npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01796-7

Research and implementation of mural classification based on lightweight network

Jin Zheng[1,2], Manjun Zhang[2], Yinghui Zhang[2], Wuxin Yuan[2] & Zhao Xiaobing[1]

Dunhuang murals, as invaluable historical and cultural heritage, pose significant challenges in automatic classification due to their large volume, visual similarity, and deterioration over time. This study introduces SER-Net, a lightweight and efficient classification network optimized for real-time mural recognition on mobile devices. A specialized dataset covering nine dynasties—Early Tang, Northern Wei, Northern Zhou, Peak Tang, Sui, Late Tang, Middle Tang, Five Dynasties, and Western Wei—was manually constructed and augmented to address class imbalance. SER-Net is designed based on RepVGG and ResNet18, and incorporates the SED-Block module, which integrates squeeze-and-excitation (SE) attention and Channel-Shuffle mechanisms to improve feature representation. Moreover, the use of depthwise separable convolution significantly reduces the model parameters while maintaining accuracy. Experimental results demonstrate that SER-Net effectively balances model size, accuracy, and computational efficiency, making it suitable for deployment in resource-constrained environments.

Dunhuang mural painting, as an important part of the history of Chinese art development, has a long and deep cultural heritage[1]. An in-depth study of Dunhuang murals is of great significance in exploring a series of cultural activities such as trade exchanges along the Silk Road, the spread of Buddhist culture, and language exchanges between China and neighboring countries. However, researchers need to spend a lot of time to organize, compare and analyze a large number of mural pictures after completing the required mural data collection, which is less efficient. Therefore, how to efficiently assist people to find fresco data is an issue worthy of in-depth research.

In recent years, both domestic and international scholars have increasingly focused on the application of intelligent information processing technologies in the classification of mural images. With the continuous advancement of computer technologies, the digitization of mural artworks has been effectively realized through methods such as 3D laser scanning and three-dimensional reconstruction, significantly facilitating their preservation, analysis, and research. In particular, laser technologies have been widely adopted in the field of cultural heritage conservation, offering high-precision, non-destructive means for data acquisition and restoration support[2]. Wei Daoquan[3] used long and short-term memory neural networks to extract nonlinear spectral features of mural paintings and combined them with convolutional neural networks for nonlinear mapping, and finally achieved more than 97% of OA coefficient results, which effectively improved the accuracy of the pigment classification of mural painting images. Zeng Ziming et al. [4] used bag-of-words model to construct the underlying feature matching algorithm, which firstly extracts the image shape features using SIFT, then generates the visual dictionary by K-means and maps it into TF-IDF vectors, and then calculates the similarity to match the image sorting by inner product, and conducts the semantic search using the subject words. Aiming at the characteristics of large differences between classes and noise of mural paintings. Although traditional learning methods are able to extract mural features to a certain extent, simple feature extraction has greater limitations due to the subjectivity and unique cultural background of murals. Therefore, deep learning methods have been widely used. Kumar et al.[5] used pre-trained AlexNet and VGGNet models to extract mural features and combined them with support vector machines for classifier fusion to successfully classify thangka images into eight categories of Indian art forms.Caspari et al.[6] constructed a three-layer CNN network for detecting early Iron Age tombs of Google Earth open-source optical satellite data. Cao et al.[7] designed an Inception-v3 network incorporating migration learning for ancient mural painting dynasty identification classification, and added color histograms with local binary patterns (LBP) to better extract the artistic features of the mural paintings. Ding Yunle et al.[8] proposed a 3D cavity convolution residual neural network fusing multi-scale features for pigment classification of multi-spectral images of frescoes, which finally achieved an average accuracy of 96.89%.

In summary, a deep exploration of mural features and the adoption of appropriate classification methods are key to improving the efficiency of mural management and preservation. As deep learning becomes

[1]Minzu University of China, Beijing, China. [2]Huanggang Polytechnic College, Huanggang, China. e-mail: nmzxb_cn@163.com

increasingly prevalent in mural recognition, deploying lightweight yet effective models on mobile devices has become a pressing need. Traditional convolutional neural networks are often too large and computationally intensive for efficient mobile deployment. Although CNN pruning is one way to compress models, it typically requires multi-stage training, retraining, and additional parameter tuning, which complicates practical deployment and may lead to performance degradation. Recent advancements also include the application of multi-scale diffusion models for mural restoration[9] and the use of neural architecture search for mural dynasty classification[10]. In addition, researchers have explored enhanced U-Net architectures for improving restoration transparency[11], and proposed residual attention hybrid networks to optimize contour extraction[12]. These works further confirm the potential of integrating deep learning with mural-specific feature modeling for improved classification and restoration accuracy.

To address these challenges directly, this study proposes a lightweight mural classification network named SER-Net (Squeeze-and-Excitation RepVGG[13] Network), which is custom-designed based on RepVGG and ResNet18[14]. Instead of post-hoc compression, SER-Net is inherently lightweight and efficient. Specifically, it introduces the SED-Block module, which integrates a channel attention mechanism and a Channel-Shuffle structure to enhance feature fusion. The reparameterization-based RepVGG structure improves inference speed, while depth-wise separable convolutions mitigate the parameter overhead introduced by attention modules. This architecture enables SER-Net to achieve real-time classification performance with fewer parameters and lower computational cost, making it highly suitable for mobile applications.

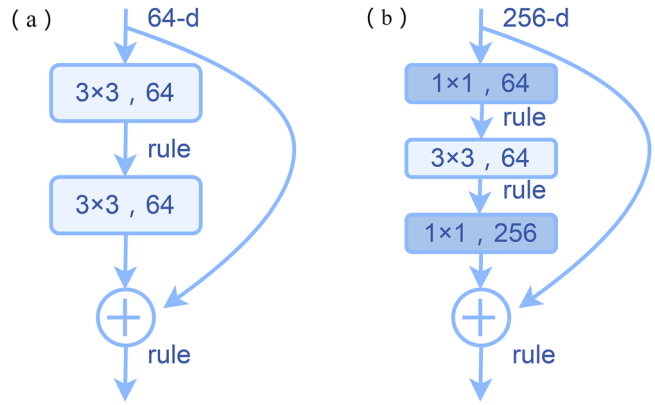

Fig. 1 | Residual network structure diagram[14]. **a** Residual Structure. **b** Residual Bottleneck Structure.

Methods

Residual network structure

The ResNet family, a network architecture proposed by Microsoft Labs in 2015, In previous networks, an increase in the number of layers led to an increase in the error rate due to the gradient degradation or gradient explosion problem. To solve this problem, ResNet introduced the residual network structure (shown in Fig. 1), which allows the network to obtain better results by increasing the depth.

RepVGG network structure

RepVGG is an optimized variant of ResNet that overcomes its limitations, such as constraints on input/output dimensions and challenges in model compression. Introduced in 2021. Figure 2 shows that RepVGG utilizes VGG's 3 × 3 convolutions and single-branch structure, enhancing flexibility and reducing memory usage for lightweight devices. It integrates residual branches and 1 × 1 convolutions during training but uses a single-branch structure for efficient inference, ensuring effective feature extraction and optimized deployment.

Backbone network design

This paper proposes a backbone network structure that employs jump connections to link feature information of different layers, RepVGG to fuse information features of each layer, and parallel processing of convolutions of different sizes to fully extract feature information of each layer. The reparametric structure is utilized to accelerate the model inference process. The framework first employs RepVGG to efficiently extract low-level image

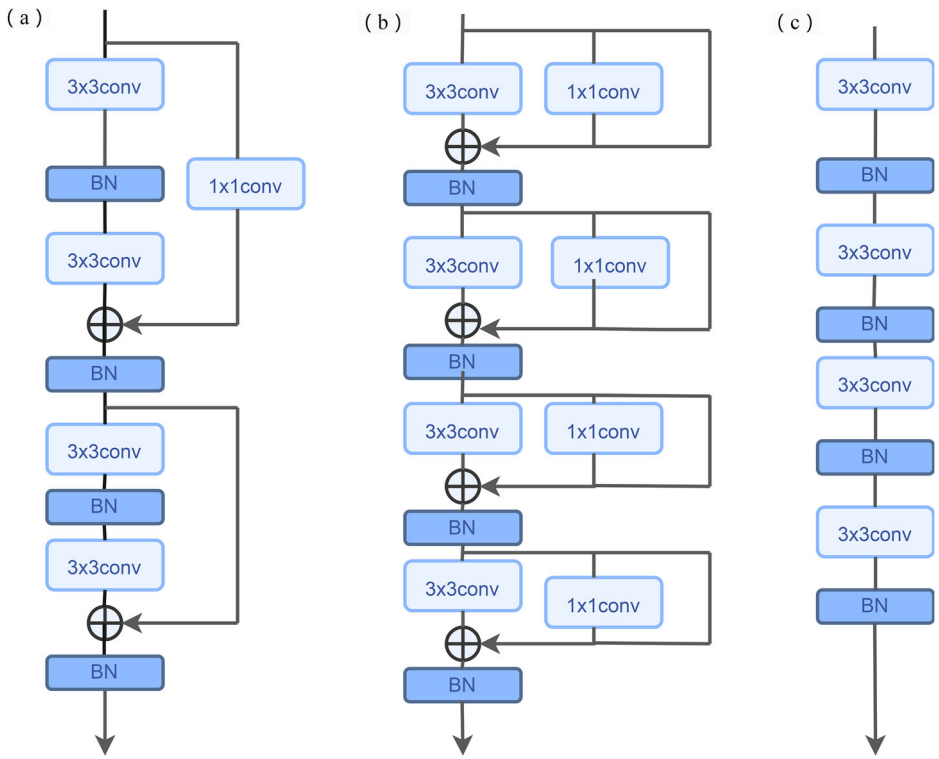

Fig. 2 | Sequence-to-sequence keyword extension model based on attention mechanism[13]. **a** ResNet Structure. **b** RepVGG Train Structure. **c** RepVGG Inference Structure.

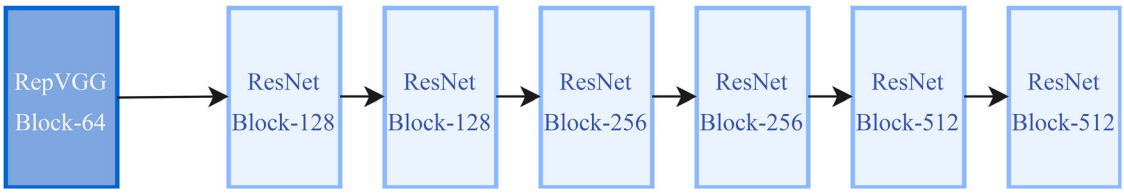

Fig. 3 | Modular structure of the designed backbone network.

features from murals and then utilizes ResNet to capture high-level information features. Figure 3 illustrates the SED-Block, the backbone module we designed.

Figure 3 illustrates that the first module of the network, RepVGG, operates in parallel to extract texture and color features from the base layer of the fresco using distinct convolutional operations. RepVGG replaces the initial convolutional part (Stem) in ResNet18, specifically the 7×7 Conv. The residual structure within the module enables the reuse of feature information between the bottom and top layers, enhancing the feature extraction capability of the fresco image while maintaining network flexibility.

Residual network-based design for Dunhuang murals classification

In this paper, we aim to extract effective features of mural images and meet the real-time requirements of mobile. To this end, a comparative analysis is first carried out, and by introducing different attention mechanisms and channel obfuscation techniques, the Channel Attention Mechanism[15] (Squeeze-and-Excitation, SE) and the Channel-Shuffle mechanism[16] are finally selected to improve the feature extraction and reuse capabilities of the network. However, the attention mechanism increases the number of model parameters, for this reason, in this paper, depth-separable convolution is used instead of ordinary convolution to control the number of model parameters and computation. Based on this, the SER-Net network is proposed, which consists of one RepVGG module and six SED-Block modules.

Among them, the RepVGG module uses several modules in parallel to improve the information extraction capability of the bottom layer of the network, and the remaining layers use SED-Block modules to extract high-level image features. The fresco image classification network SER-Net proposed in this paper contains an input part, RepVGG module, SED-Block module and an output part. Among them, the SED-Block module consists of two branches, each branch consists of 3×3 DW-Conv, BN layer and 1×1 Conv, and uses the channel attention mechanism to enhance the feature extraction ability of the network. The output results of each branch undergo a Concat operation to recover the original number of channels, and the Channel-Shuffle mechanism is used in this paper in order to better fuse the channel information and improve the model's ability to reuse feature information. The whole network structure is shown in Fig. 4:

In the study of frescoes, this paper observes that some neighboring dynasties have extremely similar styles, leading to unsatisfactory classification results. To enhance the network's ability to extract information from mural image features, this paper adds SE Attention and Channel Shuffle mechanisms, as shown in Fig. 5. The SE Attention mechanism helps the model learn the importance of different channel information and allows the network to pay attention to the importance of different channels of the mural, resulting in improved mural classification. Additionally, the added SE module is lightweight, meeting the practical requirements of deploying the model to mobile terminals.

To mitigate the increase in model parameters caused by the SE attention mechanism, this study replaces standard convolution with depthwise separable convolution[17], effectively reducing computational complexity while maintaining feature extraction capability.

Deep separable convolution has fewer parameters compared to normal convolution. The number of parameters in a feature map with M channels and a convolutional kernel size D_k is $M \times D_k \times D_k$. In contrast, a normal convolutional layer with N such convolutional kernels outputs multiple channel features, resulting in $N \times M \times D_k \times D_k$. Using depth-separable convolution significantly reduces the computation. Although depth-separable convolution has the same computation as normal convolution, it uses N convolution kernels for point-by-point convolution on features that output multiple channels, reducing the overall computation. The formula for the computational amount of deep separable convolution, denoted by F, is provided in Eq. (1).

$$F = D_k \times D_k \times M \times D_F \times D_F + M \times N \times D_F \times D_F \tag{1}$$

Where D_k is the convolution size, the input image size is D_F, N is the number of convolution kernels, and M is the number of channels. Convolution compares these two parametric quantities to get the result of Eq. (2), and it can be found that the parametric quantity of the depth separable convolution is much smaller than that of the standard convolution.

$$A = \frac{D_k \times D_k \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_k \times D_k \times M \times N + D_F \times D_F} = \frac{1}{N} + \frac{1}{D_k^2} \tag{2}$$

Where, A represents the ratio of parameters between depth separable convolution and standard convolution. This paper processes mural pictures with a size of 256×256, 3 channels, and a 3×3 convolution kernel. Through ablation experiments, the proposed method achieves significant reductions in network parameters and computation while maintaining model accuracy.

Results

The Dunhuang mural painting dataset used in this paper is sourced from electronic scanned versions of The Complete Collection of Dunhuang Mural Paintings in China[18], The Complete Collection of Dunhuang Cave Paintings[19], and Dunhuang Pattern Facsimiles[20]. A total of 3221 original mural images were selected and labeled based on the corresponding dynasties. The study focuses on nine labeled dynasties, as depicted in Fig. 6, where the horizontal axis represents the dynasties and the vertical axis represents the number of mural paintings for each dynasty.

Initial experiments revealed several issues with the dataset: (1) a small sample size, which increases the risk of overfitting; (2) variations in mural preservation resulting in long-tailed distribution and increased difficulty; and (3) non-uniformity in the format and quality of digital resources, requiring manual processing. To address these issues and mobile shooting scenarios, this paper employs data augmentation techniques, such as flipping, color random dithering, scaling and cropping, rotation, Gaussian noise, masking, and panning. These methods are illustrated in Fig. 7 and improve the model's fitting performance.

To address the issue of a long-tailed distribution in the Dunhuang mural painting dataset, this paper samples the dataset based on the sample distribution, achieving a balanced distribution across all samples. By applying varying levels of data augmentation, the distribution of each dynasty in the dataset transforms from a long-tailed distribution to a class-balanced distribution. The enhanced mural painting data for each dynasty is visualized in Fig. 8.

To validate the contribution of each component in SER-Net, we conducted ablation experiments. We sequentially removed the SED-Block, depthwise separable convolution, and attention mechanism to observe their

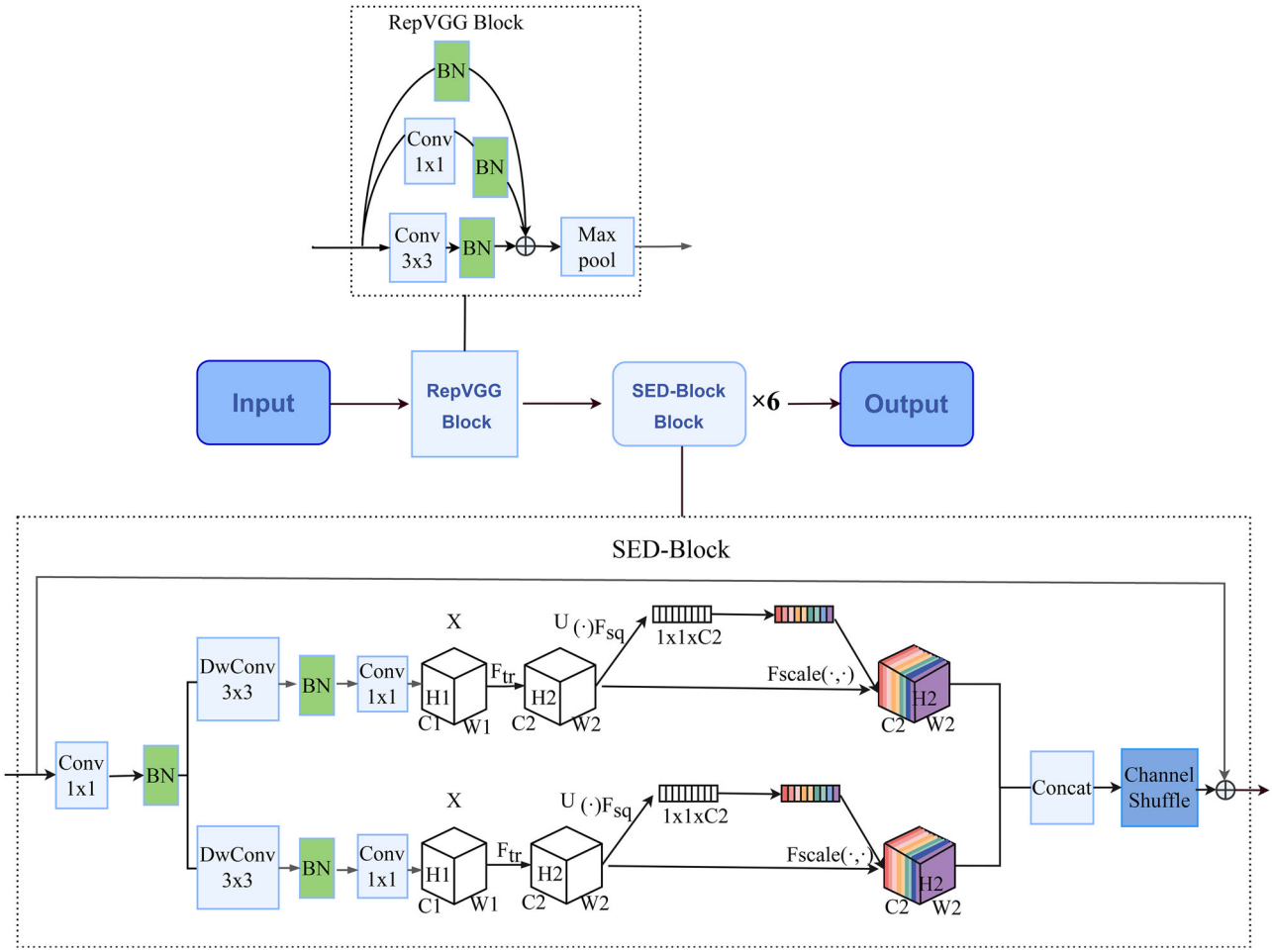

Fig. 4 | Sequence-to-sequence keyword extension model based on attention mechanism.

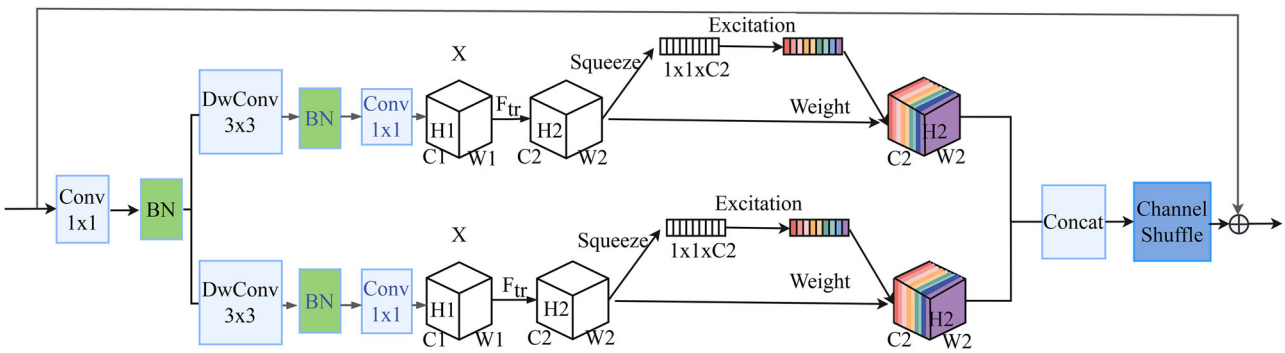

Fig. 5 | A schematic diagram of the model incorporating SE attention and Channel Shuffle mechanisms.

individual impact on classification performance. The comparison experiments show that each module has a significant contribution to the overall accuracy and efficiency.

(1) Combining RepVGG module

This study presents a novel approach that synergistically integrates the RepVGG and ResNet modules to augment the classification backbone. The ResNet module effectively addresses the challenge of gradient vanishing in deep convolutional neural networks by incorporating skip connections for the fusion of multi-layer features. Conversely, the RepVGG module enhances the feature extraction capability of single-layer convolutions through the parallelization of multiple convolutions within a single module, while ensuring model lightweightness. Therefore, Fig. 9 shows that through the combined utilization of these two modules, the performance of the classification backbone is significantly advanced, capitalizing on the multi-layer feature fusion capability of the ResNet module and the efficient feature extraction ability of the RepVGG module.

This paper investigates different approaches to improve the Stem part, including the 7×7 Conv used in ResNet, a 4×4 convolution with step size 4 in the ConvNeXt[21] network, and three 3×3 convolutions proposed in the literature[22]. An ablation comparison experiment is conducted on the Dunhuang mural dataset to evaluate these approaches under the same experimental conditions. Three control groups are established for the Stem part: ResNet18's 7×7 Conv, ConvNeXt's 4×4

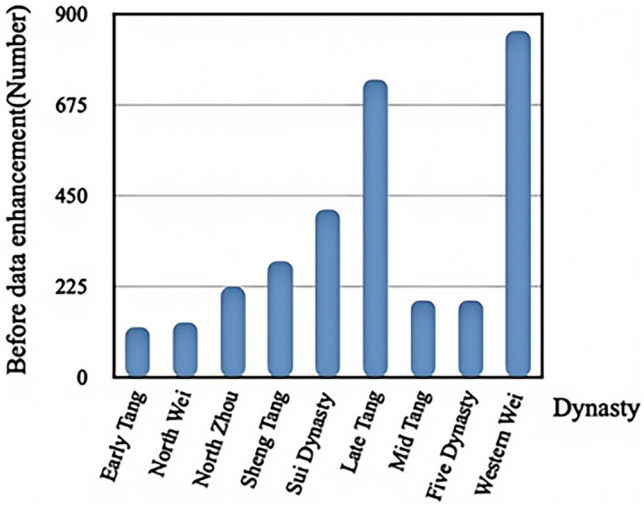

Fig. 6 | Histogram of original mural painting data by dynasty.

Conv with step size 4, and the three 3 × 3 Conv approaches from the literature. By solely modifying the Stem part while keeping the other modules unchanged, three sets of experimental results are obtained and shown in Table 1 The results indicate that replacing the Stem part with the RepVGG module optimizes all metrics when compared to the other three experiments.

(2) SE attention mechanism

In this study, we compare three attention mechanisms, namely the SE attention mechanism, Coordinate Attention (CA)[23], and spatial attention mechanism[24]. Experimental results of these comparisons are presented in Table 2, and to better understand the feature extraction effect, visualization of the experimental results is performed using the Grad-CAM algorithm[25], as depicted in Fig. 10. Our findings show that the SE attention mechanism outperforms the other two mechanisms in terms of improving the mural classification effect while having a lighter structure that meets algorithm implementation requirements. By focusing on different channels, the SE attention mechanism enables the model to learn the importance of information from different channels and activate relevant channels for extracted features.

Fig. 7 | Schematic diagram of image data enhancement for each dynasty. (**a**) Scaling (**b**) Vertical Flip (**c**) Horizontal Flip (**d**) Increase Brightness (**e**) Salt-and-Pepper Noise (**f**) Decrease Brightness.

The SER-Net model was trained on the Dunhuang mural dataset using a 6:2:2 split for training, validation, and testing. This ratio was selected after preliminary experiments demonstrated that it offered a balanced trade-off between training depth and generalization performance on a relatively small dataset. Although a comprehensive comparison table is not included, trials with alternative splits such as 7:1.5:1.5 and 8:1:1 resulted in either slight overfitting or unstable validation accuracy. Hence, 6:2:2 was adopted as the optimal configuration.

To optimize performance under limited data conditions, we set the batch size to 16 and the number of data-loading workers to 4, ensuring efficient GPU utilization. The initial learning rate was set to 0.0001. Confusion matrix analysis (Fig. 11) and manual inspection reveal that the model often misclassifies the Middle Tang, Late Tang, and Five Dynasties periods. This confusion is largely due to stylistic continuities across these dynasties. The Five Dynasties murals inherited vivid colors and red-green contrast from the Late Tang, while the Middle Tang featured earthy red tones mixed with yellow or black, contrasting with the Late Tang's preference for earthy yellow or white. These subtle visual overlaps contribute to classification ambiguity across the three periods. A summary of preliminary experiments with different data splits is provided in Table 3 for reference.

Additionally, the evaluation metrics for each dynasty classification were calculated, and the results are shown in Table 4. Considering the reliance of image classification algorithms on pre-trained models, this study first pre-trains on the ImageNet dataset[26] to obtain better initialization parameters, followed by fine-tuning on the Dunhuang mural dataset. The accuracy achieved after fine-tuning is considered the final performance of the model.

We compared and analyzed the confusion matrix of the model before and after pre-training by selecting various network structures, including ResNet series, RepVGG18, classic lightweight networks, and MobileVit-S network structure using traditional CNN and Transformer in reference[27]. These comparisons were evaluated based on multiple indicators such as parameter count, computation cost, accuracy, and runtime. The specific experimental results are shown in Table 5, and the proposed SER-Net algorithm achieved an accuracy of 90.1%. The results showed that the SER-Net model proposed in this article had fewer parameters and higher accuracy than the ResNet series and RepVGG18 baseline networks. Compared with classic lightweight networks, the SER-Net model not only had higher accuracy but also ran faster, which verified the theory mentioned at the beginning of the article that the model's computation cost and runtime are independent. Moreover, compared with MobileVit-S, the SER-Net model

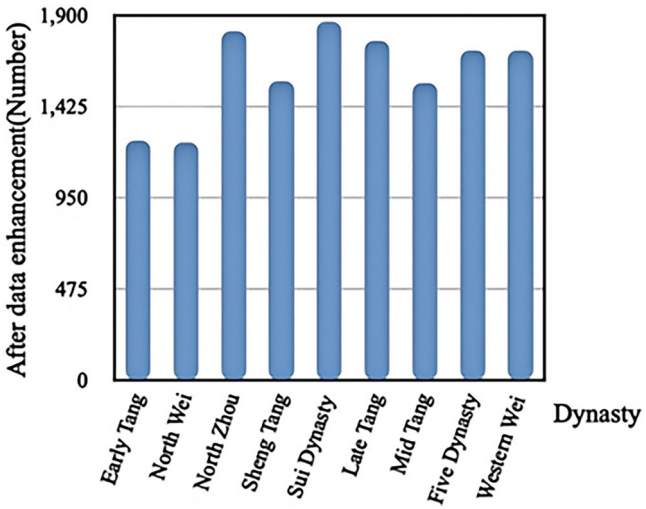

Fig. 8 | Histogram of mural painting data by dynasty after data enhancement.

Table 2 | Comparison of evaluation metrics for experimental results of replacing Stem layers

Network	Params	Flops	Accuracy	Runtime	Frame rate
Original Network	11.69 M	1.82 G	80.6%	0.09 s	11.11
Incorporate the CA Attention Mechanism	18.68 M	1.71 G	82.5%	0.33 s	3.03
Incorporate the Spatial Attention Mechanisms	13.53 M	1.96 G	83.6%	0.27 s	3.70
Incorporate the SE attention mechanism	12.10 M	1.89 G	85.2%	0.15 s	6.67

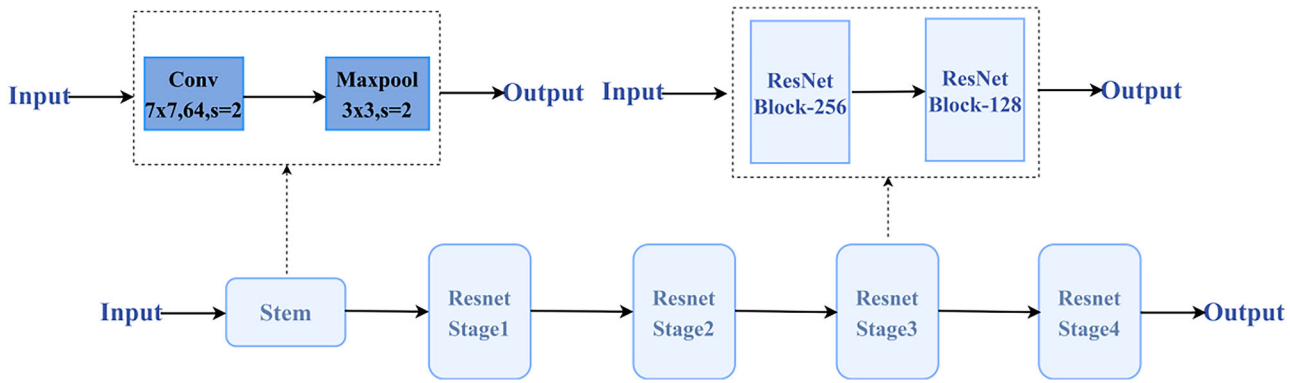

Fig. 9 | Schematic diagram of the structure of ResNet18.

Table 1 | Comparison of evaluation metrics for experimental results of replacing Stem layers

Stem	Params	Flops	Accuracy	Runtime	Frame rate
Original Convolution	11.69 M	1.82 G	80.6%	0.09 s	11.11
3×3 + 3 × 3 + 3 × 3	70.76 M	5.89 G	83.9%	0.12 s	5.23
4 × 4(stride = 4)	11.68 M	1.71 G	82.5%	0.23 s	3.67
RepVGG Module	11.53 M	1.26 G	85.6%	0.07 s	14.28

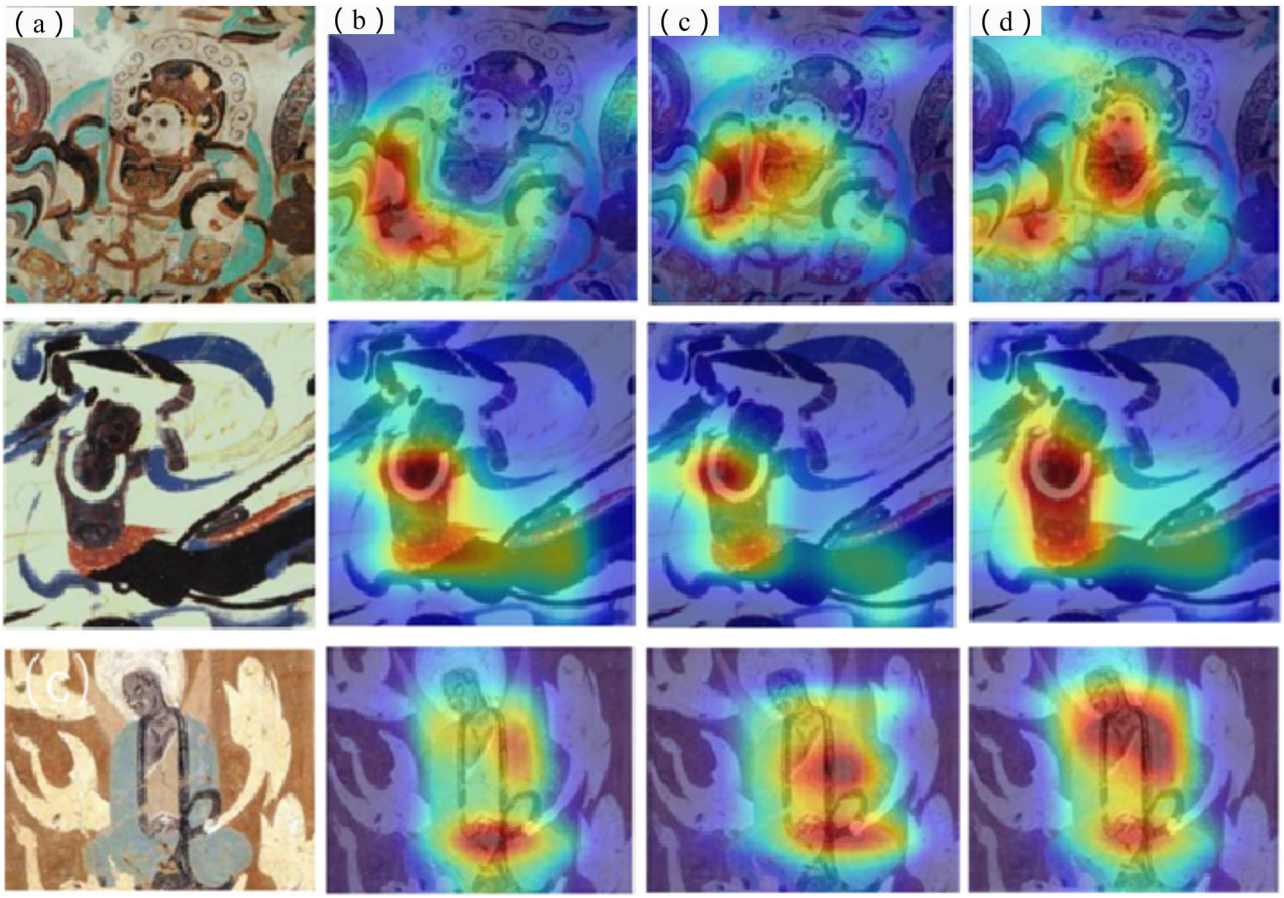

Fig. 10 | Schematic diagram of the structure of ResNet18. a Original Network (**b**) CA Attention (**c**) Spatial Attention (**d**) SE Attention.

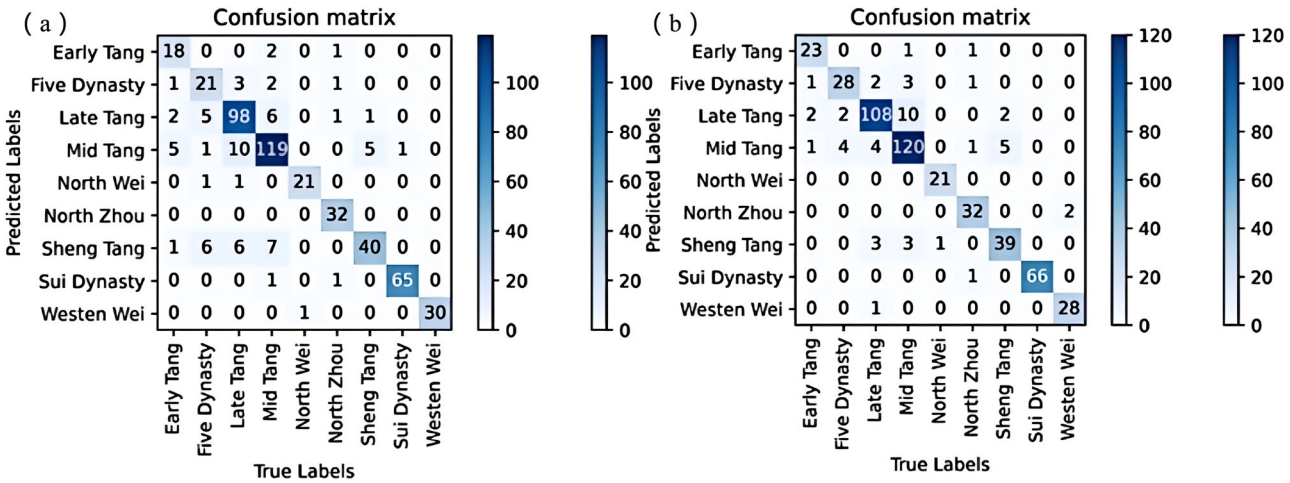

Fig. 11 | Comparison of confusion matrix plots for models before and after loading pre-training. a Confusion Matrix Plot Before Pre-Training Model Loading (**b**) Confusion Matrix Plot After Pre-Training Model Loading.

Table 3 | Performance comparison under different train-validation-test splits

Split Ratio (Train:Val:Test)	Accuracy	Observations
6:2:2	86.4%	Balanced performance
7:1.5:1.5	83.6%	Mild over-fitting observed
8:1:1	85.2%	Validation accuracy unstable

far exceeded its running speed in actual testing, demonstrating the high efficiency of the entire network structure designed in this article. Overall, the SER-Net model performed the best in mural classification accuracy, outperforming other networks, indicating that the SER-Net model has a good fitting effect on the Dunhuang mural dataset and can effectively extract features from mural images.

In order to comprehensively evaluate the performance of the proposed SER-Net, we compare it with several classical lightweight networks,

Table 4 | Indicators for assessing the categorization of each mural painting dynasty

Dynasty	Precision	Recall	Specificity
Early Tang	92%	85.2%	99.6%
Five Dynasty	80%	82.4%	98.5%
Late Tang	87.1%	91.5%	96%
Middle Tang	88.9%	87.6%	96%
Northern Wei	100%	95.5%	100%
Northern Zhou	94.1%	88.9%	99.6%
Sheng Tang	84.8%	84.8%	98.5%
Sui Dynasty	98.5%	100%	99.8%
Western Wei	96.8%	93.3%	99.8%

Table 5 | Comparison of metrics for evaluating the results of the various network experiments

Network	Params	Flops	Accuracy	Runtime	Frame rate
Resnet18	11.69 M	1.82 G	87.5%	0.107 s	10.31
Resnet34	21.80 M	3.67 G	88.3%	0.191 s	5.23
RepVGG18	11.50 M	1.80 G	86.8%	0.091 s	10.99
ShufflenetV2	**2.28 M**	**0.15 G**	79.8%	0.096 s	10.41
MoblienetV1	5.10 M	0.57 G	80.1%	0.147 s	6.80
MoblienetV2	3.50 M	0.32 G	81.0%	0.098 s	10.20
PeleeNet	2.80 M	0.52 G	82.7%	0.126 s	7.93
Efficientnetv2_b0	21.46 M	2.87 G	82.1%	0.129 s	7.74
MobileVit-S	5.56 M	1.42 G	85.5%	0.51 s	1.96
LightVit-T	9.4 M	0.73 G	89.1%	0.18 s	5.55
SER-Net	8.22 M	0.95 G	**90.1%**	**0.09 s**	**11.11**

The bold values represent the best (highest) performance metrics among the compared methods, highlighting the networks that achieved the top results for each evaluation metric.

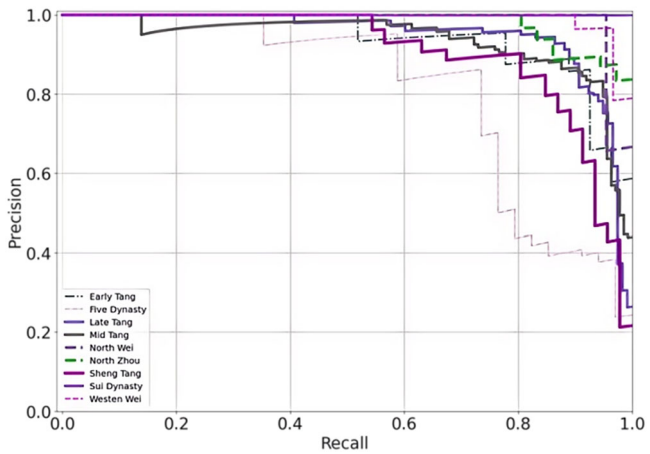

Fig. 12 | The PR curves for each category.

including ShufflenetV2[28], MobileNetV1, MobileNetV2[29], PeleeNet[30], EfficientNetV2-B0[31], MobileViT-S, and LightViT-T[32], which have been widely used in mobile vision tasks due to their compact structure and efficient inference.

To gain a more intuitive understanding of the classification performance for each dynasty, this study calculated the accuracy and recall rate for each category and plotted the PR curve of the model's prediction results. From Fig. 12, it can be observed that as the recall value increases, the PR curves for different periods show a decreasing trend at varying rates. The recall values for most dynasties are around 0.8, while the precision remains above 0.8. It can be seen from Table 4 that the precision of the Five Dynasties period, which is 80.0%, and the recall rate, which is 82.4%, are significantly lower than those of other periods. By analyzing Fig. 12 and Table 4 together, it is found that images from the Five Dynasties period are most likely to be incorrectly predicted as images from other dynasties. Furthermore, from the confusion matrix in Fig. 11, it can be concluded that images from the Mid-Tang period are primarily misclassified as images from the Late Tang period, while images from the Late Tang period and the Five Dynasties period are primarily misclassified as images from the Mid-Tang period.

As shown in Fig. 12, misclassified samples provide insights into the model's challenges. To further interpret the classification errors, we analyzed the dynasties with lower recognition accuracy, specifically the Five Dynasties, Early Tang, and Sheng Tang periods. The misclassifications are primarily attributed to stylistic similarities across adjacent dynasties and intra-dynastic variations. For example, murals from the Five Dynasties often inherit vivid red-green color contrasts from the Sheng Tang period, while some Early Tang images display bright tonal features commonly seen in later periods. Additionally, inconsistent lighting conditions and occasional image blurring reduce the model's ability to extract color and texture features accurately. These overlapping visual characteristics and image quality issues contribute to the challenges in distinguishing between these specific periods.

Discussion

This study constructs a dedicated Dunhuang mural image dataset and proposes a dynasty classification network that integrates ResNet18, RepVGG18, and classic lightweight network design principles. Multiple attention mechanisms were compared, and the channel attention mechanism (SE) was ultimately selected for its superior feature extraction capability. To further control model complexity and computation cost, depthwise separable convolutions and channel-shuffling techniques were employed, improving feature utilization efficiency.

Experimental results show that the proposed method not only improves classification accuracy but also significantly reduces model parameters and inference time. Compared to the Baseline model, it achieves a 2.6% improvement in accuracy, a 3.47 MB reduction in model size, and a 7 ms decrease in runtime. Compared to RepVGG18, it improves accuracy by 2.1%, reduces parameters by 3.28 MB, and shortens runtime by 20 ms. These results demonstrate not only the effectiveness of the proposed approach but also its advantages in parameter compression and deployment-friendliness, making it highly suitable for mobile or resource-constrained real-time applications.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 11 December 2024; Accepted: 14 May 2025;
Published online: 03 October 2025

References

1. Dong, H. & Yang, Y. The connection between Dunhuang murals and Dunhuang. *Adv. Humanit. Res.* **11**, 5–12 (2024).
2. Wang, C. L. et al. Application of laser technology in cultural relics protection. *Laser Optoelectron. Prog.* **59**, 1700003 (2022).
3. Wei, D. Q. & Wang, H. Q. Pigment classification method of mural sparse multi-spectral image based on space spectrum joint feature. *Acta Photonica Sin.* **51**, 0430002 (2022).
4. Zeng, Z. & Sun, S. Research on mobile visual search model for Dunhuang murals. *Inf. Sci. Serv.* **42**, 104–112 (2021).
5. Kumar, S. & Tyagi, A. Indian art form recognition using convolutional neural networks. *Proc. 5th Int. Conf. Signal Process. Integr. Netw. (SPIN)* 800–804 (2018).

6. Caspari, G. & Crespo, P. Convolutional neural networks for archaeological site detection – Finding "princely" tombs. *J. Archaeol. Sci.* **110**, 104998 (2019).
7. Cao, J. F., Yan, M. M., Jia, Y. M. & Tian, X. D. Application of Inception-v3 model with transfer learning in the identification of ancient mural dynasties. *Comput. Appl.* **41**, 3219–3227 (2021).
8. Ding, Y. L., Wang, H. Q., Wang, K., Wang, Z. & Zhen, G. A 3D-CNN classification method for mural multispectral image pigments based on multi-scale feature fusion. *Prog. Laser Optoelectron.* **59**, 369–377 (2022).
9. Han, H., Wang, X., Zhao, J. & Li, Y. DiffuMural: Multi-scale Diffusion Model for Missing Part Restoration in Dunhuang Murals. arXiv preprint arXiv:2504.09513 (2025).
10. Zhou, L., Meng, H. & Zhang, J. Dynastic Classification of Ancient Murals via Discrete Stochastic Neural Architecture Search. *Herit. Sci.* **12**, 27 (2024).
11. Jia, Y., Chen, Q. & Wang, B. Dunhuang Murals Restoration with Visualized Process using Enhanced U-Net Architecture. *Appl. Sci.* **15**, 1422 (2025).
12. Liu, Y., Zeng, Y. & Xiao, Z. RamixNet: Residual Attention and Convolution Hybrid Network for Dunhuang Mural Contour Generation. arXiv preprint arXiv:2212.00935 (2022).
13. Ding, X. et al. RepVGG: Making VGG-style ConvNets great again. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. 13733–13742 (2021).
14. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. 770–778 (2016).
15. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. 7132–7141 (2018).
16. Zhang, X., Zhou, X., Lin, M. & Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. 6848–6856 (2018).
17. Howard, A. G. et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017).
18. Chinese Dunhuang Mural Complete Works Committee. *The Complete Works of Chinese Dunhuang Murals*. Tianjin People's Fine Arts Publishing House, Tianjin (2006) (in Chinese).
19. Dunhuang Academy. The Complete Works of Dunhuang Caves. Shanghai People's Publishing House, Shanghai (2001) (in Chinese).
20. Fan, J. S. *Copy Drawings of Dunhuang Patterns*. Jiangsu Ancient Books Publishing House, Jiangsu (2000) (in Chinese).
21. Liu, Z. et al. A ConvNet for the 2020s. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. 11976–11986 (2022).
22. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *Proceedings of the International Conference on Machine Learning (ICML)*, 1–12 (2015).
23. Hou, Q. et al. Coordinate Attention for Efficient Mobile Network Design. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* **2021**, 13713–13722 (2021).
24. Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. Coordinate Attention for Efficient Mobile Network Design. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 13713–13722 (2021).
25. Selvaraju, R. R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
26. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
27. Mehta, S. & Rastegari, M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint* arXiv:2110.02178 (2021).
28. Ma, N., Zhang, X., Zheng, H. T. & Sun, J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. *Proc. Eur. Conf. Comput. Vis*. 116–131 (2018).
29. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. C. MobileNetV2: Inverted residuals and linear bottlenecks. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. 4510–4520 (2018).
30. Wang, R., Li, X., Ling, C. & Lin, Z. Pelee: A real-time object detection system on mobile devices. *Adv. Neural Inf. Process. Syst.* **31**, 1963–1972 (2018).
31. Tan, M. & Le, Q. V. EfficientNetV2: Smaller models and faster training. *Proc. Int. Conf. Mach. Learn.* **139**, 10096–10106 (2021).
32. Huang, T. et al. LightViT: Towards light-weight convolution-free vision transformers. *arXiv preprint* arXiv:2207.05557 (2022).

Author contributions

Z.Z. and Z.W. were responsible for drafting the main manuscript text. Z.Z. also prepared all figures and tables. Additionally, Z.W. contributed to the overall structure and clarity of the manuscript. All authors participated in reviewing and revising the manuscript to ensure its accuracy and coherence. X.Z. is the corresponding author and is responsible for manuscript submission, correspondence during the review process, and handling all communication related to the publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Zhao Xiaobing.