remote sensing

MDPI

Article

SAR and Multi-Spectral Data Fusion for Local Climate Zone Classification with Multi-Branch Convolutional Neural Network

Guangjun He [1], Zhe Dong [2], Jian Guan [3], Pengming Feng [1], Shichao Jin [1] and Xueliang Zhang [4,*]

[1] State Key Laboratory of Space-Ground Integrated Information Technology, Space Star Technology Co., Ltd., Beijing 100095, China
[2] School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China
[3] Group of Intelligent Signal Processing, College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
[4] School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China
* Correspondence: zxl@nju.edu.cn

Abstract: The local climate zone (LCZ) scheme is of great value for urban heat island (UHI) effect studies by providing a standard classification framework to describe the local physical structure at a global scale. In recent years, with the rapid development of satellite imaging techniques, both multi-spectral (MS) and synthetic aperture radar (SAR) data have been widely used in LCZ classification tasks. However, the fusion of MS and SAR data still faces the challenges of the different imaging mechanisms and the feature heterogeneity. In this study, to fully exploit and utilize the features of SAR and MS data, a data-grouping method was firstly proposed to divide multi-source data into several band groups according to the spectral characteristics of different bands. Then, a novel network architecture, namely Multi-source data Fusion Network for Local Climate Zone (MsF-LCZ-Net), was introduced to achieve high-precision LCZ classification, which contains a multi-branch CNN for multi-modal feature extraction and fusion, followed by a classifier for LCZ prediction. In the proposed multi-branch structure, a split–fusion-aggregate strategy was adopted to capture multi-level information and enhance the feature representation. In addition, a self channel attention (SCA) block was introduced to establish long-range spatial and inter-channel dependencies, which made the network pay more attention to informative features. Experiments were conducted on the So2Sat LCZ42 dataset, and the results show the superiority of our proposed method when compared with state-of-the-art methods. Moreover, the LCZ maps of three main cities in China were generated and analyzed to demonstrate the effectiveness of our proposed method.

Keywords: SAR; multi-spectral; data fusion; local climate zone; multi-branch CNN

Citation: He, G.; Dong, Z.; Guan, J.; Feng, P.; Jin, S.; Zhang, X. SAR and Multi-Spectral Data Fusion for Local Climate Zone Classification with Multi-Branch Convolutional Neural Network. *Remote Sens.* **2023**, *15*, 434. https://doi.org/10.3390/rs15020434

Academic Editors: Lionel Bombrun and Domenico Velotto

Received: 12 October 2022
Revised: 21 December 2022
Accepted: 6 January 2023
Published: 11 January 2023

1. Introduction

The local climate zone (LCZ) scheme, originally proposed to provide an interdisciplinary taxonomy, plays an important role in urban heat island (UHI) effect studies [1]. With the deepening of research, it can also provide a foundation for other urban-oriented studies, such as population density estimation, economic development monitoring [2], urban climatology [3], infrastructure planning [4], navigation application of disaster mitigation [5], etc. Recently, with the resolution of satellite images becoming higher, and the swath becoming wider, LCZ classification in remote sensing images has attracted more attention due to its broad vision in urban-oriented applications [6]. The 17 LCZ classes, including 10 urban types and 7 natural types, as shown in Figure 1 [7], have been developed to document climate-related surface properties at a local scale in the studies of the urban environment.

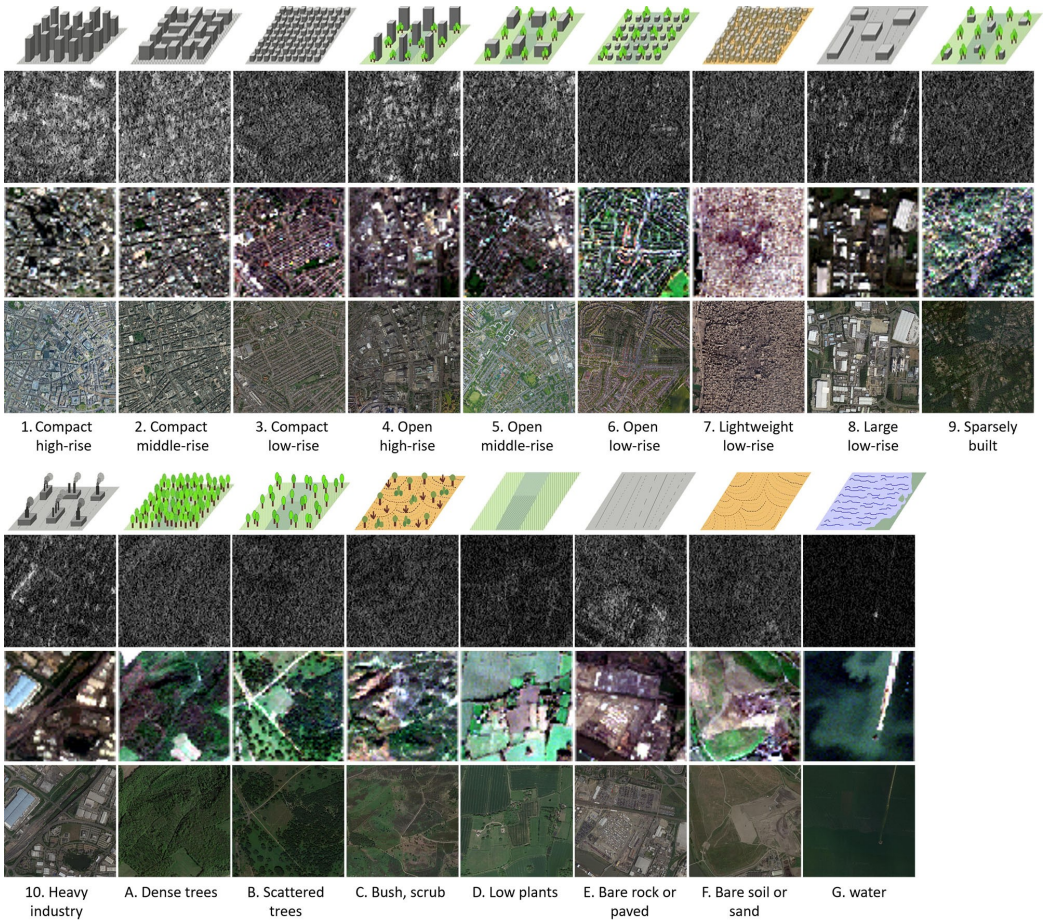


Figure 1. Visualization examples of Sentinel-1 and Sentinel-2 image scenes of the 17 LCZ classes, where the top row is the intensity (in dB) of the Sentinel-1 scene, the middle row is the corresponding Sentinel-2 scene in RGB, and the bottom row is the high-resolution image from Google Earth as a reference.

To obtain large-scale LCZ maps for urban-oriented studies, multi-spectral (MS) data and synthetic aperture radar (SAR) remote sensing data have attracted much attention [8] due to their large imaging width. MS images are known to contain information with reflective and missive characteristics that can provide rich details, which makes MS imageries relatively easier to interpret [9]. However, MS images are often influenced by the presence of clouds or atmospheric conditions when imaging, resulting in difficulty meeting the temporal requirements for LCZ classification [10]. Different to MS, SAR imaging is sensitive to the backscatter of terrain and object characteristics (e.g., target structure, geometry, and material characteristics), and has the characteristic of a coherent imaging capability (both amplitude and phase signals) [11]. In addition, SAR can provide the source of illumination with longer wavelengths that can penetrate through all weather conditions and collect images at any time of the day. However, due to the limited band numbers as well as the effects caused by speckle noises, slant-range imaging, foreshortening, layover, and shadows, SAR images are not intuitive for interpreting and are relatively difficult to interpret when compared to MS images [12].

Considering the complementarity of MS and SAR data, multi-source data fusion has been applied for LCZ classification [13]. Nevertheless, due to the significant differences between the imaging mechanisms of SAR and MS images, when fused by using the simple concatenation methods, the gray value differences between MS and SAR images are obvious [14,15] and cause substantial color distortion in the fused images [14]. In other words, available information in MS and SAR images is hardly developed thoroughly

when dealing directly with LCZ classification due to the characteristics of the two different physical imaging mechanisms. Therefore, how to better fuse MS data and SAR data and eliminate the difference between the two extracted features is of great importance for LCZ classification.

In order to solve the problems mentioned above, we designed a novel data-grouping method to arrange the multi-source data according to the guidance of band characteristics and an MsF-LCZ-Net framework for high-precision LCZ classification, where a multi-branch convolution neural network (CNN) structure is proposed for feature extraction and feature fusion from MS and SAR data, followed by a self-attention-based classifier to enhance the feature aggregation and combine spatial and channel attention efficiently.

In summary, the contributions of this work can be listed as follows:

(1) A data-grouping strategy is proposed to arrange the fusion of MS and SAR data into band groups according to spectral characteristics, achieving a sufficient fusion of multi-source data.
(2) A multi-branch fusion CNN is proposed to perform LCZ classification, where a multi-branch structure is introduced to extract and fuse features, with residual learning and self channel attention combined into the proposed classifier to accomplish the LCZ mapping.
(3) We conducted experiments on the So2Sat LCZ42 dataset as well as real scenarios, and the experimental results demonstrate the effectiveness and robustness of our proposed method.

The remainder of the paper is organized as follows: Section 2 reviews the related works. Section 3 presents the details of the proposed method. Experiments are conducted in Section 4 to demonstrate the effectiveness of our method, and Section 5 gives the performance analysis for LCZ classification. Finally, we conclude this paper and give the potential future work in Section 6.

2. Related Work

As a unique classification scheme, LCZ mapping provides a large scale of essential information for urban-oriented studies. Due to their effectiveness and accessibility, MS data have become one of the most widely used data sources for LCZ mapping. In order to make full use of the temporal and spectral information, multi-season Sentinel-2 images were utilized to perform LCZ classification with a ResNet-based [16] architecture [17]. Moreover, a majority voting strategy was applied to make an accuracy improvement. In [18], a simple and lightweight model named Sen2LCZ-Net-MF was proposed for large-scale LCZ classification by fusing multi-level features, and their experiments proved that multi-level fusion can make the lightweight model achieve a better performance than relative heavyweight models, whereas, in [19], authors presented a parcel-based idea to map LCZ from Sentinel-2 images, which took the shape and size of the real footprint of LCZs into consideration to improve the mapping accuracy. In addition, considering the scale variance problem, multi-scale features from different channels were extracted and fused to achieve feasible LCZ classification [20].

However, due to the influence from weather and the absence of height information, when MS images are applied to perform LCZ classification, classes with spectral characteristics similar to others are not properly classified [21]. Though this problem could be potentially solved by SAR images, it would result in a disappointing LCZ classification performance because of information loss caused by speckle noises.

Taking into account the complementarity of MS and SAR images, multi-source data fusion has proved to be a promising method for improving the stability and accuracy of LCZ classification. In [22], the authors examined the effect of merging SAR data for generating LCZs, and the experimental results showed that combining MS and SAR data can improve the overall performance. In [23], with the help of the Google Earth Engine (GEE) platform and freely available datasets, the authors demonstrated that multi-source data fusion can help to generate more accurate LCZ mapping on a large scale. In [24],

CNNs were used to extract and fuse features from MS and SAR images to perform LCZ classification. However, these methods ignore the differences in heterogeneous data and the correlation between bands in MS and SAR data. In our previous work [25,26], we adopted an embranchment CNN and dynamic filter network to integrate different bands in SAR and MS data, which could enhance the fusion performance in LCZ classification, but how to combine the advantages of MS and SAR images and mitigate the difference between heterogeneous features is still the challenge in LCZ classification.

3. Methodology

The framework of the proposed MsF-LCZ-Net is shown in Figure 2. Our LCZ classification framework followed a pipeline approach. Firstly, when dealing with MS and SAR data, we designed a data-grouping method to generate independent band groups. Then, all of these band groups were transported to a multi-branch model, which consisted of feature extraction branches and feature fusion layers. Finally, we performed LCZ classification using the fused features.

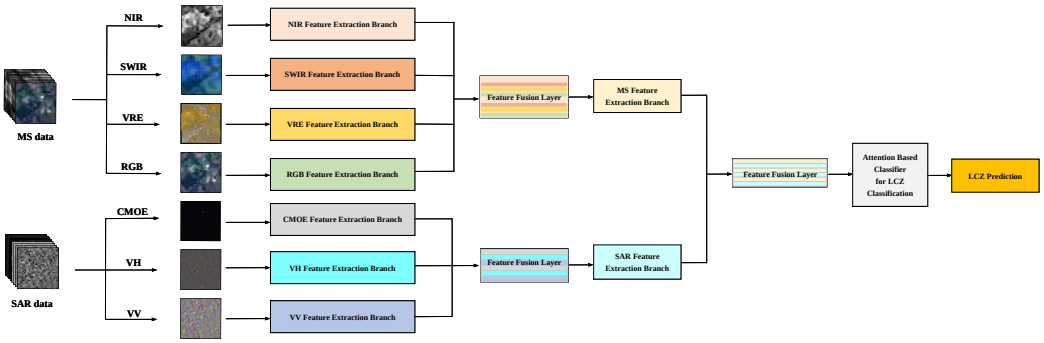


Figure 2. The overall flowchart of the proposed method (MsF-LCZ-Net).

In this work, SAR and MS data from Sentinel-1 and Sentinel-2, respectively, were employed for LCZ classification. Specifically, following [7], the VV-VH dual polarization single-look complex (SLC) Level-1 data from Sentinel-1 were used. After applying a series of pre-processing steps, such as radiometric calibration polarimetric speckle reduction and terrain correction, eight real-valued bands were obtained from the Sentinel-1 data. On the other hand, cloud-free Sentinel-2 images were obtained from Google Earth Engine (GEE), which contain bands B2, B3, B4, and B8 with a 10 m ground sampling distance (GSD), bands B5, B6, B7, B8a, B11, and B12 with a 20 m GSD, and bands B1, B9, and B10 with a 60 m GSD. Since bands B1, B9, and B10 mainly contain data related to the atmosphere, which bear little relevance to LCZ classification, they were not considered in this work. In this way, the employed Sentinel-2 data consisted of 10 real-valued bands. To keep the consistency of Sentinel-2 data, the 20 m GSD bands were up-sampled to a 10 m GSD using the bilinear interpolation method. The details of band information are described in Table 1.

For SAR data, different polarization modes show various sensitivities to the various types of ground surface. Similarly, there are different dependencies between the spectral bands of MS data. To extract features from multi-source data more effectively and facilitate the feature fusion, SAR data were divided into three band groups: VV, VH, and a covariance matrix off-diagonal element (CMOE), according to polarization methods. Similarly, we divided the MS data into four band groups based on the correlation between the bands. Bands B2, B3, and B4 were grouped into RGB to reflect the color of the ground surface. Bands B5, B6, B7, and B8a were grouped as vegetation red edge (VRE) to detect vegetation information. Bands B11 and B12 were combined as the group of short-wave infrared (SWIR), which is less susceptible to the influence of clouds. In addition, band 8 was regarded as the group of near-infrared (NIR) to detect water on the ground. The grouping situation is shown in Table 2.

Table 1. Selected features from Sentinel-1 (band 1–8) and Sentinel-2 (band 9–18) data.

Number	Band	Number	Band
1st band	Real part of the unfiltered VH channel	9th band	B2
2nd band	Imaginary part of the unfiltered VH channel	10th band	B3
3rd band	Real part of the unfiltered VV channel	11th band	B4
4th band	Imaginary part of the unfiltered VV channel	12th band	B5 (upsampled to 10 m from 20 m GSD)
5th band	Intensity of the refined Lee-filtered VH signal	13th band	B6 (upsampled to 10 m from 20 m GSD)
6th band	Intensity of the refined Lee-filtered VV signal	14th band	B7 (upsampled to 10 m from 20 m GSD)
7th band	Real part of the refined Lee-filtered covariance matrix off-diagonal element	15th band	B8
		16th band	B8a (upsampled to 10 m from 20 m GSD)
8th band	Imaginary part of the refined Lee-filtered covariance matrix off-diagonal element	17th band	B11 (upsampled to 10 m from 20 m GSD)
		18th band	B12 (upsampled to 10 m from 20 m GSD)

Table 2. Band-grouping situation.

Group	Band	Group	Band
VV	1st band 2nd band 5th band	RGB	9th band 10th band 11th band
VH	3rd band 4th band 6th band	VRE	12th band 13th band 14th band 16th band
CMOE	7th band 8th band	SWIR	17th band 18th band
		NIR	15th band

3.1. Multi-Branch CNN for Feature Fusion

To fully utilize the complementary information of MS and SAR data while alleviating their disadvantages, the feature-level fusion strategy was employed. Since different band groups contain specific spatial and spectral information, a split–fusion-aggregate strategy was adopted to strengthen the expression of characteristics. Specifically, a multi-branch structure was proposed to extract local low-level features from each band group. The fusion layers were then applied to combine different hierarchical features into a higher-order feature set, thus achieving the ability of learning the complementary relationship between different features.

As shown in Figure 3, the proposed multi-branch CNN consisted of seven shallow feature extraction branches, two deep feature extraction branches, and three feature fusion layers. In detail, the shallow branches were employed to capture low-level features from each band group, which contain rich texture and boundary information. Taking a three-band input (RGB) as an example, the corresponding feature extraction branch structure is shown in Figure 3a. Note that the feature extraction branches for other groups share the same structure. It can be seen that a 3×3 convolution was firstly employed to linearly combine pixels on different channels of the RGB band group. Then, a double-branch feature extractor (FE) block (Figure 3c) was designed to obtain multi-scale features by applying down-sample operations with different depths and kernel sizes. In addition, referring to the residual learning idea, a shortcut connection was added between the first convolution layer and ReLU so that the gradients learned from backpropagation could be conveyed efficiently, hence easing the training process. The residual learning behind ResNet [16] aims to make the shallower architecture deeper by adding the identity mapping from the previous layer to the current layer and then applying appropriate non-linear activation. Adding skip connections in the network helps a larger gradient flow to the previous layers,

through which, the problem of degradation is solved effectively. Specifically, defining the input RGB band group as I_{RGB}, the extracted RGB features F_{RGB} can be expressed as:

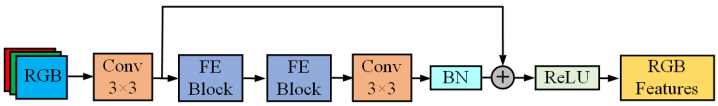
$$F_{RGB} = \delta\left(\phi(I_{RGB}) + \hat{f}_1(\phi(I_{RGB}))\right) \tag{1}$$

where ϕ represents the 3×3 convolution layer, $\hat{f}_1$ denotes the residual function (two FE blocks, and a 3×3 convolution layer followed by a batch normalization (BN)), and δ is the ReLU activation operation.

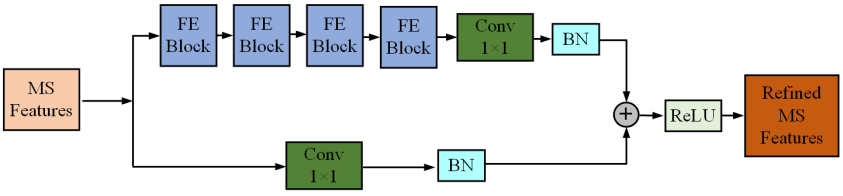
Then, a feature fusion layer was applied to fuse the extracted RGB, VRE, SWIR, and NIR features into MS features, and to obtain SAR features in the same way. The obtained MS features F_{MS} can be expressed as follows:

$$F_{MS} = [F_{RGB} \| F_{VRE} \| F_{SWIR} \| F_{NIR}] \tag{2}$$

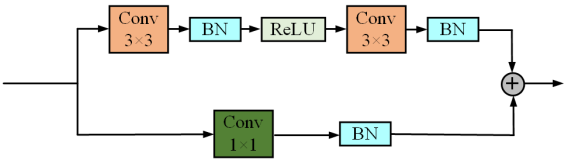
where '$\|$' refers to the channel concatenation, and F_{VRE}, F_{SWIR}, F_{NIR} represent the extracted VRE, SWIR, and NIR features, respectively.



(a) Shallow Feature Extraction Branch



(b) Deep Feature Extraction Branch



(c) Feature Extractor (FE) Block

Figure 3. Structure of the feature extraction branches, where (**a**–**c**) represent shallow feature extraction branch, deep feature extraction branch and feature extractor block, respectively.

After the shallow feature extraction branch, two deep feature extraction branches were introduced to perform feature refinement and extract semantic features, the structure of which can be seen in Figure 3b. Similar to shallow branches, shortcut connections were applied to solve the gradient divergence problem of the model. In addition, we used 1×1 convolution layers to reduce the dimensionality of the feature map, and improved the expressive ability of the model by adding nonlinearity. Taking MS features as an example, the refined MS features $\overline{F_{MS}}$ can be expressed as:

$$\overline{F_{MS}} = \delta\left(\hat{f}_2(F_{MS}) + \gamma(F_{MS})\right) \tag{3}$$

where $\hat{f}_2$ represents the residual function (four FE blocks and a 1×1 convolution layer followed by a BN), and γ denotes the 1×1 convolution layer followed by a BN.

Finally, a feature fusion layer was proposed to fuse the refined features $\overline{F_{MS}}$ and $\overline{F_{SAR}}$ by using channel concatenation. After this, three consequent 3 × 3 convolutions were applied to reduce the dimensionality of the fused features and map them to the original space size, which can be expressed as follows:

$$F_f = \rho_3(\rho_2(\rho_1([\overline{F_{MS}}\|\overline{F_{SAR}}]))) \quad (4)$$

where F_f is the fused features, and ρ_1, ρ_2, and ρ_3 denote three 3 × 3 convolution layers.

3.2. Self Channel Attention for LCZ Classification

Even though identity-based shortcut connections have proved to be able to learn considerably deeper and stronger networks, all feature maps in the bridge connections are considered to be equally important. Under this circumstance, the squeeze-and-excitation (SE) block [27], proposed to compute explicit associations between feature channels, has increased the representational power of conventional residual modules effectively. Although SE blocks use the global context to calibrate the weights of different channels and thus adjust the channel dependence, the feature fusion using weight recalibration can hardly make full use of the global context. To solve this problem, a self-channel attention (SCA) block, which can establish long-range spatial and inter-channel dependencies through a not-full-squeeze way, was proposed, the structure of which is shown in Figure 4.

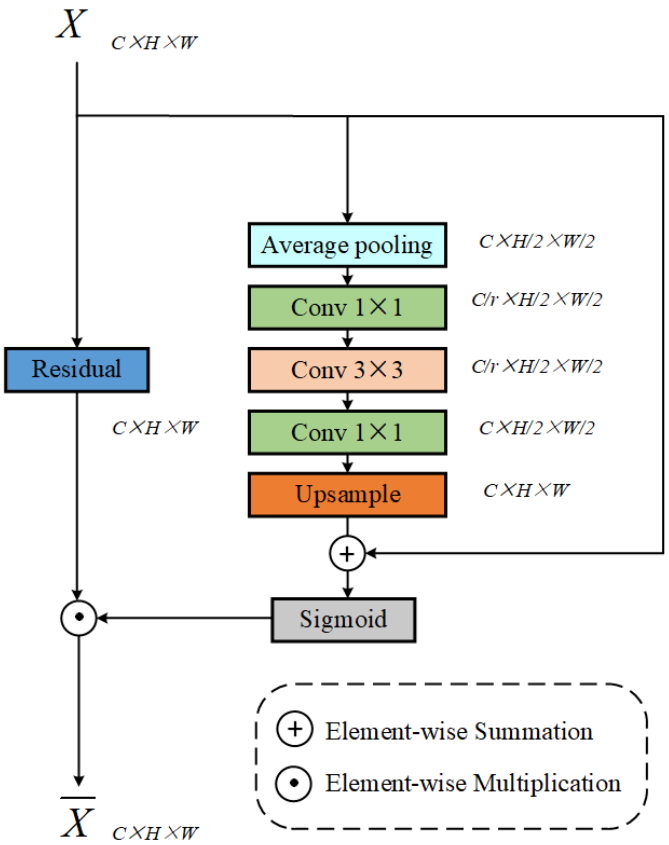


Figure 4. Schema of the self channel attention (SCA) block, where *C*, *H*, and *W* represent the number of channels, height, and width of the feature map, respectively. *r* is a reduction ratio used to vary the capacity. For simplicity, we denote convolution as Conv while omitting normalization and activation layer after each convolution layer.

The proposed SCA block aims to learn representative features through a re-weighting mechanism that accounts for both local and global aspects. Specifically, in SCA, a residual branch was applied to facilitate gradient flow to earlier layers, thereby addressing the degra-

dation problem. In addition, an attention branch was proposed to half-down-sample the feature map to produce non-local spatial and channel attention efficiently, where identity mapping was applied to adjust the channels' weight. The SCA can be formulated as:

$$\bar{X} = X_{res} \cdot (\varsigma(X_{att} + X)) \tag{5}$$

where X and $\bar{X}$ are the input and output feature maps, X_{res} is the residual representation of X, X_{att} represents the output of the attention branch $B_{att}(\cdot)$, and ς refers to the sigmoid activation operation.

Similar to the SE block, our SCA block has two operations, viz., squeeze and excitation. Features were first passed to the squeeze operation to produce a descriptor by aggregating along each of their spatial dimensions, which allows information from the global receptive field of the network to be used by all of its layers. To preserve the spatial information and establish long-range dependencies, the global average pooling was applied to half-down-sample the input feature map X. Then, convolutions with the same cascaded structure as the residual block were employed to further extract context features. The squeeze operation was followed by an excitation operation, where the produced embedding was utilized to obtain a collection of modulation weights for the feature maps. In this work, the up-sampling operation was introduced to recover the feature map to its original spatial size. The function of the attention branch, $B_{att}(\cdot)$, can be expressed as follows:

$$B_{att}(\cdot) = Up(Conv_{1-3-1}(Apool(\cdot))) \tag{6}$$

where $Apool(\cdot)$ is an average pooling layer used to perform the not-full-squeezed operation, $Conv_{1-3-1}$ represents the three cascaded convolutions, and $Up(\cdot)$ refers to the up-sampling mapping operation. Essentially, SCA blocks pay more attention to those informative features by establishing spatial-channel interdependencies in an efficient way.

To achieve high-precision LCZ mapping, an SCA-based classifier was designed for LCZ classification, the structure of which is shown in Figure 5. Concretely, our classifier contains three residual modules, each of which consists of an SCA block and (N-1) residual blocks. Hence, the depth of each residual module is N. Average pooling follows the stack of residual modules, and the final layer is a fully connected layer followed by a softmax activation, which predicts a specific LCZ class for the input image patch.

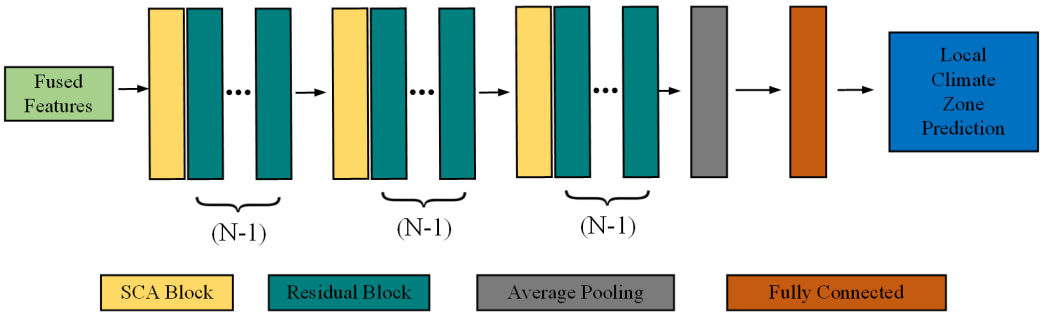


Figure 5. Structure of the classifier for LCZ classification, where N is the depth of each residual module.

4. Experiments

4.1. Dataset

Though a large number of publicly available datasets exist for scene classification, few datasets contain a large scale of multi-source images for LCZ classification. The So2Sat LCZ42 dataset [28] is a benchmark dataset for global LCZ that is funded by the European Research Council (ERC). The dataset was labeled by a group of domain experts in remote sensing following a carefully designed labeling workflow similar to that in world urban database and access portal tools (WUDAPT) [29]. Afterward, a rigorous quality assessment was conducted with independent label voting by domain experts who had not labeled the areas in the labeling stage.

The dataset contains 400,673 pairs of corresponding Sentinel-1 SAR and Sentinel-2 multi-spectral image patches with LCZ labels, which were selected from 42 urban agglomerations and 10 additional smaller areas over all of the inhabited continents (except for Antarctica) around the world. There are a total of 17 class image patches in the So2Sat LCZ42 dataset, where 10 are urban classes and 7 are natural classes. The dimension of the image patches in the dataset is 32 by 32 pixels, corresponding to a physical dimension of 320 by 320 m. For machine learning purposes, the So2Sat LCZ42 dataset was split into a training set, a test set, and a validation set, which consist of 352,366, 24,188, and 24,119 pairs of image patches, respectively. Note that all three sub-sets are geographically separated from each other, despite having drawn the test and validation sets from the same list of cities. The sample distributions of the training dataset and test dataset are shown in Figure 6.

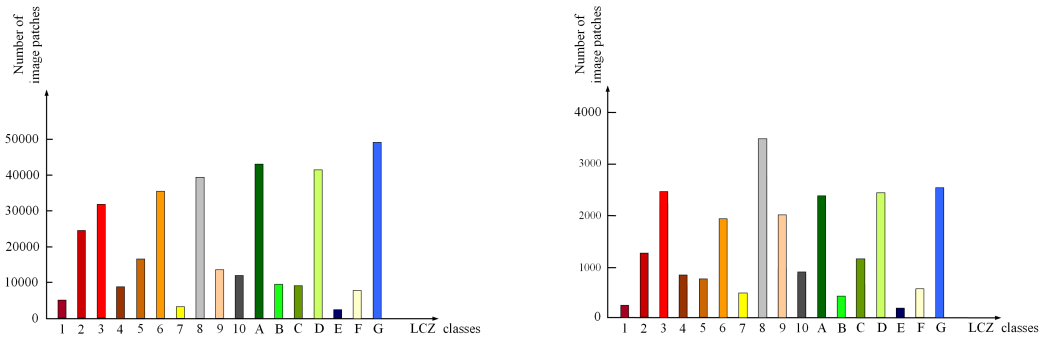


Figure 6. Sample distribution of LCZ labels in training set (**a**) and test set (**b**).

4.2. Experimental Setup

In this work, the experiments were conducted on Python 3.6 using the Pytorch framework. Two NVIDIA GeForce 2080Ti GPUs, with 12GB RAM for each GPU, were employed in the training process, and the compute unified device architecture (CUDA was used to improve the speed. We used the cross entropy loss as the training loss, where momentum was set to be 0.9 and weight decay was set to be 0.0001. The Adam optimizer [30] was chosen for model optimization. The learning rate was initially set as 1×10^{-3}, and dynamically adjusted with the 'Poly' learning rate scheduler [31], where the power was set to 0.9. The batch size was fixed as 32, and the model was trained for 80 K iterations.

4.3. Performance Metrics

To evaluate the performance of different models for LCZ classification, the overall accuracy (OA), average accuracy (AA), F_1 score, and Kappa coefficient were employed as the performance metrics.

As one of the most commonly used performance measurements, OA is the proportion of correctly classified samples to the total samples, which is an assessment of the overall accuracy. The calculation formula is given as follows:

$$\mathrm{OA} = \frac{TP}{TP + FP} \tag{7}$$

where TP represents true positive samples and FP represents false positive samples.

AA is the average of the ratios between the correct predictions of each class and the total number of each class, calculated as follows:

$$\mathrm{AA} = \frac{\sum_{K=1}^{N} \mathrm{recall_K}}{N} \tag{8}$$

where N represents the total number of categories and recall_K represents the recall value of the Kth class.

The F_1 score is the harmonic average of accuracy and recall, which is used to comprehensively reflect the performance of the model [32], and is calculated as follows:

$$F_1 = \sum_{K=1}^{N} \frac{2 \times \text{precision}_K \times \text{recall}_K}{\text{precision}_K + \text{recall}_K} / N \quad (9)$$

where precision_K represents the precision value of the Kth class.

The confusion matrix was also utilized to reflect the overall effect of LCZ classification. Based on this, the kappa coefficient [33] was calculated to reflect the overall bias of the model.

4.4. Performance Evaluation

With the test dataset of the Sot2Sat LCZ42 dataset, we obtained an OA of 67.87%, an AA of 59.56%, and a kappa of 0.6476. The confusion matrix is presented in Figure 7. It can be seen that most predicted labels match the ground truth labels (especially for LCZ 8, A, D, and G), but several LCZ data types have a relatively low prediction accuracy (especially for LCZ C). Due to bush and low plants having similar texture characteristics, most of the LCZ C are misclassified as LCZ D. In contrast, LCZ G is no doubt the easiest to classify. In addition, other confusions are also caused by similar textures, materials, and heights between LCZ classes. However, it is difficult to confused between urban classes (LCZ 1-10) and natural classes (LCZ A-G).

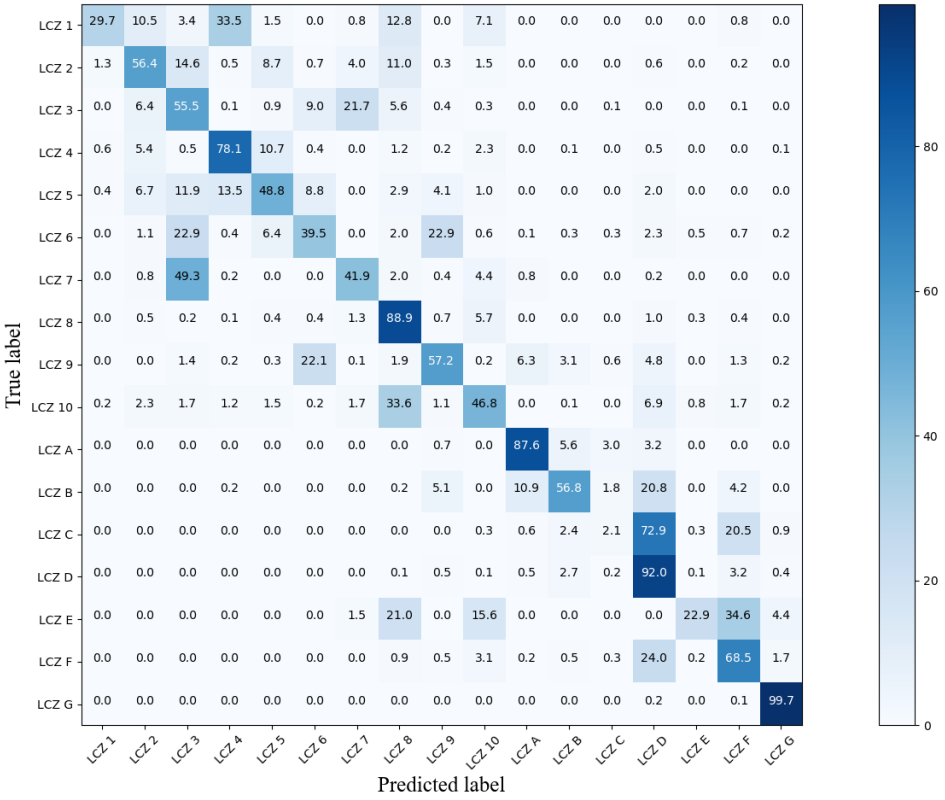


Figure 7. Confusion matrix of classification results from MsF-LCZ-Net (N = 5). For ease of presentation, the results are normalized by one percent of the total number of samples in each LCZ class.

The prediction results for each LCZ class are given in Table 3. We calculated the harmonic mean value- F_1 score of each LCZ class to analyze the classification results more objectively. Concretely, the vast majority of LCZ classes have high F_1 score values, except for LCZ 7, C, and E. In addition, we also found that the proportion of each LCZ class

in the training set is positively correlated with that in the predicted results, from which, we can deduce that the balance of the dataset has an important impact on the output of the network. When the characteristics of the two LCZ classes are similar, the network is accustomed to predicting the class as the one with more training samples. Moreover, the smaller the ratio of a class in the training set, the greater the difference between precision and recall in the test set, and the more likely this class is to be misclassified.

Table 3. Prediction results for each LCZ class on the test set.

Class	Precision(%)	Recall(%)	F_1(%)	Ratio(%)
LCZ 1: compact high-rise	29.7	74.5	42.5	1.44
LCZ 2: compact mid-rise	56.4	67.1	61.3	6.93
LCZ 3: compact low-rise	55.5	57.0	56.3	8.99
LCZ 4: open high-rise	78.1	74.5	76.2	2.46
LCZ 5: open mid-rise	48.8	49.1	48.9	4.68
LCZ 6: open low-rise	39.5	50.1	44.1	10.02
LCZ 7: lightweight low-rise	41.9	24.4	30.8	0.93
LCZ 8: large low-rise	88.9	79.8	84.1	11.16
LCZ 9: sparsely built	57.2	66.5	61.5	3.86
LCZ 10: heavy industry	46.8	53.7	50.0	3.39
LCZ A: dense trees	87.6	91.2	89.4	12.18
LCZ B: scattered trees	56.8	45.2	50.4	2.70
LCZ C: bush and scrub	2.1	18.2	3.7	2.60
LCZ D: low plants	92.0	61.1	73.5	11.74
LCZ E: bare rock or paved	22.9	57.3	32.8	0.68
LCZ F: bare soil or sand	68.5	44.6	54.1	2.24
LCZ G: water	99.7	98.1	98.9	14.00

4.5. Performance Comparison with State-of-the-Art Methods

To illustrate the effectiveness of our proposed method, the performance of the proposed method was compared with nine other state-of-the-art deep-learning-based LCZ classification methods using MS and SAR multi-source data: MSPPF-NETS [20], LCZ-MF [18], EB-CNN [25], DenseNet-DFN [26], FusionNet [34], LCZNet [35], ResNext29_8_64 [24], MCFUNet-LCZ [36], and RSNNet [37]. In order to facilitate a comparison, the settings and environment of all competing methods in the training phase were set as the same. Due to some models not using pre-trained backbones in the feature extraction step, pretrained backbones were abandoned in all methods for a fair comparison. All parameter settings refer to the above experiment settings. The depth of the residual module in the classifier N was set to be 5, with which, the best classification results can be obtained. The influence of depth on the results will be explained in the next section. The results are given in Table 4, where we can see that our method achieves the best overall performance in terms of OA, AA, and kappa.

4.6. Large-Scale LCZ Maps

To further prove the value of our work, we selected three mega cities of China to obtain their city maps, among which, the prediction maps of Beijing and Guangzhou are shown in Figure 8, and the LCZ maps of Shanghai in different periods are shown in Figure 9. We obtained the Sentinel-1 VV-VH dual polarization single-look complex (SLC) Level-1 data and cloud-free Sentinel-2 images from the Copernicus Open Access Hub using SentinelSat Python API, and refer to the method in [28] to preprocess the MS and SAR data, respectively. Since MS images and SAR images are not in the same reference coordinate system, we used the Image Registration Workflow from ENVI to register them. After that, a sliding window method was applied to divide the registered multi-source images into 32×32 image patches, and were predicted by our trained model in turn, finally obtaining the LCZ map.

Table 4. Classification performance comparison with other state-of-the-art methods on the So2Sat LCZ42 dataset.

	OA(%)	AA(%)	Kappa(×100)
MSPPF-NETS [20]	62.05	51.32	58.54
LCZ-MF [18]	65.66	50.63	62.21
EB-CNN [25]	61.11	44.37	57.07
DenseNet-DFN [26]	64.07	50.59	60.49
FusionNet [34]	64.57	52.17	57.45
LCZNet [35]	66.23	57.76	63.15
ResNext29_8_64 [24]	64.91	54.05	61.47
MCFUNet-LCZ [36]	65.74	53.18	61.94
RSNNet [37]	64.15	51.66	60.79
MsF-LCZ-Net (N = 5)	**67.87**	**59.56**	**64.76**

From Figure 8, we can find that Beijing shows a compact urban structure, which is concentrated in the center of the whole map. The central areas are mainly classified as LCZ 1 compact high-rise and LCZ 2 compact middle-rise. LCZ B scattered trees and LCZ A dense trees are located in the western and northern area, whereas the eastern areas are mainly classified as LCZ 8 large low-rise. In contrast, Guangzhou's city structure is relatively dispersed.

In addition, we also compared the LCZ maps of the same area at different periods, as shown in Figure 9. As one of the fastest-growing cities in the world, the development speed of Shanghai attracts worldwide attention. We compared the LCZ classification results of the Shanghai area in 2015 and 2021. From Figure 9, we can see that, compared to six years ago, Shanghai's urban structure has become more compact. Compact buildings replace bare soil and large low-rise as the dominant classes in Shanghai. The development of Shanghai in the past six years is mainly concentrated in the central and eastern areas. Moreover, we can also deduce that Shanghai's economic center of gravity has a tendency to move eastward.

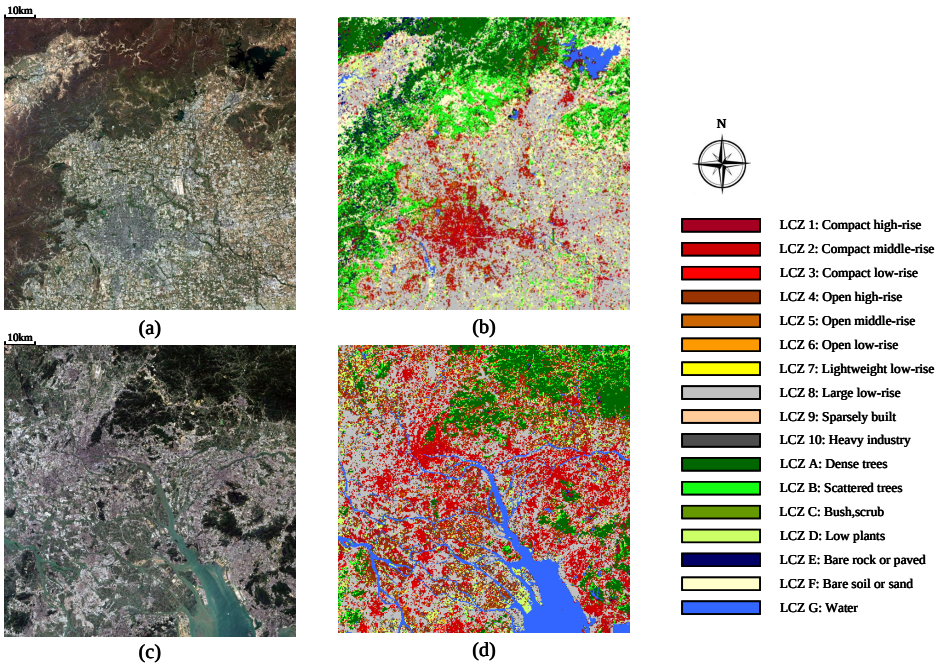


Figure 8. Two examples of large-scale LCZ classification, where (**a**,**b**) are the Sentinel-2 satellite image of Beijing, China and the corresponding LCZ map, and (**c**,**d**) are the Sentinel-2 satellite image of Guangzhou, China and the corresponding LCZ map.

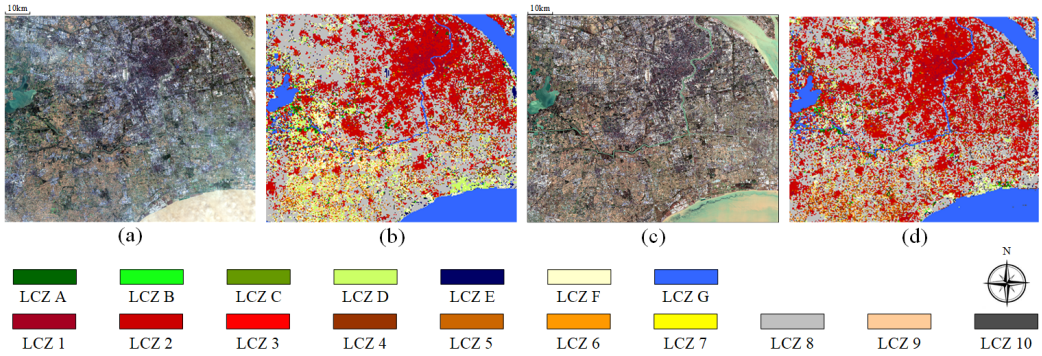


Figure 9. Changes in Shanghai, China in 6 years, where (**a**,**b**) are the Sentinel-2 satellite image of Shanghai in 2015 and the corresponding LCZ map, and (**c**,**d**) are the Sentinel-2 satellite image of Shanghai in 2021 and the corresponding LCZ map.

5. Discussion

5.1. Effect of Network Depth

Our proposed classifier is composed of three connected residual modules, and the depth of each part is N. In order to analyze the effect from the architecture, depth N in the LCZ classification was explored to illustrate the change in accuracy according to the model complexity.. The evaluation results for the training set and test set are presented in Figure 10. With the depth N increasing, all evaluation metrics (OA, AA, and kappa) increase as expected. For the training set, MsF-LCZ-Net (N = 7) obtains the best performance, with an OA of 97.88%, an AA of 97.19%, and a Kappa of 0.9766. This is because a deeper network corresponds to a larger size of the receptive field, and hence more high-level features can be obtained. In addition, the shortcut connections in residual learning can effectively prevent the problem of gradient vanishing caused by the network being too deep. In contrast, for the test set, MsF-LCZ-Net (N = 5) obtains the best performance, with an OA of 67.87%, an AA of 59.56%, and a kappa of 0.6476.

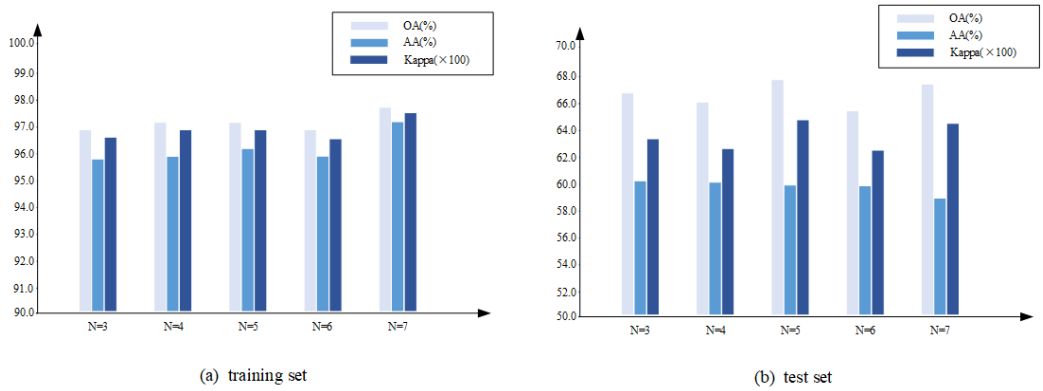


Figure 10. Mean values of OA, AA, and kappa based on three tests about network depth N, evaluated on the training set (**a**) and test set (**b**) in the So2Sat LCZ42 dataset, respectively.

5.2. Effect of Data Fusion

To find out the effect from data fusion on LCZ classification, we conducted an experiment with SAR data, MS data, and MS&SAR data. The obtained results are given in Table 5, where the depth N was set to be 5. It can be seen that the classification results of SAR data on the test set are disappointing. This is because there are many speckle noises in the SAR data, which will influence the analysis and interpretation of the SAR images. In addition, we also found that, when using our method to fuse MS data and SAR data, the fused data achieve the best performance in both the training set and the test set. This shows that the fusion method of multi-source remote sensing data can combine the advantages of different source data while alleviating their respective disadvantages so as to

achieve better LCZ classification results. Considering that the phase of the SLC source data may act as a noise in the classification algorithm, we further conducted LCZ classification by only using the filtered intensity (SAR (Lee-filtered)) motivated by the effectiveness of the degree of polarization [38]. The results show that the OA, AA, and kappa of the MS&SAR (Lee-filtered) data increase by 0.15%, 0.71%, and 0.39, respectively, compared with the MS&SAR data, indicating that it could be a better choice to use filtered intensity for LCZ classification.

Table 5. Mean values of OA, AA, and kappa evaluated on the So2Sat LCZ42 dataset using different data sources, where SAR refers to using all of the eight bands and SAR (Lee-filtered) refers to using only the fifth and sixth bands as described in Table 1.

Dataset	OA (%)	AA (%)	Kappa (×100)
SAR	32.08	22.14	26.54
SAR(Lee-filtered)	46.74	39.75	41.37
MS	65.41	53.96	60.72
MS & SAR	67.87	59.56	64.76
MS & SAR(Lee-filtered)	68.02	60.27	65.15

5.3. Ablation Study

In this section, we designed an ablation study to verify the effectiveness of SCA blocks and FE blocks. The LCZ classification results of different structures are shown in Figure 11, where the depth of the network was also set to be 5. It can be seen that, when the SCA blocks are abandoned in the classifier, the OA, AA, and kappa drop by 2.67%, 3.02%, and 0.0305, respectively, in the test set, with which, it can be deduced that SCA blocks are beneficial for preserving spatial information and aggregating non-local extracted features upon the main channel. Hence, when the SCA blocks are not applied in the classifier, the complex correlation between channels will decrease. When the FE blocks are not employed in the multi-branch feature extraction structure, the OA, AA, and kappa drop by 2.29%, 1.89%, and 0.0249, respectively, in the test set. This proves that our proposed FE blocks are beneficial for the extraction and fusion of multi-scale features.

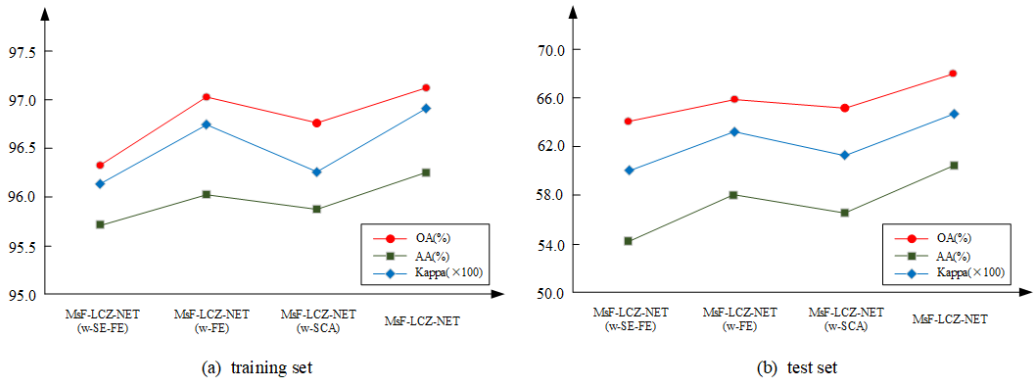


Figure 11. Mean values of OA, AA, and kappa based on three tests about different structures, evaluated on the training set (**a**) and test set (**b**) in the So2Sat LCZ42 dataset, respectively. Here, we use 'w-' to indicate without a specific module.

6. Conclusions

In this study, we proposed an LCZ classification framework that included a data-grouping method and an MsF-LCZ-Net framework for conducting LCZ mapping, where a multi-branch CNN was designed to capture multi-level information and enhance the feature representation, and SCA blocks were introduced to the classifier to establish spatial-channel dependencies, thereby enhancing the LCZ classification results. We evaluated our method on the So2Sat LCZ42 dataset, and experiments showed that our MsF-LCZ-Net

outperformed all of the comparing methods. In addition, we performed LCZ mapping on three mega cities in China, and also explored the influence of the network depth, data fusion, and different architecture on the LCZ classification. When the network depth was set to 5, the MsF-LCZ-net had the best generalization ability. The fusion of the multi-source data can improve the LCZ classification performance. In addition, the ablation study proved the effectiveness of SCA and FE blocks. In the future, we are committed to solving the impact of imbalanced data samples on LCZ classification. In addition, investigating how to better distinguish between similar LCZ classes is also the focus of our future work.

Author Contributions: Conceptualization, G.H.; methodology, Z.D.; validation, P.F.; formal analysis, J.G.; investigation, G.H.; resources, X.Z.; data curation, G.H.; writing—original draft preparation, Z.D.; writing—review and editing, P.F. and S.J.; visualization, Z.D.; supervision, J.G.; project administration, G.H.; funding acquisition, C.J. All authors have read and agreed to the published version of the manuscript.

Funding: The study of this paper is funded by the National Natural Science Foundation of China (NSFC) under grant contracts No. 41801291, No. 61806018, and No. 42071297.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this work, cloud-free Sentinel-2 images employed for training and validation were obtained from Google Earth Engine (GEE) https://earthengine.google.com/ (accessed on 6 July 2020), and the So2Sat LCZ42 dataset are available on http://doi.org/10.14459/2018mp1483140 (accessed on 2 July 2020).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kotharkar, R.; Bagade, A. Evaluating urban heat island in the critical local climate zones of an Indian city. *Landsc. Urban Plan.* **2018**, *169*, 92–104. [CrossRef]
2. Stewart, I.D.; Oke, T.R. Local Climate Zones for Urban Temperature Studies. *Bull. Am. Meteorol. Soc.* **2012**, *93*, 1879–1900. [CrossRef]
3. Alexandri, E.; Jones, P. Temperature decreases in an urban canyon due to green walls and green roofs in diverse climates. *Build. Environ.* **2008**, *43*, 480–493. [CrossRef]
4. de Munck, C.; Pigeon, G.; Masson, V.; Meunier, F.; Bousquet, P.; Tréméac, B.; Merchat, M.; Poeuf, P.; Marchadier, C. How much can air conditioning increase air temperatures for a city like Paris, France? *Int. J. Climatol.* **2013**, *33*, 210–227. [CrossRef]
5. Bechtel, B.; Daneke, C. Classification of local climate zones based on multiple earth observation data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2012**, *5*, 1191–1202. [CrossRef]
6. Xu, Y.; Ren, C.; Cai, M.; Wang, R. Issues and challenges of remote sensing-based local climate zone mapping for high-density cities. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, United Arab Emirates, 6–8 March 2017; pp. 1–4.
7. Schmitt, M.; Hughes, L.; Zhu, X.X. The SEN1-2 dataset for deep learning in SAR-optical data fusion. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *4*, 141–146. [CrossRef]
8. Mills, G.; Ching, J.; See, L.; Bechtel, B.; Foley, M. An introduction to the WUDAPT project. In Proceedings of the the 9th International Conference on Urban Climate, Toulouse, France, 20–24 July 2015; pp. 20–24.
9. Bechtel, B.; Foley, M.; Mills, G.; Ching, J.; See, L.; Alexander, P.; O'Connor, M.; Albuquerque, T.; de Fatima Andrade, M.; Brovelli, M.; et al. CENSUS of Cities: LCZ Classification of Cities (Level 0)—Workflow and Initial Results From Various Cities. In Proceedings of the ICUC9-9th International Conference on Urban Climate Jointly with 12th Symposium on the Urban Environment, Toulouse, France, 20–24 July 2015.
10. Seo, D.; Kim, Y.; Eo, Y.; Lee, M.; Park, W. Fusion of SAR and Multispectral Images Using Random Forest Regression for Change Detection. *ISPRS Int. J.-Geo-Inf.* **2018**, *7*, 401. [CrossRef]
11. Oliver, C.; Quegan, S. *Understanding Synthetic Aperture Radar Images*; SciTech Publishing: Raleigh, NC, USA, 2004.
12. Blaes, X.; Vanhalle, L.; Defourny, P. Efficiency of crop identification based on optical and SAR image time series. *Remote. Sens. Environ.* **2005**, *96*, 352–365. [CrossRef]
13. Xu, G.; Zhu, X.; Tapper, N.; Bechtel, B. Urban climate zone classification using convolutional neural network and ground-level images. *Prog. Phys. Geogr. Earth Environ.* **2019**, *43*, 410–424. [CrossRef]
14. Hong, G.; Zhang, Y.; Mercer, B. A wavelet and IHS integration method to fuse high resolution SAR with moderate resolution multispectral images. *Photogramm. Eng. Remote Sens.* **2009**, *75*, 1213–1223. [CrossRef]

15. Liu, G.; Li, L.; Gong, H.; Jin, Q.; Li, X.; Song, R.; Chen, Y.; Chen, Y.; He, C.; Huang, Y.; et al. Multisource remote sensing imagery fusion scheme based on bidimensional empirical mode decomposition (BEMD) and its application to the extraction of bamboo forest. *Remote. Sens.* **2016**, *9*, 19. [CrossRef]
16. He, K.; Zhang, X.; Ren, S.; Jian, S. Identity Mappings in Deep Residual Networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
17. Chunping, Q.; Schmitt, M.; Lichao, M.; Xiaoxiang, Z. Urban local climate zone classification with a residual convolutional neural network and multi-seasonal Sentinel-2 images. In Proceedings of the 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), Beijing, China, 19–20 August 2018; pp. 1–5.
18. Qiu, C.; Tong, X.; Schmitt, M.; Bechtel, B.; Zhu, X.X. 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.* **2020**, *13*, 2793–2806. [CrossRef]
19. Zhou, Y.; Wei, T.; Zhu, X.; Collin, M. A Parcel-Based Deep-Learning Classification to Map Local Climate Zones From Sentinel-2 Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 4194–4204. [CrossRef]
20. Yang, R.; Zhang, Y.; Zhao, P.; Ji, Z.; Deng, W. MSPPF-Nets: A Deep Learning Architecture for Remote Sensing Image Classification. In Proceedings of the IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3045–3048.
21. Fiaschi, S.; Holohan, E.P.; Sheehy, M.; Floris, M. PS-InSAR analysis of Sentinel-1 data for detecting ground motion in temperate oceanic climate zones: A case study in the Republic of Ireland. *Remote Sens.* **2019**, *11*, 348. [CrossRef]
22. Bechtel, B.; See, L.; Mills, G.; Foley, M. Classification of local climate zones using SAR and multispectral data in an arid environment. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 3097–3105. [CrossRef]
23. Shi, L.; Ling, F. Local Climate Zone Mapping Using Multi-Source Free Available Datasets on Google Earth Engine Platform. *Land* **2021**, *10*, 454. [CrossRef]
24. Yan, Z.; Ma, L.; He, W.; Zhou, L.; Lu, H.; Liu, G.; Huang, G. Comparing Object-Based and Pixel-Based Methods for Local Climate Zones Mapping with Multi-Source Data. *Remote Sens.* **2022**, *14*, 3744. [CrossRef]
25. Feng, P.; Lin, Y.; Guan, J.; Dong, Y.; He, G.; Xia, Z.; Shi, H. Embranchment CNN Based Local Climate Zone Classification Using Sar And Multispectral Remote Sensing Data. In Proceedings of the IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 6344–6347.
26. Feng, P.; Lin, Y.; He, G.; Guan, J.; Wang, J.; Shi, H. A Dynamic End-to-End Fusion Filter for Local Climate Zone Classification Using SAR and Multi-Spectrum Remote Sensing Data. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 4231–4234.
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
28. Zhu, X.X.; Hu, J.; Qiu, C.; Shi, Y.; Kang, J.; Mou, L.; Bagheri, H.; Haberle, M.; Hua, Y.; Huang, R.; et al. So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [Software and Data Sets]. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 76–89. [CrossRef]
29. Bechtel, B.; Alexander, P.J.; Beck, C.; Böhner, J.; Brousse, O.; Ching, J.; Demuzere, M.; Fonte, C.; Gál, T.; Hidalgo, J.; et al. Generating WUDAPT Level 0 data–Current status of production and evaluation. *Urban Clim.* **2019**, *27*, 24–45. [CrossRef]
30. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–13.
31. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
32. AlBeladi, A.A.; Muqaibel, A.H. Evaluating compressive sensing algorithms in through-the-wall radar via F1-score. *Int. J. Signal Imaging Syst. Eng.* **2018**, *11*, 164–171. [CrossRef]
33. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Med. Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
34. Gawlikowski, J.; Schmitt, M.; Kruspe, A.; Zhu, X.X. On the fusion strategies of Sentinel-1 and Sentinel-2 data for local climate zone classification. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 2081–2084.
35. Liu, S.; Shi, Q. Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China. *ISPRS J. Photogramm. Remote Sens.* **2020**, *164*, 229–242. [CrossRef]
36. Ji, W.; Chen, Y.; Li, K.; Dai, X. Multicascaded Feature Fusion-Based Deep Learning Network for Local Climate Zone Classification Based on the So2Sat LCZ42 Benchmark Dataset. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 449–467. [CrossRef]
37. Zhou, L.; Shao, Z.; Wang, S.; Huang, X. Deep learning-based local climate zone classification using Sentinel-1 SAR and Sentinel-2 multispectral imagery. *Geo-Spat. Inf. Sci.* **2022**, *25*, 1–16. [CrossRef]
38. Chang, J.G.; Shoshany, M.; Oh, Y. Polarimetric radar vegetation index for biomass estimation in desert fringe ecosystems. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7102–7108. [CrossRef]