

Article

Classification and Model Explanation of Traditional Dwellings Based on Improved Swin Transformer

Shangbo Miao ¹, Chenxi Zhang ², Yushun Piao ^{1,*} and Yalin Miao ^{3,*}¹ School of Architecture and Urban Planning, Shenyang Jianzhu University, Shenyang 110168, China; miaoshangbo@stu.sjzu.edu.cn² School of Mechanical Engineering, Southwest Jiaotong University, Chengdu 610031, China; chenxi_zhang@my.swjtu.edu.cn³ School of Printing, Packaging and Digital Media, Xi'an University of Technology, Xi'an 710048, China

* Correspondence: ufo3076@126.com (Y.P.); myl@xaut.edu.cn (Y.M.)

Abstract: The extraction of features and classification of traditional dwellings plays significant roles in preserving and ensuring the sustainable development of these structures. Currently, challenges persist in subjective classification and the accuracy of feature extraction. This study focuses on traditional dwellings in Gansu Province, China, employing a novel model named Improved Swin Transformer. This model, based on the Swin Transformer and parallel grouped Convolutional Neural Networks (CNN) branches, aims to enhance the accuracy of feature extraction and classification precision. Furthermore, to validate the accuracy of feature extraction during the prediction process and foster trust in AI systems, explainability research was conducted using Grad-CAM-generated heatmaps. Initially, the Gansu Province Traditional Dwelling Dataset (GTDD) is established. On the constructed GTDD dataset, the Improved Swin Transformer attains an accuracy of 90.03% and an F1 score of 87.44%. Comparative analysis with ResNet-50, ResNeXt-50, and Swin Transformer highlights the outstanding performance of the improved model. The confusion matrix of the Improved Swin Transformer model reveals the classification results across different regions, indicating that the primary influencing factors are attributed to terrain, climate, and cultural aspects. Finally, using Grad-CAM-generated heatmaps for explaining classifications, it is observed that the Improved Swin Transformer model exhibits more accurate localization and focuses on features compared to the other three models. The model demonstrates exceptional feature extraction ability with minimal influence from the surrounding environment. Simultaneously, through the heatmaps generated by the Improved Swin Transformer for traditional residential areas in five regions of Gansu, it is evident that the model accurately extracts architectural features such as roofs, facades, materials, windows, etc. This validates the consistency of features extracted by the Improved Swin Transformer with traditional methods and enhances trust in the model and decision-making. In summary, the Improved Swin Transformer demonstrates outstanding feature extraction ability and accurate classification, providing valuable insights for the protection and style control of traditional residential areas.

Keywords: traditional dwellings; feature extraction and classification; CNN; Swin Transformer; explainability; Grad-CAM



Citation: Miao, S.; Zhang, C.; Piao, Y.; Miao, Y. Classification and Model Explanation of Traditional Dwellings Based on Improved Swin Transformer. *Buildings* **2024**, *14*, 1540. <https://doi.org/10.3390/buildings14061540>

Academic Editor: Yasser Mohamed

Received: 28 February 2024

Revised: 24 April 2024

Accepted: 22 May 2024

Published: 25 May 2024



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

1. Introduction

Traditional dwellings constitute an important architectural category, influenced by specific natural environment, economic, and social factors, embodying cultural value and regional characteristics [1]. The stylistic formation of traditional dwellings is shaped by factors such as geography, geology, climate, religion, society, politics, and history [2]. These structures encapsulate rich cultural connotations, bearing memories of nostalgia, and stand as a significant representation of traditional Chinese culture within the Chinese nation [3].

After the Industrial Revolution, the development of urbanization led to increasingly similar architectural styles in cities. This trend resulted in the gradual disappearance of the

traditional architectural characteristics of urban dwellings, while in rural areas, although traditional features remain identifiable, erosion by nature and improper restoration have blurred the unique elements of traditional dwellings [4]. Therefore, excavating the distinctive features of traditional dwellings and preserving their unique individuality [5] not only aids in understanding and recognizing traditional dwellings but also supports the conservation of architectural heritage and traditional villages.

In the past, research on architectural forms or the characteristics of traditional dwellings primarily relied on “Architectural Morphology” (Architectural Morphology: It refers to the external manifestation of architectural space, which is a human-created material form. It encompasses the visual elements of architecture, such as shape, size, color, texture, position, orientation, and visual inertia. It is the collective perception formed by people about various architectural forms, representing the overall impression left by the existence of architecture) [6], “Architectural Typology” (Architectural Typology: It is a new way of thinking about architecture. It includes three aspects: 1. Typology inherits architectural forms from history; 2. Typology inherits specific architectural fragments and outlines; 3. Typology attempts to reassemble these fragments in a new context. It achieves continuity and harmony in urban morphology by exploring the selection and transformation of urban spatial types, thus maintaining spatial order in the city) [7], and “Semiotics” (Semiotics: In semiotic terms, the appearance, materials, and functions of architecture are abstracted from their respective utilitarian purposes to acquire cultural significance beyond architecture, thus forming a signifying system akin to a language symbol system) [8]. However, these approaches were heavily based on subjective viewpoints and qualitative assessments. With the rapid advancement of artificial intelligence, particularly in image recognition technology, deep learning techniques have found widespread application in identifying and classifying local architectural features [9,10], yielding promising results. However, to date, there has been limited use of deep learning techniques for holistic learning and feature extraction to classify entire buildings.

For this purpose, researching a model that classifies traditional dwelling features based on their overall characteristics, to validate the accuracy of feature extraction during the prediction process and foster trust in AI systems, explainability research was conducted using Grad-CAM-generated heatmaps, holds significant research value and importance for the preservation of traditional dwellings and rural revitalization.

In this study, a novel classification model for traditional dwellings is proposed based on deep learning. Utilizing AI explainability techniques, an analysis is conducted on the strengths and weaknesses of feature extraction across different models. Simultaneously, an analysis of the features extracted by the proposed model from traditional residential architecture is performed. This is completed to verify human–machine consistency and establish robust human–machine trust. The overall contributions can be summarized as follows:

- Gansu Province was partitioned into five geographic regions, conducting surveys on traditional residences within each area, culminating in the establishment of an image dataset pertaining to traditional residences in Gansu Province (GTDD);
- Considering the substantial similarity among traditional residential structures across various regions in Gansu Province, a novel classification model has been proposed to enhance the accuracy of feature extraction and classification. This model, based on the Swin Transformer, implements a parallel arrangement of grouped convolutional neural network (CNN) branches for classification purposes;
- To validate the superior performance of our proposed deep learning network in classifying traditional dwellings, the GTDD dataset was inputted into four distinct deep learning classification models. The evaluation of these models involved metrics such as Accuracy, F1-score, confusion matrices;
- Utilizing the Grad-CAM method to generate heatmaps for explainability analysis. Explainability aids in discussing and analyzing the variations in the morphology of

traditional dwellings. Additionally, it contributes to validating the effectiveness of different classification models and establishing good human–machine trust.

The structure for the remaining sections of the paper is outlined as follows: Section 2 provides an overview of the relevant research in the areas of traditional dwellings, machine learning for architectural classification, and Grad-CAM. Section 3 delves into the construction of the dataset and introduces our developed classification model, the Improved Swin Transformer classification model. Section 4 details the experimental setup, presents the results, and conducts discussions based on the outcomes. Finally, the paper concludes by presenting the findings and summarizing the key insights obtained.

2. Review of Literature

We established the following standards for selecting previous research. For studies on traditional village residences or vernacular architecture, we chose earlier classic and representative research. For architectural classification research based on machine learning, we mainly referred to studies on architectural image classification using CNN in the past five years. For Transformer research, we selected studies on Transformer in other image classifications from the last three years.

2.1. Research on Traditional Dwelling

Traditional dwellings, also known as vernacular architecture or regional architecture, are not only buildings but also important material symbols that carry regional culture and artistic value [11,12]. They evoke deep emotional memories of their hometown and past times. In the 1960s, Bernard Rudofsky proposed the concept of “architecture without architects” and brought vernacular architecture into the field of research, which opened up the study of traditional dwellings abroad [13].

Research on traditional villages in China traces back to the 1940s when the concept of traditional rural dwellings as a distinct architectural type was introduced [14], initiating systematic studies in this field. Liu [15] provided a systematic discussion on the evolution of Chinese housing and offered a detailed introduction to the characteristics of Ming and Qing dynasty dwellings. Wang et al. [16] suggested that the differentiation of traditional Chinese rural architecture is influenced by a combination of factors such as the natural environment, society, culture, and population, leading to the classification of traditional rural dwellings into 12 types. Sha et al. [17] proposed that the similarities and differences in the form, style, and layout of traditional Chinese dwellings stem from the influences of natural elements like geology, landforms, hydrology, and vegetation. Liu et al. [18] conducted a systematic classification study of traditional rural dwellings based on natural zoning, administrative divisions, cultural affiliations, and landscape regionalization.

2.2. Research on Machine Learning for Architectural Classification

In recent years, some researchers have explored the use of image classification techniques to categorize architectural images [19].

Initially, identifying architectural styles heavily relied on conventional image classification techniques [20]. Mathias et al. [21] employed SIFT [22] to extract local features from architectural facade images for automatic recognition of architectural facade styles, yet relying solely on singular visual features could not completely discern buildings. Consequently, Goel et al. [23] improved mining algorithms by using the Word Mining algorithm to position entire images of buildings, aiming to retain spatial context. However, these methods failed to effectively capture the morphological features of architectural styles. Zhang et al. [24] used image blocks to represent fundamental architectural elements and proposed a hierarchical sparse coding algorithm to describe spatial relationships for effectively matching image blocks to capture morphological features. However, as highlighted by Vondrick et al. [25], high similarity in feature space information among different architectural elements might lead to challenges in accurately determining architectural style categories. To address this, Jiang et al. [26] proposed a CSCAE encoder to extract multiple

low-level features from local image regions, enhancing feature recognition. However, these methods face challenges in capturing high-level semantic features and complex content, exhibiting limited generalization capabilities.

With the continuous evolution of deep learning, classic models like LeNet-5 [27], AlexNet [28], VGG [29], ResNet [30], MobileNet [31], and EfficientNet [32] have been introduced. These models have been applied in various domains such as image classification [33,34], object detection [35], semantic segmentation [36], facial recognition [37], medical image analysis [38], autonomous driving [39], and natural language processing [40], achieving outstanding results.

Given their outstanding image classification and feature extraction capabilities, researchers have applied them to the field of architectural type recognition and classification, yielding commendable results [41]. Seung-Yeul et al. [42] employed a Convolutional Neural Network (R-CNN) and a YOLO model to classify and locate architectural types and structural components in videos and images of East Asian traditional buildings from South Korea, Japan, and China. Gonzalez et al. [43] utilized a Convolutional Neural Network (CNN) to annotate Google Street View photos, automatically identifying building materials and architectural structure types from building facades. Han et al. [44] trained a CNN model with a dataset of ancient Chinese architectural images, successfully identifying ancient buildings in four regions of Hubei Province. Quantitative results indicate consistency between the recognized architectural features and those of buildings in the Hubei region. Lamas et al. [45] represented the correlation between European architectural styles and building elements using a structural classification method. They used a Faster R-CNN model to input new architectural images and employed a ResNet-101 model to extract architectural element features from these new input images, thus identifying architectural styles. Zhang et al. [46] employed an EfficientNet neural network structure to automatically classify various feature indicators of residential buildings, including floors, styles, quality, and materials.

In recent years, owing to the exceptional performance of Transformer [47] in natural language processing, some scholars have started applying it to image classification, exploring its potential in visual tasks [48]. Dosovitskiy et al. [49] introduced the Vision Transformer (ViT), which marked the first application of a Transformer in image classification tasks, demonstrating its effectiveness on large-scale image data. Meng et al. [50] proposed a Transformer-based edge detector that integrates both overall image context and local details to extract clear object boundaries and meaningful edges. Li et al. [51] introduced a hybrid network model that combines CNN and Transformer for precise image localization. Zhang et al. [52] employed a novel rice disease identification method based on the Swin Transformer [53], effectively enhancing the accuracy of rice disease detection.

In summary, CNN and Transformer each have their strengths in image feature recognition and classification. CNN extracts image features through convolutional kernels, emphasizing spatial structures and local features while maintaining the translational and rotational invariance of images. In contrast, Transformer relies on self-attention mechanisms, enabling the capture of more diverse image features. Studying the similarities and differences in residential architectural features involves not just variations in individual architectural elements but also the distinctions among multiple elements. Therefore, this paper combines a CNN model and a Swin Transformer model to classify the features of traditional dwellings in different regions of Gansu Province.

2.3. Grad-CAM Research

As mentioned earlier, although deep models demonstrate excellent performance across various domains, their end-to-end nature often leads to limited transparency and explainability [54]. The explainability of deep learning not only aids in enhancing the robustness, reliability, and trustworthiness of models, thereby allowing for wider applications, but also establishes user trust in model decisions [55]. Additionally, it assists in more effectively validating and adjusting algorithms [56] to optimize models more reasonably. Wang et al. [57]

used Grad-CAM to explain the features learned from individual images in chest CT scans, aiming to enhance transparency in medical decision-making. Han et al. [58] proposed a method for banknote recognition and counterfeit detection systems, showcasing model decisions through heatmap visualization. Omeiza et al. [59] introduced an approach to explain autonomous driving behavior, enabling the assessment of the explainability of explanations generated by autonomous driving systems.

As mentioned above, the Grad-CAM heatmap method has been applied in industries like medicine, autonomous driving, and finance, providing substantial data for their decision-making processes. Currently, there is limited research on the application of the Grad-CAM heatmap method in architectural style classification. The utilization of the Grad-CAM heatmap method in architectural classification not only verifies whether the focal points of deep learning models align with traditional feature extraction but also establishes a level of trust between humans and machines. It also helps in identifying features previously overlooked, enabling more profound research.

3. Materials and Methods

3.1. Data Collection

3.1.1. Introduction of Object Area

Gansu Province is located in the northwest of China, between latitude $32^{\circ}11'$ to $42^{\circ}57'$ north and longitude $92^{\circ}13'$ to $108^{\circ}46'$ east. Its terrain is long and narrow, stretching 1655 km from east to west and 530 km from north to south [60]. It shares borders with Xinjiang, Shaanxi, Sichuan, Qinghai, Ningxia, Inner Mongolia, and Mongolia. The region encompasses diverse terrains including mountains, plateaus, plains, river valleys, deserts, and Gobi, resulting in a varied and complex landscape. The province features diverse climatic types, ranging from tropical monsoon, temperate monsoon, and temperate continental, to high-altitude frigid climates from the south to the north. Gansu Province is divided into 14 prefecture-level divisions and is home to 16 ethnic groups, including the Han, Hui, Tibetan, and Tu, making it a multi-ethnic province. Renowned for its rich historical and cultural heritage, the province is traversed by both the Yangtze River and the Yellow River, nurturing cultures such as the Fu Xi Culture, Silk Road Culture, Dunhuang Culture, Yellow River Culture, and Frontier Culture. The province's diverse terrains, rich cultures, and numerous ethnicities have given rise to a variety of traditional dwellings across different regions within the province. Hence, this paper selects Gansu Province as the research subject to study the features and classification of traditional dwellings across its diverse regions.

Based on different regions and cultures, residential architecture in China can be categorized into northern, northwestern, southern, southwest, and Lingnan architectures [61]. The architectural styles in each region are influenced by local geographical environments, climate conditions, cultural traditions, and other factors, presenting unique characteristics and styles that reflect China's extensive history and diverse cultural heritage [16].

The subject of this study falls within the category of northwestern residential architecture. Gansu Province's extensive historical and cultural heritage, diverse geographical landscapes and climates, as well as its diverse ethnicities and cultures, have led to distinct ethnic and regional features in traditional dwelling across different regions [62]. Therefore, as depicted in Figure 1, we have delineated the research area into five geographic units: Gannan Plateau Region, Hexi Corridor Region, Longdong Loess Plateau Region, Longnan Mountain Region, and Longzhong Loess Plateau Region [63].

- Gannan Plateau Region comprises: Gannan Tibetan Autonomous Prefecture;
- Hexi Corridor Region comprises: Jiuquan, Jiayuguan, Zhangye, Jinchang, Wuwei;
- Longdong Loess Plateau Region comprises: Pingliang, Qingyang;
- Longnan Mountain Region comprises: Tianshui, Longnan;
- Longzhong Loess Plateau Region comprises: Lanzhou, Dingxi, Baiyin, Linxia Hui Autonomous Prefecture.

Gansu Province

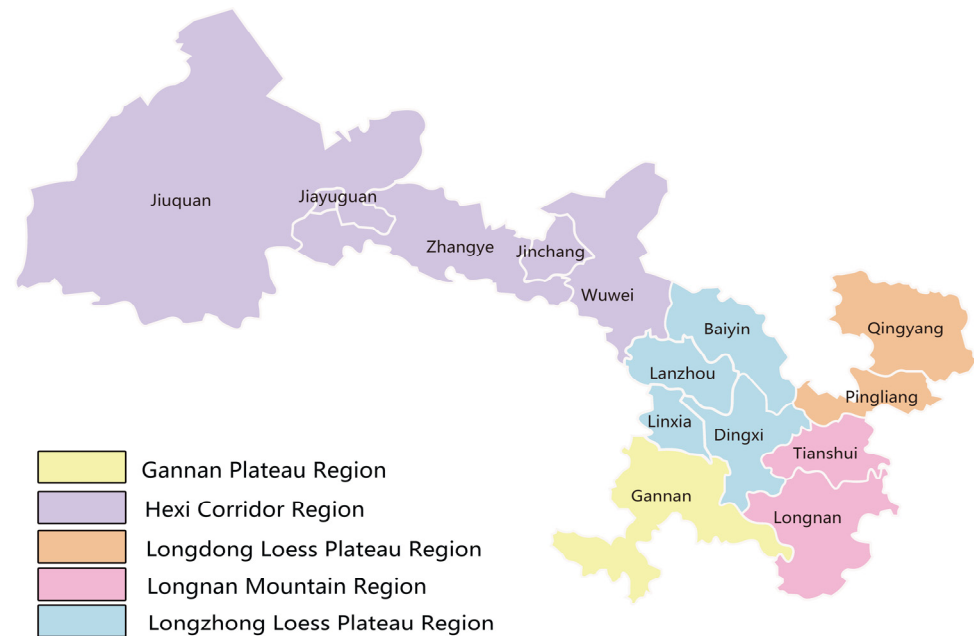


Figure 1. Regional Division of Gansu.

3.1.2. The Characteristics of Traditional Dwelling in Different Regions of Gansu Province

Diverse economic, cultural, topographical, and climatic conditions have led to the formation of unique ethnic characteristics and regional disparities in traditional dwellings across different regions. These variations are evident in spatial arrangements, roof forms, building materials, and architectural decorations [62].

(1) Gannan Plateau Region

The Gannan Plateau Region is situated in the transitional zone between the Qinghai–Tibet Plateau and the Loess Plateau, characterized by numerous valleys and limited flat areas with an elevation decreasing from west to east. It is predominantly inhabited by Tibetan and Han ethnic groups, serving as a cultural convergence between Tibetan and Han cultures. Therefore, the traditional dwelling in the Gannan region is primarily represented by the Gannan Tibetan “Tamping Houses” and adobe flat-topped houses [64]. The structure is “Through Mortise-and-Tenon Structure” or “Raised-beam Style”. The primary building materials are earth, wood, and stone; and the roof forms are flat or dual-sloped stone roofs [62].

(2) Hexi Corridor Region

The Hexi Corridor Region features strong solar radiation, flat terrain, and numerous deserts, serving as the convergence point between Central Plains culture and Western Regions culture [65]. These aspects influence the characteristics of traditional dwelling in this region. Dwellings here typically have a large depth but small width, with relatively small building areas, and low courtyard walls almost aligning with the roofs. This architectural style reduces exposed surfaces, mitigates intense sunlight, resists wind and sand erosion, and historically served as a defense against external invasions [66]. The primary building materials are earth and wood. The structure is “Through Mortise-and-Tenon Structure” or “Raised-beam Style”; roofs are either flat with mud or slightly raised in the center [67].

(3) Longdong Loess Plateau Region

The Longdong Loess Plateau Region is located in the eastern part of Gansu Province, bordering Ningxia and Shaanxi. The traditional dwelling in this region primarily consists of cave dwellings and traditional courtyard houses. Before the Qing Dynasty, 95% of farmers lived in cave dwellings [68]. However, due to migration and economic development, cave dwellings are now rarely seen. Over the past century, the predominant traditional dwelling in this area has remained courtyard houses. The primary building materials are earth and wood; the structure is “Through Mortise-and-Tenon Structure” or “Raised-beam Style”. Unlike other courtyard houses, there is a “Small Tall House” constructed in the southeast or southwest corner of the courtyard, saving land area and providing an elevated position, serving as a lookout and defense against bandits [64].

(4) Longnan Mountain Region

The Longnan Mountainous Region encompasses the Qinling Mountains and Min Mountains, where the Yangtze River and Yellow River basins converge, resulting in highly intricate topography. The predominant traditional dwelling in this area comprises “Tamping Houses” and adobe-tile houses [69]. Tamping Houses typically are two-story wooden-framed houses with earth walls and double-sloped roofs made of wooden boards rather than tiles [70]. To enhance drainage and prevent rainwater erosion [71], houses in this region have steeper slopes compared to other areas in the province, mostly featuring double-sloped roofs and “Overhanging Mountain-Style” walls.

(5) Longzhong Loess Plateau Region

The climate in the Longzhong Loess Plateau Region primarily falls under a temperate continental monsoon climate, mainly semi-humid to semi-arid, with a gradual decrease in rainfall from south to north. The traditional dwelling in this area is predominantly represented by traditional courtyard houses [72]. The primary building materials include earth, wood, or brick and wood; the structure is constructed in mortise and tenon or post-and-lintel styles. The roof forms are mostly single-sloped, with very few having double-sloped roofs [73].

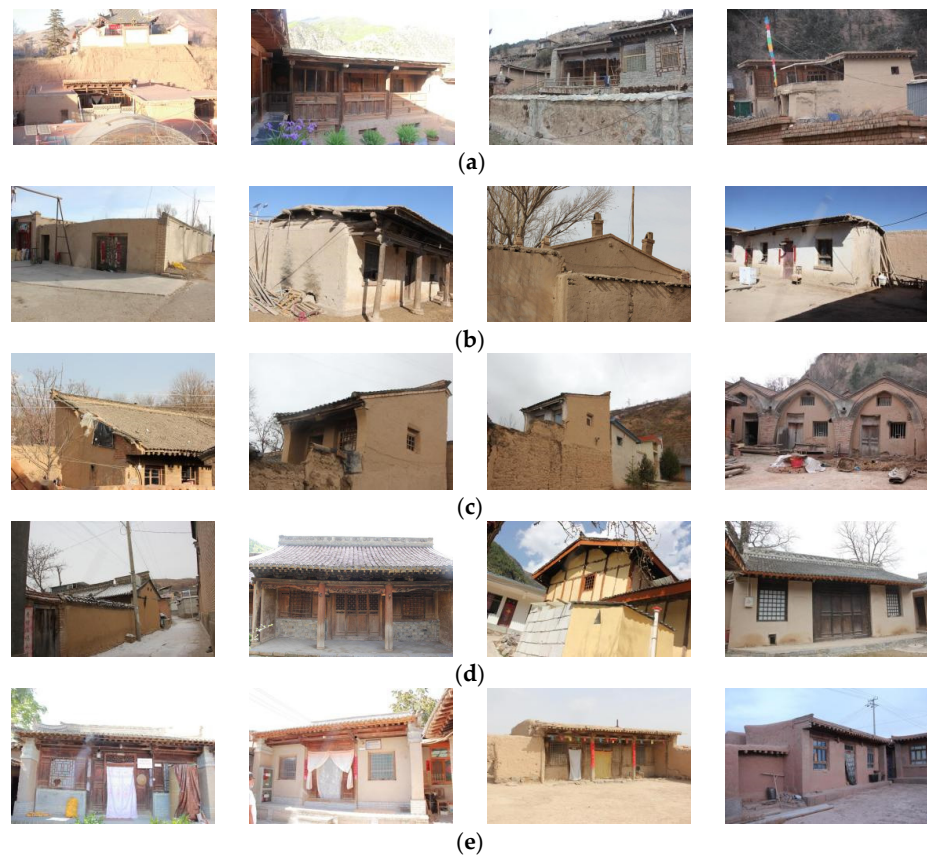
3.1.3. Gansu Traditional Dwellings Dataset (GTDD)

The architectural style of traditional dwellings is a crucial component in managing the rural landscape and preserving cultural heritage structures [9,74]. Therefore, this study primarily focuses on establishing a dataset by collecting images of traditional dwellings within traditional villages in Gansu Province.

Currently, there is no existing image dataset available directly applicable to our research. Hence, we aim to build an image dataset containing traditional dwellings found in villages across Gansu Province. Due to the lack of open resources pertaining to traditional residences in Gansu Province and for the sake of accuracy in our collected image data, our team conducted field surveys on six batches of 112 nationally recognized traditional villages within Gansu Province between 2012 and 2022. To ensure effectiveness and representativeness, we established specific criteria for selecting traditional dwellings in Gansu’s traditional villages, as follows: (1) Selection of houses constructed before 1980, based on onsite research; (2) Houses should reflect regional or ethnic characteristics through architectural style, building materials, and construction techniques; (3) Selected houses should maintain their original features and structural integrity; (4) Houses still utilized for traditional production or living. To eliminate the impact of different lighting conditions on the training results, the dataset collection time includes morning, noon, and afternoon, involving three lighting conditions: front-light, side-light, and back-light. It also covers both sunny and cloudy weather conditions. We ensured diversity in lighting conditions during data collection to minimize the impact of changes in lighting environments on training results. Following these criteria, as illustrated in Table 1, we conducted onsite photography in 72 traditional villages. As shown in Figure 2, we amassed a total of 5228 images of traditional dwellings in Gansu Province. The 5228 images were divided into a training set of 4185 images (80%) and a test set of 1043 images (20%).

Table 1. Selection of villages in the study area and the number of images in the training and test datasets.

Geographical Division	The Number of Villages	Names of Villages	Number of Training Samples/Number of Testing Samples
Gannan Plateau Region	7	Yangbu, Boyu, Niba, Cirina, Hongbaoz, Laligou, Pingding,	243/60
Hexi Corridor Region	14	Tieren, Dongqu, Shuixia, Zehu, Baozi, Nuanquan, Tiancheng, Longshou, Wangfu, Gaosier, Gucheng, Xiakou, Maying, Gaomiao	1020/254
Longdong Loess Plateau Region	8	Zhengping, Luochuan, Jihong, Zhangqu, Wanyan, Zhangaopo, Gaozheng, Tandian	347/86
Longnan Mountain Region	20	Jieting, Xiamiao, Kemengdao, Dongyu, Meijiang, Chaijiashe, Huoshaozhai, Caoheba, Qingni, Chouchi, Daoping, Zhujiagou, Baiguo, Zhangba, Hanan, Xinzhai, Rugongshan, Shimenggou, Qiangqu, Hujiadazhuang,	969/242
Longzhong Loess Plateau Region	23	Geligo, Hekou, Huangjiazhuang, Kuangou, Liancheng, Longwan, Weiquan, Sanhe, Yongtai, Jinya, Xiangshui, Pingbao, Wuliushu, Dahuangwan, Jiangtan, Zhonghe, Chenghe, Yongfeng, Gucheng, Xianglin, Muchang, Wenfeng, Huangjiazhuang,	1606/401
Total	72		4185/1043

**Figure 2.** Partial image data of traditional dwelling in Gansu. (a) Gannan Plateau Region; (b) Hexi Corridor Region; (c) Longdong Loess Plateau Region; (d) Longnan Mountain Region; (e) Longzhong Loess Plateau Region.

3.2. Framework of the Model

3.2.1. Swin Transformer

The Transformer has not only gained popularity in Natural Language Processing (NLP) but has also propelled explorations in the field of computer vision by leveraging long-range dependencies between pixels. Considering the limitations of the Vision Transformer (ViT) in achieving fine-grained image recognition and its high computational complexity [75], this paper adopts the Swin Transformer. By introducing a method that computes self-attention within nonoverlapping window regions and using a movable window approach, it achieves global modeling capabilities while significantly reducing computational requirements. The Swin Transformer architecture resembles that of CNN. The Swin Transformer consists of Patch Partition, Linear Embedding, and four stages, with each stage composed of multiple Patch Merging and Swin Transformer Block structures. Following each Patch Merging, the output feature map's height and width are halved, while the depth is doubled [47].

The advantage of Swin Transformer lies in performing self-attention computation only within a small window of 7×7 patches, reducing algorithmic complexity and computational requirements. Moreover, this paper employs padding techniques to ensure that the image is divisible by the window. Unlike traditional Multi-Head Self-Attention (MSA) modules, Swin Transformer is constructed based on shifted windows, replacing the conventional MSA with W-MSA [76]. As illustrated in Figure 3, the Swin Transformer inputs feature maps into Swin Transformer blocks, passing through LayerNorm (LN) layers, W-MSA, and MLP layers sequentially, followed by another LN layer, SW-MSA, and MLP. Compared to W-MSA, SW-MSA excels in executing a moving window approach, facilitating information exchange between windows that were originally nonoverlapping.

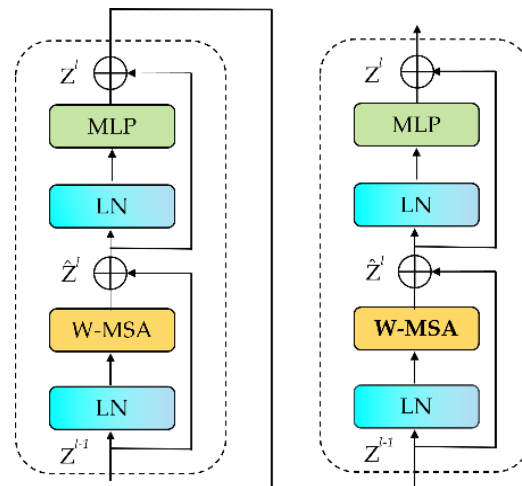


Figure 3. Swin Transformer Block.

3.2.2. Grouped Convolution

A general unit of a neural network can be represented by the following formula:

$$F(x) = \sum_{i=1}^C T_i(x) \quad (1)$$

In the realm of ResNeXt, the transformation T above can take any form, such as a series of convolution operations, totaling C independent transformations. The term C , termed as the cardinality by the authors, is highlighted as more crucial to the outcome compared to the width and depth.

Combining the identity mapping of ResNet, the structure with residual can be represented by the following equation:

$$y = x + \sum_{i=1}^C T_i(x), \quad (2)$$

in Equation (2), the variable “ x ” is the structure with residuals.

ResNeXt introduces a strategy situated between conventional convolutional kernels and depth-wise separable convolutions: grouped convolutions. This approach balances the two strategies by controlling the number of groups (cardinality). Grouped convolution draws inspiration from Inception. Unlike Inception, which requires manual design for each branch, ResNeXt maintains an identical topology across its branches. Finally, when combined with residual networks, this approach culminates in the ultimate ResNeXt model.

As depicted in Figure 4, the residual connection corresponds to x'' in the formula, while the remaining consists of 16 independent convolution transformations of the same structure, subsequently fused together, adhering to the split–transform–merge paradigm.

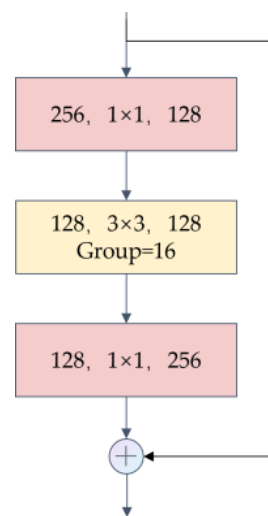


Figure 4. Grouped Convolution Modules. Note: In Figure 4, the grouped convolution module first performs dimensionality reduction on the input features using a 1×1 convolution layer, reducing the number of channels from 256 to 128. Then, it processes the features using grouped convolution with a 3×3 kernel and 16 groups, while keeping the output channels equal to 128. Next, the features are upsampled using a 1×1 convolution. Finally, the output is added to the input to obtain the final output.

3.2.3. The Overall Architecture of The Network

Considering the significant similarity among different categories of architectural images across various regions in Gansu Province, to enhance recognition accuracy by enabling the classification network to simultaneously learn distinctive local features and global characteristics, this paper introduces an end-to-end dual-path architecture for architectural classification. Illustrated in Figure 5, the network is built upon the Swin Transformer and deploys parallel CNN branches to enhance the network’s capability to extract local features. In the end, fully connected layers are employed, with the number of output neurons corresponding to the number of categories for classification. The CNN branch comprises three stages of ResNeXt Block16x64d and other components, each stage consisting of 3 blocks. In “ResNeXt Block16x64d”, the “16” denotes the number of groups in grouped convolutions, i.e., the cardinality, indicating the quantity of grouped convolutions utilized within each basic residual block. The “64” signifies the number of convolutions per branch. This structural design aims to enhance the model’s representational and feature learning abilities. The network primarily consists of local information branches, global information branches, a fusion layer, and an output layer.

The architecture involves feeding building images into distinct components like the ResNeXt Block and convolutional modules, forming a local information extraction branch to strengthen local feature extraction and enhance network performance. Similarly, a global

information branch uses the Swin Transformer to acquire long-range semantic details from building images, compensating for deficiencies in local features to enable comprehensive image recognition. Subsequently, these two branches undergo a cascaded operation to integrate the extracted features, ultimately culminating in category outputs through fully connected layers.

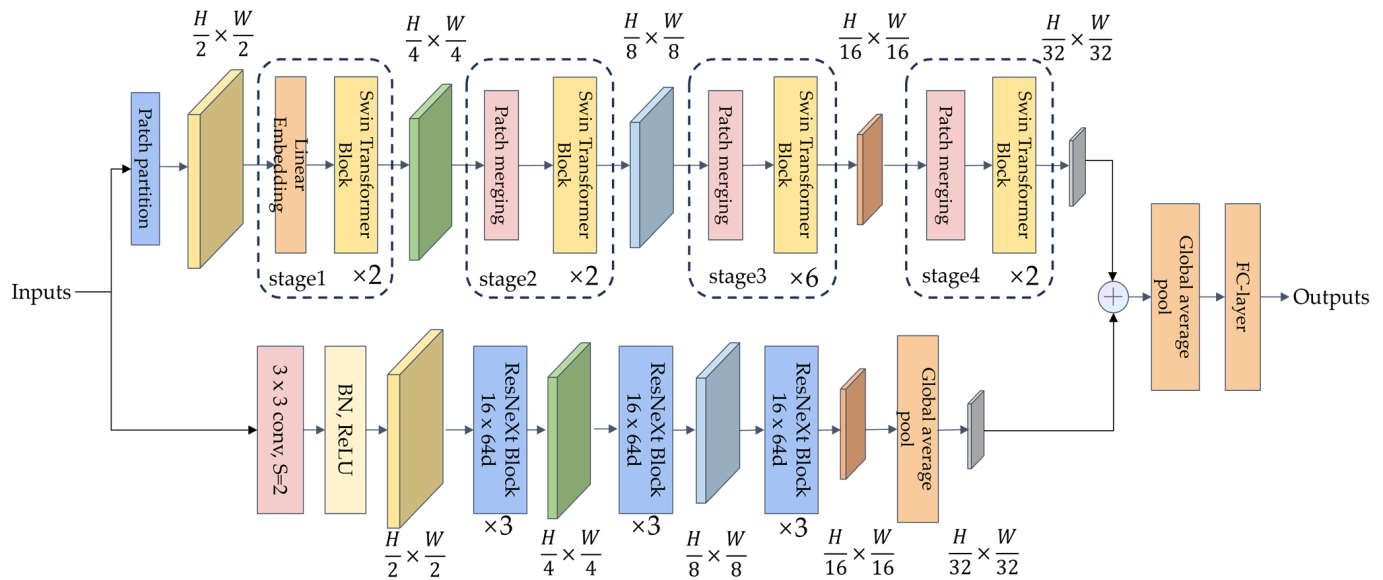


Figure 5. The overall architecture of the network.

4. Results and Discussion

In this section, we present an analysis of the experimental data obtained through the experimental methods proposed in this paper. Specifically, we applied deep learning techniques to classify and visualize traditional dwellings in Gansu Province. Our study of the architectural features of traditional residences in Gansu Province is based on geographical divisions. Due to the non-open-source nature of the model and the limited research on similar work, for a fair comparison, we only compared open-source methods on our self-built GTDD dataset. Therefore, employing four different deep learning models: ResNet-50 [30], ResNeXt-50 [77], Swin Transformer [47], and an Improved Swin Transformer, classified traditional residences in different regions. This allowed us to discern the varying performance of these deep learning models in classifying traditional residences within the region.

Subsequently, employing the explainability model Grad-CAM [55], we generated heatmaps for the four aforementioned deep learning models during the classification of traditional residences. These heatmaps were used to validate the performance of different deep-learning models in feature extraction.

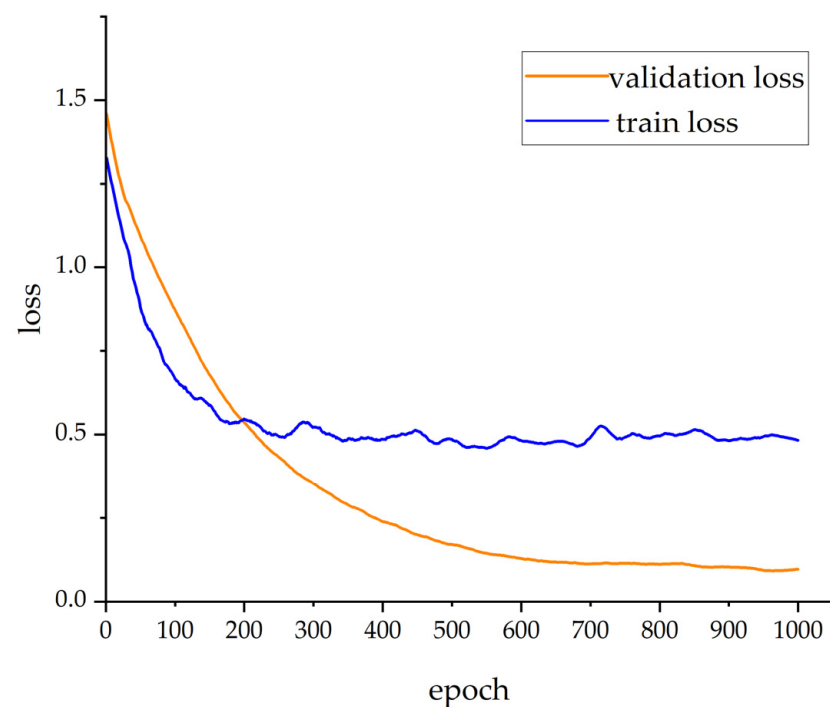
Finally, an analysis was conducted on the features and discrepancies in traditional residences based on the heatmaps generated by the Improved Swin Transformer model.

4.1. Experimental Environment

In order to ensure the fairness and accuracy of the experiment, the same experimental environment was used, as shown in Table 2. And set the same initial training parameters. All input image data were resized to 256×256 pixels, and the batch size was set to 32. All models underwent training for 1000 epochs. Using the SGD optimizer, the initial learning rate was set to 1×10^{-2} , and the final learning rate was reduced to 0.01 times the maximum learning rate. The learning rate was decayed using cosine annealing. Figure 6 shows the entire training process of the Improved Swin Transformer model, wherein the model was trained to converge around 700 epochs.

Table 2. Experimental environment.

Environment	Versions or Model Number
CPU	Intel Xeon(R) W-2235 CPU @ 3.80 GHz × 12 (Chandler, AZ, USA, Purchased in Xi'an, China)
GPU	NVIDIA RTX 3090, 24 GB memory (Santa Clara, CA, USA, Purchased in Xi'an, China)
OS	Ubuntu 18.04
CUDA	11.3
PyTorch	1.8.1
Python	3.9.2

**Figure 6.** Improved Swin Transformer loss curve.

4.2. Model Evaluation Metrics

In order to objectively evaluate the effectiveness of the proposed model in identifying traditional residential structures within traditional villages in Gansu Province, this paper primarily adopts four metrics Accuracy, Precision, Recall, F1-score and Inference Time as classification criteria to accurately assess the model's performance.

Accuracy: It is the most prevalent evaluation metric, signifying the ratio of correctly predicted samples to the total predicted samples, irrespective of whether the predicted samples are positive or negative. Computed as follows:

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\%, \quad (3)$$

Precision: It measures accuracy and reflects the ratio of correctly predicted positive samples to all samples predicted as positive, essentially revealing how many predicted positive samples are genuinely positive. Computed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \quad (4)$$

Recall: Also termed sensitivity, it illustrates the ratio of correctly predicted positive samples to the total number of true positive samples. It illuminates the ability to correctly identify positive samples from this sample set. Computed as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \quad (5)$$

F1-score: At times, Precision and Recall metrics may contradict each other. In such scenarios, the F-score metric, which is the harmonic mean of Precision and Recall, is called upon. It provides a balanced assessment of the algorithm, ranging from a minimum of 0 to a maximum of 1. Computed as follows:

$$\text{F1 - score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\%, \quad (6)$$

In the aforementioned equations, TP (True Positive) represents correctly classified positive instances; TN (True Negative) represents accurately classified negative instances; FP (Falsely Positive) represents falsely classified positive instances; FN (Falsely Negative) represents erroneously classified negative instances.

4.3. Experimental Results and Discussion

4.3.1. Comparison among Different Models

To validate the effectiveness of our proposed approach in recognizing and classifying traditional dwellings in Gansu, we conducted experiments on the GTDD dataset. We compared the Improved Swin Transformer model proposed in this paper with other models, including the ResNet-50, ResNeXt-50, and Swin Transformer models.

(1) Different Models' Classification Results

The experimental results are shown in Table 3. Each model selects the weight file with the highest accuracy. As depicted in the table, our method achieved good classification performance on the GTDD dataset, with the five metrics being 90.03%, 85.88%, 89.32%, 87.44%, and 0.22 s, respectively.

Table 3. Comparative experimental results of different models.

Models	Accuracy/%	Recall/%	Precision/%	F1-Score/%	Inference Time/S
ResNet-50	86.67 (−3.36)	80.99 (−4.89)	83.30 (−6.02)	82.06 (−5.38)	0.18 (−0.04)
ResNeXt-50	87.54 (−2.49)	82.27 (−3.61)	89.49 (+0.17)	85.12 (−2.32)	0.17 (−0.05)
Swin Transformer	80.73 (−9.3)	70.54 (−15.34)	79.30 (−10.02)	73.40 (−14.04)	0.16 (−0.06)
Improved Swin Transformer	90.03	85.88	89.32	87.44	0.22

Note: In Table 3, the plus or minus signs within the table parentheses indicate the difference of the value from the value of the Improved Swin Transformer model. A negative value indicates it is less than the value of the Improved Swin Transformer model, while a positive value indicates it is greater than the value of the Improved Swin Transformer model.

ResNet-50 and ResNeXt-50 are two types of pure CNN models. Between the two, ResNeXt-50 demonstrates superior performance compared to ResNet-50. ResNeXt-50 is an enhancement of the ResNet model, differing in its use of a concept called “groups” instead of multiple smaller pathways. These “groups” internally consist of multiple parallel pathways, each dedicated to learning distinct features. This design enables the network to more effectively acquire a greater number of features, thereby enhancing its representational capacity [78,79]. According to Table 3, compared to the ResNet-50 model, the ResNeXt-50 model exhibits an increase in Accuracy by 0.87%, F1-score by 3.06%, Precision by 6.19%, Recall by 1.28% and Inference Time by 0.01 s. This indicates that increasing the number of “groups” contributes to enhancing model performance. While CNN holds an advantage in capturing local features, their slightly smaller convolutional kernels result in a slight deficiency when capturing global features.

The Swin Transformer model represents an approach that utilizes sliding window computation for self-attention, showcasing advantages in global information interaction. However, the necessity of processing all pixels leads to a substantial amount of information, potentially overshadowing local feature details [53]. Furthermore, the Swin Transformer requires a significant amount of data during training to ensure higher performance [80]. As indicated in Table 3, the Swin Transformer model exhibits the lowest performance across all four metrics.

Due to various factors such as natural surroundings, historical heritage, and religious influences, traditional dwellings exhibit multifaceted differences in their structural forms across different regions. These disparities encompass aspects like eaves, doors, windows, materials, and colors [16,81]. As a result, our proposed classification network can concurrently extract a broader array of architectural features, thereby enhancing the model's accuracy. Consequently, the incorporation of the CNN model into the Swin Transformer architecture addresses deficiencies in global feature acquisition exhibited by the CNN network, demonstrating a complementary advantage [82]. As depicted in Table 3, the Improved Swin Transformer model achieves an accuracy 3.36%, 2.49%, and 9.3% higher than ResNet-50, ResNeXt-50, and the Swin Transformer models, respectively. Its recall also outperforms the other three models by 4.89%, 3.61%, and 15.34%, while its F1-score surpasses them by 5.38, 2.32, and 14.04, respectively. Although the Precision of the ResNeXt-50 model slightly surpasses that of the Improved Swin Transformer model, the latter's F1-score outperforms the former by 2.32%. F1-score serves as an evaluation metric for classification tasks, representing the harmonic mean of Precision and Recall and thus providing a comprehensive assessment of the model's performance. Based on inference time, the Improved Swin Transformer takes the longest time to infer one image among the four models, at 0.22 s. This is slower than ResNet-50, ResNeXt-50, and Swin Transformer by 0.04 s, 0.05 s, and 0.06 s, respectively, but this difference is acceptable. Since the purpose of our work is to classify and extract features from one or several images later on, this time difference does not significantly impact future work. Therefore, considering Accuracy, Precision, Recall, F1-score, and Inference Time collectively, our proposed Improved Swin Transformer model emerges as the superior choice for traditional dwelling classification.

(2) Explainability Comparison of Different Models.

Precision is not the sole evaluation metric for algorithms; good explainability is also an indispensable characteristic of an excellent algorithm. [54]. Utilizing the Grad-CAM method to generate heatmaps for explainability analysis allows for a direct reflection of the distinguishing features that various classification models focus on in these different regional images [77]. This, in turn, validates the feature extraction capabilities and trustworthiness of different models. Figure 7 illustrates the heatmaps of traditional dwellings across different regions by various models.

In Figure 7, as indicated by a4, a5, b4, c5, d4, e4, f5, h5, and j5, the features extracted by the ResNeXt-50 and ResNet-50 models appear relatively singular, whereas the Swin Transformer model exhibits a more diverse focus on feature points, such as f3, g3, h3, and j3. Illustrated by a3, b5, c4, d4, f4, g3, g4, h4, i3, i4, and j4, the Swin Transformer, ResNeXt-50, and ResNet-50 models predominantly concentrate on the roof and gable features of traditional dwellings. However, the Improved Swin Transformer model not only attends to these roof and gable features (e.g., d2, g2, h2) but also captures facade elements such as doors, windows, and columns (e.g., a2, b2, c2, e2, f2, i2, j2). Consequently, the Improved Swin Transformer can identify and discover more architectural characteristics. Regarding the influence of the environment on the models, as depicted by b3, c3, d3, d5, e3, and e5, environmental factors impact the Swin Transformer, ResNeXt-50, and ResNet-50 models in the extraction of traditional dwelling features. These factors lead to the identification of sky, ground, trees, and other environmental elements rather than the accurate extraction of features specific to traditional dwellings. Conversely, the Improved Swin Transformer can accurately pinpoint the architectural elements even amidst varied environmental contexts. Overall, the Improved Swin Transformer demonstrates the most

precise extraction of traditional dwelling features while being least affected by external environmental influences.

Therefore, through comparative analysis of Accuracy, Precision, Recall, F1-score, and heatmaps across different models, the Improved Swin Transformer model excels in the identification and classification of features in traditional dwellings compared to the other three models.

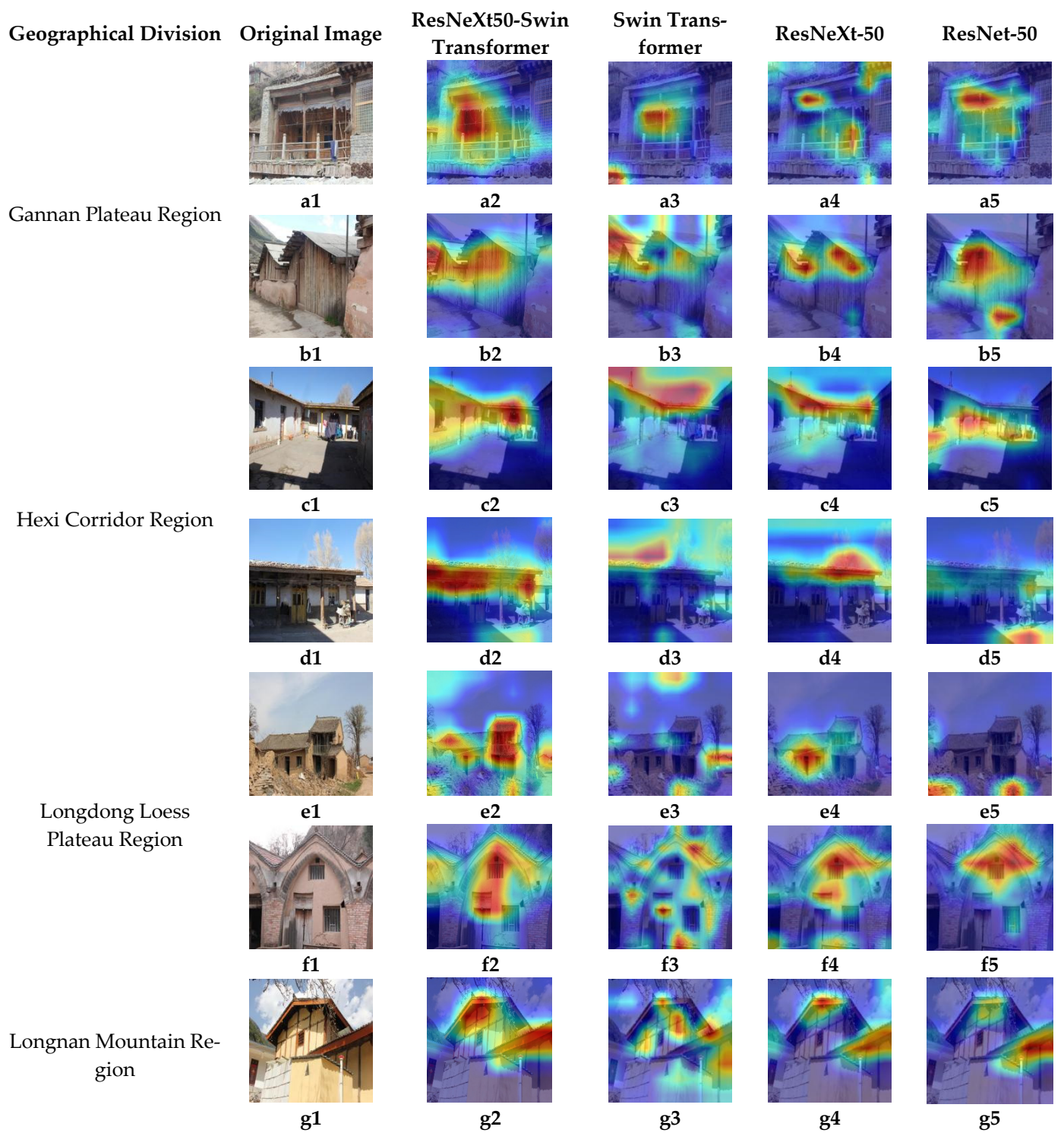


Figure 7. Cont.

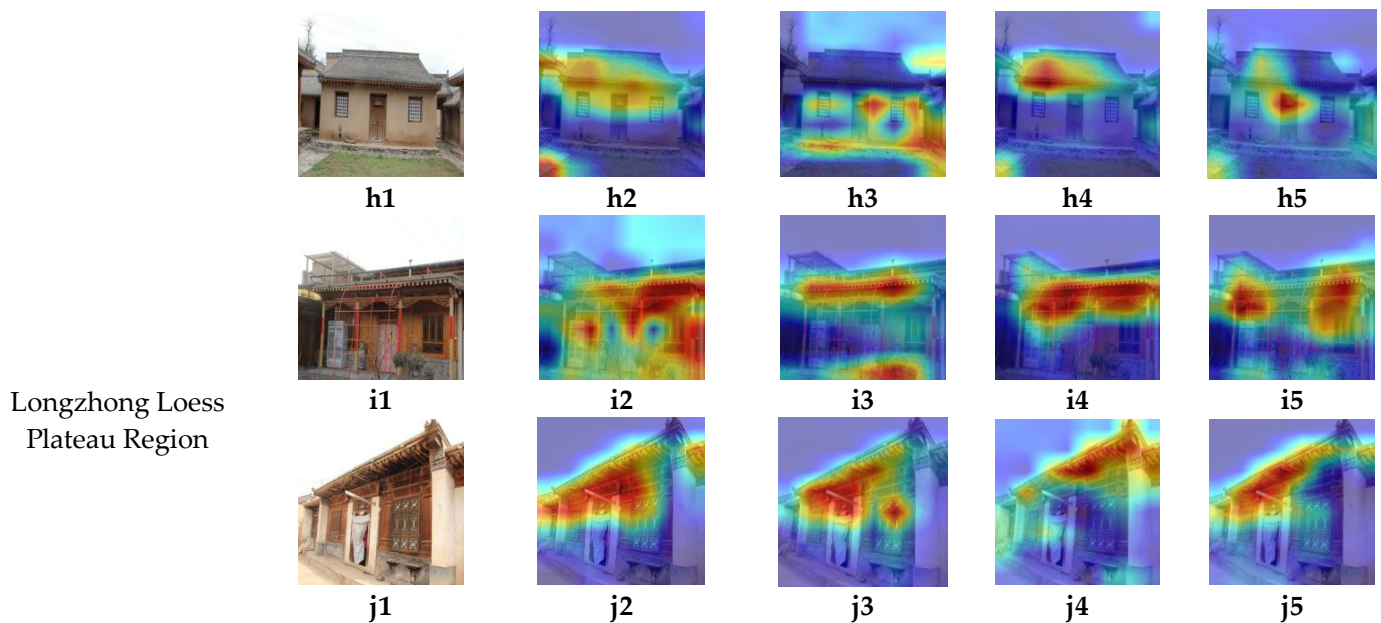


Figure 7. Heatmaps of focus points on traditional dwellings across different regions by various models. In Figure 7, (a–j) refers to traditional residential buildings in different regions, where numbers 1 refer to the original photos taken, and 2–5 refer to the heat map results generated by different models.

4.3.2. Confusion Matrix

In order to better understand how our Improved Swin Transformer model performs in classifying different regions, we have generated a confusion matrix for the Improved Swin Transformer model. Figure 8 depicts the confusion matrix of the Improved Swin Transformer model, outlining the recognition accuracy and the number of photos for five regions, along with the probability and quantity of misclassifications into other regions. From Figure 8, it is evident that the recognition accuracy for traditional dwellings in Gannan is 75%, with a relatively high probability of being misclassified as the Loess Plateau region at 11.7%. The Hexi Corridor region demonstrates a recognition accuracy of 95.7%, yet it is more prone to misidentification than the Loess Plateau at 3.9%. The Loess Plateau East region exhibits a recognition accuracy of 64.0%, with the highest likelihood of misidentification as the Loess Plateau area, reaching 22.1%. The Loess Plateau South region shows a recognition accuracy of 87.2%, with the highest probability of misidentification as the Loess Plateau at 10.7%. Finally, the Loess Plateau Central region achieves a recognition accuracy of 89.5%, yet it is most likely to be misidentified as the Hexi Corridor region at 6.5%. Considering the terrain and climatic background, these observations are reasonably explicable.

As depicted in Figure 9, the northwestern part of Longnan is adjacent to Zhouqu County in Gannan, resulting in striking similarities in their traditional dwellings that pose a challenge for human differentiation. In Figure 10, the southeastern region of Hexi borders Baiyin City in Longzhong, exhibiting identical climatic and topographic features, consequently leading to remarkable resemblances in their traditional dwelling. Figure 11 illustrates that Longdong, belonging to the Loess Plateau, shares exceedingly similar terrain, climate, and cultural aspects with Longzhong. Without a sufficient array of study samples, identification errors are highly probable. Longzhong, situated in the central part of Gansu Province, interfaces with various other regions, fostering a diverse range of architectural styles in traditional dwellings; however, those at the boundaries showcase remarkably high similarities.

As evidenced above, geography and culture emerge as primary factors contributing to misclassification within the model. Furthermore, the size of the training samples also

serves as an influencing factor; the limited quantity of training samples impedes the model’s ability to learn nuanced distinctions.

Actual Label	GanNan	45 75%	2 3.3%	0 0%	7 11.7%	6 10.0%
	HeXi	0 0%	243 95.7%	1 0.4%	0 0%	10 3.9%
	LongDong	0 0%	5 5.8%	55 64.0%	7 8.1%	19 22.1%
	LongNan	1 0.4%	4 1.7%	0 0%	211 87.2%	26 10.7%
	LongZhong	2 0.5%	26 6.5%	4 0.9%	10 2.5%	359 89.5%
		GanNan	HeXi	LongDong	LongNan	LongZhong
Predicted Label						

Figure 8. Confusion matrix of the Improved Swin Transformer model subsection.



Figure 9. The comparison of traditional dwelling between the Gannan Plateau Region and the Mountainous Region of Longnan. (a) Traditional dwelling in the Gannan Plateau Region; (b) traditional dwelling in the Mountainous Region of Longnan.



Figure 10. The comparison of traditional dwelling between the Hexi Corridor Region and the Mountainous Region of Longnan. (a) Traditional dwelling in the Hexi Corridor Region; (b) traditional dwelling in the Longzhong Loess Plateau Region.



Figure 11. The comparison of traditional dwelling between the Longdong Loess Plateau Region and the Longzhong Loess Plateau Region. (a) Traditional dwelling in the Longdong Loess Plateau Region; (b) traditional dwelling in the Longzhong Loess Plateau Region.

4.3.3. Explainability Analysis of Improved Swin-Transformer

Conducting explainability analysis on Grad-CAM-generated heatmaps from the output of the classification model is beneficial for understanding which features the model extracts to identify objects [76]. This process aids in validating the model's feature extraction capabilities.

Concerning traditional dwelling, this approach not only provides a clear view of the model's primary focus during feature extraction, allowing for comparisons with previous conventional feature extraction methods, but also reveals features previously overlooked in traditional extraction techniques.

Analysis of the heatmaps generated from the Improved Swin Transformer model outputs reveals. In Figure 12, the model primarily focuses on the features of traditional dwellings in the Gannan Plateau region, concentrating on aspects like eaves and materials. Gannan boasts abundant timber resources, serving as the primary building material. Except for the surrounding walls, which are constructed using earth (or stone), the rest of the structure comprises wood. Beams, pillars, rafters, and more often employ durable materials like Scots pine or cypress to create a wooden load-bearing structure. Upon completion of the roof, the upper, lower, and surrounding areas of the house are entirely finished with wooden panels [62], as depicted in Figure 12a,b. In forested areas, roofs are designed as pitched "Tamping Houses", shown in Figure 12c. Conversely, areas farther from the forest use mud for flat roofs [64], as seen in Figure 12d.

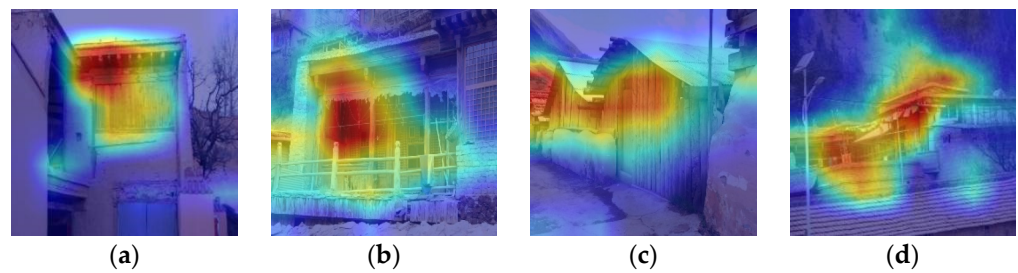


Figure 12. (a–d) Heatmaps of traditional dwelling in the Mountainous Region of Gannan.

Figure 13 illustrates the focus on characteristic features in the Hexi Corridor region, notably the low, flat-roofed houses, as depicted in Figure 13a. Due to the region's arid, less rainy climate, severe winter conditions, high winds, and constraints in building materials, houses typically have either flat roofs without tiles, as shown in Figure 13b, or slightly raised flat roofs in the center, as shown in Figure 13c. These architectural styles not only facilitate sand prevention, insulation, and deterrence against intruders but also take drainage into consideration [67].

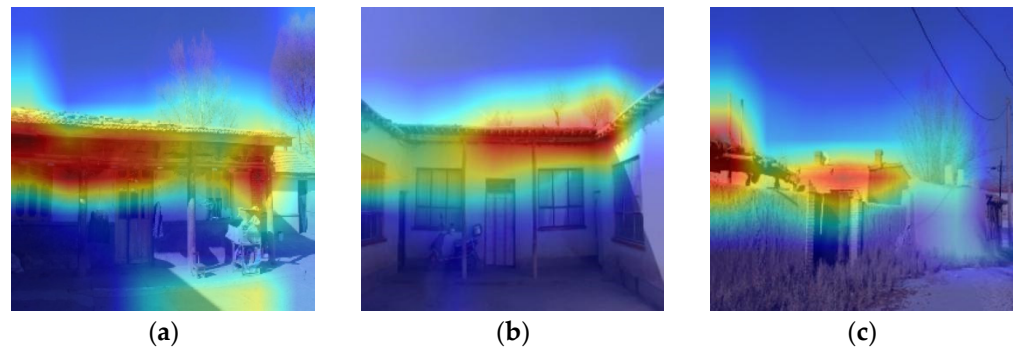


Figure 13. (a–c) Heatmaps of traditional dwelling in the Hexi Corridor Region.

As depicted in Figure 14, the focus in the Longdong Loess Plateau region revolves around architectural forms. Illustrated in Figure 14a, the “Small Tall House” stands as a distinctive characteristic of residential buildings in this area, present in nearly every household. Initially established to deter intruders, although it no longer serves this purpose [64], locals believe it still possesses a protective role against evil spirits. In Figure 14b, the model also emphasizes the unique feature of “Guo Yao” specific to the Longdong region. Due to the scarcity of timber in this area, compared to wooden structures, “Guo Yao” is more efficient in terms of manpower, resources, and finances [83].

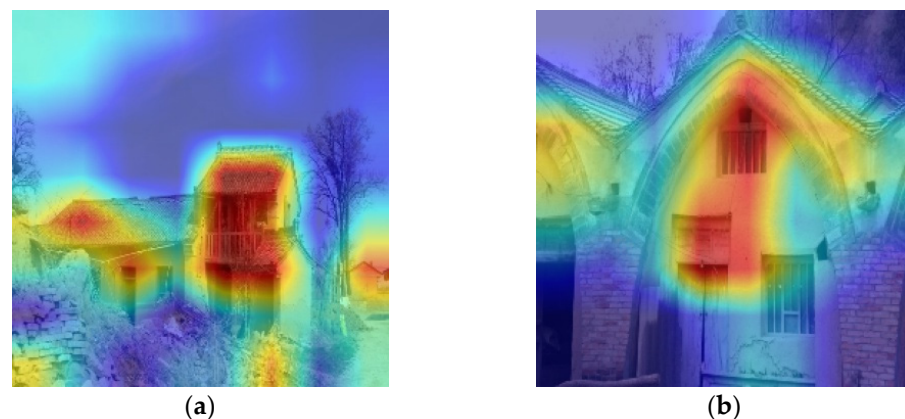


Figure 14. (a,b) Heatmaps of traditional dwelling in the Longdong Loess Plateau Region.

In Figure 15, the primary focus of the mountainous region of Longnan lies in the eaves and gables. The southern part of Longnan experiences a humid climate with abundant precipitation, ample timber resources, and a terrain characterized by numerous peaks and valleys. Consequently, residential buildings adopt a broad layout with shallow depths, as seen in Figure 15a,b. Their gables predominantly follow the overhanging mountain style, featuring wooden double-sloped roofs [69]. In the northeastern region, with a relatively flat terrain, courtyard-style residences prevail, as depicted in Figure 15c. These houses have double-sloped roofs, but their slopes are steeper, ranging from 5% to 10% compared to the other four regions [84].

In Figure 16, the primary focus in the Loess Plateau region of Longzhong lies on the doors, windows, and eaves of residential buildings. Compared to the other three regions, the traditional dwellings in this area exhibit a façade featuring either “four doors and eight windows” as shown in Figure 16a or “four doors and four windows” as depicted in Figure 16b. Influenced by precipitation, the eaves in this region predominantly feature single-sloped roofs as shown in Figure 16c, with fewer double-sloped roofs [73], characterized by deeper eaves in Figure 16d. The roof slopes range between 3% to 5%, and the gable style tends to be rigid mountain style [84].

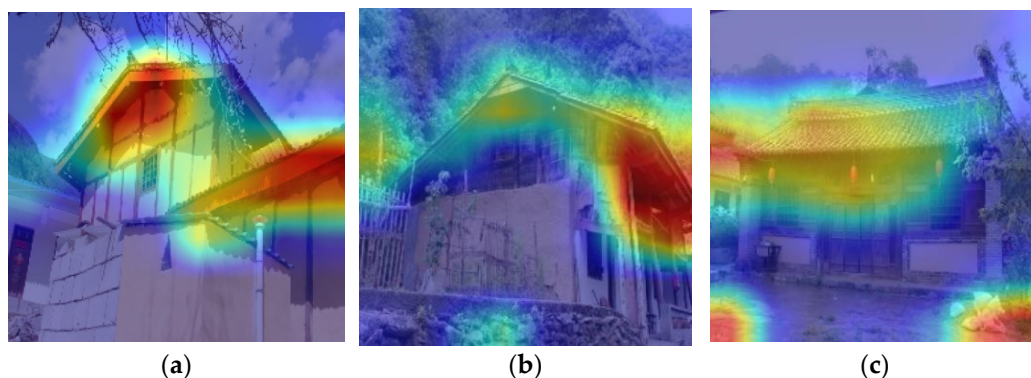


Figure 15. (a–c) Heatmaps of traditional dwelling in the Mountainous Region of Longnan.

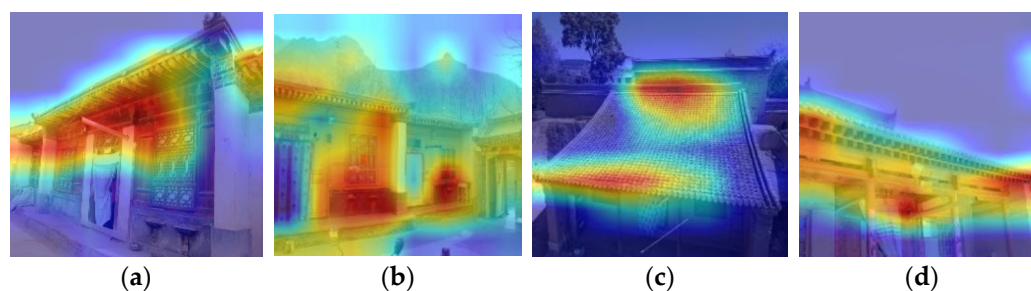


Figure 16. (a–d) Heatmaps of traditional dwelling in the Longzhong Loess Plateau Region.

Through explainability analysis of heatmaps generated by Grad-CAM, our proposed Improved Swin Transformer model exhibits a substantial consistency with the features extracted by traditional methods concerning the characteristic focus points of traditional dwelling in Gansu. This correspondence underscores the effectiveness of this approach in extracting and identifying the features of traditional dwellings. It helps establish human–machine trust and may offer new avenues for future research on traditional dwelling.

5. Conclusions

The study focuses on traditional dwelling in Gansu Province, utilizing deep learning techniques to classify the characteristic features of traditional dwelling across different

regions within Gansu Province. Additionally, explainability analysis of the classification results is conducted through heatmaps generated by Grad-CAM.

To address subjectivity, low accuracy, difficulty in quantification, and indistinct feature extraction in classifying characteristic features of traditional dwellings across different regions, we employed an Improved Swin Transformer model. To demonstrate the advantages of this Improved Swin Transformer model, under the same environment and parameters, we compared it with the ResNet-50, ResNeXt-50, and the Swin Transformer models, utilizing our established GTDD dataset. The results indicate that our proposed Improved Swin Transformer model exhibits a higher accuracy than ResNet-50, ResNeXt-50, and the Swin Transformer by 3.36%, 2.49%, and 9.3%, respectively. In terms of F1-score, our Improved Swin Transformer model achieved 87.44%, surpassing the other three models by 5.38%, 2.32%, and 14.04%, respectively. Based on these metrics, our proposed Improved Swin Transformer model demonstrates superior performance. Additionally, from the generated heatmaps, compared to the other models, our Improved Swin Transformer model is least affected by environmental factors, accurately pinpointing architectural features.

Upon examining the confusion matrix generated by the Improved Swin Transformer model, the model exhibits the highest identification accuracy in the Hexi Corridor region, reaching 95.7%. Following this, the accuracy rates are 89.5%, 87.2%, and 75% for the Longzhong Loess Plateau, Longnan mountainous region, and Gannan Plateau regions, respectively. The model displays the lowest identification accuracy in the Longdong Loess Plateau region, at 64%.

Explainability research on the heatmaps generated by the Improved Swin Transformer model indicates that the model can accurately extract distinctive features of traditional dwellings in different regions. These features align closely with those highlighted by traditional methods, validating the scientific effectiveness of the model. This consistency also establishes a strong level of trust between humans and the model.

In conclusion, this study serves not only as a tool for extracting the morphological features and classification of traditional dwellings in Gansu Province but also offers a theoretical basis for the preservation and utilization of traditional village dwellings and architectural heritage. Simultaneously, it provides insights into improving the living environment and controlling architectural style in traditional villages. Moreover, this methodology can be applied to identify and classify features of traditional residential and heritage buildings in other regions.

In the future, we plan to further refine the model, increase the dataset size, and enhance the model's ability to recognize and classify features of traditional dwellings. This endeavor aims to contribute insights into the preservation and sustainability of traditional dwellings and villages.

Author Contributions: Conceptualization, S.M. and C.Z.; Methodology, Y.P. and S.M.; Software, S.M.; Validation, S.M. and C.Z.; Formal Analysis, S.M.; Investigation, S.M. and C.Z.; Resources, S.M.; Data Curation, S.M.; Writing—Original Draft Preparation, S.M.; Writing—Review and Editing, Y.P., Y.M. and C.Z.; Visualization, S.M.; Supervision, Y.P. and Y.M.; Project Administration, Y.M.; Funding Acquisition, Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number 62076200) and by the Key Research and Development Project of Shaanxi Province (grant number 2023-YBGY-149).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Acknowledgments: Thanks to the School of Printing, Packaging and Digital Media, Xi'an University of Technology for providing model training and data analysis.

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

References

1. Yan, Z.; Jin, T.; Da, D.H.; Da, W.X. The Research on Traditional Dwelling Culture Geography. *South Archit.* **2013**, *1*, 83–87.
2. De, Q.S. *From Traditional Houses to Regional Buildings*; China Building Materials Industry Press: Beijing, China, 2004; pp. 106–108, ISBN 978-7-80159-604-8.
3. Li, L. Research on the Protection of the Residential Buildings in Traditional Village from the Cultural Prespective: A Case of Wanjian Village in Anhui. *Urban. Archit.* **2023**, *20*, 94–98+167.
4. Banister, F. *A History of Architecture on the Comparative Method*; The MIT Press: Cambridge, MA, USA, 1922; ISBN 978-1-172-80684-3.
5. Pan, Z. Research to Traditional Civil Building and Regional Culture. *Shanxi Archit.* **2014**, *40*, 15–17.
6. Ya, K.C. Research on adaptation of building forms in geographic environment. *Shanxi Archit.* **2015**, *41*, 9–10.
7. Rossi, A. *The Architecture of the City*; Oppositions Books; The MIT Press: Cambridge, MA, USA, 1984; ISBN 978-0-262-68043-1.
8. Yang, M.C. The meaning of studying architectural semeiology on the regional architectural design. *Shanxi Archit.* **2009**, *35*, 33–34.
9. Xia, B.; Li, X.; Shi, H.; Chen, S.; Chen, J. Style Classification and Prediction of Residential Buildings Based on Machine Learning. *J. Asian Archit. Build. Eng.* **2020**, *19*, 714–730. [[CrossRef](#)]
10. Wu, Y. Classification of Ancient Buddhist Architecture in Multi-Cultural Context Based on Local Feature Learning. *Mob. Inf. Syst.* **2022**, *2022*, 8952381. [[CrossRef](#)]
11. Yan, H.; Sheng, C.; Wei, C.; Chang, Z.C. The Concept and Cultural Connotation of Traditional Villages. *Urban Dev. Stud.* **2014**, *21*, 10–13.
12. Huan, Z.L. Study on the Hollowing of Traditional Villages in Hunan Province. Master's Thesis, Hunan Normal University, Changsha, China, 2016.
13. Xue, L. Re-understanding and Evaluation of vernacular Architecture: Interpreting Architecture Without an Architect. *Architect* **2005**, *3*, 105–107.
14. Zhi, L.W. Introduction to Chinese Traditional Dwellings (Part 1). *Archit. J.* **1994**, *11*, 52–59.
15. Dun, Z.L. *Chinese Housing Overview: Traditional Residence*; Department of Philosophy and Writing, Huazhong University of Science and Technology Press: Wuhan, China, 2018; ISBN 978-7-5680-3889-8.
16. De, G.W.; Qing, Y.L.; Yong, F.W.; Zi, Q.F. The characteristic of regional differentiation and impact mechanism of architecture style of traditional residence. *J. Nat. Resour.* **2019**, *34*, 1864–1885.
17. Run, S. The natural view and origin of Chinese Traditional Dwelling culture. *Hum. Geogr.* **1997**, *3*, 29–33.
18. Pei, L.L.; Chun, L.L.; Yun, Y.D.; Xiu, Y.S.; Bo, H.L.; Zui, H. Landscape Division of Traditional Settlement and Effect Elements of Landscape Gene in China. *Acta Geogr. Sin.* **2010**, *65*, 1496–1506.
19. Grilli, E.; Remondino, F. Classification of 3D Digital Heritage. *Remote Sens.* **2019**, *11*, 847. [[CrossRef](#)]
20. Starzyńska, M.B.; Roussel, R.; Jacoby, S.; Asadipour, A. Computer Vision-Based Analysis of Buildings and Built Environments: A Systematic Review of Current Approaches 2022. *ACM Comput. Surv.* **2023**, *55*, 284–309.
21. Mathias, M.; Martinovic, A.; Weissenberg, J.; Haegler, S.; Van Gool, L. Automatic Architectural Style Recognition. The International Archives of the Photogrammetry. *Remote Sens. Spat. Inf. Sci.* **2012**, *38*, 171–176.
22. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
23. Goel, A.; Juneja, M.; Jawahar, C.V. Are Buildings Only Instances?: Exploration in Architectural Style Categories. In Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing, Mumbai, India, 16–19 December 2012; ACM: Mumbai, India, 2012; pp. 1–8.
24. Zhang, L.; Song, M.; Liu, X.; Sun, L.; Chen, C.; Bu, J. Recognizing Architecture Styles by Hierarchical Sparse Coding of Blocklets. *Inf. Sci.* **2014**, *254*, 141–154. [[CrossRef](#)]
25. Vondrick, C.; Khosla, A.; Malisiewicz, T.; Torralba, A. Hoggles: Visualizing Object Detection Features. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1–8.
26. Jiang, S.; Shao, M.; Jia, C.; Fu, Y. Learning Consensus Representation for Weak Style Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2906–2919. [[CrossRef](#)]
27. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
28. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
29. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition 2015. *arXiv* **2014**, arXiv:1409.1556.
30. 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.
31. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications 2017. *arXiv* **2017**, arXiv:1704.04861.
32. Tan, M.; Le, Q. Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
33. Mittal, U.; Chawla, P.; Tiwari, R. EnsembleNet: A Hybrid Approach for Vehicle Detection and Estimation of Traffic Density Based on Faster R-CNN and YOLO Models. *Neural Comput. Appl.* **2023**, *35*, 4755–4774. [[CrossRef](#)]

34. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 7 December 2015; pp. 1–9.
35. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
36. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
37. Siddiqi, M.H.; Khan, K.; Khan, R.U.; Alsirhani, A. Face Image Analysis Using Machine Learning: A Survey on Recent Trends and Applications. *Electronics* **2022**, *11*, 1210. [\[CrossRef\]](#)
38. Al-Masni, M.A.; Al-Antari, M.A.; Choi, M.-T.; Han, S.-M.; Kim, T.-S. Skin Lesion Segmentation in Dermoscopy Images via Deep Full Resolution Convolutional Networks. *Comput. Methods Programs Biomed.* **2018**, *162*, 221–231. [\[CrossRef\]](#)
39. Ishihara, K.; Kanervisto, A.; Miura, J.; Hautamaki, V. Multi-Task Learning with Attention for End-to-End Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2902–2911.
40. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. Available online: <https://webofscience.clarivate.cn/wos/alldb/full-record/WOS:000534424305072> (accessed on 30 December 2023).
41. Dautov, E.; Astafeva, N. Convolutional Neural Network in the Classification of Architectural Styles of Buildings. In Proceedings of the 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, 26–29 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 274–277.
42. Ji, S.Y.; Jun, H.-J. Deep Learning Model for Form Recognition and Structural Member Classification of East Asian Traditional Buildings. *Sustainability* **2020**, *12*, 5292. [\[CrossRef\]](#)
43. Gonzalez, D.; Rueda-Plata, D.; Acevedo, A.B.; Duque, J.C.; Ramos-Pollan, R.; Betancourt, A.; Garcia, S. Automatic Detection of Building Typology Using Deep Learning Methods on Street Level Images. *Build. Environ.* **2020**, *177*, 106805. [\[CrossRef\]](#)
44. Zou, H.; Ge, J.; Liu, R.; He, L. Feature Recognition of Regional Architecture Forms Based on Machine Learning: A Case Study of Architecture Heritage in Hubei Province, China. *Sustainability* **2023**, *15*, 3504. [\[CrossRef\]](#)
45. Lamas, A.; Tabik, S.; Cruz, P.; Montes, R.; Martínez-Sevilla, Á.; Cruz, T.; Herrera, F. Monu MAI: Dataset, Deep Learning Pipeline and Citizen Science Based App for Monumental Heritage Taxonomy and Classification. *Neurocomputing* **2021**, *420*, 266–280. [\[CrossRef\]](#)
46. Chun, M.Z.; Ren, S.T.; Chen, M.S.; Dang, S.Z. Research on Quantitative Measurement of Automatic Classification of Residential Buildings Under Deep Learning. *J. Southwest China Norm. Univ. (Nat. Sci. Ed.)* **2023**, *48*, 1–11.
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [\[CrossRef\]](#)
48. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding 2019. *arXiv* **2018**, arXiv:1810.04805.
49. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. Available online: <https://arxiv.org/abs/2010.11929v2> (accessed on 30 December 2023).
50. Pu, M.; Huang, Y.; Liu, Y.; Guan, Q.; Ling, H. Edter: Edge Detection with Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18 June 2022; pp. 1402–1412.
51. Li, Z.; Wang, W.; Xie, E.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P.; Lu, T. Panoptic Segformer: Delving Deeper into Panoptic Segmentation with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18 June 2022; pp. 1280–1289.
52. Zhang, Z.; Gong, Z.; Hong, Q.; Jiang, L. Swin-Transformer Based Classification for Rice Diseases Recognition. In Proceedings of the 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI), Kunming, China, 19 September 2021; pp. 153–156.
53. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10 October 2021; pp. 10012–10022.
54. Peng, B.Y.; Ji, T.S.; Biao, Z.; Yao, B.Z.; Jian, Y. A review of research on interpretability of depth models for image classification. *J. Softw.* **2023**, *34*, 230–254.
55. 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 October 2017; pp. 618–626.
56. Wen, K.H.; Fei, T.; Zi, D.W.; Li, F. Image Segmentation Based on Deep Learning: A Survey. *Comput. Sci.* **2023**, *11*, 107–116.
57. Wang, Y.; Feng, C.; Guo, C.; Chu, Y.; Hwang, J.-N. Solving the Sparsity Problem in Recommendations via Cross-Domain Item Embedding Based on Co-Clustering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, New York, NY, USA, 11–15 February 2019; ACM: Melbourne, VIC, Australia, 2019; pp. 717–725.

58. Han, M.; Kim, J. Joint Banknote Recognition and Counterfeit Detection Using Explainable Artificial Intelligence. *Sensors* **2019**, *19*, 3607. [[CrossRef](#)] [[PubMed](#)]
59. Omeiza, D.; Web, H.; Jirotko, M.; Kunze, L. Towards Accountability: Providing Intelligible Explanations in Autonomous Driving. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 11–17 July 2021; pp. 231–237.
60. Yong, Q.S.; Bao, X.Z. *Geography of Gansu Province*; Gansu Education Press: Lanzhou, China, 1990; ISBN 978-7-5423-0164-2.
61. Xue, L. The localism of Chinese regional culture and architecture. *J. Tianjin Univ. (Sci. Technol.)* **1997**, *30*, 548–554.
62. Yu, P.; Wang, F.Z.; Yong, J.Z.; Dong, D. Regional Differentiation of the Construction Monomer Plane Shape of Traditional Dwellings in Gansu Province. *Areal Res. Dev.* **2019**, *38*, 158–164.
63. Ben, T.L.; Xiao, J.Z.; Li, X.J. *Traditional Village in Gansu*; Southeast University Press: Nanjing, China, 2018; pp. 1–2.
64. Xiao, Q.G. A Geographical Study of Traditional Folk Houses in Ganqing. Ph.D. Thesis, Shaanxi Normal University, Xi'an, China, 2018.
65. Jun, N. Analysis of the Blending of Multi-ethnic Cultures in the Hexi Corridor. *J. Southwest Minzu Univ. (Humanit. Soc. Sci. Ed.)* **2018**, *39*, 34–39.
66. Wei, W. Fort Building in Hexi Corridor Area. Master's Thesis, Xi'an University of Architecture and Technology, Xi'an, China, 2010.
67. Ying, Y.H. Study on Defensive Village Settlements and Residential Buildings in Hexi Corridor Area. Master's Thesis, Xi'an University of Architecture and Technology, Xi'an, China, 2023.
68. Zhong, B.W.; Guo, X.H. *Gansu Folklore Overview*; Nationalities Publishing House: Beijing, China, 2006; ISBN 978-7-105-07577-5.
69. Ming, H.Y.; Xing, Y.L.; Xiang, W.M. Study on the Geographical Differentiation of Plane Form of Traditional Dwellings in Longnan Area. *J. Gansu Sci.* **2022**, *34*, 81–89.
70. Xiang, W.M.; Jing, L. Research on Shape Characteristic of Traditional Dwellings in Longnan County. *Tradit. Chin. Archit. Gard.* **2016**, *3*, 51–56.
71. Qiu, F.H. The Study of Ming and Qing Folk Houses in Tianshui, Gansu Province. Master's Thesis, Xi'an University of Architecture and Technology, Xi'an, China, 2006.
72. Xiang, W.M.; Ming, H.Y.; Hong, R.S. Analysis on the status quo and characteristics of traditional residential houses in Lanzhou. *Dev. Small Cities Towns* **2012**, *3*, 88–92.
73. Xiang, W.M.; Ming, H.Y. The Living Fossil of Ancient Vernacular Architecture in Northwest of China: Study on the Dwelling Architecture in Qingcheng Town, Lanzhou City in Gansu Province. *Hua Zhong Archit.* **2009**, *27*, 106–109.
74. Shan, L.; Zhang, L. Application of Intelligent Technology in Facade Style Recognition of Harbin Modern Architecture. *Sustainability* **2022**, *14*, 7073. [[CrossRef](#)]
75. Kenton, J.D.M.-W.C.; Toutanova, L.K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, p. 2.
76. Song, K.; Yang, J.; Wang, G. A Swin Transformer and MLP Based Method for Identifying Cherry Ripeness and Decay. *Front. Phys.* **2023**, *11*, 1278898. [[CrossRef](#)]
77. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21 July 2017; pp. 1492–1500.
78. Qi, Q.Z.; Zhen, L.; Ya, N.Z.; Jia, L.Z. Global-Local-Aware conditional random fields-based building extraction for high spatial resolution remote sensing images. *J. Remote Sens.* **2021**, *25*, 1422–1433.
79. Tanwar, S.; Singh, J. ResNext50 Based Convolution Neural Network-Long Short-Term Memory Model for Plant Disease Classification. *Multimed. Tools Appl.* **2023**, *82*, 29527. [[CrossRef](#)]
80. Chen, J.; Yuan, G.; Zhou, H.; Tan, C.; Yang, L.; Li, S. Classification of Solar Radio Spectrum Based on Swin Transformer. *Universe* **2023**, *9*, 9. [[CrossRef](#)]
81. Yue, C.; Fei, H. Research on Defensive Traditional Folk Houses under the Influence of Regional Culture: Taking Hexi Region of Gansu Province as an Example. *Archit. Cult.* **2023**, *4*, 235–237.
82. Yuan, W.; Zhang, X.; Shi, J.; Wang, J. LiteST-Net: A Hybrid Model of Lite Swin Transformer and Convolution for Building Extraction from Remote Sensing Image. *Remote Sens.* **2023**, *15*, 1996. [[CrossRef](#)]
83. Yu, L. Study on Rural Human Settlement Environment in Shaanxi-Gansu-Ningxia Ecologically Fragile Area. Ph.D. Thesis, Xi'an University of Architecture and Technology, Xi'an, China, 2010.
84. Xiang, W.M.; Li, Z.; Jun, W.; Yi, B.J. Study on the zoning of traditional dwellings in the multi-cultural interleaving area. *Archit. J.* **2020**, *S2*, 1–7.

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.