SOCIAL INFORMATION FUSED URBAN FUNCTIONAL ZONES CLASSIFICATION NETWORK

Weipeng Lu, Chao Tao*, Qi Ji, Haifeng Li

School of Geosciences and Info-Physics, Central South University, Changsha 410083, China - weipeng_lu@163.com

Commission III, WG III/6

KEY WORDS: Urban Functional Zones, Point of Interest, Remote Sensing, Classification

ABSTRACT:

Fast-changing cities need efficient management. Accurate classification of urban functional zone (UFZ) can provide important reference for cities management. Remote sensing imagery (RSI) is large scale, high resolution and fast update, which can provide massive data for UFZ extraction. However, UFZ are more concerned with social attributes such as industrial production and commercial activities, while images can only provide visual features, which is not enough for an elaborate UFZ classification. To solve this problem, in this paper, we combine RSI and point of interest (POI) data together for UFZ classification, and propose a Social Information Fused Urban Functional Zones Classification Network (SIF-Net). For RSI, we simply use a Xception CNNs network extract the visual information. For POI data, we first build a coarse heatmap for each type of POI (e.g. retail, apartment. . .), and then combine them as a POI tensor. Afterward, we use a channel attention module (CAM) based CNN model to fuse heatmaps from each type of POI, and then build a fine distribution of UFZ as the social information. Finally, we fuse the visual information extracted from RSI and social information extracted from POI by concatenating them. By fusing this two complementary information, our method makes up for the shortcomings of extracting UFZ based on RSI and general CNNs only. Compared with current state-of-the-art methods, experiments show that the proposed SIF-Net can significantly improve the UFZ classification result.

1. INTRIODUCTION

Rapid development brings a more complex city. Urban functional zones (UFZ) like residential zone, commercial zone and industrial zone as the basic cells in the city, play an important role for city operation(Song et al., 2019). Detecting UFZ quickly and accurately can provide a valuable reference for urban management(Wang et al., 2017).

In recent decades, thanks to the improvement of satellites technology, very high-resolution (VHR) remote sensing imagers (RSI) are available, which provides massive data for urban research. Differ from the low-resolution imageries that contain several objects in one pixel, objects in VHR imageries are much clearer. Thus, many groups have proposed methods for VHR image analysis. Object-based analysis is a stage of image processing that focus on features like color, outline, shape, texture etc. of an object in image(Lowe, 2004, Yang, Newsam, 2010, Luo et al., 2013). Based on this view, many methods for UFZ classification were proposed(Zhang et al., 2018b, Zhang et al., 2018a). Usually, a UFZ is composed by several specific objects. For example, a residential area will contain buildings, some green plants and road, while besides buildings, large parking lots and cars may appear in commercial area like a supermarket. Zhang et al. proposed hierarchical semantic cognition by mining the scene information contained in objects combinations and distribution (Zhang et al., 2018a). However, to extract feature, variety handcrafted operators were designed which costs a lot of time and these fixed operators can only extract specific feature, whose robustness is unsatisfied(Cheng et al., 2017).

In 2012, achievement of AlexNet brings deep learning, especially convolutional neural networks (CNNs), which do well in visual information extraction, back to image processing(Krizhevsky et al., 2012). Since then, a large amount of work was carried out. Two factors were mainly focused. First is represented by VGG (Simonyan, Zisserman, 2014) and ResNet (He et al., 2016) that committed to solving the structure and training problems of general network models(Szegedy et al., 2015, Ioffe, Szegedy, 2015, Szegedy et al., 2016, Szegedy et al., 2017, Chollet, 2017, Long et al., 2015). Second is represented by methods proposed in(Liu et al., 2018, Cheng et al., 2018, Wang et al., 2018),that adapted the network to a specific task such as UFZ classification from RSI. Liu et al. found multi-scale features of remote sensing images can help distinguish scenes of different sizes, so multi-layered pyramids are introduced in CNNs to analyze the images(Liu et al., 2018). And in (Cheng et al., 2018, Wang et al., 2018), two special loss functions are used to train the network to distinguish small differences between different scenes that are visually similar.

Although the previously method has improved on UFZ classification to some extent, it can still be clearly seen that there are a lot of misclassification on commercial area, industrial area etc. in NWPU-RESISC45 data set(Cheng et al., 2017) compared with other categories(Cheng et al., 2018). Finding that these methods are focus on objects relationship and visual feature in an imagery, it is easy to confuse scenes like commercial area and industrial area who are very similar visually and compositionally. When it comes to UFZ, no matter which part of a city, visual and object are similar. Therefore, it is necessary to find other unique characteristics of different UFZ.

UFZ is a descriptions of land use or in another word, how people use a zone. When people are involved, social activities such as commercial activities, industrial production, and residential life are also involved. Unlike visual features, social features are difficult to obtain from RSI. In past several

* Corresponding author

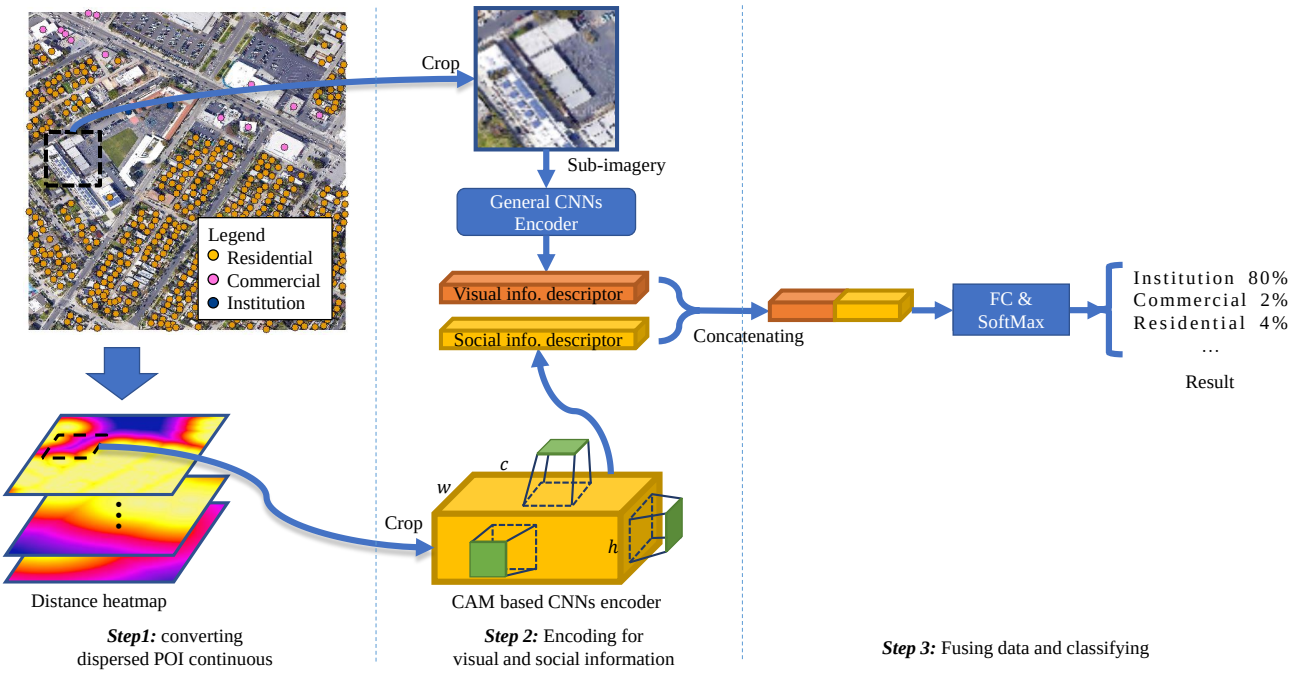

Figure 1. Flowchart of SIF-Net. There are three main steps. Step one, converting dispersed POI data continuous. Step two, crop POI and RSI data and encoding them. Step three, fusing two data by concatenating them and classifying.

years, geographical data with social attributes like point of interests (POI), social media log-in location, population heat map etc. were introduced into RSI analysis(Hu, Han, 2019a, Chen et al., 2018a, Chen et al., 2018b, Hu, Han, 2019b, Song et al., 2018, Yang et al., 2018, Ge et al., 2019). With POI only, Yang et al. proposed a unsupervised methods to establish UFZ, in which shared nearest neighbour(Ertöz et al., 2003) was utilised to build clusters of POI, and K-means is used to classify these clusters taking the frequency of POI in a cluster as the descriptor of this cluster(Yang et al., 2018). With POI and RSI, Song et al. carried out an experiment in Xiamen, China and proposed a method that introduce POI data to extract UFZ based on remote sensing images(Song et al., 2018). This team conducted a survey on POI and UFZ, in which questionnaire was used to determine the weight of various POI in determining UFZ. Moreover, Cao et al. introduced, besides POI, human heat map, social media data and night-time imagery into RSI analysis(Hu, Han, 2019a).

POI is updated by clients and administers with rich social attribution, and it is getting more and more attention. Distribution and combination of different POI (e.g. retail and apartment) will suggest UFZ class. For instance, centralized apartments and scattered retail stores are likely to be residential area. And in this paper, we consider POI as the source of social information. Previous methods for POI analysis regard POI as dispersed data and did not take the spatial influence of a POI into consideration. For instance, a vast airport corresponds to one POI. It is possible that an airport will be divided into several sub-images. Consequently, only one sub-image will contain this POI, while others cannot share POI information. Therefore, it is worth converting POI continuous. Besides, the skill for processing POI data and RSI are still based on surveys and statistics, which cannot mine deeper information contained in POI(Xia et al., 2017). It is considerable to analyse POI data with deep learning skills like CNNs.

In this paper, we propose a Social Information Fused Urban Functional Zones Classification Network (SIF-Net) that can fuse POI and RSI data for UFZ classification. Particularly, to make discrete data continuous, we convert POI into heatmaps based on distance by type of POI and concatenate them into a tensor. POI's distribution is shown on each heatmap, and its combination of different kinds of POI is contained in channels. Another problem is CNNs is not good at extracting information between channels (Zhong et al., 2019, Woo et al., 2018). POI does not have as many visual features as images for CNNs to extract, and forasmuch, it is necessary to find a way to mine information from the perspective of channels. We design a channel attention module (CAM) based on CNNs to mine POI combination from channels of tensor. A general CNNs encoder and a CNNs encoder with CAM will processing RSI and POI separately into two descriptor vectors, and a FC layer with SoftMax attaching to the end of SIF-Net will predict class of UFZ from these two vectors. Several experiments are carried out, proving that the introduction of POI data and channel attention mechanism can effectively improve the UFZ classification accuracy based on RSI with increasing rate of 11.96% on kappa.

2. METHODOLOGY

RSI has great visual information like color and shape, which do help in land cover analysis, while to achieve elaborate UFZ classification, visual feature is not enough. As an important data of geographic information system, POI is uploaded by administers and clients and contain much social information. Thus, we introduce POI to the UFZ classification task, hoping to improve classification accuracy through its social attribution. It is well known that POI is point data, and UFZ is a region. Just like Figure 1 shows, to avoid that some image of region has no POI data like mentioned in section I, in step 1, SIF-Net will first make the POI data continuous by building distance heatmap. In step 2, we design a channel attention module based on CNNs to encode information contained in heatmap from the channel and visual aspects into a social information descriptor.

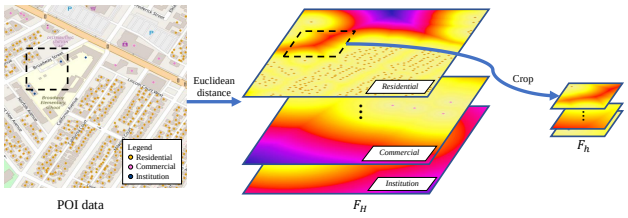

Figure 2. An example of building heatmap tensor according to Euclidean distance.

At the same time, a pretrained deep CNN model will extract visual information in imagery and covert imagery into a visual information descriptor. In step 3, social information and visual information descriptors will be fused into a new descriptor by concatenation, and then this new descriptor will be classified by fully connected (FC) layer.

2.1 Continuous heatmap converted from dispersed POI

As we all know that a POI, taking airport for example, is not mean the pixel that this POI located in is airport, but a vast region around this POI is airport. However, a POI cannot distribute on two no-overlap sub-images, but an airport is likely be divided into more than one sub-image, which result that there are always some sub-images that have no POI of airport in it. To avoid such phenomenon, it is necessary to make POI continuous. Finding that POI has a strong influence on the near, but a weak influence on the far, distance can describe this influence.

For each category of POI in a large region with size of $M \times N$, we build a distance heatmap tensor like Figure 2. Assuming that there are totally c types of POI in research region, for i^{th} type, the value of pixels that POI located in is zero and value of rest is the Euclidean distance between them and the POI closest to them. In this way, a heatmap $H_i \in \mathbb{R}^{M \times N}$ is built, in which value of each pixel is the distance between it and POI closest to it.

Overlaying all c heatmaps, a 3-dimensional tensor $F_H = [H_1, H_2, \cdots, H_c] \in \mathbb{R}^{c \times M \times N}$ is established. Finally, region of imagery with size of $m \times n$ in F_H is cropped and noted as $F_h \in \mathbb{R}^{c_1 \times m_1 \times n_1}$.

2.2 Encoding for visual and social information

RSI contains visual information and POI contains social information. The organizational structure of the two is different, so it is necessary to encode them separately before fusing them. It is believed that a vector can describe an image (Xia et al., 2017). And two CNNs will convert RSI and POI into vectors $v_r \in \mathbb{R}^{d_r}$ and $v_p \in \mathbb{R}^{d_p}$.

2.2.1 General CNNs for visual information encoding For RSI, there are a number of encoders based on general CNNs for feature extraction. For instance, Xception (Chollet, 2017) make RGB image $I \in \mathbb{R}^{3 \times 299 \times 299}$ into 2048-dimension vector $v_r \in \mathbb{R}^{2048}$.

2.2.2 CAM based CNNs for social information encoding In recent year, channel attention mechanism is introduced into image processing to extract channel relationship in high dimension feature map(Zhong et al., 2019),(Woo et al., 2018). Usually, a convolutional layer does convolution from the height-width direction. Its kernel size is always three, five or seven. To do such convolution, each layer of the convolution kernel

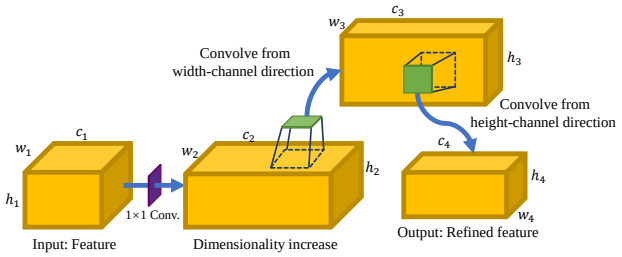

Figure 3. Architecture of CAM. There are three convolutional layers in a CAM. A convolutional layer with size of one is for dimensionality increase, and two layers for extract feature from channel direction.

is convolved with the corresponding channel, and then all results are added. This convolution extracts visual feature of each channel firstly, and then fuses these features, mining a weak channel relationship. Besides POI distribution contained in each correspond channel of F_h, it is worth thinking that different POI's combination also indicates UFZ class. This combination of different kinds of POI is embedded in channels relationship. Thus, CAM (Figure 3) is proposed to enhance channel relationship mining. CAM considers two aspects: adapting kernel with size of one and convolving from height-channel direction and width-channel direction.

Firstly, let feature map F_h as the input of CAM, a 1×1 convolution layer will increase F_h into a dimensionality increasing feature map $F_{di} \in \mathbb{R}^{c_2 \times h_2 \times w_2}$, where $h_2 = h_1$ and $w_2 = w_1$, according to discipline of convolution when padding of layer is zero and stride is equal to one. On the one hand, 1×1 convolution layer will not consider information around kernel centre, which means it extract channel relationship directly. On the other hand, more dimension in F_{di} can provide more channel information for mining.

Secondly, rather than from height-width direction, two convolutional layers will convolute F_{di} from width-channel direction and height-channel direction sequentially. In this way, more detail in adjacent channels will be observed, and refined feature map F_{out} is obtained, whose size depends on the parameters of two channel direction convolutional layers above.

CAM is a block that can be attached to any part of the CNNs. Differ from CNNs for RSI, SIF-Net will use a CNNs encoding with CAM to encode social information.

2.3 Fusion for POI and RSI and classification

SIF-Net has transformed POI and RSI into v_p and v_r separately. Finally, SIF-Net fuses two descriptor by concatenating them as $v_f = [v_p, v_r] \in \mathbb{R}^{d_p + d_r}$, and classifying v_f by a FC layer and SoftMax function.

3. EXPERIMENT AND ANALYSIS

3.1 Experiment

In order to test the effectiveness of SIF-Net, we designed the following experiments.

We chose CSU-RSISC10 (`https://github.com/SalviaL/CSU-RSISC10-DATASET`) as the data set that has 10 categories (road, commercial, industrial, residential, redeveloped area,

Model	F1 score										κ
	Commercial	Industrial	Residential	Redeveloped area	Institutional	Harbor	Water	Open space	Airport	Road	
Xception	0.3922	0.3541	0.8640	0.3793	0.2838	0.8615	0.9380	0.8340	0.8421	0.8131	0.7638
SPP_Net+KML	0.4293	0.4680	0.8734	0.3750	0.1746	0.8265	0.9109	0.8260	0.8566	0.8133	0.7624
SIF-Net without CAM	0.5996	0.8204	0.8953	0.6667	0.7548	0.9065	0.9587	0.8803	0.9650	0.7832	0.8213
SIF-Net	0.6879	0.8424	0.9057	0.6842	0.7529	0.8991	0.9641	0.8804	0.9804	0.8000	0.8552

Table 1. Evaluation for experiment of four methods on CSU-RSISC10

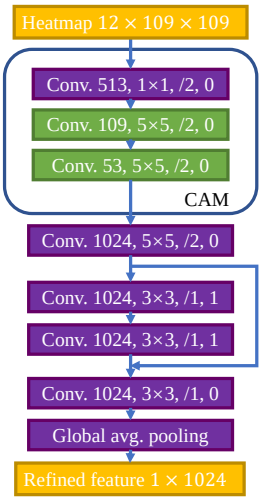

Figure 4. Architecture of encoder for POI with CAM. Purple box is operator from height-width direction and green box is from channel direction as described in section 2.2.2 Numbers in boxes are output channels, kernel size, stride size and padding size.

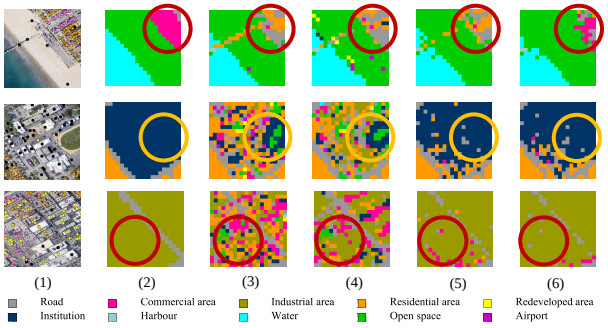

Figure 5. UFZ classification results using CSU-RESISC10 dataset. (1) Images and POI. (2) Ground truth. (3) Xception. (4) SPP-Net with MKL (5) SIF-Net without CAM. (6) SIF-Net. Small circles in (1) are POI data. Orange circles represent houses and apartments. Pink represent hotel, retail, and commercial. Yellow, blue, and black represent industrial, institution and unclassified. Conclusion

institutional, harbour, water, open space and airport), over one million sample with size of 100×100 pixels and is built from a continuous region. We divide data into training set and validation set by region for convenience of visualization. And it is also easy for us to download POI data from OpenStreetMap (OSM). Some kinds of POI are with small amount and we merge them into another similar categories. The following 12 type are POI we obtained from OSM. They are unclassified, hotel, retail, commercial, apartment, house, residential, industrial, institution, nature, public service and airport. For RSI, we chose Xception(Chollet, 2017) as the encoder and transform an imagery into $v_r \in \mathbb{R}^{2048}$.

For POI, accordingly, 12 distance heatmap are built and cropped out $F_h \in \mathbb{R}^{12\times100\times100}$ by region of imagery. To avoid the loss of central point and edge information of F_h, before information extraction, we resize F_h into size of $12 \times 109 \times 109$ (Wu et al., 2019).

A CNNs with CAM shown in Figure 4 is designed for this experiment. It is composed by a CAM, several convolution layer, a residual block(He et al., 2016), and a global average pooling layer. Through this CNNs, a heatmap F_h will be converted into $v_p \in \mathbb{R}^{1024}$.

Let $v_f = [v_r, v_p] \in \mathbb{R}^{3072}$. And a FC layer attached by a SoftMax will classify v_f into a 10-dimension vector, which represent probability distributions of input imagery of 10 classes in CSU-RSISC10.

To train SIF-Net, cross entropy loss and Adam optimizer are used. There hyperparameters for training are as following: set batch size as 16, learning rate as 1e-5, learning descent as 0.98 per 50 epochs. And we also verify advantages of POI data and CAM in our method through comparative experiments with methods of Xception(Chollet, 2017), SPP-Net with MKL(Liu et al., 2018) and a SIF-Net without CAM. The result of experiments is evaluated by F1 score and kappa (κ)

3.2 Analysis

From Table 1, it is obviously that both SIF-Net without CAM and SIF-Net achieves better result on κ. Comparing to methods only using RSI, κ of POI involved methods has an increase rate over 7.52% and 11.96%. Focusing on each category, two SIF-Net gets a better result on most of categories than other two methods. And a dramatic increasing is occurred on commercial, industrial, institutional area with growth rates of 75.4%, 137.9% and 165.3% compared with SIF-Net and Xception which has no POI data. We also find that SIF-Net get 0.8000 on road which is not as good as previous methods. We suppose that there is not a class of POI related to road, which might result to this decreasing on F1 score. When it comes to CAM, according to κ, SIF-Net performs better overall with an increasing rate of 4.1% than model without CAM. Especially, industrial and redevelop area have an increasing rate over 2.5%. More remarkable is growth rate of F1 score on commercial over 14.7%. It is proved that the introduction of CAM is effective. From the visualized image (Figure 5), accuracy increasing on commercial area, industrial area and institution are much more obvious. It is difficult to distinguish whether a building belongs to a residential, commercial or industrial area, because the similarity on shape and color, while SIF-Net fuses social information that point out what this building is used to. Similarly, it is difficult to judge whether the green space belongs to open space (e.g. park) or institution (e.g. school). For other two methods without POI data, the result in red circle is quite a chaos, and in row two of Figure 5 circled in yellow, playgrounds in school is misclassified into open space, while SIF-Net avoids this to some extent. We suppose that the social information provided by POI makes contribution to this improvement.

4. CONCLUSION

The social information provided by POI data can improve the performance of UFZ classification task. In this paper, we proposed Fused Urban Functional Zones Classification Network (SIF-Net) that fuses social information from POI and visual information from RSI. Firstly, SIF-Net create continuous heatmap tensor converted from dispersed POI to make sure every part of imagery can receive social information. Then, RSI and POI is encoded by general CNNs and CNNs with CAM into two descriptors. Finally, by concatenating this two descriptors, a new descriptor will be classified by FC layer and SoftMax. It is shown that with the help of social information and channel relationship excavation a dramatic improvement is occurred on visually similar UFZ like commercial, industrial and institution. As for further work, it is worth considering spatial information reasoning. Methods that handle sub-imagery one by one always ignore spatial relationship among sub-imageries. Another aspect is data screening. Although various platforms will check the uploaded data, it does not rule out the occurrence of erroneous data. This mistake might cause the network to learn the wrong information while training.

REFERENCES

Chen, W., Huang, H., Dong, J., Zhang, Y., Tian, Y., Yang, Z., 2018a. Social functional mapping of urban green space using remote sensing and social sensing data. *Isprs Journal of Photogrammetry and Remote Sensing*, 146, 436–452.

Chen, Y., Ge, Y., An, R., Chen, Y., 2018b. Super-resolution mapping of impervious surfaces from remotely sensed imagery with points-of-interest. *Remote Sensing*, 10(2), 242.

Cheng, G., Han, J., Lu, X., 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10), 1865-1883.

Cheng, G., Yang, C., Yao, X., Guo, L., Han, J., 2018. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5), 2811-2821.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 1800–1807.

Ertöz, L., Steinbach, M., Kumar, V., 2003. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. *Proceedings of the 2003 SIAM international conference on data mining*, SIAM, 47–58.

Ge, P., He, J., Zhang, S., Zhang, L., She, J., 2019. An Integrated Framework Combining Multiple Human Activity Features for Land Use Classification. *ISPRS international journal of geo-information*, 8(2), 90.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 770–778.

Hu, Y., Han, Y., 2019a. Identification of Urban Functional Areas Based on POI Data: A Case Study of the Guangzhou Economic and Technological Development Zone. *Sustainability*, 11(5), 1-15.

Hu, Y., Han, Y., 2019b. Identification of Urban Functional Areas Based on POI Data: A Case Study of the Guangzhou Economic and Technological Development Zone. *Sustainability*, 11(5), 1-15.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (eds), *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., 1097–1105.

Liu, Q., Hang, R., Song, H., Li, Z., 2018. Learning Multiscale Deep Features for High-Resolution Satellite Image Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1), 117-126.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Luo, B., Jiang, S., Zhang, L., 2013. Indexing of Remote Sensing Images With Different Resolutions by Multiple Features. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(4), 1899-1912.

Simonyan, K., Zisserman, A., 2014. VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. *arXiv: Computer Vision and Pattern Recognition*.

Song, J., Lin, T., Li, X., Prishchepov, A. V., 2018. Mapping Urban Functional Zones by Integrating Very High Spatial Resolution Remote Sensing Imagery and Points of Interest: A Case Study of Xiamen, China. *Remote Sensing*, 10(11), 1737.

Song, J., Tong, X., Wang, L., Zhao, C., Prishchepov, A. V., 2019. Monitoring finer-scale population density in urban functional zones : A remote sensing data fusion approach. 190.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. *Thirty-First AAAI Conference on Artificial Intelligence*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Wang, L., Fang, F., Yuan, X., Luo, Z., Liu, Y., Wan, B., Zhao, Y., 2017. Urban function zoning using geotagged photos and openstreetmap. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 815–818.

Wang, Y., Zhang, L., Tong, X., Nie, F., Huang, H., Mei, J., 2018. LRAGE: Learning Latent Relationships With Adaptive Graph Embedding for Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2), 621-634.

Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., 2018. Cbam: Convolutional block attention module. V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds), *Computer Vision – ECCV 2018*, Springer International Publishing, Cham, 3–19.

Wu, S., Wang, G., Tang, P., Chen, F., Shi, L., 2019. Convolution with even-sized kernels and symmetric padding. *Neural Information Processing Systems*, 1192–1203.

Xia, G., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7), 3965-3981.

Yang, J., Cao, J., He, R., Zhang, L., 2018. A unified clustering approach for identifying functional zones in suburban and urban areas. *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 94–99.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, GIS '10, Association for Computing Machinery, New York, NY, USA, 270–279.

Zhang, X., Du, S., Wang, Q., 2018a. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. *Remote Sensing of Environment*, 212, 231–248.

Zhang, X., Du, S., Wang, Q., Zhou, W., 2018b. Multiscale Geoscene Segmentation for Extracting Urban Functional Zones from VHR Satellite Images. *Remote Sensing*, 10, 281.

Zhong, X., Gong, O., Huang, W., Li, L., Xia, H., 2019. Squeeze-and-excitation wide residual networks in image classification. *2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, China, 395–399.