ISPRS Journal of Photogrammetry and Remote Sensing 181 (2021) 345–366

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Local climate zone classification using a multi-scale, multi-level attention network

Minho Kim [a], Doyoung Jeong [a], Yongil Kim [a,*]

[a] Department of Civil and Environmental Engineering, Seoul National University, South Korea

ARTICLE INFO

Keywords:
Local climate zone
Scene classification
Multi-scale multi-level attention
Convolutional block attention module
Context aggregation
Remote sensing
Urban climate

ABSTRACT

Local Climate Zones (LCZ) offer a climate-aware and standardized classification scheme composed of 17 urban and natural landscape classes. Recent deep learning-based LCZ classification studies have adopted a scene classification approach with computer vision-inspired models. In light of these advancements, this study introduces a multi-scale, multi-level attention network (MSMLA-Net) for deep learning-based LCZ classification. MSMLA-Net integrates a multi-scale (MS) module to generate multi-scale features from the input data and a novel multi-level attention (MLA) module as a branch unit from the model's main ResNet backbone. MLA uses the convolutional block attention module (CBAM) at multiple stages to generate multi-level spatially and spectrally enhanced features for context aggregation. This study presents comprehensive model-based experiments on model depth, the individual and combined influence of MS and MLA modules, and the addition of attention mechanisms. Furthermore, data-based experiments are conducted to determine optimal Sentinel-2 spectral bands, while OpenStreetMap (OSM) building data, ALOS World 3D DSM height information, and a national land cover map are included as ancillary bands. LCZ classification tests are conducted on six major cities in South Korea. With regards to model-based performance, MSMLA-Net was built using a modified SE-ResNet50 backbone (MSMLA-50) and obtained the best classification results using 48 by 48-pixel input patches with a combination of all Sentinel-2 and ancillary bands, outperforming three state-of-the-art LCZ classification models. In particular, only MSMLA-50 reached over 70% built-up overall accuracy and maintained high accuracy even when tested on completely unseen areas in a "citywise" leave-one-out sampling strategy. For data-based results, the combination of Sentinel-2 Red Edge bands were helpful, mainly due to the greater number of available bands. Training only ancillary data generated up to 75.0% for built-up overall accuracy, albeit at the cost of low natural overall accuracy. Using all available bands and ancillary data produced the best results, but a combination of only OSM and Sentinel-2 bands also generated comparable accuracy. Ultimately, the proposed MSMLA-Net bridges advanced computer vision techniques such as attention mechanisms and multi-level context aggregation to improve LCZ classification. The MS and MLA modules can be applied on different backbones, sophisticated sampling schemes, and data inputs to flexibly achieve improved classification results.

1. Introduction

More than half of the world's population currently live in urban areas and the United Nations (UN) projects this estimate to escalate to 65% by 2050 – a striking number when considering that only around 1–3% of the Earth is defined as built-up area (DESA, 2018; Mills, 2007). As global urbanization rates rise, artificial paving over natural landscapes increases the absorption of solar radiation, while reducing the sky view factor and emitting heat. The resulting thermal variations can impose adverse implications such as air pollution and harmful thermal-related risks like extreme heat waves (Tan et al., 2010; Li and Bou-Zeid, 2013; Heaviside et al., 2017), which target the majority of the global population currently living in cities. The impact of these risks is expected to only intensify with the continuous exacerbation of climate change effects (Li and Bou-Zeid, 2013), underscoring the need to monitor and better understand urban climates.

There are three main limitations regarding conventional urban-related data. First, the growth of modern cities into urban agglomerations has rendered urban extents nearly obsolete. A systematic and standardized method is required to improve the classification of urban

* Corresponding author at: Department of Civil and Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea.
E-mail address: yik@snu.ac.kr (Y. Kim).

https://doi.org/10.1016/j.isprsjprs.2021.09.015
Received 15 January 2021; Received in revised form 16 August 2021; Accepted 17 September 2021

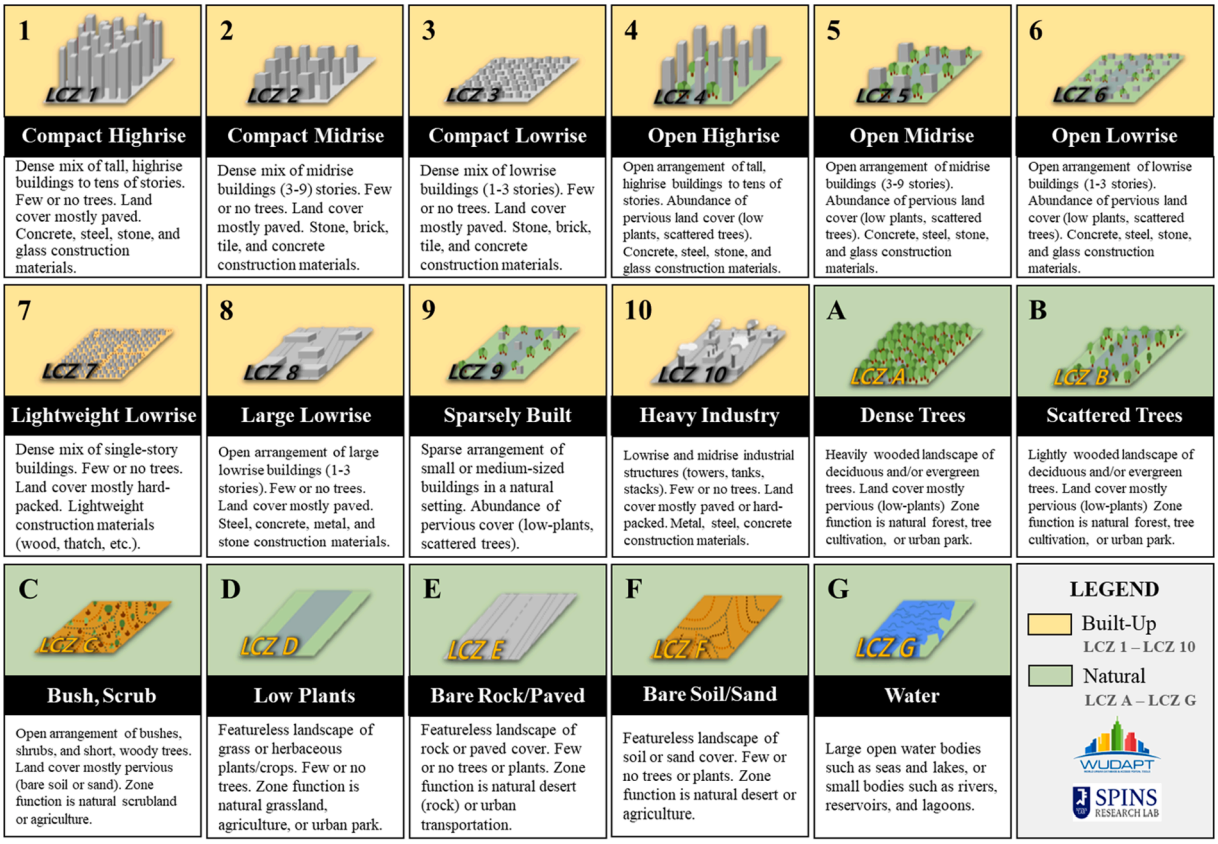

Fig. 1. LCZ classification scheme with ten urban (yellow) and seven natural (green) classes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

climates from the urban–rural dichotomy. Second, many existing urban datasets typically have one class related to built-up or impervious areas. For example, high-resolution urban Geographical Information System (GIS) layers such as the Global Urban Footprint (GUF) or the Global Artificial Imperviousness Area (GAIA) only contain one category representing urban regions. Third, traditional land cover datasets, such as land use/land cover (LULC) maps, only represent urban form and function without consideration of key climatic properties. These datasets may be used to distinguish urban surface cover but are not representative of thermal and climate-related characteristics. Hence, existing urban datasets and LULC maps lack the information on climatic properties needed to characterize micro-climates and to obtain a comprehensive understanding of the urban climate.

To address these limitations, Local Climate Zone (LCZ) scheme was introduced as a climate-based classification scheme for urban temperature studies (Stewart and Oke, 2012; Stewart et al., 2014). The LCZ scheme is composed of 17 urban and natural classes based on uniform regions of surface cover (pervious and impervious), surface structure (height and density), building material (heavy and light-weight), and anthropogenic activity (heat output) that span hundreds of meters to kilometers at horizontal scale as shown in Fig. 1 (Stewart and Oke, 2012). Early LCZ classification studies applied point or grid-based data obtained from simulations of climatic properties (Stewart and Oke, 2012; Oke, 2004; Stewart and Oke, 2009; Stewart and Oke, 2010). With the proliferation of remote sensing (RS) and GIS data, studies employed supervised LCZ classification methods with machine learning classifiers (Bechtel, 2011; Bechtel and Daneke, 2012; Bechtel et al., 2015; Bechtel et al., 2016). In particular, the World Urban Database and Access Portal Tools (WUDAPT) is a community-based platform for users to create LCZ maps at 100 m resolution using Landsat 8 images and the random forest (RF) classifier (Bechtel et al., 2019; Mills et al., 2015). Around 100 global cities have been mapped, but the resulting LCZ map data suffered from moderate classification accuracy (average overall accuracy of 74.5%) (Bechtel et al., 2015; Bechtel et al., 2019; Zhu et al., 2020; Yoo et al., 2019).

Recent studies have shown that convolutional neural network (CNN) based models demonstrate higher classification accuracy and better generalization ability over machine learning classifiers like RF (Qiu et al., 2018; Rosentreter et al., 2020; Yoo et al., 2019; Zhu et al., 2020). Further, deep learning-based LCZ classification studies have adopted a scene-based approach by applying fixed-size image patches rather than a pixel-based approach to utilize contextual information in neighboring pixels. For instance, Yoo et al. (2019) used Landsat-8 images from the 2017 IEEE GRSS data contest to compare pixel and patch-based CNN models with RF. Even when RF utilized hand-crafted features, the patch-based CNN outperformed CNN for all LCZ classes and tested cities. Rosentreter et al. (2020) adopted a similar model architecture to (Yoo et al., 2019) inspired by VGGNet (Simonyan and Zisserman, 2014) and confirmed that CNN generated higher overall accuracy over RF when mapping LCZ classes in German cities. Liu and Shi (2020) proposed LCZNet which was built using residual learning (He et al., 2016), a basic inception module, and six squeeze-excitation (SE) block modules (Hu et al., 2018) to create LCZ maps for cities in metropolitan China. LCZNet demonstrated the importance of input data patch sizes to provide sufficient contextual information and highlighted the significance of using SE block attention with deep models. Qiu et al. (2020) presented a benchmark model for LCZ classification called Sen2LCZ which benefitted from using double pooling layers and a multi-level feature fusion layer (Qiu et al., 2020). Sen2LCZ displayed the significance of fusing multi-level feature maps together for LCZ classification. The study also confirmed optimal model depths and widths, while preset computer vision-based models were determined to be unsuitable for LCZ scene classification.

In general, larger input image patches can exploit surrounding local

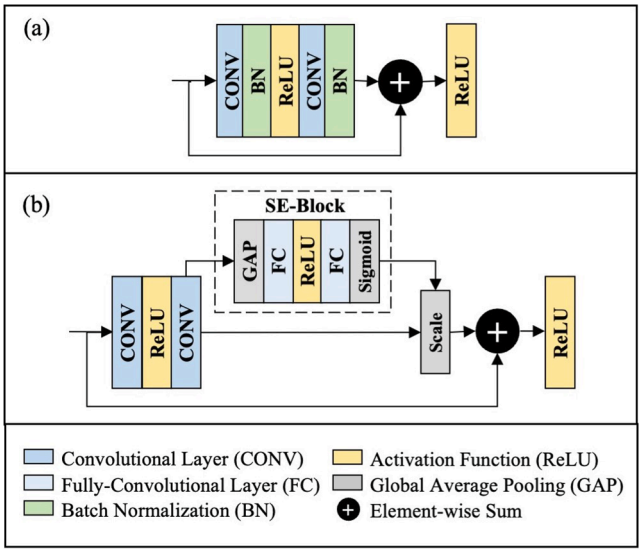

Fig. 2. (a) ResBlock and (b) SE-ResBlock modules.

pixel information to improve classification accuracy (Zhu et al., 2020; Rosentreter et al., 2020; Liu and Shi, 2020), whereas pixel-based classification methods are notorious for salt-and-pepper noise. Local contextual information is essential to capture the heterogeneous urban and surrounding areas (Zhang et al., 2019; Verdonck et al., 2017). Most related works applied 32 by 32-pixel image patches at 10 m spatial resolution (Zhu et al., 2020; Rosentreter et al., 2020). Liu and Shi (Liu and Shi, 2020) compared numerous patch sizes and concluded that larger patch sizes between 32 by 32 to 64 by 64 pixels (corresponding to 320 m by 320 m and 640 m by 640 m patches, respectively) were most helpful in improving classification accuracy, while exceedingly larger sizes up to 96 by 96-pixel patches lowered accuracy (Liu and Shi, 2020).

To fully utilize contextual information in the input patches, multi-scale/multi-branch modules and attention mechanisms can be used in a scene classification model. Previous studies have shown promising improvements when using multi-scale layers (Liu and Shi, 2020), attention mechanisms (Zhu et al., 2020; Qiu et al., 2020; Liu and Shi, 2020), and multi-level feature aggregation methods (Qiu et al., 2020). In light of these recent developments, this study proposes a novel model architecture called the multi-scale, multi-level attention network (MSMLA-Net). The model integrates a simple multi-scale (MS) module followed by a multi-level attention (MLA) module which branches from the main backbone to utilize the convolution block attention module (CBAM) at multiple levels (Woo et al., 2018). The main role of MLA is to combine multi-level spatially and spectrally enhanced local and global information for context aggregation. This study demonstrates the effectiveness of MSMLA-Net through a comprehensive series of model and data-based experiments. For this study, LCZ classification is carried out on six major cities in South Korea using 10 m resolution S2 images and resampled ancillary data to produce 100 m resolution LCZ maps. From a model-based perspective, this study explores the model depth, addition of attention mechanisms, and the individual and combined influence of MS and MLA modules. From a data-based viewpoint, MSMLA-Net is fed with numerous Sentinel-2 spectral band combinations, OpenStreetMap (OSM) building data, ALOS World 3D (AW3D30) DSM height information, and a national land cover map to determine optimal input band combinations. Lastly, a city-wise classification is conducted to evaluate model robustness and performance.

2. CNN models and attention modules

2.1. Convolutional neural networks

CNNs for image classification are generally composed of convolutional layers and fully-connected (FC) layers as well as activation functions, batch normalization, and pooling layers. Filters are used in the convolutional layer to extract feature maps from the input data via the convolution operation, processing the pixel values in the image (spatial domain) and across numerous bands (spectral domain). The spatial local connectivity of an input image along spectral bands is fused within the local receptive field to produce high-level image representations and extract useful features.

2.2. Residual learning

To help ease the training of deeper neural networks, residual learning was introduced by adding shortcut connections using identity mapping layers (He et al., 2016). Residual learning can be expressed by the following formula:

$$x_{m+1} = F(\varphi(x_m), w_m) + x_m \tag{1}$$

where x_m and x_{m+1} correspond to the input and output vectors of the considered layer m, w_m denotes the parameters associated with layer m, φ is the activation function, and F is the residual function. $F + x_m$ denotes the shortcut connection and element-wise addition. Since identity shortcut connections do not add extra complexity to the model, the entire network can be deepened and trained more efficiently via back-propagation (He et al., 2016; Szegedy et al., 2017).

2.3. Squeeze excitation networks

Attention-based modules are added onto CNNs to learn more meaningful information from the image's features or patterns of interest. In general, attention modules are composed of a 2D convolutional layer, multi-layer perceptron (MLP), and a sigmoid function to generate a refined attention map. Despite the use of skip connections to facilitate deep layers, ResNets and other computer vision-based models have shown poor performance for LCZ classification (Qiu et al., 2020), indicating the need for further modifications aside from naïve connections of deep convolutional layers. One solution is adding SE-blocks to help model training and propagate information through deep layers. The squeeze phase is needed for global information embedding and the excitation phase uses this embedding to produce per-channel weights to apply to feature maps (Hu et al., 2018). Given an intermediate feature map F with dimensions H, W, C for height, width, and bands, the squeeze and excitation functions can be expressed as:

$$F_{sq}(F_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c(i,j) \tag{2}$$

$$F_{ex}\left(F_{sq}(Fc), F_c\right) = \sigma\{FC_2\delta\left(FC_1\left[F_{sq}(F_c)\right]\right)\} \tag{3}$$

where F_c is the feature map at band c, F_{sq} and F_{ex} refer to the squeeze and excitation functions, respectively, σ denotes the sigmoid activation function and δ is the Rectified Linear Unit (ReLU) activation function. The resulting output of the SE block is generated by channel-wise multiplication (F_{scale}) between a scalar value and the original input feature map F_c, and can be written as:

$$a_c = F_{scale}(F_c, s_c) \tag{4}$$

Residual learning and SE blocks are integrated into ResNet models and are used as modules in the form of ResBlocks and SE-ResBlocks, as shown in Fig. 2. Shallow ResNet variants such as ResNet18 use the original ResBlock architecture; however, for ResNet50 and larger

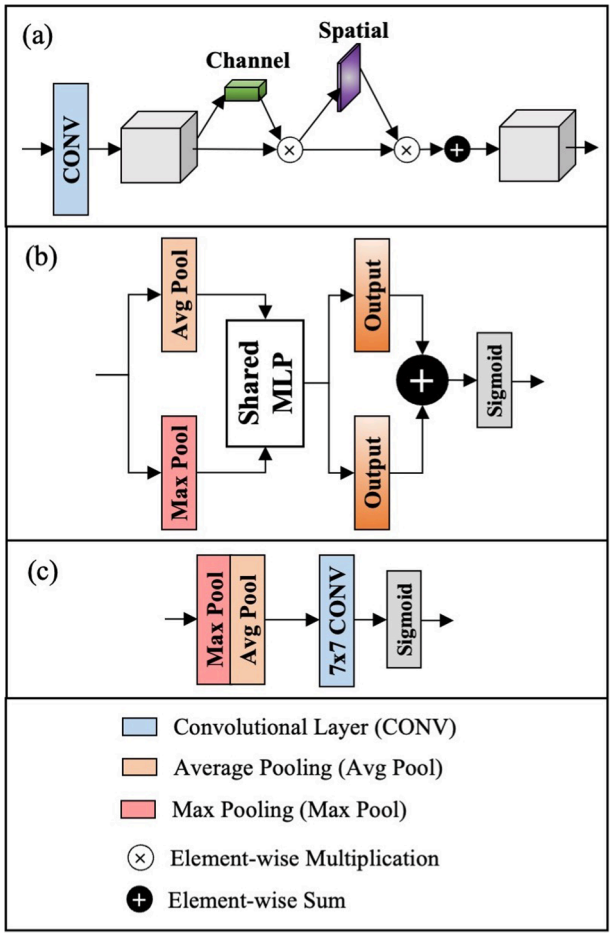

Fig. 3. CBAM architecture showing channel-spatial attention. (a) Module architecture, (b) Channel (spectral) attention module, (c) Spatial attention module.

variants, ResBlocks use a bottleneck architecture with 1 × 1 convolutional layers to reduce the number of parameters and matrix multiplications.

2.4. Convolutional block attention module

CBAM was introduced as an efficient unit consisting of the channel and spatial attention mechanisms to enhance the performance of the SE block and representation power by exploiting the inter-relationships along the channels and spatial axes using an attention mechanism to extract meaningful features (Woo et al., 2018). Compared to the SE block, CBAM uses max-pooling to retrieve more distinctive channel features. Simply, the channel attention module focuses on detecting the feature, while the spatial attention module locates the feature (Woo et al., 2018). In the context of LCZ classification, CBAM can be applied to exploit the spatial and spectral resolutions of multispectral satellite images. CBAM can be expressed using the following two expressions:

$$F' = M_{channel}(F) \otimes F \tag{5}$$

$$F'' = M_{spatial}(F') \otimes F' \tag{6}$$

where F is an input feature map from an intermediate layer, $M_{channel}$ is the inferred 1D channel attention map, $M_{spatial}$ is the inferred 2D spatial attention map, $\otimes$ symbolizes element-wise multiplication, and F'' is the final CBAM output at the original input feature map's dimensions. Hence, the intermediate feature map F is inputted into the channel attention module, the channel-refined feature map F' into the spatial attention module, and the CBAM result is F''.

For channel attention, the input feature map's dimension is squeezed and aggregated using both average pooling (AvgPool) and max pooling (MaxPool) to generate two different spatial context descriptors that are subsequently processed in the shared MLP. The simultaneous use of average and max pooling helps to infer distinctive object features, which are useful to distinguish heterogeneous features in urban areas (Woo et al., 2018). The processed descriptors are then merged using element-wise summation. For spatial attention, global average pooling (GAP) and max pooling are performed along the channel axis and concatenated to produce the resulting feature descriptor. The channel and spatial attention maps can be represented as:

$$M_{channel}(F) = \sigma\{MLP(AvgPool(F)) + MLP(MaxPool(F))\} \tag{7}$$

$$M_{spatial}(F) = \sigma(f^{(7,7)}\{[AvgPool(F); MaxPool(F)]\}) \tag{8}$$

where σ is the sigmoid activation function, $f^{(7,7)}$ denotes a filter with a size of 7 by 7 pixels. The entire CBAM module process and each attention path are shown in Fig. 3.

3. Proposed model and modules

3.1. Multi-scale (MS) module

Multi-scale (MS) modules are suitable for scene classification to exploit local information by fusing multiple perspectives of the input image at different scales. Smaller receptive fields can extract fine features from local information, while larger receptive fields can delineate spatial contextual features from global information (Agnes et al., 2020). Since urban landscapes are typically very heterogeneous in dense cities, the MS module can interpret the fine-grained, contextual features. For LCZ classification, this MS strategy has only been applied in one related study which uses an Inception-based module on LCZNet with residual learning and SE blocks (Liu and Shi, 2020). However, the effect of the MS module was not tested, although multi-scale layers show potential to improve LCZ classification of finer-scaled features. This study thus implements this MS module at the beginning of the proposed model after the input layer. The MS module consists of three convolutional kernels of 5, 3, and 1 pixel with filter sizes of 16, 32, and 16, respectively. The three outputs from the filters are concatenated together into a stacked feature map for further processing.

3.2. Multi-level attention (MLA) module

CNNs extract various low to high-level features that can be used as contextual information. Context aggregation is effective for semantic segmentation of remote sensing images to help the recognition of objects and regions (Zhang et al., 2020; Cheng et al., 2019). Contextual information is quintessential in deep learning-based LCZ classification since some LCZ classes are not easily differentiable in heterogeneous and complex urban and rural environments. Sen2LCZ applied multi-level feature fusion to combine features after convolution and double-pooling (average and maximum pooling) layers, generating multi-level feature maps (Qiu et al., 2020). CBAM was also integrated into Sen2LCZ after each convolutional block but demonstrated near-negligible improvements in classification accuracy (Qiu et al., 2020); however, CBAM was placed in Sen2LCZ's main backbone and processed in sequence with other convolution and pooling layers. While CBAM is recommended to be placed after each convolution layer (Woo et al., 2018), if CBAM is inputted in the model's main backbone, this placement may not be effective for small-sized scene classification due to the 7 × 7 convolutional kernel in CBAM's spatial attention and the large reduction ratio.

Inspired by these studies, this study introduces the multi-level attention (MLA) module as a branched unit, which implements CBAM

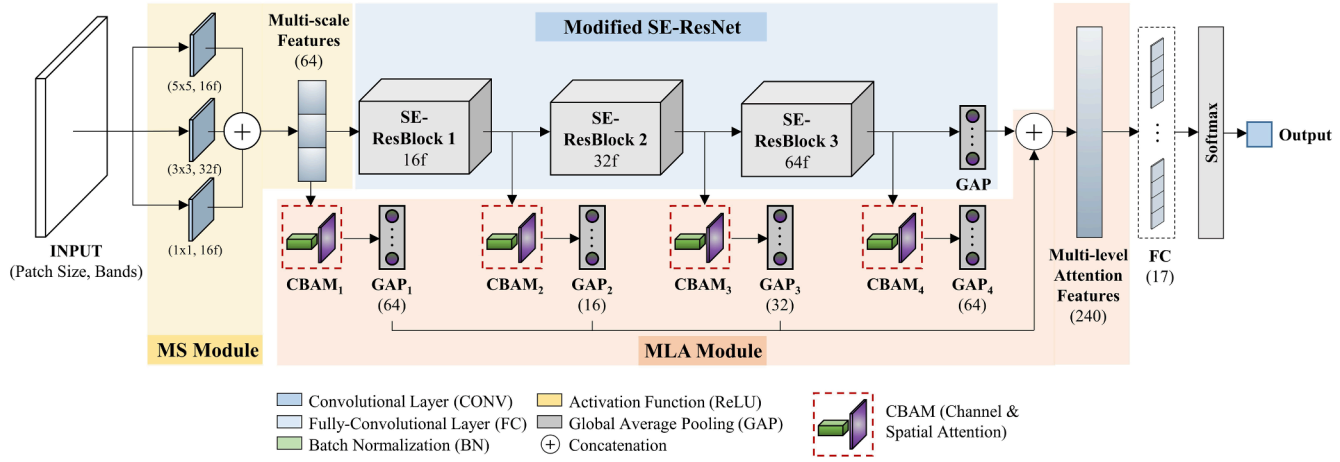

Fig. 4. Model architecture of proposed MSMLA-Net with MS (yellow) and MLA (red) modules on a SE-ResNet backbone. When only adding the MS module, the resulting 64-channel, multi-scale feature map is fed directly into the SE-ResNet backbone. At the end of the last SE-ResBlock, the output feature map has a window of 8 × 8 with 64 channels. GAP is processed on this feature map, followed by a softmax classifier layer for LCZ classification. When only adding the MLA module, the 7 × 7 convolutional layer in the original ResNet is modified to a 1 × 1 convolution, which acts to project the input layer data to the SE-ResNet backbone. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

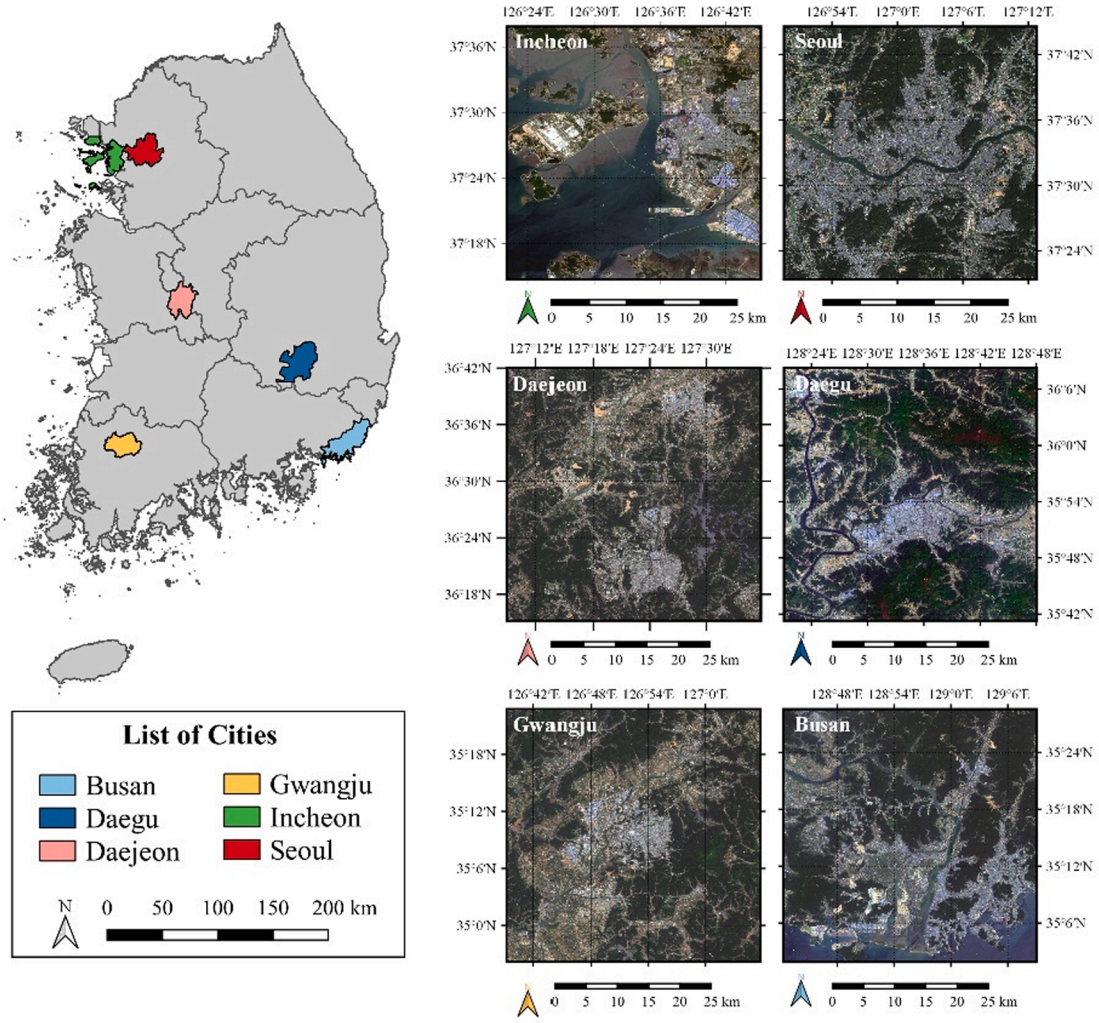

Fig. 5. Overview of test sites and their geographic locations with S2 images.

Table 1
S2 spectral bands and original specifications.

Band	Band Names	Central Wavelength (μm)	Spatial Resolution (m)
2	Blue	0.490	10
3	Green	0.560	
4	Red	0.665	
5	Vegetation Red-Edge1	0.705	20
6	Vegetation Red-Edge2	0.740	
7	Vegetation Red-Edge3	0.783	
8	NIR	0.842	10
8A	Vegetation Red-Edge	0.965	20
11	SWIR1	1.610	
12	SWIR2	2.190	

Table 2
List of acquired S2 input images.

City	Acquisition Date	City	Acquisition Date
Busan	2018/04/25 2018/06/24	Incheon	2018/05/28 2018/09/25
Daegu	2017/04/30 2017/06/09	Gwangju	2018/05/23 2018/09/25
Daejeon	2018/05/23 2018/08/01	Seoul	2017/05/03 2017/09/20

and GAP after each block of convolutional layers. In comparison to previous studies that used CBAM in LCZ classification models, the two novel aspects of MLA are: (1) the grouped usage of CBAM and GAP after each ResBlock to produce representative features, and (2) the placement of CBAM outside of the main ResNet backbone to avoid overly convoluting the feature maps. The by-products from CBAM and GAP groups are concatenated together to create multi-level CBAM-based features, which consist of local and global features that are more spatially and spectrally "aware". The multi-level features can then serve as contextual information and are aggregated with the feature map produced via the model's main ResNet backbone. In this study, CBAM uses a reduction rate of 8 and a 7×7 convolution kernel for spatial attention.

3.3. Proposed classification model: MSMLA-Net

This study proposes MSMLA-Net as a lightweight and effective LCZ classification model which integrates the MS and MLA modules into a modified SE-ResNet backbone. SE-ResNet18 and SE-ResNet50 are used as shallow and deep backbones and are specifically modified to be more efficient for LCZ classification. First, the fourth ResBlock from the original ResNet structure is omitted to ensure that the receptive field does not decrease to a negligible extent; second, GAP is added as the final pooling layer; third, the initial 7×7 convolution and max pooling layer in original ResNets are removed to avoid overly reducing feature map sizes. Moreover, MSMLA-Net uses deep ResNet backbones but is still lightweight since the MS and MLA modules do not add much computational load and GAP is helpful to extract representative features whilst minimizing the required number of parameters.

The two proposed modules (MS and MLA) can be added selectively to any backbone structure, and both are effective in improving LCZ classification accuracy. The structure of MSMLA-Net and the configuration of the two modules are shown in Fig. 4. Starting from the input image, MSMLA-Net consists of a three-layer MS module, three SE-ResBlocks, the MLA module, and the classification head. The MS module produces a concatenated 64-channel, multi-scale feature map which is fed into the SE-ResNet backbone as well as the MLA module. First, the multi-scale feature map generated by the MS module is sent to the SE-ResNet backbone, which consists of three SE-ResBlocks with filter sizes that double after every block. Filter sizes of 16, 32, and 64 are used to maintain a lightweight model and avoid unnecessary overhead. The final feature map at the end of the third SE-ResBlock is produced with a size of 8×8 pixels and 64 channels. Second, the multi-scale feature map is branched out into the MLA module. In more detail, the feature map is fed into a series of CBAM and GAP, generating an output with a dimension size of 64. This CBAM and GAP group is processed after each SE-ResBlock, producing three by-products with dimension sizes of 16, 32, and 64. In total, one multi-scale and multi-level CBAM by-product and three multi-level CBAM by-products are generated and are concatenated with the GAP product from the SE-ResNet backbone to create an aggregated feature map with a dimension size of 240. Subsequently, the classification head accepts the multilevel attention features via an FC layer and the softmax function retrieves class-wise probabilities to output a single LCZ class.

4. Test sites and input data

4.1. Test sites

Six cities were selected as test sites based on the city's morphology, population, and geographic location: Busan, Daegu, Daejeon, Gwangju, Incheon, and Seoul. The geographic location of each test site and an S2 image of each city is displayed in Fig. 5. The satellite images shown in the figure were acquired during the Spring to Autumn season. The area of each city is color-coded and matched to the north arrows below each satellite image for the reader's convenience.

Busan is situated in the southeastern corner of the peninsula, harboring one of the world's busiest ports. The city is mixed with natural landscapes including low mountains and rivers. Most of Busan's urbanized areas such as Haeundae are located near the southeast coast and are filled with numerous skyscrapers, high-rise buildings, and compactly built urban complexes. There are also many large industrial complexes within Busan and near the ports. Daegu is in southeastern Korea and is surrounded by numerous mountains and smaller hills. To the east and northeast of the city, there are large industrial complexes for manufacturing, textiles, and mining. High-rise buildings and urban complexes can be found at the heart of the city center. Daejeon is situated in central South Korea and is composed of numerous administrative and government facilities. The city is also home to numerous universities as well as research and development complexes. In Daejeon's S2 image used in this study, another city called Chungju is located to the northeast of Daejeon and is another important center for agriculture. Gwangju is in southwestern Korea and is known as a regional agricultural and commercial hub. The city has a large proportion of paddy fields and dry fields, making Gwangju a key area for grain and rice production. Incheon is in northwestern South Korea and is a coastal city with numerous ports. Like Busan, Incheon also consists of numerous industrial complexes that are situated near ports, while major urban areas are located near the city's center and to the east. Seoul is the capital city of South Korea with a population of approximately 10 million people and a very high population density. Large and dense urban areas are filled with numerous skyscrapers and high-rise buildings both north and south of the Han River, which flows through the heart of the city.

4.2. Input data

(1) S2 multispectral satellite images (Main input data)

S2 satellite images are distributed at up to 10 m in spatial resolution and offer 12 spectral bands. Although Landsat 8 images were mainly used in the WUDAPT protocol and in many previous studies, most recent deep learning-based studies incorporated S2 images (Qiu et al., 2018,

Table 3
Clustering of Level-2 NLCM classes to match LCZ classes.

Level-2 NLCM Class	LCZ Class	Level-2 NLCM Class	LCZ Class
Residential	Urban	Deciduous Forest	A
Industrial	10	Coniferous Forest	A
Commercial	Urban	Mixed Forest	A
Communication	Urban	Natural Grassland	C
Transportation	Urban	Non-Natural Grassland	C
Public Utilities	Urban	Inland Wetland	-
Paddy Field	D	Coastal Wetland	-
Non-Irrigated Land	D	Natural Barren Land	F
Protected Cultivation	D	Non-Natural Barren Land	E
Orchard	D	Inland Water	G
Other Cropland	D	Seawater	G

2019, 2020; Rosentreter et al., 2020; Yoo, 2020). Considering this recent trend in the literature, this study used S2 images with band specifications shown in Table 1. Coastal aerosol, water vapor, and cirrus bands (bands 1, 9, and 10) were omitted due to their minor significance with LCZ classification.

The images specified in Table 2 were acquired as S2 Level-1C top-of-atmosphere radiance products during favorable atmospheric conditions since LCZ classes are best observed over dry surfaces on calm, clear nights (Stewart and Oke, 2012). Two images of 5,010 by 5,010 pixels (relative to 10 m spatial resolution) were acquired for each city during a period between late Spring and early Autumn. The images were downloaded from the Copernicus Open Access Hub, but are also available as cloud-masked mosaics using Google Earth Engine (Schmitt et al., 2019) and via the *Sentinelsat* API. A common scaling factor was applied to all the S2 images, and all of the bands were then resampled to 10 m spatial resolution using bilinear interpolation and co-located to the same extent.

(2) Ancillary data (Optional input data)

Three sets of ancillary data were included as individual input layers in the model's training dataset to integrate information on feature height, building footprint, and LULC. An example of each dataset is shown in the Appendix in Fig. A1. First, building footprint data from OSM was used, which also provides information on the density of the features. For instance, LCZ 1–3 and LCZ 4–6 can differ by feature distribution and spacing. Many related studies integrated OSM data to enhance overall LCZ classification accuracy (Fonte et al., 2019; Qiu et al., 2018) and to extract building fractions at different scales (Bechtel et al., 2019). Second, given the potential of building height information from a DSM (Yoo et al., 2020; Zhou et al., 2020), this study utilized the AW3D30 DSM generated by the Panchromatic Remote-sensing Instrument for Stereo Mapping which offers global height information at 30 m in spatial resolution (Tadono et al., 2016). Further, AW3D30 was found to be the most robust and stable source of global digital elevation among freely available sources (Uuemaa et al., 2020). Third, the level-2 national land cover map (NLCM) developed by the Ministry of Environment in 2019 was used as LULC information to enhance the classification of natural LCZ classes. The NLCM contains 23 different classes and is provided at a map scale of 1:25,000. In general, the level-2 NLCM is useful for regional-level land-use observations. Although LULC maps may not consider climatic properties, the land use classes provide valuable information on the form and function of urban and natural landscapes. LULC maps are generally maintained to a high standard and can be a helpful proxy in LCZ classification (Wang et al., 2018).

The 2019 version of the NLCM was implemented in this study because the map offered national coverage, whereas maps from previous years did not contain data in specific cities. The level-2 NLCM classes were aggregated to corresponding LCZ classes based on similar landscape characteristics and are shown in Table 3. Since all the NLCM classes did not directly match with LCZ classes, certain classes, such as inland and coastal wetland, were omitted. Also, since the NLCM data does not include building height, urban function-related classes were clustered into a single LCZ class.

4.3. LCZ point sample labeling

To generate high-quality training samples, this study combined good sampling protocol for LCZ classification from various studies. The rigorous three-step LCZ sample labeling methodology used to create the So2Sat dataset was adapted for this study (Zhu et al., 2020). The main difference of this study's method with So2Sat (Zhu et al., 2020) is the use of LCZ point samples and verification using high-quality reference data.

First, in the "learning" phase, the self-assessment test by HUMINEX is helpful for users to familiarize themselves with the LCZ classification scheme (http://77.69.20.19/dev/driver/training.php). This test is composed of 357 test images, shown using Google or Bing Maps, and asks the user to find the correct LCZ class. The user's test score is recorded and can provide a benchmark evaluation of labeling quality. This step is crucial since, without proper knowledge of the LCZ classes and their characteristics, users may produce biased and dissimilar LCZ labels (Bechtel et al., 2017). The image patches in this study were sampled following successful completion of the driver test and by considering similar class labels from an earlier study on the LCZ classification of Korean cities (Kim and Eum, 2017). Example patch images of each LCZ class are shown in the Appendix in Fig. A2.

Second, in the "labeling" phase, the S2 image patches (training data) and LCZ class patches (ground truth labels) were manually labeled as points with the support of high-quality reference data. In more detail, a region of homogeneous LCZ characteristics is first identified based on the LCZ classification scheme and the required scale of the input image (10 m spatial resolution). Then, the LCZ samples are labeled as points, instead of polygons like in WUDAPT. The advantage to this method is that point sampling can be less cumbersome than meticulously drawing edges of polygons and is more efficient to develop patches for scene classification.

Third, in the "validation" phase, three S2 bands (RGB) served as a basemap. Google Earth and GSV were compared with this basemap to distinguish similar and confusing LCZ classes based on WUDAPT's sampling protocol (Bechtel et al., 2019; Mills et al., 2015). To develop high quality training samples, additional GIS layers were included in accordance with our previous study (Kim et al., 2020). NDVI was used as an indicator of perviousness and to assist with the classification of natural LCZ classes. In particular, to aid with the interpretation of elevation-related LCZ classes 1–6, the Master Architectural Building Plan (MABP) dataset provided by the Electronic Architectural Administration Information System (EAIS) was used as high-quality building information reference data. The building plan dataset consists of various types of building information for each major city in Korea such as building height, number of floors, building area, and geographic location. The building height thresholds for MABP were set based on the LCZ definitions for height features (Low: 0–3 floors, Mid: 3–10 floors, High: >10 floors). To note, MABP was only used in this study to aid with LCZ sampling as shown in the Appendix in Fig. A3 and was not included in the training dataset. Detailed information on sample labeling is shown by the modified flowchart in the Appendix in Fig. A4.

4.4. Point-to-patch extraction

Once labeling is complete, the LCZ point samples act as centroids to extract non-overlapping image patches from the input data. This step is referred to as "point-to-patch" extraction. The point sample's LCZ class is then allocated to a ground truth label patch with the same size as that of the image patch. In this study, all the input image data are cropped to patch sizes of 32 by 32 pixels (320 m by 320 m) and 48 by 48 pixels (480 m by 480 m). For simplicity, patch sizes of 32 by 32 pixels and 48 by 48 pixels are henceforth abbreviated as "P32" and "P48", respectively. P32 is most widely used in LCZ classification studies and the global So2Sat

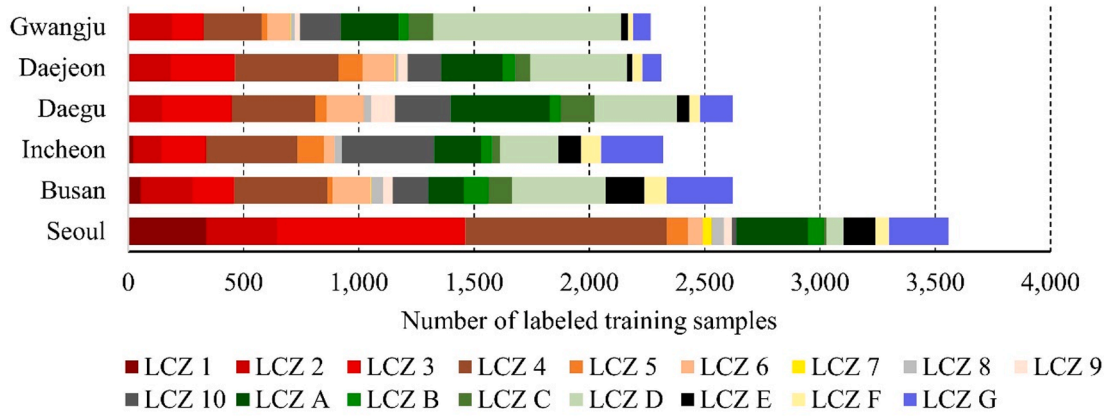

Fig. 6. Distribution of the total number of labeled LCZ samples per city.

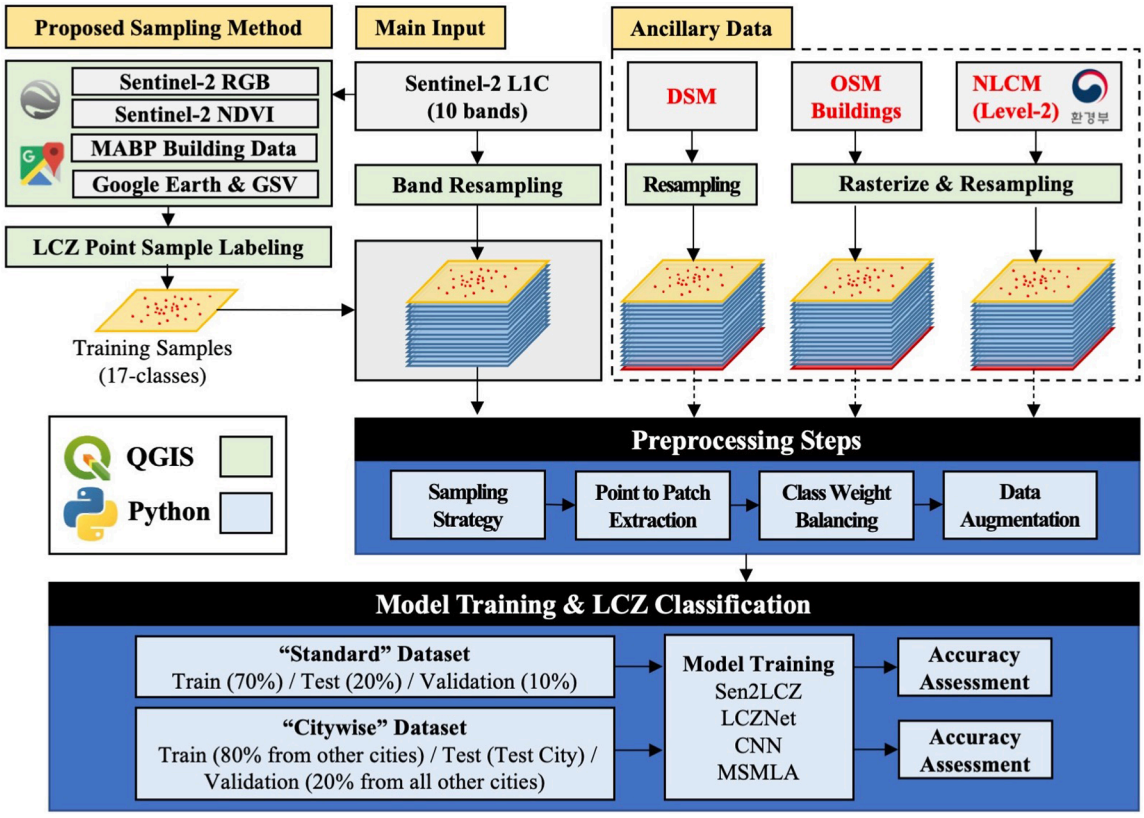

Fig. 7. Overall workflow showing the implementation of the proposed framework.

dataset, whereas P48 demonstrated optimal classification accuracy with LCZNet.

To generate LCZ maps, the dense feature maps are obtained through a sliding window approach, similar to the methodology in many previous LCZ classification studies (Yoo et al., 2019; Rosentreter et al., 2020; Qiu et al., 2020; Liu and Shi, 2020; Yoo et al., 2020). All the input data used for LCZ classification are resampled to 10 m resolution. Zero-padding is applied to the outside boundaries of the input data based on the patch size. Using P32 as an example, (32–10)/2 zero-value pixels are added to each side of the 5010 by 5010-pixel input image, creating a padded 5032 by 5032-pixel image. The sliding window traverses through the entire padded image using a step size of 10 pixels to assign pixel-level predictions for each patch. The final product is generated as an LCZ map resized to 501 by 501 pixels to match a spatial resolution of 100 m. The distribution of labeled training samples for each city is shown in Fig. 6 and the number of samples per class is specified in the Appendix in Table A1.

5. Experimental setup

5.1. Sampling strategies and implementation details

Two sampling strategies are used to split the input dataset into training, test, and validation sets. First, the "Standard" sampling method is based on using points extracted from a grid with equal pixel spacing to select spatially non-overlapping patches. In more detail, grids are created for each of the six cities and grid cells are randomly selected using an 80 to 20 ratio to designate training and test cells, respectively. For each city, the training and test cells are then overlaid on the sampled LCZ label points specified in Section 4.3 and intersecting points within

Table 4
Classification accuracy metrics.

Category	Metric	Formula	#
Overall	Overall Accuracy (OA)	$\frac{1}{N}\sum_{i=1}^{17} N_i^C$	(9)
	Weighted Accuracy (WA)	$\frac{1}{N}\sum_{i=1}^{17} w \bullet N_i^C$	(10)
	F1-Score (F1)	$2\left(\frac{Precision \times Recall}{Precision + Recall}\right)$	(11)
Class-Aggregated	Built-up OA (OA_{BU})	$\frac{1}{N_{built-up}}\sum_{i=1}^{10} N_i^{C,built-up}$	(12)
	Natural OA (OA_N)	$\frac{1}{N_{natural}}\sum_{i=11}^{17} N_i^{C,natural}$	(13)

the superimposed cell are extracted. These points are then used for point-to-patch extraction to create training and test patches. Subsequently, the set of training patches is split to generate a validation set. In sum, training, test, and validation sets are created in approximately 70:20:10 proportions, respectively. Second, the "Citywise" sampling method is based on sampling data from each city in a leave-one-out manner. In essence, test sets are created from points sampled from one city, while training sets include points from the remaining cities. The training set is then split to create a validation set using an 80 to 20 percent ratio. A total of six training, test, and validation sets are created for each of the six cities. For training datasets with large and diverse sample distributions such as So2Sat, geographic sampling based on non-overlapping geographic regions (such as East and West halves of each test city) may be considered more robust. For custom datasets of local regions, however, geographic sampling may not divide training and test datasets with ample variance, such that the model would learn from areas with similar characteristics and would inevitably yield biased classification results. Sampling strategies must therefore be selected in accordance with the characteristics of the sampled training dataset.

Since the urban morphology and landscape composition of different cities may vary, manual sampling of ground truth LCZ labels may result in an imbalanced distribution of LCZ classes, as shown in Fig. 6. This class imbalance can result in a biased classification for specific classes. To alleviate this issue, a class weighting method from the *scikit-learn* Python library (Pedregosa, 2011) is applied to adjust the LCZ class prior probabilities (Rosentreter et al., 2020). Furthermore, data augmentation from the *albumentations* Python library (Buslaev et al., 2020) is implemented to apply horizontal flips, vertical flips, and rotations to increase the size of the dataset and reduce overfitting. All experiments were performed using an Intel Core i7-6700 CPU at 3.40 GHz and an NVIDIA GeForce RTX 2070 Super Graphics Processor Unit (GPU) with 8 GB of memory. Python 3.7.9 was used with Tensorflow 2.3.0. For training hyperparameters, an early stop of 15 epochs, a learning rate of 0.002, and a decay factor of 0.004 were used. The adaptive moment estimation (adam) optimizer was chosen to minimize the cross-entropy loss function. Filter weights were initialized using "He normal" initialization (He et al., 2015). All models in the experiments were trained from scratch. An overall workflow of the entire LCZ classification process is shown in Fig. 7 for the reader's convenience.

5.2. Accuracy assessment

Overall accuracy (OA), built-up OA, (OA_{BU}), and natural OA (OA_N) have been used in many related studies (Bechtel et al., 2015; Zhu et al., 2020; Yoo et al., 2019; Rosentreter et al., 2020; Qiu et al., 2020; Liu and Shi, 2020; Demuzere et al., 2019). For class-wise OA metrics, OA_{BU} is an average of OA of each class in LCZ 1 to 10 ($N_i^{C,built-up}$), while OA_N uses LCZ A to G ($N_i^{C,Natural}$). However, since OA can be unreliable for imbalanced datasets, weighted accuracy (WA) was also computed, where weights (w) from a similarity matrix designed based on climatic properties such as openness, height, cover, and thermal inertia are used to consider intra-class similarities for LCZ classification (Bechtel et al., 2017). To account for the imbalanced dataset, F1-score was calculated as a weighted value based on the number of samples in each LCZ class. The specific formula for each accuracy metric is listed in Table 4.

In general, the classification of natural LCZ classes is relatively easier compared to built-up LCZ classes for highly heterogeneous and dense cities. Natural LCZ classes tend to be more homogenous and occupy larger expanses, whereas built-up LCZ classes include many tiny objects and fine-grained features. OA_N results are therefore typically very high in accuracy and may drive the OA and F1 results upward. Based on this reasoning, OA_{BU} can be considered as a stricter indicator of LCZ classification model performance for modern cities with a high level of landscape heterogeneity.

6. Experimental results and discussion

6.1. LCZ classification with SOTA models

The SOTA models referenced in this study are implemented with the following specifications. First, Sen2LCZ uses 17 convolutional layers with 3 × 3 kernels and 16 filters (Sen2LCZ f16D17 in the original study) (Qiu et al., 2020) based on the model's optimal performance in the original study. Second, LCZNet is composed of a 3-layer MS layer followed by six SE-ResBlocks with the number of filters doubling for every two residual blocks. The initial SE-ResBlock uses 16 filters and a max pooling layer is added after the second and fourth SE-ResBlock, while a GAP layer is added after the sixth SE-ResBlock. A 3 × 3 convolutional kernel is used throughout the model. Third, CNN is a VGG-style model with four convolutional blocks, each containing a convolutional layer, a batch normalization layer, a ReLU activation function, and a max pooling layer. The first block uses 16 filters, which is doubled for every subsequent block. Similarly, a 3 × 3 convolutional kernel is used throughout the model. After the four convolutional blocks, an FC layer and a dropout layer are added.

All models were trained using "All S2" and "All + Ancillary" datasets and tested using the "Standard" sampling strategy. The classification results for "All S2" and "All + Ancillary" are shown in Tables 5 and 6, respectively, while confusion matrices are provided in Figs. 8 and 9.

(1) Results from the "All S2" Dataset

The proposed MSMLA-Net models outperformed all other models for both patch sizes, with MSMLA-50 recording the highest accuracy for all classification metrics. While deeper models are known to improve model performance, increasing the number of trainable parameters did not always translate to higher classification accuracy, as evidenced by the

Table 5
Classification accuracy results of SOTA models using "All S2" data.

Model	Trainable Parameters	Accuracy for P32 (%)					Accuracy for P48 (%)				
		OA	WA	OA_{BU}	OA_N	F1	OA	WA	OA_{BU}	OA_N	F1
Sen2LCZ	793,348	82.3	75.1	54.7	87.9	80.4	81.0	73.9	54.4	84.8	78.7
LCZNet	3,181,809	82.9	81.3	66.3	89.1	81.9	83.0	79.8	63.8	86.5	82.2
CNN	235,105	84.6	80.2	65.2	89.2	83.7	83.4	78.1	61.1	89.0	82.2
MSMLA-18	181,867	84.4	82.2	**68.4**	88.5	83.5	84.4	80.4	64.7	89.4	83.5
MSMLA-50	808,913	**86.1**	**82.8**	**68.4**	**91.1**	**85.6**	**87.1**	**85.0**	**72.4**	**91.8**	**86.5**

Table 6
Classification accuracy results of SOTA models using "All + Ancillary" data.

Model	Trainable Parameters	Accuracy for P32 (%)					Accuracy for P48 (%)				
		OA	WA	OA_{BU}	OA_{N}	F1	OA	WA	OA_{BU}	OA_{N}	F1
Sen2LCZ	793,348	82.2	74.9	56.6	87.1	80.5	81.5	78.8	63.4	85.0	79.3
LCZNet	3,181,809	84.2	81.1	65.2	89.6	83.7	84.1	79.8	62.6	89.7	83.3
CNN	235,105	84.1	81.7	66.7	89.4	83.3	83.6	82.2	67.0	90.0	82.8
MSMLA-18	181,867	85.1	83.4	70.4	89.7	84.5	85.2	81.9	66.6	89.8	84.6
MSMLA-50	808,913	**87.0**	**84.6**	**70.7**	**91.9**	**86.6**	**87.4**	**85.6**	**73.5**	**91.5**	**87.0**

Table 7
Ablation study on MS and MLA modules using ResNet models.

Models	Modules			Accuracy (P32)				Accuracy (P48)			
	SE	MS	MLA	OA	OA_{BU}	OA_{N}	F1	OA	OA_{BU}	OA_{N}	F1
ResNet18				82.2	60.6	87.4	81.0	82.8	62.0	87.1	81.7
+SE	✓			82.0	58.8	87.2	80.7	82.5	59.8	87.9	81.3
+SE+MS	✓	✓		84.0	68.2	88.1	83.2	**84.4**	64.6	**89.4**	**83.7**
+SE+MLA	✓		✓	**84.8**	66.9	**89.1**	**84.2**	83.8	**66.4**	89.3	82.7
+SE+MS+MLA	✓	✓	✓	84.4	**68.4**	88.5	83.5	**84.4**	64.7	**89.4**	83.5
ResNet50				80.4	58.0	85.0	79.2	81.5	63.1	86.7	80.4
+SE	✓			82.6	66.7	87.1	81.6	83.3	67.0	88.4	82.3
+SE+MS	✓	✓		86.0	68.7	89.6	85.5	86.7	71.1	91.4	86.2
+SE+MLA	✓		✓	**86.2**	**70.4**	90.1	**85.8**	86.5	68.8	90.3	85.9
+SE+MS+MLA	✓	✓	✓	86.1	68.4	**91.1**	85.6	**87.1**	**72.4**	**91.8**	**86.5**

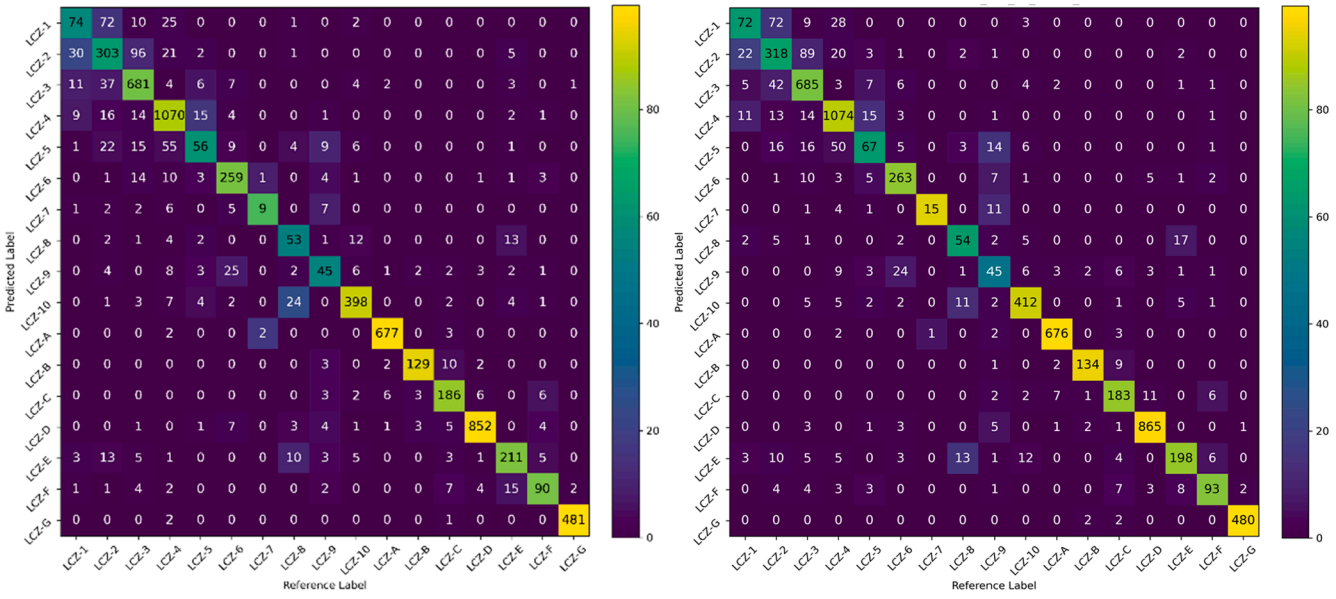

Fig. 8. Confusion matrix for MSMLA-50 with "All S2". Left: P32. Right: P48.

results of LCZNet in Table 5. Previous studies also noted that large models, particularly computer vision-based models, recorded low to negligible contribution to the classification accuracy (Zhu et al., 2020; Qiu et al., 2020; Liu and Shi, 2020). The significance of MS and MLA is emphasized when comparing MSMLA-Net with similarly sized models. In contrast to the Sen2LCZ benchmark's accuracy results, MSMLA-50 recorded improvements with P32 (OA: +3.8%, F1: +5.2%) and P48 (OA: +6.1%, F1: +7.8%), despite using nearly the same number of parameters. Even the lightweight MSMLA-18 recorded higher accuracy over all three SOTA models, reinforcing the effectiveness of MS and MLA. MSMLA-18 and CNN reached similar OA and F1 results, but CNN may have suffered from over-fitting due to the large FC layer at the end of the model. Instead of relying on FC layers, MSMLA-Net uses GAP to aggregate output feature maps which reduces the model's size substantially without forfeiting model performance. MSMLA-18 is lighter and yielded higher WA than CNN for P32 (ΔWA: +2.0%) and P48 (ΔWA: +2.3%), showing MSMLA-18's ability to classify difficult, ambiguous LCZ classes more effectively.

In terms of built-up LCZ classification, MSMLA-50 surpassed 70% OA_{BU} when using P48. Although MSMLA-18 was not as effective, the lightweight model still outperformed all other models in OA_{BU} for both patch sizes. For instance, MSMLA-18 recorded higher OA_{BU} compared to CNN for P32 (ΔOA_{BU}: +3.0%) and P48 (ΔOA_{BU}: +3.6%). The shallow MSMLA-18 can generate sufficiently accurate classification results, while the deeper MSMLA-50 returns optimal performance, particularly for built-up LCZ classification.

As modern cities grow denser and more complex, LCZ classification of urban agglomerations will need to identify ambiguous LCZ classes more accurately (Yoo et al., 2020). The confusion matrices in Fig. 8 exhibit intra-class confusion among built-up LCZ classes within the LCZ

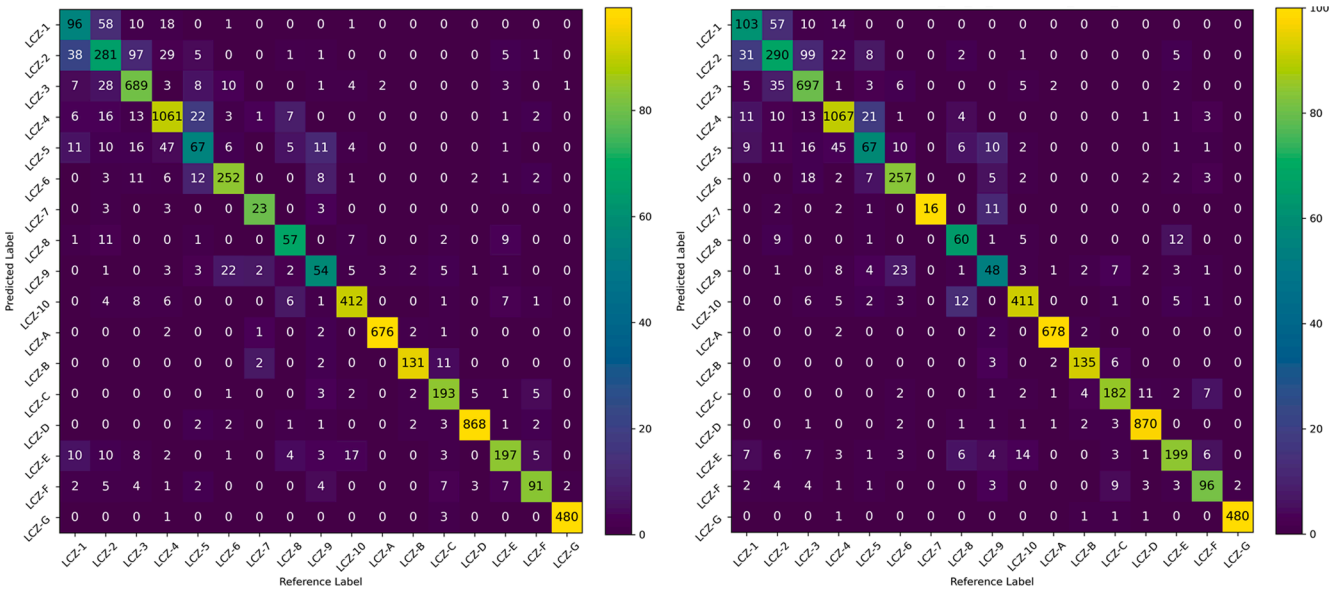

Fig. 9. Confusion matrix for MSMLA-50 with "All S2 + Ancillary". Left: P32. Right: P48.

classes 1 to 6 range due to the scarcity of height information. These classes are considered to be more difficult to classify in recent studies using remote sensing data (Qiu et al., 2020). LCZ classes 6 (Open Low-rise) and 9 (Sparsely Built) were commonly confused due to the similar characteristic of low features in open areas. LCZ E (Bare Rock/Paved) was also mistaken for LCZ 8 (Large Lowrise) since the spectral response of large, paved areas and large low-rise features may overlap.

(2) Results from the "All + Ancillary" Dataset

For "All + Ancillary", the proposed MSMLA-Net models recorded superior classification accuracy for both patch sizes and both input datasets. The integration of ancillary bands tended to have a greater effect in enhancing classification accuracy across all models compared to inputting a larger patch size. Increasing patch size was helpful for MSMLA-Net models, only if the models were able to learn the additional contextual data. This observation can also be attributed to the higher WA results when using "All + Ancillary" data for both patch sizes suggesting that the contextual information was more helpful for the classification of ambiguous classes. MSMLA-50 outperformed all models again, reaching over 87% in both OA and F1 when using P48. MSMLA-50 also returned the highest OA_{BU} (P32: 70.7% and P48: 73.5%), revealing that the deep variation of the proposed model was the most effective at using the additional contextual information.

Similar to the results from "All S2", most misclassification occurred for LCZ classes 1 to 6. Based on the confusion matrices in Fig. 8, however, adding ancillary data was beneficial for classifying LCZ classes 1 (Compact Highrise) and 8 (Large Lowrise). The addition of height information in the DSM helped to identify LCZ 1, while OSM data and the exclusive category in NLCM helped differentiate buildings from paved surfaces.

With regards to built-up LCZ classification, only LCZNet experienced a significant reduction in OA_{BU} when using ancillary bands with respect to the "All S2" for P32 (ΔOA_{BU}: −1.1%) and P48 (ΔOA_{BU}: −1.2%). This result, in conjunction with that of "All S2", highlights the importance of finding an optimal model depth that can support the implementation of MS and MLA modules. While this study used ResNet as a backbone for three ancillary bands, future studies may need to consider different model depths or backbones to accommodate additional ancillary band combinations. An in-depth evaluation of MSMLA-Net is therefore provided in the subsequent section and the specific influence of ancillary bands is investigated in more detail in Section 6.3.

6.2. Model-based experiments

Model-based experiments were designed in the form of an ablation study to investigate the influence of model depth, the MS and MLA modules, and two patch sizes (P32 and P48). The ablation study used modified structures of ResNet18 and ResNet50 to compare shallow and deep models for LCZ classification and the results are provided in Table 7. The highest accuracy values for each metric are emphasized in bold text and the best result for all cases is highlighted. To note, the additions of SE, MS, and MLA (+SE + MS + MLA) shown in Table 7 are synonymous with the MSMLA-Net models proposed in this study. The number of parameters for all variations of MSMLA-Net and additions of MS and MLA in this study are provided in the Appendix in Table A2.

6.2.1. Effect of model depth

In general, the model depth of the ResNet backbones was found to be correlated with classification accuracy. Since SE blocks tended to improve accuracy for both model depths, the MS and MLA modules were only added to the SE-ResNet backbone. However, the classification accuracy of the SE-ResNet plateaued at SE-ResNet50, indicating that excessive model depth was unnecessary for LCZ classification. As a result, SE-ResNet18 and SE-ResNet50 were used as the shallow and deep backbones, respectively. A comprehensive evaluation of ResNet and SE-ResNet backbones at varying depths is provided in the Appendix in Tables A3 and A4.

The results from ablation studies in Table 7 revealed that deeper models connected with SE blocks yielded higher classification accuracy. Classification results using only ResNet backbones (without SE, MS, MLA) demonstrated that the shallow model recorded higher accuracy for P32 and slightly higher OA, OA_N, and F1 for P48. However, adding SE blocks to ResNet18 resulted in a slight reduction in OA and F1 and a considerable drop in OA_{BU}, since the inclusion of SE blocks actually decreased the number of trainable parameters in SE-ResNet18 (165,549 parameters) compared to the modified ResNet18 (185,137 parameters) used in this study. On the other hand, although the modified ResNet50 model recorded the lowest accuracy results when used alone, adding SE blocks to the backbone helped improve accuracy significantly for P32 (ΔOA: +2.15%, $\Delta F1$: +2.45%) and P48 (ΔOA: +1.85% and $\Delta F1$: +1.92%), indicating the importance of using SE blocks to help propagate information in deeper models. This result agrees with a recent study (Liu and Shi, 2020) which highlighted the significance of adding SE-blocks to LCZNet for LCZ classification.

Table 8
Classification results from adding CBAM in the main backbone after every convolutional layer in contrast to CBAM as a branched unit (MLA).

Patch Size	Backbone	Placement of CBAM	OA (%)	OA_{BU} (%)	OA_N (%)	F1 (%)
P32	SE-ResNet18	Backbone	81.9	57.2	86.1	80.3
		MLA (Branched)	**84.8**	**66.9**	**89.1**	**84.2**
	SE-ResNet50	Backbone	82.1	61.1	85.9	81.1
		MLA (Branched)	**86.2**	**70.4**	**90.1**	**85.8**
P48	SE-ResNet18	Backbone	83.1	64.0	87.1	82.4
		MLA (Branched)	**83.8**	**66.4**	**89.3**	**82.7**
	SE-ResNet50	Backbone	83.7	66.0	89.2	83.0
		MLA (Branched)	**86.5**	**68.8**	**90.3**	**85.9**

6.2.2. *Effect of adding MS and MLA*

Recall that the MS layer is introduced after the model's input layer to generate multi-scaled feature maps and can encapsulate the heterogeneity and multi-sized characteristics of features in complex landscapes. Further, the MLA module is added as a branched unit to avoid over-convoluting feature maps and combines multi-level by-products of CBAM and GAP from this branch for context aggregation of local and global features. Based on these descriptions, the ablation studies also presented the influence of using the MS and MLA modules individually and simultaneously. MS and MLA modules were integrated into SE-ResNet backbones given the improvements with SE blocks shown in the previous sub-section.

First, the individual addition of MS and MLA produced vastly higher accuracy results compared to when using only SE-ResNet backbones. This trend was consistent across both aggregated and classwise metrics for both patch sizes, confirming the effectiveness of each module. Second, both modules were most effective when using SE-ResNet50 as the backbone. MS and MLA returned the highest aggregated accuracy (OA: 86.7% and 86.5%, F1: 86.2% and 85.9%) when using SE-ResNet50 with P48. Third, the MS module tended to improve aggregated accuracy metrics (OA and F1) with P48 data, whereas the MLA module was typically more effective with P32 data. Fourth, the MS and MLA modules were especially effective in raising OA_{BU}; however, the contribution of each module to OA_{BU} was ambiguous. To elaborate, MS recorded higher OA_{BU} for the shallow model with P32 and the deeper model with P48, while MLA returned higher OA_{BU} for the deeper model with P32 and the shallow model with P48. Lastly, while individual additions of MS and MLA helped improve accuracy, the combined addition of MS and MLA using SE-ResNet50 (identical to MSMLA-50) yielded the highest accuracy for all metrics, reaching up to 87.1% OA and 86.5% F1. In particular, MSMLA-50 with P48 generated the highest class-aggregated OAs, highlighting the synergy of the two modules. Ultimately, MSMLA-50 with P48 was determined to be the optimal model for deep learning-based LCZ classification and the model can serve as a good benchmark for future LCZ classification research, particularly when mapping heterogeneous and complex areas.

6.2.3. *Effect of CBAM placement*

Attention mechanisms such as the SE block and CBAM can be used to link deeper models more efficiently (Hu et al., 2018; Woo et al., 2018). Previous studies on deep learning-based LCZ classification opted to add CBAM in the main backbone at the end of every convolutional block and noted insignificant improvements in accuracy (Qiu et al., 2020). However, the low classification accuracy may have occurred since feature maps after multiple convolution and pooling layers may be overly reduced or complex for CBAM to process effectively. To address this shortcoming, this experiment was designed to compare the effect of adding CBAM in a multi-level branched unit instead of the main backbone.

In more detail, CBAM was inserted in MSMLA-Net after each SE-ResBlock in the main backbone and the classification results were compared with those of SE-ResNet18 + MLA and SE-ResNet50 + MLA in Table 8. When using CBAM in the main backbone, the deeper backbone and larger P48 data led to higher classification accuracy; nevertheless, these accuracy improvements were marginal when compared with the results from adding the MLA modules. MLA was especially more effective when comparing the OA_{BU} between the two module placements at P32 for SE-ResNet18 (ΔOA_{BU}: 9.7%) and SE-ResNet50 (ΔOA_{BU}: 9.3%).

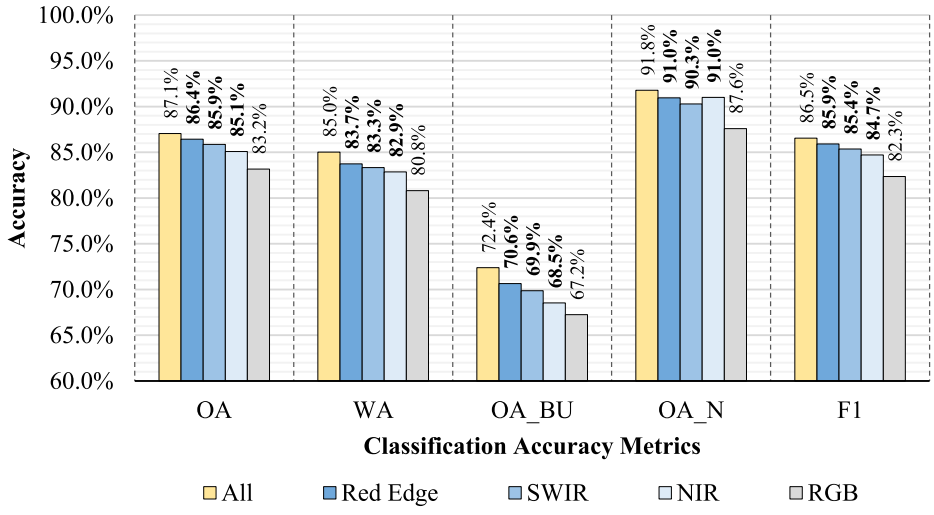

Fig. 10. Comparison of various S2 band combinations on classification accuracy.

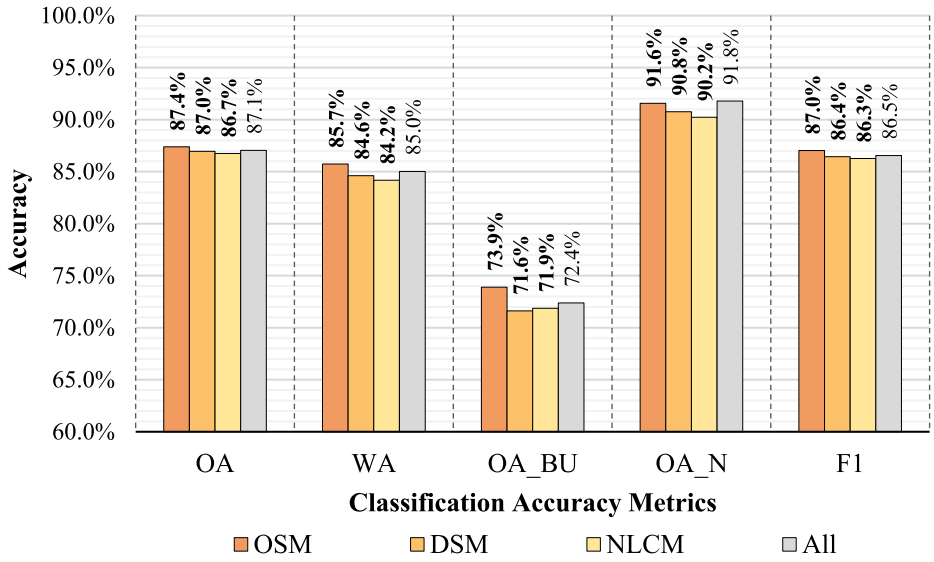

Fig. 11. Comparison of ancillary band combinations for LCZ classification.

Table 9
MSMLA-Net using only ancillary bands with P48.

Model	OA (%)	WA (%)	OA_{BU} (%)	OA_N (%)	F1 (%)
MSMLA-18	69.7	67.7	53.9	68.1	67.5
MSMLA-50	**77.6**	**81.1**	**75.0**	**75.0**	**76.8**

6.3. Data-based experiments

6.3.1. Influence of S2 spectral band combinations

The influence of the S2 spectral bands was evaluated in this experiment using MSMLA-50 and P48. S2 RGB bands (Bands 2, 3, 4) were used as the baseline input dataset, while combinations of NIR (Bands 2, 3, 4, 8, 8A), Red Edge (Bands 2, 3, 4, 5, 6, 7), and SWIR (Bands 2, 3, 4 11, 12) were compared to investigate the influence of spectral resolution. As displayed in Fig. 10, Red Edge bands were determined to be the most influential on classification accuracy, followed by SWIR and NIR. The results indicate that classification accuracy is proportional to the number of bands and amount of data in the input dataset. Based on this reasoning, the Red Edge bands may have scored a higher accuracy since the combination contains one more band compared to SWIR and NIR. Nevertheless, this observation also suggests that all the Red Edge bands were influential for LCZ classification. NIR bands exhibited the lowest OA, OA_{BU}, and F1, which may be due to the redundant spectral information in bands 8 and 8A. Yet the NIR bands recorded the highest OA_N, suggesting that NIR bands are particularly useful in identifying natural landscapes. SWIR bands recorded the lowest OA_N, which may explain their poor performance in previous studies (Qiu et al., 2018a,b).

6.3.2. Influence of individual ancillary bands

The influence of each ancillary band was compared using MSMLA-50 with P48 data. Each ancillary band was combined with the ten S2 bands to prepare input datasets of 11 bands and the results are shown in Fig. 11. OSM returned the highest accuracy for all metrics, followed by DSM and NLCM. Moreover, OSM attained the highest OA_{BU} of 73.90%, which is 5.51% higher than when using only S2 bands. This improvement can be linked to the usefulness of building footprint data in OSM, which has also been shown to improve LCZ classification accuracy in previous studies (Fonte et al., 2019). DSM and NLCM also helped increase OA_{BU} by a slight margin; however, only the addition of OSM helped improve OA_N with respect to the combination of all ten S2 bands, while DSM and NLCM decreased the accuracy. Despite the inclusion of

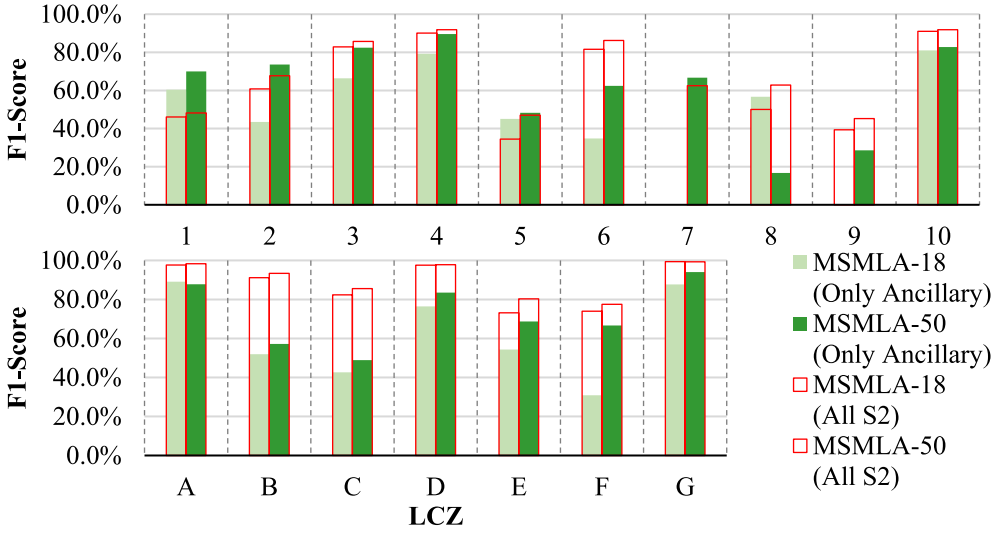

Fig. 12. Individual F1 scores for MSMLA-Net comparing "All S2" and "Only Ancillary" datasets at P48.

Table 10
MSMLA-Net when using a combination of S2 and ancillary bands with P48.

Model	OA (%)	WA (%)	OA_{BU} (%)	OA_{N} (%)	F1 (%)
MSMLA-18	85.2	81.9	66.6	89.8	84.6
MSMLA-50	**87.4**	**85.6**	**73.5**	**91.5**	**87.0**

height information, the addition of DSM yielded the lowest OA_{BU}, likely due to its coarse spatial resolution (30 m) in comparison to S2 images and the lack of height information of fine-grained features in highly dense areas. NLCM recorded a slightly higher OA_{BU} compared to DSM, suggesting that allocating a separate class for industrial regions as LCZ 10 may be beneficial. However, since LULC classes do not correspond directly with LCZ classes, more sophisticated ways to select and cluster specific categories in LULC map data should be considered for use in LCZ

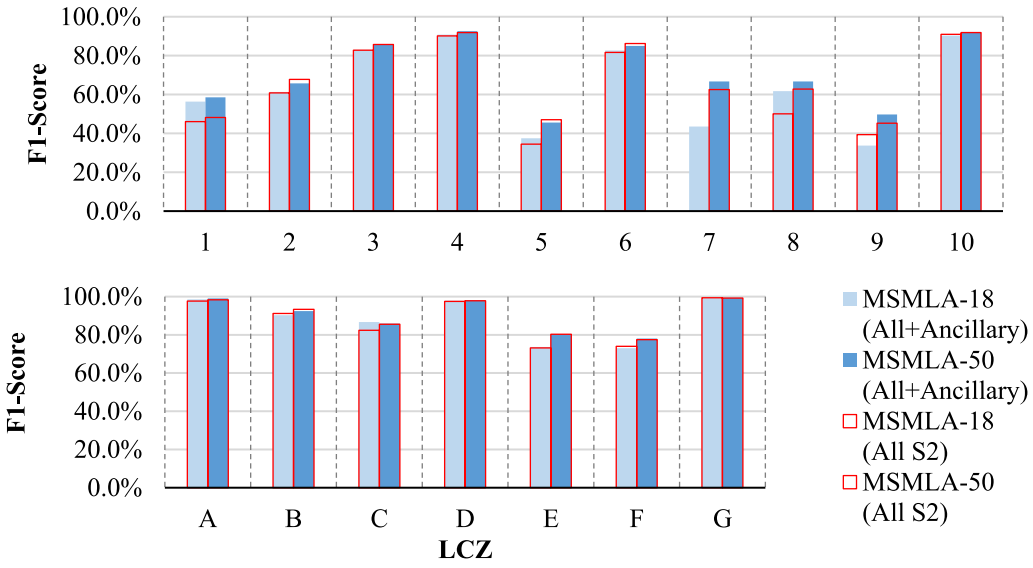

Fig. 13. Individual F1 scores for MSMLA-Net comparing "All S2" and "All + Ancillary" datasets at P48.

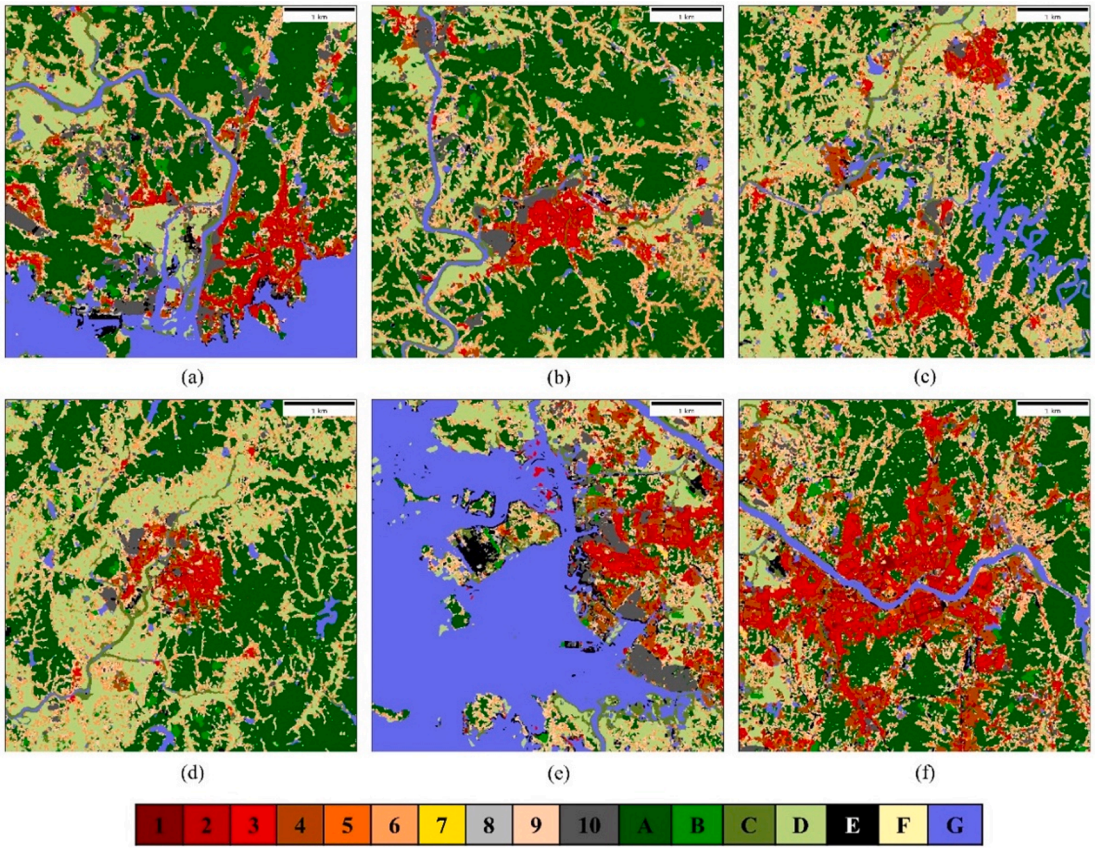

Fig. 14. LCZ maps were generated for the six cities via MSMLA-50 using "S2 All" input data at P48 under the "Standard" strategy. (a) Busan (2018/06/24), (b) Daegu (2018/06/09), (c) Daejeon (2018/08/01), (d) Gwangju (2018/09/25), Incheon (2018/09/25), (f) Seoul (2017/09/20).

Table 11
Classification accuracy results for SOTA and proposed models shown as an average of all cities.

Patch Size	Model	OA (%)	WA (%)	OA_{BU} (%)	OA_N (%)	F1 (%)
P32	Sen2LCZ	80.4	69.1	52.9	80.9	79.0
	LCZNet	80.6	69.0	51.4	81.3	79.9
	CNN	80.8	69.5	50.9	81.9	80.0
	MSMLA-18	80.5	69.5	53.3	79.9	79.8
	MSMLA-50	**82.0**	**71.1**	**54.4**	**82.5**	**81.5**
P48	Sen2LCZ	80.3	69.2	51.0	82.4	78.6
	LCZNet	80.4	68.1	48.4	82.3	79.9
	CNN	80.8	68.7	49.5	82.0	80.0
	MSMLA-18	80.7	70.2	52.0	82.2	79.8
	MSMLA-50	**82.4**	**70.9**	**52.8**	**83.9**	**81.7**

classification. For further accuracy enhancements, MABP or other high-quality building plans with elevation information, DSMs with higher resolution, and LiDAR data can be considered as higher quality ancillary data.

6.3.3. *Influence of only ancillary bands*

OSM, DSM, and NLCM were combined into one input dataset ("Only Ancillary") and trained again on both MSMLA-18 and MSMLA-50 with P48. This experiment aimed to determine the true capacity of information in the ancillary data and the results are shown in Table 9. MSMLA-50 returned higher OA_{BU} (75.0%) when using only ancillary bands compared to when using All S2 (72.4%) or any of the previous input band combinations shown above. On the one hand, this result suggests that some S2 bands may be superfluous for LCZ classification due to their coarse spatial resolution or unnecessary spectral information. On the other hand, using only ancillary bands severely reduces OA_N, highlighting the need for spectral information from S2 for natural LCZ classification. In addition, MSMLA-50's WA was found to be 3.47% greater than OA, suggesting the model's potential in classifying more difficult LCZ classes when using the three ancillary bands. This result highlights the effectiveness of ancillary data and the margin for improvement in LCZ classification accuracy, particularly for ambiguous built-up LCZ classes.

A closer inspection of F1 scores for each LCZ class in Fig. 12 revealed more insight behind the improved built-up accuracy and reduced natural accuracy when compared to the results from All S2. Results from all ten S2 bands ("All S2") were added as a comparison benchmark for "Only Ancillary". Built-up LCZ classes 1, 2, and 5 experienced improvements in F1, which agrees with MSMLA-50's higher WA result in Table 9. Furthermore, the deeper MSMLA-50 outperformed MSMLA-18 for all LCZ classes, except for LCZ classes 8 and A. MSMLA-18 had difficulty learning LCZ classes 7 and 9 due to the model's shallow depth and its inability to learn the scarce number of training samples. In contrast, MSMLA-50 recorded high accuracy for LCZ classes 1 to 4, suggesting the importance of deeper models to learn ancillary data more effectively for built-up LCZ classification. LCZ 1 returned the greatest growth in accuracy using "Only Ancillary", scoring an even higher F1 score than "All S2". However, the ancillary bands demonstrated a relatively smaller influence on natural LCZ classification across all classes, as evidenced by the low OA_{NA} in Table 9. This result can be attributed to the lack of spectral information in the ancillary data that are typically used to distinguish natural landscapes such as forests, crops, and water.

6.3.4. *Influence of all S2 and ancillary bands*

All ten S2 bands and the three ancillary bands were combined into a 13-band input dataset (All + Ancillary). The "All + Ancillary" dataset was trained on MSMLA-18 and MSMLA-50 for P48, and the results are shown in Table 10. In comparison to the results from "All S2" (Refer to "+SE + MS + MLA" in Table 7), the addition of the ancillary bands in "All + Ancillary" helped increase aggregated accuracy metrics for MSMLA-50. Only OA_N experienced a slight drop in accuracy. On the other hand, MSMLA-18 encountered a considerable drop across all aggregated metric results, since the model's shallow depth was unsuitable to learn the ancillary data. Furthermore, MSMLA-50 experienced a

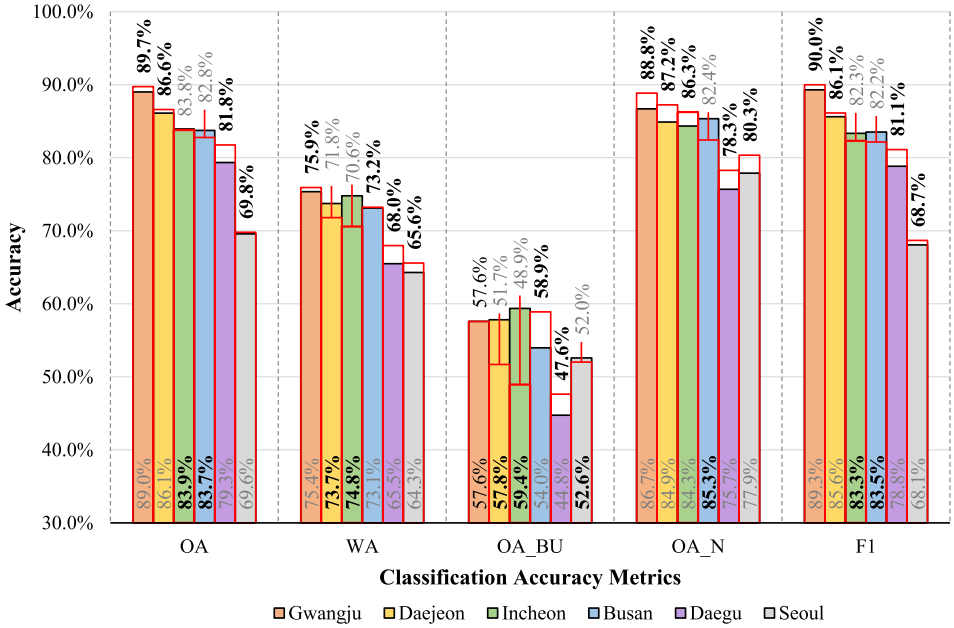

Fig. 15. Overall and class aggregated classification accuracy results from MSMLA-50 plotted for each city. Red outlines and text above the bars correspond to P32 results. Colored bars and text located at the base of the plot correspond to P48 results. For each city, the result with higher accuracy is highlighted in bold text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

strong increase in OA, WA, and F1 in comparison to the results from "Only Ancillary", which is likely due to the increase in OA_N from the inclusion of S2 bands. OA_{BU} decreased by 1.5%. This reduction in OA_{BU} suggests that not all input bands may be useful for built-up LCZ classification.

Fig. 13 displays the F1 scores of each LCZ class for "All S2" and "All + Ancillary" datasets with P48. The F1 scores of natural LCZ classes for both models using both datasets were found to be relatively similar, as shown by their high OA_N in Table 10. For built-up LCZ classes, LCZ classes 1 (Compact Highrise), 8 (Large Lowrise), 9 (Sparsely Built) experienced a notable increase in accuracy, with LCZ 1 undergoing the largest increase. MSMLA-18 also experienced increased accuracy for the same classes, albeit to a smaller extent.

6.4. LCZ classification in Korea

The "All S2" data at P48 was used to train MSMLA-50 based on the "Standard" sampling strategy to generate LCZ maps for the six cities as shown in Fig. 14. LCZ maps from only one of the acquired S2 images are shown in this study. The maps display how MSMLA-50 can classify the highly dense and heterogeneous regions in all the cities.

6.5. Citywise evaluation

The stringent sampling standards using the "Citywise" strategy ensured that the training dataset was composed of spatially and geographically non-overlapping samples and that the test dataset of each city was not included in the training dataset and had never been seen before by the trained model. Furthermore, this experiment was conducted to investigate the generalization ability of the classification models, especially when trained with imbalanced datasets. Classification results of the SOTA models for P32 and P48 are presented in Table 11.

Despite the difficult training environment, MSMLA-50 still maintained its superior performance and recorded the highest accuracy values across all metrics. Furthermore, MSMLA-50 recorded a significant improvement over MSMLA-18 for both path sizes, reinforcing the importance of a deeper SE-ResNet backbone when integrating MS and MLA. Compared to Sen2LCZ, MSMLA-50 improved for P32 (ΔOA: +1.6% and ΔF1: +2.5%) and P48 (ΔOA: +2.1% and ΔF1: +3.1%). MSMLA-18 recorded similar overall accuracy metrics (OA and F1) and higher OA_{BU} compared to the three SOTA models. Considering the lightweight structure of MSMLA-18, this result suggests the effectiveness of MS and MLA particularly for built-up LCZ classes.

In contrast to the "Standard" sampling strategy, the "Citywise" strategy results displayed lower accuracy results, which was also noted in previous studies that tested for model generalization (Yoo et al., 2019; Rosentreter et al., 2020). In addition, all models returned a minimum accuracy of 80.0% OA and approximately 79.0% F1 for both patches. Aside from MSMLA-50, the similar results in MSMLA-18 and the SOTA models were likely due to the strict sampling standards and the lack of training data for the models to learn the various LCZ classes in each city. This baseline accuracy was likely due to the easier classification of natural LCZ classes, as evidenced by the high OA_N. An interesting observation is that OA_{BU} tended to decrease and OA_N increase with larger input patch sizes. This result suggests that the additional contextual information in P48 may increase complexity and ultimately be detrimental for the model to classify fine-grained features such as areas with small buildings.

One concern is that WA results were lower by about 10% and only the proposed models MSMLA-18 and MSMLA-50 returned over 70% WA. The relatively lower WA compared to OA indicates the difficulty of classifying ambiguous LCZ classes particularly with less training data. Collecting more high-quality training samples is a potential solution but would require more manual work. Alternatively, large LCZ datasets like So2Sat can provide additional training data; however, transfer learning techniques may be required to compensate for the domain shift and potential bias in label operators (Liu and Shi, 2020; Kim et al., 2020). For generalization tasks using never before seen datasets such as with the "Citywise" sampling method, few-shot or zero-shot learning also offer potential solutions for effective LCZ classification (Qiu et al., 2020; Rußwurm et al., 2020).

Following the model comparison results, LCZ classification results for each city using MSMLA-50 are shown in Fig. 15. MSMLA-50 was selected because it was the only model that demonstrated a salient improvement in accuracy for all metrics. Gwangju was the easiest city to classify, reaching up to 89.7% OA and 90.0% F1, due to the city's large, homogeneous natural LCZ classes such as LCZ A (Dense Trees) and D (Low Plants). Daejeon also produced high classification accuracy, given the greater proportion of LCZ D in the S2 image. The remaining four cities have a bigger population and larger urban surface area. Incheon and Busan produced similar OA and F1 results, but the former produced higher WA and OA_{BU}. This result is due to the numerous large industrial complexes (LCZ E) and urban areas such as Incheon international airport in Incheon (LCZ 10). Conversely, Seoul was the most difficult because the megacity contains a much greater proportion of high-rise and highly dense features. Further, most samples of LCZ 1 (Compact Highrise) and 4 (Open Highrise) were inevitably collected from Seoul, and since the "Citywise" strategy omits samples from the city being classified, the lack of such high-rise LCZ samples severely reduced the OA and F1 scores for Seoul. However, Daegu suffered the lowest class aggregated accuracy, despite recording higher overall accuracies (OA and F1) than Seoul. The low accuracy result can be linked to the LCZ composition of Daegu, which consists of many training samples from LCZ classes 6 (Open Lowrise), 9 (Sparsely Built), and C (Bush, Scrub) which are often confused with one another (Refer to Table A1 in the Appendix).

6.6. Limitations and future steps

First, intra-class similarity for elevation-based built-up LCZ classes (LCZ 1 to 6) still proved difficult to differentiate. Even with the guidance of high-quality building information, more improvement is needed in built-up, heterogeneous regions. Other common intra-class confusion trends included LCZ classes 8 and 10 as well as 6 and 9, which were also prevalent in previous studies. Ground-based observations using GSV (Xu et al., 2019) or LiDAR measurements can be used to enhance classification accuracy in these areas. Second, the output LCZ maps were produced at 100 m; however, outputting LCZ maps at even finer spatial resolutions (50 m) may be considered for highly dense and complex regions (Yoo et al., 2020). As cities grow even denser and more heterogeneous, higher resolutions may be better suited to characterize the complexity and fine detail. From a model-based perspective, deeper and wider models can be explored if they are connected using efficient architecture to ensure the effective propagation of information. Third, the proportion of scarce classes such as LCZ 7 in the input dataset were relatively very low and tended to engender underestimated results. To overcome scarce LCZ classes, conducting more rigorous sampling is one option. Semi-supervised or meta-agnostic meta-learning of multi-modal input data can help to improve sampling procedures (Rußwurm et al., 2020). Fourth, since LCZ classes are not mutually disjoint, certain features may not be represented by the classification scheme. New subclasses may be required to properly characterize the features and their

local regions. Fifth, this study only used open-source data to increase the study's accessibility. Instead, commercial products such as high spatial resolution DSMs and very high resolution satellite images can provide height information at a finer spatial resolution to further improve classification accuracy (Collins and Dronova, 2019; Simanjuntak et al., 2019). The use of very high-resolution satellite imagery and object-based image analysis may help to improve LCZ classification at finer scales. In the case of Korea, the Korea Multi-Purpose Satellite (KOMPSAT) 3 and 3A satellites, as well as the Compact Advanced Satellite 500-1 and 500-2 missions, can be valuable resources to help improve LCZ classification.

7. Conclusion

Deep learning-based LCZ classification is progressing urban climate monitoring towards highly accurate and large-scale LCZ mapping. Instead of relying on global datasets, this study conducted supervised classification anew to generate high-quality LCZ training samples and produce highly accurate LCZ maps. For this purpose, this study developed MSMLA-Net as a lightweight and powerful LCZ classification model based on a modified SE-ResNet backbone, which integrates the MS module to exploit multi-scale features in concert with the MLA module for the multi-level context aggregation of spectrally and spatially enhanced features.

First, this study combined good practices of LCZ classification from recent studies. Highly precise building information from MABP, vegetation cover information, Google satellite, and GSV data were referenced to ensure the generation of high-quality LCZ training samples. The sampling methodologies and established training datasets can be used to extend LCZ classification studies using regional datasets. Second, MSMLA-Net was evaluated using an exhaustive ablation study, which compared the effect of SE blocks, MS, and MLA modules. Moreover, the study compared both P32 and P48 inputs, given their effective usage in recent studies. In addition, the MLA module presents an effective way to incorporate CBAM into LCZ classification models as a branched unit. Third, this study evaluated multiple SOTA models (Sen2LCZ, LCZNet, CNN) proposed in recent studies on deep learning-based LCZ classification. This study conducted comprehensive model and data-based experiments. From a model-based perspective, model depth, the addition of attention mechanisms, and the influence of MS and MLA were investigated. From a data-based viewpoint, this study conducted classification tests with various band combinations to find the influence and full potential of input bands. Numerous S2 bands (RGB, NIR, Red Edge, SWIR, All S2), individual ancillary data combinations, and the combination of all bands ("All + Ancillary") were compared. Fourth, this study also included two sampling methodologies ("Standard" and "Citywise") to verify model performance and robustness.

The proposed MSMLA-50 model returned the best classification results for virtually every case, outperforming all SOTA models for both patch sizes, all Sentinel-2 spectral band and ancillary band combinations, and both sampling strategies. MSMLA-50 recorded the best results when trained under the "Standard" sampling strategy with "All + Ancillary" data at P48, demonstrating the model's effectiveness in exploiting the additional contextual information for LCZ classification. In particular, MSMLA-50 enhanced OA_{BU} over 72.4% and 73.5% for "All S2" and "All + Ancillary" data at P48. Furthermore, integrating the MS and MLA modules generated a substantial improvement over ordinary computer vision models when supported with SE blocks as evidenced by the vast improvements in accuracy for both MSMLA-18 and MSMLA-50. Moreover, implementing CBAM with GAP as a branched unit via MLA was determined to be more effective as opposed to adding CBAM in sequence after every convolutional layer.

Data-based experiments were conducted using the best-performing MSMLA-50. Classification accuracy was found to be correlated with the number of bands in the input dataset. However, the NIR band combination was especially effective when classifying natural LCZ classes. Out of the three ancillary bands used in this study, OSM was the most influential, recording notable increases in OA_{BU} and OA_{N}. DSM and NLCM were less effective, but the 13-band combination of "All + Ancillary" data nonetheless produced the best accuracy results, as noted in the model-based experiments. To discover the full capacity of the ancillary data and verify that their influence was not solely driven by the additional increase in the amount of data, classification of only the three ancillary bands revealed the highest recorded OA_{BU} (75.0%) but at the cost of a low OA_{N}. Although this study only used open-source data to ensure the accessibility of the experiments and results, future studies may consider adding higher resolution ancillary data or higher-level products.

MSMLA-50 maintained its superior performance under more rigorous sampling conditions with the "Citywise" evaluation. Given that the perceived baseline accuracy for LCZ classification is around 80% OA, MSMLA-50 was the only model that yielded a considerable improvement in accuracy. LCZ maps for the six test cities in Korea displayed MSMLA-50's ability to interpret the high level of heterogeneity and identify small, fine-grained features. Nevertheless, the model suffered from common intra-class confusion, particularly with built-up LCZ classes 1 to 6. Given the continuous urbanization trends and development of more complex and compact cities, classifying ambiguous LCZ classes will be doubly important for higher resolution LCZs.

The methods and results produced in this study can be extended further. The proposed MSMLA-Net and the MS and MLA modules can be applied flexibly for future custom-made LCZ dataset in different regions. Also, the generated high-quality LCZ training samples and LCZ classification maps can be used to advance urban climate research in South Korea. While the scope of this study focuses on level-0 LCZ classification, on a higher level, accurate classification of LCZs can be instrumental to monitor urban microclimates and the everchanging climate in an effort to help target global initiatives pursued by the United Nations such as SDG11 (Sustainable Cities and Communities) and SDG13 (Climate Action).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2019R1I1A2A01058144). This research was supported by a grant (20009742) of Ministry-Cooperation R&D program of Disaster-Safety, funded by Ministry of Interior and Safety (MOIS, Korea). The Institute of Engineering Research at Seoul National University provided research facilities for this work.

Appendix

1. *Sampling methodology*

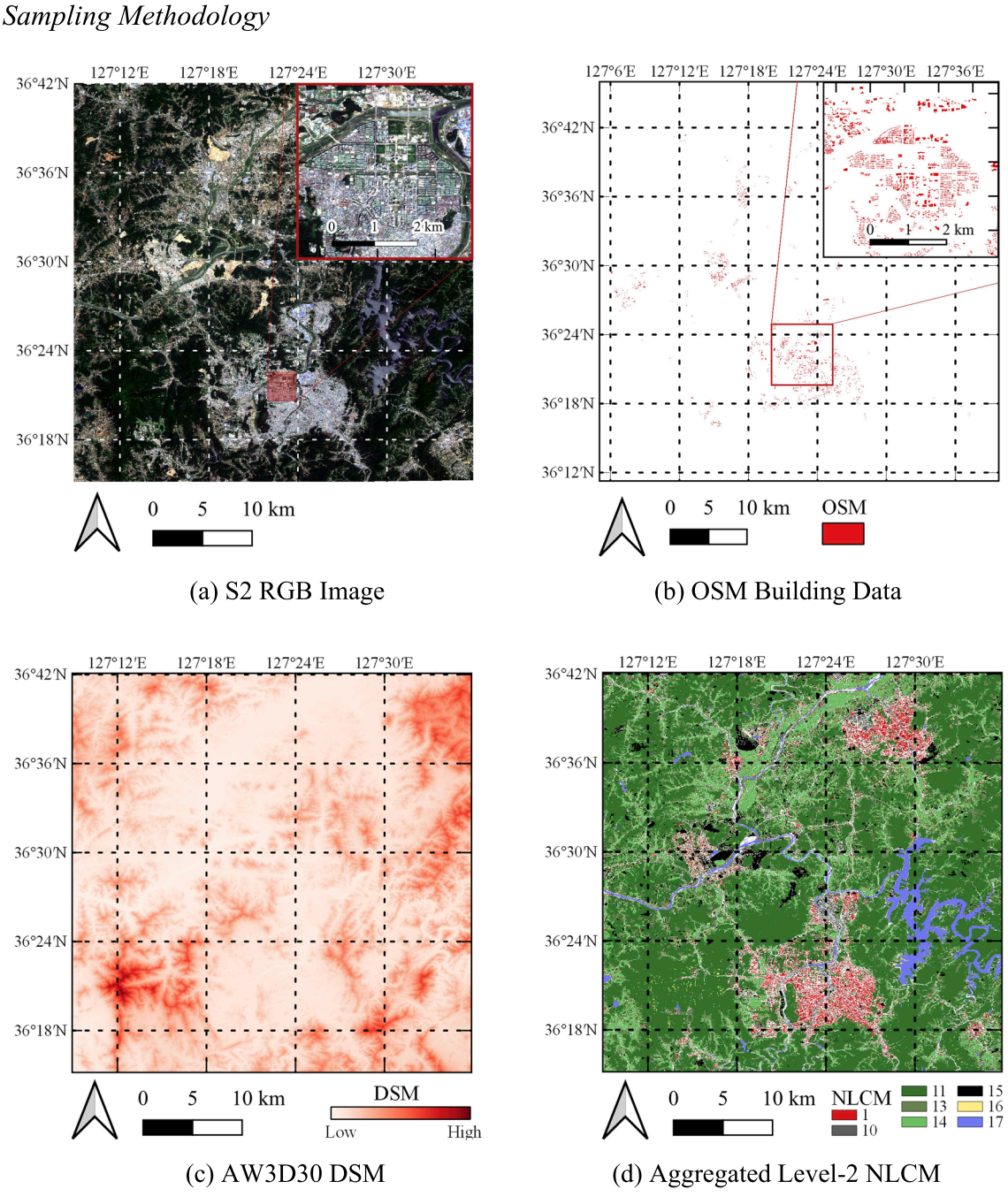

Fig. A1. Examples of ancillary data taken for Gwangju. OSM, DSM, and NLCM are used as actual inputs in the input dataset for model training.

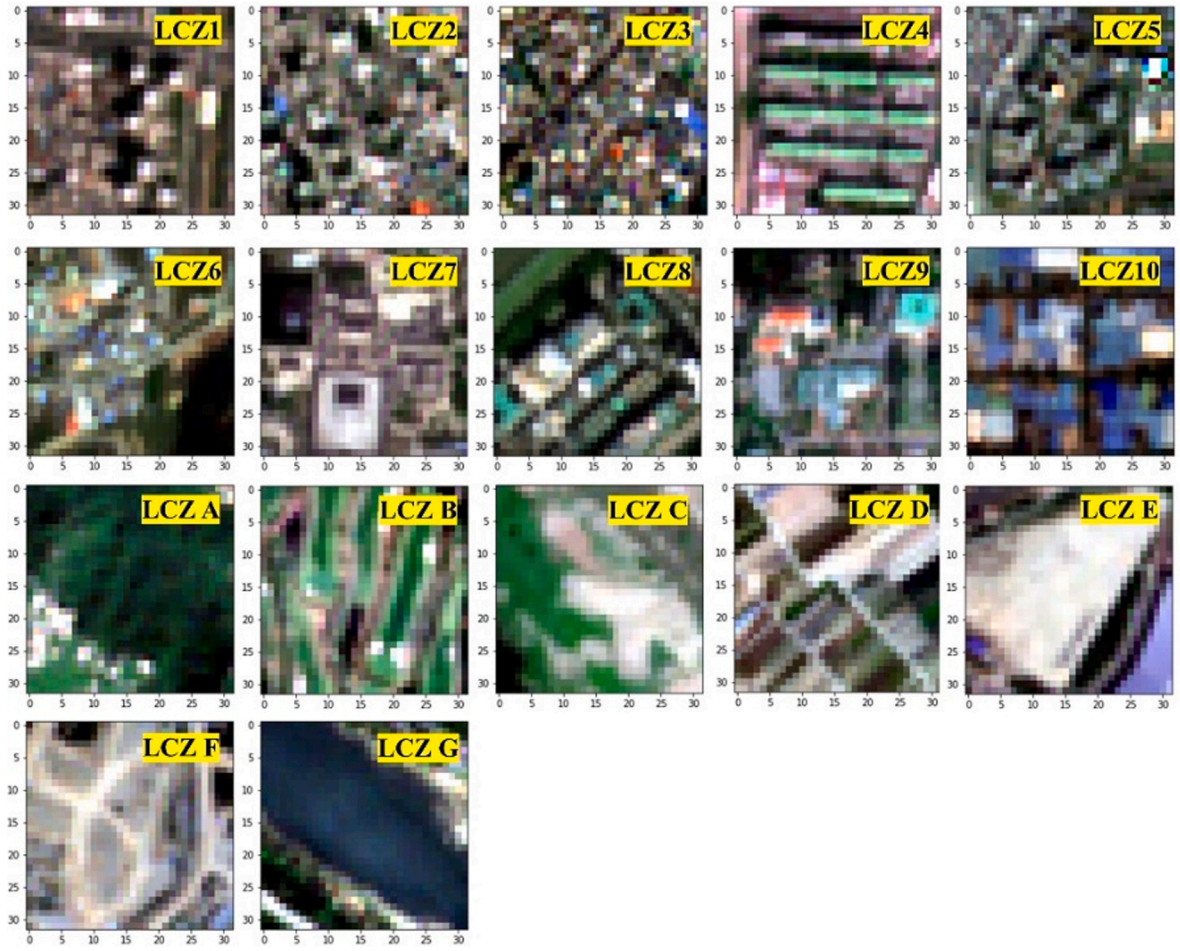

Fig. A2. 32 by 32-pixel-pixel patches of S2 images shown for each LCZ class as an example of the LCZ classification scheme for use in South Korea.

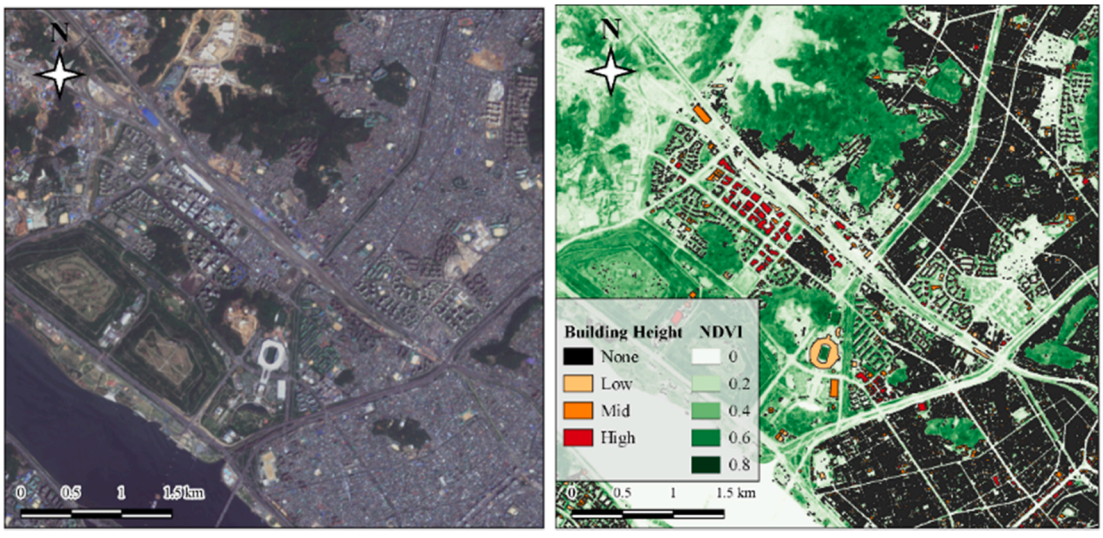

Fig. A3. Stacking of an S2 RGB image with NDVI and MABP to help manual labeling of LCZ sample points.

2. *LCZ training samples*

Table A1
Number of "point" LCZ samples per city organized for each LCZ class.

LCZ Class		Busan	Daegu	Daejeon	Gwangju	Incheon	Seoul	Total
1		54	7	7	4	22	338	432
2		224	140	177	184	121	300	1146
3		181	302	279	139	194	810	1905
4		405	361	449	252	400	870	2737
5		22	52	106	26	116	92	414
6		166	161	136	98	55	63	679
7		2	0	4	2	2	38	48
8		51	31	13	18	31	60	204
9		43	102	41	22	7	36	251
10		155	243	146	177	400	20	1141
A		152	430	266	252	203	309	1612
B		106	44	53	40	50	67	360
C		103	149	67	110	35	12	476
D		407	358	419	813	254	71	2322
E		167	54	24	31	101	137	514
F		97	46	43	21	91	60	358
G		288	141	82	77	270	259	1117
Total		**2623**	**2621**	**2312**	**2266**	**2352**	**3542**	**15716**

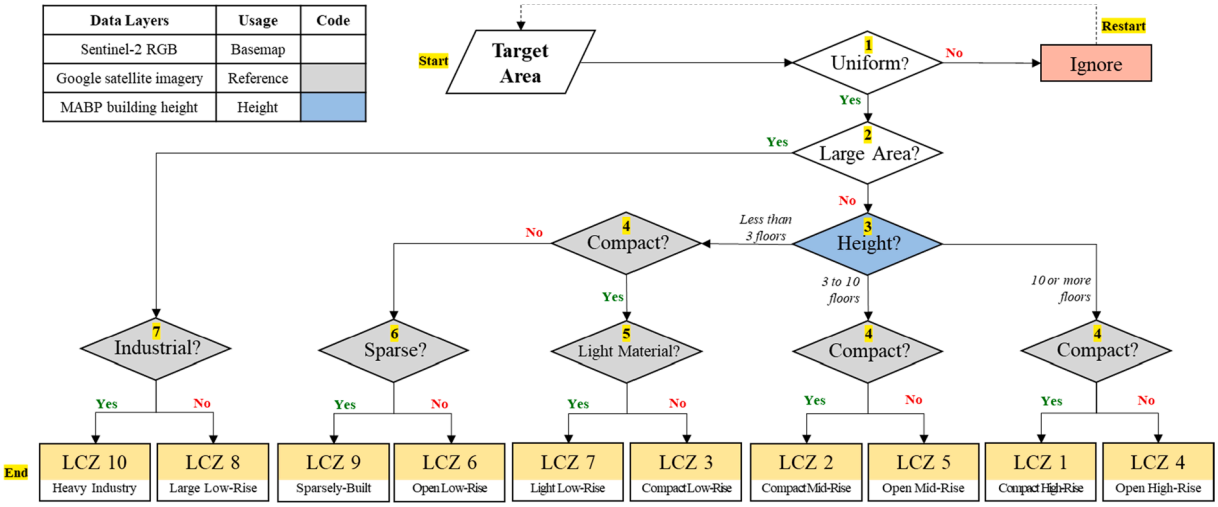

Fig. A4. Detailed decision rule workflow for built-up LCZ class labeling adapted from Zhu et al. (Zhu et al., 2020).

3. *Model experiments*

The workflow starts with the "Learning" phase. There are seven main decisions based on Zhu et al. (Zhu et al., 2020): (1) Uniform features recommended as at least five homogeneous 100 m resolution pixels; (2) Large building footprints; (3) Comparison of the number of building floors and height using MABP; (4) Compactness assessed based on building surface fraction of target scene; (5) Features built with light materials; (6) Sparsely distributed features based on compactness and perviousness; (7) Presence of industrial features and facilities. Building

surface fraction can be estimated by using a 100-by-100 m polygon in Google Earth and by comparing features according to the target patch size.

Table A2
Number of parameters for model variations used in the ablation studies.

Model	SE	MS	MLA	Parameters
ResNet18				185,137
+SE	✓			165,549
+SE + MS	✓	✓		166,269
+SE + MLA	✓		✓	166,961
+SE + MS + MLA	✓	✓	✓	181,867
ResNet50				562,785
+SE	✓			683,065
+SE + MS	✓	✓		686,089
+SE + MLA	✓		✓	791,703
+SE + MS + MLA	✓	✓	✓	808,913

Table A3
Comparison of accuracy results using various backbone models with P32. Results show that SE-ResNet50 reaches the highest accuracy results.

Model	OA	WA	UA	NA	F1
ResNet18	82.20%	77.38%	60.60%	87.36%	80.99%
ResNet34	81.95%	77.11%	60.48%	**87.37%**	80.70%
ResNet50	80.42%	74.87%	57.99%	84.97%	79.18%
ResNet101	78.52%	71.96%	54.51%	81.75%	76.74%
SE-ResNet18	81.95%	76.11%	58.75%	87.20%	80.69%
SE-ResNet34	81.98%	76.24%	59.05%	86.64%	80.43%
SE-ResNet50	**82.57%**	**80.77%**	**66.68%**	87.13%	**81.63%**
SE-ResNet101	81.23%	75.81%	58.34%	86.60%	79.84%

Table A4
Comparison of accuracy results using various backbone models with P48. Results show that SE-ResNet50 reaches the highest accuracy results.

Model	OA	WA	UA	NA	F1
ResNet18	82.80%	77.97%	61.99%	87.06%	81.65%
ResNet34	82.93%	77.69%	60.97%	87.76%	81.92%
ResNet50	81.46%	78.97%	63.09%	86.69%	80.40%
ResNet101	79.40%	72.87%	53.81%	83.62%	77.65%
SE-ResNet18	82.51%	77.19%	59.78%	87.85%	81.32%
SE-ResNet34	82.85%	77.21%	60.13%	88.04%	81.85%
SE-ResNet50	**83.31%**	**80.99%**	**66.95%**	**88.37%**	**82.32%**
SE-ResNet101	82.54%	78.42%	60.17%	88.24%	81.58%

References

Agnes, S. Akila, Anitha, J., Pandian, S. Immanuel Alex, Peter, J. Dinesh, 2020. Classification of mammogram images using multiscale all convolutional neural network (MA-CNN). J. Med. Syst. 44 (1) https://doi.org/10.1007/s10916-019-1494-z.

Bechtel, Benjamin, 2011. Multitemporal Landsat data for urban heat island assessment and classification of local climate zones. 2011 Joint Urban Remote Sensing Event. IEEE.

Bechtel, Benjamin, Alexander, Paul, Böhner, Jürgen, Ching, Jason, Conrad, Olaf, Feddema, Johannes, Mills, Gerald, See, Linda, Stewart, Iain, 2015. Mapping local climate zones for a worldwide database of the form and function of cities. ISPRS Int. J. Geo-Inf. 4 (1), 199–219.

Bechtel, Benjamin, See, Linda, Mills, Gerald, Foley, Micheal, 2016. Classification of local climate zones using SAR and multispectral data in an arid environment. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (7), 3097–3105.

Bechtel, Benjamin, Demuzere, Matthias, Sismanidis, Panagiotis, Fenner, Daniel, Brousse, Oscar, Beck, Christoph, Van Coillie, Frieke, Conrad, Olaf, Keramitsoglou, Iphigenia, Middel, Ariane, Mills, Gerald, Niyogi, Dev, Otto, Marco, See, Linda, Verdonck, Marie-Leen, 2017. Quality of crowdsourced data on urban morphology—the human influence experiment (HUMINEX). Urban Sci. 1 (2), 15. https://doi.org/10.3390/urbansci1020015.

Bechtel, Benjamin, Alexander, Paul J., Beck, Christoph, Böhner, Jürgen, Brousse, Oscar, Ching, Jason, Demuzere, Matthias, Fonte, Cidália, Gál, Tamás, Hidalgo, Julia, Hoffmann, Peter, Middel, Ariane, Mills, Gerald, Ren, Chao, See, Linda, Sismanidis, Panagiotis, Verdonck, Marie-Leen, Xu, Guang, Xu, Yong, 2019. Generating WUDAPT level 0 data–current status of production and evaluation. Urban Clim. 27, 24–45.

Bechtel, Benjamin, Daneke, Christian, 2012. Classification of local climate zones based on multiple earth observation data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5 (4), 1191–1202.

Buslaev, Alexander, Iglovikov, Vladimir I., Khvedchenya, Eugene, Parinov, Alex, Druzhinin, Mikhail, Kalinin, Alexandr A., 2020. Albumentations: fast and flexible image augmentations. Information 11 (2), 125. https://doi.org/10.3390/info11020125.

Cheng, Wensheng, Yang, Wen, Wang, Min, Wang, Gang, Chen, Jinyong, 2019. Context aggregation network for semantic labeling in aerial images. Remote Sens. 11 (10), 1158. https://doi.org/10.3390/rs11101158.

Collins, Jed, Dronova, Iryna, 2019. Urban landscape change analysis using local climate zones and object-based classification in the Salt Lake Metro Region, Utah, USA. Remote Sens. 11 (13), 1615.

Demuzere, Matthias, Bechtel, Benjamin, Mills, Gerald, 2019. Global transferability of local climate zone models. Urban Clim. 27, 46–63.

DESA, U., 2018. Revision of world urbanization prospects. Population Division of the UN Department of Economic and Social Affairs, UN, New York. https://population.un.org/wup.

Fonte, Cidália, Lopes, Patrícia, See, Linda, Bechtel, Benjamin, 2019. Using OpenStreetMap (OSM) to enhance the classification of local climate zones in the framework of WUDAPT. Urban Clim. 28, 100456.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Heaviside, Clare, Macintyre, Helen, Vardoulakis, Sotiris, 2017. The urban heat island: implications for health in a changing environment. Curr. Environ. Health Rep. 4 (3), 296–305.

Hu, Jie, Shen, Li, Sun, Gang, 2018. Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Kim, Kwon, Eum, Jeong-Hee, 2017. Classification of local climate zone by using WUDAPT protocol-a case study of Seoul, Korea. J. Korean Inst. Landscape Arch. 45 (4), 131–142.

Li, D., Bou-Zeid, E., 2013. Synergistic interactions between urban heat islands and heat waves: the impact in cities is larger than the sum of its parts. J. Appl. Meteorol. Climatol. 52 (9), 2051–2064.

Liu, Shengjie, Shi, Qian, 2020. Local climate zone mapping as remote sensing scene classification using deep learning: a case study of metropolitan China. ISPRS J. Photogramm. Remote Sens. 164, 229–242.

Mills, Gerald, 2007. Cities as agents of global change. Int. J. Climatol.: J. Royal Meteorol. Soc. 27 (14), 1849–1857.

Kim, M., Jeong, D., Choi, H., & Kim, Y., 2020. Developing High Quality Training Samples for Deep Learning Based Local Climate Zone Classification in Korea. *arXiv preprint arXiv:2011.01436*.

Oke, T.R., 2004. Initial guidance to obtain representative meteorological observations at urban sites.

Mills, Gerald, Bechtel, Benjamin, Ching, Jason, See, Linda, Feddema, Johan, Foley, Michael, Alexander, Paul, O'Connor, Martin, 2015. An introduction to the WUDAPT project. Proceedings of the 9th International Conference on Urban Climate.

Pedregosa, F., et al., 2011. Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Qiu, Chunping, Schmitt, Michael, Ghamisi, Pedram, Mou, Lichao, Zhu, Xiao Xiang, 2018. Feature importance analysis of Sentinel-2 imagery for large-scale urban local climate zone classification. IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE.

Qiu, Chunping, Schmitt, Michael, Mou, Lichao, Ghamisi, Pedram, Zhu, Xiao, 2018. Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets. Remote Sens. 10 (10), 1572. https://doi.org/10.3390/rs10101572.

Qiu, Chunping, Mou, Lichao, Schmitt, Michael, Zhu, Xiao Xiang, 2019. Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network. ISPRS J. Photogramm. Remote Sens. 154, 151–162.

Qiu, Chunping, Tong, Xiaochong, Schmitt, Michael, Bechtel, Benjamin, Zhu, Xiao Xiang, 2020. Multilevel feature fusion-based CNN for local climate zone classification from sentinel-2 images: benchmark results on the So2Sat LCZ42 dataset. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 2793–2806.

Qiu, C., Tong, X., Schmitt, M., Bechtel, B., Zhu, X.X., 2020. Multilevel feature fusion-based CNN for local climate zone classification from sentinel-2 images: Benchmark results on the So2Sat LCZ42 dataset. IEEE J. Select.Topics Appl. Earth Observ. Remote Sens. 13, 2793–2806.Chicago.

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 Sens. Environ. 237, 111472.

Rußwurm, Marc, Wang, Sherrie, Korner, Marco, Lobell, David, 2020. Meta-Learning for Few-Shot Land Cover Classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Schmitt, Michael, Hughes, Lloyd, Qiu, Chunping, Zhu, Xiao Xiang, 2019. Aggregating cloud-free sentinel-2 images with google earth engine. ISPRS Ann. Photogram. Remote Sens. Spatial Inform. Sci. 4.

Simanjuntak, Royger M., Kuffer, Monika, Reckien, Diana, 2019. Object-based image analysis to map local climate zones: The case of Bandung, Indonesia. Appl. Geogr. 106, 108–121.

Stewart, I., Oke, T., 2009. Classifying urban climate field sites by "local climate zones": The case of Nagano, Japan. IN: Seventh International Conference on Urban Climate.

Stewart, I.D., Oke, T.R., 2010. Thermal differentiation of local climate zones using temperature observations from urban and rural field sites. Ninth Symposium on Urban Environment.

Stewart, I.D., Oke, T.R., 2012. Local climate zones for urban temperature studies. Bull. Am. Meteorol. Soc. 93 (12), 1879–1900.

Stewart, I.D., Oke, T.R., Krayenhoff, E.S., 2014. Evaluation of the 'local climate zone'scheme using temperature observations and model simulations. Int J. Climatol. 34 (4), 1062–1080.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. February). Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Thirty-first AAAI conference on artificial intelligence*. AAAI.

Tadono, T., Nagai, H., Ishida, H., Oda, F., Naito, S., Minakawa, K., Iwamoto, H., 2016. Generation of the 30 M-mesh global digital surface model by ALOS PRISM. Int. Arch. Photogram. Remote Sens. Spatial Inf. Sci. 41.

Tan, Jianguo, Zheng, Youfei, Tang, Xu, Guo, Changyi, Li, Liping, Song, Guixiang, Zhen, Xinrong, Yuan, Dong, Kalkstein, Adam J., Li, Furong, Chen, Heng, 2010. The urban heat island and its impact on heat waves and human health in Shanghai. Int. J. Biometeorol. 54 (1), 75–84.

Uuemaa, Evelyn, Ahi, Sander, Montibeller, Bruno, Muru, Merle, Kmoch, Alexander, 2020. Vertical accuracy of freely available global digital elevation models (ASTER, AW3D30, MERIT, TanDEM-X, SRTM, and NASADEM). Remote Sens. 12 (21), 3482. https://doi.org/10.3390/rs12213482.

Verdonck, Marie-Leen, Okujeni, Akpona, van der Linden, Sebastian, Demuzere, Matthias, De Wulf, Robert, Van Coillie, Frieke, 2017. Influence of neighbourhood information on 'Local Climate Zone'mapping in heterogeneous cities. Int. J. Appl. Earth Obs. Geoinf. 62, 102–113.

Wang, Ran, Ren, Chao, Xu, Yong, Lau, Kevin Ka-Lun, Shi, Yuan, 2018. Mapping the local climate zones of urban areas by GIS-based and WUDAPT methods: a case study of Hong Kong. Urban Clim. 24, 567–576.

Woo, Sanghyun, Park, Jongchan, Lee, Joon-Young, Kweon, In So, 2018. Cbam: Convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV).

Xu, Guang, Zhu, Xuan, Tapper, Nigel, Bechtel, Benjamin, 2019. Urban climate zone classification using convolutional neural network and ground-level images. Prog. Phys. Geogr.: Earth Environ. 43 (3), 410–424.

Yoo, Cheolhee, Lee, Yeonsu, Cho, Dongjin, Im, Jungho, Han, Daehyeon, 2020. Improving local climate zone classification using incomplete building data and sentinel 2 images based on convolutional neural networks. Remote Sens. 12 (21), 3552.

Yoo, Cheolhee, Han, Daehyeon, Im, Jungho, Bechtel, Benjamin, 2019. Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. ISPRS J. Photogramm. Remote Sens. 157, 155–170.

Zhang, Guichen, Ghamisi, Pedram, Zhu, Xiao Xiang, 2019. Fusion of heterogeneous earth observation data for the classification of local climate zones. IEEE Trans. Geosci. Remote Sens. 57 (10), 7623–7642.

Zhang, Jing, Lin, Shaofu, Ding, Lei, Bruzzone, Lorenzo, 2020. Multi-scale context aggregation for semantic segmentation of remote sensing images. Remote Sens. 12 (4), 701. https://doi.org/10.3390/rs12040701.

Zhou, Xilin, Okaze, Tsubasa, Ren, Chao, Cai, Meng, Ishida, Yasuyuki, Mochida, Akashi, 2020. Mapping local climate zones for a Japanese large city by an extended workflow of WUDAPT Level 0 method. Urban Clim. 33, 100660.

Zhu, Xiao Xiang, Hu, Jingliang, Qiu, Chunping, Shi, Yilei, Kang, Jian, Mou, Lichao, Bagheri, Hossein, Haberle, Matthias, Hua, Yuansheng, Huang, Rong, Hughes, Lloyd, Li, Hao, Sun, Yao, Zhang, Guichen, Han, Shiyao, Schmitt, Michael, Wang, Yuanyuan, 2020. So2Sat LCZ42: a benchmark data set for the classification of global local climate zones [Software and Data Sets]. IEEE Geosci. Remote Sens. Mag. 8 (3), 76–89.