# Local Climate Zone Mapping by Coupling Multilevel Features With Prior Knowledge Based on Remote Sensing Images

Xinrun Zhong, Huifang Li, *Member, IEEE*, Huanfeng Shen, *Senior Member, IEEE*, Meiling Gao, Zhihua Wang, and Jinqiang He

***Abstract*— Local climate zone (LCZ) mapping can explore the variability of the impact of urban form on the thermal environment in different urban contexts, and large-scale LCZ mapping can help us to better understand the spatial and temporal dynamics of the climate in urban areas around the world. Studies have indicated that deep learning-based methods can effectively perform the LCZ classification. However, the accuracy of LCZ classification on large-scale datasets is still unsatisfactory, mainly due to the fact that the traditional convolutional neural networks are not good at mining contextual information, which is crucial for fully understanding remote sensing (RS) scenes. In this article, to solve this problem, we propose an LCZ mapping method based on RS images by coupling multilevel features mined from global and local ranges with prior knowledge, named LCZ-MFKNet. The global and local features are extracted through Swin Transformer and space-maintained ResNet (SM-ResNet) model branches, respectively, and then fused through an improved squeeze-and-excitation (iSE) module. The prior knowledge studied from the theoretical definition and experimental tests is that two typical sets of LCZ categories are easily confounded in multiclass classification but separable in two-class classification. Experiments are conducted on the large publicly available So2Sat LCZ42 dataset, where the proposed LCZ-MFKNet method achieved the highest LCZ mapping accuracy. Moreover, six megacities were selected globally for LCZ mapping, and the results verified the accuracy and the general applicability of the proposed LCZ-MFKNet method in large-scale LCZ mapping.**

***Index Terms*— Local climate zone, local–global feature fusion, multilevel features, prior knowledge, transformer.**

Manuscript received 15 November 2023; revised 10 January 2024; accepted 28 January 2024. Date of publication 31 January 2024; date of current version 12 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFF1301103, in part by the National Natural Science Foundation of China under Grant 42371366, and in part by the Key Research and Development Program of Hubei Province under Grant 2023BAB066. *(Corresponding author: Huifang Li.)*

Xinrun Zhong and Huifang Li are with School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China (e-mail: 2021202050019@whu.edu.cn; huifangli@whu.edu.cn).

Huanfeng Shen is with the School of Resource and Environmental Sciences and the Collaborative Innovation Center for Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: shenhf@whu.edu.cn).

Meiling Gao is with the College of Geological Engineering and Geomatics, Chang'an University, Xi'an 710054, China (gaomeiling@chd.edu.cn).

Zhihua Wang is with the School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ 85287, USA (zhwang@asu.edu).

Jinqiang He is with the School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China (e-mail: 08205101@cumt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3360522

## I. Introduction

THE concept of the local climate zone (LCZ) was proposed by Stewart and Oke [1] in 2012 in response to the urban heat island (UHI) problem, where they defined an LCZ as an area with uniform surface cover, surface structure, materials, and human activity, on a horizontal scale of a few hundred meters to a few thousand meters. Based on standardized descriptions of surface structure and land cover, Stewart and Oke [1] classified the landscape into 17 LCZ categories, including ten building categories (LCZ 1–10) and seven natural cover categories (LCZ A–G), as shown in Fig. 1. The differences between rural and urban building structures, as well as ground cover, are one of the main reasons for the formation of UHIs, and a quantitative relationship between urban form and local climate can be constructed based on LCZs. Large-scale mapping of LCZs can help us to better understand the spatial and temporal dynamics of LCZs in cities at a global scale and to explore the variability of the impact of urban morphology on the thermal environment in different urban contexts. Therefore, the LCZ classification system is very important. However, high-precision LCZ maps that can be generalized globally are still lacking, which hinders the macroscale and deep understanding of the urban climate.

Currently, there are two main types of LCZ classification methods: geographic information system (GIS)-based methods and remote sensing (RS)-based methods. The GIS-based methods recognize LCZ categories by calculating the different surface parameters required for LCZ classification [2], but this approach suffers from the problem of the many input geographic data sources and the difficulty in obtaining some of them. The RS-based methods classify the images by analyzing the similarities and differences of the spatial, spectral, and textural features in the images, which are abundant and easily obtained. As a result, the RS-based methods are now becoming the mainstream methods for LCZ classification. In 2015, the World Urban Database and Access Portal Tools (WUDAPT) project [3], [4] was proposed to provide free access to Landsat satellite imagery and the System for Automated Geoscientific Analyses (SAGA) GIS software with the random forest (RF) classifier for LCZ classification. WUDAPT provides a simple and easy-to-understand platform to produce free LCZ data that can be applied by users without spatial analysis or urban climate expertise. However, the overall accuracy (OA) of the

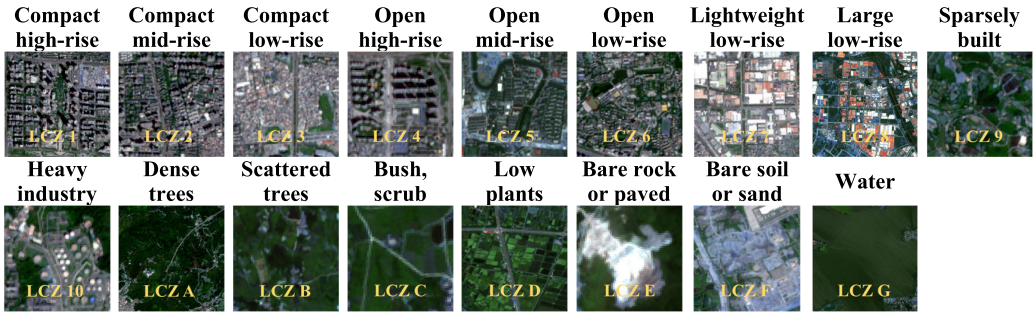



Fig. 1. 17 classes of LCZs and corresponding scene examples. The high-resolution images of the scene examples are all Sentinel-2 imagery.

WUDAPT LCZ maps is around 50% [5], [6], [7], which is not high enough for many applications. WUDAPT is a pixel-based classification and does not capture the horizontal heterogeneity of urban landscapes, and the RF classifier has difficulty in handling image spatial context information and lacks a hierarchical feature extraction capability, resulting in the poor accuracy of the WUDAPT product. As a result, a series of LCZ classification methods based on convolutional neural networks (CNNs) have emerged, which are capable of automatically learning and effectively capturing the spatial features in images. Compared to the pixel-based RF methods, CNN models are able to obtain the contextual information of the input image, and it has been reported that a CNN model was able to significantly improve the LCZ classification accuracy to more than 80% in eight German cities [8]. The MSMLA-Net method enhances the spectral and spatial characterization by integrating a multilevel attention module, and it outperformed three advanced LCZ classification models in tests on six cities in South Korea [9]. LCZNet uses a convolutional layer with multiscale filters to obtain the spatial features and integrates the channel features through SE-Residual blocks, and the model achieved a nearly 20% higher accuracy than WUDAPT in three economic regions of China [10]. It has been found that CNNs can achieve a significantly higher accuracy than WUDAPT. However, in fact, the training of a CNN is highly dependent on the dataset, and the datasets used in the above studies all had a small range of study area, which leads to poor model transferability, and the datasets used were not the same, which leads to poor comparability of the accuracy among the different CNN models.

Currently, LCZ datasets can be divided into two types according to their size. One type is the small datasets, which are constructed with data from only a few specific cities, such as Dongguan city [11], eight German cities [8], six Korean cities [9], and three Chinese economic regions [10]. Correspondingly, based on these datasets, the DRSNet [12], MSMLA-Net [9], and LCZNet [10] methods were proposed, and very high accuracies were obtained. However, these models have the problem of poor generalization ability when applied to other cities outside of the dataset. The idea of domain adaptation was therefore introduced to alleviate the generalization difficulty, to some extent [10], [13], [14]. A high-resolution domain adaptation network (HighDAN) [15] utilizes adversarial learning to effectively reduce the disparities between RS images from different urban environments, thus greatly eliminating interclass differences. The other type is the large datasets. The largest publicly available LCZ dataset is the So2Sat LCZ42 dataset [16]. So2Sat LCZ42 covers 42 cities with approximately 500 000 patches worldwide. The global training data that are geographically nonoverlapping enable the model to learn image features from different urban backgrounds, thus enhancing the generalization ability of the model. However, the large variation in the data domain caused by the large amount of data makes it difficult to train the model, resulting in a lower accuracy for the models trained on this dataset [17], [18]. Currently, there are models trained on So2Sat LCZ42 dataset, such as Sen2LCZ-Net [17] and MCFUNet-LCZ [18]. MCFUnet-LCZ is currently the best-performing model in classification performance on the So2Sat LCZ42 dataset, obtaining an OA and Kappa coefficient of 0.700 and 0.680, respectively.

In this article, we focus on large datasets to solve the challenge of low LCZ classification accuracy, in order to realize globally usable LCZ mapping. The current main factors limiting the mapping accuracy include two aspects. First, the feature mining is not sufficient. CNNs can effectively mine the local information of the imagery, while the local characteristics of the convolutional layer limit the network's ability to capture global contextual information, and CNNs cannot fully explore the geometric structure and spatial distribution of the features contained in high-resolution RS images. Second, LCZ mapping is a kind of multiclass recognition task, in which the class samples are not balanced, and it is difficult to realize the optimal recognition accuracy for each class under the global optimization objective.

In feature mining, Transformer models [19], which can capture a longer range of dependencies through a self-attention mechanism, have developed rapidly in recent years, due to their powerful global modeling capability and interpretability. Transformer models have excelled in natural language processing and computer vision tasks, such as the Swin Transformer [20] and Extended Vision Transformer (ExViT) [21]. In an RS foundation model named SpetralGPT [22], the utilization of transformers enables the comprehensive exploitation of RS big data with varying sizes, resolutions, time series, and regions. For multicategorization tasks, scene prior knowledge can help the model to distinguish similar categories, thus improving the classification accuracy [23], [24], [25], [26], [27]. This is because there are often large differences between the same scene categories, and such internal differences can make it difficult for the model to fit the category features. Therefore, in addition to using the model

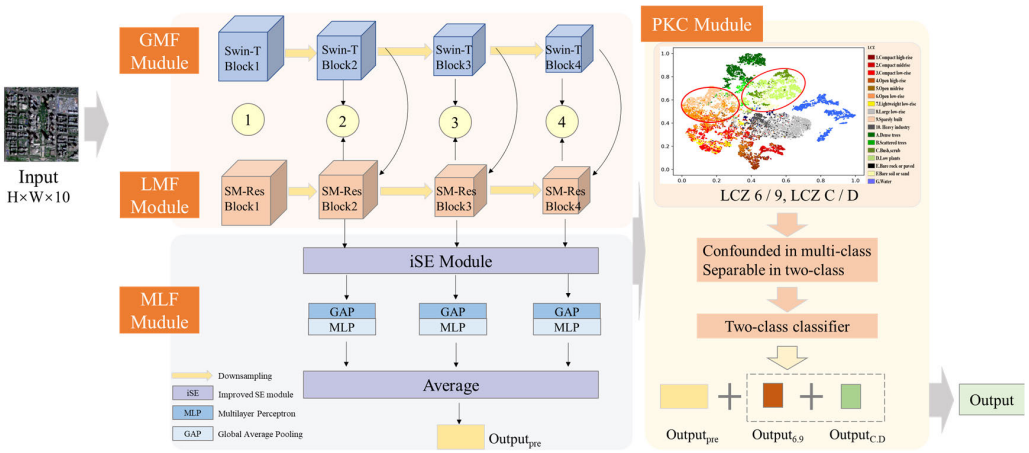



Fig. 2. Overall architecture of LCZ-MFKNet. The GMF module and the LMF module are used to extract global and local features, respectively. The MLF module is used to efficiently fuse the global and local features and achieve multilevel feature fusion. The features extracted by the MLF module are input into the PKC module for refinement.

to directly extract LCZ features, additional prior knowledge is needed to guide the model in distinguishing confusable categories. Prior knowledge can be preexisting background information, experience, or knowledge prior to performing a task [28]. Multilabel classification results generated by an auxiliary task can be used as prior information to guide the model to ultimately generate a more accurate description [29]. Topography and land-cover relationships, and also vegetation growth differences, can be used as prior knowledge to constrain the classifier's decision-making, thus improving the classification accuracy [30]. The use of land-use data as prior knowledge can effectively improve the classification accuracy of the model for Landsat 8 Operational Land Imager (OLI) imagery [31]. Moreover, the traditional numerical solution of low-rank representation (LRR) has been successfully used as prior knowledge to guide the parameter optimization of the deep neural network [32]. In summary, for accurate LCZ mapping, it is necessary to accumulate useful prior knowledge from large-scale benchmark datasets or other relevant data [33], [34], [35].

Therefore, an LCZ mapping method coupling multilevel features with prior knowledge for RS images—LCZ-MFKNet—is proposed in this article. The global and local features are first extracted by a transformer and a residual network, respectively, and then aggregated through a newly constructed multilevel feature fusion (MLF) module. The prior knowledge of the LCZs is learned through comprehensively analyzing the category properties and the confusion matrices, based on the multiclass training. The prior knowledge is finally coupled with the multilevel features to obtain the LCZ mapping results. The proposed LCZ-MFKNet method was compared with the state-of-the-art classification networks and advanced LCZ mapping methods on the So2Sat LCZ42 dataset. The experimental results demonstrate that LCZ-MFKNet can obtain highly accurate LCZ mapping results and is superior to the comparison methods.

The main contributions of this article can be summarized as follows: 1) an MLF module is constructed to integrate the global and local features, which can fully exploit the global–local features and the multiscale information in a scene; 2) the prior knowledge of LCZs is studied and coupled with the multilevel features to build the LCZ-MFKNet model on the large So2Sat LCZ42 dataset, and a high mapping accuracy is achieved; and 3) the LCZ maps for six megacities around the world obtained using LCZ-MFKNet are visually accurate, verifying the generalization ability of LCZ-MFKNet.

The rest of this article is organized as follows. The proposed LCZ-MFKNet method is described in detail in Section II. In Section III, the experimental results, the comparison results, and the LCZ maps for the six megacities are presented. Finally, our conclusions are drawn in Section IV.

## II. Methods

LCZ mapping is a kind of scene classification task. It is important to fully explore the spatial structure and distribution of features in the imagery and recognize the subtle differences between the different categories for LCZ mapping. In addition, prior knowledge of the categories is essential for improving the mapping precision. Therefore, the LCZ-MFKNet method proposed in this article focuses on the extraction of the global and local multiscale features of a scene, which is guided by a prior knowledge of the two-class discrepancy. The overall framework of the proposed LCZ-MFKNet method is shown in Fig. 2. LCZ-MFKNet consists of four main modules: the global multiscale feature extraction (GMF) module, the local multiscale feature extraction (LMF) module, the multilevel feature fusion (MLF) module, and the prior knowledge coupling (PKC) module. The multilevel features consisting of multiscale features extracted from both global and local scopes are fused through the MLF module to support the preliminary prediction of the categories. The PKC module is then used to guide the preliminary results to achieve the final LCZ classification results.

### A. Multilevel Feature Extraction and Fusion

*1) Global Multiscale Feature Extraction:* In LCZ scene classification, obtaining the global distribution of the features and mining the global contextual information of the RS scene are conducive to fully understanding the RS scene.

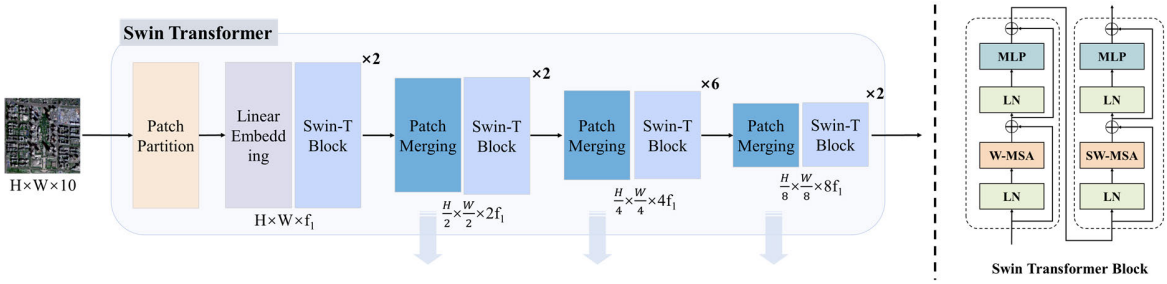



Fig. 3. GMF based on Swin Transformer. The input of the GMF module is the original image patch and the output is the features at the last three scales of the GMF module. The main component block of Swin Transformer is Swin Transformer block, which performs feature extraction through a multihead self-attention mechanism.

The Swin Transformer [20] has been shown to be effective in constructing long-distance dependencies between buildings and natural features in patch images and can model multiscale features. Therefore, in this study, a Swin Transformer based on window multihead self-attention (W-MSA) is used as the global feature extraction module.

The overall framework of the GMF module is shown in Fig. 3. The Swin Transformer effectively achieves global feature extraction by shifted window attention. The input patch image $I \in R^{H\times W\times C}$ is first reduced to a suitable range by a patch partition module. Each patch is then mapped to a higher dimensional feature space through a linear embedding module to improve the model's expressive ability. In the Swin Transformer, there are four scales, and each scale except for the first scale is realized by a patch merging layer to reduce the feature scale, which increases the receptive field of the next window attention operation on the input patch image and realizes multiscale feature extraction for the input patch image. In the third scale level, the network is deepened by stacking multiple Swin Transformer blocks to better capture the cross-window features of the input image and reduce the loss of feature information. The output of the global feature extraction module consists of the output feature $G_i(i = 2, 3, 4)$ from each patch merging layer.

*2) Local Multiscale Feature Extraction:* Although global contextual information is important for LCZ scene classification, the texture differences between different building categories in the LCZ classification system are relatively small, so deep mining of the rich spatial information contained in the local details also requires attention. ResNet is a commonly used basic model for LCZ classification tasks [9], [36], [37] as it effectively increases the feature utilization efficiency by preserving intermediate features through a shortcut connection, thus deeply mining image detail features. The traditional ResNet model, in order to accommodate the classification of large image sizes, is designed with a static layer containing a convolutional layer and a max-pooling layer, to reduce the input image size, thus reducing the computational effort. However, this static layer tends to lead to the loss of a large amount of detail and structural information, which is not conducive to the classification of RS imagery with dense feature distributions and small sizes. Therefore, in this article, we propose a space-maintained ResNet (SM-ResNet) model based on ResNet 34, which is adapted to the LCZ classification task.

SM-ResNet has 34 layers and uses downsampling by convolution to expand the receptive field and build multiscale structures. SM-ResNet extracts image features through a residual block, which consists of two convolutional layers and a shortcut connection. The shortcut connection retains more gradient information and feature information. In this study, the static layer was optimized and replaced with a convolutional layer with a kernel size of $1 \times 1$, as shown in Fig. 4, to enlarge the nonlinear features while keeping the input image size unchanged, so as to achieve interaction between the channel information to obtain the features of the channel dimension $M$. The extracted channel dimension features $M$ are then fed into the hierarchical structure of the network to obtain feature maps $L_i(i = 2, 3, 4)$ at different scales, which are used for the MLF.

*3) Multilevel Feature Fusion:* In the proposed approach, the multilevel features include global and local scope features at the same scale and multiscale features after multiscope feature fusion. MLF is the process of combining features from different scopes and scales to obtain more accurate classification results, and it is worth exploring how to fuse multiscale global and local features to achieve the best representation of LCZs. The structure of the MLF module is shown in Fig. 5.

The squeeze-and-excitation (SE) module [38] is a classical channel attention module that can be used for feature fusion. However, the direct use of the SE module in the RS image classification task does not effectively improve the accuracy [39]. Therefore, in this study, an improved SE module (iSE) [39] is used to aggregate the global and local scope features to obtain more comprehensive and rich semantic information and to improve the model's comprehensive recognition ability for the overall structure and local details of the image.

The global features $G_i$ and local features $L_i$ of the corresponding scales are concatenated to achieve feature compression through the global average pooling layer, and the weights of the channel dimension are obtained by two fully connected layers. The iSE module adds a batch normalization layer after each fully connected layer, which normalizes the input value to a distribution with a mean of 0 and a variance of 1. This ensures the effectiveness of the gradient during training. Meanwhile, the activation function is replaced with the HardTanh function. The HardTanh function is implemented by a simple thresholding operation, which is more computationally efficient and has a stronger nonlinear representation ability to better fit complex feature distributions.

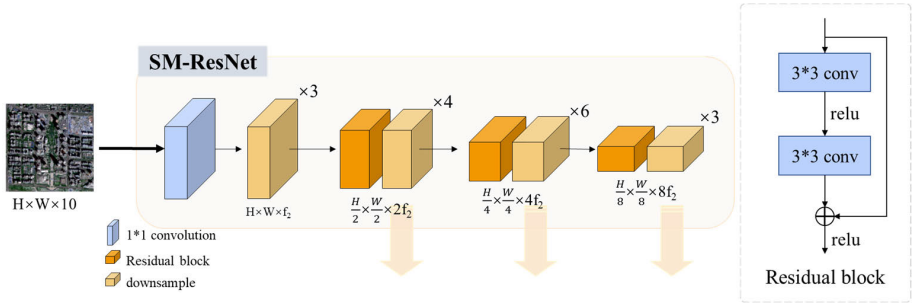



Fig. 4. SM-ResNet: extractor for the LMF module. The input of the LMF module is the original image patch and the output is the features at the last three scales of the LMF module. SM-ResNet based on ResNet 34 adds a convolutional layer with a kernel size of 1 × 1 to the first layer, which preserves spatial information while acquiring channel features and adequately extracts the features of small-size images.

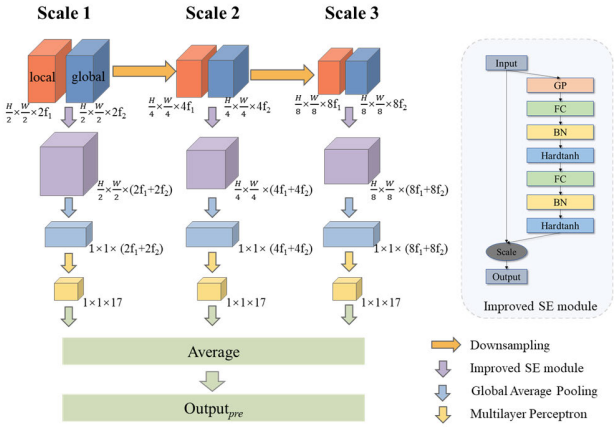



Fig. 5. Structure of the MLF module.

The fusion features $GL_i$ are obtained by mapping the channel weights to the concatenated features. The fusion of multiscope features, both global and local, reduces the shortcomings of individual features and improves the performance of the model.

Multilayer cascading can avoid, as much as possible, the loss of effective feature information caused by the layer transformation process [18], [40]. In the proposed approach, MLF is realized by combining the multiscope features and cascading the multiscale fusion features, which effectively complements the local and global feature advantages and preserves the feature information at different scales.

The fused multiscope features $GL_i$ of the three scales are mapped to a one-dimensional (1-D) space through a global average pooling layer. The mapped 1-D features are subject to complex linear transformations through a multilayer perceptron to fit the classification features and obtain the corresponding hierarchical classification results $F_i$. In order to take full advantage of the prediction results of each level, the results of the four levels are averaged and used for the loss calculation and optimization. The average result is the initial prediction $O_{\text{pre}}$. The multilevel feature cascade allows the model to capture local and global information from multiple scales of the input image for a more comprehensive understanding and representation of the input image. Furthermore, the final result comes from the feature mapping of all scales, rather than a single output, to complete the classification, and the involvement of multilevel features makes the model more accurate

$$F_i = \text{MLP}(\text{AvgP}(\text{GL}_i)), \quad (i = 2, 3, 4) \tag{1}$$

$$O_{\text{pre}} = \text{Average}(F_2, F_3, F_4) \tag{2}$$

where $GL_i$ represents the fused multiscope features at the $i$th scale, $F_i$ is the classification results at the $i$th scale, and $O_{\text{pre}}$ is the initial prediction results. AvgP is the average pooling layer and MLP is a multilayer perceptron layer.

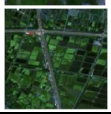
### B. Prior Knowledge Mining and Coupling

*1) Prior Knowledge Mining for LCZs:* According to the definition of LCZs, the physical attributes of some categories are very close and their presentations are very similar in RS images. We studied the 17 LCZ categories in depth and found the most similar sets, which are open low rise (LCZ 6) versus sparsely built (LCZ 9) and bush (LCZ C) versus low plants (LCZ D), as shown in Table I. It can be seen that open low rise and sparsely built have the same height of roughness elements as well as the same terrain roughness, and their sky view factors are very similar, with bush and low plants identical in terms of building surface fraction, impervious surface fraction, and pervious surface fraction, and they are similar in terms of aspect ratio and height of roughness elements. These small differences in characteristics make it difficult to discriminate these two similar sets from the 17 categories. The last column of Table I shows the Sentinel-2 images for these two sets of categories, where the similarity of these two sets in the RS images can be seen.

In addition to qualitatively analyzing the difficulty in distinguishing these two sets of categories, quantitatively categorizing these two similar sets from 17 categories verifies the difficulty of separating them. The discriminability of these two LCZ sets was examined by using four state-of-the-art classification networks based on the large So2Sat LCZ42 dataset. The confusion matrices for these four networks are shown in Fig. 6. The accuracies for sparsely built and bush are very low, and all models except the Swin Transformer do not reach a 10% accuracy on the bush class, which seriously lowers the OA of the LCZ classification. The significantly higher accuracy for bush with the Swin Transformer suggests that the global features extracted by the Swin Transformer are better suited to differentiate bush from low plants. The low accuracy for sparsely built and bush is mainly due to the fact that sparsely built is misclassified as open low rise in

TABLE I

VALUES OF THE GEOMETRIC AND SURFACE COVER PROPERTIES OF THE LCZS [1]. ALL PROPERTIES ARE UNITLESS, EXPECT HEIGHT OF ROUGHNESS ELEMENTS (m)

| Local climate zone (LCZ) | Sky view factor | Aspect ratio | Building surface fraction | Impervious surface fraction | Pervious surface fraction | Height of roughness elements | Terrain roughness class | Sentinel-2 patches |
|---|---|---|---|---|---|---|---|---|
| LCZ 6 Open low-rise | 0.6–0.9 | 0.3–0.75 | 20–40 | 20–50 | 30–60 | 3–10 | 5–6 |  |
| LCZ 9 Sparsely built | >0.8 | 0.1–0.25 | 10–20 | <20 | 60–80 | 3–10 | 5–6 |  |
| LCZ C Bush, scrub | 0.7–0.9 | 0.25–1.0 | <10 | <10 | >90 | <2 | 4–5 |  |
| LCZ D Low plants | >0.9 | <0.1 | <10 | <10 | >90 | <1 | 3–4 |  |

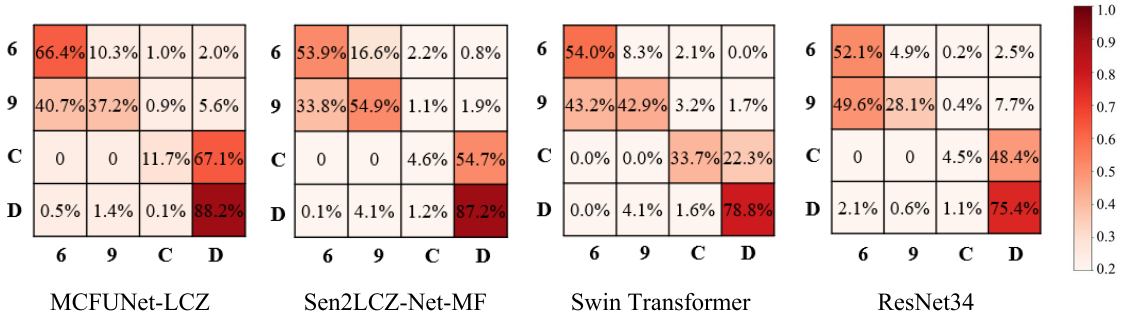
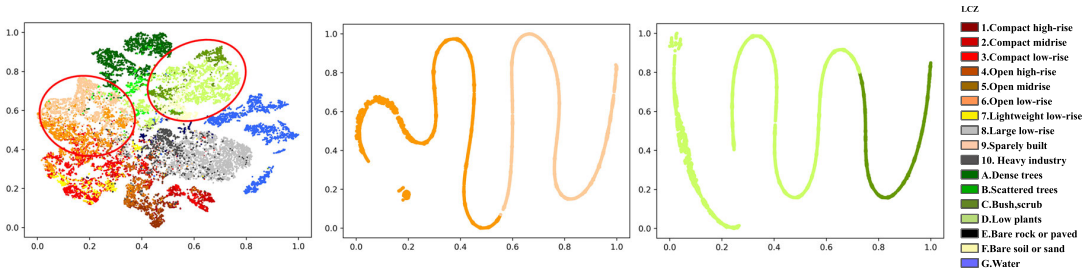



Fig. 6. Confusion matrices for MCFUNet-LCZ, Sen2LCZ-Net-MF, Swin Transformer, and ResNet34 on the So2Sat LCZ42 dataset. The horizontal and vertical coordinates of the confusion matrix represent the LCZ categories, where 6 represents open low rise, 9 represents sparsely built, C represents bush, and D represents low plants.




(a) (b) (c)

Fig. 7. Visualization of features before the fully connected layer by t-SNE. (a) Feature visualization for the 17 LCZ classes. (b) Feature visualization for the two-class classification of LCZs 6 and 9. (c) Feature visualization for the two-class classification of LCZs C and D. It can be observed that the feature points of LCZ C versus LCZ D and LCZ 6 versus LCZ 9 have a lot of overlap and are indistinguishable in the 17-class feature visualization but are distinguishable in the two-class classification.

upward of 40% of the cases, and bush is misclassified as low plants in even more than 50% of the cases, which illustrates the difficulty of distinguishing these two sets of categories in the 17-class classification for the deep learning model. It can also be seen from the visualization of the features in Fig. 7 that these two sets of categories are easily confused, but the differentiation between these two sets of categories is relatively clear in the separate two-category classifier.

Therefore, in the proposed approach, this property, which is easily confounded in multiclass classification but separable in two-class classification, is used as prior knowledge to guide the two difficult sets of classes using a two-class classifier to achieve the category accuracy.

*2) Coupling Strategy:* The specific PKC strategy is shown in Fig. 8. In order to ensure the training efficiency of the model, we use the lightweight ResNet34 model as the two-class classifier. First, the PKC module extracts the indices belonging to these two sets of categories from the model's initial prediction results $O_{\text{pre}}$. The images corresponding to these two categories are then obtained by masking from the original images based on the indices and repredicted by the trained two-class ResNet34. Finally, the repredicted categories

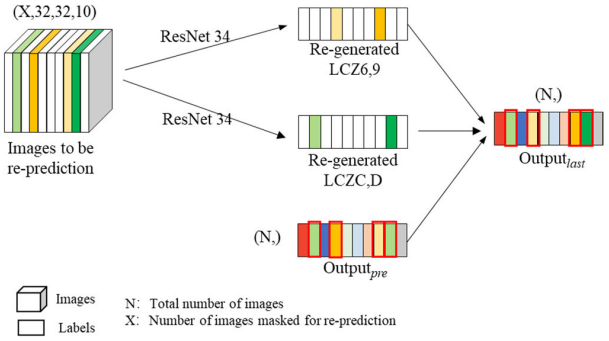



Fig. 8. Specific coupling details for the PKC module.

and the original initial prediction categories $O_{\text{pre}}$ are integrated to obtain the final prediction result $O_{\text{last}}$. The PKC module is utilized to individually extract features that can be easily distinguished between these categories, allowing the model to better distinguish between the easily confounded categories, resulting in more accurate LCZ classification results.

### C. Loss Function

As mentioned before, a variety of objects are contained in an LCZ category, and assigning hard labels to each patch can easily mislead the model. In order to avoid the model having too strong trust for the original labels, we use a label smoothing loss function to solve the problem. Label smoothing [41] is a regularization method that adds random noise to the original one-hot labels. Using label smoothing can train the model more robustly and at the same time works better in multiple classification tasks. Label smoothing regularization (LSR) has been used in many image classification models [42], [43], [44], [45].

Label smoothing transforms hard labels into soft labels, which makes the network optimization smoother. The label smoothing loss can be expressed as follows:

$$H(q', p) = (1 - \varepsilon)H(q, p) + \varepsilon H(u, p) \tag{3}$$

where $p$ represents the predicted value, $q$ refers to the original labeled value, $q'$ refers to the smoothed labeled value, $u$ is the uniform distribution, $H(q, p)$ represents the cross-entropy loss, and $\varepsilon$ is the smoothing factor. The loss function is the cross-entropy loss when the smoothing parameter $\varepsilon$ is 0.

The LSR loss smooths the assignment of true labels to other categories, thus reducing the over-reliance on individual categories and improving the generalization ability and robustness of the model. In addition, the LSR loss allows for more compact aggregation between samples of the same class, increasing the interclass distance and decreasing the intraclass distance, which allows the classifier to better distinguish the differences between different classes and obtain more discriminative features, thus achieving a higher classification accuracy.

## III. Experiments

### A. Dataset and Experimental Settings

The So2Sat LCZ42 dataset is one of the few publicly available large datasets for LCZ classification, which contains approximately 500 000 Sentinel-2 image patches and their LCZ labels for 42 urban agglomerations around the world (plus ten additional small regions). The So2Sat LCZ42 dataset was labeled by a group of domain experts following a workflow and evaluation process that was similar to WUDAPT. The So2Sat LCZ42 dataset also underwent a rigorous quality assessment. The overall confidence level is 85%, making this dataset a benchmark dataset for high-quality LCZ classification. It has been shown that image sizes ranging from 32 × 32 to 64 × 64 help deep learning models to obtain better LCZ classification results [10]. The patch size in the So2Sat LCZ42 dataset is 32 × 32, which is a very suitable design in terms of image size. In this study, we choose the version [16] (http://doi.org/10.14459/2018mp1483140) that is divided by cities as the basic unit so that the training set and the test set are completely separated geospatially.

In terms of parameterization, we experimentally set the window size to 4 and the patch size to 1 to save spatial information. The number of Swin Transformer blocks in each scale was set to 2, 2, 6, and 2. The batch size is set to 64. The initial learning rate was set to 1e−3. A cosine-warm-up function was used as the learning rate decay function, in which eight epochs were set for the warm-up. The learning rate decay function can make the model converge faster and obtain better results. In order to control the training time and avoid overfitting, we used the method of early stopping, which stops the training if there is no increase in the verification accuracy for 40 epochs. After training, we take the weights with the highest validation accuracy as the final weights.

In order to quantitatively evaluate the performance of the model, we adopt the precision evaluation metrics of OA, Kappa, and confusion matrix. For the qualitative analysis, eight cities around the world are selected for the LCZ mapping: Wuhan, the Guangzhou–Foshan area, Beijing, Shanghai, Tokyo, New York, Los Angeles, and Sydney. Among these cities, the first two are used as visual comparisons between our proposed method and four advanced methods, thus visually evaluating the proposed method.

### B. Comparative Experiments

In the comparative experiments, 12 networks are selected as benchmark methods, namely, Sen2LCZ-Net [17], LCZ-CNN [10], MCFUNet-LCZ [18], DenseNet [46], SM-ResNet, DeepLabv3+ [47], U-Net [48], Xception [49], Swin Transformer [20], Vision Transformer [50], and ExViT [21], as shown in Table II, the first three of which are networks specifically designed for LCZ mapping. The experimental results show that the LCZ-MFKNet method proposed in this article can reach 0.738 in OA and 0.712 in Kappa, which is the highest accuracy obtained on the So2Sat LCZ42 dataset.

The classification effect of SM-ResNet is significantly better than the other compared CNN models, which is mainly because preserving spatial information as much as possible is very beneficial for small-size image classification. However, SM-ResNet does not extract global features, resulting in limited classification accuracy in the small-size scene image classification. ExViT, the best-performing model of the compared Transformer architectures extends the processing of RS

TABLE II
PERFORMANCE COMPARISON AMONG THE 12 NETWORKS ON THE SO2SAT LCZ42 DATASET

| | Method | OA | Kappa |
|---|---|---|---|
| CNN | DenseNet [46] | 0.663 | 0.631 |
| | SM-ResNet | 0.672 | 0.649 |
| | DeepLabv3+ [47] | 0.657 | 0.626 |
| | U-Net [48] | 0.614 | 0.572 |
| | Xception [49] | 0.654 | 0.604 |
| Transformer | Swin Transfomer [20] | 0.659 | 0.625 |
| | Vision Transformer [50] | 0.588 | 0.553 |
| | ExViT [21] | 0.663 | 0.630 |
| LCZ | Sen2LCZ-Net [17] | 0.677 | 0.646 |
| | LCZ-CNN [10] | 0.651 | 0.614 |
| | MCFUNet-LCZ [18] | 0.700 | 0.680 |
| Proposed | LCZ-MFKNet | **0.738** | **0.712** |

images by incorporating separable convolution modules on top of the vision Transformer. This approach fully leverages spatial and modal channel information, resulting in a 7.5% higher accuracy compared to the Vision Transformer. The accuracy of the Swin Transformer is 7.1% higher than that of the Vision Transformer, which is because the multiscale features extracted by the Swin Transformer help the LCZ classification. The advanced LCZ classification networks of Sen2LCZ-Net and MCFUNet-LCZ obtain better classification results on the So2Sat LCZ42 dataset than all the compared CNN models. Sen2LCZ-Net and MCFUNet-LCZ both adopt the strategy of multilayer feature cascading, which is an effective way to strengthen the cascade of features between network layers. The LCZ-MFKNet method proposed in this article also draws on these findings, using both global and local feature extraction networks with a multilevel structure and cascading the multilevel features in the feature fusion module. In addition to multilevel feature extraction, the optimization of the model with prior knowledge is also very significant for accuracy improvement. Overall, LCZ-MFKNet obtains the best classification performance, compared to the other benchmark methods.

The confusion matrices for SM-ResNet, ExViT, Sen2LCZ-Net, and LCZ-MFKNet are shown in Fig. 9. The accuracy of SM-ResNet, ExViT, and Sen2LCZ-Net in LCZ 9 and LCZ C is clearly lower than that of the proposed LCZ-MFKNet method, and the accuracy for LCZ C is even less than 10%. The accuracy of the proposed LCZ-MFKNet method in LCZ 9 reaches 71.9%, and the accuracy for LCZ C reaches 69.8%. This shows that the prior knowledge module can better extract features that distinguish similar categories, thus improving the classification accuracy. Swin Transformer achieves an accuracy of only 19.5% in LCZ 1. The proposed model combines the advantages of SM-ResNet and Swin Transformer to achieve a higher accuracy in LCZ 1. In all four models, the categories of compact low rise (LCZ 3), large low rise (LCZ 8), sparsely built (LCZ 9), dense trees (LCZ A), low plants (LCZ D), bare soil (LCZ F), and water (LCZ G) all show a high classification accuracy of over 70%. However, lightweight low rise (LCZ 7) has a low accuracy, which is mainly because the limitation of the image resolution makes the subtle difference between the classes difficult to recognize, even with a two-class classifier, and the annotations for LCZ 7 in So2Sat LCZ42 are also not sufficiently confident.

TABLE III
ABLATION STUDY FOR EACH MODULE OF LCZ-MFKNET

| GMF | LMF | MLF | PKC | OA | Kappa |
|---|---|---|---|---|---|
| √ | | | | 0.669 | 0.637 |
| | √ | | | 0.684 | 0.658 |
| √ | √ | | | 0.693 | 0.662 |
| √ | √ | √ | | 0.710 | 0.680 |
| √ | √ | √ | √ | 0.738 | 0.712 |

* GMF: global multi-scale feature extraction module; LMF: local multi-scale feature extraction module;
MLF: multi-level feature fusion module;
PKC: prior knowledge coupling module.

### C. Ablation Study

An ablation study is conducted for each module in LCZ-MFKNet: GMF, LMF, MLF, and PKC, to analyze the accuracy contribution of each module. The specific results are listed in Table III. The ablation experiments are trained using the label smoothing loss function.

The first and second rows show the results for the GMF and LMF modules, with Kappa coefficients of 0.637 and 0.658, respectively. The Swin Transformer is able to capture a

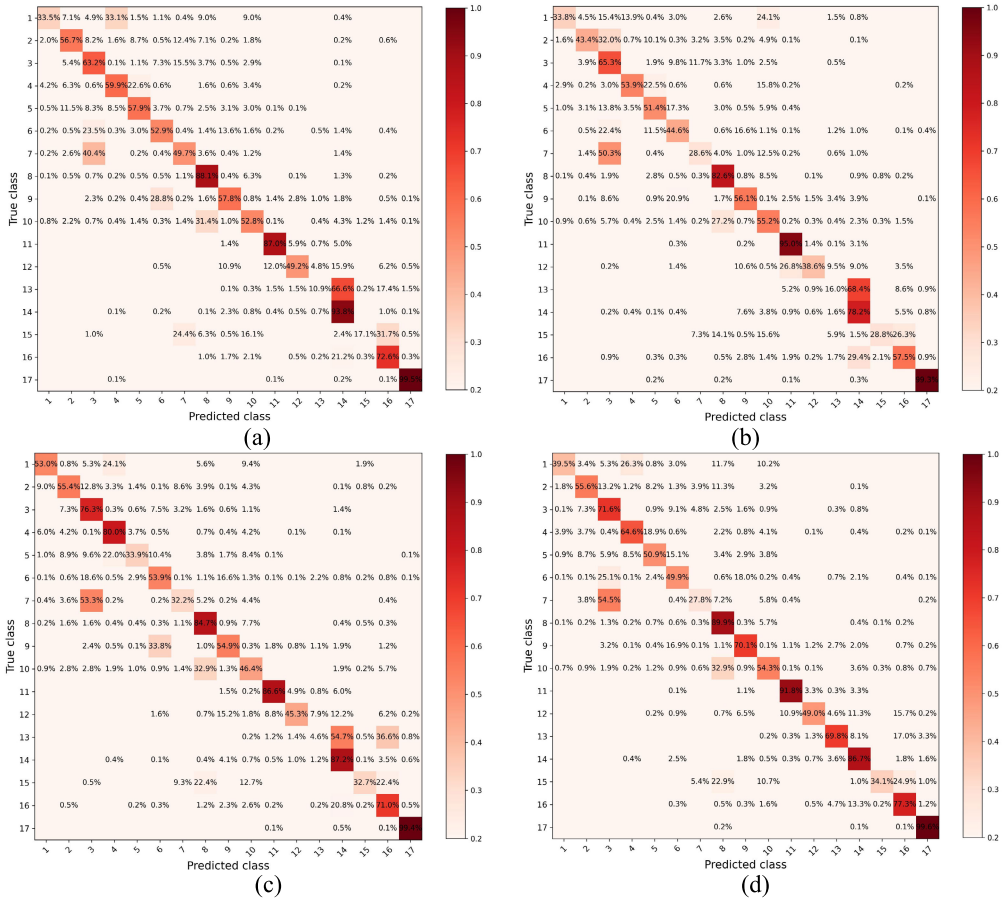



Fig. 9. Confusion matrices of the classification results of (a) SM-ResNet, (b) ExViT, (c) Sen2LCZ-Net, and (d) LCZ-MFKNet. The results are normalized by the total number of samples per LCZ.

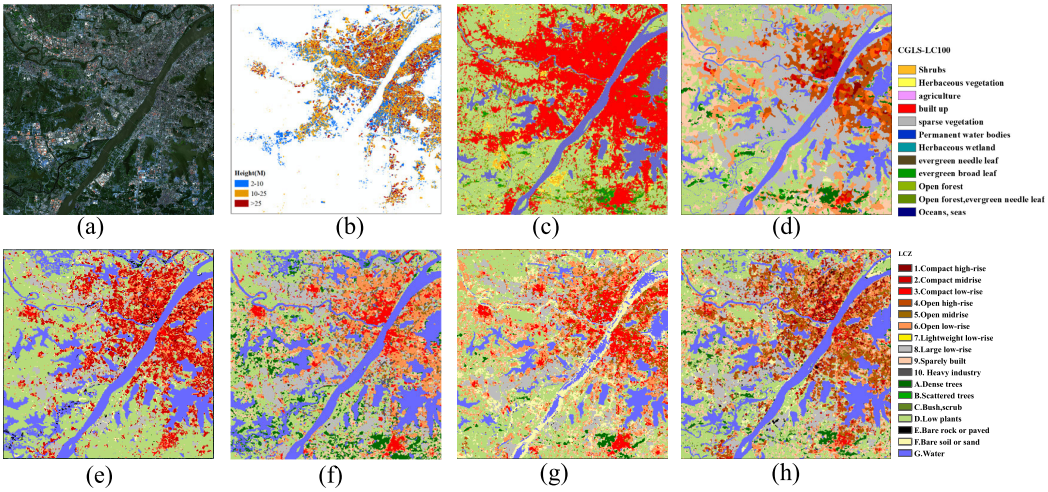



Fig. 10. Qualitative comparative analysis for Wuhan. (a) Real Sentinel-2 image. (b) and (c) Building height data and land classification data, respectively, which are qualitative references for the classification of LCZs in terms of building categories and natural categories. (d)–(h) Classification results of WUDAPT, LCZ-MFKNet, SM-ResNet, ExViT, and Sen2LCZ-Net in Wuhan, respectively.

wide range of contextual information and has an advantage in grasping the global features of an image. SM-ResNet makes full use of the spatial information of the image and is stronger in perceiving the details and local features of the image. The third and fourth rows show the accuracies obtained by normal superposition of the GMF and LMF modules and fusion by the MLF module, respectively, where it can be seen that the addition of the MLF module improves the accuracy by 1.7% over the simple superposition of the GMF and LMF modules, with an OA and Kappa of 0.710 and 0.680, respectively. This shows that MLF allows the global information and local features to be more effectively combined. The fusion of multiscope features and the comprehensive utilization of multilevel features enables the model to extract richer semantic information, which enhances the expressive ability of the model and facilitates LCZ classification. In order to verify the effectiveness of the PKC module, the fourth and fifth rows compare the accuracy before and after the use of the PKC

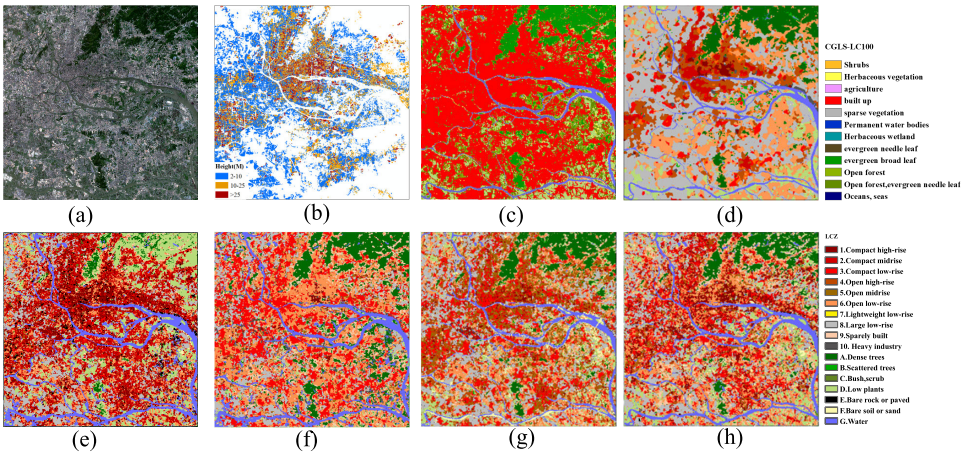



Fig. 11. Qualitative comparative analysis for Guangzhou–Foshan. (a) Real Sentinel-2 image. (b) and (c) Building height data and land classification data, respectively, which are qualitative references for the classification of LCZs in terms of building categories and natural categories. (d)–(h) Classification results of WUDAPT, LCZ-MFKNet, SM-ResNet, ExViT, and Sen2LCZ-Net in Guangzhou–Foshan, respectively.

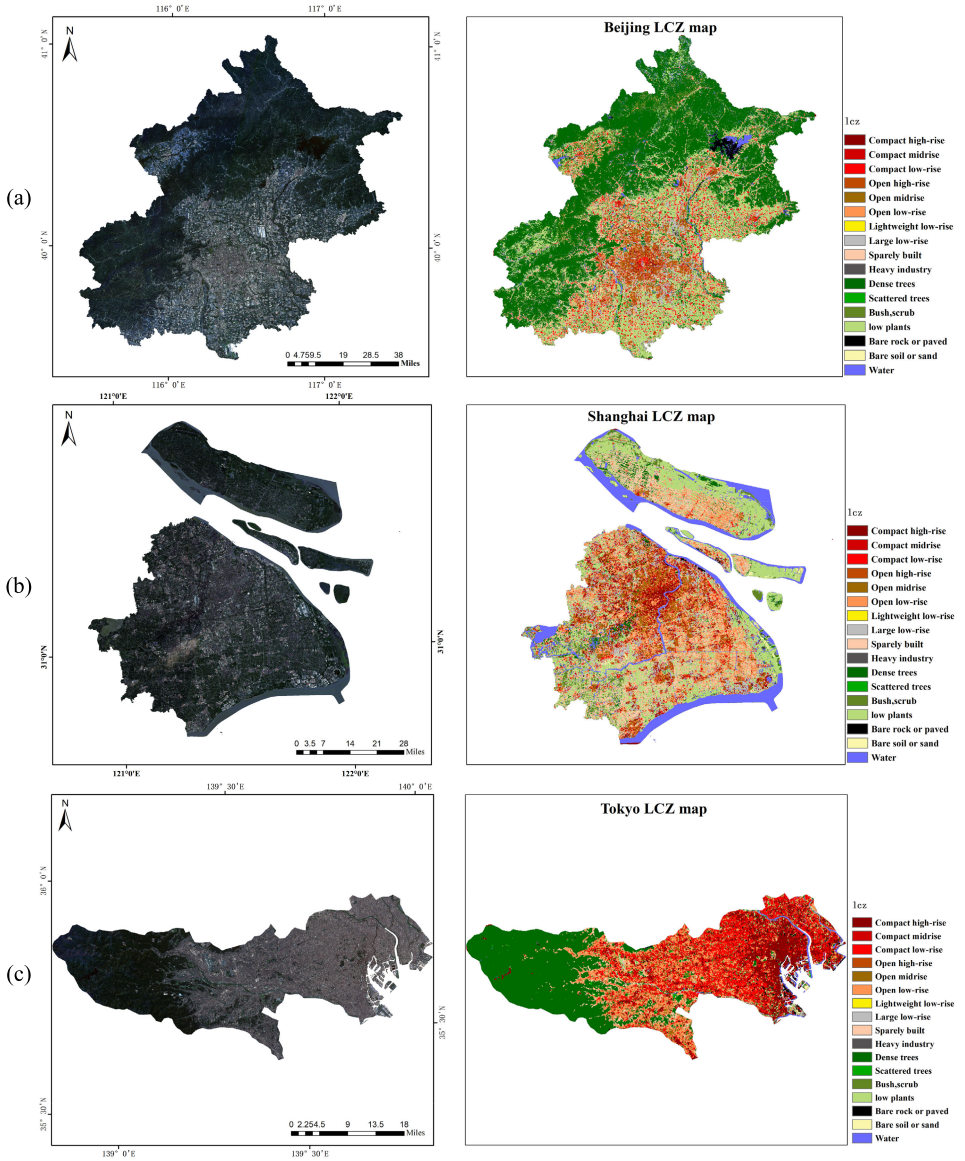



Fig. 12. LCZ maps of (a) Beijing, (b) Shanghai, and (c) Tokyo. The left column of images is the Sentinel-2 images, and the right is the corresponding LCZ maps.

module. The results show that the PKC module improves the Kappa by 3.2% and the OA by 2.8%. The introduction of prior knowledge guides the model to further differentiate against similar categories with poor classification accuracy,

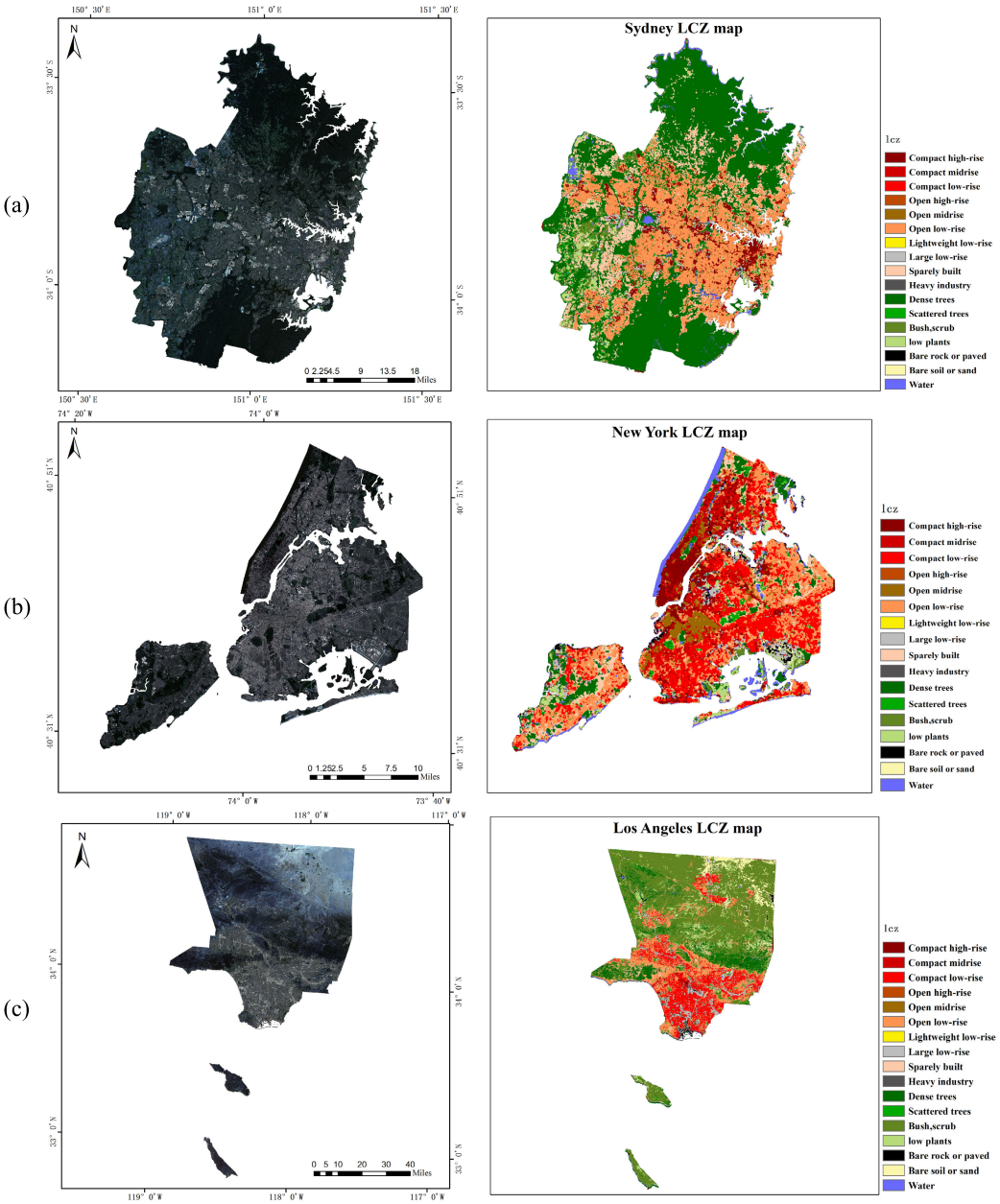



Fig. 13. LCZ maps of (a) Sydney, (b) New York, and (c) Los Angeles. The left column of images is the Sentinel-2 images, and the right is the corresponding LCZ maps.

which results in a significant accuracy improvement in the LCZ classification.

### D. LCZ Mapping

*1) Qualitative Analysis:* SM-ResNet and ExViT are the highest accuracy models among the compared CNN and Transformer networks, respectively. Sen2LCZ-Net is the classical LCZ classification method and MCFUNet-LCZ has no public code. Therefore, we compared the prediction results of the proposed method with those of WUDAPT, LCZ-MFKNet, SM-ResNet, ExViT, and Sen2LCZ-Net on two real Sentinel-2 images. We also analyzed in detail the effect of classifying the building categories and natural categories in the LCZ classification results by building height data from OpenStreetMap (OSM) and CGLS-LC100 data (global Land Cover product at 100-m spatial resolution).

Fig. 10 shows the model classification results for Wuhan. The obvious contrast in the building categories is the high-rise building categories. A building height greater than 25 m (dark red) in the building height data corresponds to high rise in the LCZ classification system, and based on the degree of clustering of buildings, it is clear that most of the high-rise area is compact high rise. SM-ResNet does not predict any high-rise areas, and it easily misclassifies high rise as mid rise. ExViT recognizes open low rise more accurately, but the high-rise buildings are not identified correctly. Sen2LCZ-Net misclassifies compact high rise as compact low rise. LCZ-MFKNet predicts almost the same distribution of compact high rise as that of high-rise building in the building height data,

indicating that the proposed method obtains the best results for compact high rise. Sen2LCZ-Net easily predicts high rise as low rise. WUDAPT can identify compact high rise, but many areas are misclassified as large low rise, which does not match the building height data. In terms of the natural categories, SM-ResNet does not recognize any dense trees and misclassifies low plants as water. Sen2LCZ-Net misclassifies some water as bare rock. The proposed LCZ-MFKNet method obtains better prediction results on both water and dense trees. Overall, in Wuhan, the proposed method achieves the best LCZ prediction results, in both the built-up areas and natural categories.

Fig. 11 shows the model classification results for Guangzhou–Foshan. Since Guangzhou–Foshan is not in the training study area, the prediction effect in this image can reflect the generalization ability of the model. In the built-up area, SM-ResNet recognizes many of the mid rise as compact high rise, and there are instances where buildings are misclassified as bare rock. Sen2LCZ-Net and LCZ-MFKNet are both better predictors in terms of building categories. In terms of the natural categories, ExViT performs well in dense tree and water, SM-ResNet tends to misclassify dense trees as low plants, and Sen2LCZ-Net misclassifies the water at the bottom of the image as bare rock. Overall, the proposed LCZ-MFKNet method obtains the best prediction results for Guangzhou–Foshan, indicating that the model has a strong generalization ability and good potential for large-scale LCZ mapping.

*2) LCZ Mapping of Megacities:* We selected six large cities from around the world as examples of LCZ mapping of megacities. Sentinel-2 images and the corresponding mapping results of LCZ-MFKNet for Beijing, Shanghai (China, Asia), Tokyo (Japan, Asia), New York, Los Angeles (USA, North America), and Sydney (Australia, Oceania) are presented in Figs. 12 and 13.

The distribution of Beijing's buildings is a circular pattern. The urban area of Beijing is dominated by compact low rise (LCZ 3) in the center of the city. There are increased building heights on the periphery of the city center, dominated by open high rise (LCZ 4), open middle rise (LCZ 5), and compact high rise (LCZ 1). The area along the Huangpu River in the center of Shanghai is an area of compact high rise (LCZ 1), and the areas far away from the city center have lower building density and height, with open middle rise (LCZ 5) and open low rise (LCZ 6). Tokyo is one of the most populous cities in the world and therefore has a high density of buildings, as land is limited and buildings are mostly hig -rise, so the built-up area is dominated by compact high rise (LCZ 1) and compact low rise (LCZ 3). Even though Kyoto is a high-density city, it also focuses on urban green space, so the natural category is dominated by dense trees (LCZ A).

The city center of Sydney is the area with the highest concentration of high-rise buildings, which is dominated by compact high rise (LCZ 1). Suburban areas have relatively low building heights and densities and are dominated by open low rise (LCZ 6) and sparsely built (LCZ 9). New York is known for its high-density built-up areas. The Manhattan area has a large number of high-rise buildings, predominantly compact high rise (LCZ 1), with a significantly higher density of buildings than other boroughs. The central Los Angeles area is densely built and dominated by compact high rise (LCZ 1) and compact low rise (LCZ 3). The suburban areas have relatively low building heights and densities and are dominated by open low rise (LCZ 6) and large low rise (LCZ 8).

Overall, the LCZ classification mapping results for the six megacities are consistent with our perception, with good visualization and fine and accurate classification results.

## IV. Conclusion

Large-scale accurate mapping of LCZs is useful for studying climate differences across cities. Most of the existing LCZ classification methods are trained based on small datasets with an inadequate generalization ability, while LCZ classification methods based on large datasets cannot easily achieve a high accuracy. Therefore, in this article, we propose an LCZ classification model coupling multilevel features with prior knowledge based on the available large So2Sat LCZ42 dataset. The multilevel feature extraction and fusion framework allows both global and local features of the imagery to be acquired and multiscale features to be modeled, and these fully extracted features lead to high accuracy. Based on the analysis of the physical properties of the 17 classes of LCZs and the results of a large number of quantitative experiments, we find that some similar categories, although difficult to classify in 17 classes, are classable in two classes, and we used such attributes as prior knowledge to optimize the model. The experimental results fully proved the effectiveness of the prior knowledge in the improvement of the accuracy. The good visualization results obtained mapping six megacities at a global scale demonstrate the generalization capability of the proposed LCZ-MFKNet method, which can realize high-precision LCZ mapping at a large scale.

In the future, the LCZ mapping capability can also be improved by making full use of richer and more diverse information by combining multimodal data, in addition to only using optical RS datasets. The anticipation for a higher precision global coverage LCZ dataset is worth it.

## Acknowledgment

The numerical calculations in this article have been done on the supercomputing system at the Supercomputing Center, Wuhan University.

## References

[1] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Amer. Meteorological Soc.*, vol. 93, no. 12, pp. 1879–1900, Dec. 2012.

[2] E. Lelovics, J. Unger, T. Gál, and C. Gál, "Design of an urban monitoring network based on local climate zone mapping and temperature pattern modelling," *Climate Res.*, vol. 60, no. 1, pp. 51–62, May 2014.

[3] M. Demuzere et al., "A global map of local climate zones to support Earth system modelling and urban-scale environmental science," *Earth Syst. Sci. Data*, vol. 14, no. 8, pp. 3835–3873, Aug. 2022.

[4] B. Bechtel et al., "Mapping local climate zones for a worldwide database of the form and function of cities," *ISPRS Int. J. Geo-Information*, vol. 4, no. 1, pp. 199–219, Feb. 2015.

[5] C. Ren et al., "Assessment of local climate zone classification maps of cities in China and feasible refinements," *Sci. Rep.*, vol. 9, no. 1, p. 18848, Dec. 2019.

[6] B. Bechtel et al., "Generating WUDAPT level 0 data–current status of production and evaluation," *Urban Climate*, vol. 27, pp. 24–45, Mar. 2019.

[7] F. Huang et al., "Mapping local climate zones for cities: A large review," *Remote Sens. Environ.*, vol. 292, Jul. 2023, Art. no. 113573.

[8] J. Rosentreter, R. Hagensieker, and B. Waske, "Towards large-scale mapping of local climate zones using multitemporal sentinel 2 data and convolutional neural networks," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111472.

[9] M. Kim, D. Jeong, and Y. Kim, "Local climate zone classification using a multi-scale, multi-level attention network," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 345–366, Nov. 2021.

[10] S. Liu and Q. Shi, "Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China," *ISPRS J. Photogramm. Remote Sens.*, vol. 164, pp. 229–242, Jun. 2020.

[11] F. H. Chen and J. Y. Tsou, "Mapping urban form and land use with deep learning techniques: A case study of Dongguan city, China," *Int. J. Oil Gas Coal Technol.*, vol. 29, pp. 306–328, Jun. 2022.

[12] F. Chen and J. Y. Tsou, "DRSNet: Novel architecture for small patch and low-resolution remote sensing image scene classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 104, Dec. 2021, Art. no. 102577.

[13] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018.

[14] A. Ma, C. Zheng, J. Wang, and Y. Zhong, "Domain adaptive land-cover classification via local consistency and global diversity," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606317.

[15] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856.

[16] X. X. Zhu 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. Replaces Newsletter*, vol. 8, no. 3, pp. 76–89, Sep. 2020.

[17] C. Qiu, X. Tong, M. Schmitt, B. Bechtel, and X. X. Zhu, "Multilevel feature fusion-based CNN for local climate zone classification from Sentinel-2 images: Benchmark results on the So2Sat LCZ42 dataset," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2793–2806, 2020.

[18] W. Ji, Y. Chen, K. Li, and X. Dai, "Multicascaded feature fusion-based deep learning network for local climate zone classification based on the So2Sat LCZ42 benchmark dataset," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 449–467, 2023.

[19] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[20] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[21] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.

[22] D. Hong et al., "SpectralGPT: Spectral foundation model," 2023, *arXiv:2311.07113*.

[23] X. Ding et al., "Prior knowledge-based deep learning method for indoor object recognition and application," *Syst. Sci. Control Eng.*, vol. 6, no. 1, pp. 249–257, Jan. 2018.

[24] X. Liu, J. Yuan, and T. Wang, "Interior human action recognition method based on prior knowledge of scene," *Comput. Sci.*, vol. 49, pp. 225–232, Mar. 2022.

[25] S. Baier, Y. Ma, and V. Tresp, "Improving visual relationship detection using semantic modeling of scene descriptions," in *Proc. Semantic Web ISWC*, 2017, pp. 53–68, doi: 10.1007/978-3-319-68288-4_4.

[26] S. Baier, Y. Ma, and V. Tresp, "Improving information extraction from images with learned semantic models," in *Proc. Twenty-Seventh Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 5214–5218.

[27] S. Sharifzadeh, S. M. Baharlou, and V. Tresp, "Classification by attention: Scene graph classification with prior knowledge," in *Proc. 35th AAAI Conf. Artif. Intell., 33rd Conf. Innov. Appl. Artif. Intell. 11th Symp. Educ. Adv. Artif. Intell.*, 2021, pp. 5025–5033.

[28] Y. Xue, X. Li, Z. Li, and C. Cao, "Prior knowledge-based retrieval and validation of information from remote-sensing data at various scales," *Int. J. Remote Sens.*, vol. 33, no. 3, pp. 665–673, Feb. 2012.

[29] X. Ye et al., "A joint-training two-stage method for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4709616.

[30] A. Li, J. Jiang, J. Bian, and W. Deng, "Combining the matter element model with the associated function of probability transformation for multi-source remote sensing data classification in mountainous regions," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, pp. 80–92, Jan. 2012.

[31] C. Xu, Z. Chen, and R. Hou, "Deep learning classification method of Landsat 8 OLI images based on inaccurate prior knowledge," *J. Comput.*, vol. 40, pp. 3550–3557, Dec. 2020.

[32] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513412.

[33] T. Wu, J. Luo, L. Xia, Z. Shen, and X. Hu, "Prior knowledge-based automatic object-oriented hierarchical classification for updating detailed land cover maps," *J. Indian Soc. Remote Sens.*, vol. 43, no. 4, pp. 653–669, Dec. 2015.

[34] H. Ji, Z. Gao, Y. Zhang, Y. Wan, C. Li, and T. Mei, "Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625513.

[35] S. Danfeng, L. Hong, and L. Pei, "The application of the method combine prior knowledge and fuzzy adaptive resonance theory map in remote sensing classification," in *Proc. Int. Conf. Info-Tech Info-Net.*, 2001, pp. 326–331.

[36] C. Qiu, M. Schmitt, L. Mou, P. Ghamisi, and X. Zhu, "Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets," *Remote Sens.*, vol. 10, no. 10, p. 1572, Oct. 2018.

[37] Q. Chunping, M. Schmitt, M. Lichao, and Z. Xiaoxiang, "Urban local climate zone classification with a residual convolutional neural network and multi-seasonal Sentinel-2 images," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens. (PRRS)*, Aug. 2018, pp. 1–5.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[39] D. Yu, H. Guo, Q. Xu, J. Lu, C. Zhao, and Y. Lin, "Hierarchical attention and bilinear fusion for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6372–6383, 2020.

[40] S. Sukhanov, I. Tankoyeu, J. Louradour, R. Heremans, D. Trofimova, and C. Debes, "Multilevel ensembling for local climate zones classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 1201–1204.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[42] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based DenseNet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, 2020.

[43] B. Hou et al., "Panchromatic image land cover classification via DCNN with updating iteration strategy," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 1472–1475, doi: 10.1109/IGARSS39084.2020.9323700.

[44] W. Wang, Y. Chen, and P. Ghamisi, "Transferring CNN with adaptive learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5533918.

[45] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–10.

[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[47] L. C. E. Chen, Y. K. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Comput. Vis. ECCV*, 2018, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[49] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

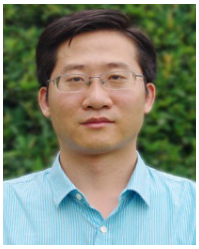
[50] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

**Xinrun Zhong** received the B.S degree in geographic information science from Lanzhou University, Lanzhou, China in 2021. He is currently pursuing the M.S degree in human geography with the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China.

Her research interests include deep learning, computer vision, and local climate zone classification.

**Huifang Li** (Member, IEEE) received the B.S. degree in geographical information science from the China University of Mining and Technology, Xuzhou, China, in 2008, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013.

She is currently a Professor with the School of Resources and Environmental Science, Wuhan University. She focuses on the study of urban remote sensing, including the complicated radiometric correction of optical remote sensing images, spatial–temporal continuous land surface temperature reconstruction, and urban heat island monitoring and analysis.



**Huanfeng Shen** (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

He is currently a Distinguished Professor with Wuhan University, where he is also the Dean of the School of Resource and Environmental Sciences. He was or is the Principal Investigator (PI) of two projects supported by the National Key Research and Development Program of China and six projects supported by the National Natural Science Foundation of China. He has authored or coauthored more than 150 peer-reviewed international journal articles, where over 60 appeared in IEEE journals, and published four books as a Chief Editor. His research interests include remote sensing image processing, multisource data fusion, and intelligent environmental sensing.

Dr. Shen is a fellow of the Institution of Engineering and Technology (IET), an Education Committee Member of the Chinese Society for Geodesy Photogrammetry and Cartography, and a Theory Committee Member of the Chinese Society for Geospatial Information Society. He was a recipient of the First Prize in Natural Science Award of Hubei Province in 2011, the First Prize in Nature Scientific Award of China's Ministry of Education in 2015, and the First Prize in Scientific and Technological Progress Award of Chinese Society for Geodesy Photogrammetry and Cartography in 2017. He is also a Senior Regional Editor of the *Journal of Applied Remote Sensing* and an Associate Editor of Geography and Geo-Information Science and *Journal of Remote Sensing*.

**Meiling Gao** received the B.Sc. degree in geographical information science from Chang'an University, Xi'an, China, in 2014, and the Ph.D. degree in cartography and geographic information system from Wuhan University, Wuhan, China, in 2019.

She is now a Lecturer with the School of Geology Engineering and Geomatics, Chang'an University, Xi'an, China. Her main research interests include spatiotemporal data reconstruction and urban thermal environment.

**Zhihua Wang** received the B.Eng and M.Eng degrees in civil and environmental engineering from Nanyang Technological University, Singapore, in 2002 and 2004, respectively, and the Ph.D. degree in civil and environmental engineering from Princeton University, NJ, USA, in 2011.

He joined the School of Sustainable Engineering and the Built Environment, Arizona State University, as an Assistant Professor in 2012, where he has been an Associate Professor since 2018.

Dr. Wang is a member of the American Physical Society, the American Meteorological Society, the American Geophysical Union, the American Society and Civil Engineers, and the International Association for Urban Climate.

**Jinqiang He** is currently studying at the China University of Mining and Technology, Xuzhou, China. His research interests include remote sensing and deep learning.