

Article

FGC-ConvNeXt: Frequency-Guided Collaborative Modeling of Spatial and Frequency Features for Han Stone Relief Image Classification

Hua Wei ¹, Junxiang Diao ¹, Wenlin Jin ², Hongsheng Liu ³, Zhihua Diao ^{4,*}, Lijuan Zhang ¹ and Shuang Liang ¹

¹ School of Art and Design, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 2002066@zzuli.edu.cn (H.W.); 2017013@zzuli.edu.cn (J.D.); 2009030@zzuli.edu.cn (L.Z.); 2023021@zzuli.edu.cn (S.L.)

² Jinjia Jun Kiln Co., Ltd., Yuzhou 461670, China; juncijwl@163.com

³ Hengde Jun Porcelain Co., Ltd., Yuzhou 461670, China; ryg1975111@163.com

⁴ College of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China

* Correspondence: 2010018@zzuli.edu.cn

Abstract

Han stone reliefs contain rich historical and artistic information, but their images often exhibit severe weathering, rubbing noise, and complex carving textures, making fine-grained classification challenging. To address the strong coupling between global semantic structures and local material textures, this paper proposes Frequency-Guided Collaborative ConvNeXt (FGC-ConvNeXt), a dual-branch spatial–frequency classification framework. The spatial branch extracts multi-scale semantic features using ConvNeXt-Tiny, while the frequency branch applies Fast Fourier Transform to model complementary structural and texture-sensitive information. A fixed frequency threshold of $\tau = 0.22$ is used for frequency decomposition, and both branches are projected into a 512-dimensional feature space. The Spatial–Frequency Gated Fusion module dynamically adjusts their contributions at the sample level, while frequency-domain auxiliary supervision with a loss weight of $\lambda = 0.4$ prevents branch degradation during joint optimization. Experiments were conducted on a self-constructed four-class dataset containing 6368 Han stone relief images with an input resolution of 224×224 pixels. Over three independent runs, FGC-ConvNeXt achieved an accuracy of $97.6 \pm 0.2\%$, a precision of $97.8 \pm 0.2\%$, a recall of $97.4 \pm 0.3\%$, and an F1-score of $97.6 \pm 0.2\%$, outperforming the ConvNeXt-Tiny baseline by 4.1 percentage points in accuracy. The proposed model contains 29.8 M parameters. Visualization analyses provide qualitative support for more separated feature distributions and more concentrated activation patterns, while the robustness experiments indicate improved classification stability under the evaluated conditions.

Keywords: stone relief classification; ConvNeXt; frequency domain features; Fourier transform; multi-scale attention; gated fusion

Academic Editor: Silvia Liberata Ullo

Received: 16 July 2026

Revised: 3 August 2026

Accepted: 5 August 2026

Published: 6 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 \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As an intuitive microcosm of the material civilization and spiritual world of the Han Dynasty, Han stone reliefs have become an important interdisciplinary research object connecting archaeology, art history, history, and cultural anthropology [1,2]. These relief

images, carved on the surfaces of tombs, ancestral halls, and Que towers, systematically encapsulate the social production, ritual norms, belief systems, and aesthetic trends of the Han Dynasty. Depicting diverse and rich themes that range from material production and daily social life to auspicious symbols and ascension to immortality, the completeness of their imagery and the density of their cultural information stand unique among the visual remains of the era. They provide irreplaceable physical evidence for interpreting the continuous development of Chinese civilization and hold inestimable value in academic research and the preservation of cultural heritage [3].

With the rapid iteration of high-precision 3D scanning, multispectral imaging, and digital storage technologies, the traditional research paradigm for Han stone relief images, which heavily relies on physical rubbings, is undergoing a fundamental transformation. The systematic collection and standardized processing of massive relief rubbings and original-stone photographs into digital databases have broken through the physical bottlenecks of artifact preservation and dissemination [4]. This shift has catalyzed emerging research paradigms such as “digital archaeology” and “smart museology” [5,6]. In this context, utilizing computer vision and deep learning technologies to achieve efficient and precise automatic classification of massive Han stone relief images has become a core issue in solving the paradox of “high data volume, low analysis efficiency” in cultural heritage digitalization [7], emerging as a frontier hotspot in digital humanities and intelligent archaeology [8].

Existing classification methods for Han stone relief images primarily adopt mature deep convolutional neural networks (DCNNs) from the natural image domain, such as ResNet, YOLO-based unified frameworks, and purely spatial domain modeling methods like ConvNeXt [9–11]. While these methods perform excellently on natural image datasets by relying on the synergistic distribution of color, texture, and semantic information, Han stone relief images present unique challenges. They are affected by superimposed factors including stone textures, carving techniques (e.g., line engraving, low relief), rubbing processes, and preservation environments [12]. Consequently, they exhibit a distinct dual visual characteristic of “explicit semantic structure and complex material texture”. At the macroscopic level, low-frequency components contain global semantic compositional information—such as architectural outlines and character postures—forming the core basis for categorization [13,14]. At the microscopic level, high-frequency components—composed of granular stone textures, carving blade marks, and ink density variations—are crucial for distinguishing similar categories [15]. Repeated convolution and downsampling may reduce the representation of some fine-scale texture cues, particularly when these cues are weak relative to dominant semantic structures. This may limit the discrimination of visually similar Han stone relief categories, although the extent of this effect depends on the network architecture and training strategy [16].

To overcome the limitations of pure spatial domain modeling, researchers have introduced frequency-domain analysis into visual tasks to mine the discriminative value of different frequency components [17]. For instance, FNet achieves efficient global information modeling using Fourier transforms, while frequency attention mechanisms enhance sensitivity to fine-grained differences [18,19]. Furthermore, introducing high-frequency feature supervision has been proven to significantly strengthen the ability to identify minute structures in texture-sensitive tasks like medical and remote sensing imaging [20]. Despite these advantages, existing frequency methods suffer from three core deficiencies: first, a lack of targeted decoupling modeling for the inherent “structure-texture” separation characteristics of Han stone relief images; second, static feature fusion methods that fail to dynamically adapt to the varying visual demands of different categories (e.g., architectural images rely more on structural information, whereas motifs depend on texture details); and third, the vulnerability of the frequency branch in dual-stream networks

to be easily suppressed by the main spatial branch, failing to fully exploit the discriminative value of high-frequency information [21].

Recent studies have further demonstrated the value of frequency-domain modeling in visual recognition. FcaNet introduces multi-spectral frequency attention [22], while GFNet and AFNO perform efficient global feature interaction in the Fourier domain [23,24]. Wavelet-based visual models also preserve high-frequency details that may be lost during spatial downsampling [25]. In fine-grained recognition, methods such as TransFG and progressive multi-granularity learning improve discrimination by emphasizing subtle local differences [26]. Deep learning has also been increasingly applied to artwork and archaeological image classification. However, most existing methods are designed for natural images or general cultural heritage analysis and do not explicitly address the strong coupling between global semantic structures and material-sensitive textures in Han stone relief images. This gap motivates the proposed task-oriented spatial–frequency collaborative framework [27].

To address these issues, this paper proposes FGC-ConvNeXt (Frequency-Guided Collaborative ConvNeXt), a task-oriented spatial–frequency collaborative ConvNeXt framework for Han stone relief image classification [28]. Rather than claiming novelty from the isolated use of ConvNeXt, FFT, gated fusion, or auxiliary supervision, the proposed method focuses on the coordinated adaptation of these components to the unique visual characteristics of Han stone reliefs. Different from general spatial–frequency fusion methods, FGC-ConvNeXt is designed around the strong coupling between macroscopic semantic structures and microscopic rubbing textures in Han stone relief images. The spatial branch is responsible for modeling global semantic layouts such as architectural outlines and human postures, whereas the frequency branch explicitly captures high-frequency cues such as carving marks, stone granularity, rubbing noise, and ink-density variations [29]. Moreover, the Spatial–Frequency Gated Fusion module adaptively adjusts the contribution of the two branches at the sample level, and the frequency-domain auxiliary supervision prevents the frequency branch from being weakened during joint optimization. Therefore, the contribution of this work lies not in proposing a completely new basic vision operator, but in constructing a task-specific and interpretable spatial–frequency collaborative framework for fine-grained cultural heritage image classification.

2. Materials and Methods

2.1. Dataset Collection and Construction

The dataset contains 6368 Han stone relief images collected from published academic image materials and publicly accessible cultural heritage image archives. These resources include digitized relief rubbings and artifact images from different regions of China, covering diverse themes, carving styles, and visual compositions. During dataset construction, the images were digitized, cropped, renamed, deduplicated, and reorganized into a unified database. Consequently, complete title-level source records, source-wise image counts, and artifact- or site-level metadata cannot be reliably reconstructed for every image. To avoid reporting potentially inaccurate provenance information, only the verified overall and category-wise statistics are provided in this study. In this study, the term “Han stone relief image” is used as a general term covering both digitized rubbing images and photographs of original stone reliefs. The two image modalities were included in the same dataset and processed using the same preprocessing and classification protocol. They were mixed during model training rather than being used to train separate modality-specific models, and modality labels were not used as supervisory information.

All images were used solely for non-commercial academic research. Copyright and related rights of the original images remain with the corresponding publishers, institutions, or other rights holders. The authors do not claim ownership of the original visual

materials and do not possess unrestricted rights to redistribute the complete raw-image dataset.

During preprocessing, non-relief regions, including captions, page borders, background layouts, textual annotations, and other irrelevant visual elements, were removed to ensure that the retained regions focused only on effective relief content [30,31]. Dataset quality control was then conducted through rule-based manual visual inspection rather than unrestricted subjective selection. Images with severe blur, substantial occlusion, overexposure, incomplete relief content, duplicated visual content, or large irrelevant regions were excluded according to predefined criteria. Near-duplicate samples were further checked using visual similarity, source-page information, and filename records.

Each image was assigned a single image-level label according to its dominant thematic content. Images dominated by buildings or architectural layouts were labeled as Architecture; images mainly depicting human activities were labeled as Human Figures; images dominated by animals, mythical creatures, or symbolic motifs were labeled as Auspicious Symbols; and repetitive geometric, floral, or non-narrative motifs were labeled as Decorative Patterns. For images containing multiple types of content, the dominant subject, occupied visual area, narrative focus, and available descriptions in the source materials were considered jointly. For example, an image containing both people and buildings was labeled as Architecture when the architectural composition dominated the scene, but as Human Figures when human activities formed the main narrative content.

Two researchers independently annotated all images. Inconsistent labels were jointly reviewed and resolved through discussion and cross-checking with relevant archaeological, art-historical, and source descriptions. Only image-level category labels were created; no pixel-level segmentation masks or object-contour annotations were included.

Each image was manually labeled into one of four categories based on its visual content and thematic characteristics: architecture, human figures, auspicious symbols, and decorative patterns, as illustrated in Figure 1. The labeling process was conducted with reference to textual descriptions in the source literature and validated through multiple rounds of cross-checking to ensure consistency and accuracy.



Figure 1. Example images of the four categories in Han stone relief image dataset.

The final dataset consists of 1552 architecture images, 1686 human figure images, 1547 auspicious symbol images, and 1583 pattern images, forming a well-balanced and representative classification benchmark. Visually similar but non-duplicate samples from different categories were retained as challenging cases, while near-duplicate images were grouped within the same subset to reduce data leakage. To enhance model generalization

and robustness, various data augmentation techniques were applied during training, including random cropping, horizontal flipping, rotation, and brightness/contrast adjustment.

To reduce the risk of potential data leakage, we further introduced a source-aware and near-duplicate-controlled splitting procedure. Since some historical Han stone relief images do not contain complete provenance metadata, such as exact artifact ID, rubbing ID, plate number, excavation site, or acquisition record, it is difficult to guarantee a fully strict artifact-level or site-level split for every sample. Therefore, we adopted a conservative leakage-control strategy based on the available provenance information and manual visual inspection.

For samples with reliable metadata, images from the same publication, plate/page, artifact record, rubbing source, or archaeological site were assigned to the same group. Samples belonging to the same group were placed exclusively in one subset, namely the training set, validation set, or test set. For samples with incomplete provenance information, we further performed near-duplicate screening by jointly considering visual similarity, source page information, filename records, and manual checking. Images with highly similar visual content or suspected common artifact/rubbing origins were grouped together whenever possible before dataset splitting.

It should be noted that the dataset split in this study is not a simple augmented-image-level random split. Data augmentation was conducted only after the original images had been divided into training, validation, and test sets, and it was applied exclusively to the training set. The validation and test sets retained their original images and did not participate in any augmentation operation. This procedure prevents augmented variants of the same original image from appearing across different subsets and reduces the risk of inflated test performance caused by near-duplicate leakage.

Rather than applying a simple image-level random split, source groups were constructed before dataset partitioning. Images associated with the same identifiable publication, plate or page, artifact record, rubbing source, or archaeological site were assigned to the same group, and each group was placed exclusively in the training, validation, or test set. For samples with incomplete provenance information, visual similarity, source-page information, filename records, and manual inspection were jointly used to identify suspected near-duplicates or common artifact/rubbing origins. The resulting source groups were divided into training, validation, and test sets at an approximate ratio of 7:2:1. The final split contained 4458 training images, 1274 validation images, and 636 test images. Group-level stratification was applied to maintain a comparable category distribution across the three subsets. Because complete artifact-level and site-level metadata were unavailable for some historical images, this procedure reduces, but cannot completely eliminate, the risk of source-level data leakage. The main details of dataset construction, quality control, annotation, and splitting are summarized in Table 1.

Table 1. Summary of dataset construction and quality-control procedures.

Item	Description
Data sources	Published academic materials and curated cultural heritage image resources
Dataset size	6368 images in four categories
Class distribution	Architecture: 1552; Human Figures: 1686; Auspicious Symbols: 1547; Decorative Patterns: 1583
Quality control	Rule-based manual visual inspection
Exclusion criteria	Severe blur, occlusion, overexposure, incomplete content, duplicates, and large irrelevant regions

Annotation	Independent annotation by two researchers using the dominant-theme rule
Annotation agreement	Inconsistent labels were jointly reviewed and resolved by consensus
Provenance limitation	Complete source-wise and artifact- or site-level metadata were unavailable for some historical images
Dataset split	Source-aware, group-level stratified training/validation/test split containing 4458/1274/636 images, respectively, corresponding approximately to a 7:2:1 ratio.
Leakage control	Images from the same identifiable publication, page, artifact, rubbing source, or site were assigned to one subset; additional near-duplicate screening was performed for samples with incomplete provenance.
Data availability	Raw images are not publicly redistributed; labels, split files, scripts, and model configurations are available under appropriate conditions

2.2. Proposed Method

This paper proposes a task-oriented spatial–frequency collaborative framework (FGC-ConvNeXt) to address the challenge of strong coupling between macroscopic semantic structures and microscopic rubbing textures in Han stone relief image classification. The contribution of FGC-ConvNeXt does not lie in introducing new basic operators, but in coordinating an asymmetric spatial–frequency dual-branch design, sample-adaptive gated fusion, and frequency-branch auxiliary supervision for the structure–texture characteristics of Han stone relief images. As illustrated in Figure 2, the proposed network mainly consists of a dual-stream architecture with parallel feature extraction branches (i.e., a spatial branch and a frequency-domain branch), an adaptive spatial–frequency gated fusion module (SFGF), and a joint loss function with frequency-domain auxiliary supervision.

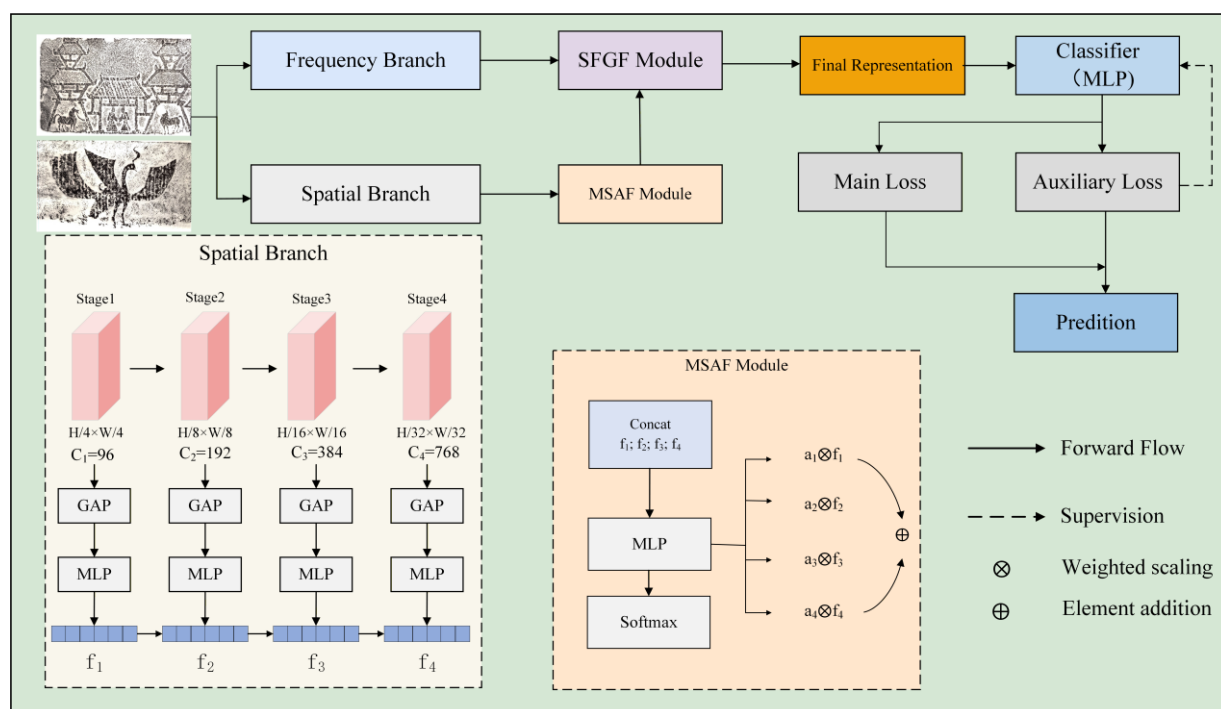


Figure 2. Overall architecture of FGC-ConvNeXt.

2.2.1. Spatial Feature Extraction Branch Based on ConvNeXt

The spatial branch is designed to extract deep geometric structures and global semantic information from Han stone relief images, such as building contours and human poses. Due to the high complexity of these images, this paper adopts the lightweight and efficient ConvNeXt-Tiny as the backbone network to balance computational cost and feature extraction capability.

Given an input Han stone relief image, it is fed into the ConvNeXt network. To fully capture structural information at different receptive fields, we extract feature maps from the four stages of the network (Stage 1 to Stage 4). Let the feature map extracted from the i -th stage be S_i . To eliminate scale discrepancies and achieve feature alignment, adaptive average pooling is applied to each feature map, followed by independent linear projection layers to reduce the dimensionality to a unified size d . Subsequently, a multi-scale feature aggregation strategy based on an attention mechanism is employed to weight and fuse these features. Finally, a refinement module is used to produce the spatial feature vector. In this study, the feature dimension d is set to 512.

Given an input image $X \in R^{B \times 3 \times 224 \times 224}$, the four stages of ConvNeXt-Tiny produce feature maps $S_i \in R^{B \times C_i \times H_i \times W_i}$, where $(C_1, C_2, C_3, C_4) = (96, 192, 384, 768)$ and the corresponding spatial resolutions are 56×56 , 28×28 , 14×14 , and 7×7 , respectively. Each feature map is converted into a channel descriptor using global adaptive average pooling and independently projected into a 512-dimensional feature space:

$$Z_i = GELU(LN(W_i GAP(S_i) + b_i)), Z_i \in R^{B \times 512} \quad (1)$$

The importance of each stage is calculated using an attention-based aggregation mechanism:

$$\alpha_i = \frac{\exp(w_a^T Z_j)}{\sum_{j=1}^4 \exp(w_a^T Z_j)}, Z = \sum_{i=1}^4 \alpha_i Z_i \quad (2)$$

The aggregated feature is subsequently processed by a refinement layer consisting of a 512-dimensional linear layer, Layer Normalization, GELU activation, and dropout with a rate of 0.2:

$$F_s = Dropout(GELU(LN(W_r Z + b_r))) \quad (3)$$

where $F_s \in R^{B \times 512}$ denotes the final spatial feature.

2.2.2. Branch for Frequency-Domain Feature Decoupling Based on Fast Fourier Transform

Traditional convolutional neural networks tend to learn low-frequency smooth regions while frequently ignoring unique textures on Han stone reliefs, including weathered mottling, rubbing ink traces and delicate carving marks left by ancient craftsmen. To explicitly capture such high-frequency texture information, this paper proposes a frequency-domain feature decoupling branch based on Fast Fourier Transform (FFT) [32] (Figure 3).

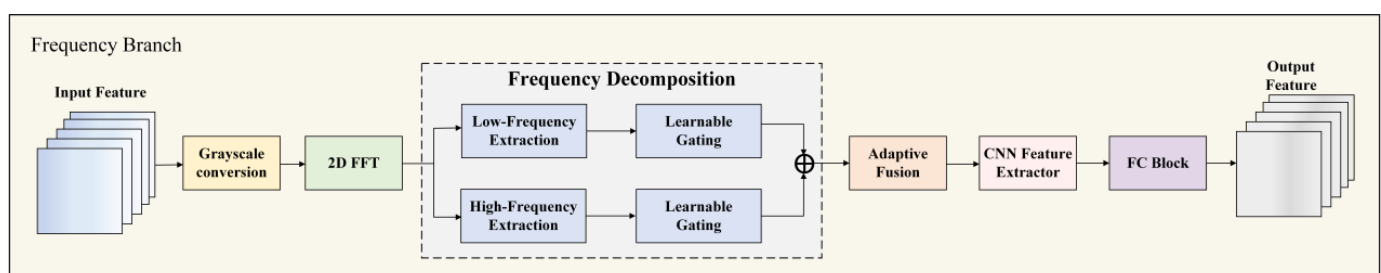


Figure 3. Structure of the frequency-domain feature decoupling branch.

First, the input image X is converted into a single-channel grayscale image. This design is based on the visual characteristics of Han stone relief images, whose discriminative cues mainly arise from intensity-based carving contours, rubbing textures, stone granularity, weathering traces, and ink-density variations rather than rich chromatic differences. Importantly, grayscale conversion removes inter-channel chromatic redundancy but does not discard luminance stratification. Intensity variations associated with rubbing ink, surface weathering, carving depth, and material degradation are retained in the grayscale signal and subsequently represented in the frequency amplitude spectrum. Therefore, grayscale frequency decomposition preserves the principal texture-sensitive cues while reducing the computational and representational redundancy introduced by channel-wise RGB decomposition.

In implementation, the input image is resized to 224×224 pixels, and grayscale conversion is performed using $X_{gray} = 0.299 X_R + 0.587 X_G + 0.114 X_B$. A centered 2D FFT is then applied, and the logarithmic amplitude spectrum is calculated as $M = \log(1 + |F|)$. A two-dimensional discrete Fourier transform (2D DFT) is utilized to transform it from the spatial domain to the frequency domain, and a centering operation is applied to shift the zero-frequency component to the center of the spectrum. The formula for calculating the logarithmic amplitude spectrum $M(u, v)$ is as follows:

$$M(u, v) = \log \left(1 + \left| \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} X_{gray}(x, y) e^{-j2\pi(\frac{ux}{H} + \frac{vy}{W})} \right| \right) \quad (4)$$

The amplitude spectrum is adopted as the primary frequency-domain representation because it directly reflects the energy distribution of different frequency components. The amplitude spectrum is used because it directly represents the energy distribution of different frequency components. We acknowledge that Fourier phase contains important spatial and structural information and should not be regarded as uninformative. In the present framework, phase is omitted mainly to maintain a lightweight frequency branch and to reduce representational overlap with the spatial branch, which already models semantic structure and spatial layout.

To separate structural and textural information in the frequency domain, a normalized radial distance matrix is constructed as:

$$D(u, v) = \sqrt{\bar{u}^2 + \bar{v}^2}, \bar{u}, \bar{v} \in [-1, 1] \quad (5)$$

A fixed threshold-based hard mask, rather than an adaptive mask, is then defined to separate the amplitude spectrum into low-frequency and high-frequency components. Given the normalized frequency distance $D(u, v)$, the low-frequency mask and high-frequency mask are defined as follows:

$$Mask_{low}(u, v) = \begin{cases} 1, & D(u, v) \leq \tau \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

$$Mask_{high}(u, v) = 1 - Mask_{low}(u, v) \quad (7)$$

where τ denotes a fixed frequency threshold used to control the separation between low-frequency structural information and high-frequency texture information. The threshold τ remains unchanged during both training and inference and does not contain learnable parameters. In this study, τ is set to 0.22 according to the sensitivity analysis reported in the ablation experiments. It should be noted that the frequency mask itself is threshold-based rather than adaptive. The adaptive modeling ability of FGC-ConvNeXt is mainly provided by the subsequent Spatial-Frequency Gated Fusion (SFGF) module, which dynamically adjusts the contribution of spatial and frequency-domain features at the sample

level. The low- and high-frequency components are concatenated along the channel dimension to form a two-channel tensor. This tensor is processed by four convolutional blocks with output channels of 32, 64, 128, and 256. Each block contains a 3×3 convolution with stride 1 and padding 1, followed by Batch Normalization, GELU activation, and 2×2 max pooling. Global average pooling, linear projection, and dropout are finally applied to obtain a 512-dimensional frequency-domain feature vector.

2.2.3. Adaptive Spatial–Frequency Gated Fusion Module (SFGF)

Different types of Han stone reliefs rely on spatial and frequency-domain information to varying degrees. For instance, the “architecture” category is more dependent on spatial structures, whereas the “pattern” category leans heavily on high-frequency features in the frequency domain. Simply adopting feature Concatenation or Addition is highly likely to result in information overlap and mutual interference among heterogeneous features. To address this issue, this paper proposes an adaptive Spatial–Frequency Gated Fusion (SFGF) module [33].

First, the spatial feature F_s and frequency-domain feature F_f are concatenated along the channel dimension. The concatenated features are then fed into a multi-layer perceptron (MLP) consisting of two linear layers and Layer Normalization, followed by the Sigmoid activation function to generate the gating weight vector G .

$$G = \sigma(W_2(\text{GELU}(\text{LN}(W_1[F_s, F_f]))) \quad (8)$$

where $[\cdot, \cdot]$ denotes the feature concatenation operation; W_1 and W_2 are learnable weight matrices, and σ stands for the Sigmoid activation function, the generated gating vector $G \in (0,1)^d$.

Before fusion, the spatial feature F_s and the frequency-domain feature F_f are projected into the same d -dimensional feature space. In this study, d is set to 512. Therefore, $F_s \in R^d$ and $F_f \in R^d$ for a single sample. For mini-batch training, the corresponding tensor dimensions are $F_s, F_f \in R^{B \times d}$, where B denotes the batch size. The gating vector G generated by the MLP has the same dimensionality as the two feature vectors, i.e., $G \in (0,1)^d$ for a single sample and $G \in (0,1)^{B \times d}$ for a mini-batch.

$$F_{fused} = G \odot F_s + (1 - G) \odot F_f \quad (9)$$

where $\odot$ denotes Hadamard element-wise multiplication. Specifically, each element G_i independently controls the fusion weight of the i -th feature dimension. When G_i approaches 1, the fused representation relies more on the spatial feature $F_{s,i}$; when G_i approaches 0, it relies more on the frequency-domain feature $F_{f,i}$. Therefore, the SFGF module performs dimension-wise adaptive soft selection between spatial and frequency-domain features.

For visualization, the 512-dimensional gating vector of each sample is reduced to a scalar by averaging across the feature dimensions:

$$g_n = \frac{1}{512} \sum_{i=1}^{512} G_{n,i} \quad (10)$$

where g_n denotes the sample-level mean gating weight. The reduction is first performed across the 512 feature dimensions for each sample, and the resulting scalar values are then grouped according to the ground-truth categories for statistical visualization.

Although spatial–frequency fusion has been explored in previous computer vision studies, most existing methods are designed for natural image classification, restoration, or enhancement tasks, where frequency information is usually used as a general complementary representation. In contrast, the proposed FGC-ConvNeXt is specifically designed

for Han stone relief images, whose classification depends on both semantic composition and subtle material/rubbing textures. Therefore, the two branches in FGC-ConvNeXt are not treated as symmetric feature extractors. The spatial branch focuses on category-level semantic structures, while the frequency branch is constrained to learn texture-sensitive discriminative cues. Furthermore, the proposed gated fusion mechanism performs adaptive feature selection at the sample level, and the auxiliary supervision explicitly maintains the effectiveness of the frequency branch. These designs make the proposed framework different from simple spatial–frequency concatenation, addition, or generic gated reweighting strategies.

2.2.4. Frequency-Domain Auxiliary Supervision and Joint Loss Function

During the joint training of multimodal or two-stream networks, the ConvNeXt-based spatial branch possesses strong fitting capacity and tends to dominate gradient backpropagation at the early training stage. Consequently, the frequency branch fails to be fully optimized and degenerates into an ineffective dormant branch.

To force the frequency branch to learn high-frequency texture features beneficial for the classification task, an independent Auxiliary Classifier is appended to the end of the frequency branch. During training, the network yields not only primary classification results from fused features $\hat{Y}_{main}$ but also auxiliary predictions derived solely from frequency-domain features $\hat{Y}_{freq}$.

This paper adopts the standard Cross-Entropy Loss function:

$$L_{CE}(Y, \hat{Y}) = - \sum_{c=1}^C Y_c \log(\hat{Y}_c) \quad (11)$$

where Y denotes the one-hot ground-truth label of samples, and C is the total number of categories. To achieve stable end-to-end optimization, the following joint loss function is formulated in this work:

$$L_{total} = L_{CE}(Y, \hat{Y}_{main}) + \lambda L_{CE}(Y, \hat{Y}_{freq}) \quad (12)$$

where λ is the balance factor for auxiliary loss. With this design, the frequency branch receives not only joint gradients backpropagated from the fusion module but also direct supervision signals from the auxiliary classifier, which effectively prevents gradient starvation. In our experimental setup, λ is empirically set to 0.4. During testing and inference, the auxiliary classifier is discarded without introducing extra computational overhead.

2.3. Evaluation Metrics

To quantitatively evaluate the classification performance of the proposed FGC-ConvNeXt model, four commonly used metrics are adopted in this study, including Accuracy, Precision, Recall, and F1-score [34]. These metrics are calculated based on the confusion matrix and can comprehensively reflect the overall recognition performance and category-level classification ability of the model.

For the multi-class classification task, Accuracy is calculated as:

$$Accuracy = \frac{\sum_{k=1}^C TP_k}{N} \quad (13)$$

where C denotes the number of categories, TP_k is the number of correctly classified samples in the i -th category, and N is the total number of test samples.

Precision, Recall, and F1-score are calculated using macro-averaged definitions:

$$\text{Precision} = \frac{1}{C} \sum_{i=1}^C \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \quad (14)$$

$$\text{Recall} = \frac{1}{C} \sum_{i=1}^C \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \quad (15)$$

$$\text{F1} = \frac{1}{C} \sum_{i=1}^C \frac{2\text{P}_i\text{R}_i}{\text{P}_i + \text{R}_i} \quad (16)$$

where TP_i , FP_i , and FN_i denote the number of true positive, false positive, and false negative samples for the i -th category, respectively. P_i and R_i represent the precision and recall of the i -th category. Macro-averaged metrics are adopted to ensure that each category contributes equally to the final evaluation.

3. Experiment and Result Analysis

3.1. Experimental Environment and Training Setup

The training environment is a workstation equipped with an NVIDIA RTX 4090 GPU, running on the Windows operating system. The software configuration includes CUDA 11.8, Python 3.8, and PyTorch 2.0.0. To ensure fairness and reproducibility, all experiments in this study are conducted under a unified data processing protocol.

All images were first resized to 256×256 pixels using bicubic interpolation. During training, a fixed-size 224×224 crop was randomly sampled from each resized image, corresponding to 87.5% of the image side length and approximately 76.6% of the image area; no variable random-resized-crop scale was used. Random horizontal flipping was applied with a probability of 0.5. Random rotation was performed within $[-10^\circ, 10^\circ]$ using bilinear interpolation. Brightness and contrast factors were independently sampled from $[0.8, 1.2]$. The augmented images were subsequently normalized using the ImageNet channel-wise mean and standard deviation. During validation and testing, images were resized to 256×256 pixels using bicubic interpolation and center-cropped to 224×224 pixels without random augmentation. Because the four categories have a relatively balanced sample distribution, standard shuffled mini-batch sampling was used, and no class-balanced sampler or class weighting was applied. The cross-entropy loss was calculated without label smoothing, i.e., the label-smoothing coefficient was set to 0.

To prevent GPU memory overflow, the batch size for all models is set to 16. In addition, due to the limited scale of the Han stone relief image dataset, which is insufficient to train large deep networks from scratch to full convergence, the backbone of the spatial branch (ConvNeXt-Tiny) is initialized with pretrained weights from the ImageNet dataset. All models are trained for 100 epochs, and a cosine annealing learning rate scheduler is employed to achieve smooth decay of the learning rate.

In terms of optimization strategy, the AdamW optimizer with decoupled weight decay is adopted. Considering the dual-stream architecture, a differential learning rate scheme is applied: a smaller initial learning rate is assigned to the pretrained backbone to preserve its low-level feature extraction capability, while relatively larger learning rates are used for the classification head and the frequency-domain decoupling branch to facilitate rapid adaptation to the specific frequency-domain texture characteristics of Han stone relief images. Furthermore, to prevent gradient explosion in the early stages of training, gradient clipping with a maximum norm of 5.0 is applied. To improve statistical reliability, each model was independently trained three times using different random seeds, namely 2023, 2024, and 2025. The reported results are presented as the mean $\pm$ standard deviation over three independent runs using random seeds 2023, 2024, and 2025. The

same random seeds were used for the compared models to maintain matched experimental conditions. Because only three runs were conducted, the results are interpreted descriptively rather than through formal statistical hypothesis testing. The individual seed-level results are provided in Appendix A.

For all comparison and ablation experiments, the weighting coefficient of the frequency-domain auxiliary supervision loss is strictly fixed. The detailed hyperparameter settings of the proposed model (FGC-ConvNeXt) are summarized in Table 2. The frequency threshold τ and auxiliary-loss weight λ were selected exclusively according to the macro-F1 score on the validation set. The test set was not used for hyperparameter selection and was evaluated only after the final configuration had been determined.

Table 2. Core hyperparameter configurations for FGC-ConvNeXt.

Category	Hyperparameter	Value
Optimization	Optimizer	AdamW
	Base LR (ConvNeXt Backbone)	5×10^{-5}
	Head LR (Frequency & Classifier)	5×10^{-4}
	LR Scheduler	Cosine Annealing
	Weight Decay	1×10^{-4}
Training Setup	Gradient Clip (Max Norm)	5
	Batch Size	16
	Total Epochs	100
	Aux Loss Weight (λ)	0.4
	Input Resolution	224×224
Model Configuration	Spatial Feature Dimension	512
	Frequency Input	Grayscale image
	Spectrum Processing	Centered 2D FFT + log amplitude
	Frequency Separation	Radial mask, $\tau = 0.22$
	Frequency Encoder	Four Conv–BN–GELU–Max–Pool blocks
	Encoder Channels	32–64–128–256
	Frequency Output	GAP + Linear projection to 512
	Dropout Rate	0.2
	Number of Classes	4

3.2. Ablation Experiment Analysis

To provide a more concise and comprehensive evaluation of the proposed FGC-ConvNeXt framework, we consolidate the relevant analyses into a unified table and present the ablation study in a simplified form. In addition to classification accuracy, recall and F1-score are also reported to offer a more complete assessment of model performance.

To evaluate the contributions of the frequency branch, fusion strategy, and auxiliary supervision, controlled ablation experiments were conducted under the same dataset split and training settings. The results are presented in Table 3.

Table 3. Ablation study of the core components and fusion strategies of FGC-ConvNeXt.

Model Variant	MS Attention	Refinement	Freq. Branch	Fusion Method	Aux. Loss	Params (M)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline	×	×	×	—	×	28.6	93.5	93.4	93.2	93.1
Variant A	✓	✓	✓	Concatenation	×	29.6	95.1	95	94.8	94.9
Variant B	✓	✓	✓	Addition	×	29.4	94.7	94.6	94.5	94.5

Variant C	✓	✓	✓	SFGF	×	29.8	96.5	96.4	96.2	96.1
Ours	✓	✓	✓	SFGF	✓	29.8	97.6	97.8	97.4	97.6

Note: ✓ indicates that the component is included; × indicates that it is not included; — indicates not applicable.

As shown in Table 3, introducing the frequency branch improves model performance under all evaluated fusion strategies, indicating that frequency-domain features provide complementary information to the spatial representation. Among the fusion methods, element-wise addition obtains the lowest performance, while concatenation preserves more information and achieves better results. SFGF obtains the best performance among the evaluated fusion strategies by performing sample-dependent, dimension-wise adaptive fusion. Furthermore, introducing frequency-domain auxiliary supervision increases the accuracy and F1-score to 97.6%, demonstrating the benefit of directly constraining the frequency branch during joint optimization. Element-wise addition and feature concatenation serve as static fusion controls, whereas SFGF performs adaptive vector-level fusion. Scalar gating and static channel-wise gating were not exhaustively evaluated and will be investigated in future work.

Table 4 evaluates the impact of frequency components and auxiliary supervision. The high-frequency branch consistently outperforms the low-frequency one, indicating that fine-grained texture features are more discriminative for stone relief classification. Combining both components yields the best performance, demonstrating their complementary roles in capturing global structure and local details. As the auxiliary weight λ increases from 0 to 0.4, the validation performance initially improves, indicating that moderate auxiliary supervision benefits frequency-branch optimization. The best evaluated result is obtained at $\lambda = 0.4$. Further increasing λ reduces performance, suggesting that excessive auxiliary weighting may disturb the balance between the main and auxiliary objectives.

Table 4. Validation-set analysis of frequency components and auxiliary-loss weights.

Setting	Configuration	Accuracy (%)	Recall (%)	F1-Score (%)
Low Frequency Only	LF	94.8	94.5	94.6
High Frequency Only	HF	95.3	95	95.1
Both (Ours)	LF + HF	97.6	97.4	97.6
$\lambda = 0.0$	No Aux	96.5	96.2	96.1
$\lambda = 0.2$	Weak Aux	97.1	96.9	96.8
$\lambda = 0.4$	Optimal	97.6	97.4	97.6
$\lambda = 0.6$	Strong Aux	96.8	96.5	96.4
$\lambda = 0.8$	Over-regularized	95.9	95.6	95.5

Note: The results are obtained from the representative run with random seed 2024. The “LF + HF (Both/Ours)” and “ $\lambda = 0.4$ ” rows correspond to the same final configuration and therefore report identical metrics.

To further justify the design of the frequency-domain branch, we conducted a concise analysis of the frequency threshold and representative frequency representations. As shown in Table 5, $\tau = 0.22$ achieves the best performance. When τ is smaller, excessive high-frequency noise is introduced; when τ is larger, useful mid-frequency rubbing and carving details are partially suppressed. Therefore, $\tau = 0.22$ is adopted in the final model.

Table 5. Validation-set sensitivity analysis of the frequency threshold τ .

Threshold τ	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
0.14	96.5 ± 0.3	96.3 ± 0.3	96.7 ± 0.3	96.5 ± 0.3
0.18	97.1 ± 0.2	96.9 ± 0.3	97.3 ± 0.2	97.1 ± 0.2

0.22	97.6 ± 0.2	97.4 ± 0.3	97.8 ± 0.2	97.6 ± 0.2
0.26	97.2 ± 0.3	97.0 ± 0.3	97.4 ± 0.3	97.2 ± 0.3
0.3	96.7 ± 0.3	96.5 ± 0.4	96.9 ± 0.3	96.7 ± 0.3

In addition, Table 6 compares several representative frequency representations. Among the frequency representations evaluated under the current experimental setting, the grayscale FFT amplitude spectrum achieves the highest accuracy and F1-score. This indicates that the main discriminative cues in Han stone relief images are more closely related to intensity-based rubbing textures and carving traces than to RGB channel differences. Although phase-aware and learnable frequency representations may further improve modeling flexibility, they also introduce additional complexity and are left for future work.

Table 6. Comparison of representative frequency representations.

Frequency Representation	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Params (M)
Grayscale FFT amplitude spectrum	97.6 ± 0.2	97.4 ± 0.3	97.8 ± 0.2	97.6 ± 0.2	29.8
RGB-channel FFT amplitude spectrum	97.2 ± 0.3	97.0 ± 0.3	97.4 ± 0.3	97.2 ± 0.3	30.4
DCT-based frequency representation	94.4 ± 0.4	94.2 ± 0.4	94.6 ± 0.4	94.3 ± 0.4	29
DWT-based frequency representation	95.2 ± 0.3	95.0 ± 0.3	95.4 ± 0.3	95.1 ± 0.3	29.1

Because the radial coordinates are normalized to $[-1, 1]$, τ is defined relative to the spectrum radius rather than in absolute pixels. This reduces its direct dependence on image resolution under the fixed 224×224 input setting. Nevertheless, the optimal threshold may vary with image resolution, object scale, and acquisition conditions. Soft or learnable masks, amplitude–phase joint representations, and localized frequency modeling were not exhaustively evaluated in this study and will be investigated in future work.

To further examine the influence of several implementation choices, additional controlled experiments were conducted on spectrum processing, frequency-encoder architecture, projection dimension, and backbone initialization. In each comparison, only the investigated setting was changed, while the dataset split, training schedule, optimizer, fusion module, and auxiliary-loss weight were kept unchanged. The results are presented in Table 7.

Table 7. Additional sensitivity analysis of implementation choices.

Investigated Factor	Setting	Params (M)	FLOPs (G)	Latency (ms)	Accuracy (%)	F1-Score (%)
Spectrum Processing	Raw amplitude	29.8	6.1	5.85	96.9 ± 0.3	96.9 ± 0.3
	Log amplitude (Ours)	29.8	6.1	5.88	97.6 ± 0.2	97.6 ± 0.2
Frequency Encoder	GAP + Linear projection	28.9	4.8	4.97	95.8 ± 0.3	95.7 ± 0.3
	Four-block CNN (Ours)	29.8	6.1	5.88	97.6 ± 0.2	97.6 ± 0.2
Projection Dimension	256	29.2	5.9	5.62	97.0 ± 0.3	97.0 ± 0.3
	512 (Ours)	29.8	6.1	5.88	97.6 ± 0.2	97.6 ± 0.2
	768	30.5	6.4	6.14	97.3 ± 0.3	97.3 ± 0.3

Backbone Initializa- tion	Training from scratch	29.8	6.1	5.88	94.8 ± 0.4	94.7 ± 0.4
	ImageNet pretrain- ing (Ours)	29.8	6.1	5.88	97.6 ± 0.2	97.6 ± 0.2

As shown in Table 7, the log-amplitude transform outperforms the direct use of raw amplitude, indicating that logarithmic compression helps reduce the influence of extremely large spectral responses. Replacing the four-block frequency CNN with global average pooling and linear projection causes a clear performance decrease, demonstrating the importance of local frequency-pattern modeling. A projection dimension of 512 achieves the best trade-off between representation capacity and computational cost; reducing the dimension to 256 limits feature representation, whereas increasing it to 768 introduces additional complexity without further improvement. ImageNet initialization also provides a clear advantage over training from scratch, which is particularly beneficial under the relatively limited scale of the current dataset.

3.3. Comparative Experiments

To comprehensively and objectively evaluate the practical performance and robustness of the proposed FGC-ConvNeXt network in Han stone relief image classification, we compared it with seven representative mainstream image classification models under the same hardware and software platform, dataset split, and basic data preprocessing settings. The compared baseline models include the lightweight convolutional network EfficientNet-B0, the densely connected DenseNet-121, the classical deep convolutional network ResNet-50, SE-ResNet-50 with channel attention, Swin-T based on hierarchical Vision Transformer, ViT-Base with global self-attention, and ConvNeXt-Tiny, which serves as the pure spatial-domain baseline model of the proposed method. Tables 8 and 9 include representative externally published CNN, Transformer, and frequency-domain methods, together with parameter-controlled internal variants.

Table 8. Quantitative comparison with mainstream image classification models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Params (M)	FPS	FLOPs (G)
EfficientNet-B0	86.2 ± 0.4	86.5 ± 0.5	85.8 ± 0.6	86.1 ± 0.5	5.3	450	0.39
DenseNet-121	87.6 ± 0.3	87.9 ± 0.4	87.1 ± 0.5	87.5 ± 0.4	8	275	2.9
ResNet-50	88.4 ± 0.5	88.1 ± 0.6	88.6 ± 0.5	88.3 ± 0.5	25.6	360	4.1
SE-ResNet-50	89.7 ± 0.4	89.9 ± 0.5	89.3 ± 0.5	89.6 ± 0.4	28.1	335	4.2
Swin-T	91.2 ± 0.3	91.5 ± 0.4	90.8 ± 0.5	91.1 ± 0.4	28.3	220	4.5
ViT-Base	91.5 ± 0.4	91.8 ± 0.5	91.2 ± 0.4	91.4 ± 0.4	86.6	185	17.6
ConvNeXt-Tiny	93.5 ± 0.3	93.2 ± 0.4	93.7 ± 0.3	93.4 ± 0.3	28.6	213	4.5
FGC-ConvNeXt (Ours)	97.6 ± 0.2	97.8 ± 0.2	97.4 ± 0.3	97.6 ± 0.2	29.8	170	6.1

Table 9. Comparison with related frequency-domain and spatial–frequency methods.

Method	Frequency Mechanism	Fusion Strategy	Frequency Auxiliary Su- pervision	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Params (M)
ConvNeXt-Tiny	None	None	No	93.5 ± 0.3	93.4 ± 0.3	93.7 ± 0.3	93.2 ± 0.4	28.6
ConvNeXt + DCT branch	DCT-based branch	Static fusion	No	94.4 ± 0.4	94.2 ± 0.4	94.6 ± 0.4	94.3 ± 0.4	29

ConvNeXt + FcaNet	Multi-spectral frequency attention	Channel re-weighting	No	94.9 ± 0.3	94.7 ± 0.4	95.1 ± 0.3	94.8 ± 0.3	28.9
ConvNeXt + Wavelet branch	DWT-based branch	Static fusion	No	95.2 ± 0.3	95.0 ± 0.3	95.4 ± 0.3	95.1 ± 0.3	29.1
GFNet-Tiny	Global Fourier filtering	Single-stream filtering	No	95.0 ± 0.4	94.8 ± 0.4	95.2 ± 0.4	95.0 ± 0.4	25
FFT branch + Addition	FFT amplitude decomposition	Addition	No	95.5 ± 0.3	95.3 ± 0.3	95.7 ± 0.3	95.5 ± 0.3	29.4
FFT branch + Concatenation	FFT amplitude decomposition	Concatenation	No	96.1 ± 0.3	95.9 ± 0.3	96.3 ± 0.3	96.1 ± 0.3	29.6
FGC-ConvNeXt w/o Aux. Loss	FFT amplitude with radial separation	SFGF	No	96.8 ± 0.2	96.6 ± 0.3	97.0 ± 0.2	96.8 ± 0.2	29.8
FGC-ConvNeXt	FFT amplitude with radial separation	SFGF	No	97.6 ± 0.2	97.4 ± 0.3	97.8 ± 0.2	97.6 ± 0.2	29.8

All models were independently trained three times using random seeds 2023, 2024, and 2025, and the mean and standard deviation of each evaluation metric are reported. The same seeds and evaluation protocol were used across the compared models. Given the limited number of runs, the comparisons are reported as descriptive performance differences rather than formal claims of statistical significance. The quantitative comparison results are shown in Table 8. For a fair comparison, all comparison models were trained using the same dataset split, input resolution, data augmentation strategy, optimizer, training epochs, and evaluation protocol. However, the baseline models were not forced to directly reuse all hyperparameters of FGC-ConvNeXt. Instead, key hyperparameters, including learning rate, weight decay, optimizer type, and batch size, were adjusted within a limited validation-based search range according to the characteristics of each architecture and commonly used training settings. This strategy avoids disadvantaging baseline models by directly applying hyperparameters designed for the proposed dual-branch model. Unless otherwise specified, ImageNet pretrained weights were used for model initialization. Each experiment was repeated three times with different random seeds, and the mean and standard deviation were reported. Computational efficiency was evaluated under the same hardware and software environment. FLOPs were calculated using an input resolution of 224×224 pixels. Inference throughput and single-image latency were measured on an NVIDIA RTX 4090 GPU with a batch size of 1 and FP32 precision. After 100 warm-up iterations, the average runtime was calculated over 1000 forward passes with CUDA synchronization. Image loading and preprocessing time were excluded, and the auxiliary classifier was removed during inference.

Across the three runs, FGC-ConvNeXt achieved an accuracy of $97.6 \pm 0.2\%$, compared with $93.5 \pm 0.3\%$ for ConvNeXt-Tiny, corresponding to an average improvement of 4.1 percentage points. The relatively small standard deviations indicate stable results across the three tested random seeds. However, these results are treated as descriptive evidence because the limited number of runs does not support a sufficiently robust inferential statistical analysis. These results suggest that the performance gain of the proposed model is stable rather than caused by random initialization. A detailed analysis of the experimental results reveals significant performance differences among different network architectures when processing Han stone relief images. EfficientNet-B0 and DenseNet-121 obtain relatively lower accuracy under the current dataset and training protocol. Their performance

may be related to differences in architectural design, feature representation, model capacity, and optimization settings; however, the present experiments do not isolate the contribution of each factor.

Transformer-based models, including Swin-T and ViT-Base, achieve accuracies of 91.2% and 91.5%, respectively, demonstrating the advantages of self-attention mechanisms in capturing global features. Nevertheless, the large parameter scale of ViT-Base, reaching 86.6 M, introduces a high computational burden. Because the larger model capacity may require a longer training schedule and more architecture-specific regularization than the standardized protocol adopted in this study. Therefore, the reported result represents its performance under the unified experimental setting rather than its fully optimized upper bound. The baseline model ConvNeXt-Tiny achieves an accuracy of 93.5% due to its modernized macro-architecture design and strong spatial feature extraction capability.

Under the current dataset and experimental protocol, FGC-ConvNeXt achieves the highest accuracy of 97.6% and an F1-score of 97.6% among the evaluated models, with a small standard deviation of ± 0.2 , indicating its effectiveness and stability for Han stone relief image classification. It is worth noting that, compared with SE-ResNet-50 and the advanced Swin-T model, FGC-ConvNeXt achieves a substantial performance improvement with a comparable parameter scale of 29.8 M. This result indicates the effectiveness of adaptively integrating Fourier frequency-domain features with spatial-domain features. These results demonstrate a clear trade-off between improved fine-grained classification performance, robustness, and additional computational cost.

The computational-efficiency results further quantify the overhead introduced by the frequency-domain branch. Compared with ConvNeXt-Tiny, FGC-ConvNeXt increases the parameter count from 28.6 M to 29.8 M, corresponding to an increase of approximately 4.2%. The computational complexity increases from 4.5 GFLOPs to 6.1 GFLOPs, while the single-image latency increases from 4.69 ms to 5.88 ms and the throughput decreases from 213 FPS to 170 FPS. Meanwhile, classification accuracy improves from 93.5% to 97.6%, representing a gain of 4.1 percentage points. These results show that the FFT-based frequency branch and additional feature-mapping modules introduce measurable computational overhead but provide a substantial improvement in fine-grained classification performance. Therefore, the proposed method presents a quantifiable trade-off between recognition performance and computational efficiency.

To further demonstrate that the improvement of FGC-ConvNeXt does not merely come from using a stronger backbone, we additionally compare the proposed method with several frequency-domain and spatial-frequency related baselines. The original comparison with general-purpose CNN and Transformer models mainly evaluates the overall classification ability of different architectures, but it does not directly verify whether the proposed frequency-domain branch and gated fusion strategy are superior to other frequency-related designs. Therefore, additional baselines, including DCT-based frequency modeling, frequency channel attention, DWT-based wavelet modeling, Fourier global filtering, and fixed spatial-frequency fusion strategies, are introduced for comparison.

For a controlled comparison, all methods are trained and evaluated under the same dataset split, input resolution, data augmentation strategy, training epochs, and evaluation protocol. All ConvNeXt-based variants have comparable parameter counts and were evaluated under the same experimental protocol, indicating that the improvement is not solely caused by increased model capacity. As shown in Table 9, all frequency-enhanced variants outperform the pure spatial-domain ConvNeXt-Tiny baseline, indicating that frequency-domain information is beneficial for Han stone relief image classification. However, fixed frequency decomposition and static fusion strategies still show limited perfor-

mance because they cannot dynamically adapt to the category-dependent visual requirements of Han stone relief images. In contrast, FGC-ConvNeXt achieves the best performance by combining explicit frequency-domain texture modeling, sample-level gated fusion, and frequency-domain auxiliary supervision.

To evaluate the robustness of the proposed method under common image-quality variations, additional tests were conducted by applying Gaussian blur, JPEG compression, brightness variation, Gaussian noise, and random cropping to the original test images. The same trained models were directly evaluated under these conditions without additional fine-tuning. The clean test set results are also reported as a reference, as shown in Table 10.

Table 10. Robustness evaluation under fixed single-severity image degradations.

Test Condition	ConvNeXt-Tiny Acc. (%)	FGC-ConvNeXt Acc. (%)	Improvement (%)
Clean test set	93.5	97.6	4.1
Gaussian blur	86.8	93.2	6.4
JPEG compression	88.5	94.1	5.6
Brightness variation	89.7	94.8	5.1
Gaussian noise	85.9	92.6	6.7
Random cropping	87.6	93.5	5.9
Average under degraded conditions	87.7	93.6	5.9

Note: “Average under degraded conditions” denotes the unweighted arithmetic mean of the accuracies under Gaussian blur, JPEG compression, brightness variation, Gaussian noise, and random cropping. The clean test-set result is not included in this average. The source-aware split is treated separately as a dataset-partition protocol and is not included in the degraded-condition average.

To evaluate robustness under common image-quality degradations, the trained models were directly tested on perturbed versions of the original test set without additional fine-tuning. A fixed moderate severity was used for each degradation: Gaussian blur employed a 5×5 kernel with $\sigma = 1.5$; JPEG compression used a quality factor of 50; brightness variation used a multiplicative factor of 0.7; Gaussian noise was sampled from $N(0, 0.05^2)$ after scaling pixel values to $[0, 1]$; and random cropping retained 80% of the image area before resizing to 224×224 pixels. Each condition was evaluated once for every test image. Gaussian noise and random cropping were generated using a fixed random seed of 2024, making the reported perturbations reproducible and deterministic.

As shown in Table 10, the performance of both models decreases under image degradation, but FGC-ConvNeXt consistently maintains higher accuracy than ConvNeXt-Tiny. The average improvement under the five degraded conditions is 5.9 percentage points, indicating that spatial–frequency collaborative modeling improves classification stability under common image-quality variations. These results reflect robustness to simulated degradations rather than independent cross-source generalization. These results evaluate robustness at one fixed moderate degradation level and should not be interpreted as a complete characterization across all corruption severities. Evaluation using multiple severity levels and corresponding degradation curves will be considered in future work.

3.4. T-SNE Feature Distribution Visualization Analysis

To qualitatively examine the feature distributions learned by ConvNeXt-Tiny and FGC-ConvNeXt, t-SNE was used to project the high-dimensional test features into a two-dimensional space [35]. This visualization is intended only to illustrate the projected fea-

ture-distribution patterns and is not regarded as quantitative evidence of semantic decoupling. As shown in Figure 4, different colors and markers represent the four major semantic categories in the self-constructed Han stone relief dataset, namely architecture, auspicious symbols, human figures, and patterns.

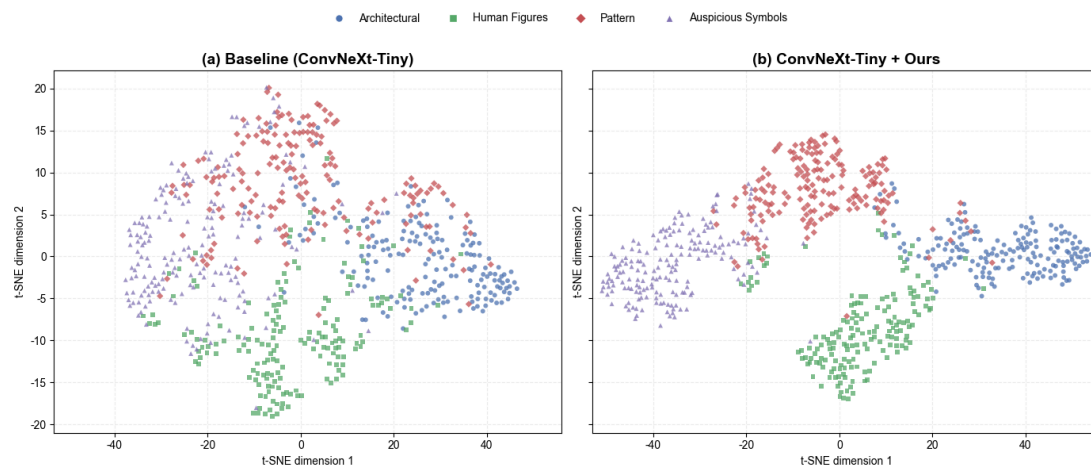


Figure 4. t-SNE visualization of feature distributions generated by the baseline ConvNeXt-Tiny and the proposed FGC-ConvNeXt model.

The 512-dimensional features before the final classifier were extracted from the same test samples for both models. The features were L2-normalized before projection. The t-SNE settings were fixed as follows: two output dimensions, a perplexity of 30, PCA initialization, 1000 iterations, and a random seed of 2024. The same sample set and projection settings were used for both models to ensure a consistent qualitative comparison.

For the baseline model, as shown in Figure 4a, the features extracted by the conventional pure spatial-domain model ConvNeXt-Tiny exhibit obvious confusion and overlap in the two-dimensional feature space. In particular, the decision boundaries between “Auspicious Symbols” and “Human Figures”, as well as those between “Patterns” and “Architecture”, are relatively blurred. In the displayed projection, the baseline shows visible overlap between several categories. However, this observation is qualitative and does not establish the underlying cause of the classification errors. These factors make it difficult for conventional convolutional neural networks to effectively distinguish highly coupled deep semantic structures from superficial texture noise, thereby limiting the discriminative ability of the model when processing cross-domain features of similar categories.

In contrast, the feature distribution of the proposed model is shown in Figure 4b. After introducing the spatial–frequency collaborative modeling architecture, the feature distribution extracted by FGC-ConvNeXt is consistently improved. Samples belonging to the same category are more compactly clustered in the feature space, showing stronger intra-class compactness. Meanwhile, the boundaries among different categories become clearer, exhibiting larger inter-class separation. Samples that were severely confused in the baseline model are effectively separated by the proposed model.

Compared with ConvNeXt-Tiny, the feature distribution produced by FGC-ConvNeXt shows tighter intra-class clusters and clearer boundaries between visually similar categories. This change is consistent with the quantitative improvement reported in the comparative experiments. The spatial–frequency gated fusion module preserves complementary structural and texture cues, while frequency-domain auxiliary supervision prevents the frequency branch from being suppressed during joint optimization. Their combined effect reduces feature overlap, particularly between Human Figures and Auspicious Symbols and between Architecture and Decorative Patterns.

3.5. Visualization Analysis of Model Interpretability

To evaluate the interpretability of the proposed method, this study adopts the improved Gradient-weighted Class Activation Mapping algorithm, namely Grad-CAM++ [36], to visualize class activation maps. The results are shown in Figure 5. The activation responses of the baseline model are scattered and broadly distributed across both foreground and background regions. In contrast, the activation regions generated by the proposed method are more concentrated and highly consistent with the semantic structures of the target objects. This indicates that the baseline model tends to rely on global texture features, whereas the proposed model can extract more discriminative and effective features.

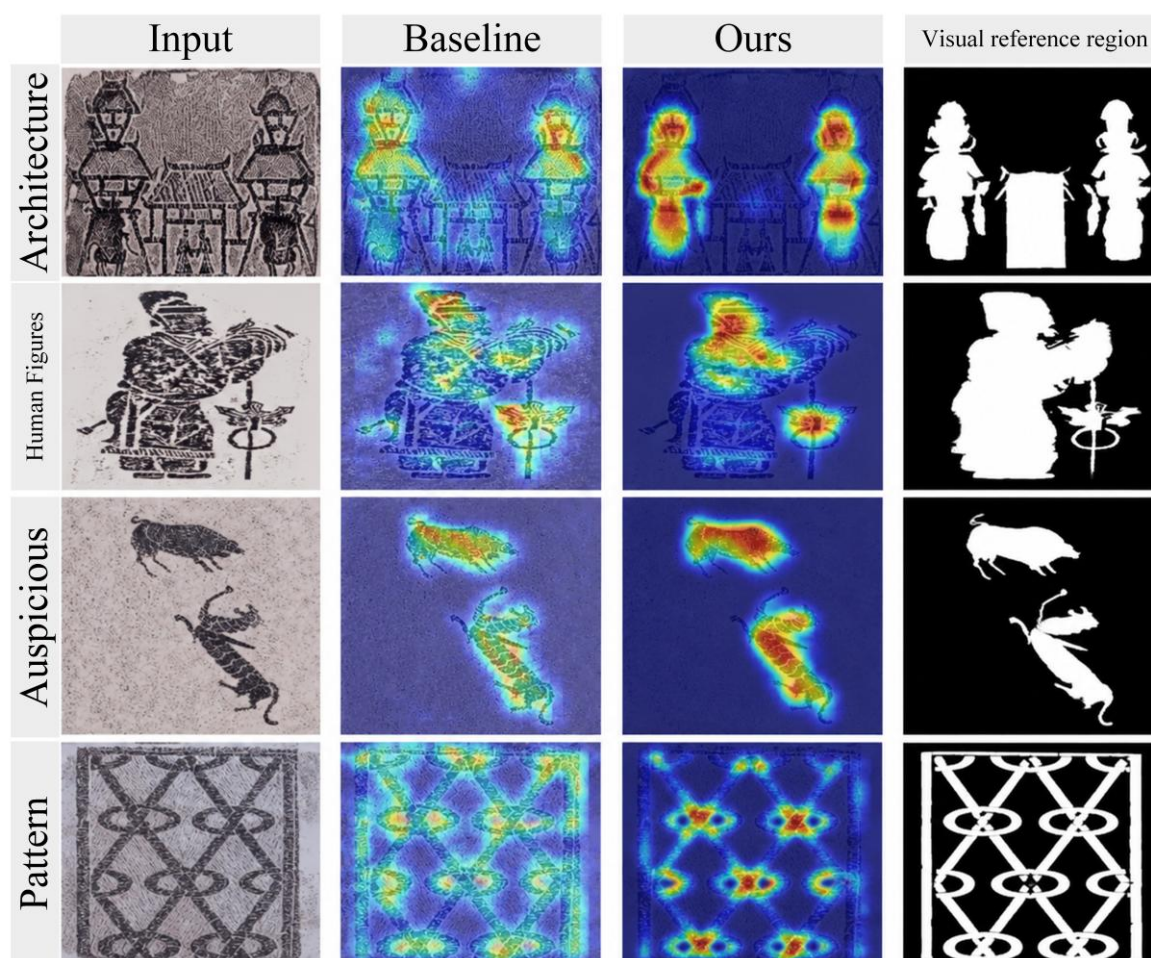


Figure 5. Grad-CAM++ visualization comparison between the baseline ConvNeXt-Tiny and the proposed FGC-ConvNeXt model. The last column shows manually outlined visual reference regions for qualitative interpretation only. These regions are not ground-truth segmentation masks and are not used for model training or quantitative evaluation.

It is noteworthy that the class activation maps generated by the proposed method exhibit a clear distribution pattern characterized by central focus and edge attenuation. The most discriminative regions are highlighted with stronger activation responses, while the activation intensity gradually decreases toward the surrounding areas. This characteristic effectively suppresses irrelevant background interference and reduces the influence of redundant texture noise, which is particularly important for Han stone relief images with complex patterns and large intra-class variations. In this way, the model produces stronger activation responses around visually relevant elements, such as human poses, animal contours, and structural nodes in geometric patterns. The dataset used in

this study only contains image-level category labels and does not include pixel-level segmentation annotations. Therefore, the binary regions shown in Figure 5 are not ground-truth segmentation masks. They are manually outlined visual reference regions based on the visible relief content and are used only for qualitative interpretation of the Grad-CAM++ results. These reference regions are not involved in model training, loss calculation, or quantitative evaluation.

To further evaluate the model's ability to distinguish visually similar samples, additional Grad-CAM++ comparisons were conducted on representative inter-class hard cases from the test set. The selected samples exhibit similar global structures, semantic layouts, or texture distributions, while their category differences mainly depend on subtle local details. The results are presented in Figure 6.

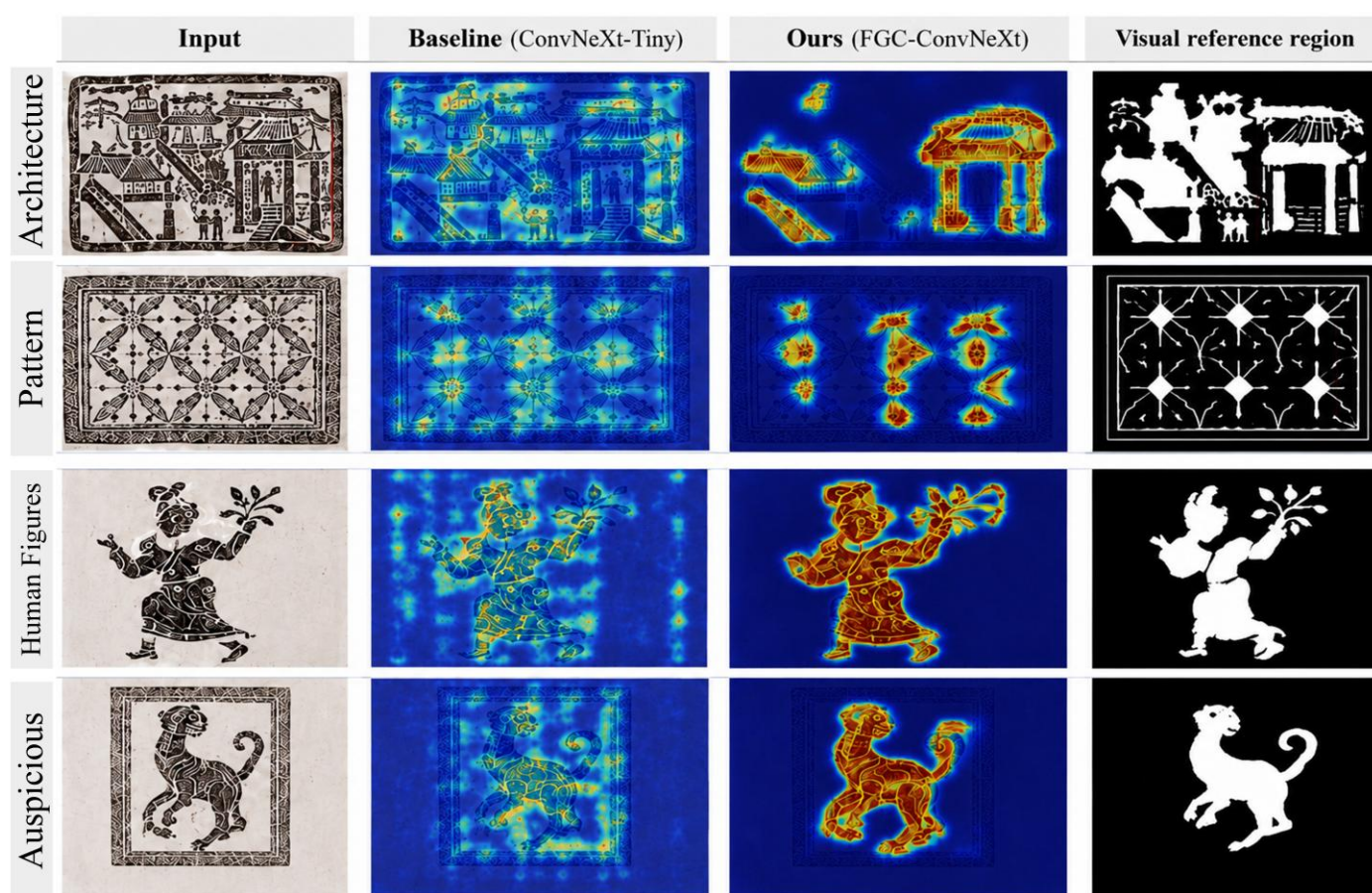


Figure 6. Grad-CAM++ comparison between ConvNeXt-Tiny and FGC-ConvNeXt on representative visually similar samples from different categories. Compared with the baseline, FGC-ConvNeXt produces more concentrated responses around category-specific structural and texture details.

As shown in Figure 6, the selected samples share comparable global contours or texture distributions, making them challenging for conventional spatial-domain classification. ConvNeXt-Tiny tends to generate dispersed responses over shared foreground structures or background textures. In contrast, FGC-ConvNeXt focuses more consistently on category-specific local details. For Architecture, the proposed model emphasizes roof structures, columns, and structural boundaries, whereas for Decorative Patterns, it highlights repeated motif units and texture transitions. For Human Figures, the activation responses concentrate on body posture, limbs, clothing contours, and held objects, while for Auspicious Symbols, the model focuses more clearly on animal heads, limbs, tails, and

body contours. These observations provide additional qualitative evidence that the frequency-domain branch supplies complementary microscopic cues for distinguishing visually similar samples.

Relative to the baseline, FGC-ConvNeXt produces more localized activation responses around human postures, animal contours, architectural boundaries, and repeated decorative motifs, while reducing responses to irrelevant background textures. The comparison provides qualitative evidence that the frequency branch contributes complementary local-detail cues to the spatial representation. Because Grad-CAM++ offers post-hoc visualization rather than pixel-level localization measurements, these results are interpreted as supporting evidence for the model's decision patterns rather than as quantitative proof of localization accuracy.

3.6. Analysis of the Adaptive Gating Weight Distribution Mechanism

To investigate the feature-weighting behavior of SFGF, the 512-dimensional gating vector of each test sample was first averaged across feature dimensions to obtain one sample-level scalar gating weight. The resulting scalar values were then grouped by the ground-truth category and visualized using violin plots. Figure 7 was generated from the representative run with random seed 2024 rather than by pooling the results of multiple runs. A value closer to 1 indicates greater weighting of spatial features, whereas a value closer to 0 indicates greater weighting of frequency-domain features.

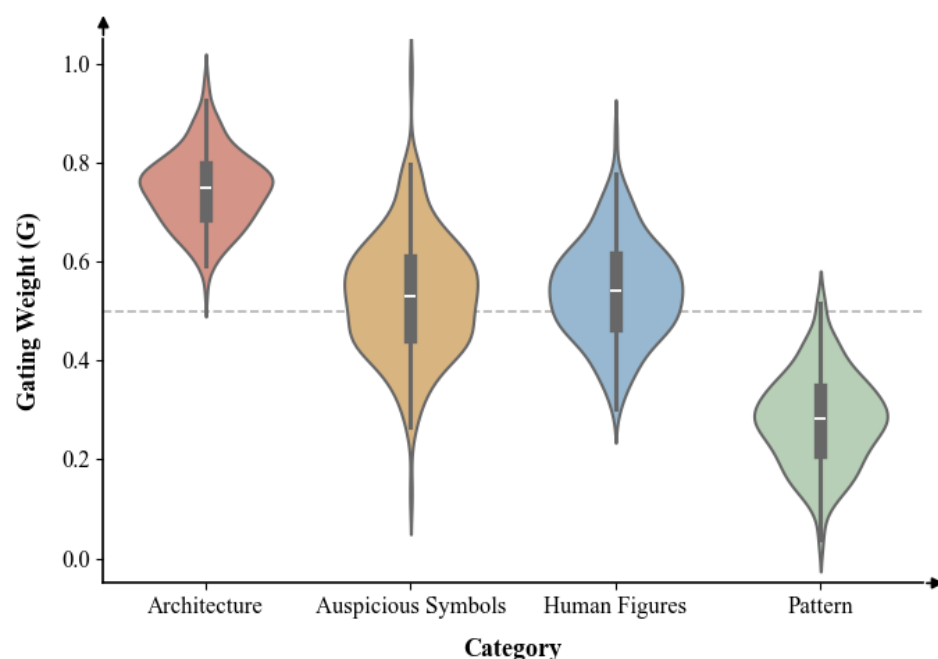


Figure 7. Category-wise gating weight distributions. Values closer to 1 indicate greater reliance on spatial features, whereas values closer to 0 indicate greater reliance on frequency-domain features. The dashed horizontal line at $G = 0.5$ indicates equal weighting of the spatial and frequency-domain features.

From the overall distribution trend, different categories exhibit distinct feature dependencies. Specifically, the mean gating weight of the Architecture category is approximately 0.73, and its distribution is relatively concentrated, indicating a clear preference for the spatial branch. This suggests that, for images with explicit geometric structures and macroscopic compositional features, the model tends to rely more on spatial-domain information for classification.

In contrast, the mean gating weight of the Pattern category is approximately 0.31, and the overall distribution shifts toward the frequency branch. This indicates that the high-frequency texture information extracted by the frequency branch, such as carving details and repetitive decorative patterns, provides stronger discriminative capability for this category. In addition, the gating weights of the Human Figures and Auspicious Symbols categories are mainly distributed around 0.5, showing a relatively balanced dependence on both spatial and frequency features. This demonstrates that the model can adaptively adjust the fusion ratio between spatial and frequency-domain features at the sample level according to the coupling degree of macroscopic structures and microscopic textures in each image.

Overall, Figure 7 provides a qualitative view of the sample-dependent fusion behavior learned by SFGF. Different categories exhibit different gating-weight distributions, suggesting that the model adaptively adjusts the relative contributions of spatial and frequency-domain features. However, these distributions are descriptive observations from one representative run and should not be interpreted as causal evidence of category-specific feature dependence.

3.7. Confusion Matrix and Fine-Grained Classification Performance Analysis

As shown in Table 11, FGC-ConvNeXt achieves consistently high class-wise Precision, Recall, and F1-score across all four categories. Architecture obtains a Recall of 98.1% and an F1-score of 97.7%, indicating that the model can effectively recognize stable architectural contours and structural layouts. Human Figures achieves a Precision of 98.2%, a Recall of 97.0%, and an F1-score of 97.6%. Its slightly lower Recall can mainly be attributed to substantial variations in human poses, clothing contours, and interactions between figures.

Table 11. Class-wise classification performance of FGC-ConvNeXt on the test set.

Category	Test Samples	Correctly Classified	Precision (%)	Recall (%)	F1-Score (%)
Architecture	155	152	97.4	98.1	97.7
Human Figures	169	164	98.2	97	97.6
Auspicious Symbols	155	151	96.8	97.4	97.1
Decorative Patterns	158	155	98.1	98.1	98.1
Overall	637	622	—	—	Accuracy: 97.6

Auspicious Symbols obtains a Precision of 96.8%, a Recall of 97.4%, and an F1-score of 97.1%. Its relatively lower Precision indicates that a small number of samples from other categories are incorrectly predicted as auspicious symbols, mainly because some symbolic motifs share similar local contours and carving textures with human figures and decorative patterns. Decorative Patterns achieves the highest F1-score of 98.1%, demonstrating the effectiveness of the frequency-domain branch in capturing repetitive motifs, carving traces, and fine-grained texture structures.

To more comprehensively and precisely evaluate the classification decision behavior of the model across specific subcategories, this study plots and compares the confusion matrix heatmaps of the baseline model ConvNeXt-Tiny and the proposed FGC-ConvNeXt on the test set (Figure 8). From the confusion matrix of the baseline model, it can be observed that conventional pure spatial-domain modeling is prone to misclassification among similar subcategories when processing highly coupled Han stone relief images. Specifically, there is relatively severe mutual confusion between the Auspicious Symbols and Human Figures categories. In addition, some Pattern samples are easily misclassified as Architecture. One possible explanation is that the spatial baseline may represent some fine-scale carving and rubbing cues less effectively than the proposed dual-branch model.

However, the confusion matrix alone cannot establish the specific mechanism responsible for these errors. And to ensure consistency between the class-wise performance table and the confusion matrix, both results were recalculated from the same set of predictions obtained from a representative experimental run. The confusion matrix was normalized by the number of ground-truth samples in each category. Therefore, the diagonal entries correspond to the class-wise Recall values reported in Table 10. All percentages are rounded to one decimal place for display.

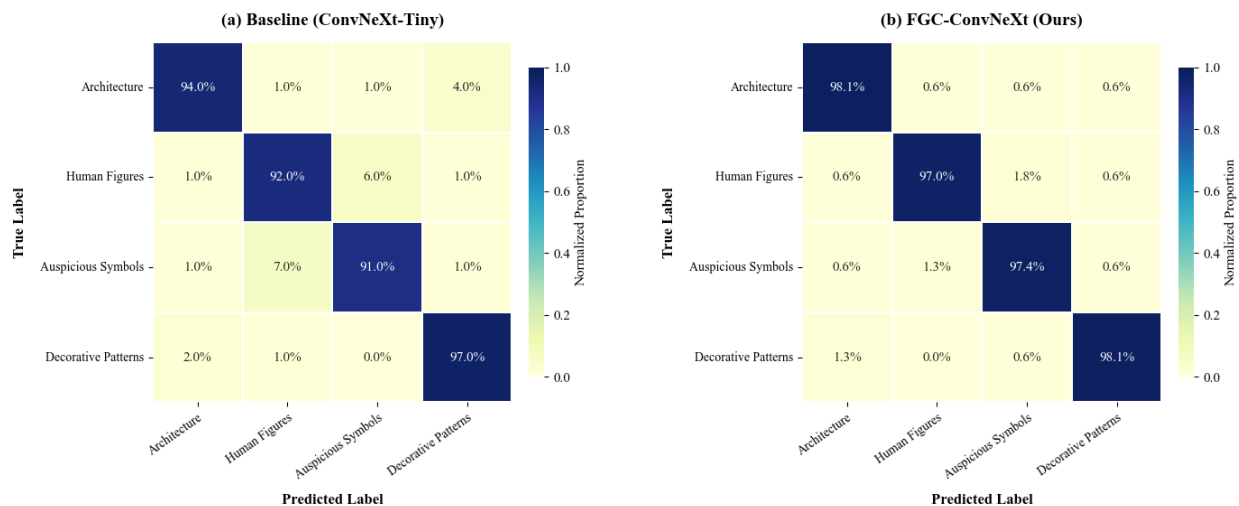


Figure 8. Comparison of confusion matrices between the ConvNeXt-Tiny and FGC-ConvNeXt on the test set.

In contrast, the diagonal elements in the confusion matrix of the proposed FGC-ConvNeXt model are close to ideal values, and the classification accuracy of each subcategory is consistently improved. The Human Figures–Auspicious Symbols pair and the Pattern–Architecture pair, which are heavily confused in the baseline model, show reduced confusion under the proposed method. This result suggests that the frequency-domain branch may provide complementary discriminative cues when the model encounters fine-grained samples with similar macroscopic contours but different microscopic techniques.

4. Discussion

4.1. Effectiveness and Mechanism Analysis

This study focuses on the core challenge of the strong coupling between structural information and texture information in Han stone relief images, and proposes the FGC-ConvNeXt model that integrates spatial-domain and frequency-domain features. The experimental results show that the proposed method achieves higher overall performance than the evaluated convolutional and Transformer-based models under the current experimental protocol. This performance improvement is not caused by a single module alone, but rather results from the collaborative optimization of multiple mechanisms.

First, from the perspective of feature modeling, traditional convolutional neural networks mainly rely on the spatial domain to progressively extract semantic information. When processing cultural heritage images with strong texture interference, some fine-scale texture cues may become less prominent after repeated feature transformations and downsampling, although the extent of this effect depends on the architecture and training strategy. In contrast, the proposed method introduces a frequency-domain branch to explicitly decompose the input image into low-frequency structural components and high-

frequency texture components, enabling the model to learn feature representations in parallel across different frequency scales. The ablation results show that high-frequency information contributes more significantly to classification performance than low-frequency information. This finding essentially indicates that the discriminative cues of Han stone relief images depend more on high-frequency features such as carving details, blade mark directions, and rubbing textures. Meanwhile, when low-frequency and high-frequency information are jointly modeled, the model achieves optimal performance, suggesting that the two types of information are clearly complementary in semantic representation.

Second, in terms of feature fusion strategy, the proposed SFGF module effectively avoids information redundancy and interference commonly found in traditional feature concatenation methods. By introducing a dynamic weight allocation mechanism, the model can adaptively adjust the contribution ratio of the spatial and frequency branches according to the semantic characteristics of each input sample. The gating weight distribution experiment shows that different categories exhibit distinct feature dependencies. Images in the Architecture category, which contain clear geometric structures, rely more heavily on spatial information, whereas images in the Pattern category, which contain complex textures, depend more strongly on frequency-domain information. The Human Figures and Auspicious Symbols categories reflect a joint effect of structural and texture information. Such a fine-grained, sample-level dynamic adjustment mechanism provides the model with stronger representational flexibility and generalization ability.

Third, frequency-domain auxiliary supervision provides additional optimization guidance to the frequency branch. The results show that moderate auxiliary weighting improves classification performance, whereas excessively large weights reduce it. This suggests that an appropriate balance between the main and auxiliary objectives is important. However, when the auxiliary weight becomes too large, model performance declines, indicating that overly strong auxiliary constraints may disrupt the balance between the main task and the auxiliary task. This phenomenon further demonstrates that the weight allocation among different loss terms is crucial for model performance in a multi-task learning framework.

In addition, the t-SNE projections and Grad-CAM++ activation maps provide qualitative observations of the learned feature distributions and decision-related activation patterns. These visualizations are sensitive to parameter settings, random initialization, and sample selection, and therefore should not be interpreted as proof of semantic decoupling, causal feature dependence, or localization accuracy. They are used only as supplementary qualitative evidence together with the quantitative classification results. In the displayed projections and activation maps, FGC-ConvNeXt shows visually more compact feature groupings and more concentrated activation responses than the baseline. However, these observations remain qualitative and should not be interpreted as quantitative evidence of feature separability, localization accuracy, or causal feature dependence.

4.2. Study Limitations and Future Directions

Despite the promising results, several limitations should be acknowledged. The dual-branch architecture, FFT operations, and gated fusion module introduce additional computational and memory costs, which may hinder deployment on resource-constrained devices. Future work will explore lightweight frequency branches, model compression, and more efficient frequency-domain operators.

The experiments were mainly conducted on a self-constructed Han stone relief dataset. Although the dataset contains images from multiple sources, some raw images cannot be fully released because of copyright and cultural heritage data-management restrictions. This may limit reproducibility and direct comparison with other studies. To

improve transparency, labels, data splits, preprocessing scripts, model configurations, and experimental protocols will be provided under appropriate conditions.

The generalization ability of the proposed model also requires further validation. Because rubbing images and original-stone photographs differ in contrast, surface texture, and acquisition conditions, a modality-specific evaluation may provide additional insight. Such an analysis was not conducted in the present study and will be considered in future work. Although source-aware grouping, near-duplicate screening, and robustness tests were adopted, incomplete provenance information makes it difficult to guarantee a fully artifact-disjoint or site-disjoint split. In addition, a fully independent external benchmark is still lacking. Future studies will therefore evaluate the model on independently collected data from different institutions, regions, materials, and digitization conditions. The baseline comparison also adopted a standardized training protocol without exhaustive architecture-specific hyperparameter optimization. Although this design controls the data, training budget, and evaluation conditions, it may not provide the optimal configuration for every architecture, particularly large Transformer models such as ViT-Base. Future work will conduct broader model-specific searches over learning rate, regularization, augmentation strength, and training duration.

The current frequency branch relies on amplitude-based representation and a fixed threshold mask, which may limit its adaptability to diverse image conditions. Moreover, the framework remains dependent on manually annotated static images and does not incorporate archaeological context such as textual descriptions, excavation records, chronological information, and artifact metadata. Future research will investigate adaptive frequency modeling, self-supervised or semi-supervised learning, and multimodal integration of visual and archaeological knowledge.

5. Conclusions

To address the problems of strong coupling between structural and texture information, significant fine-grained differences, and susceptibility to noise interference in Han stone relief images, this study proposes FGC-ConvNeXt, a classification model based on spatial–frequency collaborative modeling. The proposed method introduces a frequency-domain branch to perform frequency decomposition on input images, combines an adaptive gated fusion module to dynamically integrate spatial and frequency-domain features, and uses an auxiliary supervision mechanism to enhance the learning ability of the frequency branch. As a result, the model constructs feature representations with stronger discriminability and robustness.

Extensive experimental results show that the proposed model achieves competitive performance across accuracy, precision, recall, and F1-score, and obtains the best overall results among the evaluated models under the current experimental protocol. On the self-constructed four-category dataset containing 6368 images, FGC-ConvNeXt achieved an accuracy of $97.6 \pm 0.2\%$, a precision of $97.8 \pm 0.2\%$, a recall of $97.4 \pm 0.3\%$, and an F1-score of $97.6 \pm 0.2\%$ across three independent runs. Compared with ConvNeXt-Tiny, the proposed model improved accuracy from $93.5 \pm 0.3\%$ to $97.6 \pm 0.2\%$, corresponding to an increase of 4.1 percentage points, while the parameter count increased from 28.6 M to 29.8 M. Under the same inference environment, the computational complexity increased from 4.5 GFLOPs to 6.1 GFLOPs, while FGC-ConvNeXt maintained an inference speed of 170 FPS compared with 213 FPS for ConvNeXt-Tiny. Visualization analyses provide supplementary qualitative observations of the projected feature distributions and activation patterns, while the robustness experiments indicate improved classification stability under the evaluated conditions.

The reported results establish the effectiveness of spatial–frequency collaborative modeling for the present four-category Han stone relief dataset. FGC-ConvNeXt improves

fine-grained classification performance while maintaining a practical inference speed, although the frequency branch introduces additional computational cost. Its broader applicability still requires validation on fully independent datasets collected from different institutions, archaeological sites, materials, and digitization conditions. Future research can be further extended in the following directions: first, exploring more lightweight spatial–frequency fusion structures to reduce computational cost; second, introducing self-supervised or semi-supervised learning strategies to reduce dependence on large-scale annotated data; and third, incorporating multimodal information for higher-level semantic understanding, thereby promoting the intelligent analysis and digital preservation of cultural heritage.

Author Contributions: Conceptualization, H.W., J.D., W.J., L.Z. and S.L.; methodology, H.W. and J.D.; software, H.W. and Z.D.; validation, W.J., H.L. and L.Z.; formal analysis, W.J., H.L. and S.L.; resources, Z.D.; data curation, H.W.; writing—original draft preparation, J.D.; supervision, J.D., W.J. and Z.D.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Education Department of Henan Province, grant number 2026SJGLX394, 2025-JCZD-16 and HZC2026KT014. The APC was funded by No. 2026SJGLX394, 2025-JCZD-16 and HZC2026KT014.

Data Availability Statement: Due to cultural heritage data-management requirements and source-specific restrictions, the raw Han stone relief images used in this study are not publicly available. The category labels, training/validation/test split files, preprocessing scripts, model configurations, and experimental protocols are available from the corresponding author upon reasonable request. Processed image samples may be shared where permitted by the relevant source institutions and data-management requirements.

Conflicts of Interest: Author Wenlin Jin was employed by the company Jinjia Jun Kiln Co., Ltd. Author Hongsheng Liu was employed by the company Hengde Jun Porcelain Co., Ltd. The remaining 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.

Appendix A. Individual Results of Different Random Seeds

Table A1. Individual results of ConvNeXt-Tiny and FGC-ConvNeXt under three random seeds. The seed-level results are provided for transparency. No formal hypothesis test was conducted because only three independent runs were available.

Model	Random Seed	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ConvNeXt-Tiny	2023	93.2	93.1	93.4	92.8
	2024	93.5	93.4	93.7	93.2
	2025	93.8	93.7	94	93.6
	Mean $\pm$ SD	93.5 $\pm$ 0.3	93.4 $\pm$ 0.3	93.7 $\pm$ 0.3	93.2 $\pm$ 0.4
FGC-ConvNeXt	2023	97.4	97.6	97.1	97.4
	2024	97.6	97.8	97.4	97.6
	2025	97.8	98	97.7	97.8
	Mean $\pm$ SD	97.6 $\pm$ 0.2	97.8 $\pm$ 0.2	97.4 $\pm$ 0.3	97.6 $\pm$ 0.2

References

- Shih, H. Han stone reliefs from Shensi Province. *Arch. Chin. Art. Soc. Am.* **1960**, *14*, 49–64.
- Zhou, X. An archaeological interpretation of han dynasty brick and stone pictorial reliefs. *Chin. Stud. Hist.* **2018**, *51*, 276–299. <https://doi.org/10.1080/00094633.2018.1553457>.

3. Chan, W.-Y.; Ma, S.-Y. Heritage preservation and sustainability of China's development. *Sustain. Dev.* **2004**, *12*, 15–31. <https://doi.org/10.1002/sd.224>.
4. Pavlidis, G.; Koutsoudis, A.; Arnaoutoglou, F.; Tsioukas, V.; Chamzas, C. Methods for 3D digitization of cultural heritage. *J. Cult. Herit.* **2007**, *8*, 93–98. <https://doi.org/10.1016/j.culher.2006.10.007>.
5. Morgan, C. Current digital archaeology. *Annu. Rev. Anthropol.* **2022**, *51*, 213–231. <https://doi.org/10.1146/annurev-anthro-041320-114101>.
6. Morgan, C. Avatars, monsters, and machines: A cyborg archaeology. *Eur. J. Archaeol.* **2019**, *22*, 324–337. <https://doi.org/10.1017/eaa.2019.22>.
7. Fiorucci, M.; Khoroshiltseva, M.; Pontil, M.; Traviglia, A.; Del Bue, A.; James, S. Machine learning for cultural heritage: A survey. *Pattern Recognit. Lett.* **2020**, *133*, 102–108. <https://doi.org/10.1016/j.patrec.2020.02.017>.
8. 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. <https://doi.org/10.3390/app7100992>.
9. 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.
10. 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, 27–30 June 2016; pp. 779–788. <https://doi.org/10.1109/CVPR.2016.91>.
11. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11966–11976. <https://doi.org/10.1109/CVPR52688.2022.01167>.
12. Pintus, R.; Pal, K.; Yang, Y.; Weyrich, T.; Gobbetti, E.; Rushmeier, H. A survey of geometric analysis in cultural heritage. *Comput. Graph. Forum.* **2016**, *35*, 4–31. <https://doi.org/10.1111/cgf.12668>.
13. Vaish, P.; Wang, S.; Strisciuglio, N. Fourier-basis functions to bridge augmentation gap: Rethinking frequency augmentation in image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 17763–17772. <https://doi.org/10.1109/CVPR52733.2024.01682>.
14. Bo, D.; Wang, X.; Shi, C.; Shen, H. Beyond low-frequency information in graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 3950–3957. <https://doi.org/10.1609/aaai.v35i5.16514>.
15. Peigney, M. A Fourier-based machine learning technique with application in engineering. *Int. J. Numer. Methods Eng.* **2021**, *122*, 866–897. <https://doi.org/10.1002/nme.6565>.
16. Qi, J.; Zhang, X.; Ye, D.; Ruan, Y.; Guo, X.; Wang, S.; Li, H. SFNet: Fusion of Spatial and Frequency-Domain Features for Remote Sensing Image Forgery Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 28563–28578. <https://doi.org/10.1109/jstars.2025.3621474>.
17. Liu, S.; Hu, J. Visual saliency based on frequency domain analysis and spatial information. *Multimed. Tools Appl.* **2016**, *75*, 16699–16711. <https://doi.org/10.1007/s11042-016-3903-3>.
18. Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; Ontanon, S. Fnet: Mixing tokens with fourier transforms. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 4296–4313. <https://doi.org/10.18653/v1/2022.naacl-main.319>.
19. Lin, S.; Zeng, Y.; Gong, Y. Learning of time-frequency attention mechanism for automatic modulation recognition. *IEEE Wirel. Commun. Lett.* **2022**, *11*, 707–711. <https://doi.org/10.1109/LWC.2022.3140828>.
20. Zhou, Y.; Hu, X.; Han, J.; Wang, L.; Duan, S. High frequency patterns play a key role in the generation of adversarial examples. *Neurocomputing* **2021**, *459*, 131–141. <https://doi.org/10.1016/j.neucom.2021.06.078>.
21. Xu, Z.Q.J.; Zhang, Y.; Luo, T. Overview frequency principle/spectral bias in deep learning. *Commun. Appl. Math. Comput.* **2025**, *7*, 827–864. <https://doi.org/10.1007/s42967-024-00398-7>.
22. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 783–792. <https://doi.org/10.1109/ICCV48922.2021.00082>.
23. Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; Zhou, J. Global filter networks for image classification. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 980–993.
24. Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; Catanzaro, B. Adaptive Fourier neural operators: Efficient token mixers for transformers. *arXiv* **2024**, arXiv:2111.13587.

25. Fu, Y.; Xie, Y.; Fu, Y.; Chen, J.; Jiang, Y.G. Wave-san: Wavelet based style augmentation network for cross-domain few-shot learning. *arXiv* **2022**, arXiv:2203.07656. <https://doi.org/10.48550/arXiv.2203.07656>.
26. Jiang, Z.; Dong, Z.; Fan, J.; Yu, Y.; Xian, Y.; Wang, Z. Breast TransFG Plus: Transformer-based fine-grained classification model for breast cancer grading in Hematoxylin-Eosin stained pathological images. *Biomed. Signal Process. Control* **2023**, *86*, 105284. <https://doi.org/10.1016/j.bspc.2023.105284>.
27. Charvillat, V.; Tonazzini, A.; Van Gool, L.; Nikolaidis, N. Image and video processing for cultural heritage. *EURASIP J. Image Video Process.* **2010**, *2009*, 163064. <https://doi.org/10.1155/2009/163064>.
28. Chen, Y.; Liang, H.; Li, H.; Song, S. A lightweight time-frequency-space dual-stream network for active sonar-based underwater target recognition. *IEEE Sens. J.* **2025**, *25*, 11416–11427. <https://doi.org/10.1109/JSEN.2025.3538688>.
29. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. <https://doi.org/10.1109/ICCV48922.2021.00986>.
30. Liu, J.; Lin, L.; Cai, Z.; Wang, J.; Kim, H.-J. Deep web data extraction based on visual information processing. *J. Ambient Intell. Humaniz. Comput.* **2024**, *15*, 1481–1491. <https://doi.org/10.1007/s12652-017-0587-0>.
31. Dehbozorgi, P.; Ryabchykov, O.; Bocklitz, T. A systematic investigation of image pre-processing on image classification. *IEEE Access* **2024**, *12*, 64913–64926. <https://doi.org/10.1109/ACCESS.2024.3395063>.
32. Brigham, E.O.; Morrow, R.E. The fast Fourier transform. *IEEE Spectr.* **1967**, *4*, 63–70. <https://doi.org/10.1109/MSPEC.1967.5217220>.
33. Zhou, M.; Huang, J.; Yan, K.; Hong, D.; Jia, X.; Chanussot, J.; Li, C. A general spatial-frequency learning framework for multi-modal image fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 5281–5298. <https://doi.org/10.1109/TPAMI.2024.3368112>.
34. Sujon, K.M.; Hassan, R.; Choi, K.; Samad, A. Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *J. Big Data* **2025**, *12*, 268. <https://doi.org/10.1186/s40537-025-01313-4>.
35. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 1–27.
36. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Piscataway, NJ, USA, 2018; pp. 839–847.

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.