Urban Land Use Classification Based on Aerial and Ground Images

Rui Cao
International Doctoral Innovation Centre
University of Nottingham Ningbo China
Ningbo, China
rui.cao@nottingham.edu.cn

Guoping Qiu
College of Information Engineering, Shenzhen University
Shenzhen, China
School of Computer Science, University of Nottingham
Nottingham, UK
guoping.qiu@nottingham.ac.uk

***Abstract*—Urban land use is key to rational urban planning and management. Traditional land use classification methods rely heavily on domain experts, which is both expensive and inefficient. In this paper, we explore to utilise deep neural network based approaches to label urban land use at pixel level using high-resolution aerial images and ground-level street images. We use a deep neural network to extract semantic features from sparsely distributed street images and interpolate them in the spatial domain to match the spatial resolution of the aerial images, which are then fused together through a deep neural network for classifying land use categories. We test our methods on a large publicly available aerial and street images dataset of New York City, and the results show that using aerial images alone can achieve relatively high classification accuracy and the ground-level street views contain useful information for urban land use classification. Fusing street image features with aerial images can improve classification accuracy to some extent but the improvement is somewhat limited.**

***Index Terms*—Land use classification, semantic segmentation, aerial images, street views, data fusion**

I. Introduction

Urban areas account for less than 2% of the earth land surface, but accommodate more than half of the world population [1], [2]. Unprecedented urbanisation leads to rapid changes of urban surface, it is therefore of great significance to monitor our urban land so as to provide essential information to decision makers to better manage our cities.

Urban land use and land cover (LULC) maps are very important tools to understand and monitor our cities, they reflect the macro properties of the urban surface. Specifically, land cover indicates the physical attributes of landscapes, such as forestry and water body, while land use documents how people use the land with social-economic purposes, such as residential, commercial, and recreational purposes.

In remote sensing area, earth observation data such as multi-spectral satellite images have long been utilised to classify different land covers in terms of spectral reflectance characteristics of different objects [3]. However, they are insufficient for categorising urban land use types of different social-economic properties, and land use classification still relies heavily on labour-intensive land survey [1], which are inefficient and expensive.

With the development of geospatial technologies, we are able to acquire very high resolution (VHR) satellite and aerial images, which enables us to acquire more details from overhead images than before. Nevertheless, despite great increase in spatial resolution, urban land use classification from overhead images is regarded as a extremely difficult task [4], because only the top of the cities can be captured from nadir view, the lack of ground-level details makes it hard to predict social-economic usage purposes of urban land.

Fortunately, the situation is changing. On the one hand, great progresses have been made on semantic segmentation task with the development of deep neural networks (DNNs) [5], which facilitates the pixel-level urban land use classification problem [6]. On the other hand, the growing accessibility to different sources of geo-tagged data makes it possible to fuse data of different modalities and observations [7].

To help with the difficult problem of urban land use classification, we experiment with deep-learning based approaches on high-resolution aerial images and ground-level street views respectively, and also develop a method for the integration of the two sources of data.

The paper is organised as follows. In Section 2, we review related work on land cover and land use classification and current research progress on semantic segmentation using deep neural networks. In Section 3, we formulate urban land use classification as a probabilistic optimisation problem. Section 4 describes the method we use to construct ground feature maps from street views, and integrate overhead and ground-level images for urban land use classification problem. In Section 5, we test our methods on a publicly available dataset of large aerial and street images, and analyse the results. Finally, we conclude in Section 6.

II. Related work

A. Land use and land cover classification

Remote sensing: Land use and land cover classification via satellite images have long been a difficult research problem in remote sensing area. Most related work has engaged with land cover classification [1], [2], [4], and normally, the inference of specific land cover types more relies on spectral reflectance

978-1-5386-7021-7/18/$31.00

characteristics of spatial objects via multispectral satellite images [3], because the spatial resolution of satellite images in visible bands is limited. With the development of geospatial technologies, very high resolution (VHR) satellite and aerial images become more available, which enables us to analyse more spatial patterns via these images [3], [8], [9].

Proximate sensing: Traditional land use map is provided by land survey [1] which is labour-intensive, time-costing, and expensive. To alleviate the situation, researchers have tried to infer land use from proximate sensing data. Pei et al. [10] use aggregated mobile phone data to conduct land use classification in mesh grid level. Tu et al. [11] couple mobile phone data and social media check-in data to infer urban land function zones. Zhu et al. [4] use ground-level geo-referenced images from Flickr to do land use mapping based on land parcel map. Kang et al. [12] utilise street images to classify building functions given building footprints.

Multimodal data fusion: Remote and proximate sensing data include macro overhead and micro ground-level information respectively. The integration of them is believed to be able to capture both information and therefore provide more insights into the understanding of urban land use distribution than just use one data source alone. Tu et al. [1] and Jia et al. [2] integrate satellite images and mobile phone positioning data to generate urban land use maps. Liu et al. [13] and Hu et al. [14] combine satellite images and POIs (Points of Interest) to classify urban land parcels, showing that social media data have the potential for augmenting LULC classification. Some researches also try to fuse data of different views in terms of physical appearance of urban surface [7]. For example, Workman et al. [15] incorporates ground-level street views into aerial images to classify land use types and predict building ages and functions. Zhang et al. [16] integrates airborne LiDAR, high resolution aerial imagery, and street views data to classify urban land parcels.

B. DNN-based semantic segmentation

With the unprecedented success of deep neural networks, many computer vision applications have seen great breakthroughs, including image classification, object detection, and semantic segmentation.

Fully Convolutional Network (FCN) [17] is regarded as a milestone for DNN-based semantic segmentation. Ever since it is proposed to solve the pixel-level classification problem, more and more semantic segmentation researches have focused on deep neural networks methods. The network changes the architecture of normal deep convolutional neural networks for classification by replacing fully connection layers with convolutional layers which enable it to make dense pixel-level predictions, this paradigm is adopted by many DNN-based semantic segmentation methods followed [5]. However, FCN has its drawbacks. The most significant problem is its pooling layer, which can aggregate information and extract spatial-invariant features. But spatial information is crucial for semantic segmentation problems since pixel-level predictions are to be made. To address the problem, two main architectures are proposed, the first one is encoder-decoder architecture, such as SegNet [18], and the other is dilated convolutions [5].

Most breakthrough in DNN-based semantic segmentation happened on natural images [5]. However, remote sensing images are very different from ordinary natural images. Some efforts and progresses have been made on satellite and aerial image segmentation using deep learning approaches [6]. For example, Kampffmeyer et al. [9] have proposed a convolutional neural network for land cover mapping using very high resolution aerial images with high accuracy. Audebert et al. [8] have proposed DNN-based models to fuse multi-modal and multi-scale remote sensing data. Because of the simplicity and effectiveness of SegNet [18] on both natural and aerial images, our work has adopted it as the basic network.

III. Problem statement

The problem of urban land use classification using different sources of images can be formulated as follows:

$$L^*(x_i) = \arg \max_{L(x_i) \in \forall L} (P(L(x_i)|I_1(x_i), I_2(x_i), ..., I_n(x_i)))$$

where, x_i is the location of a pixel, $I_k(x_i)(k = 1, 2, ..., n)$ is the corresponding value of the input image k at the location, $L(x_i)$ is the semantic label of the pixel, i.e. urban land use category of the location. The problem is to maximise the probability of predicting the accurate land use type, given $n(n \geq 1)$ input images of the specified location.

IV. Methodology

To utilise ground-level GSVs for urban land use classification, we propose an approach to constructing ground feature maps. We first extract semantic features from GSVs and then construct ground feature maps by interpolating those features in spatial domain. Furthermore, to fuse aerial images and ground-level street images, we take both aerial images and ground feature maps as inputs to our proposed deep convolutional neural network, which is able to fuse the two sources of data from different views.

A. Ground feature map construction

In order to align ground-level GSVs with overhead aerial images in pixel level, we construct ground feature maps. The construction process is illustrated in Figure 1. Basically, there are two major steps, GSV feature extraction and spatial interpolation.

1) GSV feature extraction: To get the semantic features from GSVs, we firstly extract GSV features using PlacesCNN which is trained on the Places-365 dataset [19] and is used for ground-level scene recognition. For each location with GSVs, there are four images facing different directions. We first use pretrained PlacesCNN (without the last fully connection layer) to extract a 512-dimensional feature vector for each image, and then concatenate the extracted four feature vectors into a 2048-dimensional feature vector for each location. After that, PCA is used to reduce the dimension to 50.

Fig. 1. Construction of ground feature map. (1) GSV feature extraction. (2) Spatial interpolation.

2) Spatial interpolation: Sparse street views cover nearby visual areas of the urban surface, instead of just single dots in the space. Thus, it is important to project the semantic information of GSVs to their covered pixels from top-down viewpoint. To form a dense ground-level feature map from sparsely distributed GSV features, we use Nadaraya-Watson kernel regression (see Equation (1)) to interpolate GSV features in the spatial domain.

$$f(x) = \sum_{i=1}^{k} w_i f(x_i) / \sum_{i=1}^{k} w_i \tag{1}$$

where, in our case, $f(x)$ is the value of a pixel x, $f(x_i)$ is the value of nearby GSV point x_i, the impact of x_i on x is measured by the weight w_i, k is the number of nearby points.

To estimate the impact of nearby GSVs on a pixel, we use Gaussian kernel to calculate weights. Considering the limited visual coverage of GSVs, we set a distance threshold to cut off far away GSVs. The kernel to calculate the weights is show in Equation (2).

$$w_i = \begin{cases} \exp(-d(x,x_i)^2/h^2), & d(x,x_i) \le h \\ 0, & d(x,x_i) > h \end{cases} \tag{2}$$

where, w_i is the weight that the GSV point x_i impacts on the pixel x, $d(x,x_i)$ is the distance between them, and h is the bandwidth of the Gaussian kernel, and is also used as cutoff distance threshold.

B. DNN-based data fusion

The overall architecture of our proposed network is shown in Figure 2. The network is based on SegNet [18] which is composed of two major components, encoder and decoder. The encoder uses the architecture of the first 13 convolutional layers of VGG-16 [20], and the decoder is symmetric to the encoder counterpart, only to use max unpooling to replace max pooling layers. The last layer is a Softmax layer to make the final pixel-level predictions.

However, our proposed network has an extra encoder. Input aerial images and ground feature maps are fed into the two encoders separately, and then the outputs from the encoders are stacked together as input fed into the decoder to upscale and make the final predictions.

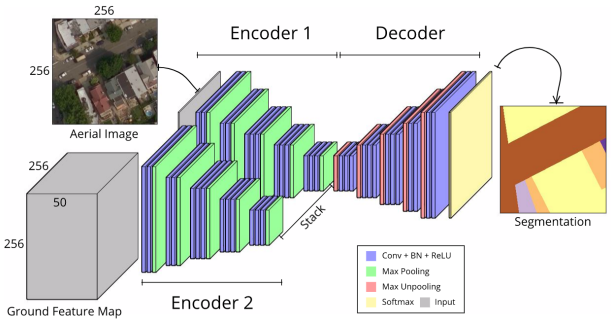

Fig. 2. Overview of our proposed network architecture. The network is composed of two encoders and one decoder. The composition of *Encoder1* and *Decoder* is SegNet [18]. To fuse two sources of data, we add an extra encoder. The aerial images and ground feature maps are fed into the two encoders separately, and then the outputs of the two encoders are stacked as input to the decoder to get the final segmentation results.

V. Experiments

A. Datasets

New York City (shown in Figure 3) is one of the most densely populated cities in the US. It has a land area of 783.84 km^2 with more than 8 million population. The land use of New York City is highly diversified which poses great challenges for land use classification. In the experiments, we use a public available dataset of New York City from [15].

The datasets consist of three types of data: high-resolution aerial images, corresponding land use maps, and sparsely sampled street views. An example of the data pairs are shown in Figure 4. The aerial images are from Bing Map with ground resolution of about 0.3 metres. The ground-level images come from Google Street Views, with four images from different heading directions at each place. The land use maps are from New York City Department of City Planning, and are categorised into 11 categories (see Table I) documenting primary land use in tax lot level. To accommodate the missing data and unlabelled areas, two extra categories are added, i.e. *unknown* and *background*.

The dataset contains two subsets. The Brooklyn dataset covers the whole area of Brooklyn borough with 73,921 aerial image tiles in total (a large portion of them are over water which are discarded), 39,244 of them are used as training data, and 4,361 randomly selected tiles are used as validation data. There are 557,308 GSVs within the area in total. The Queens dataset covers a squared area in Queens borough which are used as test set, there are 10,044 aerial image tiles and 154,412 GSVs in total.

B. Comparative studies

To explore the effectiveness of aerial images and street views, we have conducted three groups of experiments, i.e. segmentation using aerial images only, with street views only, and integrating aerial and street images respectively.

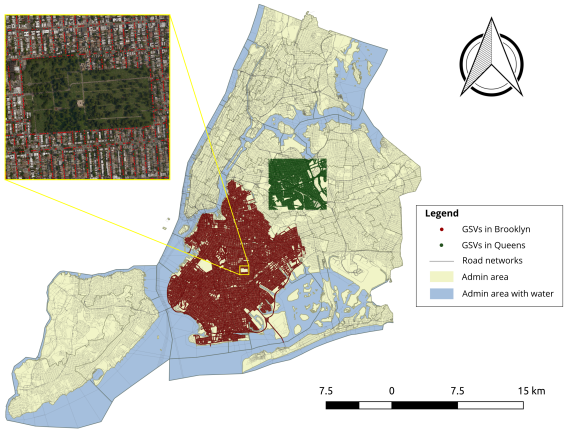

Fig. 3. Overview of the study area. The datasets we used cover Brooklyn and part of Queens borough in New York City. Red and green dots symbolise the locations where Google street views are sampled in the two boroughs respectively. In addition, aerial images are also available in the study area as illustrated in the yellow box.

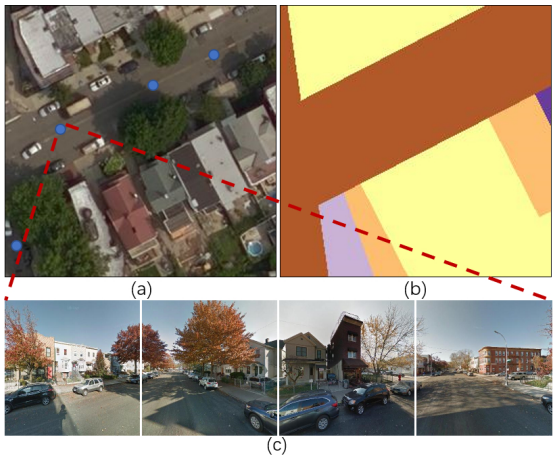

Fig. 4. An example of study data. (a) Aerial image. (b) Segmentation label. (c) Google street views at the pointed position in the aerial image, with different heading directions.

TABLE I
LAND USE CATEGORIES OF NEW YORK CITY.

Code	Land use type	Abbreviation
1	One & two family buildings	FB-1&2
2	Multi-family walk-up buildings	FB-WU
3	Multi-family elevator buildings	FB-E
4	Mixed residential & commercial buildings	Mix.
5	Commercial & office buildings	Com.
6	Industrial & manufacturing	Ind.
7	Transportation & utility	Trans.
8	Public facilities & institutions	Public
9	Open space & outdoor recreation	Open
10	Parking facilities	Parking
11	Vacant land	Vacant

1) Aerial images only: In this group of study, the input data only include aerial images, and original SegNet is used to conduct the segmentation task.

2) Street views only: In this experiment, we first extract semantic features from GSVs, and then interpolate them in the spatial domain to acquire ground feature maps. Next, we use the spatially densified ground feature maps as inputs to SegNet by modifying the input filters' shape to match the dimensions of input ground feature maps.

3) Integrating aerial images and street views: In this study, we try to fuse aerial images and ground feature maps constructed from GSVs. We use a deep convolutional neural network which adds an extract encoder to SegNet, the architecture of the network is shown in Figure 2. The aerial images and ground feature maps are fed into the network together, and then get the final segmentation results.

C. Implementation details

In the experiments, models are implemented based on the PyTorch framework. For training phase, we use Stochastic Gradient Descend optimisation algorithm with a initial learning rate of 0.01, a momentum of 0.9, a weight decay of 0.0005, and a batch size of 16. Learning rate is divided by 10 after the epoch of 15, 25, and 35 epochs. In addition, cross entropy is used as loss function, and the encoders are initialised by VGG-16 weights pretrained on ImageNet. For street views, we use pretrained ResNet-18 based PlacesCNN to extract features and set the cutoff threshold as 30 metres.

D. Evaluation metrics

To evaluate the pixel-level classification results, we adopt overall pixel accuracy, mean IoU, and Kappa coefficient as our evaluation metrics which are frequently used in semantic segmentation and remote sensing area:

1) Pixel accuracy: $p_0 = \sum_{i=1}^{n} x_{ii}/N$

2) Mean IoU: $mIoU = (1/n)\sum_{i=1}^{n} IoU_i$
$(IoU_i = x_{ii}/(\sum_{j=1}^{n} x_{ij} + \sum_{j=1}^{n} x_{ji} - x_{ii}))$

3) Kappa coefficient: $K = (p_0 - p_e)/(1 - p_e)$
$(p_e = \sum_{i=1}^{n}(\sum_{j=1}^{n} x_{i,j} \sum_{j=1}^{n} x_{j,i})/N^2)$

where x_{ij} is the element in confusion matrix, N is the total pixel numbers, and n is the number of classes.

E. Results

We have trained and validated the networks on Brooklyn dataset, and tested them on Queens dataset. For each group of experiment, i.e. aerial images only (*aerial*), ground feature maps only (*ground*), and integration of the two sources of data (*fused*), we have conducted five times of training, and then take the average of test results on those models as the final results. In addition, we ignore the *unknown* category for the final evaluation of classification results.

1) Overall results: The results of the three comparative experiments are listed in Table II. We can see that:

The segmentation results using overhead aerial images alone can already achieve a relatively high pixel accuracy of 77.62% and 74.02%, mean IoU of 47.39% and 38.57% on Brooklyn validation and Queens test sets respectively. Ground feature maps alone can reach an accuracy of 54.31% and 32.94% of pixel accuracy on the two evaluation sets, which proves that the ground feature maps constructed from GSVs contain information that may improve land use classification results. In

addition, the Kappa coefficients of 40.62% and 13.15% shows that the results are fairly consistent and are better than random.

Moreover, our proposed method to fuse overhead and ground-level images achieves an overall accuracy of 78.26% and mIoU of 48.43% on Brooklyn validation set, which are both slightly higher than using aerial images alone; whereas the corresponding evaluation scores on Queens test set witness a slight decrease in accuracy. The results imply that there are redundant information between aerial and street images, the ground-level information may help with urban classification results, but the improvement is limited.

It is interesting that our results of pixel accuracy and mIoU using aerial images alone have achieved much better results than [15], however, the fused results are just a little better than theirs. One explanation to this is that the network we use can already extract enough information thus the contributions from ground-level street views are shadowed.

TABLE II
OVERALL EVALUATION RESULTS.

		Brooklyn valid. set		Queens test set	
		Ours	Workman [15]	Ours	Workman [15]
Aerial	accuracy	**77.62**	69.63	**74.02**	69.27
	kappa	72.50	-	68.01	-
	mIoU	**47.39**	31.70	**38.57**	28.46
Ground	accuracy	**54.31**	44.66	32.94	**47.40**
	kappa	40.62	-	13.15	-
	mIoU	16.11	**18.04**	6.42	**15.04**
Fused	accuracy	**78.26**	77.40	**72.95**	70.55
	kappa	73.29	-	66.86	-
	mIoU	**48.43**	45.54	**37.68**	33.48

2) Per-class results: The overall results reflect the average accuracy of classification among all classes. In order to figure out the variation regarding different land use categories, we compare the IoU metric of each class. The validation and test results of Brooklyn and Queens datasets are shown in Table III and IV respectively. It can be seen from the tables that:

Specific land use types, such as *background*, *one and two family buildings*, and *multi-family elevator buildings* show significantly higher IoU values compared with the average, which may be related to their high percentage of areas and more distinguishable physical appearances. This is consistent with the fact that, in both datasets, *background* class (roads, water area, and etc) accounts for the largest portion of land, followed by easy recognisable residential areas, especially *one and two family buildings* which are usually low-floor villas with big gardens. On the other hand, *parking facilities* and *vacant land* show considerably lower values than the mean IoU which may be caused by the low percentage of pixel numbers of these categories.

It should be noted that, for some categories, the evaluation results vary significantly between Brooklyn and Queens boroughs, for example, *open space and outdoor recreation* achieves an accuracy of more than 15% higher in Queens than that in Brooklyn, which is related to the different urban landscapes of the two areas since the land use area in Queens are significantly larger than that of Brooklyn. This demonstrates that DNN models are dependent on data, their performances are associated with different datasets.

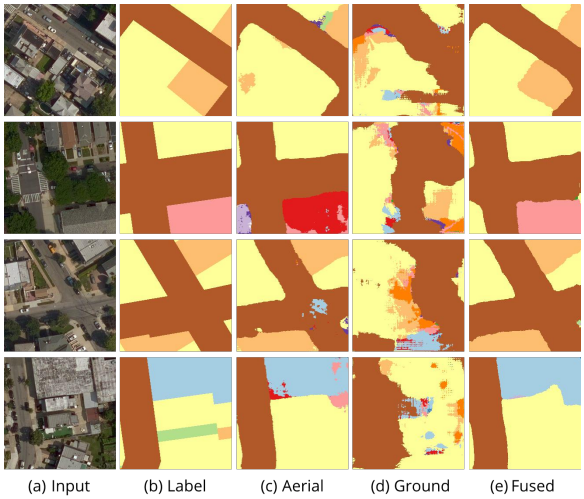

Fig. 5. Segmentation results of the three comparative studies. The first two rows are the evaluation results on Brooklyn validation set, and the others are results on Queens test set.

Typical examples of segmentation results of the three comparative studies are shown in Figure 5. As we can see, ground feature map based segmentation is considerably distorted, while the results of using aerial images alone can achieve much better results. However, the fusion of ground-level information to aerial images helps to refine the segmentation results in these cases.

F. Discussion

From our experiment, we can see that the classification problem using aerial images alone can achieve a good segmentation result, this shows that DNNs have the ability to learn the mapping between different land use types and their inner spatial arrangement and patterns. Instead of selecting features manually, deep learning methods learn the representational features from given data automatically.

Furthermore, the ground feature maps constructed from GSVs also include urban land use information. Thus, we have expected more improvement in accuracy when we try to integrate ground-level street views with aerial images, however, the results are lower than our expectation in spite of some slight improvement on validation set. Possible reasons for this may be, (1) the two sources of data contain duplicated information, and the aerial images already contain much of what there is in the GSVs, (2) the simple approach we used to fuse the two sources may be inefficient and needs further study.

From the results above, we argue that street views can provide ground-level details which overhead aerial images are lacking, but the contributions they make should not be over-exaggerated. Moreover, the proper way to fuse those two kinds of data also matters which remains an open problem for further studies.

TABLE III
PER-CLASS RESULTS ON BROOKLYN VALIDATION SET.

	Background	FB-1&2	FB-WU	FB-E	Mix.	Com.	Ind.	Trans.	Public	Open	Parking	Vacant	mIoU
Aerial	90.27	72.69	46.57	63.52	32.97	35.80	49.19	47.42	44.79	**45.77**	17.12	22.52	47.39
Ground	69.93	40.70	17.20	15.00	10.38	3.41	26.52	0.03	5.16	5.00	0.00	0.00	16.11
Fused	**90.72**	**73.09**	**47.57**	**64.36**	**35.42**	**37.35**	**51.25**	**49.84**	**45.70**	45.38	**17.62**	**22.87**	**48.43**

TABLE IV
PER-CLASS RESULTS ON QUEENS TEST SET.

	Background	FB-1&2	FB-WU	FB-E	Mix.	Com.	Ind.	Trans.	Public	Open	Parking	Vacant	mIoU
Aerial	**78.28**	70.92	29.54	**61.35**	**18.87**	**33.39**	**25.40**	35.64	**28.09**	**62.64**	10.26	**8.41**	**38.57**
Ground	45.13	16.50	6.47	0.41	1.51	1.08	4.13	0.01	1.00	0.75	0.00	0.00	6.42
Fused	76.64	**71.11**	**30.21**	57.23	18.65	29.47	24.63	**38.21**	25.40	61.82	**10.72**	8.11	37.68

VI. CONCLUDING REMARKS

Urban land use is of great significance to urban planning and management. Traditional urban land use mapping relies heavily on domain experts, which is labour-intensive and expensive. To alleviate the situation, we experiment with DNN-based models to label urban land in pixel level, using aerial images and ground-level street views. We have tested our methods on a dataset of New York City. The results show that it is possible to predict urban land use from overhead images, and the integration of ground-level street views can improve the pixel-level classification accuracy to some extent, but the improvement is limited and the fusion strategy remains an open question. In the future, we may try other fusion methods and utilise more sources of data, to further improve the classification results, and to provide more insights for the understanding of urban land use and our cities.

ACKNOWLEDGEMENT

The author acknowledges the financial support from the International Doctoral Innovation Centre, Ningbo Education Bureau, Ningbo Science and Technology Bureau, and the University of Nottingham. This work was also supported by the UK Engineering and Physical Sciences Research Council [grant number EP/L015463/1].

REFERENCES

[1] W. Tu, Z. Hu, L. Li, J. Cao, J. Jiang, Q. Li, and Q. Li, "Portraying Urban Functional Zones by Coupling Remote Sensing Imagery and Human Sensing Data," *Remote Sensing*, vol. 10, no. 1, p. 141, 2018.

[2] Y. Jia, Y. Ge, F. Ling, X. Guo, J. Wang, L. Wang, Y. Chen, and X. Li, "Urban Land Use Mapping by Combining Remote Sensing Imagery and Mobile Phone Positioning Data," *Remote Sensing*, vol. 10, no. 3, p. 446, 2018.

[3] F. Pacifici, M. Chini, and W. J. Emery, "A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification," *Remote Sensing of Environment*, vol. 113, no. 6, pp. 1276–1292, 2009.

[4] Y. Zhu and S. Newsam, "Land Use Classification Using Convolutional Neural Networks Applied to Ground-level Images," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York, NY, USA, 2015, pp. 61:1–61:4.

[5] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A Review on Deep Learning Techniques Applied to Semantic Segmentation," *arXiv:1704.06857 [cs]*, 2017.

[6] X. X. Zhu, D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[7] S. Lefèvre, D. Tuia, J. D. Wegner, T. Produit, and A. S. Nassaar, "Toward Seamless Multiview Scene Analysis From Satellite to Street Level," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1884–1899, 2017.

[8] N. Audebert, B. L. Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 20–32, 2018.

[9] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Las Vegas, NV, USA, Jun. 2016, pp. 680–688.

[10] 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," *International Journal of Geographical Information Science*, vol. 28, no. 9, pp. 1988–2007, 2014.

[11] W. Tu, J. Cao, Y. Yue, S.-L. Shaw, M. Zhou, Z. Wang, X. Chang, Y. Xu, and Q. Li, "Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns." *International Journal of Geographical Information Science*, vol. 31, no. 12, pp. 2331–2358, 2017.

[12] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018.

[13] X. Liu, J. He, Y. Yao, J. Zhang, H. Liang, H. Wang, and Y. Hong, "Classifying urban land use by integrating remote sensing and social media data," *International Journal of Geographical Information Science*, vol. 31, no. 8, pp. 1675–1696, 2017.

[14] T. Hu, J. Yang, X. Li, and P. Gong, "Mapping Urban Land Use by Using Landsat Images and Open Social Data," *Remote Sensing*, vol. 8, no. 2, p. 151, 2016.

[15] S. Workman, M. Zhai, D. J. Crandall, and N. Jacobs, "A Unified Model for Near and Remote Sensing," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy*, 2017, pp. 2707–2716.

[16] W. Zhang, W. Li, C. Zhang, D. M. Hanink, X. Li, and W. Wang, "Parcel-based urban land use classification in megacity using airborne LiDAR, high resolution orthoimagery, and Google Street View." *Computers, Environment and Urban Systems*, vol. 64, pp. 215–228, 2017.

[17] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[19] B. Zhou, . Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2018.

[20] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, 2014.