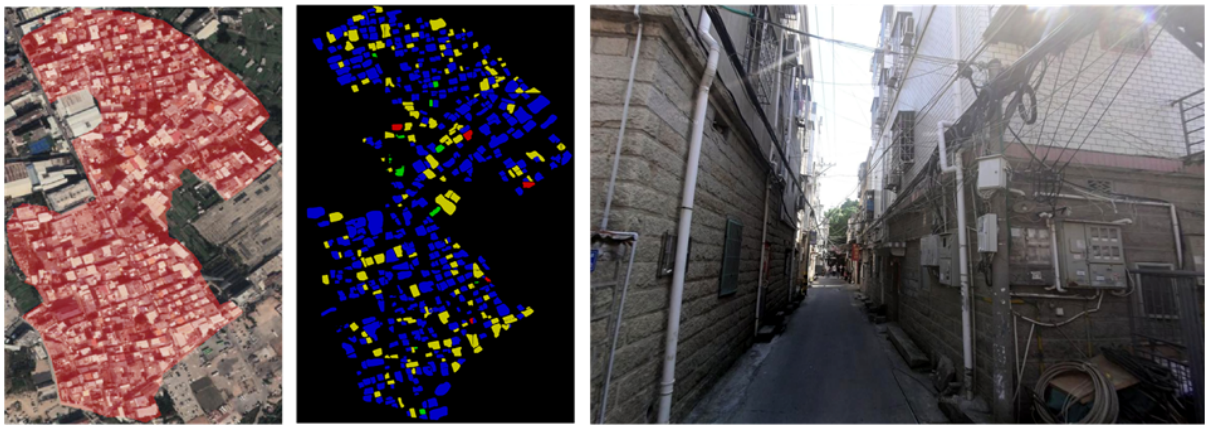
similar results which are 8,634 in Houkeng Village and 9,575 in Maohou Village. However, in fact, there is a big difference in their actual populations which are 22,333 in Houkeng Village and 6,795 in Maohou Village. The number of the resident population in Houkeng Village is around 3 times as much as that of Maohou Village, and the number of the floating population in Houkeng Village is also more 5251 than that of Maohou Village. This difference can be reflected in their different human mobility features and commercial hotness features. The daily shared bike usage in Houkeng Village is nearly 3 times as much as that in Maohou Village. Besides, Houkeng Village has 33 shops, 3 hospitals and 20 restaurants, while these POIs in Maohou Village (i.e., 16 shops, 1 hospital and 12 restaurants) is less than that in Houkeng Village. With the proposed method, the estimated resident and floating populations in Houkeng Village are 14,148 and 6,327 respectively, and the estimated resident and floating populations in Maohou Village are 7,087 and 2,176 respectively. The estimated results show that the proposed method is effective.

5.3.3 Undetected Urban Village and Failure Population Estimation. In the urban village boundary identification phase, we note that it is usually difficult for our model to identify small urban villages located close to mountain. For example, as shown in Figure 15(a), the Xibianshe Village is not detected by our model. One of the reasons is that the patch containing Xibianshe Village is largely occupied by the mountain, and the urban village boundary is relatively small. Moreover, this urban village itself is covered by flourish trees compared to other regions. These two reasons may lead to insufficiently significant features for urban village boundary identification. In the future, we plan to improve our model with automatic patch scaling method [8] to eliminate failure case.

In the population estimation phase, we note that if an urban village is recognized as a tourist attraction (e.g., fishing village), it is difficult for our model to accurately estimate its population. For example, as shown in Figure 15(b), the estimated population for Zengcuoan Village (i.e., 10,514) is around 3 times as much as the actual population (i.e., 3,772). Through analyzing the extracted features, we observe that there are a lot of bike-sharing drop-off records in Zengcuoan Village. Besides, The village presents high commercial hotness since it has many shops and restaurants in the POIs. These human activities features significantly increase the estimated population. However, as a tourist attraction, such a large increase in shared bike usage and commercial hotness are actually brought by the tourists instead of local residents. Therefore, the estimated population by our model is much higher than the real-world residential population. In this case, we suggest that the authority should exclude the number



(a) The Houkeng Village in 2017 (area: 0.15 km^2, resident population: 15,379, floating population: 6,954).



(b) The Maohou Village in 2017 (area: 0.14 km^2, resident population: 5,092, floating population: 1,703).

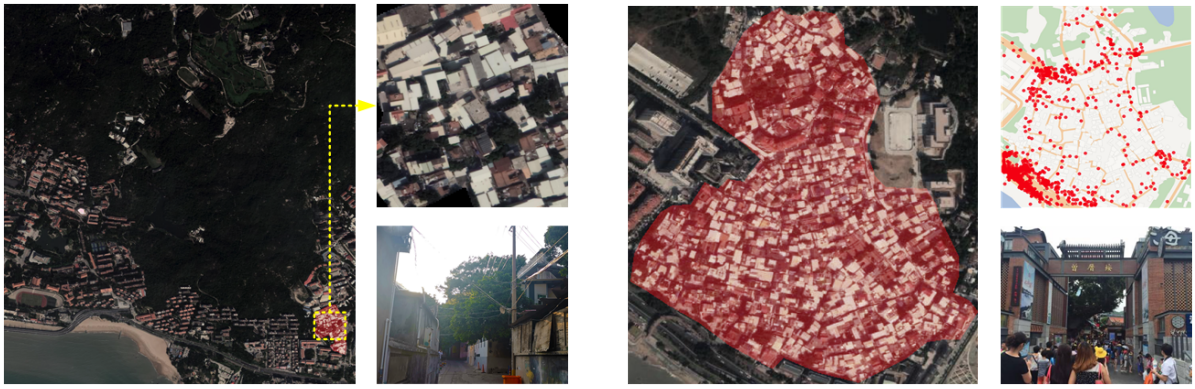
Fig. 14. Two urban villages with similar building distributions and areas but different population distributions.

of tourists from the population estimation when using our model. In the future, we plan to incorporate tourist data from local tourism bureau into our model to improve the accuracy of residential population estimation.

6 RELATED WORK

6.1 Urban Areas Segmentation Using Road Networks

The city road networks naturally segment urban areas into regions with varying sizes and shapes. Most previous works exploit two major models to represent road networks, including vector-based model and raster-based model. Vector-based model uses geometric primitives such as points, lines and polygons to denote spatial objects on the Cartesian coordinates. Zhao et al. [40] proposed a graph-theory approach to segment urban into regions based on the vector model. It turns vector data into a graph and uses dijkstra algorithm to partition the urban areas. Different from vector-based model, raster-based model quantizes an area into small discrete grid-cells indexing all the spatial objects. Generally, a raster-based map is regarded as a binary image (e.g., 0 stands for road segments and 1 stands for blank space). Yuan et al. [36] exploited morphological image processing techniques to refine rasterized road networks, and merge connected component to find individual urban regions. This method has also been used in [34, 35]. In our work, we recover the raster-based road network image from taxi



(a) The Xibianshe Village (b) The Zengcuoan Village

Fig. 15. Failure case studies on urban village boundary identification and population estimation. Figure (a) shows the satellite imagery and the street view in Xibianshe Village. Figure (b) shows the satellite imagery, the bike-sharing drop-off mapping image and the street view in Zengcuoan Village.

GPS trajectories instead of directly using the road topology. Since the road topology extracted from geological survey or recognized from remote sensing imagery is hierarchical, the patches segmented by different level road networks may be too large or too small. In contrast, taxi GPS trajectories not only successfully recover major road networks surrounding urban villages, but also avoid dividing urban village into pieces as taxis are usually not allow to enter urban villages due to their narrow roads. Therefore, using taxi GPS trajectories to extract road networks is helpful to obtain properly-sized patches for urban village boundary identification.

6.2 Object Detection and Segmentation from Satellite Imagery

With the increasing availability of high-resolution remote sensing satellite imagery, researchers have conducted a series of studies for satellite imagery. A variety of natural or man-made objects can be detected from satellite imagery, such as crop types [24], archaeological sites [1], oil spill [17], illegal deforestation [29], landslide [21], and burned area [30]. Compared with object detection, object segmentation task is a more complex challenge. It not only finds target locations, but also segments boundary of target by pixel level. Many object segmentation problems have been studied. For example, Hayatbini et al. [12] used a gradient based algorithm to segment cloud, and Berry et al. [2] proposed a stacked hourglass networks to segment parking lot. Object detection and segmentation about urban area is also a prevalent research issue. For example, Li et al. [18] extracted and analyzed the urban landscape from satellite imagery, and Sahriman et al.[28] conducted a correlation study on satellite imagery to identify urban poverty area. Urban villages, as poverty area in Chinese cities, have also been identified using satellite imagery, such as Huang et al.[15] proposed a scene representation approach with extracted buildings and vegetation to detect urban village. Due to the effective application of deep convolutional network in instance segmentation, the Mask-RCNN model is used to detect urban village and segment the boundaries of urban village from satellite imagery by Chen et al. [5]. However, few works fuse other urban data and satellite imagery to identify urban village.

6.3 Population Estimation from Heterogeneous Urban Data

The population estimation problems have been studied by many researchers using heterogeneous urban data, including static data (e.g., remote sensing satellite imagery [10, 27, 33], Lidar point cloud data [20]), dynamic data (e.g., night-time light imagery [6, 31], cellphone network accessing records [9, 34]). Using the static data to estimate population often lies on the relationship between living space and population. For example, Wang et al. [33] proposed two approaches to estimate population respectively based on residential area and dwelling count extracted from remote sensing satellite imagery, and Lu et al. [20] constructed a volume-based model leveraging Lidar point cloud data to infer population. However, these methods only estimate population from the static living space, ignoring the dynamic human activity characteristics. Especially, in the residential areas with complex household situations and large population flowing (e.g., urban village), the estimation may be inaccuracy. Therefore, some researchers considered using dynamic urban crowdsensing data for population estimation. For example, night-time light data has been used to estimate population by Sun et al. [31] and Chu et al. [6], due to its capability of indicating human activities. Under the high penetration of cellphones, many researchers conducted social intelligence researches leveraging its left digital footprints [38]. Cellphone network accessing records are often used to estimate real-time population. For example, Xu et al. [34] used a power-law distribution to model the relationship between the number of mobile users and population. Fang et al. [9] further design a MultiCell model to estimate urban populations from multiple cellphone networks. However, such methods can only estimate population for a relatively large area, which is not accurate enough for estimating population within the small urban village boundaries. Therefore, considering the complex building structures and household situations in urban village, we propose a new approach by combining the static and dynamic data to estimate population.

7 DISCUSSION

In this section, we discuss the potential applications of the proposed urban village information service system. The proposed system displays sufficient information about urban villages, including urban village boundaries, population distribution, the satellite imagery, the information of buildings and POIs for each urban village, etc., so as to provide decision support for urban management. Specifically, in terms of urban planning, the urban authority could reasonably plan the land use in urban village regions according to their different size and the information of surrounding POIs. In terms of fire safety, the urban authority could pay more attention to those urban villages with high building and population density, properly equip them with firefighting device and arrange fire inspection system, so as to reduce the safety risks such as explosion in urban villages as far as possible. Besides, many rental apartments in urban villages are found to illegally add floors and rooms. The urban authority could strengthen the supervision of these rental apartments to ensure the safety of renters. In terms of public health, urban villages with high floating population density may be at higher risk of epidemic spreading of influential diseases, the urban authority could provide actual medical services to prevent such epidemic.

8 CONCLUSION

In this work, we proposed a data fusion framework to accurately identify the boundaries and estimate the population of urban villages with heterogeneous open government data sources, which can greatly reduce the human labor and time consumption to continuously monitor the boundaries and population changes in urban villages. In boundary identification, we leverage road networks extracted from large-scale taxi trajectories to generate patches for each urban village, and augment satellite imageries with bike-sharing data to refine the boundary segmentation results. In population estimation, we extract household capacity features, human mobility features, and commercial hotness features from satellite imageries, bike-sharing drop-off records, and POI distributions, respectively. Real-world experiments on Xiamen City show that our framework accurately

identifies the boundaries and estimates the population of urban villages, outperforming the state-of-the-art baseline methods.

In the future, we plan to incorporate more contextual features (e.g., digital elevation map) to further improve the boundary identification accuracy, and explore more complicated models in population estimation (e.g., with graph neural networks).

ACKNOWLEDGMENTS

We would like to thank the reviewers and editors for their constructive suggestions. This research is supported by NSF of China No. 61802325 and No. 61872306.

REFERENCES

[1] Athos Agapiou, Dimitrios D Alexakis, Apostolos Sarris, and Diofantos G Hadjimitsis. 2013. Orthogonal equations of multi-spectral satellite imagery for the identification of un-excavated archaeological sites. *Remote Sensing* 5, 12 (2013), 6560–6586.

[2] Tessa Berry, Nicholas Dronen, Brett Jackson, and Ian Endres. 2019. Parking Lot Instance Segmentation from Satellite Imagery through Associative Embeddings. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 528–531.

[3] Tim Brindley. 2003. The social dimension of the urban village: A comparison of models for sustainable urban development. *Urban Design International* 8, 1-2 (2003), 53–65.

[4] D Chaudhuri, NK Kushwaha, Ashok Samal, and RC Agarwal. 2015. Automatic building detection from high-resolution satellite images based on morphology and internal gray variance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9, 5 (2015), 1767–1779.

[5] Longbiao Chen, Tianqi Xie, Xueyi Wang, and Cheng Wang. 2019. Identifying urban villages from city-wide satellite imagery leveraging mask R-CNN. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 29–32.

[6] Hone-Jay Chu, Chen-Han Yang, and Chelsea C Chou. 2019. Adaptive non-negative geographically weighted regression for population density estimation based on nighttime light. *ISPRS International Journal of Geo-Information* 8, 1 (2019), 26.

[7] Jifeng Dai, Kaiming He, and Jian Sun. 2016. Instance-Aware Semantic Segmentation via Multi-Task Network Cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Lucian Drăguţ, Ovidiu Csillik, Clemens Eisank, and Dirk Tiede. 2014. Automated parameterisation for multi-scale image segmentation on multiple layers. *ISPRS Journal of photogrammetry and Remote Sensing* 88 (2014), 119–127.

[9] Zhihan Fang, Fan Zhang, Ling Yin, and Desheng Zhang. 2018. MultiCell: Urban population modeling based on multiple cellphone networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–25.

[10] Stefanos Georganos, Tais Grippa, Assane Gadiaga, Sabine Vanhuysse, Stamatis Kalogirou, Moritz Lennert, and Catherine Linard. 2019. An application of geographical random forests for population estimation in Dakar, Senegal using very-high-resolution satellite imagery. In *2019 Joint Urban Remote Sensing Event (JURSE)*. IEEE, 1–4.

[11] Pu Hao, Pieter Hooimeijer, Richard Sliuzas, and Stan Geertman. 2013. What drives the spatial development of urban villages in China? *Urban Studies* 50, 16 (2013), 3394–3411.

[12] Negin Hayatbini, Kuo-lin Hsu, Soroosh Sorooshian, Yunji Zhang, and Fuqing Zhang. 2019. Effective cloud detection and segmentation using a gradient-based algorithm for satellite imagery: Application to improve PERSIANN-CCS. *Journal of Hydrometeorology* 20, 5 (2019), 901–913.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[14] Zhiyuan He and Su Yang. 2018. Multi-view Commercial Hotness Prediction Using Context-aware Neural Network Ensemble. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–19.

[15] Xin Huang, Hui Liu, and Liangpei Zhang. 2015. Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing* 53, 7 (2015), 3639–3657.

[16] Xin Huang and Liangpei Zhang. 2011. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery. *Photogrammetric Engineering & Remote Sensing* 77, 7 (2011), 721–732.

[17] Marios Krestenitis, Georgios Orfanidis, Konstantinos Ioannidis, Konstantinos Avgerinakis, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2019. Oil spill identification from satellite images using deep neural networks. *Remote Sensing* 11, 15 (2019), 1762.

[18] Zhi Li, Chenghu Zhou, Xiaomei Yang, Xi Chen, Fan Meng, Chen Lu, Tao Pan, and Wenjuan Qi. 2018. Urban landscape extraction and analysis in the mega-city of China's coastal regions using high-resolution satellite imagery: A case of Shanghai, China. *International*

journal of applied earth observation and geoinformation 72 (2018), 140–150.
[19] Yuting Liu, Shenjing He, Fulong Wu, and Chris Webster. 2010. Urban villages under China's rapid urbanization: Unregulated assets and transitional neighbourhoods. *Habitat International* 34, 2 (2010), 135–144.
[20] Zhenyu Lu, Jungho Im, and Lindi Quackenbush. 2011. A volumetric approach to population estimation using LiDAR remote sensing. *Photogrammetric Engineering & Remote Sensing* 77, 11 (2011), 1145–1156.
[21] Pilli Madalasa, Gorthi RK Sai Subrahmanyam, Tapas Ranjan Martha, Rama Rao Nidamanuri, and Deepak Mishra. 2018. Bayesian Approach for Landslide Identification from High-Resolution Satellite Images. In *Proceedings of 2nd International Conference on Computer Vision & Image Processing*. Springer, 13–24.
[22] Alberto Magnaghi. 2005. *The urban village: a charter for democracy and sustainable development in the city.* Zed books.
[23] Zhuokun Pan, Jiashu Xu, Yubin Guo, Yueming Hu, and Guangxing Wang. 2020. Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net. *Remote Sensing* 12, 10 (2020), 1574.
[24] Nan Qiao, Yi Zhao, Ruei-Sung Lin, Bo Gong, Zhongxiang Wu, Mei Han, and Jiashu Liu. 2019. Generative-Discriminative Crop Type Identification using Satellite Images. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 1–5.
[25] R. E. Woods R. C. Gonzalez. 1992. Digital image processing. , 85–103 pages.
[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.
[27] Caleb Robinson, Fred Hohman, and Bistra Dilkina. 2017. A deep learning approach for population estimation from satellite imagery. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. 47–54.
[28] Noorita Sahriman, Mohamad Zamani Zainal Abiden, Abdul Rauf Abdul Rasam, Nazirah Md Tarmizi, et al. 2013. Urban poverty area identification using high resolution satellite imagery: A preliminary correlation study. In *2013 IEEE International Conference on Control System, Computing and Engineering*. IEEE, 430–434.
[29] Zuraidah Said, Rizky Firmansyah, and Benita Nathania. 2019. The Use of Near-Real-Time Data and High-Resolution Satellite Images for Area Identification of Illegal Forest Clearing. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 6391–6393.
[30] Sara Silva, Maria J Vasconcelos, and Joana B Melo. 2010. Bloat free genetic programming versus classification trees for identification of burned areas in satellite imagery. In *European Conference on the Applications of Evolutionary Computation*. Springer, 272–281.
[31] Weichao Sun, Xia Zhang, Nan Wang, and Yi Cen. 2017. Estimating population density using DMSP-OLS night-time imagery and land cover data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10, 6 (2017), 2674–2684.
[32] Satoshi Suzuki et al. 1985. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing* 30, 1 (1985), 32–46.
[33] Shiqian Wang, Wei Li, and Jonathan Li. 2014. Using high resolution remote sensing image to help population estimation in small cities. In *2014 IEEE Geoscience and Remote Sensing Symposium*. IEEE, 3172–3175.
[34] Fengli Xu, Pengyu Zhang, and Yong Li. 2016. Context-aware real-time population estimation for metropolis. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1064–1075.
[35] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 186–194.
[36] Nicholas Jing Yuan, Yu Zheng, and Xing Xie. 2012. Segmentation of urban areas using road networks. *Microsoft Corp., Redmond, WA, USA, Tech. Rep. MSR-TR-2012-65* (2012).
[37] Jerrold H Zar. 1972. Significance testing of the Spearman rank correlation coefficient. *J. Amer. Statist. Assoc.* 67, 339 (1972), 578–580.
[38] Daqing Zhang, Bin Guo, and Zhiwen Yu. 2011. The emergence of social and community intelligence. *Computer* 44, 7 (2011), 21–28.
[39] Kang Zhao, Jungwon Kang, Jaewook Jung, and Gunho Sohn. 2018. Building Extraction From Satellite Images Using Mask R-CNN With Building Boundary Regularization.. In *CVPR Workshops*. 247–251.
[40] Si Zhao, Hongwei Wu, Lai Tu, and Benxiong Huang. 2014. Segmentation of Urban Areas Using Vector-Based Model. In *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*. IEEE, 412–416.
[41] Yu Zheng. 2015. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6, 3 (2015), 1–41.