ISSN 2616-5775 Vol. 8, Issue 9: 40-45, DOI: 10.25236/AJCIS.2025.080906

Research on Historical Building Recognition in Leshan Satellite Images Using CNN

Hongyan Fan[1,a,*]

[1]*Institute of Cultural and Tourism, Leshan Vocational and Technical College, Leshan, 614000, China*
[a]*lynnfanhy@outlook.com*
[*]*Corresponding author*

Abstract: *Historical building surveys have long faced challenges of low efficiency and high costs. This study addresses the preservation needs of historical buildings in the Leshan region by proposing a satellite image recognition framework based on an improved lightweight MobileNetV2 convolutional neural network (CNN). The model adapts to the architectural and image characteristics of the Leshan region, optimizes sample quality to enhance data usability, and incorporates an attention mechanism to improve architectural feature recognition capability. Experimental results demonstrate that the model operates with low hardware requirements, achieving 90.9% accuracy on the test set. It enables rapid identification and dynamic monitoring of historical buildings across the entire Leshan region, providing technical support for cultural heritage preservation efforts in the area.*

Keywords: *CNN, Deep Learning, Historic Buildings, Leshan Area*

1. Introduction

A Historical buildings, as key physical carriers of a city's historical evolution, serve as vital tools for preserving its distinctive character and safeguarding its collective memory[1].As a vital component of cultural heritage, historic buildings serve as tangible expressions of regional architectural styles, embodying residents' lived memories and the trajectory of urban development. Their unique cultural value and historical significance are irreplaceable. With the relentless expansion of cities, some historic structures—due to inadequate identification and protection—are demolished and rebuilt. Others, lacking ongoing monitoring, deteriorate from neglect without timely intervention, ultimately vanishing from the urban landscape. The urban cultural fabric has consequently fractured, local identity gradually fading. This issue stems from the architectural survey's reliance on specialized expertise and high costs, making preservation efforts or individual research difficult to implement.

In the past, surveying traditional historical buildings faced significant challenges, such as requiring specialized remote sensing equipment for image acquisition, professional software for data processing, and on-site verification by experts. The entire process was complex and costly, inadvertently raising high technical and financial barriers. Local units and some researchers often lack access to professional remote sensing data and cannot afford the associated equipment and software costs. This has left historical building identification efforts in their regions stuck at the stage of fragmented documentation, making it difficult to produce systematic and comprehensive results.

With the advancement of modern technology, people can increasingly access high-resolution satellite imagery with greater ease, applying it to fields such as navigation, urban environments, agricultural monitoring, and disaster prevention—bringing more convenience and security to our lives [2].Against the backdrop of open network data resources, satellite imagery is freely or low-cost accessible on online platforms, and scientific data platforms enable resource sharing. Although these types of images are not collected by professional remote sensing equipment, they can meet the needs of identifying the appearance features of historical buildings. Ordinary computers can run basic recognition models without the need for professional technical background. By preprocessing satellite images downloaded online and building models using open-source deep learning frameworks like TensorFlow or PyTorch, automated identification of historical buildings can be achieved. This combination of “online satellite imagery acquisition + lightweight deep learning” breaks through the technical and cost barriers of traditional surveys, providing a viable solution for identifying historical buildings without professional remote sensing support. It has also become a crucial approach for filling gaps in regional historical building identification.

ISSN 2616-5775 Vol. 8, Issue 9: 40-45, DOI: 10.25236/AJCIS.2025.080906

2. Current State of Research

Deep learning-based image classification has emerged as a prominent research focus in China in recent years, with numerous practical explorations integrating regional satellite imagery becoming key areas of study. In 2019, Lizhong Wang et al. focused on township buildings, establishing a dedicated sample database for townships. By employing models such as Faster-RCNN, they successfully improved recognition accuracy. In 2020, Zedi Gao et al. explored convolutional neural networks to construct deep learning models, utilizing Landsat multispectral TM imagery for building recognition, thereby validating the technology's applicability in building identification scenarios[3].As research deepens, the focus increasingly shifts toward innovative model development. In 2023, Xiong Bin et al. Optimized the pspnet architecture by integrating the Transformer attention mechanism, thereby enhancing semantic segmentation accuracy[4]. By 2025, Zhehui Li et al. Innovatively combined the SAM large model with the maskr-CNN framework to achieve high-precision recognition of specific building types[5].

The above research has achieved significant results to date, yet several challenges remain: equipment compatibility issues stemming from large model sizes; high acquisition costs for the high-quality data required for model training; and insufficient adaptability to distinctive architectural styles. Based on the issues raised above, this study did not choose a more complex model, but instead built a MobileNetV2 lightweight model based on local satellite building images. By leveraging localized enhancement and difficult-to-classify example mining, we aim to design a more efficient and cost-effective historical building recognition solution.

3. Scope of Research and Research Data

3.1 Scope of Research

Through field visits and on-site inspections, the scope of this study was defined as historical buildings within Leshan City that have stood for over a century. The research subjects encompass residential clusters in Suji Town, Luocheng Town, Qingxi Town, Jianban Town, Mucheng Town, Luomu Ancient Town, and Ancient Building Complex in the Two River Mouth of Wutong Bridge. Additionally, representative individual historical buildings such as the Former Residence of Lei Chang, the Longshen Temple in Municipal District, the Song Family Ancestral Hall in Shuikou, and the Former Residence of Guo Moruo were included. This formed a research sample combination of “ancient town clusters + independent landmarks.”

3.2 Research Data

The satellite imagery data selected for this study was sourced from the online version of the Jilin-1 2024 National Remote Sensing Image Map. During the research process, specialized image cropping tools were employed to extract satellite image segments corresponding to the selected historical building distribution areas within Leshan City. This ensured that the imagery comprehensively covered all study subjects and their surrounding critical environmental zones, providing foundational data support for subsequent spatial analysis and architectural feature extraction.

4. Research Models and Design

4.1 Model Architecture

This paper focuses on the classification of historical and modern buildings in Leshan, addressing practical challenges such as the resolution limitations of open-source satellite imagery and the difficulty in capturing architectural details. A Convolutional Neural Network (CNN) is selected as the foundational model, paired with the lightweight MobileNetV2 architecture for deep learning. At this resolution, traditional complex convolutional neural networks (CNN) often suffer from inefficient feature extraction due to redundant parameters. In contrast, mobilenetv2's depthwise convolutions effectively reduce computational load while prioritizing the capture of macro-level features like building roof materials. This approach better aligns with the practical constraints of current satellite imagery data.

To enhance the ability to capture key architectural features, a channel attention module is connected after the mobilenetV2 network[6].The spatial dimensions of multi-scale feature maps are compressed via a Global Average Pooling 2D operation, generating a single-channel feature weight vector. This vector

ISSN 2616-5775 Vol. 8, Issue 9: 40-45, DOI: 10.25236/AJCIS.2025.080906

then undergoes learning through a fully connected layer (with hidden layer nodes equal to 1/4 of the input channel count), ultimately outputting a weight matrix matching the original feature map's channel count. This module assigns higher weights to channels corresponding to morphological and textural features of historical buildings in Leshan, such as gray-tiled pitched roofs and building outlines in residential structures. Conversely, it reduces the influence of channels representing background noise like mountain shadows and water reflections, effectively resolving the confusion problem between architectural and environmental features in open-source satellite imagery. Simultaneously, to address the imbalance between the two sample categories, a dynamic sample balancing strategy is designed, ultimately achieving efficient classification based on open-source satellite imagery. The core workflow of the model can be represented as Equation (1):

$$I \xrightarrow{\text{MobileNetV2 CNN}} F \xrightarrow{\text{Channel Attention}} F' \xrightarrow{\text{GAP+FC}} P \tag{1}$$

Among them, I is a 448 × 448 × 3 standardized satellite image, F is the multi-scale feature map output by CNN, F 'is the attention optimized feature map, and P is the binary classification probability vector (P=[p historical, p modern]).

4.2 MobileNetV2 and Transfer Learning

The MobileNetV2 network model is a lightweight network model[7]that divides the traditional CNN single step multi-channel feature extraction and fusion operation into two steps: "deep convolution (single channel local feature extraction)" and "point by point convolution (1×1 convolution kernel channel fusion)". Compared to traditional convolution, this structure can significantly reduce the number of parameters and computational complexity. The compression ratio formula is (2):

$$\text{Compression Ratio} = \frac{1}{C_{\text{out}}} + \frac{1}{K^2} \tag{2}$$

When K=3 (the commonly used convolution kernel size for CNN) and C_{out}=1280 (the number of output channels for MobileNetV2), the compression ratio is about 0.11, which is only 11% of the parameters and computational complexity of traditional CNN. In this study, the CNN model had a total parameter of 3.24MB and a trainable parameter of 2.19MB. The single inference time on the CPU was only 52 seconds, fully meeting the research scenario requirements without the need for professional computing equipment.

To strike a balance between the general visual feature extraction capability of the pre-trained model and its adaptability to the architectural style of Leshan, the model employs a MobileNetV2 freeze-and-fine-tune transfer learning strategy, implementing layered control over the pre-trained layers. Initially, the plan was to freeze the first L-15 layers (where L represents the total number of layers in the network) and fine-tune only the last 15 layers. However, preliminary experiments revealed significant overfitting on the validation set, with validation accuracy peaking at 0.96 before plummeting to 0.88. Through Grad-CAM feature visualization analysis (as an auxiliary tool), it was discovered that when the top layers were overly numerous, the model tended to overlearn non-architectural features in the Leshan satellite imagery, such as the grayscale gradients of mountain shadows and the linear textures of roads. These features, unrelated to the building types, occurred frequently in the training set. Consequently, the strategy was adjusted to fine-tune only the top 20 layers, preserving the generic features of the lower layers (such as edges and textures) and enabling the top layers to concentrate on learning architectural-specific features.

During the training phase of the model, the sparse cross-entropy loss function (Equation (3)) is employed to process integer-label classification tasks (where 0 denotes historical buildings and 1 denotes modern buildings). The computation of this loss function relies on the batch size N, the true label yi of the samples, and the model's predicted probability $P_{i,\ yi}$ for the true label.

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log\left(P_{i,yi}\right) \tag{3}$$

In the configuration of hyperparameters, the learning rate η is set to 0.0001 to ensure stability during parameter updates and avoid gradient oscillations caused by excessively high learning rates. An early stopping mechanism (patience=5) is also introduced to monitor and verify accuracy, allowing the model to automatically terminate training when performance cannot be improved, preventing overfitting and ensuring stable convergence of the model.

ISSN 2616-5775 Vol. 8, Issue 9: 40-45, DOI: 10.25236/AJCIS.2025.080906

4.3 Sample Processing and Sample Balancing

There are three problems with the sample of historical buildings in Leshan, including a small number, low resolution, and the influence of climate on the images. For example,cloudy weather leads to a slightly greenish color in the images and uneven brightness, which greatly affects the feature extraction performance. To address these issues, the following data augmentation methods have been designed:

▪ Use the exposure.adjust_gamma function to control brightness adjustment within the range of 0.7–1.3, accurately simulating the light variations characteristic of Leshan's rainy and foggy weather;

▪ When enhancing the green channel, set the weighting coefficient to 1.0–1.2 and reduce the blue channel coefficient to 0.8–1.0 to match the green-dominant color characteristics of local overcast images;

▪ When slightly adjusting image angles, control perspective distortion offset within 0.1 of the image dimensions to prevent excessive deformation that could distort architectural features;

▪ Add low-intensity Gaussian noise (mean 0, standard deviation 0.02) to simulate signal interference during actual shooting, enhancing the model's robustness against disturbances.

To address the scarcity of historical building samples, this paper employs a “1:2” ratio to generate augmented variants—producing two augmented samples from each original image to directly expand the dataset. During training, a custom function constructs a classification augmentation method to control the sample ratio between historical and modern buildings. Each training batch contains two-thirds historical building samples and one-third modern building samples. Simultaneously, sample order is randomly shuffled to ensure the model learns features from both building types stably and evenly throughout training, mitigating learning bias caused by sample quantity disparities.

The data in the study is divided in an 8:1:1 ratio, encompassing the training set, validation set, and test set. During data processing, the total dataset is first split into a training set and a temporary set at an 8:2 ratio. Subsequently, the temporary set is evenly divided into a validation set and a test set to prevent data crossover and evaluation bias,ensuring fairness and accuracy in the model's evaluation. Additionally, the experiment incorporates a difficult case mining step. Based on the model's predictions, samples where "historical buildings are misclassified as modern buildings" are identified, the number of such samples is expanded threefold, and then these samples are incorporated into the training set for three rounds of targeted training.This approach aims to strengthen the model's ability to learn key features of historical buildings.

5. Experimental Results

5.1 Model Training and Validation

The experimental results show that the model training curve (Figure 1) indicates that the training accuracy ultimately approaches 1.0, the validation accuracy remains stable above 0.95, and the model training and validation losses (Figure 2) converge to below 0.05 and 0.1, respectively, without significant overfitting. This proves that the model training is stable and reliable, with good feature extraction and classification capabilities.

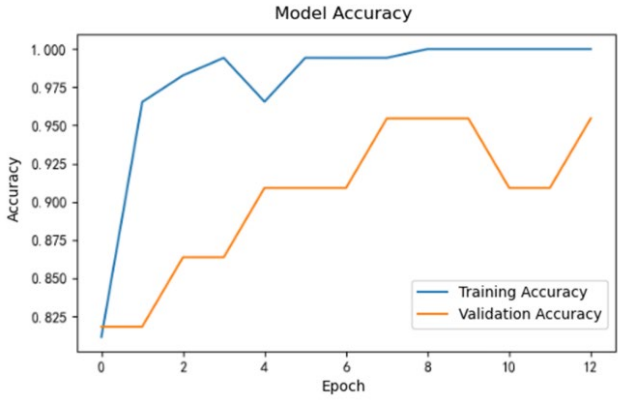

Figure 1 Training/Validation Accuracy Curve

ISSN 2616-5775 Vol. 8, Issue 9: 40-45, DOI: 10.25236/AJCIS.2025.080906

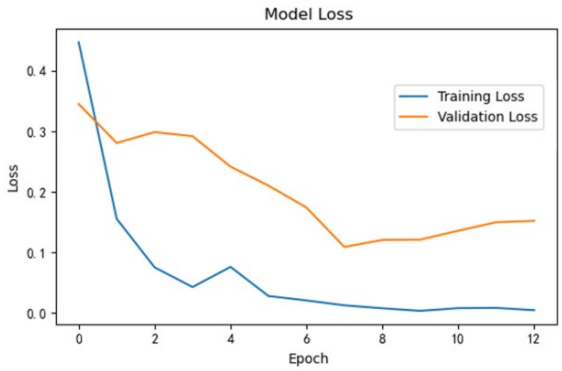

Figure 2 Training/Verification Loss Curve

5.2 Test Set Classification Results

The confusion matrix (Figure 3) indicates that among the 22 samples in the test set, all 17 historical buildings were correctly identified, while 3 out of the 5 modern buildings were accurately recognized, achieving a total accuracy rate of 90.9%. Notably, there were no omissions in the identification of historical buildings, fully satisfying the "low omission" requirement of the historical building census. In the example of prediction results (Figure 4), all three historical building samples were precisely identified, and the recognition error of modern building samples is controllable, further confirming that the model can effectively capture the core characteristics of historical buildings.

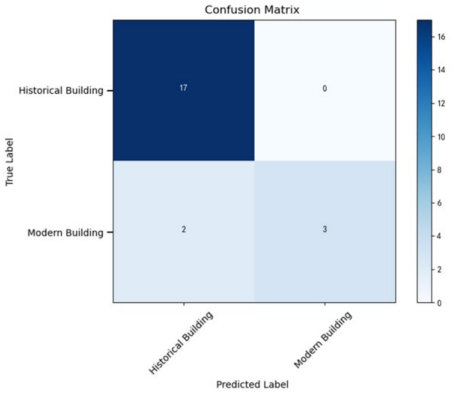

Figure 3 Confusion Matrix

Figure 4 Example of Prediction Results

ISSN 2616-5775 Vol. 8, Issue 9: 40-45, DOI: 10.25236/AJCIS.2025.080906

6. Conclusion

This study designed a deep learning model combining MobileNetV2 with an attention mechanism to address the conservation needs of historical buildings in Leshan. After expanding the excavated samples and retraining, the training logs indicate that this method reduces the average number of misclassified historical buildings by approximately 3 per batch, significantly lowering the likelihood of missed detections. It has achieved excellent results in the classification and recognition of historical and modern buildings in Leshan.

Future efforts will focus on expanding the model's application to multi-type building classification. Leshan boasts a diverse range of architectural styles, and enhancing the model's multi-category classification capabilities can help the system capture the characteristics of regional architectural styles and provide data support for conservation decisions. Additionally, we will optimize the model structure to enhance computational speed and energy efficiency. Exploring and advancing these areas will facilitate the application of models in conservation efforts, enable the deployment of monitoring and early warning systems, and offer new avenues for the contemporary preservation and utilization of Leshan's historical and cultural heritage.

References

[1] Zhongshu Zhao, Weijie Lan. Formation and Evolution of Key Concepts in the Protection System for Historic and Cultural Cities [J]. Urban Planning, 2022, 46(S2): 20-26.

[2] Lizhong Wang, Honghai Zhang, Bo Zhong, et al. Deep Learning-Based Method for Identifying Township Buildings in High-Resolution Remote Sensing Images [J]. Information Technology and Application in Scientific Research, 2019, 10(01): 88-95.

[3] Zedi Gao, Jianfeng Gao. Building Recognition Method for Landsat Satellite TM Images Based on Convolutional Neural Networks [J]. Information Technology and Informatization, 2020, (11): 196-199.

[4] Bin Xiong, Shuangde Zhang. A Semantic Segmentation Algorithm for Buildings in Satellite Remote Sensing Images Based on an Improved PSPNet [J]. Remote Sensing Information, 2023, 38(04): 73-79. DOI: 10.20091/j.cnki.1000-3177.2023.04.009.

[5] Zehui Li ,Ningning Zhu ,Chong Zhao, et al. Research on Building Type Identification Methods Based on Deep Learning of Satellite Images: The Case of Singapore's "Shophouses" [J]. New Architecture, 2025, (02): 93-98.

[6] Chenchen Liu,Xiaosan Ge,Yongbin Wu. Building Change Detection Integrating Twin Neural Networks and Mutual Attention [J]. Remote Sensing Information, 2024, 39(05): 70-77. DOI: 10.20091/j.cnki.1000-3177.2024.05.009.

[7] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted residuals and linear bottlenecks [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.