A Scene Unmixing Deep Learning Network for Local Climate Zone Mapping and Analysis Using Very High Resolution Remote Sensing Imagery

Xinji Tian [1], Jiayi Li [1,2*],Xin Huang [1,3]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, 2022202130084@whu.edu.cn
[2]Key Laboratory of Urban Land Resources Monitoring and Simulation，Ministry of Natural Resources,Wuhan 430079, zjjerica@163.com
[3]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, xhuang@whu.edu.cn

Keywords: Local Climate Zone Mapping, Scene Unmixing, Deep Learning, Very High Resolution, Remote Sensing Imagery.
.

ABSTRACT:

With the acceleration of urbanization, the environment and climate necessary for our survival have gradually deteriorated, leading to the increasing prominence of the Urban Heat Island (UHI) effect. Local Climate Zone (LCZ) classification, as a standard of urban morphology, has become an essential tool for monitoring the UHI effect and conducting temperature studies. Deep Learning (DL) models have the ability to represent high-level semantic features. Therefore, this paper proposes a mixed scene unmixing DL framework for LCZ mapping and analysis using Very High Resolution (VHR) remote sensing images. This framework consists of a two-stream deep network, including a pure scene classification network (PS-Net) and a mixed scene unmixing network (MSU-Net). We conducted random sampling tests in Wuhan, China in the experiment A. The results show that this model achieved a satisfactory accuracy with the Overall Accuracies (OAs) is 96.78% and a mixed scene unmixing Mean Absolute Error (MAE) of 0.0495. Furthermore, we applied the proposed model to generate LCZ map for five districts in Wuhan in the experiment B. The test accuracy between two experiments differs very slightly. These results demonstrate the applicability and potential of our model for LCZ mapping and urban climate analysis.

1. Introduction

With the rapid advancement of urbanization and the continuous growth of urban population, urban land and resources are facing unprecedented pressure. Meanwhile, heat emissions accompanying human activities in daily life lead to the gradual increase in urban temperatures and aggravate the Urban Heat Island (UHI) effect where temperatures in urban areas tend to be higher than those in surrounding rural or suburban areas. The UHI phenomenon is regarded as a crucial indicator for studying urban climate characteristics and has become an essential part of thermal environment research in recent years.(Fisher et al.,2006; Streutker,2003). In these studies, most researchers assess the UHI effect using straightforward classification schemes distinguishing between urban and rural areas. However, these approaches failed to depict the diversity of complex urban-rural structures. In order to better understand and assess the UHI effect, researchers urgently need a more refined and precise classification scheme.Stewart and Oke proposed the Local Climate Zone (LCZ) classification scheme in 2012(Stewart and Oke,2012), which has become the internationally recognized standard for urban morphology classification among researchers studying the UHI effect. Different from the simple urban-rural classification aforementioned, LCZs are defined by coverage spanning from hundreds of meters to several kilometers in horizontal scale, with homogeneous surface cover, structure, materials, and human activities. The LCZ scheme classifies inner urban areas based on vegetation cover, impervious surface fractions, building height, and texture and the entire LCZ classification system divided into "built" type LCZs (1-10) and "natural" type LCZs (LCZA-G), as shown in Figure 1. The LCZ concept is particularly well-suited for urban blocks, as it can precisely quantify underlying surface elements and consider the comprehensive impacts of each specific block on local climate conditions. Therefore, LCZs can help researchers identify the degree of UHI effects in different urban areas, accurately simulate climate variations within cities, and provide scientific basis for urban planning and climate adaptation(Huang et al.,2021).

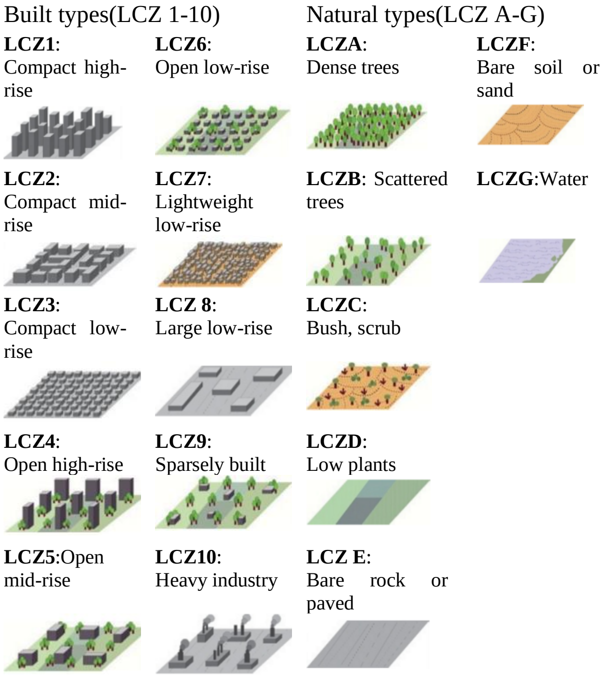
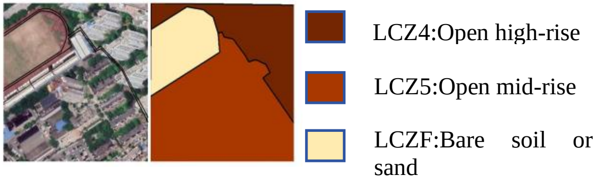


Figure1. This LCZ classification system is segmented into “built” types (LCZ 1－10) and “natural” types (LCZ A－G)(Stewart and Oke,2012).

In the past decade, some scholars have explored many computer-based methods to accurately classify LCZs, with one of the most widespread techniques being based on Geographic Information Systems (GIS) technology (Geletič and Lehnert,2016; Lelovics et al.,2014).In GIS, a large number of relevant feature parameters need to be incorporated, such as Sky View Factor (SVF), Building Surface Fraction (BSF), Fraction (ISF), Height-to-Width Ratio of Elements (HRE), and Aspect Ratio (AR)(Liu et al.,2023).However, in many cities, especially in developing ones, these required geographical feature parameters are often unavailable, thereby restricting the widespread application of GIS. Another commonly used approach involves utilizing remote sensing data from satellites and aerial platforms to provide abundant information for interpreting LCZs and generating global LCZ maps.For instance, The World Urban Database and Access Portal Tools (WUDAPT)(Ching et al.,2018) project employs Landsat satellite data, the Google Earth platform, and Random Forest (RF) classifiers to produce LCZ maps globally at a 100-meter resolution. Several researchers have achieved excellent results using the WUDAPT platform. For example, Kotharkar et al. designed an improved LCZ classification technique using WUDAPT to generate LCZ maps for Nagpur, India(Kotharkar and Bagade,2018). Likewise,Cai et al. used a WUDAPT-based approach to map LCZs in Guangzhou, China(Cai et al.,2016). Despite WUDAPT being widely applied to study the UHI phenomenon in more than 50 cities worldwide(Bechtel et al.,2019),only a limited number of cities have achieved satisfactory mapping accuracy(Wang et al.,2018).This significantly affects the subsequent applications of LCZ classification.

LCZ categories represent intricate semantic scenes. For instance, LCZ5 comprises open mid-rise buildings, low vegetation, scattered trees, and soil. However, WUDAPT could only extract low-level semantic information from Remote Sensing Images (RSIs) and falls short of capturing detailed semantic information within urban landscape space. Hence, it is necessary to apply methods more suitable for interpreting LCZs. Deep Learning(DL) models possess powerful feature learning capabilities and adaptability to complex scenes, allowing them to learn advanced land cover features and spatial information from RSIs, which are essential for producing high-precision LCZ maps. In recent years, with the continuous development of DL technology, using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for LCZ classification has become a research hotspot. Liu et al. regarded LCZ as a scene classification task and they used Sentinel-2 multispectral data to map LCZs in 15 cities in China, proposing LCZ-Net, a deep CNN consisting of residual learning and squeeze-and-excitation blocks, achieving an accuracy of 88.61%(Liu and Shi,2020).Additionally Huang et al. proposed a novel CNN (LCZ-CNN) and used Landsat data to classify LCZs in 32 cities in China, with accuracy exceeding 80% in half of these cities(Huang et al.,2021).

However, achieving an accuracy above 90% with recent LCZ classification methods based on remote sensing data and DL methods is challenging. On the one hand, this is because the resolution of the most commonly used RSIs not being high enough.Most of them are medium to low-resolution images such as Landsat and Sentinel, which are insufficient to depict the complex spatial layout of urban streets. On the other hand, current LCZ classification studies based on medium to low-resolution RSIs mainly treat LCZ mapping tasks as pixel-based classification, which directly neglects a large amount of spatial information and interactions between objects within scenes. Furthermore, in practical scene classification, pure scene classification cannot satisfy the demands of practical applications as there is often not just a pure LCZ scene. Instead, there are many mixed LCZ scenes. For example, as depicted in Figure 2, in a block belonging to a school zone, there may typically exist two or more LCZ categories, such as LCZ4 (Open high-rise), LCZ5 (Open mid-rise), and LCZF (Bare soil or sand). Considering this, unmixing scenes by quantifying the percentage of each LCZ category in each block is essential for analyzing the spatial interactions between per LCZ category.



(a)A VHR image (b)LCZ classification of (a)

Figure 2. The LCZ classification of a mixed scene in a school zone

In summary, to address the issues as mentioned above, this paper proposed a scene unmixing deep learning model, SU-Net, for LCZ mapping using Very High Resolution (VHR) images. The model thoroughly considers the characteristics of extracting LCZ categories from pure scenes to guide a more accurate scene unmixing process and is designed as a two-stream model. The dataset used in this study consists of VHR images from Google Earth, with a resolution of up to 1 meter. The study takes Wuhan as the research area and conducted experiments. The remainder of this paper is organized as follows. In Section 2, we introduce detailed information about the dataset used in the experiments. Then, in Section 3, we provide a detailed description of SU-Net, training process, and accuracy assessment metrics. In Section 4, we present and analyse the results. Finally, we draw some conclusions and offer research prospects in Section 5.

2. Study area and data

2.1 Study area

As the capital city of Hubei Province,Wuhan, located in the middle of the Jianghan Plain with a total area of approximately 8,569 square kilometers. It’s situated at the confluence of the Yangtze River and Han River, about one-fourth of its total area is covered by lakes and rivers. The climate of Wuhan falls into the subtropical monsoon climate category, characterized by four distinct seasons(Deng et al.,2022).Summers are hot and humid, winters are cold and dry, while spring and autumn are mild and pleasant, making it conducive for both agricultural production and urban development. Overall, Wuhan boasts a flat terrain and open landscape,conducive to urban construction and agricultural development. Over the past two decades, Wuhan has undergone significant economic and cultural growth, experiencing unprecedented urbanization characterized by densification and outward expansion(Jiao et al.,2021). Within this city, bustling urban centers with dense populations and various development zones have emerged, accompanied by the continuous expansion of urban construction area from approximately 6,000 square kilometres to 15,000 square kilometres. The permanent resident population has also surged from 8 million to 12.28 million. Therefore, Wuhan serves as an ideal research subject for investigating the relationship between urban landscape changes and climate variations, including phenomena such as the urban heat island effect.

2.2 Data

The VHR images used in this study were collected from the Google Earth platform (https://earth.google.com/), which has the advantage of high resolution and broad coverage. With a resolution of up to 1 meter, this platform provides sufficient detail for DL models to accurately capture surface characteristics and features, helping to improve the accuracy of classification and identification. Before conducting the experiments, we downloaded most of the images in Wuhan, a total of 507 original Google images with a size of about 3500*3500.To allow the model to learn the correct features during the training process, we manually annotated the VHR images and made secondary corrections before putting them into use. Considering the varying sizes of LCZ ranging from hundreds to thousands of square meters, Considering that the size of the LCZ ranges from hundreds to thousands of square meters, we divided the original image into 256*256 patches with a total size of 84975 after a series of coordinate transformation, cropping and other data preprocessing operations.Among these patches, the ratio of mixed scenes to pure scenes was approximately 2:1 and we randomly selected 15,000 mixed scene images and 7,500 pure images with a ratio of approximately 4:1:1 for the training set, validation set, and test set.

3. Methods

3.1 the architecture of the proposed SU-Net model

This paper proposes a scene unmixing deep learning model utilizing VHR imagery for LCZ mapping. The model fully considers the features extracted from pure scenes and the pure scene classification also guides a more accurate mix scene unmixing process. Hence, this model is designed as a two-stream model consisting of the Pure Scene Classification Network (PSC-Net) and the Mix Scene Unmixng Network (MSC-Net). PSC-Net comprises an encoder for pure scene classification, while MSC-Net consists of both an encoder and a decoder for scene unmixing and image reconstruction. Inspired by Hyperspectral Unmixing (HU) tasks, we regarded mixed scenes as mixed pixels, pure scenes as endmembers in HU tasks and the unmixing results of mixed scenes as abundances in HU tasks. Therefore, the model quantifies the percentage of each LCZ category in mixed scenes through the method of Adding Non-negative Constraint (ANC) and Adding Sum-to-one Constraint(ASC) between the encoder and decoder in MSC-Net(Hong et al.,2021). Significantly, information exchange between the two networks is maximized through parameter sharing. Specifically, PS-Net transfers information from pure LCZ scenes to MSC-Net, while MSC-Net provides detailed features of each pixel in mixed scenes to PS-Net. Thus, this network can quantify the percentage of pure scenes in each mixed scene and improve the accuracy of pure scene classification to some extent.

3.1.1 Pure Scene Classification Network (PSC-Net): The main task of this network is to classify pure images in the dataset, mainly composed of a Backbone, two constraints, and a classifier. Firstly, pure images are first segmented into fixed-size 16*16 image patches and each patch is then transformed into a vector representation through the patch embedding module. Subsequently, the Backbone extracts features from the image blocks, gradually transforming the input data into high-level feature representations. Finally, after passing through non-negative and sum-to-one constraints as in equations (1) and (2), the network ouputs the logits.

We selected Swin-V2(Liu et al.,2022) as the backbone of PSC-Net, and its network architecture is shown in Figure 3. Swin-V2 is a deep network model based on the Transformer architecture(Vaswani et al.,2017), which employs hierarchical attention mechanisms and local-window self-attention mechanisms to maintain efficiency when processing large-scale images. Compared to traditional CNNs, Swin-V2 not only captures better both global and local information in images but also enables relatively low memory consumption when handling VHR imagery. Additionally, Swin-V2 adopts a multi-level feature extraction strategy, allowing it to capture multi-scale information in images better, enhance the model's generalization ability, and apply it more efficiently to downstream tasks.

ASC and ANC are two commonly used constraint conditions in HU tasks.ANC requires that the abundance of each each endmember in the mixed pixel must be non-negative values, while ASC requires that the sum of abundances of all endmembers in the mixed pixel must be equal to one. Employing these constraints ensures the accuracy and the physical interpretability of the unmixing process and results(Keshava and Mustard,2002). Therefore, to ensure the accuracy of this study in the task of scene unmixing, we enforce the ASC and ANC in the last two blocks employing ReLU layer and Softmax layer, respectively, which are

$$ReLU(x) = \max(0, x) \quad (1)$$

$$Softmax(x_i) = \exp(x_i)\,/sum(exp(x_i) \quad (2)$$

Where x represents the output values of the encoder (Swin-V2), i represents the i-th element of the input vector.

3.1.2 Mix Scene Unmixing Network(MSU-Network): Inspired by the successful hyperspectral unmixing framework based on Auto Encoder(AE)(Palsson et al.,2018), we designed a similar MSC-Net. The overall structure is an encoder-decoder architecture, with two functionally similar basic modules to AE: quantifying the percentage of LCZ categories in mixed scene images (i.e., unmixing) and image reconstruction. Unlike AE, in our MSC-Net, the backbone is Swin-V2 introduced in Section 3.1.1, while the decoder adopts the same architecture as the decoder in Upernet,consisting primarily Pyramid Pooling Module (PPM) and Feature Pyramid Network(FPN) (Xiao et al.,2018). The general function of the decoder is mapping features back to the original image size. Compared to traditional decoder networks, the Upernet decoder has the following characteristics: first, it can dynamically adjust based on the input image size, making it directly applicable to image classification tasks of different sizes; Secondly, the decoder in UperNet introduces skip connections, which connect shallow features from the encoder with corresponding features in the decoder to better fuse features at different levels and capture details and edge information in VHR imagery more effectively.

The PPM module(Zhao et al.,2017) and FPN module(Lin et al.,2017) play vital roles in extracting and fusing multi-scale features, respectively.The PPM module employs a pyramid pooling approach, dividing the input feature map into grids of varying sizes, conducting pooling operations on each grid, and then concatenating all pooling results to generate a multi-scale feature representation. This enables the model to capture information comprehensively across different scales in the image, thereby reducing redundant computations effectively. On the other hand, the FPN module, a classic pyramid-style feature fusion network, aims to integrate and leverage features at diverse levels. Within the FPN module, the bottom-up feature extractor extracts features at various levels from the input image, while the top-down feature fusion incorporates high-level semantic information into low-level features through skip connections, forming a multi-scale feature pyramid.

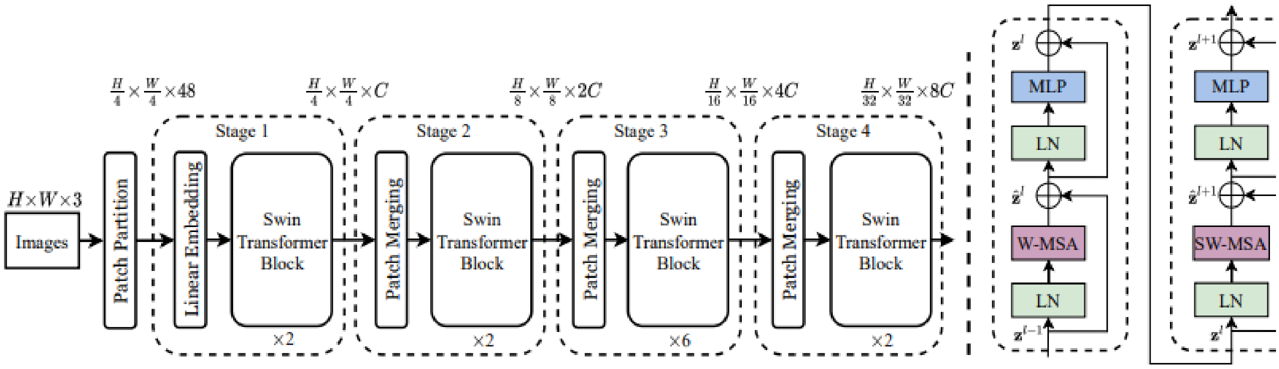


Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations, respectively(Liu et al.,2021).

3.2 The SU-Net training

The SU-Net needs to simultaneously accomplish three related but different tasks, and during the training process, it needs to optimize three loss functions. Among them, loss one and loss two are intended to optimize the results of the classification task, so they are designed as Cross-entropy loss functions which is mostly adopted in image classification tasks, while loss three aims to optimize the image reconstruction task belonging to a regression task, hence it is designed as L1 loss. To automatically balance the weights of each loss in the total loss during training, we adopt a multi-task loss function based on Gradnorm(Chen et al.,2018). The Gradnorm-based approach ensures relatively consistent gradient magnitudes for each task, reducing training instability between different tasks, and enhancing the convergence speed and performance of the model in multi-task learning, making it well-suited for the SU-Net model.

During training, to prevent overfitting, we employ the Dropout technique. Dropout randomly drops the output of some neurons in the network based on a predetermined dropout rate during each training iteration, forcing the network to learn different sub-network structures in each iteration, thereby reducing the network's reliance on specific neurons and enhancing its generalization ability(Srivastava et al.,2014).

In the training process, we choose Adam optimizer algorithm with Weight Decay (AdamW)(Loshchilov and Hutter,2017) as the optimizer algorithm for the SU-Net network, with an initial learning rate set to 1e-4 and a batch size set to 32. We chose AdamW because it is an improved algorithm based on Adam, AdamW inherites the advantages of rapid convergence and applicability of the Adam optimizer. Meanwhile, it improved the model's generalization ability and training stability by introducing a mechanism for weight decay.

3.3 Accuracy assessment

We employ Overall Accuracies (OAs),Precision and Recall calculated from the confusion matrix to measure the accuracy of pure image classification while For scene unmixing evaluation, Mean Absolute Error (MAE), Overall Pseudo Accuracy (OPA), Pseudo Precision (P-Precision), and Pseudo Recall (P-Recall) are utilized.MAE measures the average absolute error between the predicted proportions of each LCZ class in mixed scenes and the reference values and OPA,P-Precision and P-Recall are proposed to assess the accuracy between the unmixing results obtained from unmixing pure images(referred to as Pseudo classification) and the true pure classification. The calculation method parallels that of OA, Precision, and Recall for true classification, with the exception that the predicted values of the true classification are replaced with those of the pseudo-classification. The specific formulas for computing OA, Precision, and Recall for pure image classification, as well as the MAE metric for mixed image unmixing, are detailed in Table 1.

Metrics	Formulas	#
Overall Accuracies(OAs)	$\frac{\sum_{c=1}^{C}(TP_c+TN_c)}{\sum_{c=1}^{C}(TP_c+TN_c+FP_c+FN_c)}$	(3)
Precision	$\frac{\sum_{c=1}^{C}TP_c}{\sum_{c=1}^{C}(TP_c+FP_c)}$	(4)
Recall	$\frac{\sum_{c=1}^{C}TP_c}{\sum_{c=1}^{C}(TP_c+FN_c)}$	(5)
Mean Absolute Error(MAE)	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_{true}^{(i)}-y_{pred}^{(i)}\rvert$	(6)

Table1.The formula of the evaluation metrics. Where C is the number of the LCZ types.TP indicates True Positives;TN indicates True Negatives;FP indicates False Positive;FN indicates False Negatives. Where N is the number of reference labels or true labels,$y_{true}^{(i)}$ represents the true value of the i-th sample and y_{pred}^{i} represents the predicted value of the i-th sample.

4. Results and discussion

4.1 Experiment A

We conducted experiment A using 2055 pure and mixed images randomly selected from Wuhan city which were input into the SU-Net for evaluation. Table 2 presents the overall accuracy assessment of the experiment results. Our network was demonstrated its satisfactory performance in both pure image classification and mixed image unmixing. Notably, compared to other methods, there was a notable improvement in OA for pure image classification and MAE has decreased to 0.0495. Furthermore, the accuracy of pseudo-classification using the unmixing network also showed enhancement, indicating the reliable adaptation of MSU-Net to LCZ categories and the relatively precise unmixing outcomes.

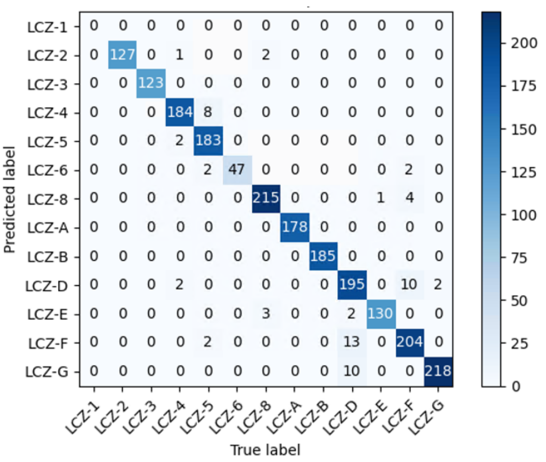


Figure 4. Confusion matrix of the pure scene classification for the Experiment A. The background color represents the number of predicted labels divided by the number of reference labels of this class (%).

The confusion matrix for the pure image classification results is shown in Figure 4. The classification achieved OAs of 96.78% and the Recall of 96.46%. It should be noted that since the manually labeled dataset for Wuhan does not include the pure LCZ-1 class, both the row and column corresponding to LCZ-1 in the confusion matrix have values of 0. For other classes, whether natural or built, the classification accuracy is above 0.92.

The accuracy metric MAE for unmixing is 0.049, indicating a small distance between the predicted and reference values. To compute Overall Pseudo Accuracy (OPA), the pure images are considered as mixed images and input into the MSU-Net within the SU-Net. The resulting pseudo-predictions for pure images are obtained, and the confusion matrix is shown in Figure 5. The classification accuracies of evaluation metrics OPA, P-Precision, and P-Recall have all shown improvement compared to the true pure image classification. This further validates that our unmixing network could be adopted to quantify the characteristics of various categories in pure images to achieve the unmixing effect.

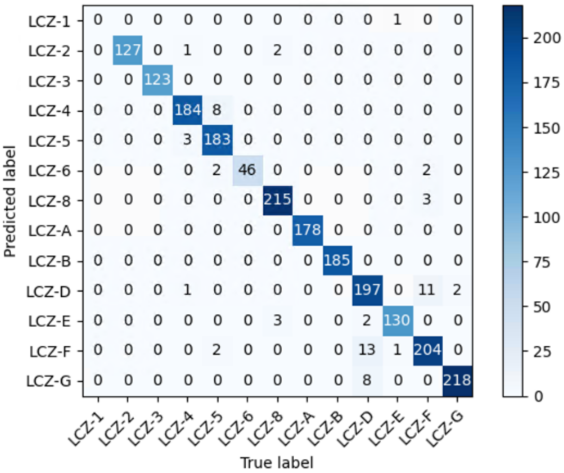


Figure 5.Confusion matrix of the pseudo pure scene classification to show the performance of the MSU-Net. The background color represents the number of predicted labels divided by the number of reference labels of this class (%)

4.2 Experiment B

To evaluate the stability and generalization ability of the SU-Net network, we conducted experiment B using all pure and mixed images from five districts in Wuhan. The confusion matrix for the classification of pure images and the confusion matrix for pseudo-classification of pure images using MSU-Net are shown in Figures 6 and 7, respectively.

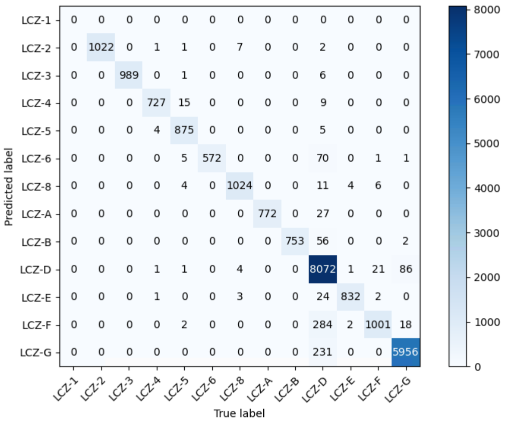


Figure 6. Confusion matrix of the pure scene classification for Experiment B. The background color represents number of predicted labels divided by the number of reference labels of this class (%).

The accuracies metrics of pure image classification in Wuhan are as follows included OA, Precision and Recall :96.09%, 96.41% and 96.09%. The Mean Absolute Error (MAE) for unmixing is calculated to be 0.0323, while OPA, P Precision and P Recall are determined to be 96.13%, 96.46%, and 96.13%, respectively. These results closely resemble the sampling test results from experiment A, albeit with a slightly higher MAE. This suggests that the SU-Net exhibits high stability and generalization ability, rendering it suitable for LCZ mapping in other cities.

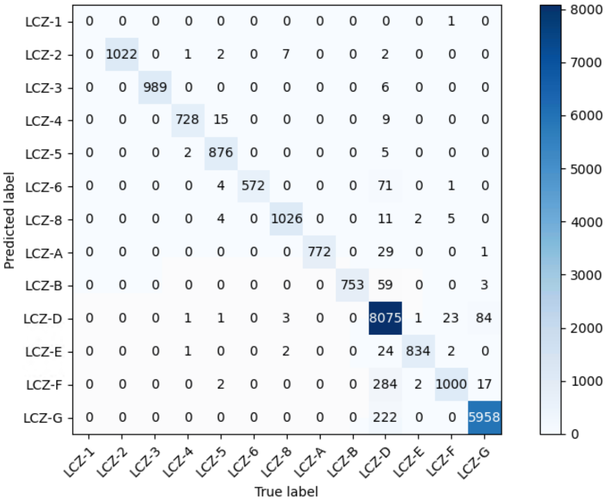


Figure 7.Confusion matrix of the pseudo pure scene classification for showing the performance of the MSU-Net. The background color represents the number of predicted labels divided by the number of reference labels of this class (%).

	Experiment1		Experiment2	
Metrics	Pure scene	Mix scene	Pure scene	Mix scene
OA	96.78%	*	96.09%	*
Precision	96.79%	*	96.41%	*
Recall	96.78%	*	96.09%	*
OPA	*	96.84%	*	96.13%
P-precision	*	96.88%	*	96.46%
P-Recall	*	96.84%	*	96.13%
MAE	*	0.0495	*	0.0323

Table 2. The results of Experiment(a) and Experiment(b)

4.3 Discussions

In most current studies that regard LCZ mapping as pure image classification, it is commonly observed that the accuracy of these methods applied to LCZ 1-6 is lower than when applied to natural classes LCZ A-G. This indicates it is challenging for these methods to learn useful spatial information, such as the height of buildings. In contrast, our network performs exceptionally well in identifying LCZ 1-5, especially with the classification OA of LCZ-3 has reached 99.99%. However, the accuracy on LCZ-6 is 0.92 which is slightly lower than other categories and is easily misclassified as LCZ-F (Bare soil or sand). Additionally, the classification accuracy of LCZ-D (Low plants) and LCZ-F (Bare soil or sand) is 0.93, as these two classes share similarities in land cover, resulting in our model having a probability of misidentifying them. We also conducted Pseudo classification experiments. Using the MSU-Net to classify pure images demonstrates our model could be adopted to unmix the mixed scenes. Furthermore, large-scale pure scene classification and unmixing experiments under large datasets demonstrated that our model not only exhibits excellent learning capabilities but also possesses high generalization abilities.

5. Conclusion

This study was inspired by the task of hyperspectral unmixing and proposed a scene unmixing deep learning model for LCZ mapping using VHR imagery. The model consists of two parallel deep networks: PSC-Net for pure image classification and MSC-Net for image unmixing. PSC-Net employs Swin-V2 as the backbone for feature extraction and outputs classification results with non-negativity and sum-to-one constraints. On the other hand, MSC-Net adopts a similar structure to autoencoders, utilizing Swin-V2 as the backbone and Upernet's decoder network structure for image reconstruction and feature fusion. Two experiments have validated the excellent performance and high generalization ability of the proposed model, providing a more convenient, efficient, and accurate method for LCZ mapping and quantitative measurement for urban heat island studies.

Currently, only Wuhan city has been tested, yet different cities exhibit variations in LCZ class distribution. Therefore, in future research, we plan to construct larger-scale datasets tailored to the characteristics and data distributions of other cities, further optimizing the model to enhance its generalization ability. Additionally, efforts will be made to establish a robust connection between LCZ mapping and real-time urban landscape analysis, integrating the model into real-time detection systems for timely LCZ map updates by continuously receiving remote sensing data.

Through these further efforts, we aim to gain a more comprehensive and accurate understanding of urban heat island effects, providing precise data support for urban planning and climate adaptation to promote sustainable urban development and improve human living environments.

Acknowledgements

The authors are grateful to the editors and anonymous reviewers for their constructive comments. The Project Supported by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation，Ministry of Natural Resources and Special Fund of the Hubei Luojia Laboratory under Grant 220100031.

References

Smith, J., 2000. Remote sensing to predict volcano outbursts. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XXVII-B1, 456-469.

Bechtel, B., Demuzere, M., Mills, G., Zhan, W., Sismanidis, P., Small, C., Voogt, J., 2019: SUHI analysis using Local Climate Zones—A comparison of 50 cities, *Urban Climate*, 28: 100451.

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

Chen, Z., Badrinarayanan, V., Lee, C.-Y., Rabinovich, A. 2018. "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks." In *International conference on machine learning*, 794-803. PMLR.

Ching, J., Mills, G., Bechtel, B., See, L., Feddema, J., Wang, X., Theeuwes, N., 2018: 0: World Urban Database and Access Portal Tools (WUDAPT), an urban weather, climate and environmental modeling infrastructure for the Anthropocene, *Bulletin of the American Meteorological Society*.

Deng, X., Cao, Q., Wang, L., Wang, W., Wang, S., Wang, L., 2022: Understanding the impact of urban expansion and lake shrinkage on summer climate and human thermal comfort in a land ‐ water mosaic area, *Journal of Geophysical Research: Atmospheres*, 127: e2021JD036131.
Fisher, J.I., Mustard, J.F., Vadeboncoeur, M.A., 2006: Green leaf phenology at Landsat resolution: Scaling from the field to the satellite, *Remote sensing of Environment*, 100: 265-79.

Geletič, J., Lehnert, M., 2016: GIS-based delineation of local climate zones: The case of medium-sized Central European cities, *Moravian Geographical Reports*, 24: 2-12.

Hong, D., Gao, L., Yao, J., Yokoya, N., Chanussot, J., Heiden, U., Zhang, B., 2021: Endmember-guided unmixing network (EGU-Net): A general deep learning framework for self-supervised hyperspectral unmixing, *IEEE Transactions on Neural Networks and Learning Systems*, 33: 6518-31.

Huang, X., Liu, A., Li, J., 2021: Mapping and analyzing the local climate zones in China’s 32 major cities using Landsat imagery based on a novel convolutional neural network, *Geo-spatial Information Science*, 24: 528-57.

Jiao, L., Dong, T., Xu, G., Zhou, Z., Liu, J., Liu, Y., 2021: Geographic micro-process model: Understanding global urban expansion from a process-oriented view, *Computers, Environment and Urban Systems*, 87: 101603.

Keshava, N., Mustard, J.F., 2002: Spectral unmixing, *IEEE signal processing magazine*, 19: 44-57.

Kotharkar, R., Bagade, A., 2018: Local Climate Zone classification for Indian cities: A case study of Nagpur, *Urban Climate*, 24: 369-92.

Lelovics, E., Unger, J., Gál, T., Gál, C.V., 2014: Design of an urban monitoring network based on Local Climate Zone mapping and temperature pattern modelling, *Climate research*, 60: 51-62.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. 2017. "Feature pyramid networks for object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117-25.

Liu, L., Lin, M., Du, Z., Liu, J., Chen, G., Du, J., 2023: Developing a CNN-based, block-scale oriented Local Climate Zone mapping approach: A case study in Guangzhou, *Building and Environment*, 240: 110414.

Liu, S., Shi, Q., 2020: Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China, *ISPRS Journal of Photogrammetry and Remote Sensing*, 164: 229-42.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L. 2022. "Swin transformer v2: Scaling up capacity and resolution." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009-19.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. 2021. "Swin transformer: Hierarchical vision transformer using shifted windows." In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012-22.

Loshchilov, I., Hutter, F., 2017: Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101*.
Palsson, B., Sigurdsson, J., Sveinsson, J.R., Ulfarsson, M.O., 2018: Hyperspectral unmixing using a neural network autoencoder, *IEEE Access*, 6: 25646-56.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014: Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research*, 15: 1929-58.

Stewart, I.D., Oke, T.R., 2012: Local climate zones for urban temperature studies, *Bulletin of the American Meteorological Society*, 93: 1879-900.

Streutker, D.R., 2003: Satellite-measured growth of the urban heat island of Houston, Texas, *Remote sensing of Environment*, 85: 282-89.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017: Attention is all you need, *Advances in neural information processing systems*, 30.

Wang, C., Middel, A., Myint, S.W., Kaplan, S., Brazel, A.J., Lukasczyk, J., 2018: Assessing local climate zones in arid cities: The case of Phoenix, Arizona and Las Vegas, Nevada, *ISPRS Journal of Photogrammetry and Remote Sensing*, 141: 59-71.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J. 2018. "Unified perceptual parsing for scene understanding." In *Proceedings of the European conference on computer vision (ECCV)*, 418-34.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. 2017. "Pyramid scene parsing network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881-90.