International Journal of Geo-Information

MDPI

Article

A Lightweight Multi-Label Classification Method for Urban Green Space in High-Resolution Remote Sensing Imagery

Weihua Lin [1], Dexiong Zhang [1], Fujiang Liu [1,*], Yan Guo [2], Shuo Chen [2], Tianqi Wu [1] and Qiuyan Hou [1]

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, China; linweihua@cug.edu.cn (W.L.); awake@cug.edu.cn (D.Z.); 1202310846@cug.edu.cn (T.W.); houqiuyan@cug.edu.cn (Q.H.)

[2] School of Computer Science, China University of Geosciences, Wuhan 430078, China; guoyan49@cug.edu.cn (Y.G.); 1202321509@cug.edu.cn (S.C.)

* Correspondence: liufujiang@cug.edu.cn

Abstract: Urban green spaces are an indispensable part of the ecology of cities, serving as the city's "purifier" and playing a crucial role in promoting sustainable urban development. Therefore, the refined classification of urban green spaces is an important task in urban planning and management. Traditional methods for the refined classification of urban green spaces heavily rely on expert knowledge, often requiring substantial time and cost. Hence, our study presents a multi-label image classification model based on MobileViT. This model integrates the Triplet Attention module, along with the LSTM module, to enhance its label prediction capabilities while maintaining its lightweight characteristic for standalone operation on mobile devices. Trial outcomes in our UGS dataset in this study demonstrate that the approach we used outperforms the baseline by 1.64%, 3.25%, 3.67%, and 2.71% in mAP, F_1, precision, and recall, respectively. This indicates that the model can uncover the latent dependencies among labels to enhance the multi-label image classification device's performance. This study provides a practical solution for the intelligent and detailed classification of urban green spaces, which holds significant importance for the management and planning of urban green spaces.

Keywords: multi-label classification; remote-sensing image; urban green space; lightweight model

check for updates

Citation: Lin, W.; Zhang, D.; Liu, F.; Guo, Y.; Chen, S.; Wu, T.; Hou, Q. A Lightweight Multi-Label Classification Method for Urban Green Space in High-Resolution Remote Sensing Imagery. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 252. https://doi.org/10.3390/ijgi13070252

Academic Editors: Wolfgang Kainz, Peng Peng, Feng Lu, Shu Wang, Maryam Lotfian and Yunqiang Zhu

Received: 16 April 2024
Revised: 1 July 2024
Accepted: 11 July 2024
Published: 13 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Urban green spaces are integral to the city's ecology, serving as the city's "purifier" and making a significant contribution to sustainable urban development [1]. Urban green spaces play a pivotal role in boosting the living quality of urban residents, preserving the natural ecological balance, and feature functions such as reducing air pollution [2–4], regulating temperature, reducing the impact of the city's heat island [5–8], and also decreasing dust, as well as noise [9,10]. Moreover, urban green spaces offer an excellent leisure platform for citizens [11,12], diminishing their stress and anxiety [13] and thereby enhancing their sense of well-being [14]. Meanwhile, urban green spaces offer citizens areas for exercise and interactive public spaces [15], contributing to improved physical and mental health [15] and enhancing social cohesion [16]. Moreover, urban green spaces provide habitats for plants and animals, aiding in the conservation of biodiversity within cities [17]. Therefore, urban green spaces play a significant protective role for both humans and other living organisms [18]. Urban green spaces also improve water quality [19–21]. An analysis of the impacts of urban green spaces, both domestically and internationally, indicates the heightened importance of their internal structure and spatial distribution when the coverage rate is less than 40% to 60%. Therefore, understanding the classifications of different types of urban green spaces is crucial. Following the requirements outlined in the "Notice on the Issuance of the (2013 Engineering Construction Standards and Regulations Formulation and Revision Plan)" by China's Ministry of Housing and Urban-Rural Development, the

"Urban Green Space Classification Standards" document was issued. This document distinctly classifies urban green spaces into five types: ancillary green space, park green space, area green space, protective green space, and square green space. This initiative enables a precise understanding of the layout of different kinds of green spaces within cities, standardizes the protection, development, and management of these spaces, and aids in enhancing the natural landscapes of both rural and urban regions, fostering sustainable development across both areas.

In the context of national urban greening projects, the detailed classification of urban green spaces represents a significant task. There are two primary methods for the detailed classification of urban green spaces: one relies entirely on visual interpretation, while the other utilizes deep learning to extract urban green spaces from remote sensing images, integrating POI (Point of Interest) and OSM (OpenStreetMap) data for detailed classification [22]. Although many researchers have made significant advancements in extracting urban green spaces using deep learning, manual intervention is still required for detailed classification. Therefore, we aim to explore the feasibility of automating the detailed classification of urban green spaces.

At the present time, deep learning is one of the most widely utilized artificial intelligence techniques in the domain of artificial intelligence, with extensive applications in image recognition, object detection, image classification, semantic segmentation, and natural language processing, among others. Multi-label image classification [23–26] represents a significant direction within deep learning, where a single image can be classified into multiple labels or categories. Unlike traditional single-label classification, multi-label classification enables an image to be allocated multiple labels, each representing different objects, scenes, or concepts present in the image. In multi-label image classification tasks, models are required to identify all possible labels present in an input image, rather than recognizing it as a single most relevant category. This approach is better suited for handling the complex and diverse content found in real-world images. Presently, multi-label image classification has been widely applied in various fields, such as protein subcellular localization [27,28], Alzheimer's disease automatic diagnosis [29], image processing of remote sensing [30], lung disease identification research [31], and all-sky aurora image processing [32]. These studies share the commonality of each image containing multiple targets to be identified, with specific connections among these targets. Given that urban green space data encompass various types of green spaces to be recognized, multi-label image classification can be effectively utilized in the research on urban green space classification. Therefore, the study introduces a deep learning framework for the refinement classification of urban green spaces, based on multi-label image classification.

To achieve an intelligent and detailed classification of urban green spaces, our study proposed a novel multi-label image classification model. This model is based on MobileViT [33] and integrates the advantages of the LSTM module [34] and the Triplet Attention module [35], enabling the high-precision detailed classification of urban green spaces while maintaining a lightweight structure. Based on the results of the experiments, the contributions made by this work can be concluded up as follows:

1. A newly formed multi-label classification model of urban green space, which incorporates the Triplet Attention module, is presented. This integration not only minimizes computational demands but also addresses the indirect correspondence between channels and weights. Furthermore, by employing an LSTM network, the model effectively minimizes interference from irrelevant information and amply utilizes effective information, capturing subtle objectives that may otherwise be overlooked. This allows for a more accurate exploration of the correlations between labels.
2. Experiments and evaluations conducted on our constructed UGS multi-label dataset prove that the presented model performs better than the existing multi-label classification methods among precision, recall, and mAP.
3. Through this study, more detailed attributes of urban green spaces can be extracted from images in an intelligent manner. This has significant implications for the plan-

ning and management of urban green spaces. The research findings can provide comprehensive decision support and multi-dimensional analysis for urban management and development, aiding in the formulation of more scientific management strategies. Consequently, the study contributes significantly to environmental protection, ecological research, and social development.

The remaining content of the paper has been arranged in the following order: Section 2 presents the sources of data, data pre-processing, and the distribution of the dataset and its true labels. Section 3 delineates the urban green space classification methodology details proposed in this document. Section 4 is dedicated to describing the evaluation criteria, associated experiments, and their outcomes. Lastly, Section 5 provides conclusions drawn from the comprehensive study conducted.

2. Related Work

2.1. Pre-Processing and Dataset

The dataset utilized in our research originates from the GF-2 satellite, featuring 0.8 m spatial resolution high-resolution remote sensing imagery. Each set of data is used by the relevant cities to declare National Garden City; therefore, the quality and precision of the data, as well as the veracity of ground truth and the comprehensiveness of land cover, are ensured, which guarantees the rigor and validity of the dataset employed in this research. The dates of the remote sensing images for different regions are as follows: RuYuan on 9 April 2020 and 27 October 2020; TongCheng on 7 April 2022; JianLi on 11 April 2022 and 16 May 2022; and HanChuan on 8 August 2022 and 12 October 2022. Upon acquisition of the raw data, preliminary processing is a requisite, which includes the orthorectification of the imagery to eliminate deformations caused by terrain or camera orientation. Subsequently, multi-spectral and panchromatic satellite image data are merged to enhance the interpretability of ground features, followed by color balancing and mosaicking to ensure the clarity and easy recognition of terrain. Afterward, all remote sensing images are cropped into samples measuring 256 × 256. Finally, data annotation is carried out, making preparations for subsequent experiments.

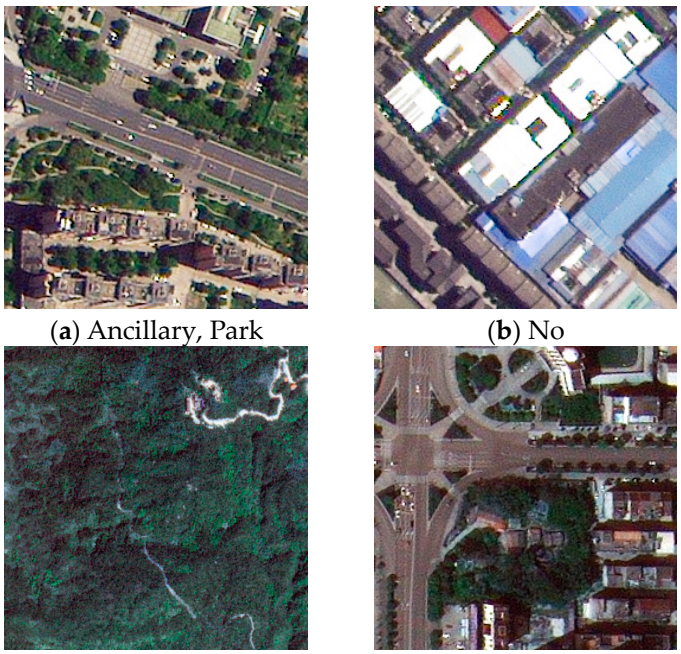
The urban green space classification dataset after data augmentation used in the study are depicted in Tables 1 and 2. Twenty percent of the data from each city, amounting to 1607 images, were randomly selected to be the test set, while the remaining 6427 images were allocated for model training. A selection of four images were chosen for illustration, as shown in Figure 1.

Table 1. Quantity of green spaces present in a picture and the quantity of images of this type.

Quantity of Green Spaces in an Image	Quantity of This Kind of Picture
1	5405
2	2143
3	460
4	26
Total	8034

Table 2. Names of green spaces and the quantity of labels for each green space.

Green Space	Quantities
Ancillary	6349
Park	1025
Square	918
Protective	824
Area	1426
No	633
Total	11,175



(**a**) Ancillary, Park (**b**) No

(**c**) Area (**d**) Ancillary, Square, Protective

Figure 1. Some sample images from the dataset.

2.2. Confusion Matrix

The confusion matrix is an essential device in the field of machine learning for evaluating the performance of classification models, illustrating the correspondence between predicted outcomes and true labels within the task. In the context of multi-label classification tasks for urban green spaces, each sample may possess labels for multiple green space categories. Therefore, the confusion matrix could provide significant information for the study. The heat map of the confusion matrix, as depicted in Figure 2, clearly shows that the "non-green space" label does not overlap with other labels, whereas overlaps are possible among the other five categories of green spaces. For instance, area green spaces and protective green spaces, together with ancillary green spaces, might frequently overlap. Consequently, this study aims to explore the potential correlations between image labels by integrating the LSTM module, with the goal of enhancing the model's classification accuracy.

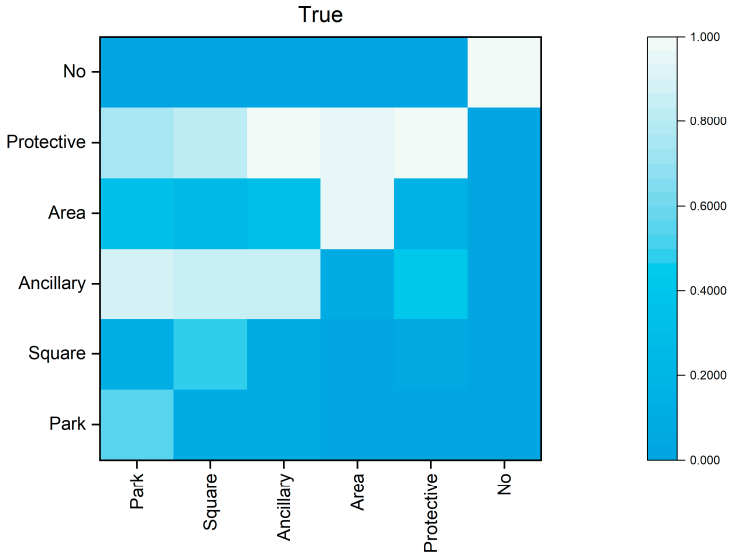


Figure 2. Heat map of the confusion matrix between different labels.

3. Methodology

Multi-label image classification can be characterized as the scenario in which a single image sample may contain multiple labels concurrently. The given sample space is X = R, where $x_i \in X$ denotes the i_{th} sample within this space, and the label space is $Y = \{y_1, y_2, \ldots, y_q\}$, with $y_i \in Y$ indicating the i_{th} category label among the collection of labels and q representing the total count of the labels in the label space. For the i_{th} image x_i in the given sample space, its corresponding labels are represented by $y_i = \{y_{i1}, y_{i2}, \ldots, y_{iq}\}$, where $y_{ij} = 1$ signifies the inclusion of label j in the image; otherwise, $y_{ij} = 0$. An end-to-end model is constructed to learn the mapping function $f : X \rightarrow Y$, facilitating the transition from images to labels. During testing, with an image provided, its multiple associated labels are predicted through the mapping function f, showcasing the model's ability to interpret the images' intricate label associations.

The article's presented model consists of three parts: MobileViT, the Triplet Attention module, and the LSTM module. The framework of the model is illustrated in Figure 3. By incorporating the Triplet Attention module, the model aims to reduce computational overhead and synthesize information across different dimensions, thereby capturing the intrinsic characteristics of the data more effectively. To minimize the interference of irrelevant information and leverage pertinent details while also detecting subtle targets that may be easily overlooked, the LSTM module is integrated. The model treats the multiple labels of an image as a sequence, initially extracting image features through MobileViT. Subsequently, it classifies features of different targets using the Triplet Attention module. Finally, the LSTM module decodes these feature maps across channels, enabling label prediction. The proposed framework performs more accurately than other models, according to experiment results.

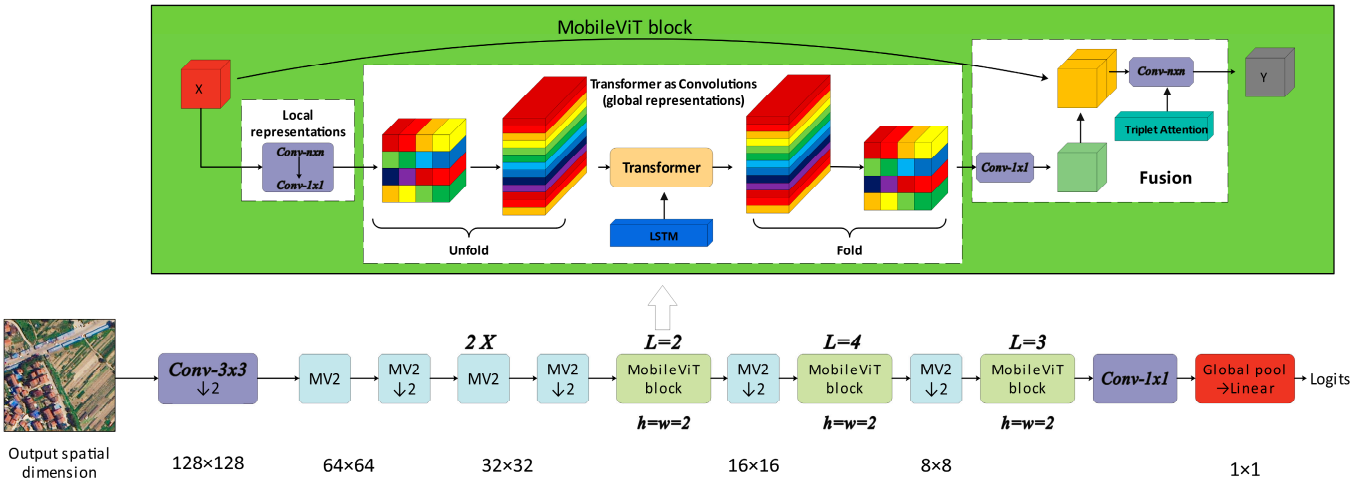


Figure 3. The structure of our model.

3.1. Feature Extraction

With computer vision developing rapidly in the past few years, numerous feature extraction networks have emerged, including ResNet [36], MobileNet-V2 [37], Vision Transformer [38], Swin_Transformer [39], and MobileViT, among others. MobileViT combines the advantages of CNNs—like spatial inductive bias and lower sensitivity to data augmentation—with the strengths of transformers, such as input-adaptive weighting and global processing. Compared to existing lightweight CNNs, MobileViT offers superior performance, generalization ability, and robustness. Hence, this paper employs MobileViT as the feature extraction network. The structure of MobileViT block is shown in Figure 4.

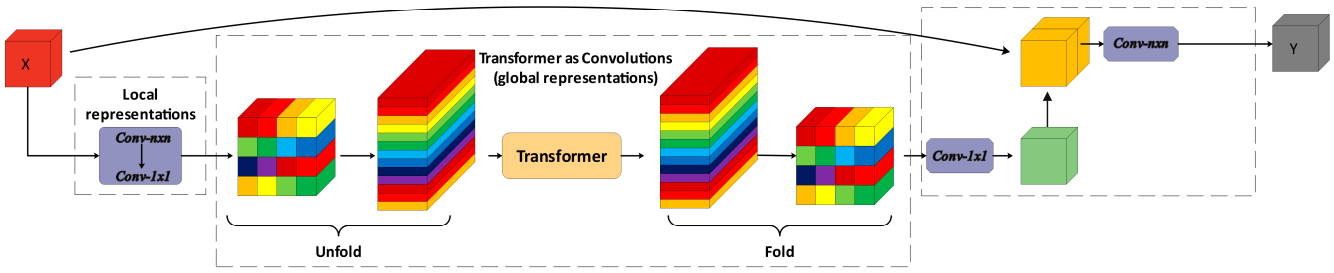


Figure 4. MoblieViT block.

For the input tensor $X \in R^{H \times W \times C}$, it is initially processed through the convolution operations of (n × n) and (1 × 1) to obtain $X_L \in R^{H \times W \times d}$. Here, the (n × n) convolution is employed to capture local information, while the (1 × 1) convolution projects the input features into a higher-dimensional space. To enable MobileViT to learn global representations with a spatial inductive bias, the "Unfold, Transformer, Fold" process for global feature modeling is used. This involves unfolding X_L into N non-overlapping segments $X_U \in R^{P \times N \times d}$, followed by modeling with a transformer to produce $X_G \in R^{P \times N \times d}$, as illustrated by the following equation:

$$X_G(p) = Transformer(X_U(p)), 1 \leq p \leq P \quad (1)$$

Subsequently, $X_G \in R^{P \times N \times d}$ is folded to derive $X_F \in R^{H \times W \times d}$, which is then subjected to a (1 × 1) convolution, resulting in C-dimensional features. This is followed by the fusion of local and global features through an (n × n) convolution, culminating in the output.

3.2. Attention Module

A tri-branch structure is used in Triplet Attention, a lightweight and effective attention mechanism, to capture cross-dimensional interactions and compute attention weights. Moreover, through rotation operations and residual transformations, it establishes inter-dimensional dependencies for the input tensor, encoding channel-wise and spatial information with very little computing overhead. Figure 5 depicts the architecture of its network.

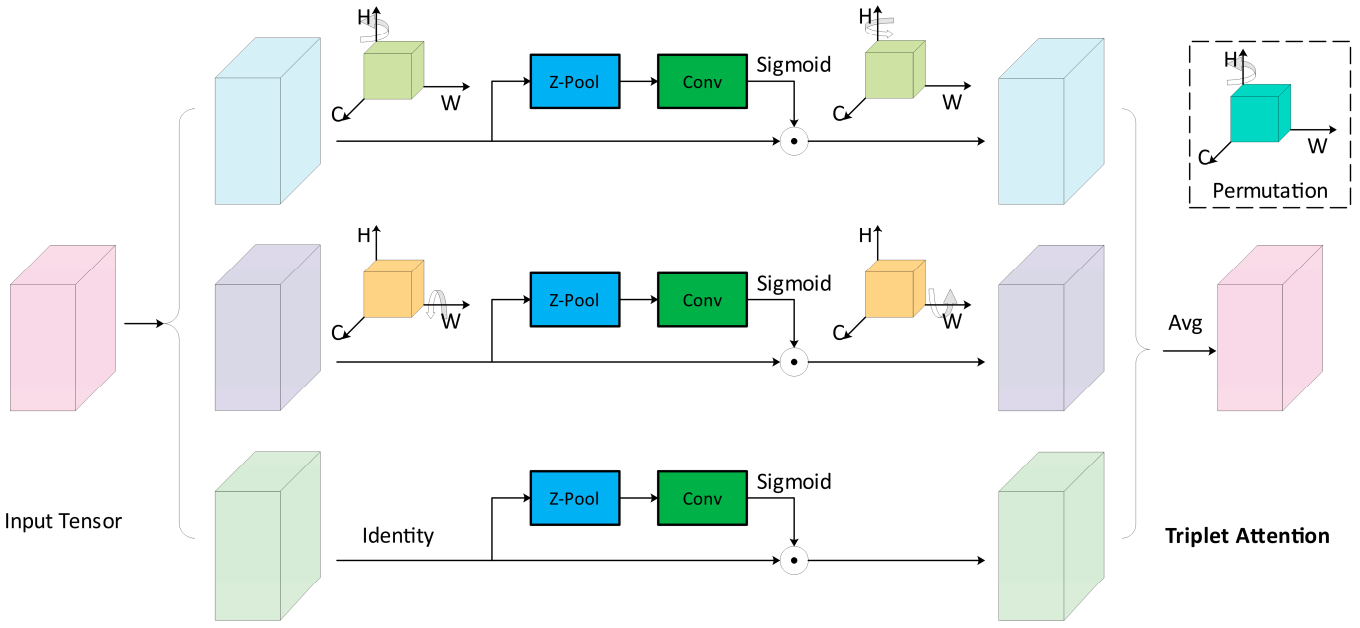


Figure 5. The illustration of the Triplet Attention module.

The Triplet Attention module is composed of three parallel branches, and each are designed to capture the dependencies among (C, H), (C, W), and (H, W), besides proposing

cross-dimensional interactions. The structure effectively addresses the challenge of calculations being independent of each other and the isolation of channel attentions and spatial attentions in CBAM, which are computed independently. When given an input tensor $X \in R^{C\times H\times W}$, the three branches of the Triplet Attention mechanism receive it immediately. Within these branches, interactions between the dimensions C, H, and W are established in pairs. Subsequently, the process incorporates batch normalization. The refined tensors $C \times H \times W$ generated from the branches are then aggregated, resulting in the output y, as depicted in the equation illustrated. The structure effectively integrates information across various dimensions, thereby enhancing the ability to capture the intrinsic characteristics of the data more effectively:

$$y = \frac{1}{3}\left(\hat{\chi_1}\sigma(\psi_1(\hat{\chi_1}^*)) + \hat{\chi_2}\sigma(\psi_2(\hat{\chi_2}^*)) + \chi\sigma(\psi_3(\hat{\chi_3}))\right) \quad (2)$$

Beyond the Triplet Attention mechanism, prevalent attention mechanisms include SE [40], CBAM [41], and SA [42], among others. Triplet Attention distinguishes itself by its reduced computational demands. The emphasis on cross-dimension interaction without dimensionality reduction eliminates the indirect correspondence between channels and weights. This approach not only enhances computational efficiency but also ensures a deeper integration of spatial and channel-wise information, thereby addressing the issue of significant spatial information loss.

3.3. LSTM for Latent Semantic Dependencies

LSTM (long short-term memory) networks represent a specialized category of recurrent neural networks (RNNs) that address the issue of long-term dependencies, wherein the state of the system at any given time can be influenced by states from much earlier in the sequence. In scenarios where the temporal interval becomes significantly large, traditional RNNs are susceptible to gradients either exploding or vanishing. LSTM circumvents these issues through its distinctive architecture, fundamentally leveraging an internal gating mechanism to selectively retain or disregard information, as illustrated in Figure 6. This mechanism incorporates the "forget gate layer" to modulate the degree to which long-term information from the cell state c_{t-1} is maintained, the "input gate layer" to regulate the extent of new information being incorporated, and the "output gate layer" to control the magnitude of information emitted as the output. The method of updating for LSTM at time step t can be delineated as follows:

$$m_t = \tanh(W_{xc}x_t + W_{hc}x_{t-1} + b_c) \quad (3)$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}x_{t-1} + b_i) \quad (4)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}x_{t-1} + b_f\right) \quad (5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}x_{t-1} + b_o) \quad (6)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes m_t \quad (7)$$

$$h_t = o_t \times tanh(c_t) \quad (8)$$

In these formulas, all instances of W and b denote parameters to be trained, while x_t denotes the input at time t. The symbols i_t, f_t, and o_t correspond to the input gate layer, forget gate layer, and output gate layer within the LSTM architecture, respectively. c_t and h_t indicate the stored and hidden states of LSTM, respectively. The function $\sigma(.)$ is indicative of the sigmoid activation function.

LSTM, through its cell states and gating mechanisms, regulates the flow of information better, effectively mitigating the impact of irrelevant data and addressing the issues of gradient decay and explosion. LSTM can also effectively preserve and update historical information via these mechanisms, which enables LSTM to capture correlations between

labels by utilizing the information encoded in the units at each time step. When making a prediction at moment c_t, integrating correlations from all previous channels' labels facilitates an enhanced ability to recognize the current predictive labels. This capability makes LSTM particularly adept at handling long-term dependency issues.

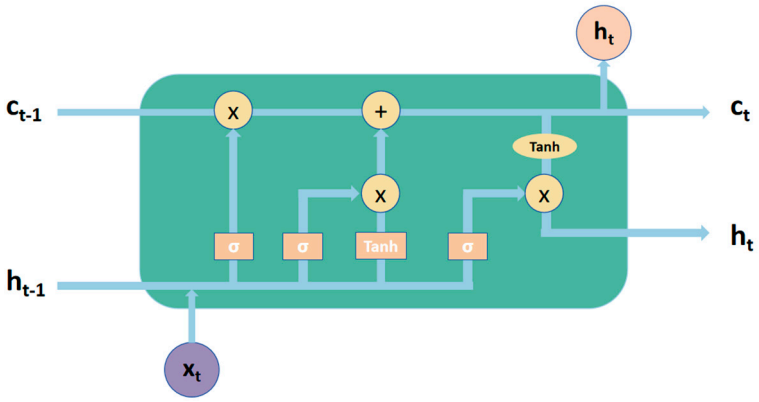


Figure 6. The illustration of the LSTM module.

In the LSTM framework, the LSTM network can be concentrated on collecting semantic relationships between labels because every channel in the feature map links to a label. Specifically, channel v_t undergoes encoding first, following which the resultant x_t is sequentially fed into the LSTM, yielding the predicted probabilities for label p_t:

$$x_t = relu(W_{vx}v_t + b_x) \tag{9}$$

$$h_t = LSTM(x_t, h_{t-1}, c_{t-1}) \tag{10}$$

$$p_t = \sigma\left(W_{hp}h_t + b_h\right) \tag{11}$$

In the above equation, the convolution parameters are W_{vx} and b_x, and the classification layer parameters are W_{hp} and b_h.

3.4. Data Augmentation

Generally, the larger the volume of sample data available for experimentation, the better the training outcomes of the model and the stronger its ability to generalize. In cases where the dataset is limited in size or the quality of samples is suboptimal, data augmentation procedures are necessitated to enhance the model's generalizability and robustness, thereby preventing overfitting. Experiments conducted without the implementation of data augmentation yield subpar results, particularly in predicting images with three or more labels and images with rare labels. Therefore, this study employs a variety of data augmentation techniques to enhance the model's performance as a whole. Unlike conventional image data, remote sensing imagery retains semantic information, even after transformations such as rotation and flipping. Consequently, this paper utilizes horizontal flipping, vertical flipping, and rotations as methods to achieve the purpose of data augmentation.

By adopting this method, samples in the dataset with fewer label categories undergo data augmentation, resulting in six training samples for each specified label category. Such transformations enhance the robustness of our model and improve its performance on rare labels.

4. Experiment

The models used in the experiments were initialized with weights trained on the ImageNet dataset and underwent training with the feature extraction layers frozen. ImageNet is one of the deep learning datasets commonly used for object detection, object

localization, and image classification, comprising over 15 million images across more than 21,000 categories. The same set of circumstances were used for every experiment in this study. The server's CPU hosting the experimental environments was an Intel(R) Xeon(R) Gold 5218 CPU (2.30 GHz), and its GPU was an Nvidia Tesla T4. The multi-label soft margin loss was employed as the loss function, and the optimizer was stochastic gradient descent (SGD), with a 0. 001 starting learning rate. To determine the best epoch for each model, an initial training period of 300 epochs was conducted.

4.1. Evaluation

In this experiment, precision, recall, F_1 and mAP were used as the evaluations. The calculation formulas are shown below:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F_1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{14}$$

where FP denotes the quantity of false positives, and TP represents the quantity of real positives. The precision is defined as its ability not to identify a negative sample as positive, and the recall is its capacity to identify every positive sample. The F_1 score can be understood as a harmonic mean of the recall and precision, with the best value of all three assessments occurring at 1 and the lowest score occurring at 0. Both recall and precision have an equal proportional contribution to the F_1 score.

mAP (mean average precision) is defined as the mean of the AP (average precision) metrics computed across various categories. The calculation of AP is based on the interpolated average precision, which is measured by the area under the $Precision - Recall$ curve. Within this formula, $n - 1$ represents the counting of detection points, while p_{interp} denotes the precision metric at a designated recall rate r. To calculate the mAP, the AP metrics are aggregatedfor every category, with K representing the entire number of classes in the analysis.

$$AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i) p_{interp}(r_{i+1}) \tag{15}$$

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{16}$$

4.2. Classification Performance

As demonstrated in Section 3 above, this study utilized a network based on Vision Transformer (ViT) or convolutional neural networks (CNNs) as our feature extraction network. To ascertain which network is better suited for image recognition of five categories of urban green space, we tested various feature extraction networks, including ResNet, MobileNet, Vision Transformer, Swin Transformer, and MobileViT. Considering the practical requirements and device limitations in classifying urban green spaces, we also used FLOPs and parameters as the evaluation metrics; thus, this research aimed for the model to achieve optimal performance metrics while minimizing FLOPs and parameters, enabling autonomous operation on mobile devices. To this end, the study utilized the THOP library to compute the FLOPs and parameters of each model, with the comprehensive results displayed in Table 3.

As indicated in Table 3, aside from ViT-B16, other models demonstrated favorable performances across various metrics. Taken together, the performance of MobileViT_XXS in the three evaluation indicators was the best. Furthermore, as shown in Table 3, MobileViT_XXS exhibited the lowest FLOPs and parameters compared to other models, making

it more suitable for practical application in this study. Consequently, this paper adopted MobileViT_XXS as the baseline for subsequent experiments.

Table 3. Different models' evaluations in the UGS dataset.

Model	mAP (%)	F1 (%)	Precision (%)	Recall (%)	FLOPs (M)	Parameters (M)
ResNet50	90.62	79.71	81.06	82.24	5398.55	25.56
MobileNet-V2	89.87	78.06	81.46	77.85	427.35	3.50
ViT-B16	54.06	28.61	22.41	45.77	21,999.71	86.42
Swin-T	91.12	80.15	86.38	78.93	7110.03	28.27
MobileViT_XXS	92.51	82.56	87.91	80.85	350.34	**1.01**

In this study, multi-label confusion matrix heatmaps were generated from predicted labels of ResNet50, MobileNet-V2, ViT-B16, Swin-T, MobileViT_XXS, the proposed model in this paper, and the true labels on the test set. The confusion matrix heatmap for our proposed model showed a closer match to the multi-label confusion matrix heatmap of the true labels, in contrast to other models, as shown in Figure 7. This indicates that the presented framework effectively predicted the interrelationships among different labels, thus substantiating the beneficial impact of the LSTM module on this research.

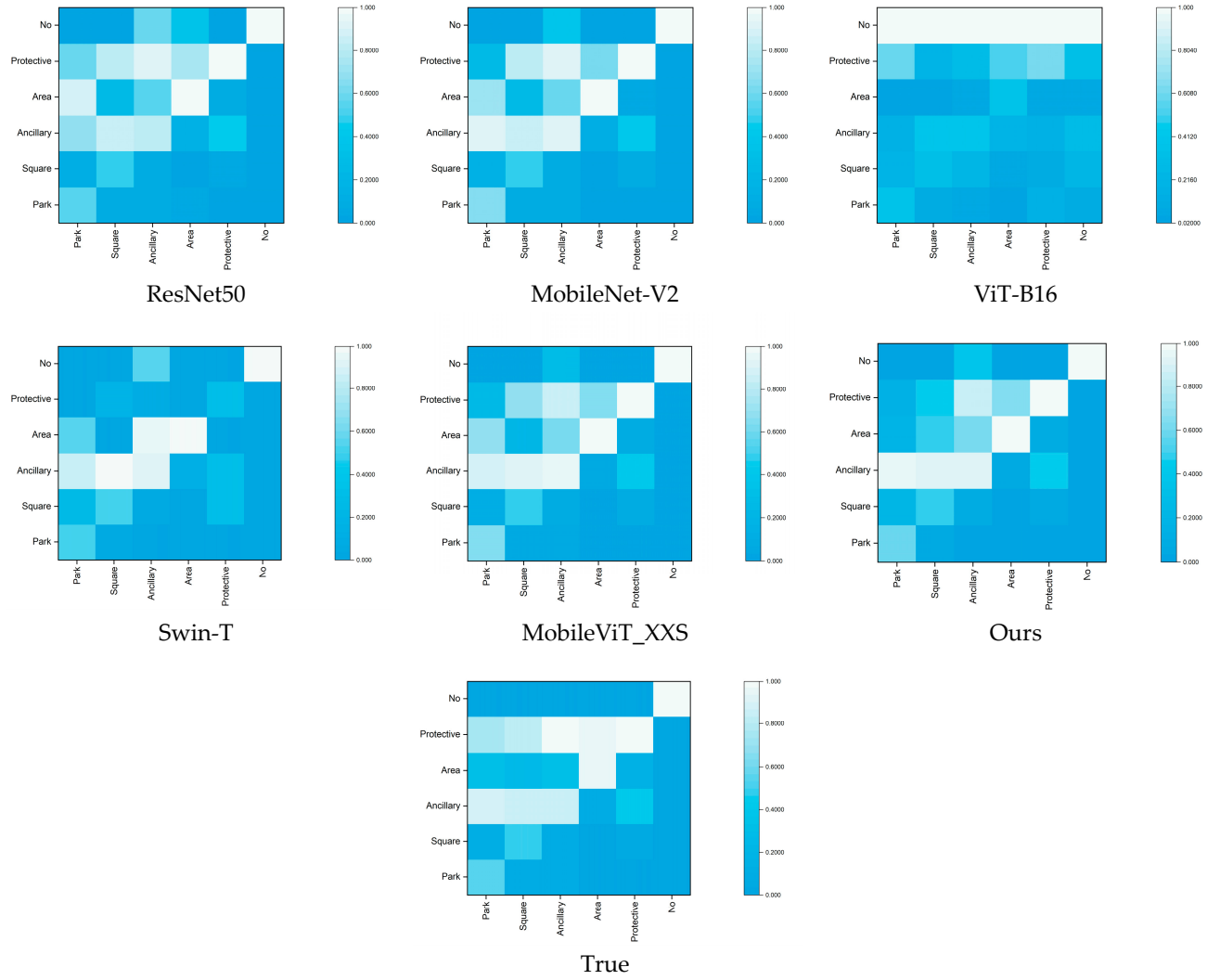


Figure 7. Heatmaps of the confusion matrix for different models in this paper.

The visualization heatmaps of the confusion matrix revealed that the introduction of the LSTM module captured semantic dependencies among labels, thereby enhancing label prediction capabilities. To prove that the incorporation of the Triplet Attention module can enhance the performance of different target classifications while maintaining the model's lightweight characteristics and to prove that the concurrent integration of both the LSTM and Triplet Attention modules further enhances model performance, ablation experiments were carried out on the model. The results indicated significant enhancements in model performance with the individual and combined introductions of the LSTM and Triplet Attention modules, compared to the original model, as depicted in Table 4. Improvements of 1.64%, 3.25%, 3.67%, and 2.71% were observed in mAP, F1, precision, and recall. The shown results show that our methodology works in considerably enhancing multi-label classification performance in urban green spaces.

Table 4. The results of ablation experiments.

Model	mAP (%)	F1 (%)	Precision (%)	Recall (%)
MobileViT	92.51	82.56	87.91	80.85
MobileViT + Triplet Attention	92.74	82.76	88.20	81.26
MobileViT + LSTM	93.77	85.32	90.45	83.49
MobileViT + LSTM + Triplet Attention	94.15	85.81	91.58	83.56

Our model was compared with other models (ResNet50, MobileNet-V2, and MobileViT_XXS) of F1, precision, and recall scores on the test dataset, as illustrated in Figures 8–10. Considering the protective green space's rarity in the dataset, as well as the high similarity to the ancillary green space, likewise, the park green space was often visually confused with other types of green space; all models demonstrated lower recall scores for these two categories, as shown in Figure 7, while our model achieved higher scores across the other four categories of green space. Additionally, as shown in Figures 8 and 9, our model also performed well with regard to precision and F1 scores. Moreover, the distribution of scores across labels in our model fit the labels' frequency distribution in the dataset. In the dataset, categories with a higher volume of data gave the model more features to learn so that it could generate better results. Therefore, the experimental results presented in this paper are reasonable and highlight the significant impacts of sample variability and label imbalance on multi-label classification.

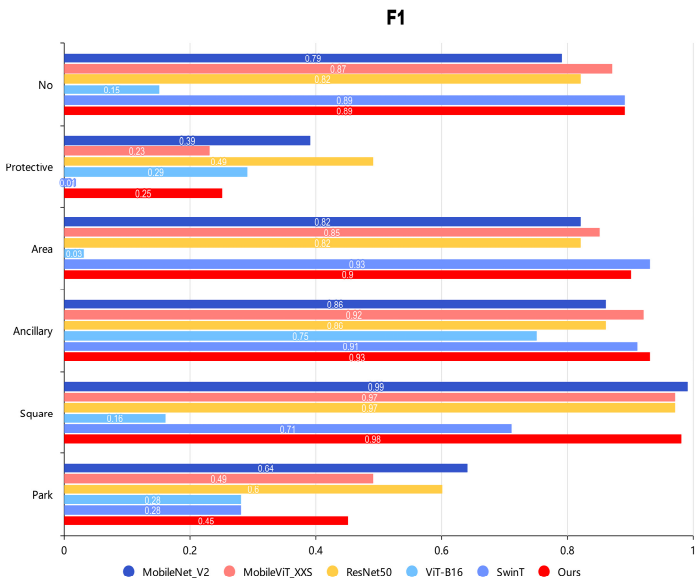


Figure 8. F1 scores for each class of model in the test set.

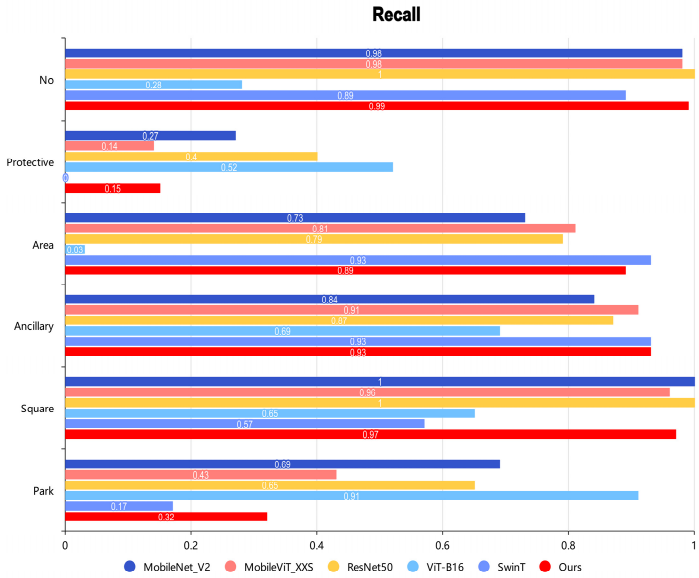


Figure 9. Recall scores for each class of model in the test set.

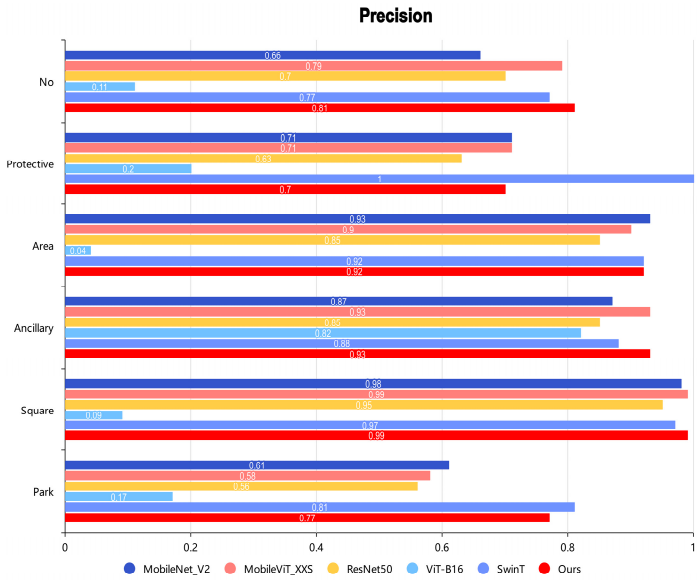


Figure 10. Precision scores for each class of model in the test set.

Figure 11 presents the training loss and mAP curves of our model over 100 training epochs. As the training epochs increased, there was a corresponding decrease in loss values and an increase in mAP values. It was evident that, compared to other models, our model exhibited less fluctuation in its loss curve and converged more rapidly, suggesting superior model fitting. The mAP curves revealed a similar trend among all models, while we obtained the highest mAP score with our model upon convergence. The comparative analysis in the two graphs demonstrated that our model outperformed others in terms of classification performance.

The mAP curves for both the training and testing datasets of our model for 100 epochs are illustrated in Figure 12. With the progression of training time, there was a steady increase in the mAP values for both datasets until convergence was reached. Notably, the mAP scores for the testing set did not significantly decline relative to those of the training set, indicating that overfitting did not occur in our model.

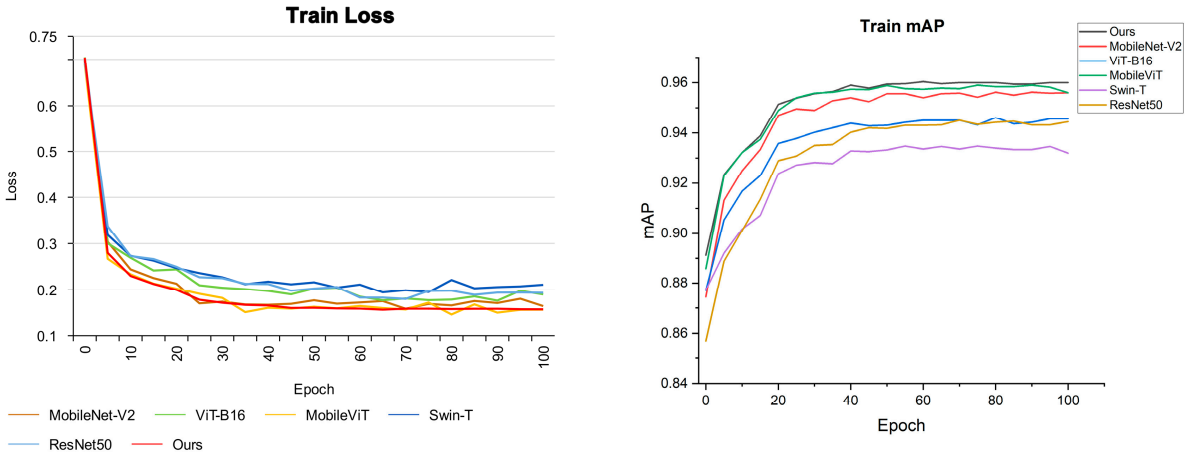


Figure 11. (**Left**): The training loss curves for each model; (**Right**): the training mAP curves for each model.

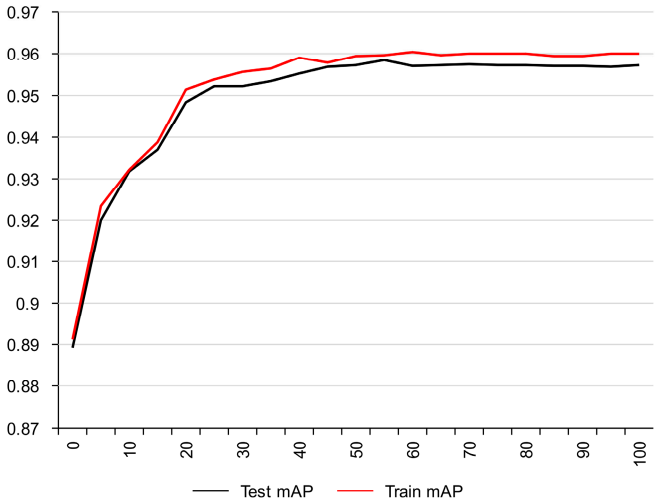


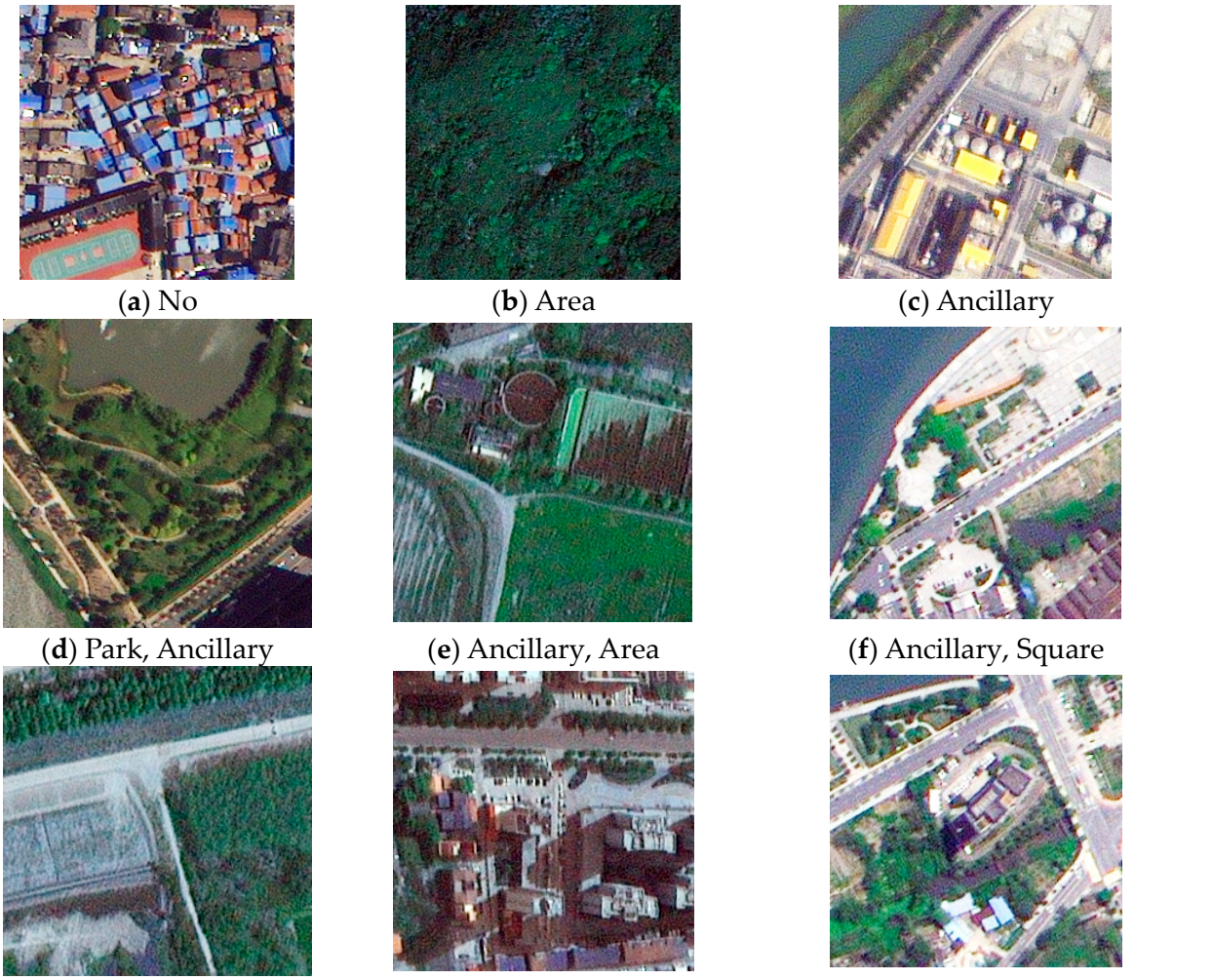
Figure 12. mAP curves on the train set and test set.

5. Discussion

Compared to existing studies, this research refined the classification of urban green spaces into five categories based on the "Urban Green Space Classification Standards". Liu et al. [43] classified urban green spaces into grasslands, forests, and agricultural land based on natural attributes, and Xu et al. [44]., utilizing HRNet with feature engineering, divided urban green spaces into deciduous trees, evergreen trees, and grasslands. Our study, however, classified urban green spaces according to their social attributes, categorizing them into five distinct types. In contrast to Huang et al.'s research, which used manual intervention for detailed classification after automated extraction, this study achieved automated and intelligent detailed classification of urban green spaces through straight deep learning. To address the need for improved classification accuracy, this study proposed a model based on MobileViT combined with the LSTM module and the Triple Attention module for the detailed classification of high-resolution remote sensing images of urban green spaces. The UGS dataset was constructed, and ablation experiments were conducted to validate the effectiveness of each module in the model. Additionally, comparative experiments with ResNet, MobileNet, Vision Transformer, Swin Transformer, and MobileViT were performed, demonstrating the robustness of our proposed model.

Beyond validating the efficacy of our proposed model, further predictive experiments were conducted to better demonstrate the model's classification performance. To verify the model's real performance, nine representative images from the UGS dataset were selected,

including three images with single-class labels, three with two-class labels, and three with three or more class labels, to showcase the model's classification capabilities. The results indicated that in Figure 13, with single- and two-class labels, the model correctly predicted all categories. However, in images with three or more class labels, some discrepancies occurred, such as failing to identify area green spaces in image (g). In the most challenging image (i) with four-class labels, the model only identified square green space and ancillary green space.



(**a**) No (**b**) Area (**c**) Ancillary

(**d**) Park, Ancillary (**e**) Ancillary, Area (**f**) Ancillary, Square

(**g**) Ancillary, Protective, Area (**h**) Park, Square, Ancillary (**i**) Park, Square, Ancillary, Protective

Figure 13. Predictions for some images on the test set. Blue labels represent unidentified.

The results of all experiments demonstrate the feasibility of using a multi-label classification model for the urban green space refined classification, providing support to urban green space research and even contributing to national urban ecological research. Additionally, there are areas for optimization in the study: given the limited number of rare labels in the dataset, the recognition of rare labels was suboptimal across these experiments. Therefore, a combination of various data augmentation techniques could be employed to enhance data robustness, and further model improvements, such as refining the loss function, could enhance performance. Secondly, the follow-up research can consider the use of the multi-source data fusion strategy, combined with social perception data and the combination of different scales of remote sensing images, in order to enhance the classification of urban green space's accuracy, as well as efficiency.

6. Conclusions

In this study, a multi-label classification method based on MobileViT is proposed to address the reliance on manual visual interpretation in the traditional detailed classification of urban green spaces. The model integrates the LSTM module and the Triple Attention module. Experimental results on the UGS dataset demonstrate the model's excellent performance in the detailed classification of urban green spaces. Additionally, the LSTM module uncovers potential dependencies between labels, while the Triple Attention module enhances classification accuracy while maintaining a lightweight model structure.

The focus of future research includes: (1) trying to use a combination of multiple data enhancement operations for data enhancement and to enhance the model's performance through further optimizing the model, such as improving the loss function and optimizing the model structure and (2) considering the use of multi-source data fusion, combining social perception data and combining different scales of remote sensing imagery of urban green spaces to carry out the study on the refinement of the classification of urban green spaces.

Author Contributions: Conceptualization, Yan Guo and Fujiang Liu; methodology, Dexiong Zhang and Weihua Lin; validation, Qiuyan Hou, Tianqi Wu, and Shuo Chen; formal analysis, Weihua Lin and Dexiong Zhang; investigation, Dexiong Zhang and Fujiang Liu; resources, Weihua Lin; data curation, Dexiong Zhang; writing—original draft preparation, Weihua Lin and Dexiong Zhang; writing—review and editing, Weihua Lin and Dexiong Zhang; visualization, Dexiong Zhang; supervision, Yan Guo and Fujiang Liu; project administration, Qiuyan Hou and Shuo Chen; funding acquisition, Weihua Lin and Fujiang Liu. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Metallogenic Patterns and Mineralization Predictions for the Daping gold deposit in Yuanyang County, Yunnan Province, grant number 2022026821; it was supported by the Open Fund of State Key Laboratory of Remote Sensing Science, grant number 6142A01210404 and the Hubei Key Laboratory of Intelligent Geo-Information Processing, grant number KLIGIP-2022-B03.

Data Availability Statement: Restrictions apply to the availability of the UGS data, and the data can be available from the authors.

Acknowledgments: We gratefully acknowledge the support of the School of Computer Science, China University of Geosciences, Wuhan, and the School of Geography and Information Engineering, China University of Geosciences, Wuhan.

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

References

1. Wu, J. Landscape sustainability science: Ecosystem services and human well-being in changing landscapes. *Landsc. Ecol.* **2013**, *28*, 999–1023. [CrossRef]
2. Lin, L.; Yan, J.; Ma, K.; Zhou, W.; Chen, G.; Tang, R.; Zhang, Y. Characterization of particulate matter deposited on urban tree foliage: A landscape analysis approach. *Atmos. Environ.* **2017**, *171*, 59–69. [CrossRef]
3. Jenerette, G.D.; Harlan, S.L.; Buyantuev, A.; Stefanov, W.L.; Declet-Barreto, J.; Ruddell, B.L.; Myint, S.W.; Kaplan, S.; Li, X. Micro-scale urban surface temperatures are related to land-cover features and residential heat related health impacts in Phoenix, AZ USA. *Landsc. Ecol.* **2016**, *31*, 745–760. [CrossRef]
4. Yan, J.; Lin, L.; Zhou, W.; Han, L.; Ma, K. Quantifying the characteristics of particulate matters captured by urban plants using an automatic approach. *J. Environ. Sci.* **2016**, *39*, 259–267. [CrossRef] [PubMed]
5. Zhou, W.; Wang, J.; Cadenasso, M.L. Effects of the spatial configuration of trees on urban heat mitigation: A comparative study. *Remote Sens. Environ.* **2017**, *195*, 1–12. [CrossRef]
6. Li, X.; Kamarianakis, Y.; Ouyang, Y.; Turner, B.L., II; Brazel, A. On the association between land system architecture and land surface temperatures: Evidence from a Desert Metropolis—Phoenix, Arizona, USA. *Landsc. Urban Plan.* **2017**, *163*, 107–120. [CrossRef]
7. Adams, M.P.; Smith, P.L. A systematic approach to model the influence of the type and density of vegetation cover on urban heat using remote sensing. *Landsc. Urban Plan.* **2014**, *132*, 47–54. [CrossRef]
8. Bowler, D.E.; Buyung-Ali, L.; Knight, T.M.; Pullin, A.S. Urban greening to cool towns and cities: A systematic review of the empirical evidence. *Landsc. Urban Plan.* **2010**, *97*, 147–155. [CrossRef]

9. Pathak, V.; Tripathi, B.; Mishra, V. Evaluation of anticipated performance index of some tree species for green belt development to mitigate traffic generated noise. *Urban For. Urban Green.* **2011**, *10*, 61–66. [CrossRef]
10. Van Renterghem, T.; Botteldooren, D. Reducing the acoustical façade load from road traffic with green roofs. *Build. Environ.* **2009**, *44*, 1081–1087. [CrossRef]
11. Xiao, R.; Zhou, Z.; Wang, P.; Ye, Z.; Guo, E.; Ji, G.C. Application of 3S technologies in urban green space ecology. *Chin. J. Ecol.* **2004**, *23*, 71–76.
12. Tu, X.; Huang, G.; Wu, J. Review of the relationship between urban greenspace accessibility and human well-being. *Acta Ecol. Sin.* **2019**, *39*, 421–431.
13. Thompson, C.W.; Roe, J.; Aspinall, P.; Mitchell, R.; Clow, A.; Miller, D. More green space is linked to less stress in deprived communities: Evidence from salivary cortisol patterns. *Landsc. Urban Plan.* **2012**, *105*, 221–229. [CrossRef]
14. Kaplan, R. The nature of the view from home: Psychological benefits. *Environ. Behav.* **2001**, *33*, 507–542. [CrossRef]
15. Maas, J.; Verheij, R.A.; Groenewegen, P.P.; De Vries, S.; Spreeuwenberg, P. Green space, urbanity, and health: How strong is the relation? *J. Epidemiol. Community Health* **2006**, *60*, 587–592. [CrossRef] [PubMed]
16. Kuo, F.E.; Sullivan, W.C. Environment and crime in the inner city: Does vegetation reduce crime? *Environ. Behav.* **2001**, *33*, 343–367. [CrossRef]
17. Savard, J.-P.L.; Clergeau, P.; Mennechez, G. Biodiversity concepts and urban ecosystems. *Landsc. Urban Plan.* **2000**, *48*, 131–142. [CrossRef]
18. Bolund, P.; Hunhammar, S. Ecosystem services in urban areas. *Ecol. Econ.* **1999**, *29*, 293–301. [CrossRef]
19. Collins, K.A.; Lawrence, T.J.; Stander, E.K.; Jontos, R.J.; Kaushal, S.S.; Newcomer, T.A.; Grimm, N.B.; Ekberg, M.L.C. Opportunities and challenges for managing nitrogen in urban stormwater: A review and synthesis. *Ecol. Eng.* **2010**, *36*, 1507–1519. [CrossRef]
20. Roy, A.H.; Wenger, S.J.; Fletcher, T.D.; Walsh, C.J.; Ladson, A.R.; Shuster, W.D.; Thurston, H.W.; Brown, R.R. Impediments and solutions to sustainable, watershed-scale urban stormwater management: Lessons from Australia and the United States. *Environ. Manag.* **2008**, *42*, 344–359. [CrossRef]
21. Barbosa, O.; Tratalos, J.A.; Armsworth, P.R.; Davies, R.G.; Fuller, R.A.; Johnson, P.; Gaston, K.J. Who benefits from access to green space? A case study from Sheffield, UK. *Landsc. Urban Plan.* **2007**, *83*, 187–195. [CrossRef]
22. Chen, W.; Huang, H.; Dong, J.; Zhang, Y.; Tian, Y.; Yang, Z. Social functional mapping of urban green space using remote sensing and social sensing data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 436–452. [CrossRef]
23. Zhang, M.-L.; Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2013**, *26*, 1819–1837. [CrossRef]
24. Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. HCP: A flexible CNN framework for multi-label image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1901–1907. [CrossRef] [PubMed]
25. Tsoumakas, G.; Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehous. Min.* **2007**, *3*, 1–13. [CrossRef]
26. Tarekegn, A.N.; Giacobini, M.; Michalak, K. A review of methods for imbalanced multi-label classification. *Pattern Recognit.* **2021**, *118*, 107965. [CrossRef]
27. Xiao, X.; Wu, Z.-C.; Chou, K.-C. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.* **2011**, *284*, 42–51. [CrossRef] [PubMed]
28. Lin, W.-Z.; Fang, J.-A.; Xiao, X.; Chou, K.-C. iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. BioSystems* **2013**, *9*, 634–644. [CrossRef]
29. Salvatore, C.; Castiglioni, I. A wrapped multi-label classifier for the automatic diagnosis and prognosis of Alzheimer's disease. *J. Neurosci. Methods* **2018**, *302*, 58–65. [CrossRef]
30. Shao, Z.; Zhou, W.; Deng, X.; Zhang, M.; Cheng, Q. Multilabel remote sensing image retrieval based on fully convolutional network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 318–328. [CrossRef]
31. Wang, J. Deep Learning-Based Lightweighting and Fusion of Label Information for Lung Disease Recognition Study. Master's Thesis, Dalian Ocean University, Dalian, China, 2024.
32. Wang, Y. Deep Learning Based Small Sample Classification and Multi-Label Classification of All-Sky Auroral Images. Master's Thesis, Shaanxi Normal University, Xi'an, China, 2022.
33. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
34. Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
35. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 3139–3148.
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
37. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

39. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
40. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
41. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
42. Zhang, Q.-L.; Yang, Y.-B. Sa-net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239.
43. Liu, W.; Yue, A.; Shi, W.; Ji, J.; Deng, R. An automatic extraction architecture of urban green space based on DeepLabv3plus semantic segmentation model. In Proceedings of the 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, 5–7 July 2019; pp. 311–315.
44. Xu, Z.; Zhou, Y.; Wang, S.; Wang, L.; Li, F.; Wang, S.; Wang, Z. A novel intelligent classification method for urban green space based on high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 3845. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.