

Article

Small-Sample Recognition and Classification of Ancient Architectural Forms in Chinese Local Gazetteer Maps Using Deep Learning Method

Lan Li ^{1,2}, Feng Kang ¹, Xue Li ³, Shujia Zhang ^{3,*}  and Zhongchao Zhou ⁴ 

¹ College of Architecture and Art, Taiyuan University of Technology, Taiyuan 030024, China; lilan02@tyut.edu.cn (L.L.); kangfeng@tyut.edu.cn (F.K.)

² Postdoctoral Research Workstation, Shanxi Construction Investment Group Co., Ltd., Taiyuan 030032, China

³ College of Architecture and Design, Shanxi Vocational University of Engineering Science and Technology, Jinzhong 030619, China; lixue1@sxgkd.edu.cn

⁴ School of Engineering, The University of Tokyo, Tokyo 113-8654, Japan; zhouzhongchao@outlook.com

* Correspondence: zhangshujia1@sxgkd.edu.cn

Abstract

Chinese local gazetteer maps contain visual evidence of historical cities and architecture, but their hand-drawn, low-texture, and small-sample characteristics make automatic recognition difficult. This study develops a transfer-learning workflow to classify gazetteer-derived architectural images and examines its usefulness for digital heritage image organization, indexing, and retrieval. A dataset of 685 images covering nine classification labels was constructed from Chinese local gazetteer maps. Five representative models, ResNet-18, MobileNetV3-Small, EfficientNet-B0, DenseNet-121, and ViT-B/16, were trained and compared under the same transfer-learning setting. Their performance was evaluated using accuracy, macro-average F1 score, class-level metrics, confusion matrices, difficult-sample analysis, and Grad-CAM visualization. CNN-based models outperformed the Vision Transformer on this task. EfficientNet-B0 achieved the best results, with a test accuracy of 87.38% and a macro-average F1 score of 86.60%; after disabling class weights, the macro-average F1 score increased to 86.81%. Misclassifications mainly occurred in visually similar or context-dependent category pairs, including gate tower and drum tower, bell tower and drum tower, and buildings inside and outside the city. The results show that compact CNNs can support classification, indexing, and retrieval of hand-drawn architectural images in gazetteer archives, although the workflow should be regarded as an auxiliary tool rather than direct evidence for specific extant buildings.

Keywords: Chinese local gazetteer maps; ancient architectural forms; hand-drawn architectural images; image recognition; small-sample classification; deep learning method; transfer learning; Grad-CAM



Academic Editor: Antonio Formisano

Received: 16 June 2026

Revised: 23 July 2026

Accepted: 29 July 2026

Published: 2 August 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The digitization of cultural heritage has moved beyond the preservation of source materials toward their computational use. Paper-based documents, historical images, and architectural heritage can enter systems of long-term management, knowledge organization, and public dissemination only when they are transformed into data that can be retrieved, classified, and reused [1,2]. In recent years, deep learning has been applied to tasks such as architectural heritage image classification, cultural heritage image recognition, ceramic

sherd classification, visual inspection of heritage assets, and the recognition of architectural heritage components [3–10]. These studies show that artificial intelligence does more than accelerate image processing; it also provides technical support for heritage image indexing, archival organization, and sustainable heritage management.

Chinese local gazetteers are important documentary sources for local history, geography, ritual institutions, urban space, and public architecture. Many gazetteers include maps and pictorial materials, such as territorial maps, walled-city plans, mountain-and-river maps, altar and temple maps, academy maps, bridge maps, scenic-view illustrations, and images of architectural ensembles. The Historical Atlas of Chinese Urban Human Settlements and related studies indicate that these images are not merely illustrative supplements but legible visual documents. They present traditional urban living environments, landscape configurations, the ordering of public architecture, and local spatial experience in a more immediate visual form [11–14]. Research tools and database-oriented approaches in local gazetteer studies further demonstrate that these materials are highly structured and have considerable computational potential [15–18]. Ancient architectural forms in local gazetteer maps therefore have documentary value as well as value for image recognition and spatial knowledge modeling. It should be noted, however, that the gazetteer sketches examined in this study are historical visual representations rather than measured drawings, field photographs, or direct records of extant buildings. Therefore, this study analyzes them as hand-drawn architectural-form images for classification, indexing, and retrieval. It does not aim to reconstruct, verify, or authenticate individual surviving buildings.

Automatic analysis of historical maps and gazetteer illustrations has become a shared concern of digital humanities, geographic information science, and computer vision. Previous studies have examined map evolution, weakly supervised segmentation, road extraction, vectorization, building-footprint reconstruction, three-dimensional reconstruction of historic cities, archaeological feature extraction, bridge segmentation, wetland identification, and semantic exploration of large map collections [19–32]. These studies show that deep models can help identify spatial elements that would otherwise require time-consuming manual annotation.

However, the architectural images examined in this study differ from standardized surveying-map or cartographic symbols. Many historical-map tasks focus on relatively regular spatial elements, such as roads, boundaries, building footprints, map symbols, and land-use units, which are usually planimetric and convention-based. By contrast, architectural forms in local gazetteer maps are pictorial, hand-drawn, low in texture, variable in scale and perspective, and often simplified into line-based representations. Their recognition therefore depends less on fixed cartographic conventions and more on recurring architectural morphology, including roof profiles, eave tiers, tower bodies, hall bodies, passageway bases, facade divisions, and overall silhouettes.

From a methodological perspective, convolutional neural networks and other deep learning models provide the basis for visual classification, object recognition, and semantic segmentation [33,34]. ImageNet-pretrained architectures, including ResNet, DenseNet, MobileNet, EfficientNet, and the Vision Transformer, have become common choices for transfer learning [35–41]. DeepLab supports segmentation tasks, and Grad-CAM helps visualize image regions that influence model decisions [42,43]. These methods make it possible to compare models and interpret their behavior in this study. Even so, the recognition of ancient architectural forms in gazetteer maps cannot be judged only by overall accuracy. It is also necessary to examine whether the model responds to meaningful architectural cues, including roof outlines, tower-body proportions, gate-tower structures, and building silhouettes.

Small sample size and class imbalance are the main challenges in this study. Heritage images are dispersed across sources, annotation requires expert judgment, and available datasets are often limited in scale. Studies on representation learning, data augmentation, small-dataset classification, class-imbalanced learning, class-balanced loss, and focal loss provide methodological support for mitigating these problems [44–50]. In related heritage contexts, transfer learning and moderate augmentation have been used for Kolkata heritage images, historic building components in Brazil, historical courtyard houses, and vernacular architecture in satellite imagery [7,9,51,52]. These findings suggest that pretrained models can be transferred to heritage recognition tasks under limited-sample conditions. For local gazetteer maps, however, augmentation strategies must be restrained: excessive rotation, cropping, or deformation may damage key architectural cues, including roof configuration, tower massing, and proportional relationships.

Recent work has produced a broader range of deep learning studies related to architectural heritage. Relevant tasks include style recognition for Cypriot architectural heritage, heritage digital twins, point-cloud datasets for dougong bracket sets, detection of roof-ridge ornaments, recognition of Lingnan ceramic-inlay images, damage detection in wangzhuang paving bricks in Suzhou gardens, disease segmentation of brick pagodas, pathology recognition in historic masonry buildings in Suzhou, multi-label image classification of Chinese ancient architecture, dynasty classification of roof-end tiles, three-dimensional reconstruction of Song-dynasty architecture from image evidence, digital regeneration of historical images of Qing-dynasty tea gardens, and crowdsourced-image-based 3D reconstruction of ancient buildings [53–65]. Collectively, these studies indicate that visual recognition in cultural heritage is shifting from generic image classification toward finer-grained architectural components, material deterioration, historical imagery, and three-dimensional semantic management.

Hand-drawn and line-based images further increase the difficulty of classification. Studies on paper-cut image classification, hand-drawn diagram recognition, sketch recognition, and sketch retrieval show that low-texture images rely more heavily on contours, local parts, keypoints, and geometric relationships [66–68]. Ancient architectural forms in local gazetteer maps share similar characteristics. Bell towers, drum towers, gate towers, pagodas and tower pavilions often share vertical compositions and multi-tiered roof forms, while pavilions, temples, and towers may also present similar roof configurations. Existing studies on residential-style recognition and image-based landscape research on Tibetan and Qiang dwellings further suggest that deep models can learn differences among architectural categories from roofs, facades, doors and windows, and overall building form [69,70]. The recognition of ancient architectural forms in local gazetteer maps therefore differs both from natural-image classification and from the extraction of standardized cartographic symbols.

Taken together, prior research has established a foundation in digital cultural heritage, local gazetteer databases, historical map interpretation, deep visual models, small-sample learning, hand-drawn image recognition, and architectural heritage image classification. However, object-level, small-sample classification of non-standardized hand-drawn architectural forms in Chinese local gazetteer maps remains insufficiently explored. This gap is reflected in four aspects. First, existing historical map studies primarily address roads, water systems, building footprints, map symbols, or land-use features, while architectural heritage image classification is usually based on site photographs, remote-sensing imagery, or three-dimensional point clouds. Annotated datasets specifically designed for hand-drawn architectural forms in local gazetteer maps therefore remain limited. Second, few studies have compared different deep learning architectures under the same small-sample setting for gazetteer-derived architectural sketches. Third, although classification studies commonly report accuracy, precision, recall, F1 score, and confusion matrices, misclassifi-

cation is seldom explained in relation to architectural morphology, such as roof profiles, eave tiers, tower bodies, arched openings, facade divisions, and overall silhouettes [3,4,7,9]. Fourth, although visual interpretation methods such as Grad-CAM are widely used in computer vision, their results are often discussed as general attention maps rather than being connected with architectural semantic cues, such as roof forms, eave tiers, tower bodies, passageway bases, facade divisions, and silhouettes [43]. To address these gaps, this study uses gazetteer-derived architectural image samples from Chinese local gazetteer maps as its object of analysis, constructs a small-sample dataset containing 685 images across nine classification labels, and compares ResNet-18, MobileNetV3-Small, EfficientNet-B0, DenseNet-121, and ViT-B/16 through transfer learning. Accuracy, macro-average F1 score, class-level metrics, confusion matrices, and Grad-CAM are then used to evaluate model performance, examine sources of misclassification, and interpret model attention in low-texture, hand-drawn architectural images.

The contributions of this study are fourfold. First, it constructs a small-sample classification dataset of hand-drawn ancient architectural forms extracted from Chinese local gazetteer maps, an image resource with significance for local history and digital architectural heritage. Second, it systematically compares multiple CNN and Transformer architectures under a unified data split and training protocol. Third, it interprets model misclassification from the perspectives of architectural morphological similarity, class imbalance, and small-sample learning. Fourth, it uses Grad-CAM to examine whether the models attend to architecturally meaningful regions, such as roof profiles, tower bodies, and building silhouettes. The findings can inform automatic annotation of local gazetteer images, digital heritage image indexing, organization of visual records of historical architecture, and human–AI collaborative cultural heritage management. By reducing the manual burden of classifying historical architectural images, the proposed workflow supports the sustainable reuse, long-term organization, and knowledge-based management of local gazetteer visual archives. Accordingly, the proposed workflow should be understood as a tool for organizing, indexing, and retrieving gazetteer-derived architectural images, rather than as evidence for establishing the exact correspondence between a sketch and a specific real building.

2. Dataset and Preprocessing

2.1. Source Maps and Extraction of Ancient Architectural Forms

Chinese local gazetteers are important classical documentary sources for recording local history, geography, administrative organization, urban space, public buildings, temples, academies, scenic sites, and territorial landscapes. In many gazetteers, maps and illustrations are placed at the beginning of the volume, including walled-city maps, territorial maps, boundary maps, city plans, and eight-scenery illustrations. Although these images are hand-drawn rather than measured architectural drawings, they provide valuable visual evidence for understanding the distribution of ancient buildings and their spatial relationships with city walls, streets, mountains, rivers, settlements, and territorial settings [11–18].

The ancient architectural forms analyzed in this study were extracted from multiple local gazetteer maps, with Figure 1 showing representative source images. They include towers, gate towers, drum towers, bell towers, pavilions, temples, tower pavilions, and building groups inside and outside the city. Compared with photographs or modern cartographic symbols, these images are simplified, low in texture, and strongly dependent on linework, silhouette, roof form, tower body, and spatial composition. Therefore, they are suitable for examining whether deep learning models can recognize ancient architectural forms from small-sample hand-drawn historical images (Figure 1).

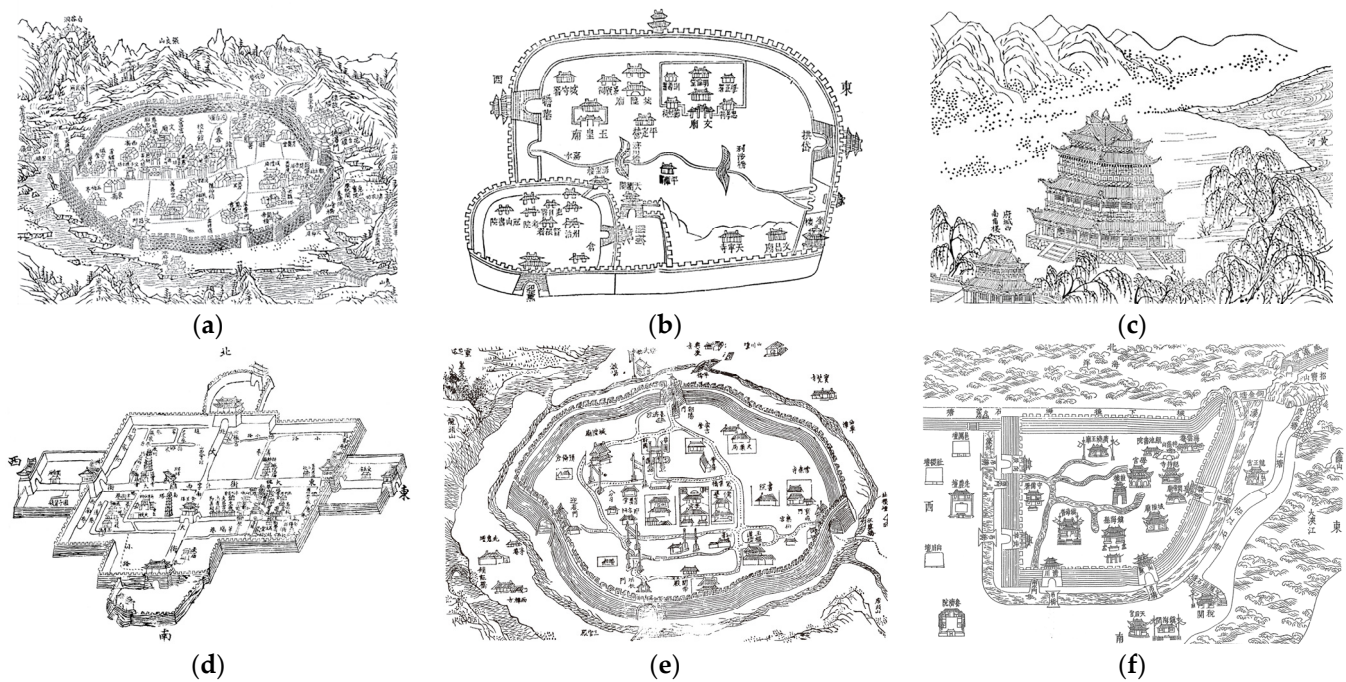


Figure 1. Representative source images from Ming- and Qing-dynasty Chinese local gazetteers used for extracting ancient architectural forms: (a) city-wall map of Jingshan County, Hubei Province; (b) city map of Pingding County, Shanxi Province; (c) Guanque Tower in Puzhou Prefecture, Shanxi Province; (d) city-wall map of Xingping, Shaanxi Province; (e) city-wall map of Anji County, Zhejiang Province; (f) county-seat map of Zhenhai County, Zhejiang Province. The Chinese characters retained in the original images are historical labels indicating place names, building names, and directions; the corresponding explanations of each subfigure are provided in the English caption.

To improve the transparency of the dataset sources, Table 1 provides representative source information for selected local gazetteer maps used in this study. The dataset was compiled from multiple county and provincial gazetteers from Shanxi, Shaanxi, Hubei, and Zhejiang. Because the extracted samples came from many gazetteer volumes, Table 1 lists representative source materials rather than an exhaustive catalogue of all source volumes. The table records the province, administrative scope, gazetteer title, period, publication year, holding institution or digital source, and representative extracted categories.

Table 1. Representative source information for selected local gazetteer maps used in dataset construction.

No.	Province	Administrative Scope	Gazetteer Title *	Period	Publication Year	Holding Institution or Digital Source	Representative Extracted Categories
1	Shanxi	Provincial	<i>Shanxi Tongzhi</i>	Qing	12th year of Yongzheng reign (1734)	National Library of China	tower; pavilion; temple; gate tower
2	Shanxi	Pingding	<i>Pingding Zhouzhi</i>	Qing	8th year of Guangxu reign (1882)	National Library of China	gate tower; buildings inside the city; pavilion
3	Shanxi	Yongji	<i>Yongji Xianzhi</i>	Qing	12th year of Guangxu reign (1886)	National Library of China	tower; temple; pavilion

Table 1. Cont.

No.	Province	Administrative Scope	Gazetteer Title *	Period	Publication Year	Holding Institution or Digital Source	Representative Extracted Categories
4	Shaanxi	Xingping	<i>Xingping Xianzhi</i>	Republican period	12th year of the Republic of China (1923)	National Library of China	drum tower; gate tower; buildings inside the city
5	Shaanxi	Provincial	<i>Shaanxi Tongzhi</i>	Qing	13th year of Yongzheng reign (1735)	National Library of China	tower; pavilion; tower pavilion
6	Hubei	Provincial	<i>Hubei Tongzhi</i>	Republican period	10th year of the Republic of China (1921)	National Library of China	gate tower; tower; buildings outside the city
7	Hubei	Jingshan	<i>Jingshan Xianzhi</i>	Qing	8th year of Guangxu reign (1882)	National Library of China	pavilion; temple; buildings inside the city
8	Hubei	Hanyang Prefecture	<i>Hanyang Fuzhi</i>	Qing	12th year of Qianlong reign (1747)	National Library of China	gate tower; drum tower; buildings inside the city
9	Zhejiang	Zhenhai	<i>Zhenhai Xianzhi</i>	Qing	17th year of Qianlong reign (1752)	National Library of China	buildings inside the city; buildings outside the city; bell tower
10	Zhejiang	Anji	<i>Anji Zhouzhi</i>	Qing	15th year of Qianlong reign (1750)	National Library of China	temple; pavilion; tower
11	Zhejiang	Provincial	<i>Zhejiang Tongzhi</i>	Qing	23rd year of Kangxi reign (1684)	National Library of China	tower; pavilion; bell tower; drum tower

* Italics indicate the titles of local gazetteers or comprehensive gazetteers from which the source images were obtained.

2.2. Dataset Description

The dataset comprises hand-drawn images of ancient architectural forms collected from Chinese local gazetteer maps for a multi-class classification task. The images were arranged in folders, with each folder corresponding to one classification label. This folder structure allowed the data to be used directly for supervised learning.

In this study, “Chinese local gazetteer maps” refers to the source material, “gazetteer-derived architectural images” refers to the cropped visual samples used for model training, and “classification labels” refers to the labels assigned to these samples. This distinction is used throughout the manuscript to avoid confusion between the historical source, the extracted image object, and the classification target.

The dataset was compiled through manual screening and annotation. Complete gazetteer maps were first reviewed to locate architectural elements or building scenes with identifiable visual features. Samples were retained when their main architectural outline, roof form, tower or hall body, passageway base, city-wall context, or scene relationship could be recognized with reasonable confidence. Images with severely incomplete outlines, heavy interference from surrounding map elements, or uncertain category attribution were not included. Initial labels were assigned according to the category definitions described below and were then checked by the research team with reference to architectural

morphology and the original gazetteer-map context. Ambiguous cases were discussed and adjusted until a consistent label was reached, following common practice in supervised image-dataset construction and heritage-image annotation [3,4,7].

The dataset contains 685 images from nine classification labels, including bell tower, buildings inside the city, buildings outside the city, tower pavilion, drum tower, gate tower, pavilion, temple, and tower. The images are mainly stored in JPG, JPEG, and PNG formats. Since the original images were obtained from different sources and may have different resolutions, aspect ratios, and color modes, they were standardized during preprocessing before model training.

Unlike natural scene images, the images in this dataset depict hand-drawn ancient architectural forms or scene-context samples. They are characterized by simplified lines, limited color information, and strong dependence on structural contours, a feature also observed in studies of hand-drawn and line-based image recognition [66–68]. Therefore, the classification task focuses more on architectural morphology and contextual cues, such as roof shape, tower body, building outline, city-wall context, and spatial composition, rather than texture or color features. This makes the dataset suitable for evaluating the ability of deep learning models to recognize structural patterns in small-sample hand-drawn architectural images, a task related to previous studies on hand-drawn diagrams, sketches, and other line-based visual materials [66–68].

Due to copyright and access restrictions related to the original gazetteer images, the complete cropped image dataset cannot be fully released at this stage. To support reproducibility as far as possible, this study provides detailed source information, category definitions, and the sample extraction workflow.

2.3. Category Definition of Ancient Architectural Forms and Scene-Context Labels

The dataset was divided into nine classification labels according to the visual, functional, and contextual characteristics of gazetteer-derived architectural images. Each category represents a specific type of architectural form or scene commonly found in hand-drawn historical architectural materials. The class labels were taken from the dataset folders and used as ground-truth labels during supervised training.

The nine categories include bell tower, drum tower, gate tower, tower, pavilion, temple, tower pavilion, buildings inside the city, and buildings outside the city (Figure 2). Among them, bell tower, drum tower, gate tower, and tower are tower-like architectural types with relatively strong vertical structures. These categories are visually similar in some cases, but they differ in architectural function, position, and structural form [71,72].

The pavilion and temple categories represent traditional building forms with distinctive roof structures and spatial layouts. Pavilions are usually lighter and more open in appearance, while temples often have more enclosed and ceremonial architectural forms [71]. Because tower pavilion overlaps visually with both pavilion and tower, its definition is further clarified below. In contrast, buildings inside the city and buildings outside the city are scene-context labels that may contain multiple buildings or environmental components.

This category definition reflects both architectural typology and visual representation. However, because the images are hand-drawn, some categories may share similar outlines or simplified structural features. Therefore, distinguishing between visually similar categories, especially tower-like forms, is an important challenge in this classification task.

This issue is also related to the structure of the label system. The nine labels used in this study do not all belong to the same semantic level. Most categories, including tower, pavilion, temple, bell tower, drum tower, gate tower, and tower pavilion, are architectural-type labels because they mainly describe the morphology, function, or recognizable form of individual architectural elements. In contrast, buildings inside the city and buildings out-

side the city are scene-context labels. These two labels do not represent strict architectural typologies; instead, they describe the spatial position of building groups in the original local gazetteer maps. Their distinction depends on contextual cues such as city walls, settlement boundaries, surrounding landscapes, and the relative position of buildings within the mapped environment.

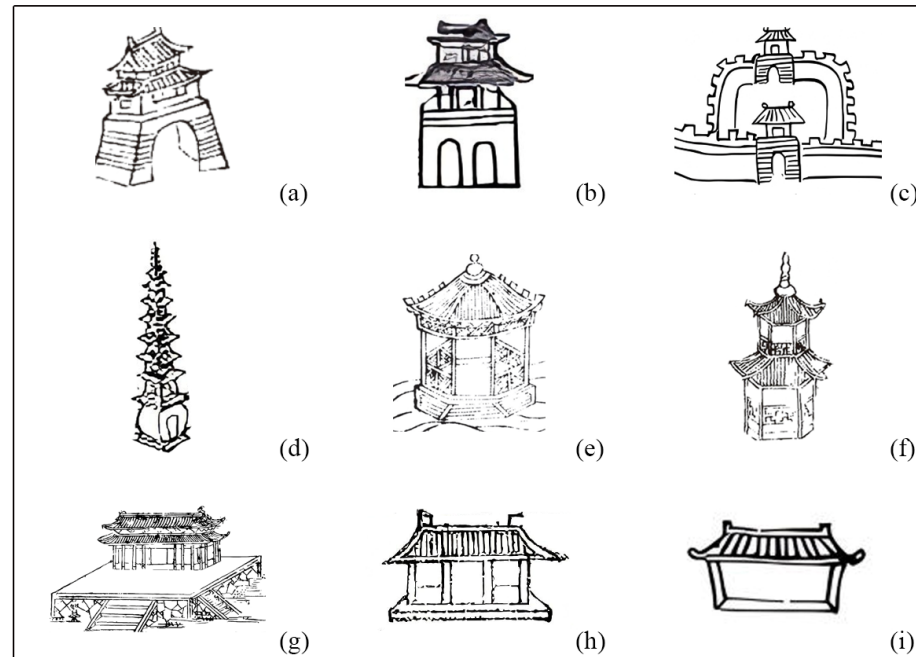


Figure 2. Representative samples of the nine categories of ancient architectural forms extracted from Chinese local gazetteer maps: (a) drum tower; (b) bell tower; (c) gate tower; (d) tower; (e) pavilion; (f) tower pavilion; (g) temple; (h) buildings inside the city; (i) buildings outside the city.

In this study, tower pavilion refers to a multi-storey pavilion-like structure with a tower-like vertical composition and repeated eave tiers. It is distinguished from pavilion by its taller and more enclosed multi-storey body, and from tower by its broader pavilion-like roof-and-floor composition rather than a narrow pagoda-like vertical silhouette [71,72].

To make the visual basis of the category definitions clearer, Figure 3 presents representative annotated samples from selected architectural categories. These examples show that the classification of gazetteer architectural images depends mainly on recurring morphological cues, such as roof form, eave tiers, tower or hall body, passageway base, city-wall body, platform base, and overall silhouette [71,72]. The figure also illustrates the main visual challenges of the dataset: low texture, simplified linework, and partial similarity among architectural forms.

To make the dataset construction process more transparent, Figure 4 illustrates the sample extraction workflow from complete local gazetteer maps to labeled image samples used for model training. The workflow includes selecting architectural regions, cropping them into object-level or scene-context samples, standardizing the cropped images as model inputs, and assigning category labels.

To ensure a consistent sample extraction standard, architectural-type samples were cropped to retain the complete architectural outline, including the main building body, roof, base, and visible structural contour where available, while minimizing unrelated background elements, decorative map lines, and adjacent buildings. For scene-context samples, such as buildings inside the city and buildings outside the city, a wider image area was retained when necessary to preserve spatial cues from the original gazetteer map, including city walls, settlement boundaries, surrounding landscape, and the relative

position of building groups. All cropped samples were then standardized in size and format before model training.

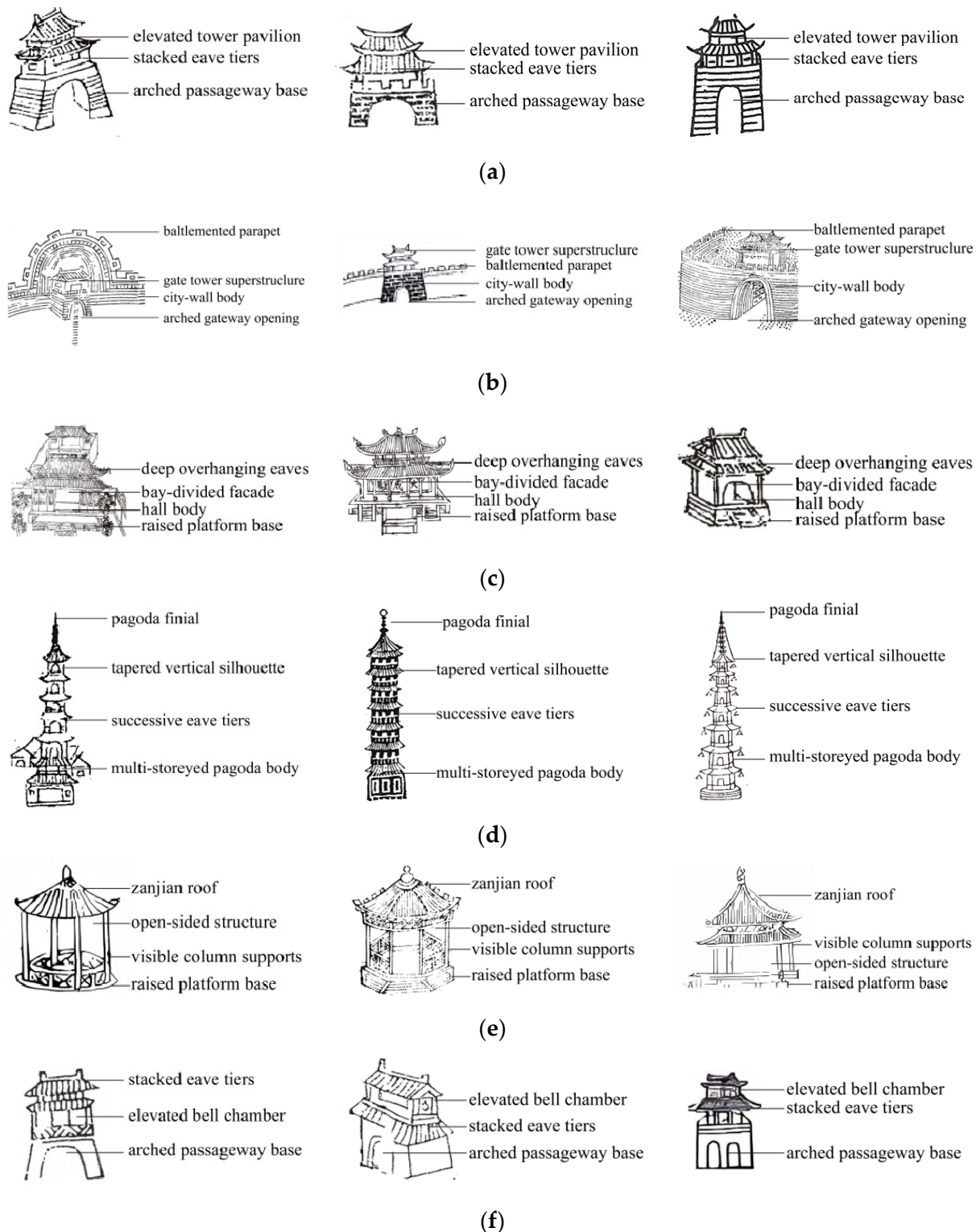


Figure 3. Representative annotated gazetteer sketches and morphological cues for selected architectural-form categories: (a) drum tower; (b) gate tower; (c) temple; (d) tower; (e) pavilion; (f) bell tower. Each row presents three representative sketches from the same category. The annotations highlight recurring visual cues within gazetteer images rather than one-to-one correspondence with specific extant buildings.

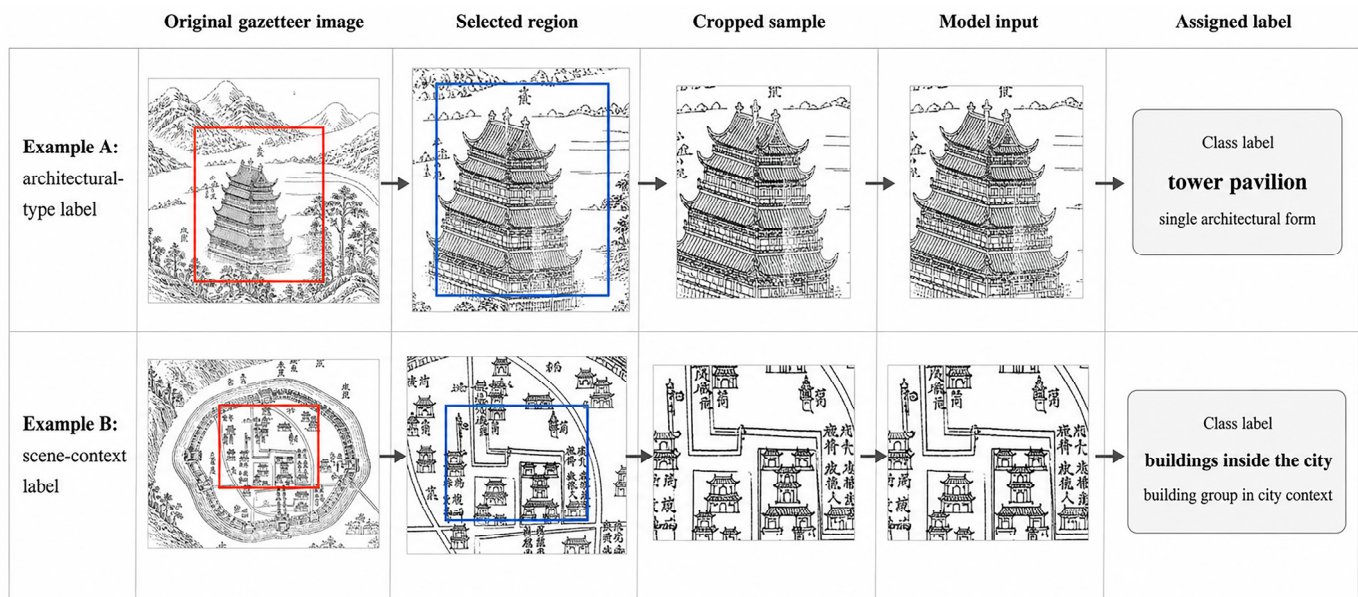


Figure 4. Sample extraction workflow from Chinese local gazetteer maps. The figure shows how architectural regions were selected from complete gazetteer images, cropped into object-level or scene-context samples, standardized as model inputs, and assigned category labels for supervised learning. The Chinese characters retained in the gazetteer image samples are original historical labels, mainly indicating place names, building names, and directions, and do not affect the interpretation of the extraction workflow.

2.4. Data Distribution and Class Imbalance

The dataset contains 685 images distributed across nine classification labels. Although the dataset was supplemented during the study, the number of images in each label remains uneven. Table 2 reports the number of samples in each class.

Table 2. Class distribution of the hand-drawn ancient Chinese architectural form dataset.

Class	Number of Images	Proportion (%)
bell tower	51	7.45
buildings inside the city	101	14.74
buildings outside the city	70	10.22
tower pavilion	68	9.93
drum tower	64	9.34
gate tower	91	13.28
pavilion	70	10.22
temple	78	11.39
tower	92	13.43
Total	685	100.00

As shown in Table 2, buildings inside the city is the largest category, containing 101 images, while bell tower is the smallest category, containing 51 images. The imbalance ratio between the largest and smallest categories is approximately 1.98:1. Although this imbalance is not extremely severe, it may still influence the learning process, especially under small-sample conditions.

Class imbalance may cause a model to learn more strongly from categories with more examples while producing less stable predictions for categories with fewer samples [48–50]. To reduce this risk, the dataset was divided by stratified sampling so that the training, validation, and test sets retained the original class proportions as closely as possible. Class-

level precision, recall, F1-score, and macro-average F1 were also reported, allowing model performance to be assessed beyond overall accuracy.

2.5. Image Preprocessing

Before model training, all images were preprocessed to ensure input consistency across different deep learning models. Since the original images were stored in different formats, including JPG, JPEG, and PNG, and some images had different color modes such as RGB, RGBA, and grayscale, all images were first converted into RGB format. This step ensured that each image had three color channels and could be processed by pretrained models designed for RGB input.

The images in the dataset also varied in resolution and aspect ratio. To provide a unified input size, all images were resized to 224×224 pixels. This size was selected because it is commonly used by ImageNet-pretrained models, including ResNet-18, MobileNetV3-Small, EfficientNet-B0, DenseNet-121, and ViT-B/16. Standardizing the image size also allowed all models to be trained and compared under the same input setting.

After resizing, pixel values were normalized with the ImageNet mean and standard deviation. This step was used because all models started from ImageNet-pretrained weights. Matching the input distribution to the pretraining setting helps the pretrained feature extractors operate more reliably.

Overall, the preprocessing procedure included RGB conversion, image resizing, tensor transformation, and ImageNet-based normalization. These steps helped reduce inconsistencies caused by different image formats, resolutions, and color modes, while ensuring that the dataset could be effectively used for transfer learning-based classification.

2.6. Data Augmentation

Data augmentation was used only on the training set to improve generalization with limited samples. Because the dataset contains relatively few drawn architectural images, training without augmentation could make the models memorize specific samples. Moderate transformations were therefore introduced to increase visual variation while preserving the essential building structure.

The augmentation operations included random horizontal flipping, slight affine transformation, limited translation, scaling, shearing, and brightness or contrast adjustment. Horizontal flipping increased spatial variation, while small affine changes encouraged the models to learn more robust structural features. Brightness and contrast adjustment simulated differences in scanning quality and drawing style.

Because the recognition of ancient architectural forms depends strongly on structural outlines and proportions, strong geometric distortion was avoided. Excessive rotation, cropping, or deformation could change the structural characteristics of the buildings and introduce misleading visual patterns. Therefore, all augmentation operations were kept within a moderate range to maintain the original architectural morphology.

Overall, augmentation reduced the risk of overfitting and made the models more robust when learning from a small hand-drawn image dataset.

2.7. Training, Validation, and Test Split

The dataset was split into training, validation, and test subsets with stratified sampling. Specifically, 70% of images were assigned to training, 15% to validation, and 15% to testing. The training subset optimized model parameters, the validation subset guided model selection and early stopping, and the test subset was reserved for final evaluation.

Stratification was used to maintain class proportions across the three subsets. This was necessary because the dataset has unequal class sizes; without stratification, minority classes might be poorly represented in validation or testing, leading to unstable evaluation.

The final division produced 479 training images, 103 validation images, and 103 test images, ensuring that all categories were represented in each subset and that the final test results reflected the classification ability of the models on unseen data. The label-level distribution of these subsets is shown in Table 3.

Table 3. Dataset split for training, validation, and testing.

Class	Training Set	Validation Set	Test Set
bell tower	36	7	8
buildings inside the city	71	15	15
buildings outside the city	49	10	11
tower pavilion	47	11	10
drum tower	45	10	9
gate tower	64	13	14
pavilion	49	11	10
temple	54	12	12
tower	64	14	14
Total	479	103	103

3. Methodology

3.1. Overall Framework

This study treats the recognition of hand-drawn ancient architectural forms in Chinese local gazetteer maps as a supervised multi-class image classification problem. The framework includes model construction, transfer-learning-based training, and performance evaluation. Five models, ResNet-18, MobileNetV3-Small, EfficientNet-B0, DenseNet-121, and ViT-B/16, were selected to compare different approaches to visual representation under identical experimental conditions.

Each model was initialized with pretrained weights, and its original classifier was replaced with a new output layer for the target categories. The models were then fine-tuned on the training set and evaluated on the independent test set using both overall and class-level metrics.

3.2. Transfer Learning Strategy

Given the limited number of samples, transfer learning was used to improve training stability and reduce overfitting. The selected architectures were initialized with ImageNet-pretrained weights, which already encode general visual features such as edges, contours, textures, and structural patterns.

For every architecture, the original classification layer was replaced by a new layer matching the nine architectural categories. The full network was fine-tuned so that both feature extraction and classification components could adapt to the drawn architectural forms in the gazetteer-map dataset.

3.3. Model Architectures

Five deep learning models were used in this study to compare different visual representation mechanisms. The selected models differ in network structure, parameter efficiency, feature reuse strategy, and global dependency modeling ability.

3.3.1. ResNet-18

ResNet-18 served as a compact CNN baseline. Its residual connections allow information and gradients to pass through the network more effectively, which helps reduce degradation in deeper networks and supports stable training.

Here, ResNet-18 provided a reliable reference model for comparison. Its original fully connected layer was replaced with a classifier for the nine architectural-form categories.

3.3.2. MobileNetV3-Small

MobileNetV3-Small was included as a lightweight CNN designed for efficient image recognition. It uses depthwise separable convolutions, inverted residual blocks, and lightweight attention modules to reduce computational cost.

The inclusion of MobileNetV3-Small allows this study to evaluate whether a lightweight model can achieve competitive performance in the classification of hand-drawn ancient architectural forms from Chinese local gazetteer maps. This is meaningful for potential applications in digital heritage platforms or mobile-assisted image annotation systems.

3.3.3. EfficientNet-B0

EfficientNet-B0 was selected as a parameter-efficient CNN. Its compound scaling strategy jointly balances network depth, width, and input resolution, rather than enlarging only one dimension of the model.

In this study, EfficientNet-B0 proved especially suitable for the classification task. Its ability to combine local structural details with broader shape information helps distinguish categories with similar appearances, such as bell towers, drum towers, gate towers, and towers.

3.3.4. DenseNet-121

DenseNet-121 was chosen for its dense connectivity. In this architecture, each layer receives features from earlier layers, encouraging feature reuse and improving information flow through the network.

This architecture may be useful for hand-drawn architectural classification because fine-grained visual cues, such as roof outlines, vertical structures, and decorative forms, need to be preserved across different levels of representation. The final classifier of DenseNet-121 was replaced with a new output layer for the target categories.

3.3.5. Vision Transformer

ViT-B/16 was used as the transformer-based comparison model. Instead of applying convolutional filters, it divides an image into fixed-size patches and processes those patches as token sequences. Self-attention is then used to model relations among different image regions.

By including ViT-B/16, the study tests whether global attention-based representation is effective for small-sample hand-drawn architectural image classification. Because transformer models often benefit from large training datasets, their behavior under limited data provides an important point of comparison.

Figure 5 summarizes the five model architectures. All were initialized with ImageNet-pretrained weights and fine-tuned on the gazetteer-derived architectural-form dataset. Their original classification heads were replaced with new layers for the nine target classes.

3.4. Loss Function and Optimization

Cross-entropy loss was used to optimize the multi-class classification task. In the first set of comparative experiments, class-weighted cross-entropy was adopted to reduce the effect of uneven class sizes. The weights were calculated from the number of samples in each class, so categories with fewer training images received higher relative weights.

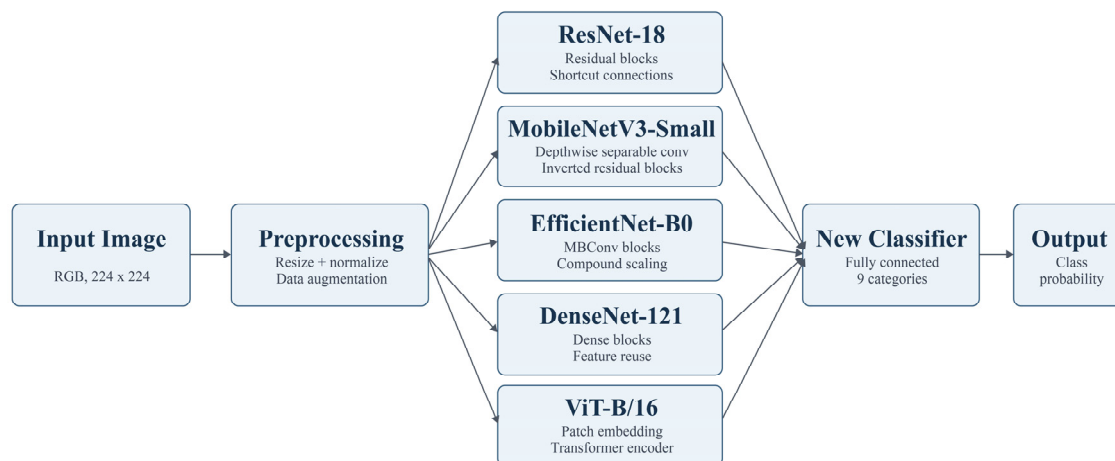


Figure 5. Model architecture overview for hand-drawn ancient architectural form classification.

After the bell tower category was supplemented, the class distribution became more balanced. An additional EfficientNet-B0 tuning experiment was therefore conducted without class weights. The result showed a small increase in macro-average F1, suggesting that weighting strategies should be selected according to the actual distribution of the dataset.

Model training used the AdamW optimizer. Unlike standard Adam, AdamW separates weight decay from gradient-based updates, which can improve regularization. A cosine annealing scheduler gradually adjusted the learning rate during training and helped the optimization process converge more smoothly.

3.5. Hyperparameter Settings

The same core settings were used for all models to make the comparison fair. Images were resized to 224×224 pixels, the batch size was 16, the initial learning rate was 3×10^{-4} , the weight decay was 1×10^{-4} , and the maximum number of epochs was 25.

Early stopping was applied according to validation performance. When the validation result failed to improve for several consecutive epochs, training was stopped and the checkpoint with the best validation performance was retained.

For the final EfficientNet-B0 setting, AdamW was used with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-4} , and class weights were disabled. Mixed-precision training was enabled on the CUDA-compatible GPU to increase computational efficiency.

3.6. Evaluation Metrics

The independent test set was evaluated with accuracy, precision, recall, F1-score, macro-average F1, and weighted-average F1. Accuracy reports the overall proportion of correct predictions, while precision, recall, and F1-score describe performance at the class level.

Because class sizes are unequal, macro-average F1 was treated as an important indicator, as it gives equal weight to each category. Weighted-average F1 was also reported to reflect overall performance while accounting for class frequency. Confusion matrices were used to examine which visually similar architectural categories were confused.

4. Experiments and Results

4.1. Experimental Environment

The experiments were run on a Windows workstation with an NVIDIA GeForce RTX 4060 Laptop GPU. The models were implemented in Python 3.12.7 with PyTorch 2.11.0, CUDA 12.6 support, and torchvision 0.26.0. GPU acceleration and mixed-precision training

were used to improve training efficiency. All experiments were conducted on a computer equipped with an NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB of video memory (NVIDIA Corporation, Santa Clara, CA, USA) under the Microsoft Windows 11 operating system (Microsoft Corporation, Redmond, WA, USA). GPU acceleration was provided by CUDA 12.6 (NVIDIA Corporation, Santa Clara, CA, USA). The models were implemented in Python 3.12.7 using PyTorch 2.11.0 and torchvision 0.26.0. The five network architectures and their ImageNet-pretrained weights were obtained through torchvision. NumPy 2.4.3 and pandas 3.0.2 were used for numerical computation and data organization; Pillow 12.1.1 was used for image loading and preprocessing; scikit-learn 1.8.0 was used to calculate classification metrics, confusion matrices, and classification reports; and Matplotlib 3.10.9 was used to generate the experimental figures. The training configuration used a fixed random seed of 42 to improve experimental reproducibility.

The dataset included 685 hand-drawn architectural images across nine categories. Stratified sampling divided them into 479 training images, 103 validation images, and 103 test images, preserving the class distribution in each subset.

For every model, images were resized to 224×224 pixels and trained with a batch size of 16. AdamW was used with an initial learning rate of 3×10^{-4} and a weight decay of 1×10^{-4} . Training was limited to 25 epochs, with early stopping based on validation results. The same split and evaluation protocol were applied to all architectures.

4.2. Overall Model Performance Comparison

Five representative architectures were trained and tested with the same data split and experimental settings to compare their classification ability. Table 4 reports the results using accuracy, macro-average F1, and weighted-average F1 as the main metrics.

Table 4. Overall performance comparison of five deep learning models on the test set.

Model	Best Epoch	Validation Accuracy	Test Accuracy	Macro-F1	Weighted-F1	Training Time (s)
EfficientNet-B0	15	0.8350	0.8738	0.8660	0.8754	311.10
DenseNet-121	4	0.8252	0.8544	0.8472	0.8546	334.76
MobileNetV3-Small	20	0.8252	0.8447	0.8299	0.8448	212.66
ResNet-18	9	0.8252	0.7767	0.7603	0.7766	61.11
ViT-B/16	24	0.6990	0.6117	0.5684	0.6025	

Table 4 shows that EfficientNet-B0 performed best overall, reaching 0.8738 test accuracy, 0.8660 macro-average F1, and 0.8754 weighted-average F1. DenseNet-121 ranked second with 0.8544 test accuracy and 0.8472 macro-average F1. MobileNetV3-Small also produced competitive results, achieving 0.8447 test accuracy with a shorter training time.

ResNet-18 performed below EfficientNet-B0, DenseNet-121, and MobileNetV3-Small. ViT-B/16 had the weakest result among the five models, with 0.6117 test accuracy and 0.5684 macro-average F1. This pattern suggests that, under the small-sample conditions of this dataset, CNNs are more effective than a transformer model that typically requires larger-scale data.

Overall, the results show that CNN architectures are better matched to the classification of hand-drawn ancient architectural forms in this dataset. EfficientNet-B0 offered the strongest balance between overall accuracy and class-level performance, so it was selected for more detailed analysis.

4.3. Classification Performance of EfficientNet-B0

Since EfficientNet-B0 achieved the best overall performance among the five evaluated models, its class-level classification results were further analyzed. Table 5 lists the precision, recall, F1-score, and support for each classification label.

Table 5. Class-level classification performance of EfficientNet-B0.

Class	Precision	Recall	F1-Score	Support
bell tower	0.7500	0.7500	0.7500	8
buildings inside the city	0.9231	0.8000	0.8571	15
buildings outside the city	0.7857	1.0000	0.8800	11
tower pavilion	1.0000	0.9000	0.9474	10
drum tower	0.7000	0.7778	0.7368	9
gate tower	1.0000	0.8571	0.9231	14
pavilion	0.9000	0.9000	0.9000	10
temple	0.7692	0.8333	0.8000	12
tower	1.0000	1.0000	1.0000	14

Table 5 indicates that EfficientNet-B0 recognized tower, tower pavilion, gate tower, and pavilion categories especially well. The tower class reached perfect test-set performance, with precision, recall, and F1-score all equal to 1.0000. Tower pavilion and gate tower also achieved high F1-scores of 0.9474 and 0.9231. These results suggest that EfficientNet-B0 can learn distinctive structural cues from hand-drawn architectural forms.

Figure 6 also shows weaker F1-scores for bell tower and drum tower, at 0.7500 and 0.7368. This lower performance is likely related to the similarity of tower-like forms, particularly when texture and color are limited. The temple class reached a moderate F1-score of 0.8000, indicating that some temple samples share visual features with other building categories.

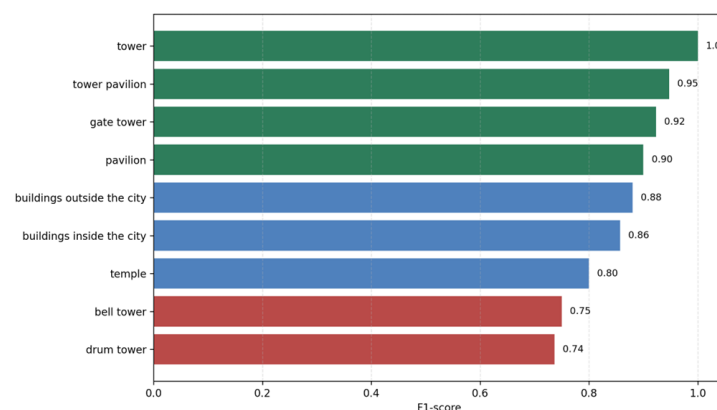


Figure 6. Class-wise F1-scores of EfficientNet-B0.

4.4. Confusion Matrix Analysis

A confusion matrix was generated from the test set to examine EfficientNet-B0 more closely. It provides a direct view of which classes were recognized correctly and which architectural forms were more often confused.

4.5. Class-Level Performance Analysis

As shown in Figure 7, EfficientNet-B0 achieved strong recognition performance for several categories. The tower category was completely correctly classified, with all 14 test samples predicted accurately. The buildings outside the city category also achieved perfect recall, with all 11 samples correctly recognized. In addition, tower pavilion, gate tower, and

pavilion showed high classification stability, indicating that the model was able to capture their distinctive visual structures.

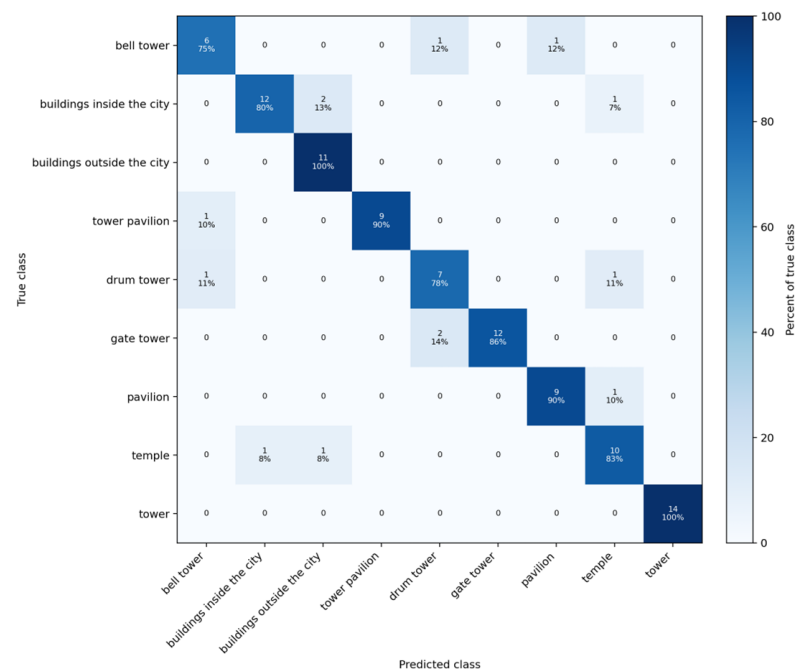


Figure 7. Confusion matrix of EfficientNet-B0.

However, several misclassification patterns were observed, as summarized in Table 6. The most evident confusion occurred between gate tower and drum tower, where two gate tower samples were incorrectly classified as drum tower. This may be because both categories often contain elevated building structures and similar tower-like outlines in hand-drawn representations. Another notable confusion occurred between buildings inside the city and buildings outside the city, where two samples of buildings inside the city were predicted as buildings outside the city. This suggests that scene-context labels may contain overlapping architectural components and contextual features.

Table 6. Main misclassification pairs of EfficientNet-B0.

True Class	Predicted Class	Number of Samples	Possible Reason
gate tower	drum tower	2	Similar tower-like structure and elevated architectural form
buildings inside the city	buildings outside the city	2	Overlapping scene-context architectural forms
bell tower	drum tower	1	Similar vertical tower form
bell tower	pavilion	1	Similar roof outline in hand-drawn style
tower pavilion	bell tower	1	Ambiguous structural contour
drum tower	bell tower	1	Similar tower-type architectural morphology
drum tower	temple	1	Shared roof and facade characteristics
pavilion	temple	1	Similar roof structure and traditional architectural style
temple	buildings inside the city	1	Contextual similarity with building groups
temple	buildings outside the city	1	Similar scene-level representation

Overall, the confusion matrix indicates that errors mainly occur between categories with similar morphology or overlapping components. This pattern matches the nature of hand-drawn ancient Chinese architectural images, where texture and color are sparse and classification relies on contour, roof shape, building proportion, and spatial layout. The errors therefore reflect both model limitations and the inherent visual closeness of some architectural types.

4.6. Hyperparameter Tuning Results

EfficientNet-B0 was selected for additional tuning after the model comparison because it produced the best initial result. The tuning tests included changing the learning rate, modifying input size, applying label smoothing, increasing weight decay, and removing class weights. The aim was to determine whether its performance could be improved under small-sample conditions.

Table 7 presents the tuning results. The initial EfficientNet-B0 model achieved 0.8738 test accuracy and 0.8660 macro-average F1. The best tuned setting removed class weights while keeping the learning rate at 3×10^{-4} and weight decay at 1×10^{-4} , maintaining 0.8738 test accuracy and increasing macro-average F1 to 0.8681.

Table 7. Hyperparameter tuning results of EfficientNet-B0.

Experiment	Best Epoch	Validation Accuracy	Test Accuracy	Macro-F1	Weighted-F1	Training Time (s)
No class weights, lr = 3×10^{-4}	12	0.8350	0.8738	0.8681	0.8746	242.39
Baseline with class weights	15	0.8350	0.8738	0.8660	0.8754	311.10
Image size 256, lr = 3×10^{-4}	4	0.8350	0.8738	0.8637	0.8756	218.30
No class weights, lr = 2×10^{-4}	12	0.8350	0.8738	0.8635	0.8743	274.87
Label smoothing = 0.02, weight decay = 3×10^{-4}	7	0.8447	0.8738	0.8611	0.8750	222.27
Label smoothing = 0.05, lr = 1×10^{-4}	13	0.8544	0.8544	0.8500	0.8560	

As shown in Figure 8 and Table 7, disabling class weights produced the highest macro-average F1 score. This result indicates that after additional samples were added to the bell tower category, the dataset became less imbalanced. Under this condition, the use of class weights may have introduced slight over-compensation for minority classes, while the unweighted loss function achieved a more balanced classification performance.

Increasing the input image size to 256 did not improve the macro-average F1 score, suggesting that the original 224×224 input size was sufficient for the present task. Similarly, label smoothing did not improve the final test performance. Although the setting with label smoothing of 0.05 achieved the highest validation accuracy, its test performance decreased, indicating weaker generalization on the independent test set.

Figure 9 presents the learning curves of the tuned EfficientNet-B0 model. The training loss decreased steadily, while the training accuracy increased rapidly, indicating that the model successfully learned discriminative visual features from the training data. The validation accuracy improved during the early epochs and reached its best performance around epoch 12. After that, the validation performance became relatively stable, while the

training performance continued to improve. This pattern suggests that early stopping was necessary to reduce the risk of overfitting.

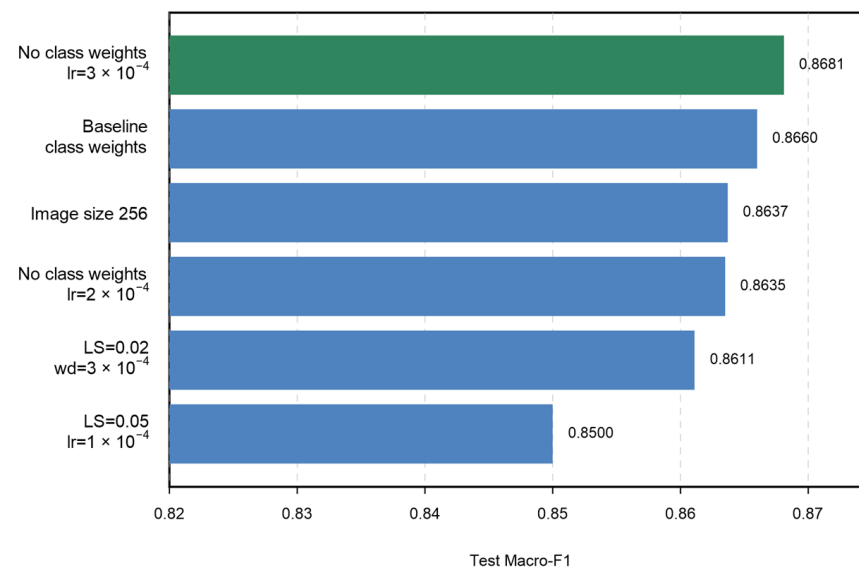


Figure 8. Macro-F1 comparison of EfficientNet-B0 tuning experiments.

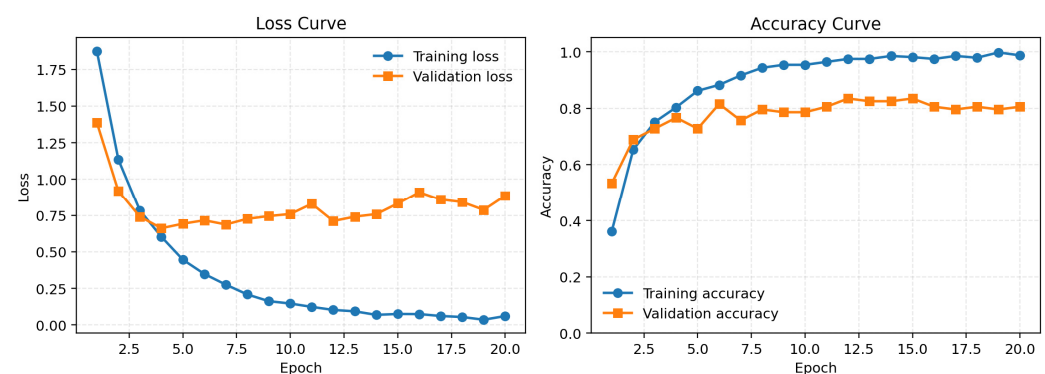


Figure 9. Learning curves of the tuned EfficientNet-B0 model.

The tuning experiment slightly increased EfficientNet-B0 macro-average F1 from 0.8660 to 0.8681. Although the improvement was modest, it demonstrates that the loss-weighting strategy should be adjusted according to the actual class distribution. Therefore, the final EfficientNet-B0 model without class weights was selected as the optimized model for this classification task.

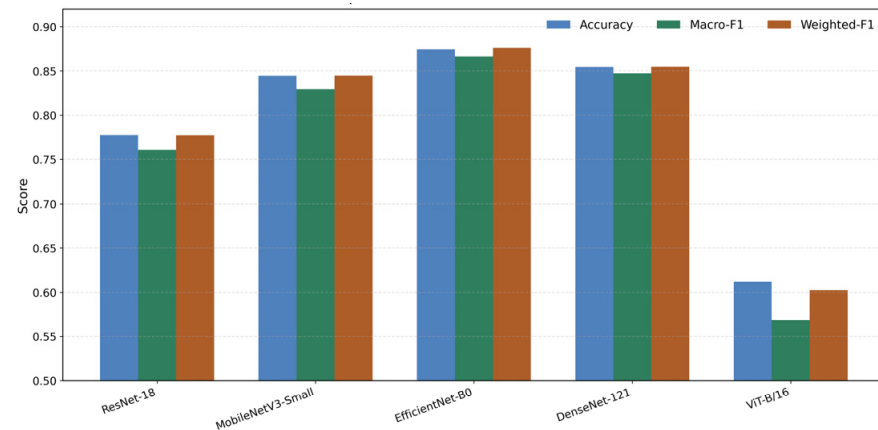
4.7. Comparison Between CNN-Based Models and Vision Transformer

To examine the effect of architecture, the four CNN models were compared with ViT-B/16. The CNN group included ResNet-18, MobileNetV3-Small, EfficientNet-B0, and DenseNet-121, while ViT-B/16 represented transformer-based visual modeling.

Table 8 and Figure 10 show that the CNN models clearly exceeded ViT-B/16 in this classification task. The average macro-F1 of the CNN group was 0.8259, while ViT-B/16 reached only 0.5684. EfficientNet-B0 achieved the highest macro-F1 score of 0.8660, whereas ViT-B/16 ranked last in accuracy, macro-F1, and weighted-F1.

Table 8. Performance comparison between CNN-based models and Vision Transformer.

Model Group	Number of Models	Mean Test Accuracy	Mean Macro-F1	Best Macro-F1
CNN-based models	4	0.8374	0.8259	0.8660
Vision Transformer	1	0.6117	0.5684	0.5684

**Figure 10.** Performance comparison between CNN-based models and ViT-B/16.

This result indicates that CNN-based architectures are more suitable for the current small-sample dataset. Hand-drawn architectural images rely heavily on local structural patterns, such as roof outlines, tower bodies, facade contours, and spatial proportions. CNNs are naturally effective at extracting such local and hierarchical visual features through convolutional kernels. In contrast, ViT-B/16 models an image as a sequence of patches and depends more strongly on large-scale training data to learn stable visual representations. Under limited data conditions, the transformer-based model may not fully capture the fine-grained architectural differences among categories.

Figure 11 compares training time with macro-F1. Although ViT-B/16 was not the slowest model, its performance was much lower than that of the CNNs, indicating that the gap was mainly related to model suitability rather than computation time. This relatively weak performance should be considered in light of the image characteristics of the dataset, rather than as evidence of a general limitation of Transformer-based models. The gazetteer samples are hand-drawn line images with sparse texture. Differences between categories are often conveyed by continuous rooflines, eave profiles, tower bodies, openings, and overall silhouettes, rather than by rich color or surface texture. When such images are divided into fixed-size patches, continuous architectural contours may be split across different patches. With limited training data, ViT-B/16 may therefore have difficulty learning stable relationships among these subtle contour-based features. CNN-based models are more effective in capturing local edges, contours, and hierarchical shape features, which helps explain the stronger performance of EfficientNet-B0 in this study.

Taken together, the experiments show that CNN models are more robust than ViT-B/16 for small-sample recognition of hand-drawn ancient architectural forms. EfficientNet-B0 produced the best overall result, suggesting that a balanced CNN can better represent structural features when training data are limited.

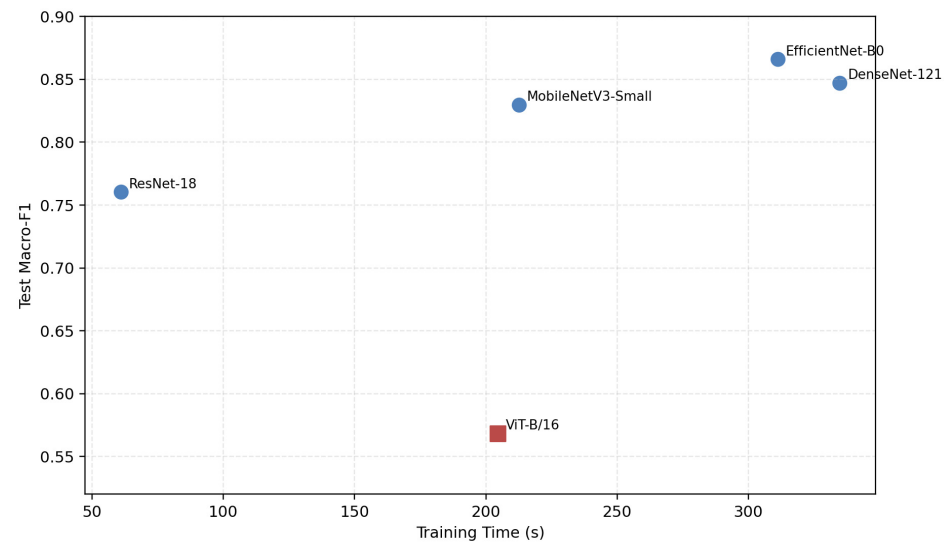


Figure 11. Training time and Macro-F1 comparison of CNN-based models and ViT-B/16.

4.8. Multi-Level Feature Representation Analysis

To further illustrate the feature extraction process of EfficientNet-B0, intermediate feature maps were visualized using several representative test images. Four correctly classified samples, including bell tower, gate tower, pavilion, and tower, were selected for visualization. Feature maps were extracted from different stages of the network, including the stem layer, early MBConv block, middle MBConv block, and final convolutional layer.

As shown in Figure 12, the shallow layers mainly respond to low-level visual patterns, such as edges, contours, and local line structures. These features are important for hand-drawn architectural images because the visual information is largely represented by outlines rather than rich color or texture.

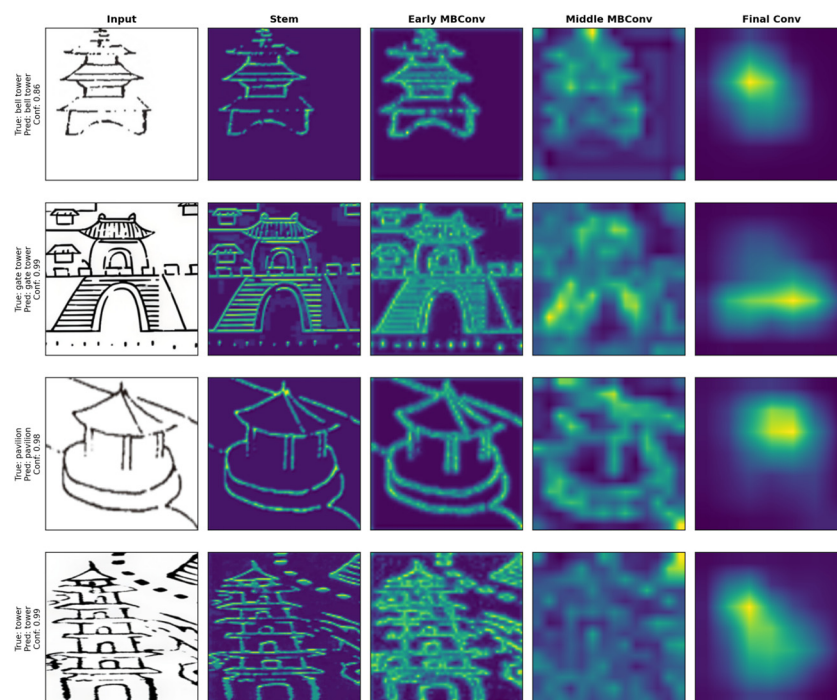


Figure 12. Feature activation maps extracted from selected EfficientNet-B0 stages for representative architectural samples.

In the early and middle MBConv layers, the activation maps become more concentrated on architectural components, including roof outlines, tower bodies, and building contours. This indicates that EfficientNet-B0 gradually transforms low-level line features into more meaningful structural representations.

In the final convolutional layer, the feature responses become more abstract and spatially focused. The model tends to emphasize the most discriminative regions of each ancient architectural form, such as the vertical tower structure in tower and gate tower, or the roof and body structure in pavilion. This visualization supports the quantitative results by showing that EfficientNet-B0 can capture hierarchical structural features from hand-drawn ancient architectural forms.

Overall, the intermediate feature visualization provides an intuitive explanation of how EfficientNet-B0 extracts visual information from architectural drawings. The model first captures basic contour information and then gradually focuses on higher-level architectural morphology, which helps explain its strong performance in the classification task.

5. Discussion

5.1. Why EfficientNet-B0 Performs Better

The results show that EfficientNet-B0 produced the strongest performance among the five models. It achieved the highest test accuracy and macro-average F1, indicating good performance in both overall prediction and class-level balance. This advantage can be understood in relation to model design, feature extraction, and the small-sample nature of the dataset.

First, EfficientNet-B0 adopts a compound scaling strategy that balances network depth, width, and input resolution. Compared with models that simply increase depth or parameter size, EfficientNet improves model capacity in a more balanced way. This design allows the network to extract richer visual features while maintaining relatively high computational efficiency. For hand-drawn ancient architectural forms, the classification depends heavily on visual cues such as roof shapes, tower silhouettes, facade structures, and spatial proportions. EfficientNet-B0 can capture these features effectively without requiring an excessively large model.

Second, EfficientNet-B0 provides a strong balance between local detail extraction and global structural representation. Hand-drawn architectural images often contain limited color and texture information, so the model must rely more on contours and structural patterns. EfficientNet-B0 can learn features from edges and lines through to higher-level building morphology, making it useful for distinguishing categories such as tower, tower pavilion, gate tower, and pavilion.

Third, EfficientNet-B0 is more appropriate for the small-sample condition of this study. Compared with larger or transformer-based models, EfficientNet-B0 has a more efficient architecture and a lower risk of overfitting when fine-tuned on a limited dataset. This may explain why it outperformed ViT-B/16, which generally requires larger-scale data to learn stable visual representations. The results suggest that, for small-scale classification of hand-drawn ancient architectural forms, a compact and well-balanced CNN architecture can be more effective than a model with stronger theoretical representation capacity but higher data demand. Table 9 summarizes these model-level differences by relating each architecture to its main model characteristic, experimental observation, and possible explanation.

Table 9. Model characteristics and performance interpretation of the evaluated architectures.

Model	Main Model Characteristic	Experimental Observation	Possible Explanation
ResNet-18	Residual learning	Moderate performance	Stable baseline but limited feature richness
MobileNetV3-Small	Lightweight convolution	Competitive performance	Efficient local feature extraction
EfficientNet-B0	Compound scaling	Best performance	Balanced depth, width, and resolution
DenseNet-121	Dense feature reuse	Second-best performance	Strong feature propagation
ViT-B/16	Patch-based self-attention	Lowest performance	Requires larger-scale data

This finding is consistent with previous heritage-recognition studies showing that transfer learning can be effective when annotated heritage images are limited [7,9,51,52]. However, the present task differs from studies based on photographs, remote-sensing images, point clouds, or three-dimensional heritage data [53–65]. In those cases, models can often use color, material texture, lighting, depth, or spatial geometry. By contrast, gazetteer-derived architectural images provide mainly sparse contours and simplified linework. The stronger performance of EfficientNet-B0 therefore suggests that compact CNNs are better suited to extracting local roof edges, eave lines, tower-body contours, and hierarchical shape features from this type of image material.

Finally, the hyperparameter tuning results further support the robustness of EfficientNet-B0. After the bell tower category was supplemented with additional images, disabling class weights slightly improved the macro-average F1 score. This indicates that EfficientNet-B0 can benefit from a more balanced training distribution and does not rely heavily on class weighting when the category imbalance is reduced. Overall, the tuning results further demonstrate the suitability of EfficientNet-B0 for recognizing hand-drawn ancient architectural forms under small-sample conditions.

5.2. Influence of Small-Sample and Class-Imbalanced Data

The dataset used in this study was characterized by both a limited sample size and an uneven distribution across classification labels. Although the total number of images reached 685 after data supplementation, the number of samples in each class was still relatively small compared with large-scale natural-image datasets. This small-sample condition increases the difficulty of training deep learning models, as the models must learn discriminative architectural features from limited visual examples.

Class imbalance is another important factor affecting classification performance. As shown in Table 10, the number of images varied across categories. The largest category, buildings inside the city, contained 101 images, while the smallest category, bell tower, contained 51 images after supplementation. Although the imbalance was not extremely severe, the uneven distribution still influenced the stability of class-level performance, especially for categories with fewer test samples.

The influence of class imbalance was reflected in the class-level results. For example, categories such as tower, tower pavilion, and gate tower achieved relatively high F1-scores, while bell tower and drum tower showed lower performance in the original EfficientNet-B0 experiment. This suggests that categories with fewer samples or greater morphological similarity may be more difficult for the model to distinguish.

The hyperparameter tuning results also demonstrate the importance of considering the actual data distribution. In the initial model comparison, class-weighted cross-entropy loss was used to reduce the influence of class imbalance. However, after additional images

were added to the bell tower category, disabling class weights slightly improved the macro-average F1 score of EfficientNet-B0. This indicates that class-weighting strategies should not be applied mechanically; instead, they should be adjusted according to the degree of imbalance in the dataset.

Table 10. Class distribution after data supplementation for EfficientNet-B0 tuning.

Class	Total Images	Training Set	Validation Set	Test Set
bell tower	51	36	7	8
buildings inside the city	101	71	15	15
buildings outside the city	70	49	10	11
tower pavilion	68	47	11	10
drum tower	64	45	10	9
gate tower	91	64	13	14
pavilion	70	49	11	10
temple	78	54	12	12
tower	92	64	14	14

Overall, the results suggest that small sample size and uneven class distribution remain important challenges in the classification of hand-drawn ancient architectural forms. Future work should continue to expand the dataset, especially for visually similar and relatively underrepresented categories, to improve the robustness and generalization ability of the classification model.

5.3. Misclassification Analysis from Architectural Morphology

The misclassification results indicate that most errors occurred between categories with similar architectural forms or overlapping visual characteristics. Since the images used in this study are hand-drawn ancient architectural forms, the model mainly relies on contours, roof shapes, vertical proportions, and spatial structures rather than rich color or texture information. Therefore, categories with similar silhouettes are more likely to be confused.

This error pattern is different from many photographic architectural heritage classification tasks, where misclassification may be strongly affected by viewpoint, illumination, material texture, weathering, or occlusion. In the present dataset, the main difficulty comes from the schematic nature of gazetteer drawings: different categories may share similar roof profiles, eave tiers, tower-like bodies, and simplified outlines. The misclassification results therefore reflect not only model limitations, but also the morphological overlap inherent in the historical visual material itself.

As shown in Figure 13, the most frequent misclassification pairs were gate tower samples misclassified as drum tower and buildings inside the city samples misclassified as buildings outside the city, with two misclassified samples in each pair. The confusion between gate tower and drum tower can be explained by their similar elevated structures, roof forms, and tower-like vertical composition. In hand-drawn images, these architectural features may be simplified or stylized, making the boundary between the two categories less distinct.

The confusion between buildings inside the city and buildings outside the city reflects another type of classification difficulty. Unlike architectural-type labels, these two categories are scene-context labels and may contain multiple buildings, walls, roofs, or spatial layouts. When contextual details are limited, the model may focus on shared building components rather than the broader spatial setting, resulting in misclassification.

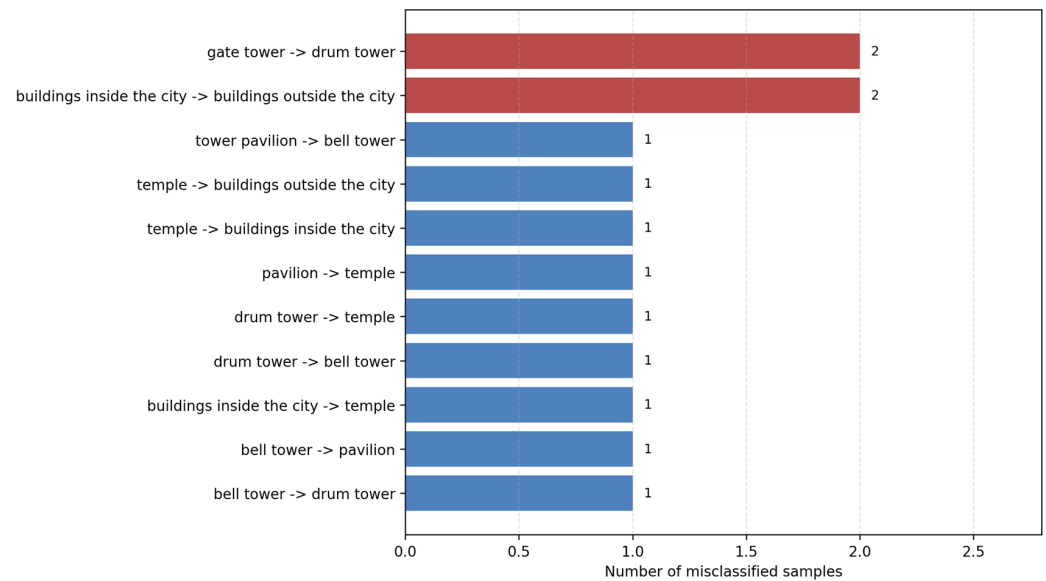


Figure 13. Major misclassification pairs of EfficientNet-B0.

To further illustrate the classification difficulty revealed by the confusion matrix, Figure 14 presents representative difficult samples from visually similar and context-dependent categories. The bell tower and drum tower samples show similar tower-like forms, stacked eaves, and arched passageway bases, which makes them difficult to distinguish in simplified line drawings. The buildings inside the city and buildings outside the city examples show that these scene-context labels depend partly on the city wall, settlement boundary, and relative spatial position in the original gazetteer map. After cropping, such contextual cues may be weakened, increasing the possibility of confusion.

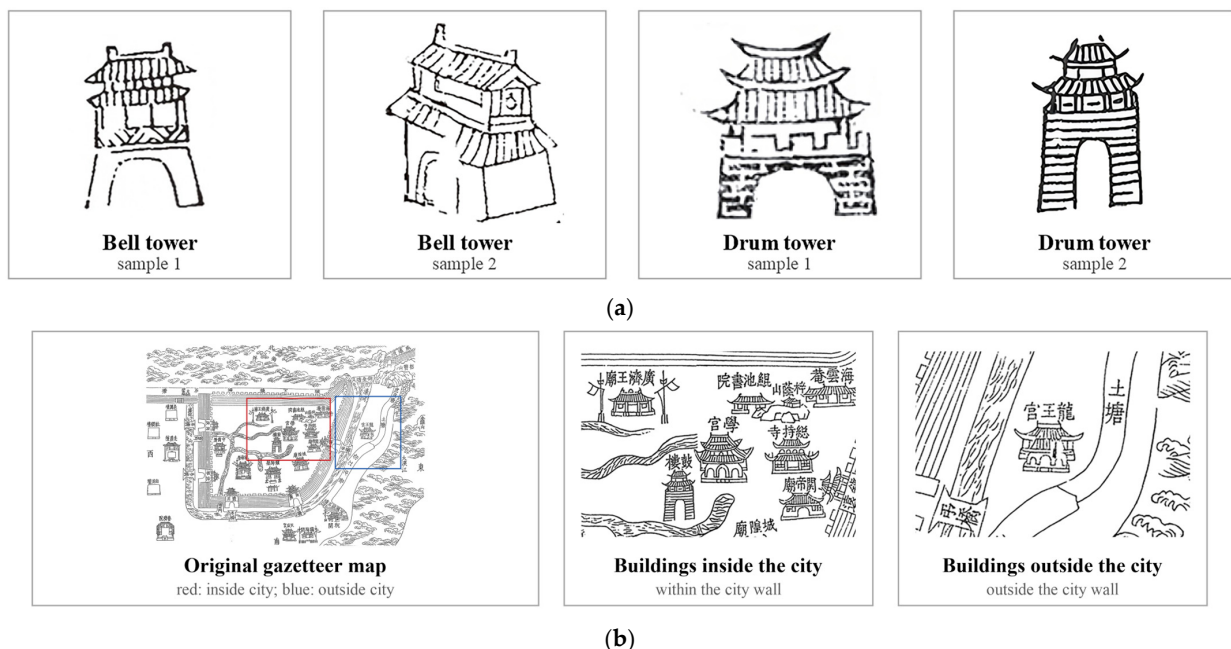


Figure 14. Representative difficult samples from visually similar and context-dependent categories: (a) bell tower and drum tower samples with similar tower-like forms; (b) buildings inside and outside the city, showing that the distinction depends partly on city-wall context and relative spatial position in the original gazetteer map. The Chinese characters retained in the gazetteer map fragments are original historical labels, mainly indicating place names, building names, and directions; the relevant category information is provided in English below each image.

Other errors, such as misclassifying bell towers as drum towers, or vice versa, and misclassifying pavilions as temples, also indicate that traditional architectural categories may share common morphological features. For example, bell towers and drum towers often have comparable vertical forms, while pavilions and temples may share similar roof structures and facade patterns. These results show that the classification difficulty is not only caused by model limitations but also by the inherent visual similarity among ancient Chinese architectural forms.

Overall, the misclassification analysis demonstrates that architectural morphology plays an important role in model performance. Future work could incorporate more fine-grained annotations, such as roof type, number of stories, tower body shape, or spatial context, to help the model distinguish visually similar categories more effectively.

5.4. Practical Implications for Digital Heritage Classification

The findings have practical value for organizing and managing digital resources related to ancient Chinese architecture, especially hand-drawn architectural images preserved in local gazetteers. The proposed deep learning workflow is intended to support semi-automatic indexing, retrieval, and expert-assisted annotation of gazetteer-derived architectural images. It should not be understood as a tool for directly validating, reconstructing, or authenticating individual surviving buildings. Rather, its value lies in generating preliminary architectural-form labels that can assist archival organization, expert review, and subsequent historical interpretation. In this sense, the model serves as an auxiliary tool for digital heritage image management rather than as independent evidence for verifying actual buildings.

As shown in Figure 15, the trained classification model can be integrated into a digital heritage workflow. First, hand-drawn architectural images are collected from historical or archival sources. After preprocessing, the images can be automatically classified by the trained EfficientNet-B0 model. The predicted categories can then be reviewed by domain experts, especially for ambiguous cases or categories with similar architectural morphology. Finally, the corrected labels can be used for digital archive indexing, retrieval, and subsequent research.

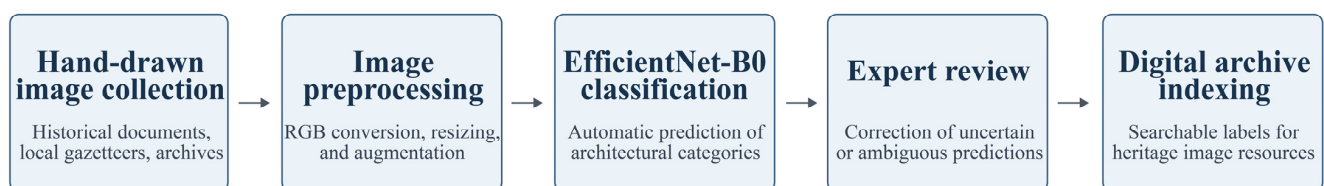


Figure 15. Potential workflow for digital heritage image classification.

This workflow can improve the efficiency of architectural image organization. Instead of manually inspecting all images from the beginning, researchers can use the model to generate preliminary labels and focus their attention on uncertain or misclassified samples. This human-in-the-loop strategy combines the efficiency of deep learning with the reliability of expert knowledge.

In addition, automatic classification can support large-scale comparative studies of ancient Chinese architecture. Once hand-drawn ancient architectural forms are systematically labeled, researchers can more easily analyze the distribution of architectural types, regional stylistic differences, and visual patterns across historical documents. Therefore, the proposed method is not only a technical classification tool but also a potential supporting method for digital humanities, architectural history, and cultural heritage preservation. In this sense, by reducing the manual burden of classifying historical architectural im-

ages, the proposed workflow supports the sustainable reuse, long-term organization, and knowledge-based management of local gazetteer visual archives.

5.5. Grad-CAM-Based Interpretability Analysis

Grad-CAM was used to interpret how EfficientNet-B0 made classification decisions. This method produces class-discriminative heatmaps from gradients in the final convolutional layer, showing which image regions contribute most to the predicted category.

As shown in Figure 16, the Grad-CAM results indicate that EfficientNet-B0 mainly focused on structurally meaningful regions of the hand-drawn ancient architectural forms. In the Grad-CAM visualizations, warmer colors indicate regions that make stronger positive contributions to the predicted class. Red and yellow regions represent relatively high model attention, green regions indicate intermediate contributions, and blue regions correspond to low or near-zero positive contributions. Because each activation map was independently normalized to the range of 0–1, the colors represent relative importance within each image and should not be interpreted as prediction probabilities or directly compared across different samples. For the tower and gate tower samples, the highlighted regions were concentrated around the vertical body and upper architectural structure. For the pavilion sample, the model paid attention to the roof and main building body. For the bell tower sample, the heatmap also covered the central tower-like structure.

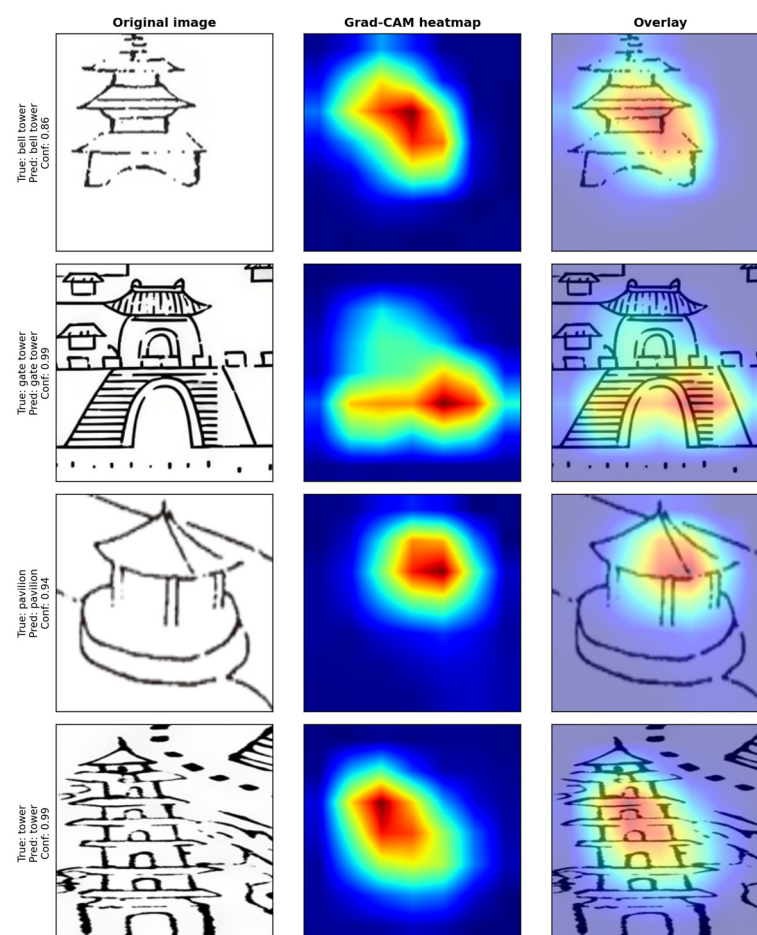


Figure 16. Grad-CAM visualization of EfficientNet-B0 on representative test samples. Warmer colors indicate stronger positive contributions to the predicted class, whereas cooler colors represent lower contributions. Each heatmap is independently normalized.

These visual results suggest that the model did not rely only on background information or irrelevant blank regions. Instead, it learned to focus on key architectural components and morphological cues, such as roof profiles, stacked eave tiers, tower bodies, hall bodies, facade outlines, arched passageway bases, platform bases, and overall structural silhouettes. These regions correspond to the architectural-form cues summarized earlier in Figure 3, indicating that the model responses are related to visually recognizable architectural features in gazetteer sketches rather than to arbitrary line patterns. This is consistent with the quantitative results, where EfficientNet-B0 achieved strong classification performance for categories with distinctive morphological features.

To examine the confusion between bell tower and drum tower in more detail, Grad-CAM was applied to representative confused samples. As shown in Figure 17, the heatmaps for both categories are concentrated on comparable architectural parts, including stacked eaves, upper tower bodies, arched passageway bases, and tower-like silhouettes. These areas are meaningful for architectural recognition, but in simplified gazetteer line drawings they are not exclusive to either category.

This helps explain why the model may confuse bell tower and drum tower even when its attention falls on relevant architectural regions. The error is therefore not simply caused by a failure to locate the building, but by the overlap of visually discriminative parts between the two categories. This reading is consistent with the earlier confusion-matrix and difficult-sample analyses.

Taken together, the representative and confused-sample Grad-CAM results suggest that the remaining errors are not mainly driven by irrelevant background regions. Rather, they arise from shared architectural features, such as eave tiers, tower bodies, arched openings, and similar silhouettes. In other words, the model can locate meaningful architectural areas, but distinguishing subtle morphological differences among visually similar categories remains difficult.

Overall, the Grad-CAM analysis improves the interpretability of the proposed classification framework. It shows that EfficientNet-B0 can identify meaningful architectural regions in hand-drawn gazetteer images, while also revealing why visually similar categories remain difficult to separate. This supports the use of the method for digital heritage image classification and expert-assisted annotation, but also indicates the need for more fine-grained architectural cues in future work.

5.6. Limitations of the Present Study

This study still has several limitations. First, although the dataset was supplemented during the experiment, the total number of images remained relatively small. This may limit the generalization ability of the trained models, especially when applied to hand-drawn architectural images from different sources or drawing styles.

Second, this study classifies gazetteer sketches as schematic and stylized historical visual representations rather than verifying their direct correspondence with actual buildings. Although these images preserve meaningful information about architectural types and morphological features, a predicted category should not be interpreted as evidence for the identity, preservation status, or original form of a specific extant building. Establishing systematic correspondence between gazetteer sketches and actual buildings requires independent architectural–historical verification based on field photographs, measured drawings, textual records, and archival materials.

Third, some categories share similar architectural forms, such as bell tower, drum tower, gate tower, and tower. These morphological similarities make classification difficult and may lead to unstable performance for certain classes. More fine-grained annotations,

such as roof type, number of stories, and structural components, could help improve future classification results.

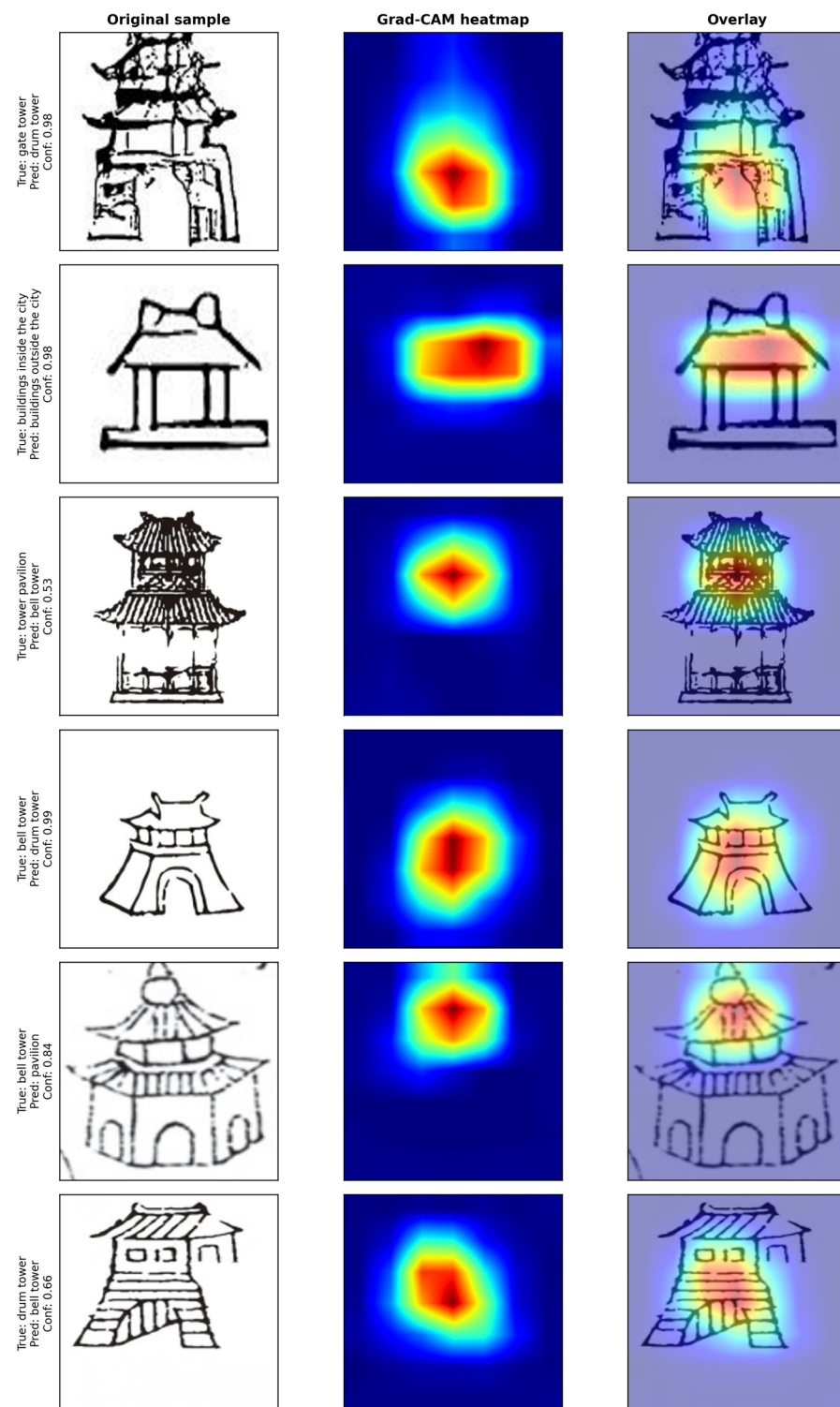


Figure 17. Grad-CAM visualization of misclassified samples.

Fourth, the mixed label structure of the dataset also introduces a limitation. While most categories are architectural-type labels, buildings inside the city and buildings outside the city are scene-context labels. These labels are useful for organizing local gazetteer map images because they preserve information about the spatial position of buildings in the original map. However, they are less visually stable than architectural-type labels after

object-level cropping. Once surrounding cues such as city walls, settlement boundaries, mountain-river environments, or spatial position are removed, the distinction between inside-city and outside-city buildings becomes less obvious. Future work could adopt hierarchical labeling or multi-scale inputs that combine cropped architectural elements with wider map context.

Fifth, although Grad-CAM was used for visual interpretation, the interpretability analysis remained qualitative and was limited to representative samples. For studies related to architectural heritage and building-form recognition, interpretability is important because model attention should correspond to architecturally meaningful features, such as roof profiles, tower bodies, facade structures, and overall silhouettes. Future work should include more systematic interpretability evaluation and attention-based comparisons to better assess whether the models identify features that are meaningful for architectural analysis.

Finally, the dataset was evaluated using an internal train–validation–test split. Future work should test the proposed method on independent external datasets from different regions, periods, and gazetteer editions to further verify its robustness and practical applicability in historical building image archives.

6. Conclusions

This study investigated the deep learning-based classification of hand-drawn ancient architectural forms extracted from Chinese local gazetteer maps under small-sample conditions. A dataset containing 685 images from nine classification labels, including architectural-type and scene-context labels, was constructed to represent historical building forms such as towers, pavilions, temples, gate towers, and buildings inside and outside city walls. Five representative deep learning models, including ResNet-18, MobileNetV3-Small, EfficientNet-B0, DenseNet-121, and ViT-B/16, were evaluated through transfer learning under the same dataset split and experimental settings.

The experiments show that CNN architectures were more effective than the Vision Transformer for this task. EfficientNet-B0 performed best, reaching 0.8738 test accuracy and 0.8660 macro-average F1 in the initial comparison. After tuning, removing class weights increased the macro-average F1 to 0.8681. These results indicate that EfficientNet-B0 can extract useful structural cues from low-texture, hand-drawn building images under limited data conditions.

The confusion matrix and class-level analysis showed that most classification errors occurred between categories with similar architectural morphology, such as gate tower and drum tower, or between scene-level categories such as buildings inside the city and buildings outside the city. This suggests that the difficulty of this task is related not only to the small dataset size but also to the inherent visual similarity among traditional building forms, including roof profiles, vertical proportions, facade structures, and overall silhouettes.

The Grad-CAM analysis further shows that EfficientNet-B0 attends to architecturally meaningful regions, such as roof profiles, tower bodies, eave tiers, arched openings, and overall silhouettes, which strengthens the interpretability of the classification results.

Overall, this study demonstrates the feasibility of applying transfer learning to the automatic classification of hand-drawn ancient architectural forms extracted from Chinese local gazetteer maps. The proposed approach can provide technical support for digital architectural heritage documentation, historical building image organization, and architectural information retrieval. However, the proposed method should be regarded as an auxiliary tool for organizing, indexing, and retrieving hand-drawn gazetteer images, and it does not by itself verify the correspondence between gazetteer sketches and specific extant buildings. Future research should expand the dataset, introduce more fine-grained architectural annotations, apply more systematic model interpretability methods, and validate

the proposed framework on independent external datasets to improve its applicability to broader building heritage image archives.

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

Funding: This research was funded by the Youth Scientific Research Project of the Basic Research Program of Shanxi Province, grant number 202303021212062; the 2025 Annual Funding Project of the Shanxi Provincial Cultural Relics Protection Fund, grant number sxswbjj020; and the Humanities and Social Sciences Research Planning Fund of the Ministry of Education of China, grant number 23YJAZH058.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: Author Lan Li is a postdoctoral researcher at the Postdoctoral Research Workstation of Shanxi Construction Investment Group Co., Ltd. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. UNESCO. *Recommendation Concerning the Preservation of, and Access to, Documentary Heritage Including in Digital Form*; UNESCO: Paris, France, 2015. Available online: <https://www.unesco.org/en/legal-affairs/recommendation-concerning-preservation-and-access-documentary-heritage-including-digital-form> (accessed on 15 June 2026).
2. Remondino, F. Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning. *Remote Sens.* **2011**, *3*, 1104–1138. [CrossRef]
3. Llamas, J.; Lerones, P.M.; Medina, R.; Zalama, E.; Gómez-García-Bermejo, J. Classification of Architectural Heritage Images Using Deep Learning Techniques. *Appl. Sci.* **2017**, *7*, 992. [CrossRef]
4. Janković, R. Machine Learning Models for Cultural Heritage Image Classification: Comparison Based on Attribute Selection. *Information* **2020**, *11*, 12. [CrossRef]
5. Cintas, C.; Lucena, M.; Fuertes, J.M.; Delrieux, C.; Navarro, P.; González-José, R.; Molinos, M. Automatic Feature Extraction and Classification of Iberian Ceramics Based on Deep Convolutional Networks. *J. Cult. Herit.* **2020**, *41*, 106–112. [CrossRef]
6. Li, R.; Geng, G.; Wang, X.; Qin, Y.; Liu, Y.; Zhou, P.; Zhang, H. LBCapsNet: A Lightweight Balanced Capsule Framework for Image Classification of Porcelain Fragments. *Herit. Sci.* **2024**, *12*, 133. [CrossRef]
7. Sasithradevi, A.; Sabarinathan, S.; Chanthini, B.; Subbulakshmi, T.; Prakash, P. MonuNet: A High Performance Deep Learning Network for Kolkata Heritage Image Classification. *Herit. Sci.* **2024**, *12*, 242. [CrossRef]
8. Mishra, M.; Lourenço, P.B. Artificial Intelligence-Assisted Visual Inspection for Cultural Heritage: State-of-the-Art Review. *J. Cult. Herit.* **2024**, *66*, 536–550. [CrossRef]
9. Ottoni, A.L.C.; Ottoni, L.T.C. A Deep Learning Approach for Cultural Heritage Building Classification Using Transfer Learning and Data Augmentation. *J. Cult. Herit.* **2025**, *74*, 214–224. [CrossRef]
10. Huang, Z. Machine Learning in the Documentation and Analysis of Historical Landmark Buildings: A Systematic Review and Practical Demonstration. *Built Herit.* **2026**, *10*, 14. [CrossRef]
11. Wang, S. *Historical Atlas of Chinese Urban Human Settlements*; Science Press: Beijing, China, 2015. (In Chinese)
12. Wang, S. Research on the construction of the academic system of Chinese indigenous urban planning. *City Plan. Rev.* **2023**, *47*, 4–9. (In Chinese)
13. Wang, S.; Li, X.; Yan, S. Traditional Chinese planning and design methods combined with large-scale landscape environments. *Chin. Sci. Bull.* **2016**, *61*, 3564–3571. (In Chinese) [CrossRef]
14. Wang, S. *Research on the Construction of Human Settlements in Historical Cities Along the Shanxi–Shaanxi Yellow River Coast*; China Architecture & Building Press: Beijing, China, 2009. (In Chinese)
15. Chen, S.-P.; Hammond, K.J.; Gerritsen, A.; Wu, S.; Zhang, J. Local Gazetteers Research Tools: Overview and Research Application. *J. Chin. Hist.* **2020**, *4*, 544–558. [CrossRef]

16. Chen, S.-P.; Yeh, C.; Wang, S.; Che, Q. Treating a Genre as a Database: A Digital Research Methodology for Studying Chinese Local Gazetteers. *Int. J. Digit. Humanit.* **2023**, *4*, 171–193. [\[CrossRef\]](#)
17. He, L.; Chen, M.; Lu, G.; Lin, H.; Liu, L.; Xu, J. An Approach to Transform Chinese Historical Books into Scenario-Based Historical Maps. *Cartogr. J.* **2013**, *50*, 49–65. [\[CrossRef\]](#)
18. Chiang, T.-C. Historical Geography in China. *Prog. Hum. Geogr.* **2005**, *29*, 148–164. [\[CrossRef\]](#)
19. Petitpierre, R.; Uhl, J.H.; di Lenardo, I.; Kaplan, F. A Fragment-Based Approach for Computing the Long-Term Visual Evolution of Historical Maps. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 363. [\[CrossRef\]](#)
20. Wu, S.; Schindler, K.; Heitzler, M.; Hurni, L. Domain Adaptation in Segmenting Historical Maps: A Weakly Supervised Approach through Spatial Co-Occurrence. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 199–211. [\[CrossRef\]](#)
21. Arundel, S.T.; Morgan, T.P.; Thiem, P.T. Deep Learning Detection and Recognition of Spot Elevations on Historic Topographic Maps. *Front. Environ. Sci.* **2022**, *10*, 804155. [\[CrossRef\]](#)
22. Avcı, C.; Sertel, E.; Kabadayı, M.E. Deep Learning-Based Road Extraction from Historical Maps. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3513605. [\[CrossRef\]](#)
23. Jiao, C.; Heitzler, M.; Hurni, L. A Fast and Effective Deep Learning Approach for Road Extraction from Historical Maps by Automatically Generating Training Data with Symbol Reconstruction. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102980. [\[CrossRef\]](#)
24. Jiao, C.; Heitzler, M.; Hurni, L. A Novel Framework for Road Vectorization and Classification from Historical Maps Based on Deep Learning and Symbol Painting. *Comput. Environ. Urban Syst.* **2024**, *108*, 102060. [\[CrossRef\]](#)
25. Chen, Y.; Chazalon, J.; Carlinet, E.; Ngoc, M.O.V.; Mallet, C.; Perret, J. Automatic Vectorization of Historical Maps: A Benchmark. *PLoS ONE* **2024**, *19*, e0298217. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Xia, X.; Zhang, T.; Heitzler, M.; Hurni, L. Vectorizing Historical Maps with Topological Consistency: A Hybrid Approach Using Transformers and Contour-Based Instance Segmentation. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103837. [\[CrossRef\]](#)
27. Heitzler, M.; Hurni, L. Cartographic Reconstruction of Building Footprints from Historical Maps: A Study on the Swiss Siegfried Map. *Trans. GIS* **2020**, *24*, 442–461. [\[CrossRef\]](#)
28. Farella, E.M.; Özdemir, E.; Remondino, F. 4D Building Reconstruction with Machine Learning and Historical Maps. *Appl. Sci.* **2021**, *11*, 1445. [\[CrossRef\]](#)
29. Garcia-Molsosa, A.; Orengo, H.A.; Lawrence, D.; Philip, G.; Hopper, K.; Petrie, C.A. Potential of Deep Learning Segmentation for the Extraction of Archaeological Features from Historical Map Series. *Archaeol. Prospect.* **2021**, *28*, 187–199. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Wong, C.-S.; Liao, H.-M.; Tsai, R.T.-H.; Chang, M.-C. Semi-Supervised Learning for Topographic Map Analysis over Time: A Study of Bridge Segmentation. *Sci. Rep.* **2022**, *12*, 18997. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Ståhl, N.; Weimann, L. Identifying Wetland Areas in Historical Maps Using Deep Convolutional Neural Networks. *Ecol. Inform.* **2022**, *68*, 101557. [\[CrossRef\]](#)
32. Hosseini, K.; Wilson, D.C.S.; Beelen, K.; McDonough, K. MapReader: A Computer Vision Pipeline for the Semantic Exploration of Maps at Scale. In Proceedings of the 6th ACM SIGSPATIAL International Workshop on Geospatial Humanities, Seattle, WA, USA, 1 November 2022; pp. 8–19. [\[CrossRef\]](#)
33. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [\[CrossRef\]](#)
34. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [\[CrossRef\]](#)
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Curran Associates: Red Hook, NY, USA, 2012; pp. 1097–1105.
37. 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.
38. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [\[CrossRef\]](#)
39. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324. [\[CrossRef\]](#)
40. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. *Proc. Mach. Learn. Res.* **2019**, *97*, 6105–6114. Available online: <https://proceedings.mlr.press/v97/tan19a.html> (accessed on 15 June 2026).
41. 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 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [\[CrossRef\]](#)

42. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626. [\[CrossRef\]](#)
43. Li, X.; Chiabrando, F.; Sammartano, G. Machine learning and deep learning for cultural heritage conservation: A bibliometric and task-oriented review. *Remote Sens.* **2026**, *18*, 628. [\[CrossRef\]](#)
44. Wang, H.; Gong, Y.; Zhang, Y.; Li, F. Artificial intelligence for sustainable cultural heritage: Practical guidelines and case-based evidence. *Sustainability* **2025**, *17*, 9192. [\[CrossRef\]](#)
45. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [\[CrossRef\]](#)
46. Brigato, L.; Barz, B.; Iocchi, L.; Denzler, J. Image Classification with Small Datasets: Overview and Benchmark. *IEEE Access* **2022**, *10*, 49233–49250. [\[CrossRef\]](#)
47. Krawczyk, B. Learning from Imbalanced Data: Open Challenges and Future Directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [\[CrossRef\]](#)
48. Ghosh, K.; Bellinger, C.; Corizzo, R.; Branco, P.; Krawczyk, B.; Japkowicz, N. The class imbalance problem in deep learning. *Mach. Learn.* **2024**, *113*, 4845–4901. [\[CrossRef\]](#)
49. Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; Belongie, S. Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9260–9269. [\[CrossRef\]](#)
50. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007. [\[CrossRef\]](#)
51. Yazdi, H.; Sad Berenji, S.; Ludwig, F.; Moazen, S. Deep Learning in Historical Architecture Remote Sensing: Automated Historical Courtyard House Recognition in Yazd, Iran. *Heritage* **2022**, *5*, 3066–3080. [\[CrossRef\]](#)
52. Monna, F.; Rolland, T.; Denaire, A.; Navarro, N.; Granjon, L.; Barbé, R.; Chateau-Smith, C. Deep Learning to Detect Built Cultural Heritage from Satellite Imagery: Spatial Distribution and Size of Vernacular Houses in Sumba, Indonesia. *J. Cult. Herit.* **2021**, *52*, 171–183. [\[CrossRef\]](#)
53. Artopoulos, G.; Maslioukova, M.I.; Zavou, C.; Loizou, M.; Deligiorgi, M.; Averkiou, M. An Artificial Neural Network Framework for Classifying the Style of Cypriot Hybrid Examples of Built Heritage in 3D. *J. Cult. Herit.* **2023**, *63*, 135–147. [\[CrossRef\]](#)
54. Pan, X.; Lin, Q.; Ye, S.; Li, L.; Guo, L.; Harmon, B. Deep Learning Based Approaches from Semantic Point Clouds to Semantic BIM Models for Heritage Digital Twin. *Herit. Sci.* **2024**, *12*, 65. [\[CrossRef\]](#)
55. Zhou, C.; Dong, Y.; Hou, M. DGPCD: A Benchmark for Typical Official-Style Dougong in Ancient Chinese Wooden Architecture. *Herit. Sci.* **2024**, *12*, 201. [\[CrossRef\]](#)
56. Hou, M.; Hao, W.; Dong, Y.; Ji, Y. A Detection Method for the Ridge Beast Based on Improved YOLOv3 Algorithm. *Herit. Sci.* **2023**, *11*, 167. [\[CrossRef\]](#)
57. Li, Y.; Zhao, M.; Mao, J.; Chen, Y.; Zheng, L.; Yan, L. Detection and Recognition of Chinese Porcelain Inlay Images of Traditional Lingnan Architectural Decoration Based on YOLOv4 Technology. *Herit. Sci.* **2024**, *12*, 137. [\[CrossRef\]](#)
58. Yan, L.; Chen, Y.; Zheng, L.; Zhang, Y. Application of Computer Vision Technology in Surface Damage Detection and Analysis of Shedthin Tiles in China: A Case Study of the Classical Gardens of Suzhou. *Herit. Sci.* **2024**, *12*, 72. [\[CrossRef\]](#)
59. Zou, J.; Deng, Y. Intelligent Assessment System of Material Deterioration in Masonry Tower Based on Improved Image Segmentation Model. *Herit. Sci.* **2024**, *12*, 252. [\[CrossRef\]](#)
60. Chen, X.; He, J.; Wang, S. Deep Learning-Driven Pathology Detection and Analysis in Historic Masonry Buildings of Suzhou. *npj Herit. Sci.* **2025**, *13*, 197. [\[CrossRef\]](#)
61. Zhang, S.; Wang, F.; Zhou, H.; Hu, L.; Yang, H.; Zhang, J.; Cai, J. A Coordinate-to-Semantic Attention Network for Multi-Label Ancient Chinese Architecture Image Classification. *npj Herit. Sci.* **2025**, *13*, 14. [\[CrossRef\]](#)
62. Liu, Z.; Hong, W.; Long, R.; Zhu, Y.; Zhang, X. A Lightweight Enhanced EfficientNet Model for Chinese Eaves Tile Dynasty Classification. *npj Herit. Sci.* **2025**, *13*, 494. [\[CrossRef\]](#)
63. Zhou, Y.; Zhang, M.; Li, L.; Huang, W. Parametric 3D Reconstructing and Interpreting Iconographic Evidence: Case of the Song Dynasty Architectural Massing Typologies. *npj Herit. Sci.* **2025**, *13*, 300. [\[CrossRef\]](#)
64. Chen, J.; Dai, X.; Zhang, X.; Zeng, Y. Multimodal Intelligent Reconstruction-Driven Digital Regeneration of Cultural Heritage in Qing Dynasty Tea Plantation Landscapes. *npj Herit. Sci.* **2026**, *14*, 40. [\[CrossRef\]](#)
65. Liu, Y.; Huo, L.; Shen, W.; Shen, T. A Method for 3D Reconstruction of Ancient Buildings Driven by Crowdsourced Images. *npj Herit. Sci.* **2026**, *14*, 81. [\[CrossRef\]](#)
66. Zhang, X.; Chen, D.; Qin, Y. Multimodal Prototype Fusion Network for Paper-Cut Image Classification. *npj Herit. Sci.* **2025**, *13*, 462. [\[CrossRef\]](#)
67. Fang, J.; Feng, Z.; Cai, B. DrawnNet: Offline Hand-Drawn Diagram Recognition Based on Keypoint Prediction of Aggregating Geometric Characteristics. *Entropy* **2022**, *24*, 425. [\[CrossRef\]](#) [\[PubMed\]](#)
68. Eitz, M.; Hays, J.; Alexa, M. How Do Humans Sketch Objects? *ACM Trans. Graph.* **2012**, *31*, 44. [\[CrossRef\]](#)

69. Yi, Y.K.; Zhang, Y.; Myung, J. House Style Recognition Using Deep Convolutional Neural Network. *Autom. Constr.* **2020**, *118*, 103307. [[CrossRef](#)]
70. Zu, X.; Gao, C.; Wang, Y. Interpreting Regional Characteristics of Tibetan-Qiang Houses in Northwestern Sichuan by Deep Learning and Image Landscape. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103865. [[CrossRef](#)]
71. Chen, D.B.; Zuo, L. Ting or Chinese Pavilion: The History of a Chinese Architectural Type and Its Translation in Transcultural Scholarship. *J. Chin. Archit. Urban.* **2025**, *7*, 4107. [[CrossRef](#)]
72. Meng, Y.; Wang, B.; Liu, G. Multi-scale Context-Guided Feature Elimination for Ancient Tower Image Classification. *J. Zhejiang Univ. (Eng. Sci.)* **2024**, *58*, 2489–2499. [[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.