Local Climate Zone Mapping as Remote Sensing Scene Classification Using Deep Learning: A Case Study of Metropolitan China

Shengjie Liu[a,b,*], Qian Shi[b,*]

[a]Department of Physics, University of Hong Kong, Pokfulam, Hong Kong SAR
[b]Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation, School of Geography and Planning,
Sun Yat-sen University, Xingang Road West, Guangzhou 510275, China

Abstract

This is the preprint version. To read the final version, please go to https://doi.org/10.1016/j.isprsjprs.2020.04.008

China, with the world's largest population, has gone through rapid development in the last forty years and now has over 800 million urban citizens. Although urbanization leads to great social and economic progress, they may be confronted with other issues, including extra heat and air pollution. Local climate zone (LCZ), a new concept developed for urban heat island research, provides a standard classification system for the urban environment. LCZs are defined by context of the urban environment; the minimum diameter of an LCZ is expected to be 400-1,000 m so that it can have a valid effect on the urban climate. However, most existing methods (e.g., the WUDAPT method) regard this task as pixel-based classification, neglecting the spatial information. In this study, we argue that LCZ mapping should be considered as a scene classification task to fully exploit the environmental context. Fifteen cities covering 138 million population in three economic regions of China are selected as the study area. Sentinel-2 multispectral data with a 10 m spatial resolution are used to classify LCZs. A deep convolutional neural network composed of residual learning and the Squeeze-and-Excitation block, namely the LCZNet, is proposed. We obtained an overall accuracy of 88.61% by using a large image (48×48 corresponding to 480×480 m^2) as the representation of an LCZ, 7.5% higher than that using a small image representation (10×10) and nearly 20% higher than that obtained by the standard WUDAPT method. Image sizes from 32×32 to 64×64 were found suitable for LCZ mapping, while a deeper network achieved better classification with larger inputs. Compared with natural classes, urban classes benefited more from a large input size, as it can exploit the environment context of urban areas. The combined use of the training data from all three regions led to the best classification, but the transfer of LCZ models cannot achieve satisfactory results due to the domain shift. More advanced domain adaptation methods should be applied in this application.

Keywords: local climate zone, convolutional neural network, scene classification, metropolitan China, urban climate

1. Introduction

The United Nations has proposed 17 Sustainable Development Goals (SDGs) for a shared blueprint for the peace and prosperity of human beings and the planet. With an emphasis on SDG 11 Sustainable Cities and Communities and SDG 13 Climate Action, urban scientists are devoted to investigating local climate in cities (Masó et al., 2019; Campbell et al., 2018; Zhu et al., 2019b), where over 55% of the world's population live in (Nations, 2015). The LCZ scheme proposed by Stewart and Oke (2012) for urban heat land study has attracted many urban scientists' attention, as it provides a standard for worldwide urban structure classification (Bechtel and Daneke, 2012; Xu et al., 2017b; Perera and Emmanuel, 2018; Liu et al., 2019; Demuzere et al., 2019a). Based on building height, impervious proportion, surface texture, etc., the LCZ scheme classifies global land covers into seventeen categories, including ten urban classes and seven natural classes.

China has gone through rapid social and economic development in the last forty years; now up to 800 million population in China live in cities (Luo et al., 2018). Urban structure in China is different from other (western) countries. Due to the large population and strict land supplies, high-density urban regions (LCZ-1 compact high-rise) are more popular in China (Güneralp et al., 2017; Huang and Wang, 2019). Hong Kong and Macau, the two special administrative regions of China, have an even higher level of urban density due to limited land resources (Lau et al., 2019). Slums are rare in China; instead, urban villages are quite common. Urban villages are extremely compact regions with a typical height of seven to eight floors (you can even shake your hands with the person in the neighboring building, as shown in the upper left of Figure 1); buildings with more than eight floors are required to install lifts. But inside the buildings, the houses are well decorated. Basic equipment, including electricity, water and internet supply, is also available. Urban villages provide the youth with a cheap living option in major cities (Kuffer et al., 2016; Wu, 2016). Although slums and urban villages are related in social science, they are often classified as LCZ-3 compact low-rise (Cai et al., 2016)

[*]Corresponding authors
Email addresses: liusj@hku.hk; sjliu.me@gmail.com (Shengjie Liu), shixi5@mail.sysu.edu.cn (Qian Shi)

and LCZ-7 lightweight low-rise, respectively, and are easily mixed in LCZ mapping (Kotharkar and Bagade, 2018). Therefore, to generate precise LCZ maps in China, special arrangements should be made to address issues like this.

The World Urban Database and Access Portal Tools (WUDAPT) initiative is proposed to use free Landsat satellite data and random forest to classify LCZs in a spatial resolution of 100 *m* worldwide. The WUDAPT method first collects Landsat satellite data and resamples them to 100 *m*. Then, the satellite and pre-collected reference data from Google Earth are used to map LCZs using a random forest classifier. It becomes the standard for LCZ mapping and is widely adopted (Bechtel et al., 2015; Wang et al., 2018a). Still, two major issues are encountered. First, an LCZ is defined by the context of urban environment, but the this method did not consider spatial information. Second, it requires local experts to collect training samples, which is labor intensive.

Due to limitation of the WUDAPT method, many studies have investigated other methods for LCZ mapping. Some tackle the availability of training samples by using transfer learning and "borrow" training samples from other regions. Xu et al. (2017a) proposed a domain-adaptation co-training approach with self-pace learning to classify LCZs, where training samples from existing cities were reused and transferred to new target cities. Qiu et al. (2019) investigated the transferability of LCZ samples in European cities by using multi-seasonal Sentinel-2 images and found cities at lower latitude were less sensitive to seasonal characteristics. Demuzere et al. (2019b) tested the global transferability of LCZ models on the Google Earth Engine platform. Although local training samples were still required, training samples from the same ecoregion considerably enhanced the classification. Ensemble learning with multisource data is highlighted in this task. In 2017, an LCZ mapping contest was hosted by the IEEE Geoscience and Remote Sensing Society, where the first-prize winner used a combination of random forests and canonical correlation forests with expert handcrafted features to earn the honor (Yokoya et al., 2018). Another research utilizing multisource data was conducted by Qiu et al. (2018). They used five kinds of data, including global urban footprint, open street map, VIIRS night lights, Sentinel and Landsat multispectral data, and showed that all data could contribute to classification. Others explored the task in spatial domain. Zheng et al. (2018) found that building surface fraction was sensitive to the geolocation of raster grids. Kotharkar and Bagade (2018) applied the LCZ scheme in Nagpur, India with the overlay technique over a grid of 250, 500 and 1000 *m*. They found seven additional LCZ subclasses as a result of mixing of two or more classes. However, to this end, existing literature mostly classify LCZs in a pixel-based manner. Exceptions are, a) the So2Sat LCZ42 dataset presented by Zhu et al. (2019a); b) a recent study presented by Rosentreter et al. (2020). These two studies both used an image size of 32×32 to generate LCZ maps in Europe, but neither of them explained the reason to use scene classification method or the choice of this particular image size, leaving two important questions: 1) why scene classification is a better solution to LCZ mapping and 2) what is the optimal scene size for this application?

Since the key factors of LCZs include sky view factor, roughness class, pervious and impervious fractions, building height and anthropogenic heat flux, the classification of LCZs significantly depends on the surrounding environment context, . In Figure 2, we present some LCZs derived from Sentinel-2 multispectral imagery in a spatial resolution of 10 *m* with image sizes of 10×10, 16×16, 32×32 and 64×64. Let's assume we'd like to map LCZs in a spatial resolution of 100 *m*. The WUDAPT solution is to resample the satellite imagery from 10 *m* to 100 *m*, in which the spatial texture is lost. Another solution is directly classifying the 10×10 image patch to an LCZ class. However, it is difficult for a person to tell the class with a small image, not to say a machine. As shown in Figures 2 (a,b,c) a, we barely recognize the class within a 10×10 image. But if we zoom out and use a larger image (e.g., 64×64) as the representation of an LCZ, we can easily recognize the regions as urban villages (Figure 2a (d)), high-rise buildings with reasonable green covers (Figure 2b (d)) and single houses (Figure 2c (d)).

Therefore, it is only reasonable to use a large image representation to classify LCZs, which directly changes the task from pixel-based mapping to scene classification. Unlike pixel-based mapping, which assigns each pixel with a class, scene classification assigns the entire image (e.g., 64×64) with only one land cover category. In this way, the spatial context of the input image plays an more important role (as it should be) in LCZ mapping. For the difference between scene classification and pixel-based mapping and for deep learning applications in remote sensing, we refer interested readers to (Zhang et al., 2016; Zhu et al., 2017; Ma et al., 2019).

In this study, our primary goal is to investigate the optimal scene size for LCZ mapping using Sentinel-2 images. We classify LCZs in 15 cities over three economic regions in China, which exhibits a unique urban structure from other regions in the world. To do so, we design a very deep CNN for LCZ mapping. We further analyze the effect of image size to individual class. Finally, we investigate the transferability of LCZ models in the three economic regions of China. The remainder of this paper is organized as follows. In section 2, we review some deep learning applications in remote sensing and go through the so-called remote sensing image classification, including pixel-based classification, semantic segmentation, and scene classification. We then introduce the data and three economic regions of China in section 3. In section 4, we describe the LCZNet in details. In sections 5 and 6, we present and analyze the results. Finally, we draw some conclusion in section 7.

2. Remote sensing image classification

In this section, we review what is called remote sensing image classification. In a boarder sense, it means three different paths to classify satellite images, whereas the conventional pixel-based (and object-based) classification is only one of them. The readers may go to one of the three review papers for a more comprehensive perspective of deep learning applications in remote sensing (Zhang et al., 2016; Zhu et al., 2017; Ma et al., 2019).

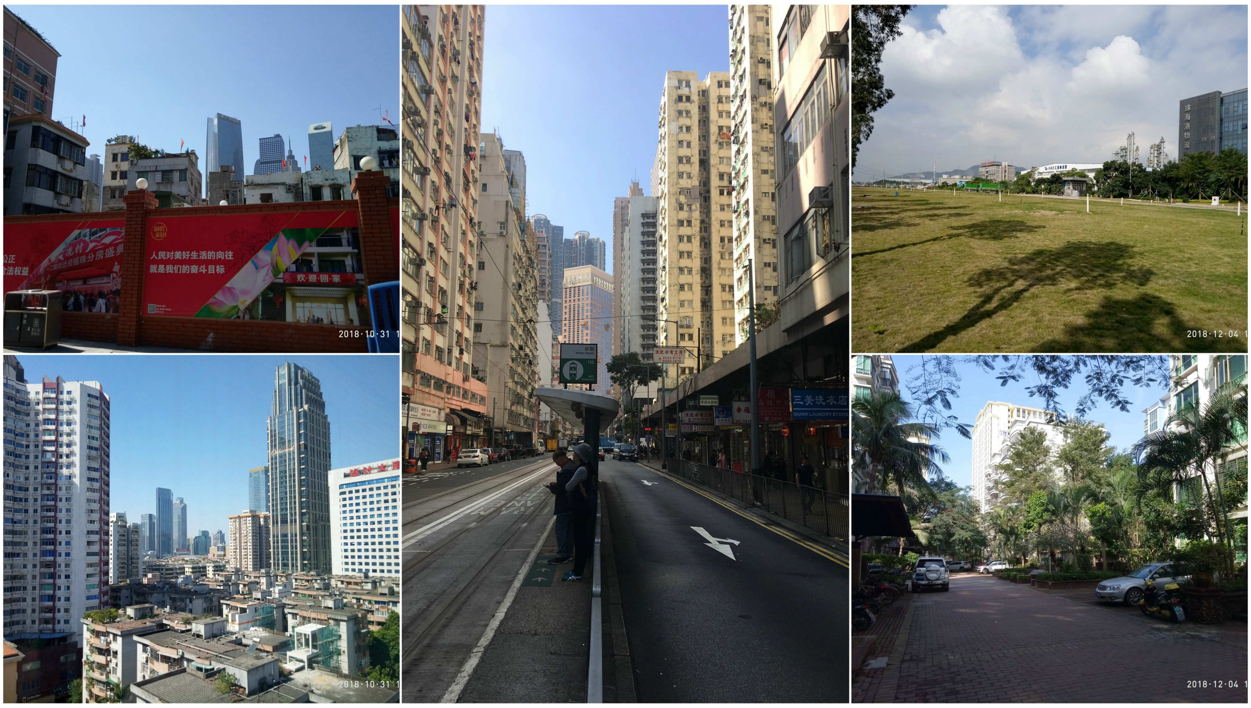


Figure 1: Photos of some of the urban landscape in China. Upper left (Guangzhou): urban villages and skyscrapers. Lower left (Guangzhou): compact mid-rise and compact high-rise. Middle (Hong Kong): extremely compact high-rise. Upper right (Zhuhai): large low-rise and/or heavy industry with sufficient low plants. Lower right (Zhuhai): open mid-rise.

Table 1: Summary of the three remote sensing image classification methods

Method		Input	Output	Reference data
Pixel-wise classification	Pixel-wise classification	1D feature vector (pixel level)	A label for a pixel	Several labels for several pixels
	Object-based image analysis	1D feature vector (object level)	A label for an object	Several labels for several objects
	Patch-based classification	3D tensor (2D feature matrix × channel)	A label for a pixel	Several labels for several pixels
Semantic segmentation		3D tensor (2D image × channel)	All labels for all pixels	All labels for all pixels
Scene classification		3D tensor (2D image × channel)	A label for an image	A label for an image

2.1. *Pixel-based Classification (LULC Mapping)*

Remote sensing image classification is a special computer vision task, which often stands for assigning each pixel of a satellite image with a land cover and land use (LULC) category. The conventional way is by treating each pixel as a sample with rich spectral features, from which we distinguish pixels from each other and label them with the most confident LULC class. Therefore, this task is often referred to as *LULC mapping* (Thenkabail et al., 2005).

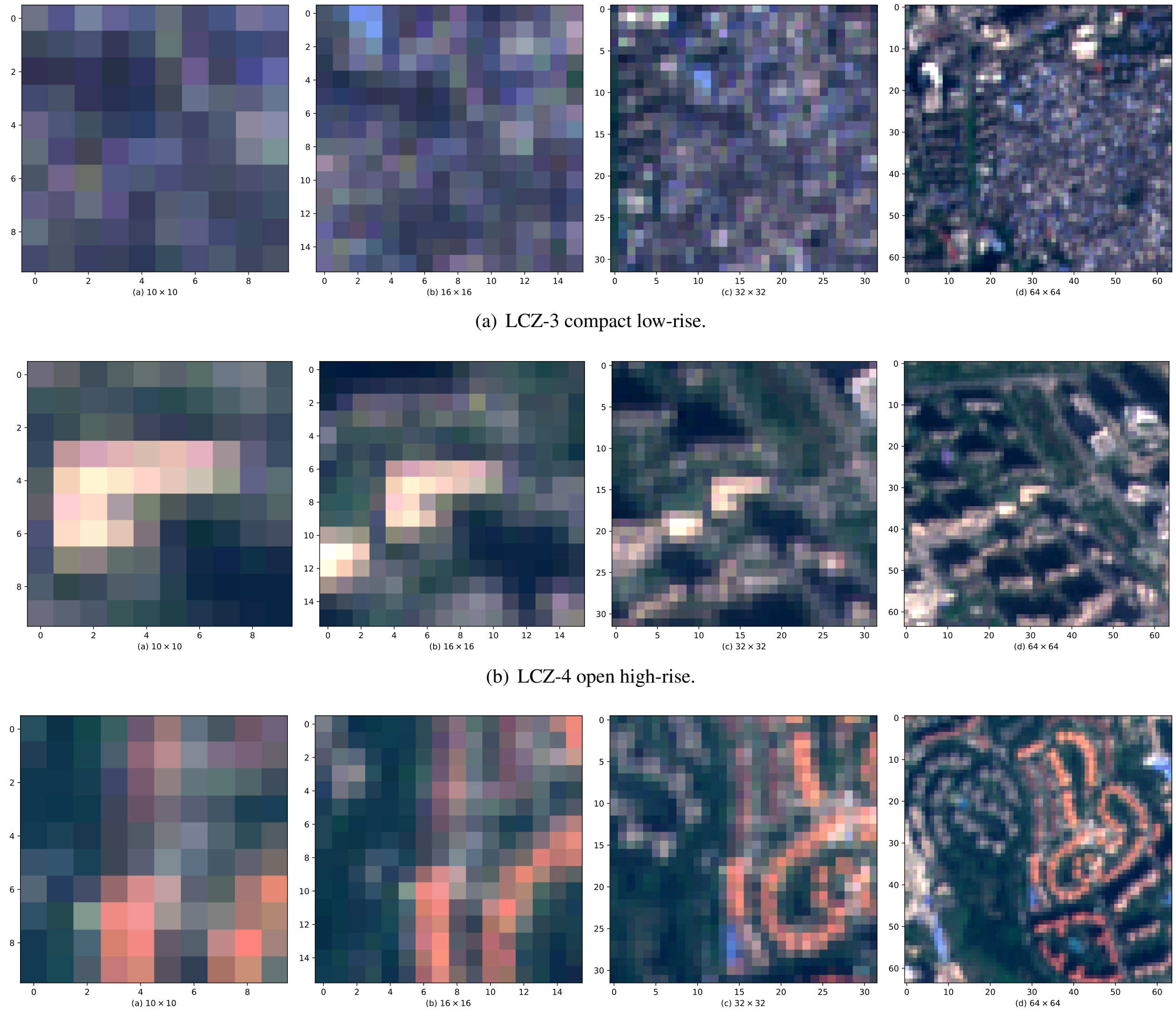
With the literature going deep, remote sensing researchers found that the usage of textural (spatial) information is helpful for *LULC mapping* (Risojević and Babić, 2012). A lot of filters (kernels or operators) are integrated into *LULC mapping* to exploit textural features in a fixed size window, such as morphological profiles and its extended version (Benediktsson et al., 2005), Gabor filters (Li and Du, 2014), and local binary patterns (Li et al., 2015). The common way is to open a fixed size window to extract the surrounding textures of a central pixel to be classified. In this manner, the size of window is often an odd number (e.g., 3×3 and 5×5). Because ground targets are irregular, a fixed size window is not the best representation to extract textural features. *Object-based image analysis (OBIA)* is then proposed to group homogeneous pixels into an object to extract textural and shape features on the object level for *LULC mapping* (Blaschke, 2010).

In the era of deep learning, remote sensing researchers use CNNs to extract textural features, as with CNNs the filters can be adaptive and are more useful for *LULC mapping* (Liu et al., 2019). Since the input sample changes from a pixel to an image patch, this method is sometimes referred to as *patch-based classification* (Sharma et al., 2017). In order to assign a label to the central pixel, the size of input images is still an odd number.

In a word, the most common remote sensing image classification is usually related to *LULC mapping*, and the core is to classify each pixel with a LULC category. The variants of *OBIA* and *patch-based classification* are two ways to extract textural features to achieve better classification, but they are still the same task because training data remain as one satellite image with some pixels labeled.

2.2. *Semantic Segmentation*

A special method called semantic segmentation for satellite image classification is developed in the era of deep learning (Audebert et al., 2016). Although the goal of semantic segmentation is also to classify each pixel with a LULC (often



(a) LCZ-3 compact low-rise.

(b) LCZ-4 open high-rise.

(c) LCZ-6 open low-rise.

Figure 2: Same regions presented as 10×10, 16×16, 32×32 and 64×64 images. One can hardly tell the class based on a small image representation, but it is easier to tell the class based on a larger image representation.

land cover) category, the approach is different. This method treats each satellite image (e.g., 256×256) as a sample instance. Through a special kind of CNNs named the fully convolutional network (FCN) (Long et al., 2015; Kampffmeyer et al., 2016), the output of semantic segmentation is directly the classification (segmentation) result of the entire image, rather than a classification summation for individual pixel. FCN is the first to address this application using deep learning. Later, many other methods were proposed, including SegNet, UNet, PSPNet, RefineNet, DeepLab, etc. We here refer interesting readers to Minaee et al. (2020) for a survey of semantic segmentation algorithms.

Semantic segmentation models often require the training data to be several images with fully annotated reference data. Some recent studies have been devoted to use imperfect annotation, which is under the scope of weakly supervised learning (Song et al., 2019; Rafique and Jacobs, 2019; Wang et al., 2020). The size of the input image is often an even number (e.g., 256×256) in line with those from computer vision tasks. Usually, only a few less than 10 land cover classes are recognized in this task (Marmanis et al., 2016), e.g., the ISPRS Vaihingen and Potsdam data (6 classes) and the DeepGlobe Land Cover Classification Challenge (7 classes), though there is some progress with up to 21 land covers (Azimi et al., 2019). This method is very effective in extracting single class of interest such as roads, building footprints, clouds, ice and water (Maggiori et al., 2016; Shi et al., 2017; Ji et al., 2018; Yang et al., 2019; Zhang et al., 2020).

2.3. Scene Classification

The third type of remote sensing image classification is often referred to as remote sensing scene classification (Cheng et al., 2017; Xia et al., 2017). This task aims at classifying the entire satellite image to only one LULC (often land use)

Table 2: Statistics of the 15 cities in the three studied regions. Type of "Urban" means the images used only cover the core city (downtown). Meanwhile, type of "ALL" means the entire administrative areas are covered in the study. The fifteen cities cover over 138 million population in China.

City	Type	Population (k)	GDP (million USD)
Guangzhou	Urban	13,501	342,200
Shenzhen	ALL	10,779	362,589
Dongguan	ALL	8,343	123,922
Foshan	Urban	7,431	148,727
Huizhou	Urban	4,727	61,422
Zhongshan	ALL	3,210	54,371
Jiangmen	Urban	4,592	43,413
Zhuhai	ALL	1,614	44,401
Macau	ALL	643	54,026
Hong Kong	ALL	7,306	358,817
Shanghai	Urban	24,183	489,206
Hangzhou	Urban	9,468	202,230
Shaoxing	Urban	4,988	81,092
Beijing	Urban	21,707	453,892
Tianjin	Urban	15,596	281,571
Total		138,088	3,101,879

category. For example, some standard satellite image datasets are developed to determine whether an image containing an airport or a stadium, or whether the satellite image is about a parking lot or a tennis court. Since this task is very similar to image recognition/classification in computer vision, models trained from daily datasets like ImageNet can be and are often transferred to this special application. As a convention, the size of the input image (e.g., 256×256) is often an even number in line with those in computer vision. This task requires the input to be several satellite images, each labeled with one category. Since multiple land covers may exist in the same scene as satellite images cover a large area, the number of identifying classes in scene classification is large, e.g., NWPU-RESISC45 (45 classes) (Cheng et al., 2017), PatternNet (38 classes) (Zhou et al., 2018), and BigEarthNet (43 classes) (Sumbul et al., 2019).

3. Study Area and Data

3.1. Study Area

In this study, we classify LCZs for fifteen cities in three economic regions of China. The three regions are the Greater Bay Area, the Shanghai-Hangzhou Metropolis and the Beijing-Tianjin Metropolis. Ten cities of the Greater Bay Area is covered in the study. The Shanghai-Hangzhou Metropolis is located in the Yangtze River Delta; the city of Shaoxing is also included in the study. Population and gross domestic product (GDP) in 2018 of these cities are shown in Table 2. The fifteen cities in this study cover a population of 138 million. Figures 3a and 3b show a 2018 night light imagery and the location of three study regions. These regions are very bright at night, showing their high-density population and vital economy.

3.1.1. The Greater Bay Area (GBA)

The Greater Bay Area locates at 23° N (Figure 3e) and consists of 10 cities, i.e., Guangzhou, Shenzhen, Dongguan, Foshan, Huizhou, Zhongshan, Jiangmen, Zhuhai, Macau, and Hong Kong, as shown in Table 2. This region covers over 60 million people. Guangzhou is the capital city of Guangdong province; Shenzhen is the third-largest city in China. The two cities consist of many urban villages (LCZ-3) that exhibit a unique urban structure in downtown. Other cities of this region are also important industrial centers (LCZ-8). For example, the main factory of one of the largest high-tech companies, Huawei, is located in Dongguan. Macau and Hong Kong are the two special administrative regions of China and the latter is one of the world's financial centers. They are with a high-density population and lack of available land, leading to their extremely high-density urban structure (LCZ-1). As some of the cities are small (e.g., Macau), it is difficult to classify LCZs within one city. Also, not all LCZs are presented in a single city. Thus, in this study, we combine the available samples in this region.

3.1.2. The Shanghai Metropolis

The second study region includes Shanghai, Hangzhou, and Shaoxing. Shanghai is the largest city in China. Hangzhou is the capital city of Zhejiang province, where the company Alibaba locates and the internet economy thrives. These three cities cover nearly 40 million population. Shanghai is a city lack of hills and mountains; the highest point is merely 98 *m* above sea level, leading to the lack of dense forests (LCZ-A).

3.1.3. The Beijing Metropolis

The third region covers the major part of Beijing and Tianjin. Beijing is the capital city in China and Tianjin is the largest industrial city of north China. The two cities consist of a population of 37 million. Due to industrial upgrades and environmental protection, heavy industrial factories (LCZ-10) were relocated from Beijing to other cities. Tianjin is similar to Shanghai and lack of dense forests (LCZ-A).

3.2. Sentinel Multispectral Imagery

Sentinel-2 multispectral data are used in this study. Sentinel is a new mission launched by the European Space Agency (ESA) to use satellite data to monitor land, ocean, and atmosphere of the Earth (Berger et al., 2012). The data consist of 13 spectral bands, including 4 bands with a ground sampling distance (GSD) of 10 *m*, 6 bands with 20 *m* GSD and 3 bands with 60 *m* GSD Drusch et al. (2012). The data are currently with the highest resolution among freely available satellite imagery. They are very suitable for large-scale LCZ mapping.

In the study, we only use the 10 *m* and 20 *m* images because images with 60 *m* GSD are not designed for classification (Drusch et al., 2012). Images with 20 *m* GSD are resampled to 10 *m* GSD using the nearest neighbor algorithm. Four sets of Sentinel multispectral data captured on 21 March 2018 were used for the Greater Bay Area. Two sets were collected in the Beijing-Tianjin Metropolis, which were captured on 20 April 2018 (Tianjin) and on 21 August 2018 (Beijing). Another two sets captured on 9 April 2018 (Hangzhou) and 19 April 2018 (Shanghai) were collected for the Shanghai-Hangzhou Metropolis.

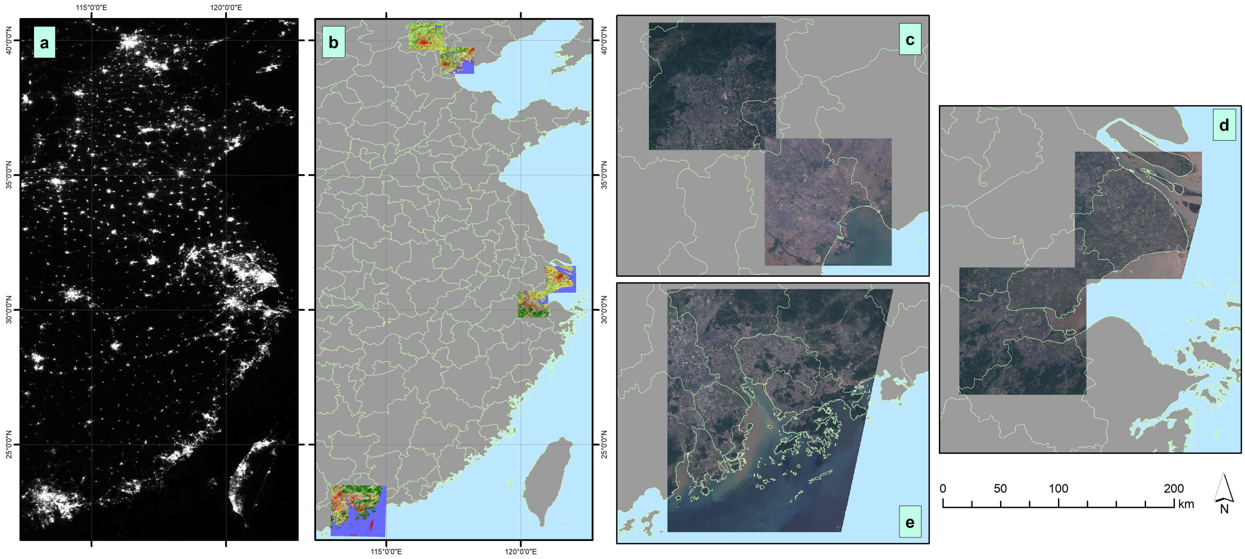


Figure 3: Location of three studied regions. (a) Annual VIIRS/DNB night lights (2018). (b) Studied regions shown with east China layout. (c) Beijing and Tianjin, referred as the Beijing metropolis. (d) Shanghai, Hangzhou and Shaoxing, referred as the Shanghai metropolis. (e) The Greater Bay Area (Guangzhou, Shenzhen, Dongguan, Foshan, Zhongshan, Jiangmen, Zhuhai, Macau, and Hong Kong).

3.3. Reference Data

Reference data were collected on Google Earth via visual interpretation. Some regions that were difficult to distinguish were checked in fields. For a standard Sentinel-2 image with a size of 10980×10980, we divided it into 100 subsets with a size of 1098×1098. The subsets were randomly chosen as training or testing parts. Labels located in the boundary (60 pixels) of each subset were discarded to ensure no overlapping between training and testing data. Then, the reference image was resampled to 100 *m* spatial resolution using the nearest neighbour method. Therefore, a pixel in the reference image correspond to a 10×10 scene in the satellite image.

Some photos are presented in Figure 1 to show the unique urban landscape in China. The compact high-rise urban landscape, which is very common in Hong Kong and other cities in China, is shown in the middle (Hong Kong). In the upper left (Guangzhou), you can find the urban villages (under reconstruction) in the front and the skyscrapers as the background. Two buildings in the front are close enough to “shake your hands”. The lower left side shows another photo taken in Guangzhou, where the compact mid-rise and compact high-rise buildings are in the same area. A large low-rise and/or heavy industry area with sufficient low plants are shown in the upper right. The landscape shown in the photo (taken in Zhuhai) is very common in China, illustrating that industrial areas can have sufficient vegetation cover. In the lower right, we show an open mid-rise photo taken in Zhuhai as well. The unique and complex landscape in China requires urban scientists to carefully assign LCZ categories.

4. Methods

4.1. The Proposed Network

In this study, we propose a network for LCZ mapping, namely the LCZNet (Figure 4). It includes an Inception module (Szegedy et al., 2017), several residual blocks (He et al., 2016a) and the Squeeze-and-Excitation blocks (Hu et al., 2018). As shown in Figure 4, at the tail of the LCZNet, it consists of a convolutional layer with multi-scale filters to extract spatial features. The extracted spatial features are then concatenated together to go through a SE-Residual block. This block has the ability to integrate channel-wise features by *squeeze* the less important features and *excite* the useful feature maps Hu et al. (2018). A total of six SE-Residual blocks are used in the LCZNet. For every two residual blocks, the number of convolutional filters doubles. At the head of the network, a global average pooling layer is applied and then the extracted high-level features are fed in a fully connected layer with the softmax function to give out the probability of each class.

4.2. Residual Learning

The idea of residual learning is to add a short cut connection so that the gradients learnt from backpropagation can convey efficiently, easing the training process (He et al., 2016b). It is defined as,

$$x_{l+1} = x_l + \mathcal{F}(\hat{f}(x_l), \mathcal{W}_l), \qquad (1)$$

where x_l and x_{l+1} are the input and output of the l-th layer, $\mathcal{W}_l$ is the parameters associated with the l-th layer, $\mathcal{F}$ is the residual function and $\hat{f}$ is an activation that only affects the $\mathcal{F}$ path (the non-skip part).

Table 3: Number of training and testing samples in the three studied regions. Each label represents a region of 10×10 in 10 *m* spatial resolution (see Appendix A for the sampling strategy).

No.	Class name	The Greater Bay Area		Shanghai Metropolis		Beijing Metropolis	
		Train	Test	Train	Test	Train	Test
LCZ-1	Compact high-rise	167	210	146	120	151	155
LCZ-2	Compact mid-rise	89	120	66	106	180	238
LCZ-3	Compact low-rise	285	442	71	113	380	317
LCZ-4	Open high-rise	302	390	231	191	297	579
LCZ-5	Open mid-rise	190	123	192	300	272	239
LCZ-6	Open low-rise	202	177	234	204	534	417
LCZ-7	Lightweight low-rise	152	179	29	36	76	64
LCZ-8	Large low-rise	333	437	333	260	366	557
LCZ-9	Sparsely built	31	76	46	24	105	186
LCZ-10	Heavy industry	106	254	231	98	152	186
LCZ-A	Dense trees	213	287	141	131	115	136
LCZ-B	Scattered trees	19	29	23	24	72	73
LCZ-C	Bush, scrub	22	23	7	7	7	4
LCZ-D	Low plants	109	128	72	64	222	262
LCZ-E	Bare rock or paved	205	196	213	165	118	179
LCZ-F	Bare soil or sand	166	543	103	95	561	1004
LCZ-G	Water	871	2111	888	761	1208	2868
	Total	3462	5725	3026	2699	4816	7464

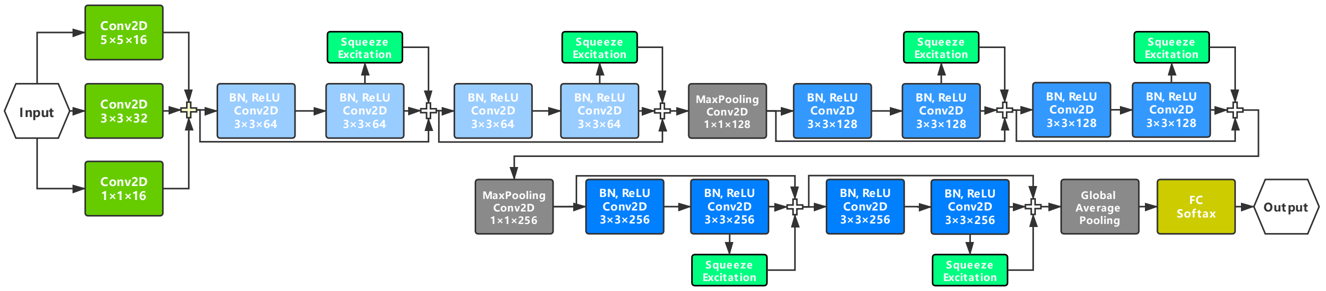


Figure 4: The network (LCZNet) used in this study.

4.3. *Squeeze-and-Excitation Blocks*

SENet was the winner of the last ImageNet competition (Hu et al., 2018), and the key to their success is the SE blocks. The SE block tries to enhance the channel relationship of feature maps learnt by CNN. As the features in later layers of a CNN tend to be abstract and class-specific, the SE block has the capability to perform dynamic channel-wise feature enhancement by assigning a weight to the feature maps. In the *squeeze* process, for a feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, we can obtain the channel-wise statistics $\mathbf{z} \in \mathbb{R}^{C}$ by using global average pooling:

$$z_c = \mathbf{F}_{sq}(\mathbf{X}_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j). \tag{2}$$

Here, $\mathbf{X}_c$ is a 2D feature map of channel c with a spatial dimension $H\times W$, $\mathbf{F}_{sq}$ is the *squeeze* process, $x_c(i,j)$ and z_c are the value of (x, y) and the channel-wise statistics of the c-th feature map. After the *squeeze* process, we obtain the channel-wise statistics $\mathbf{z}$

In the *excitation* process, we aim to fully capture the channel-wise statistics. A gating mechanism with the sigmoid function is used for this purpose,

$$s = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z})), \tag{3}$$

where σ is the sigmoid function, δ is the Rectified Linear Unit (ReLU), and $\mathbf{W}_1$ and $\mathbf{W}_2$ are two fully connected layers. The final output of *excitation* is by a channel-wise multiplication between a scalar s and the original feature map,

$$\mathbf{y}_c = \mathbf{F}_{scale}(\mathbf{x}_c, s_c). \tag{4}$$

4.4. *Network Training*

The experiments were conducted on Python 3.6 using Keras with TensorFlow backend. An Nvidia GTX 1060 6G GPU was used to accelerate the calculation. We initialized all convolutional layers with a Gaussian distribution of zero mean and 0.01 standard deviation. The AdaDelta (Zeiler, 2012) optimizer and a batch size of 16 were used in the training phase. The learning rate was set as 1.0 in the first 100 epochs and as 0.1 for another 30 epochs. If the training loss did not decrease for 5 epochs, the learning rate changed to 0.1 immediately or the training stopped when the training rate was already 0.1.

5. Results

5.1. *Classification of the Greater Bay Area*

The confusion matrix of the Greater Bay Area is shown in Figure 5. This classification achieves an OA of 90.84%, a

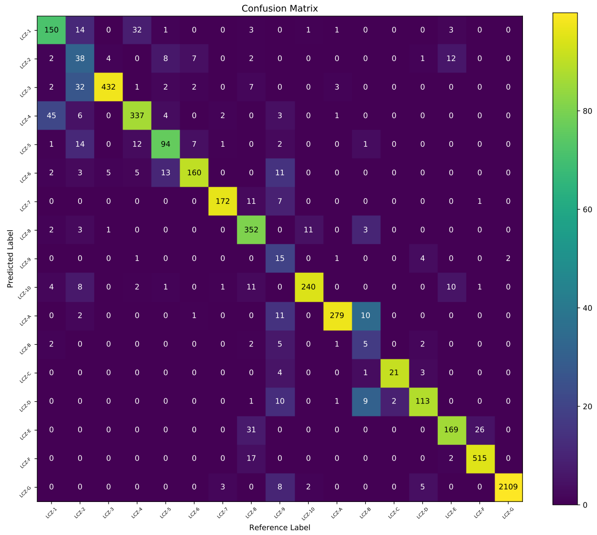


Figure 5: Confusion matrix of the Greater Bay Area. The background color represents the number of predicted labels divided by the number of reference labels of this class (%), e.g., producer accuracy for the correct class.

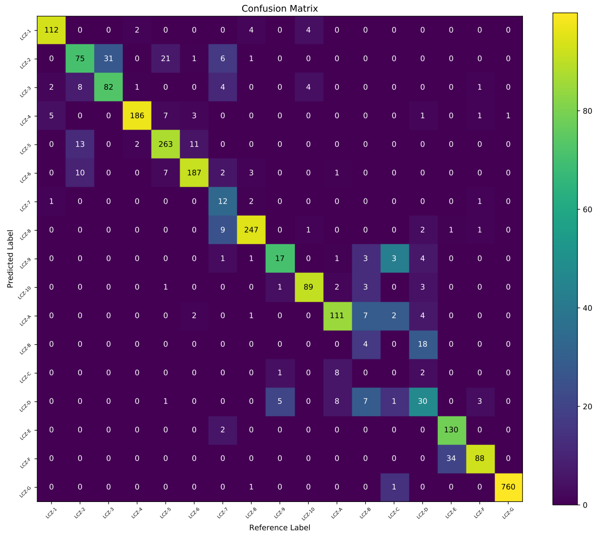


Figure 6: Confusion matrix of the Shanghai Metropolis. The background color represents the number of predicted labels divided by the number of reference labels of this class (%), e.g., producer accuracy for the correct class.

Kappa coefficient of 0.8893 and an AA of 77.64%. We can see that the OAs of LCZs are satisfactory (mostly greater than 80%) except for LCZ-9, LCZ-B, and LCZ-C (less than 50%). Urban classes are well classified, especially for LCZ-3 and LCZ-6. LCZ-3 are urban villages with special textural features, whereas LCZ-6 are mostly villas and single houses. Among the natural classes, LCZ-G water is no doubt the easiest to classify, followed by LCZ-A dense tree and LCZ-F bare soil. LCZ-B scattered trees are confused with LCZ-A dense trees.

5.2. *Classification of the Shanghai Metropolis*

Classification of the Shanghai Metropolis achieves an OA of 88.66%, a Kappa of 0.8702 and an AA of 71.93% (Figure 6). The results are similar to the GBA region. Urban classes are well classified, whereas natural classes LCZ-B scattered trees and LCZ-C bush are confused. LCZ-D low plants are also confused with other classes. Since the three natural classes are quite similar in the spectral domain, additional data should be utilized to distinguish them, e.g., high spatial resolution imagery, DEM, and/or LiDAR data.

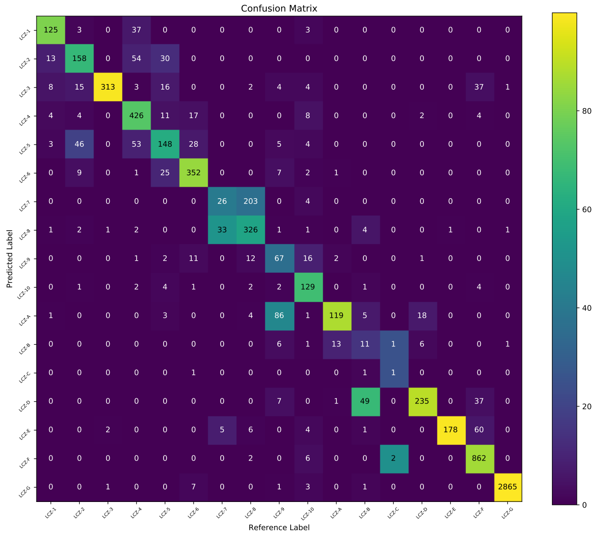


Figure 7: Confusion matrix of the Beijing-Tianjin metropolis. The background color represents the number of predicted labels divided by the number of reference labels of this class (%), e.g., producer accuracy for the correct class.

5.3. *Classification of the Beijing Metropolis*

For the Beijing Metropolis, we obtained an OA of 84.95%, a Kappa of 0.8454 and an AA of 68.98%. The confusion matrix is presented in Figure 7. For this region, the confusion between LCZ-7 lightweight low-rise and LCZ-8 large low-rise is significant. In this area, informal small factories are common, and their urban functions are similar to large factories. In a 10 *m* spatial resolution imagery, the small and large factories are easily confused as their heights and rooftops materials are similar. A possible solution to better classification may be using high spatial resolution imagery, in which the two types of factories should be easily distinguished based on their textures.

5.4. *LCZ Maps*

The LCZ maps with the corresponding satellite images are shown in Figure 8. In general, a reasonable urban structure is presented in the LCZ maps. It is easy to distinguish urban classes from natural classes and water. Airports are very easy to identify on the maps. Beijing shows a compact urban structure. The central areas are classified as LCZ-3 compact low-rise, while the entire urban regions are dominant by LCZ-2 compact mid-rise and LCZ-4 open high-rise. On the upper right side, we can see the Beijing Capital Airport is classified as LCZ-15. Some LCZ-1 compact high-rise are observed in the CBD region (Downtown East). For Tianjin, LCZ-1 compact high-rise locate in the central area. The asymmetry structure of airport is recognized as well.

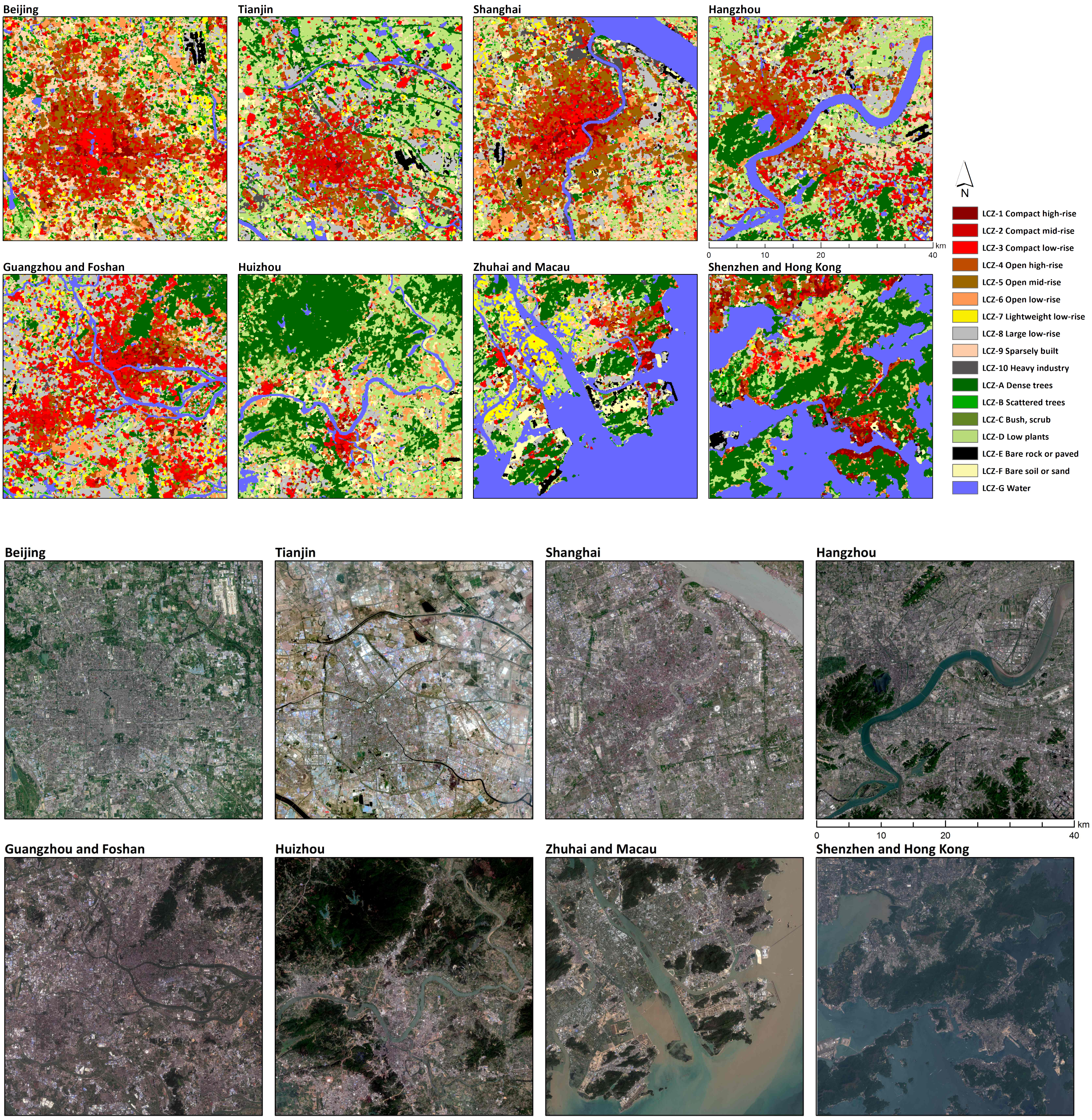


Figure 8: LCZ maps of the study area and the corresponding satellite images.

For Shanghai and Hangzhou, more LCZ-8 large low-rises are found in the suburban areas. In central Shanghai, LCZ-1 compact high-rise and LCZ-3 compact low-rise are the major classes, where LCZ-4 open high-rise and LCZ-5 open mid-rise are the dominant types in the surrounding urban areas, a result of 30-year urban expansion. The urban structure of Hangzhou is more compact compared with the aforementioned cities and is dominated by compact mid-rise and open high-rise.

As for the Greater Bay Area, more dense trees are shown in the maps due to the existence of hills and mountains. For Guangzhou and Foshan, the core urban area is smaller than Beijing. These two cities are connected to each other. The river dividing the city into two parts is recognized in the map. On the left side of Guangzhou and Foshan, a lot of LCZ-8 large low-rise are presented, a reflection of the strong industrial section in the region. Huizhou, a small city in the region, has the smallest urban area. The dominant LCZs for Huizhou are LCZ-C dense trees, LCZ-D low plants, and LCZ-F bare soil. The dominant LCZ type in peninsula of Macau is compact high-rise (on the right side of the LCZ maps with an airport on the water). From the satellite images, we can see a lot of “white” scattered between croplands in Zhuhai and Macau, which are the sheds of greenhouses. We finally check the LCZ maps in Shenzhen and Hong Kong. Compact high-rise is dominant in the Hong Kong

Table 4: Classification accuracy with different input channel. Band index can be found in Drusch et al. (2012).

Input band	OA(%)	Kappa×100	AA(%)
RGB (2,3,4)	87.04±0.60	84.48±0.78	75.29±0.66
All (2,3,4,5,6,7,8,8A,11,12)	88.61±0.48	86.58±0.57	77.17±0.61

island north, while water and dense trees are the major natural types. Low plants are observed on top of the mountains. On the upper side of the map shows Shenzhen city, where the major classes are LCZ-2 compact mid-rise and LCZ-4 open high-rise. Some LCZ-1 compact high-rises are found in the city as well.

To conclude, urban and natural classes are well distinguished in the classification map. The urban structure with rivers or mountains is well preserved. The obtained LCZ maps are visually satisfactory.

6. Discussion

6.1. Effect of Input Channels

Traditional pixel-wise land cover mapping relies on the rich spectral features to classify ground targets. As a result, remote sensing images often have tens of hundreds of channels. Adding more spectral bands should help the pixel-wise classification significantly, especially when the number of original bands is less than three (Wang et al., 2018b). On the other hand, daily RGB images only have 3 channels, but the deep learning model can still perform well on recognizing them. Remote sensing scene classification has no big difference with daily RGB image recognition. In Table 4, we show that even with only RGB channel, the network still has a very competitive performance compared to using all 10 bands of Sentinel data, indicating the scene classification nature of LCZ mapping.

6.2. Effect of the Scene Size

To find out the optimal scene size for LCZ mapping, we conducted an experiment with all available data with input sizes of 10×10, 16×16, 32×32, 48×48, 64×64, 80×80 and 96×96, corresponding to a minimum area of 100×100 m^2 and to a maximum of nearly 1,000×1,000 m^2.

The obtained results are presented in Figure 9. We achieve the best OA and Kappa using 48×48 inputs, whereas the optimal scene size in terms of AA varies among studied regions. For the GBA region, Shanghai Metropolis and Beijing Metropolis, the best AAs are achieved by using 48×48, 64×64 and 96×96 inputs, respectively. Beyond Beijing Metropolis, the difference among using the three input sizes is marginal. Although a 10×10 image size is the natural representation of a 100×100 m^2 region, it is too small and lack of context for LCZ mapping. A large scene size from 32×32 to 64×64 is found beneficial for LCZ mapping.

To analyze the influence of image size on individual class, we present the producer's accuracy (PA) and user's accuracy (UA) obtained with 10 × 10 inputs and 48 × 48 inputs in Figure 10. Compared with natural classes, urban classes benefited more from a large input size. For example, the PAs of LCZ-1, LCZ-2 and LCZ-6 increased from 57% to 84%, 26% to 63% and 75% to 84%, respectively. The UAs of these classes also increased from 57% to 77%, 44% to 60% and 70% to 88%, respectively. Since the environmental context of urban classes is more complex, a large input size helps the network to capture urban environmental features, which leads to its good performance.

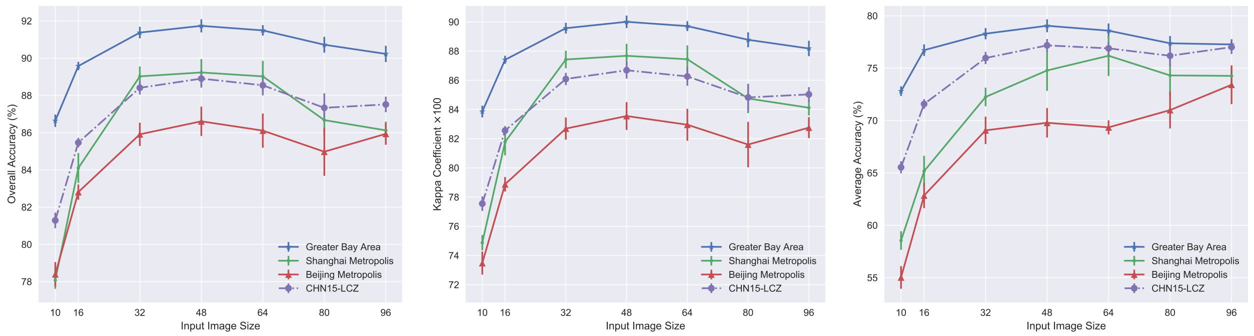
6.3. Comparison with the Competitors

In this section, we compare the classification among the proposed LCZNet, the WUDAPT method, random forest (with and without spatial features), and a recent CNN proposed by Rosentreter et al. (2020). To use spatial information in random forest, we extracted the standard deviation features from the input image. The results of these methods in terms of OA, Kappa and AA are presented in Figure 11.

We can see that the LCZNet significantly outperforms other methods, followed by the CNN proposed by Rosentreter et al. (2020). For the results obtained by random forest, they are all worse than CNN-based methods. The result obtained by the WUDAPT method (OA=67%) is the worst. The standard deviation features help random forest achieve a higher OA, about 5% higher compared to random forest with only mean spectral features. For AAs, the WUDAPT method only achieves an accuracy of 48%, whereas random forest with spatial features achieves 65% and our method 77%. The result is expected, as deep CNNs can effectively extract the urban environment context, i.e., the spatial information, without manual spatial filtering. The standard deviation filter is a simple tool to extract spatial features; more advanced methods like the morphological profiles and the Gabor filter are expected to boost the classification, but these filters are time-consuming because one has to grid-search the optimal parameters. The benefit of using CNNs is to automatically learn the proper features.

6.4. Analysis of Network Depth

Network depth is another important factor to classification. The proposed network is with six SE-Residual units. In Figure 12, we show the obtained OA, Kappa, and AA using one, two, four, and six units with different input sizes. The proposed network with six units achieves the highest classification in terms of the three evaluation metrics with input sizes greater than 32. When the inputs are small, a shallow network has a better performance. Another interesting phenomenon is that, a shallow network achieves the best classification with 32×32 inputs, whereas a deeper network achieves the best classification with 48×48 inputs. A deep network is capable to handle larger input data. This is because, the receptive field of a deep network is larger than a shallow network. For example, with only one SE-Residual unit, the receptive field of a network is 45×45 (5×3×3, determined by the size of convolutional layer in Figure 4). With the network going deep, its receptive field increases. On the other hand, although a network with two units can sense a region larger than the input size, which should be sufficient in the spatial domain, a deeper network is capable to extract high-level features, which is rather beneficial for classification.



(a)

(b)

(c)

Figure 9: Sensitivity of the scene size. The results are averaged from ten runs. (a) Overall accuracy (OA). (b) Kappa coefficient. (c) Average accuracy (AA).

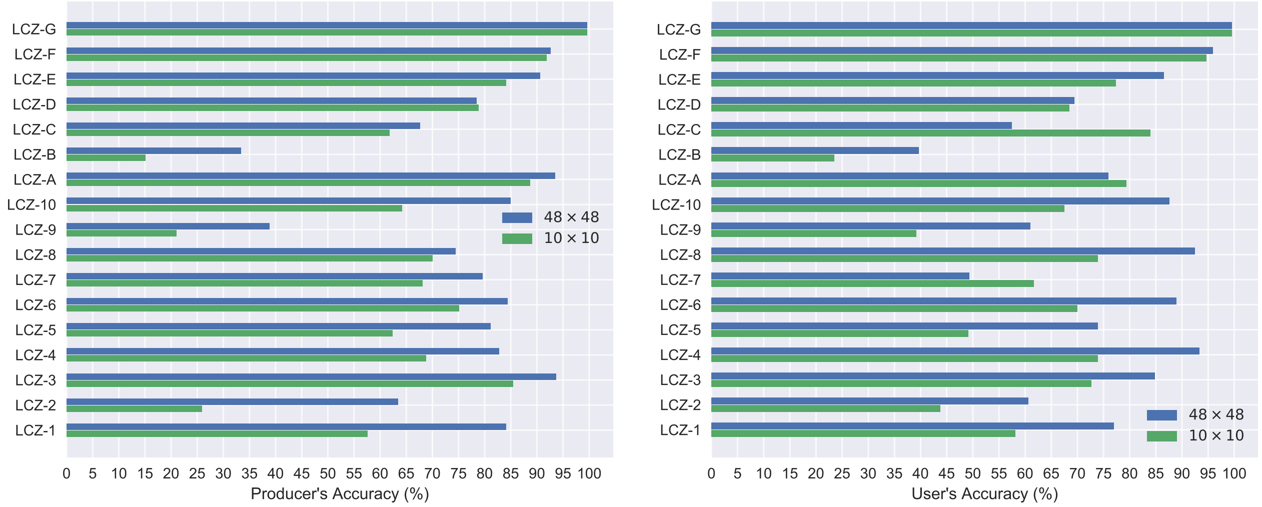
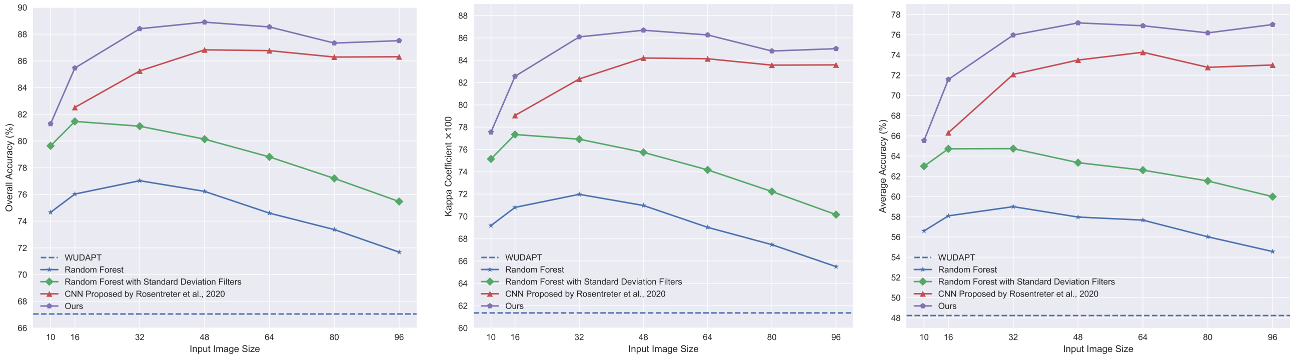


Figure 10: PA and UA obtained with 10 × 10 and 48 × 48 inputs.



(a) OA

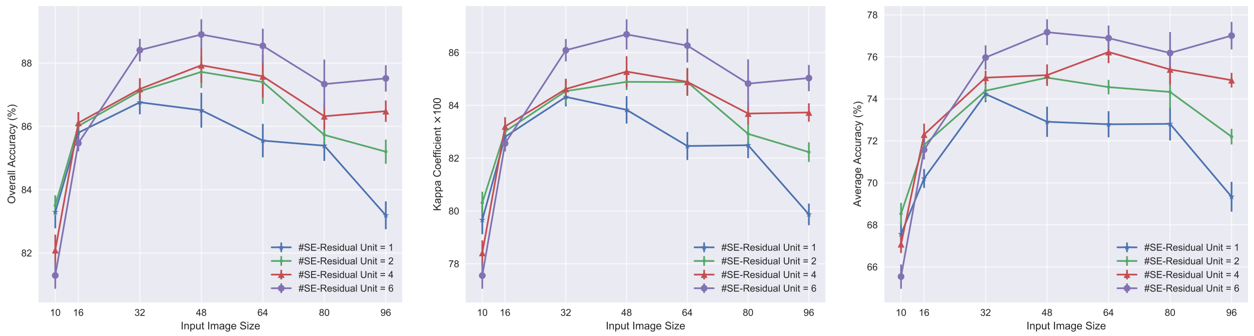
(b) Kappa

(c) AA

Figure 11: Comparison among the proposed LCZNet, random forest (with and without spatial features), the WUDAPT method, and a recent CNN proposed by Rosentreter et al. (2020). The results are averaged from ten runs. Note the CNN proposed by Rosentreter et al. (2020) is not applicable with 10×10 inputs.

6.5. *Effect of the SE Blocks*

In this section, we analyze the effect of SE blocks. The analysis is based on 64×64 input size, where the OA obtained with SE blocks is 89.44% and the OA without SE blocks is 88.76%. The F1 score of each class is shown in Figure 13. The t-SNE dimension reduction is a popular method for visualizing the latent features of a deep network (Maaten and Hinton, 2008; Zhong et al., 2017; Fang et al., 2020), where the sim-



(a) OA

(b) Kappa

(c) AA

Figure 12: Sensitivity of network depth. The results are averaged from ten runs.

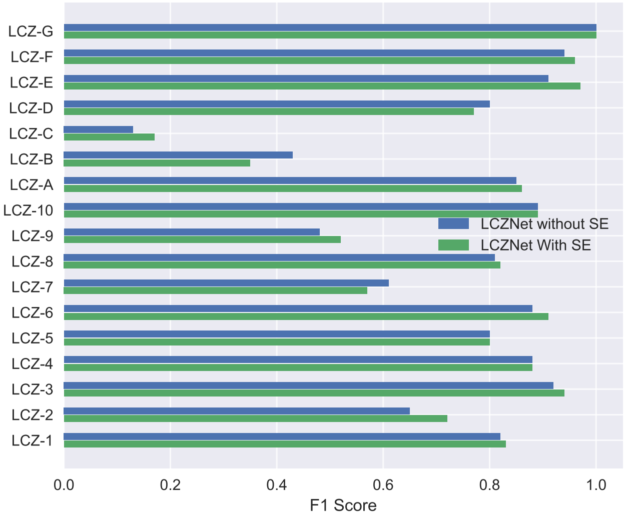


Figure 13: F1 score of each class using LCZNet w/ and w/o SE blocks.

ilar instances (same class) are closer than other instances. In our study, the used features were the output latent features before the fully connected layer, which is shown in Figure 14. As the margin is small, the F1 scores of these two networks are quite similar. With SE blocks, LCZ-2 and LCZ-E are better classified. As shown in Figure 14b with SE blocks, black triangles (LCZ-E) are more compact compared to those in Figure 14a without SE blocks. In both settings, red triangles (LCZ-2) are confused with other urban classes, and thus it is difficult to tell the difference from visualization. The compactness of yellow and green stars (LCZ-7 and LCZ-B), in which the network without SE blocks performs better, are similar from the visualization as well.

6.6. *Effect of the Training Set Size*

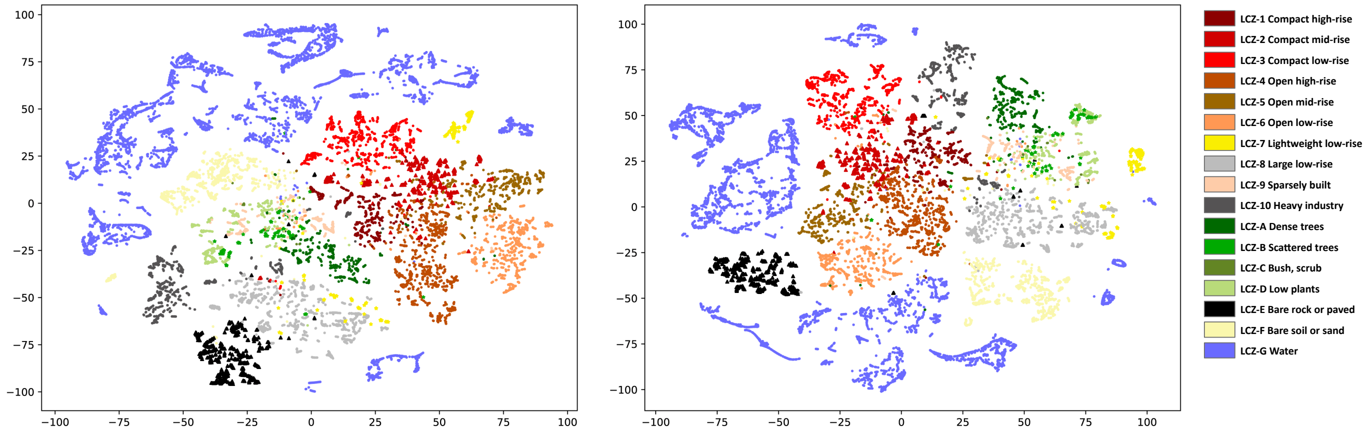
In this section, we explore the effect of training set size in LCZ mapping. The experiment is conducted on all available data with 20%, 40%, 60%, 80% and 100% of training samples with different input sizes. The results are presented in Figure 15. With the sample set enlarging, all evaluation metrics (OA, Kappa and AA) increase as expected. An interesting phenomenon is that, with limited training data, a large image representation leads to better classification compared to a small image. The gap in terms of OA between 10×10 and 64×64 inputs shrinks from 11.5% to 8.1%, whereas the gap in terms of AA between 10×10 and 48×48 inputs shrinks from 19.0% to 11.0%. A large image representation such as 32×32 and 48×48 is found more effective in LCZ mapping, especially when the training samples are limited. This is interesting, since we expected a large image representation would suffer from the curse of dimensionality and led to overfitting. But the experiment does not support the above assumption. On the other hand, it may be because the spatial features learned from a large image are more robust than those spectral-related features learned from a small image.

6.7. *Transferability*

Finally, we analyze the transferability of LCZ models. In the experiment, we trained the network with training samples from individual region and tested it in other regions. For comparative purposes, we also present the result trained on all data.

The transfer matrices of OA, Kappa and AA are presented in Figure 16. A transfer matrix is interpreted as follows. For a model trained on the dataset x (column), its classification accuracy on each dataset y (row) is presented in this column as (x,y). Take the OA as an example. The model trained on GBA obtained OAs of 89.01%, 76.77%, 56.83% and 71.81% on the GBA, Shanghai, Beijing and all the data (CHN15-LCZ). For the row denoted as Shanghai, the best classification is obtained by the model trained on all data, followed by the model trained on the same region and the model trained on GBA. We observe that the transferability of LCZ models is not satisfactory enough. The transferability of the Beijing Metropolis is the worst among the three regions. Some transferability of the Shanghai Metropolis (30° N, subtropical monsoon climate) and the Greater Bay Area (23° N, subtropical monsoon climate) can be observed since they are both located in south China with evergreen vegetation, while the Beijing Metropolis is located at a higher latitude (40° N, temperate monsoon climate).

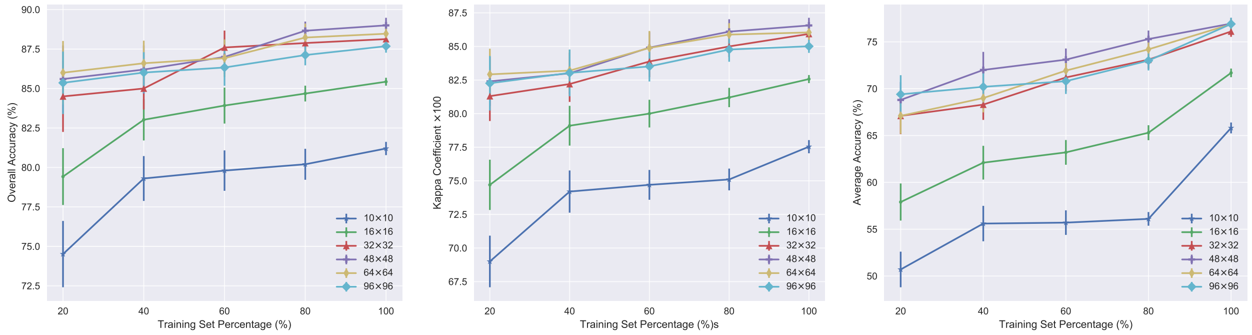
The transferability of LCZ models is a domain shift problem (Tuia et al., 2016), where each domain (city) has its unique



(a) w/o SE blocks

(b) w/ SE blocks

Figure 14: Visualization of the latent features (before the fully connected layer). Note red and black triangles are LCZ-2 and LCZ-E, where the LCZNet with SE blocks classified slightly better, whereas yellow and green stars are LCZ-7 and LCZ-B, where the LCZNet without SE blocks classified slightly better.

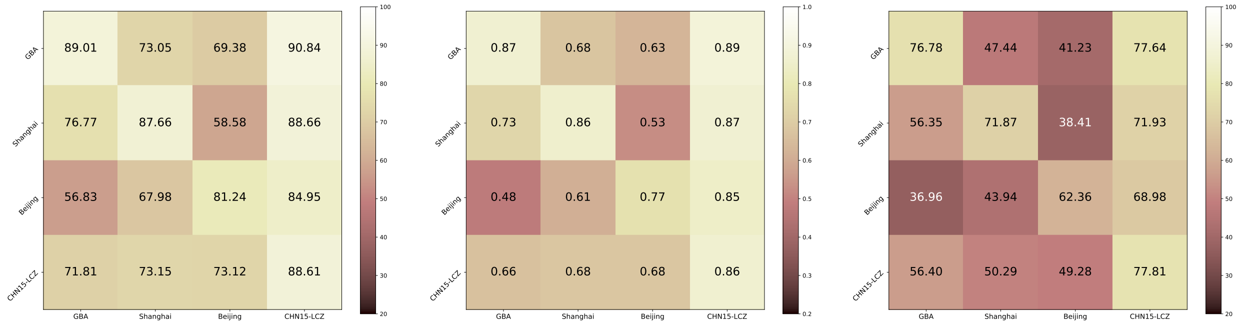


(a) OA

(b) Kappa

(c) AA

Figure 15: Effect of the training set size.



(a) OA

(b) Kappa

(c) AA

Figure 16: Transferability of the proposed model on local climate zones of different regions. For a model trained on dataset x (column x), its classification accuracy on dataset y (row y) is presented at (x,y).

features and environment, and therefore, results in difficulties in using training data from other regions. Although the topic of domain adaptation has been developed in recent years, still, no studies were conducted to tackle the domain shift problem in LCZ mapping rather than the naive model transfer. More advanced methods need to be developed or utilized for this application.

7. Conclusions

In this study, we conduct LCZ mapping on fifteen cities in three economic regions of China[1]. We first highlight the unique urban structure in China, which should be given special arrangements for LCZ mapping. Then, we review the existing literature about LCZ mapping. Most previous studies defined it as a pixel-based classification task and tried to distinguish them with spectral or multisource information. However, in the original paper, an LCZ is expected to have at least a diameter of 400-1,000 *m* so that it is possible to have an impact on the local climate. Therefore in this study, we define LCZ mapping as a remote sensing scene classification task. By doing so, the rich surrounding environment context is being considered. Since it is more reasonable to treat LCZ mapping as scene classification, we obtained very promising results on the fifteen cities (88.61%), nearly 20% higher than that obtained by the standard WUDAPT method in terms of OA. A supporting evidence that LCZ mapping should be a scene classification is that, we obtained a very competitive result (OA=87.04%) by using only RGB channel.

We also explored the suitable image size for LCZ mapping. An image size of 48×48 (480×480 m^2) was found as the optimum in terms of the three evaluation matrices, although the difference among 32×32, 48×48 and 64×64 is marginal. Larger image representation is more appropriate for LCZ mapping compared to a smaller one. The improvement of using large inputs is even more beneficial when the training samples are small.

The transferability of various models from different regions is also investigated. For the three economic regions in China, namely the Greater Bay Area (23° N, subtropical monsoon climate), the Shanghai Metropolis (30° N, subtropical monsoon climate) and the Beijing Metropolis (40° N, temperate monsoon climate), the models' transferability of the first two regions is better than the last one. The combined use of all the available data achieved the best results for all three regions. Therefore, it is recommended to combine all the available data for large scale LCZ mapping. When applying a model from a different region, researchers should compare the economic and natural environment of the two regions. However, the domain shift problem in LCZ mapping is significant. The current transfer in the literature of LCZ mapping is only a naive solution. More advanced domain adaptation techniques are expected to be applied in this application.

The ambiguity of LCZs is a major issue when generating regional and global LCZ maps. A future direction for LCZ mapping is to generate multilabel LCZ maps, where a region can be classified as several LCZ classes, i.e. the LCZ subclasses. Future studies should be conducted to tackle this issue.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61601522). The authors would like to thank the developers behind the freely accessible software used in this study, e.g., Keras and TensorFlow. They would also like to thank ESA for freely providing the Sentinel data to the public.

Appendix A. The Sampling Strategy

Figure A.17 shows an example of the sampling strategy. The reference data is with 100 *m* spatial resolution and the satellite data is with 10 *m*. Thus, when classifying the entire satellite image, the moving window moves with a stride/skip of 10. If the input size is larger than 10×10, we extended the input patch in all directions equally (e.g. 3 pixels for 16×16 and 27 pixels for 64×64). The light blue area with a text "A" (Figure A.17a) shows the current scene and the input patch (dark blue) in classification when using 64×64 input size. The surrounding dark blue region is the environment context with a 64×64 input size. After classifying this scene, we jump to the next scene (a skip of 10 pixels) and continue classification. The scene with a text "B" (Figure A.17b red and dark red region) is the 5th scene after scene A, where the overlapped areas between A and B (Figure A.17c bright blue region) are the shared context for LCZ mapping.

References

Audebert, N., Le Saux, B., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks, in: Asian conference on computer vision, Springer. pp. 180–196.

Azimi, S.M., Henry, C., Sommer, L., Schumann, A., Vig, E., 2019. Skyscapes fine-grained semantic understanding of aerial scenes, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 7393–7403.

Bechtel, B., Alexander, P., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L., Stewart, I., 2015. Mapping local climate zones for a worldwide database of the form and function of cities. ISPRS International Journal of Geo-Information 4, 199–219.

Bechtel, B., Daneke, C., 2012. Classification of local climate zones based on multiple earth observation data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 5, 1191–1202.

Benediktsson, J.A., Palmason, J.A., Sveinsson, J.R., 2005. Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Transactions on Geoscience and Remote Sensing 43, 480–491.

Berger, M., Moreno, J., Johannessen, J.A., Levelt, P.F., Hanssen, R.F., 2012. Esa's sentinel missions in support of earth system science. Remote Sensing of Environment 120, 84–90.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS journal of photogrammetry and remote sensing 65, 2–16.

Cai, M., Ren, C., Xu, Y., Dai, W., Wang, X.M., 2016. Local climate zone study for sustainable megacities development by using improved wudapt methodology–a case study in guangzhou. Procedia Environmental Sciences 36, 82–89.

Campbell, B.M., Hansen, J., Rioux, J., Stirling, C.M., Twomlow, S., et al., 2018. Urgent action to combat climate change and its impacts (sdg 13): transforming agriculture and food systems. Current opinion in environmental sustainability 34, 13–20.

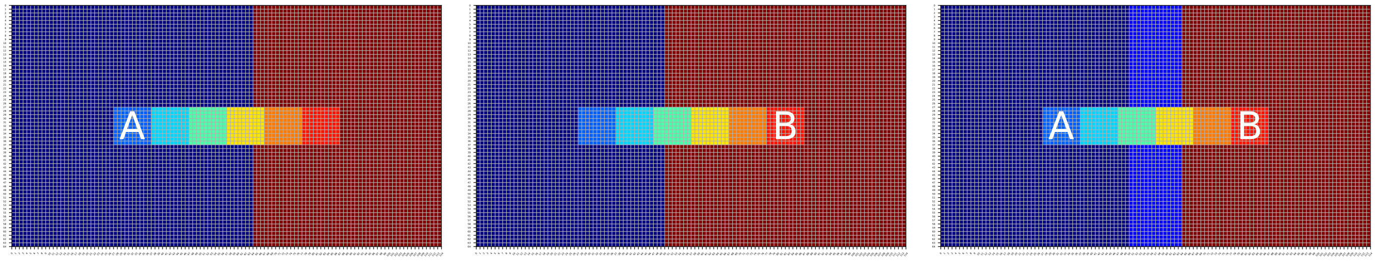
Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE 105, 1865–1883.

Demuzere, M., Bechtel, B., Middel, A., Mills, G., 2019a. Mapping europe into local climate zones. PloS one 14, e0214474.

Demuzere, M., Bechtel, B., Mills, G., 2019b. Global transferability of local climate zone models. Urban climate 27, 46–63.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al., 2012. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. Remote sensing of Environment 120, 25–36.

[1] Data and code for this study will be available at https://sjliu.me/lcz



(a)

(b)

(c)

Figure A.17: Illustration of the sampling strategy as scene classification.

Fang, B., Li, Y., Zhang, H., Chan, J.C.W., 2020. Collaborative learning of lightweight convolutional neural network and deep clustering for hyperspectral image semi-supervised classification with limited training samples. ISPRS Journal of Photogrammetry and Remote Sensing 161, 164–178.

Güneralp, B., Zhou, Y., Ürge-Vorsatz, D., Gupta, M., Yu, S., Patel, P.L., Fragkias, M., Li, X., Seto, K.C., 2017. Global scenarios of urban density and its impacts on building energy use through 2050. Proceedings of the National Academy of Sciences 114, 8945–8950.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks, in: European conference on computer vision, Springer. pp. 630–645.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.

Huang, X., Wang, Y., 2019. Investigating the effects of 3d urban morphology on the surface urban heat island effect in urban functional zones by using high-resolution remote sensing data: A case study of wuhan, central china. ISPRS Journal of Photogrammetry and Remote Sensing 152, 119–131.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on Geoscience and Remote Sensing 57, 574–586.

Kampffmeyer, M., Salberg, A.B., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 1–9.

Kotharkar, R., Bagade, A., 2018. Local climate zone classification for indian cities: A case study of nagpur. Urban climate 24, 369–392.

Kuffer, M., Pfeffer, K., Sliuzas, R., 2016. Slums from space—15 years of slum mapping using remote sensing. Remote Sensing 8, 455.

Lau, K.K.L., Chung, S.C., Ren, C., 2019. Outdoor thermal comfort in different urban settings of sub-tropical high-density cities: An approach of adopting local climate zone (lcz) classification. Building and Environment 154, 227–238.

Li, W., Chen, C., Su, H., Du, Q., 2015. Local binary patterns and extreme learning machine for hyperspectral imagery classification. IEEE Transactions on Geoscience and Remote Sensing 53, 3681–3693.

Li, W., Du, Q., 2014. Gabor-filtering-based nearest regularized subspace for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7, 1012–1022.

Liu, S., Qi, Z., Li, X., Yeh, A.G.O., 2019. Integration of convolutional neural networks and object-based post-classification refinement for land use and land cover mapping with optical and sar data. Remote Sensing 11, 690.

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

Luo, J., Zhang, X., Wu, Y., Shen, J., Shen, L., Xing, X., 2018. Urban land expansion and the floating population in china: For production or for living? Cities 74, 219–228.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS journal of photogrammetry and remote sensing 152, 166–177.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9, 2579–2605.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing 55, 645–657.

Marmanis, D., Wegner, J.D., Galliani, S., Schindler, K., Datcu, M., Stilla, U., 2016. Semantic segmentation of aerial images with an ensemble of cnns. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 3, 473.

Masó, J., Serral, I., Domingo-Marimon, C., Zabala, A., 2019. Earth observations for sustainable development goals monitoring based on essential variables and driver-pressure-state-impact-response indicators. International Journal of Digital Earth , 1–19.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2020. Image segmentation using deep learning: A survey. arXiv preprint arXiv:2001.05566 .

Nations, U., 2015. World population prospects: The 2015 revision. United Nations Econ Soc Aff 33, 1–66.

Perera, N., Emmanuel, R., 2018. A "local climate zone" based approach to urban planning in colombo, sri lanka. Urban climate 23, 188–203.

Qiu, C., Mou, L., Schmitt, M., Zhu, X.X., 2019. Local climate zone-based urban land cover classification from multi-seasonal sentinel-2 images with a recurrent residual network. ISPRS Journal of Photogrammetry and Remote Sensing 154, 151–162.

Qiu, C., Schmitt, M., Mou, L., Ghamisi, P., Zhu, X., 2018. Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets. Remote Sensing 10, 1572.

Rafique, M.U., Jacobs, N., 2019. Weakly supervised building segmentation from aerial images, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 3955–3958.

Risojević, V., Babić, Z., 2012. Fusion of global and local descriptors for remote sensing image classification. IEEE Geoscience and Remote Sensing Letters 10, 836–840.

Rosentreter, J., Hagensieker, R., Waske, B., 2020. Towards large-scale mapping of local climate zones using multitemporal sentinel 2 data and convolutional neural networks. Remote Sensing of Environment 237, 111472.

Sharma, A., Liu, X., Yang, X., Shi, D., 2017. A patch-based convolutional neural network for remote sensing image classification. Neural Networks 95, 19–28.

Shi, Q., Liu, X., Li, X., 2017. Road detection from remote sensing images by generative adversarial networks. IEEE access 6, 25486–25494.

Song, C., Huang, Y., Ouyang, W., Wang, L., 2019. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3136–3145.

Stewart, I.D., Oke, T.R., 2012. Local climate zones for urban temperature studies. Bulletin of the American Meteorological Society 93, 1879–1900.

Sumbul, G., Charfuelan, M., Demir, B., Markl, V., 2019. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 5901–5904.

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

Thenkabail, P.S., Schull, M., Turral, H., 2005. Ganges and indus river basin land use/land cover (lulc) and irrigated area mapping using continuous streams of modis data. Remote Sensing of Environment 95, 317–341.

Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE geoscience and remote sensing magazine 4, 41–57.

Wang, C., Middel, A., Myint, S.W., Kaplan, S., Brazel, A.J., Lukasczyk, J., 2018a. Assessing local climate zones in arid cities: The case of phoenix, arizona and las vegas, nevada. ISPRS journal of photogrammetry and remote sensing 141, 59–71.

Wang, Q., Zhang, F., Li, X., 2018b. Optimal clustering framework for hyperspectral band selection. IEEE Transactions on Geoscience and Remote Sensing 56, 5910–5922.

Wang, S., Chen, W., Xie, S.M., Azzari, G., Lobell, D.B., 2020. Weakly supervised deep learning for segmentation of remote sensing imagery. Remote Sensing 12, 207.

Wu, F., 2016. Housing in chinese urban villages: The dwellers, conditions and tenancy informality. Housing Studies 31, 852–870.

Xia, G.S., 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, 3965–3981.

Xu, Y., Ma, F., Meng, D., Ren, C., Leung, Y., 2017a. A co-training approach to the classification of local climate zones with multi-source data, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE. pp. 1209–1212.

Xu, Y., Ren, C., Cai, M., Edward, N.Y.Y., Wu, T., 2017b. Classification of local climate zones using aster and landsat data for high-density cities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10, 3397–3405.

Yang, J., Guo, J., Yue, H., Liu, Z., Hu, H., Li, K., 2019. Cdnet: Cnn-based cloud detection for remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing 57, 6195–6211.

Yokoya, N., Ghamisi, P., Xia, J., Sukhanov, S., Heremans, R., Tankoyeu, I., Bechtel, B., Le Saux, B., Moser, G., Tuia, D., 2018. Open data for global multimodal land use classification: Outcome of the 2017 ieee grss data fusion contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11, 1363–1377.

Zeiler, M.D., 2012. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 .

Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geoscience and Remote Sensing Magazine 4, 22–40.

Zhang, X., Jin, J., Lan, Z., Li, C., Fan, M., Wang, Y., Yu, X., Zhang, Y., 2020. Icenet: A semantic segmentation deep network for river ice by fusing positional and channel-wise attentive features. Remote Sensing 12, 221.

Zheng, Y., Ren, C., Xu, Y., Wang, R., Ho, J., Lau, K., Ng, E., 2018. Gis-based mapping of local climate zone in the high-density city of hong kong. Urban climate 24, 419–448.

Zhong, Z., Li, J., Ma, L., Jiang, H., Zhao, H., 2017. Deep residual networks for hyperspectral image classification, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE. pp. 1824–1827.

Zhou, W., Newsam, S., Li, C., Shao, Z., 2018. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS journal of photogrammetry and remote sensing 145, 197–209.

Zhu, X., Hu, J and, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Häberle, M., Hua, Y., Huang, R., et al., 2019a. So2sat lcz42: A benchmark dataset for global local climate zones classification. IEEE Geosci. Remote Sens. Mag.(submitted for publication) .

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine 5, 8–36.

Zhu, Z., Zhou, Y., Seto, K.C., Stokes, E.C., Deng, C., Pickett, S.T., Taubenböck, H., 2019b. Understanding an urbanizing planet: Strategic directions for remote sensing. Remote sensing of environment 228, 164–182.