Cao *et al. Heritage Science* (2024) 12:58
https://doi.org/10.1186/s40494-024-01172-x

Heritage Science

RESEARCH

Open Access

Ancient mural segmentation based on multiscale feature fusion and dual attention enhancement

Jianfang Cao[1,2*], Zhen Cao[1], Zhiqiang Chen[3], Fang Wang[1,2], Xianhui Wang[1] and Zhuolin Yang[1]

Abstract

To address the fuzzy segmentation boundaries, missing details, small target losses and low efficiency of traditional segmentation methods in ancient mural image segmentation scenarios, this paper proposes a mural segmentation model based on multiscale feature fusion and a dual attention-augmented segmentation model (MFAM). The model uses the MobileViT network, which integrates a coordinate attention mechanism, as the feature extraction backbone network. It attains global and local expression capabilities through self-attention, class convolution, and coordinate attention and focuses on location information to expand the receptive field and achieve improved feature extraction efficiency. An A_R_ASPP feature enhancement module is proposed for the attention-optimized residual atrous spatial pyramid pooling module. The module uses residual connections to solve the small target loss problem in murals caused by the excessive sampling rate of atrous convolution and uses a feature attention mechanism to adaptively adjust the feature map weight according to the channel importance levels. A dual attention-enhanced feature fusion module is proposed for multiscale decoder feature fusion to improve the mural segmentation effect. This module uses a cross-level aggregation strategy and an attention mechanism to weight the importance of different feature levels to obtain multilevel semantic feature representations. The model improves the mean intersection over union (MIoU) by 3.06% and the MPA by 1.81% on a mural dataset compared with other models. The model is proven to be effective at improving the segmentation details, efficiency and small target segmentation results produced for mural images, and a new method is proposed for segmenting ancient mural images.

Keywords Mural image segmentation, Attention mechanism, Atrous spatial pyramid pooling, Residual connection, Multiscale feature fusion

Introduction

Ancient murals are valuable aspects of China's long history and culture. The development of ancient Chinese murals was closely related to people's beliefs, customs and aesthetic concepts in various historical periods. These murals also reflect the political, economic, cultural, artistic and technological development levels of society at that time and have very precious historical value. Over time, ancient murals have been eroded by various diseases, such as blister disease [1] and soot disease [2], and they must be protected and restored. The existing measures, such as physical protection and manual restoration, have low efficiency and high cost levels, while the development of digital technology provides various methods for protecting ancient murals. Image segmentation technology, as a key part of image digital restoration, is also highly important for preserving ancient murals.

*Correspondence:
Jianfang Cao
caojianfangcn@163.com
[1] School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
[2] Department of Computer Science and Technology, Xinzhou Normal University, No. 1 Dunqi East Street, Xinzhou 034000, China
[3] Information Technology, SEGi University, Kota Damansara, Petaling Jaya, Selangor, Malaysia

When segmenting ancient mural images, traditional segmentation methods are mostly used, but these segmentation models do not have universal applicability. Traditional image segmentation methods divide the underlying image feature information (such as contour, edge, color, and texture information) into several nonoverlapping regions, which are segmented based on pixel discontinuity and similarity. The graph cuts algorithm proposed by Cao et al. [3] uses the wavelet transform and an adaptive feature threshold to alleviate the influence of noise in mural images and fuses the Sobel and Canny operators to extract mural contours for enhancing edge information. However, these methods tend to be affected by image complexity and specific pixel points, and segmentation errors are highly likely to occur when such approaches are used to address mural images containing noise or occlusion. Furthermore, this method also suffers from artificial intervention; i.e., artificial interaction is required when performing mural segmentation tasks. Wang et al. [4] proposed a fuzzy C-means clustering (FCM) algorithm, which can segment mural images according to a pixel aggregation process implemented in the feature space and subsequently map the images back to the original image space to obtain segmentation results. However, this algorithm does not take spatial information into account and is relatively sensitive to noise and unevenly distributed grayscales. Furthermore, it can be severely affected by the given sample distribution, which results in differences between the segmented samples and the target samples. Venkatachalam et al. [5] proposed an adaptive K-means image segmentation algorithm, which can avoid bilateral K value inputs. Gray mural images are refined via K-means clustering to obtain relatively detailed segmentation results. However, this approach is highly sensitive to the initial K value and tends to fall into local optima for murals, producing a poor overall segmentation effect; furthermore, this algorithm cannot handle mural noise and outliers. These drawbacks strongly influence the final mural segmentation effect. Traditional image segmentation methods generally do not consider spatial information and are sensitive to inhomogeneous noise and intensity levels. Problems such as blurred boundary segmentation and a lack of segmentation information are common when ancient mural segmentation methods are used.

With the continuous development of computer technology, segmentation methods based on deep learning have been widely employed in various scenes and images. A deep learning-based segmentation method uses a deeper network structure to capture complex image features, optimizes the constructed model through a large amount of training data, and further improves its segmentation accuracy. Moreover, hardware acceleration techniques such as GPUs can be used by deep learning methods to achieve improved image segmentation task efficiency. The fully convolutional neural network (FCN) proposed by Jonathan et al. [6] is a milestone in the image segmentation field. The FCN replaces the fully connected layer of the traditional convolutional neural network (CNN) with a fully convolutional layer so that the network can accept input images of any size and output dense pixel predictions of the corresponding size. Additionally, a skip connection structure is added during the upsampling process and combined with the results of different pooling layers to optimize the segmentation effect. However, the pooling layers of this method reduce the resolution of the mural feature map, and upsampling at high magnifications leads to problems such as blurred mural segmentation boundaries. Furthermore, this method does not consider the dependent relationships among the pixels in mural images. Badrinarayananan et al. [7] proposed a SegNet model with an unpooling structure that applies the maximum pooling index used in the encoder to the decoder to optimize the boundary segmentation process. However, this approach primarily focuses on pixel-level information and thus does not make full use of contextual relationships when addressing mural segmentation tasks (this type of information is crucial for correctly classifying the adjacent areas of murals, without which incorrect mural segmentation results tend to be obtained). Zhao et al. [8] proposed a PSPNet model that performs pyramid pooling operations on feature maps with different scales and obtains multiscale feature representations by aggregating multiple pieces of regional context information to improve the ability of the model to obtain global information. However, when an image contains objects with great scale differences, the effect of pyramid pooling cannot reach a level such that the features of each object can be accurately captured. In addition, the pyramid pooling operation can cause spatial information blurring to a certain extent, particularly at the edges and subtle structures of mural images, leading to unsatisfactory segmentation effects for mural edges and details. Aiming at addressing the limitations of FCNs, Chen et al. [9, 10] proposed a series of DeepLab models and designed an atrous spatial pyramid pooling (ASPP) module. The ASPP module uses atrous convolutions with multiple different expansion factors to obtain the multiscale semantic information of the input image and performs feature enhancement to optimize the segmentation effect of the model. When conducting mural segmentation, the ASPP module implements numerous atrous convolutions in parallel, which causes high computational complexity, thereby affecting the speeds of the training and inference processes. Furthermore, atrous convolutions with higher sampling rates can easily lead

to missing information for small mural targets, thereby affecting the segmentation results. Cao et al. [11] proposed the MC-DM model for mural images to extract the backbone network features with MobileNetV2, yielding improved mural segmentation efficiency and creating a new method for mural image segmentation. However, this approach may negatively influence the segmentation accuracy achieved for murals to some extent. In recent years, image segmentation methods that integrate attention mechanisms have also achieved excellent results. For example, the DANet proposed by Fu et al. [12] adds two types of attention modules to the traditional extended FCN to simulate semantic interdependence in the spatial and channel dimensions. This strategy further improves feature representations by combining the outputs of the two attention modules and obtains more accurate segmentation results than those of other methods. However, the mere reliance on an attention mechanism in this design may lead to the insufficient utilization of information acquired from small mural targets, which results in an overwhelming focus on large mural targets while the details of small mural targets are ignored. That is, the missing information problem encountered when addressing small mural targets was not effectively mitigated by this model. Zheng et al. [13] proposed the segmentation transformer (SETR) model. In this model, the input image is segmented, and position codes are subsequently added to transform the image into vector sequences. With a transformer as the encoder, this model can construct the global context at each feature learning stage, and then prediction outcomes can be obtained through step-by-step upsampling performed by the decoder. However, the employment of fixed-size blocks in the SETR results in inflexible position encodings, and the feature maps in the encoder and decoder do not use multiple scales and thus lack hierarchical feature representations. This in turn leads to edge blurring and missing information in mural segmentation tasks. Additionally, the complexity of the self-attention operations in transformers is high, and the number of model parameters is large, which is not conducive to running on resource-limited devices. To summarize, the aforementioned image segmentation methods have the following drawbacks. First, methods involving the use of traditional convolutional neural networks mainly focus on local perception domains; they lack global perception and do not assign different weights to different parts of the input. Although an attention mechanism places more emphasis on the global perception domain, the transformer method has high computational complexity, which leads to poor segmentation performance for mural images. Second, the continuous downsampling operation and high sampling rate of the atrous convolution processes in feature extraction networks can result in the loss of spatial detail information, leading to the loss and incorrect segmentation of small-scale mural targets. Furthermore, the encoder module cannot fully integrate features with different scales during the process of gradually upsampling feature maps to obtain prediction results and therefore cannot satisfactorily capture the boundaries, textures, and details of mural objects, thereby affecting the accuracy and robustness of mural segmentation models.

To address the blurred segmentation boundaries, missing details, and losses of small targets in murals during the mural segmentation process, this study proposes a multiscale feature fusion and dual attention-augmented segmentation model (MFAM) for ancient mural image segmentation based on the codec structure as its basic framework. The main contributions of this study are as follows.

(1) A lightweight transformer network called the MobileViT is introduced [14]. It is used as the main mural feature extraction method and adds a coordinate attention mechanism [15]. This approach can obtain richer mural features and contextual semantic information to improve the efficiency of the mural feature extraction process, save memory, and facilitate model deployment on resource-constrained devices.
(2) An attention-optimized residual atrous spatial pyramid pooling module (A_R_ASPP) is proposed. The residual structure is added to learn the residuals between the input and output features to enhance the feature expression ability of the model. The deep separable convolution [16] replaces the standard convolution to reduce the number of required model parameters, and an attention mechanism is used to adaptively adjust the feature map weight to improve the segmentation accuracy of the model.
(3) A dual attention-enhanced feature fusion module called the DAFM is proposed. The cross-level aggregation strategy is used to limit the number of computations, and the attention strategy is used to weight the importance of different feature levels to obtain an efficient multilevel representation. The module is lightweight and can be used repeatedly in multiscale feature fusion tasks; moreover, it can fully integrate the detailed and semantic information of features at different scales to provide enhanced mural segmentation details.

Theory

Multiscale feature fusion

Multiscale feature fusion is a process that combines feature images with different scales to obtain comprehensive

and hierarchical feature representations. Features at different scales can reflect information at different image levels. Lower-level features mainly map basic information such as the edges, texture and color of the input image; these features usually have small spatial scopes and high degrees of locality. High-level features mainly map the semantic information of an object, such as its class, posture and shape; these features have large spatial ranges and relatively low degrees of locality and can be used to process global information. Multiscale feature fusion can make full use of the information reflected by features with different scales to better understand the given image, and multiple information scales can complement and fuse with each other to obtain a more comprehensive and richer feature representation to improve the image processing performance of the model.

In visual tasks, multiscale feature fusion greatly improves the feature information loss and target scale uncertainty caused by single-scale features, which in turn enables the model to comprehensively and accurately capture the spatial hierarchies of images. ICNet [17] addresses feature maps with different resolutions by adding additional branches and combining them to fill in the missing spatial details among high-level features. It uses CFF and cascade labels to guide the high-resolution feature integration process and gradually extracts rough low-resolution semantic maps. HRNet [18] adopts a high-resolution feature fusion strategy to concatenate feature maps with different scales and performs multiple upsampling operations to finally obtain high-resolution feature representations and retain rich high-precision information. U-Net [19] uses a structure similar to that of an autoencoder to implement a skip connection between the high-resolution and low-resolution feature maps to achieve multiscale feature fusion. An FPN [20] fuses the features of the previous layer after performing upsampling with the features of the current layer possessing the same resolution by adding top-down paths and lateral connections. This approach makes use of both the strong semantic features of the top layer and the high-resolution information of the bottom layer. RANet [21] improves upon traditional methods on the basis of a two-dimensional multiscale network architecture and designs subnetworks with different input resolutions. When a current subnetwork fails to reach a given label, the higher-level subnetwork reuses and fuses the coarse-grained features of the previous subnetwork to obtain more feature representations.

The commonly used multiscale feature fusion method is similar to deep layer aggregation (DLA) [22]. It uses the bilinear interpolation method to upsample features and then splices or adds them to other features according to their channel dimensions. These methods simply aggregate the feature information acquired at multiple scales and ignore the representation gaps between features with different scales to a certain extent, which limits the effectiveness of feature information propagation and easily produces the feature aliasing effect. Although Zhang et al. [23] used gates to control the information propagation process, they still ignored the computational cost limitation while maintaining effectiveness. This paper proposes an attention-enhanced feature fusion module to compensate for the information gap between features with different scales; this approach possesses high adaptability and efficiency.

Applications of attention mechanisms in image processing models

An attention mechanism is a widely used technology in the field of machine learning and deep learning. It can help a model focus on important parts when dealing with complex data, so it can optimize the effect and accuracy of the model. In the image processing task, an attention mechanism strengthens or weakens the weight of a specific region according to the importance of the input information to selectively focus on the key image information. When segmenting mural images, it is often necessary to stack and fuse mural features with different scales to enhance the feature expression ability of the utilized model. The addition of an attention mechanism can help the model focus on the most distinguishing features and capture the features of different regions, enabling it to achieve improved segmentation accuracy. Additionally, an attention mechanism can effectively alleviate the feature redundancy caused by useless feature information stacking.

Attention mechanisms have been widely used in many computer vision models to improve the feature expression capabilities of networks. Through the weight distribution of the features extracted by the utilized neural network at different levels, the learning ability of the network is strengthened with a certain relationship. The squeeze operation of SENet [24] uses global average pooling to learn the dependencies between channels and uses multiple fully connected layers to reweight each channel through an excitation operation; thus, it can adaptively adjust the weights of different channels and improve the ability of the model to express feature information. OCNet [25] was proposed as a new semantic segmentation method that no longer performs prediction in a pixel-by-pixel manner but rather aggregates similar pixels. The proposed objective semantic pooling mechanism (OCP) uses other pixels belonging to the same category to characterize each pixel, thereby capturing long-distance dependencies in the input image. CCNet [26] utilizes a repeated cross-cross-attention module (RCCA),

which calculates the relationships between the target feature pixel and all other points in the feature map, weights the features of the target pixel, and finally obtains more effective target features.

Methods

Overall structure of the network

The mural segmentation model proposed in this paper adopts the methods of multiscale feature fusion and dual attention enhancement; its overall structure is shown in Fig. 1. The model consists of three main parts: an improved CA_MobileViT feature extraction network, an attention-optimized residual atrous spatial pyramid pooling module (A_R_ASPP), and a dual attention-enhanced multiscale feature fusion module (DAFM). f_1-f_5 represent the feature maps generated by the model at different stages. Among these, f_1, f_2 and f_3 are generated and diverted by the backbone CA_MobileViT network to be the inputs of the DAFM model for feature fusion; their sizes are $112\times112\times64$, $56\times56\times96$ and $28\times28\times128$, respectively; f_4 is the feature enhanced by the A_R_ASPP model, whose size is $14\times14\times160$; and f_5 is the feature obtained after three rounds of fusion are performed by the DAFM, and its size is $112\times112\times32$.

First, CA_MobileViT is used as the feature extraction network module of the model, which performs preliminary feature extraction on the given mural feature information. CA_MobileViT models the local and global information of the input mural image features and mitigates the high calculation cost of the ViT [27] model. During feature extraction, the features are shunted in a multiscale manner and used for multiscale feature fusion in the subsequent decoder modules. Moreover, the preliminary extracted features are sent to the A_R_ASPP module for multiscale feature enhancement to obtain higher-level semantic context information. Finally, in the decoder stage, multiscale feature fusion is performed at four different feature scales, which not only makes full use of the semantic information of the high-level features but also integrates the location information of the medium- and low-level features so that the model can accurately identify mural targets with different scales and segment them.

Improvements

CA_MobileViT feature extraction module

Due to the remarkable amount of noise contained in mural images and the unique characteristics of murals in terms of color and texture compared to those of other images, fully extracting local features from murals and thus effectively modeling long-distance information relationships for mural segmentation tasks are critical aspects of mural segmentation. Although the transformer architecture has a satisfactory effect on modeling long-distance mural information, it has drawbacks, such as its high computational complexity and the difficulty of its training process. In addition, the backbone features of traditional convolutional neural networks fail to effectively model long-distance mural feature information. Therefore, this study employs CA_MobileViT as the backbone network for extracting mural features. The CA_MobileViT feature extraction module incorporates the coordinate attention mechanism into the lightweight MobileViT network and fine-tunes it. While ensuring the

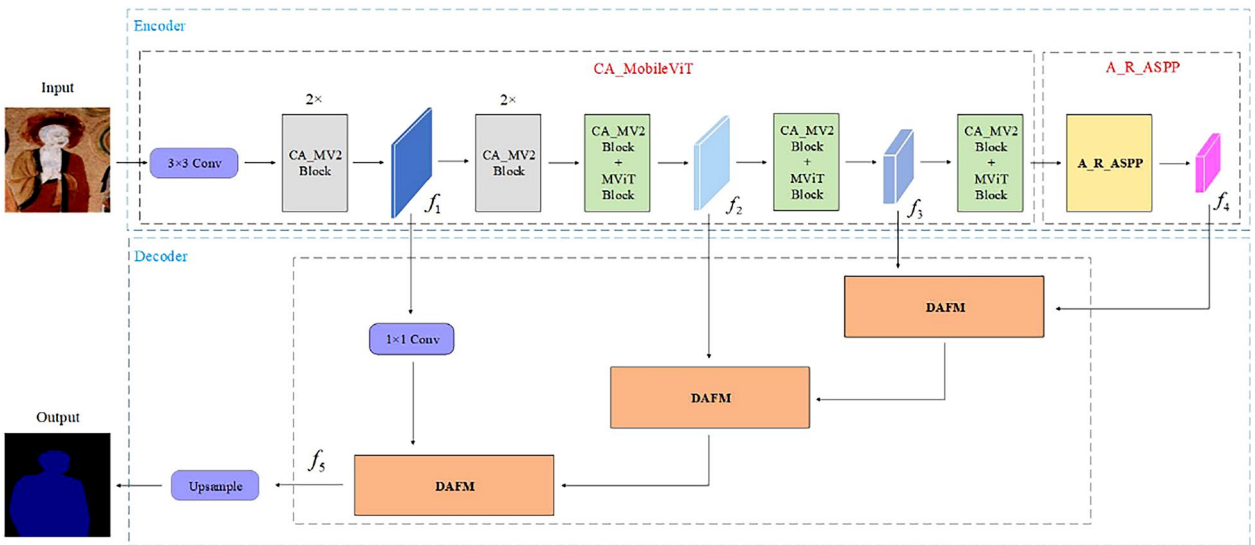

Fig. 1 The overall network structure. DAFM: the dual attention-enhanced multiscale feature fusion module

extraction effect for local and global mural features, this approach greatly reduces the number of required network parameters and improves the training speed to satisfy the adaptation requirements of the model for devices with limited computing resources. The MobileViT network uses a transformer for processing image information via convolution to construct an expression of the local and global information contained in the mural features. The coordinate attention mechanism is employed to precisely encode the location information of the mural into a neural network to model the channel relationships and long-term dependencies of the mural features. The configuration of CA_MobileViT is shown in Fig. 2.

CA_MobileViT comprises a CA_MV2 block and a MobileViT block. In the CA_MV2 block, the channel attention mechanism is decomposed into two one-dimensional feature codes, which aggregate features along two spatial feature directions. With this method, remote dependencies can be captured along one spatial direction while retaining accurate location information in the other spatial direction, and the features f^h and f^w can be obtained via Formula (1). z^h and z^w are the pooling codes of the input $x_c(i,j)$ along the horizontal and vertical coordinates, respectively. [] is the cascading operation implemented along the spatial dimension, F_1 is a 1×1 convolution transformation function, and δ is a nonlinear activation function. Then, the obtained feature map is separately encoded into a pair of direction-aware and position-sensitive attention maps g^h and g^w, namely, Formula (2), where f^h and f^w represent attention weights along the two spatial directions, σ represents the sigmoid activation function, and F_h and F_w represent 1×1 convolution transforms. Finally, the attention map is complementarily applied to the input feature map to enhance the representations of the target objects, namely, Formula (3), where $y_c(i,j)$ represents the output of the coordinate attention block.

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{1}$$

$$g^h = \sigma\left(F_h\left(f^h\right)\right), g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{2}$$

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{3}$$

The core of the MobileViT block includes local and global representations. The local representation module consists of an n×n standard convolution and a 1×1 convolution. The global characterization component of the structure is shown in Fig. 3. The input image is divided into patches with sizes of $h \times w \times d$, which are unfolded and subsequently subjected to the transformer operation to focus on global information. Finally, the concatenation operation is performed with the original input, and an n×n convolution is implemented to obtain the output. The result contains both the local and global information of the input image. The unfolding operation involves unfolding X_L into X_U, where folding is the inverse operation of unfolding.

Attention-optimized residual atrous spatial pyramid pooling module (A_R_ASPP)

The direct use of segmentation methods from other fields for mural images cannot achieve satisfactory

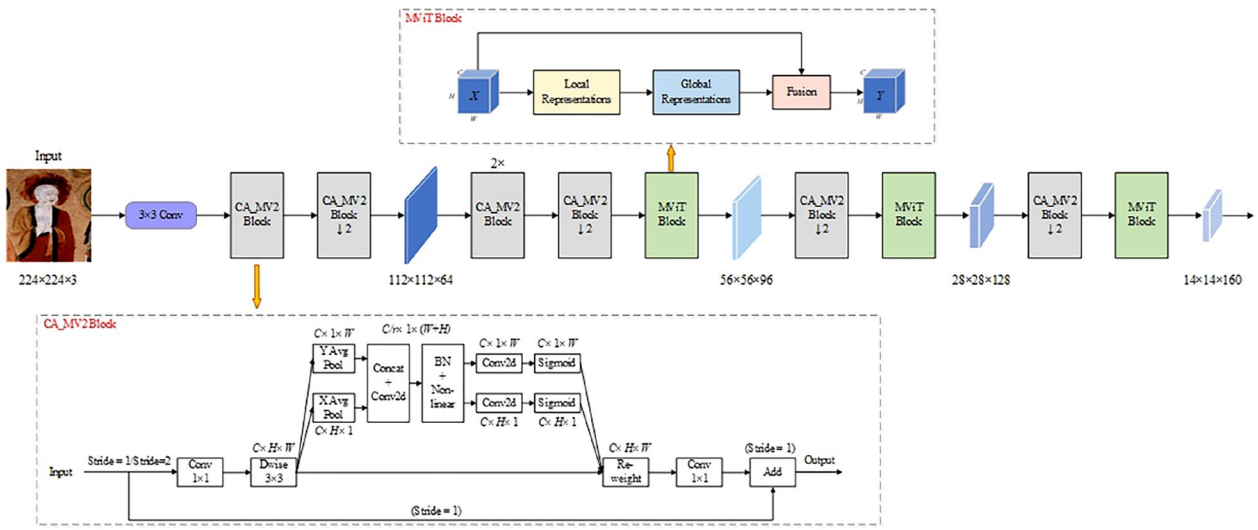

Fig. 2 The structure of CA_MobileViT

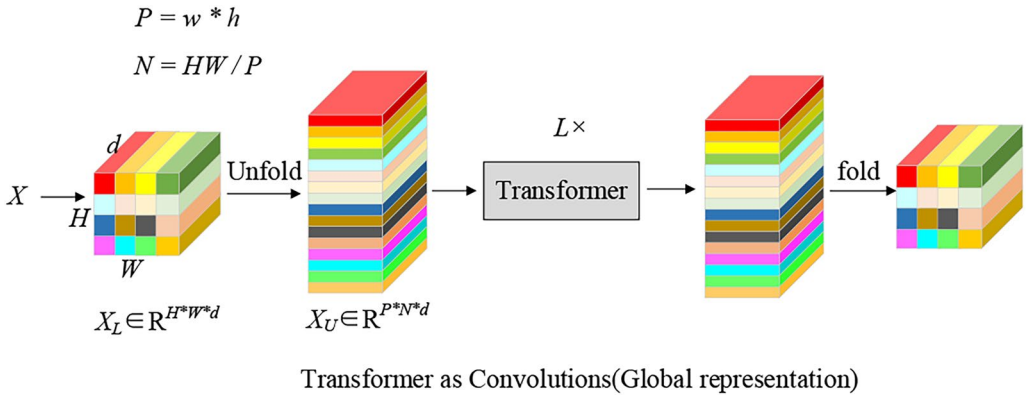

Fig. 3 The global representation structure

segmentation effects, and repeatedly conducting downsampling during the feature extraction process and the employment of atrous convolution are also likely to result in target losses for small-scale murals. To enhance the information extracted by CA_MobileViT and to solve the problems of missing targets in small-scale murals and feature information redundancy, the ASPP multiscale feature enhancement module is introduced in this study, and its structure is shown in Fig. 4.

The A_R_ASPP module consists of two parts: a residual atrous spatial pyramid pooling module (Res_ASPP module) and a feature attention module. The introduction of residual connections can mitigate the information losses associated with small target objects in murals (caused by atrous convolutions with large sampling rates) and enhance the extraction effect achieved for the target features of murals with different sizes. Additionally, the use of deep separable convolution instead of standard convolution can reduce the number of model parameters and improve the efficiency of the overall model without affecting its segmentation effect. In addition, the attention given to the channel of the multiscale features partially fused by Res _ ASPP can adaptively adjust the weight of the feature map according to the importance of different channels so that more important features can receive more attention and feature redundancy can be reduced. By analyzing the channel attention module in SENet, Wang et al. [28] found that avoiding dimensionality reduction is important for learning channel attention, and appropriate cross-channel interactions

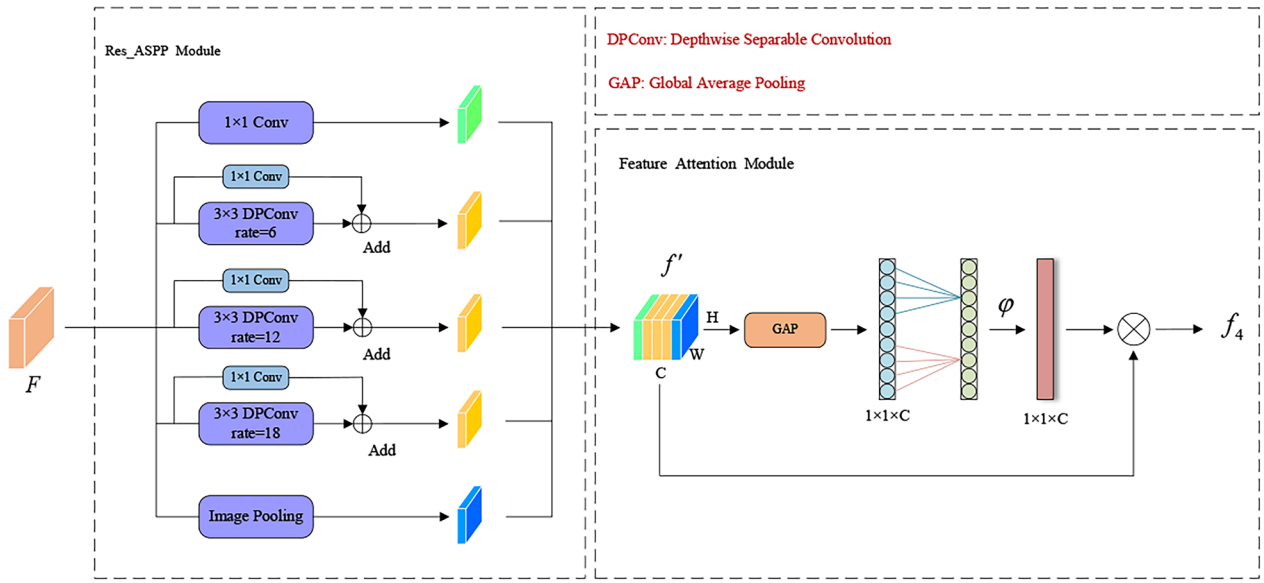

Fig. 4 The structure of the A_R_ASPP module

can significantly reduce model complexity while maintaining performance. Therefore, this paper uses an attention mechanism with a nondimensionalized local cross-channel interaction strategy in the A_R_ASPP module, namely, a feature attention module, which is realized by adaptively selecting the size of the one-dimensional convolution kernel. The processing flow of the feature attention module can be briefly described as follows. The values of the multiscale features f', which are the product of Res_ASPP, are averaged by global average pooling (GAP) for each feature channel to obtain a vector with the same number of channels. After performing feature fusion through GAP, the size of the kernel is first adaptively determined, and this step is followed by one-dimensional convolution. Then, the sigmoid function is used to learn channel attention for obtaining weights. Finally, the normalized weights and features f' are multiplied in a channel-by-channel manner to obtain a weighted feature map, as indicated in Formulas (4) and (5):

$$f_4 = f' \otimes Sigmoid\left(Conv_n\left(GAP\left(f'\right)\right)\right) \tag{4}$$

$$n = \phi(C) = \left[\frac{\log_2(C)}{\tau} + \frac{b}{\tau}\right] \tag{5}$$

where $Conv_n$ represents the one-dimensional convolution operation with n adaptive convolution kernels, C is the number of input feature channels, and b and τ are hyperparameters, which are set to 1 and 2, respectively. In this study, ⌊⌋ represents a downward rounding operation, and ϕ represents the operation of computing adaptive convolution kernels.

DAFM

As the number of network layers in the mural segmentation model increases and the receptive field gradually increases, many detailed mural features, such as boundary and location information, gradually blur after the multilayer networks are convolved. To obtain strong semantic features, traditional segmentation models for mural segmentation usually employ only the feature map of the last layer of the feature extraction network for direct localization and classification without fully utilizing the middle- or low-level features or the details contained at these levels. This treatment tends to lead to blurring in mural edge segmentation tasks, which greatly compromises the segmentation effect.

To solve the problem of feature aliasing caused by simply aggregating features with multiple scales, a dual attention-enhanced feature fusion module (DAFM) is proposed for the above problems. Fully utilizing the detailed and semantic information contained in features of different scales to optimize mural segmentation boundaries, this module can perform multiscale feature fusion on features with different scales and simultaneously solve the problem of feature aliasing from the channel and spatial perspectives of the features; this problem is typically caused by the simple aggregation of features with multiple scales. In addition, the DAFM is designed in a lightweight manner and can be used multiple times in the model. The DAFM is used as the model decoder in this paper to perform multiscale feature fusion on features with multiple scales. Utilizing this module, each pixel can select independent context information from the multilevel features in the fusion stage. The module structure is shown in Fig. 5.

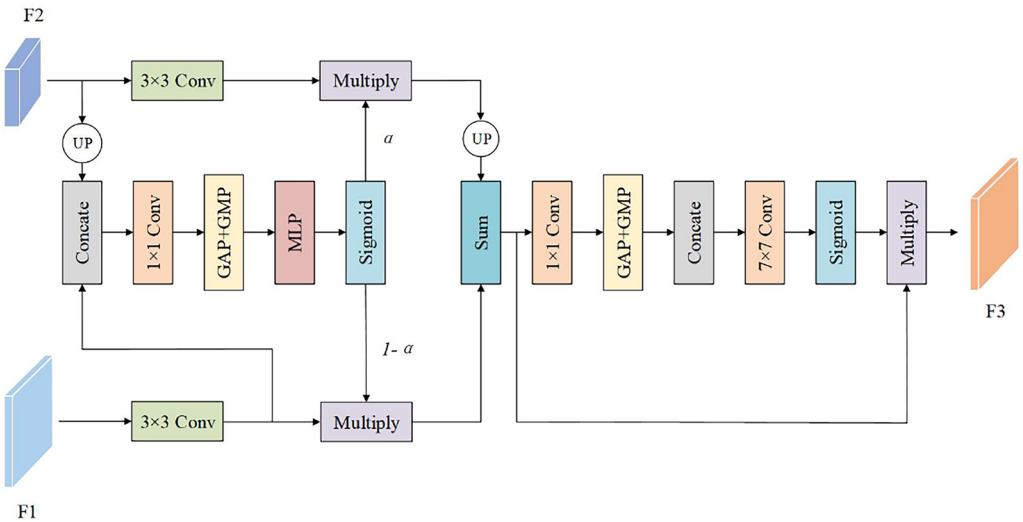

Fig. 5 The DAFM structure

The DAFM is composed of channel and spatial attention mechanisms. Channel attention uses global maximum pooling (GMP) and GAP to process mural features but omits the spatial information of mural objects. Therefore, after implementing channel attention, spatial attention is used to focus on the important areas instead of treating the entire image equally. In the channel attention section, the DAFM is connected in parallel, and the bilinear interpolation method is used to sample feature map F2 and stitch it with feature map F1. Two 1×1×C feature maps are obtained via the GAP and GMP operations. Then, they are input into a three-layer neural network (MLP), and the relative attention weights α are predicted by the sigmoid activation function. The two weights are multiplied by the channels that run in parallel and then added to obtain intermediate results. The intermediate results are focused on through the spatial attention mechanism, and finally, the DAFM output is obtained. This module uses different levels of relative attention masks to guide the fusion of the two features, which compensates for the semantic and resolution gaps between multiscale features. The combination of channel attention and spatial attention can more comprehensively and accurately capture important features.

In the feature extraction stage, the CA_MobileViT model splits the downsampled features 2, 4, 8, and 16 times. After 16 downsampling steps, the features enter the A_R_ASPP module. For the four different feature scales obtained after shunting, the DAFM is used for multiscale feature fusion. This approach can make full use of the semantic information and detailed information contained in the high-level and low-level features to optimize the effectiveness and accuracy of mural segmentation. The fusion process can be expressed by Formula (6), where M represents the operation of the DAFM:

$$f_5 = M_3\left(f_1, M_2\left(f_2, M_1\left(f_3, f_4\right)\right)\right) \tag{6}$$

where f_1-f_5 are the feature maps with different scales generated by the model during different stages.

Experimental results and analysis

Experimental design

Experimental datasets and experimental parameter configuration

The datasets used in the experiment include PASCAL VOC 2012 and an ancient Chinese mural dataset. The training set, test set and verification set of the PASCAL VOC 2012 dataset are composed of 1464, 1449 and 1456 images, respectively. This dataset contains 20 subject categories, such as cars, people, cats, cattle, horses and airplanes, and one background category.

The original images in the ancient Chinese mural dataset are obtained from scanning the images of the picture book titled "The complete collection of Dunhuang murals in China". The images are selected, and severely eroded images are removed. The images are subsequently cropped with Adobe Photoshop (2021) to the same size with a resolution of 224×224. The images are annotated with LabelMe (an image annotation tool) and transformed into JSON files, which are subsequently transformed into grayscale images in batches. The ancient Chinese mural dataset consists of cropped images and transformed grayscale images. Six different categories of images are contained in this dataset, namely, animals, architecture, auspicious clouds, believers, Buddha and other common mural images. To solve the overfitting problem caused by the use of a small dataset to train ancient mural segmentation models, data enhancement is used to expand the dataset, which contains a total of 2400 images. The ratio of the training set to the validation set is 9:1. The specific dataset information is shown in Table 1, and the data-enhanced images are shown in Fig. 6, including the original image and images impacted by inversion, Gaussian noise, and darkening.

The experiment is implemented on the Windows 11 operating system. The CPU is an Intel Core i7-12700, 16 GB of memory is utilized, and the GPU is an NVIDIA GeForce RTX 3070; the software environment includes Python 3.9 based on PyTorch-1.11.0+cuda-12.0 as the basic framework of the experiment.

Since relatively few training images are obtained from the homemade mural dataset, the model in this paper uses the pretrained weights from the ImageNet dataset to initialize the feature extraction backbone. When training on a homemade mural dataset, a stochastic gradient descent (SGD) optimization algorithm is used to update the gradient of the network. The cross-entropy (CE) loss is used as the loss function during the training process, and the cosine annealing (CosineAnnealingLR) strategy is used to adjust the learning rate.

Experimental evaluation indicators

In the experiment, the mean intersection over union (MIoU), mean pixel accuracy (MPA) and Dice coefficient are used as the evaluation indices to measure the objective

Table 1 Dataset details

Label	Animal	Build	Cloud	Disciple	Buddha	Background
Number	520	500	500	480	480	2400

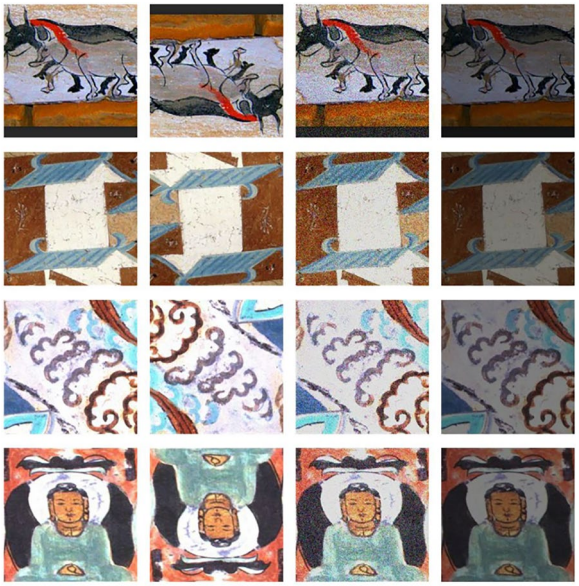

Fig. 6 Data augmentation

experimental results, and the numbers of frames per second (FPS) and parameters are used as the evaluation indices to measure the computational complexity of the network. The intersection over union (IoU) is the intersection of the real value and the predicted value of a pixel divided by the union of the real value and the predicted value of the pixel, while the MIoU is the intersection over union (IoU) of the real label and the predicted result for each class. Then, the mean IoU of all categories is calculated. The MIoU is a standard accuracy metric that reflects the degree of coincidence between a model segmentation result and the real value of the original image. The average pixel accuracy represents the average proportion of correctly classified pixels in each class in the picture. The Dice coefficient is a set similarity measurement function that is often used to calculate the similarity between two samples, while the IoU, MIoU, and MPA are defined as follows in Formulas (7), (8), and (9), (10) respectively:

$$IoU = \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij} - \sum_{j=0}^{k} P_{ji} - P_{ii}} \tag{7}$$

$$MIoU = \frac{\sum_{i=0}^{k} IoU}{k+1} \tag{8}$$

Table 2 Comparison among the effects of different segmentation models on the PASCAL VOC dataset

Model	Backbone	MIoU/%	MPA/%	Parameters/M
SegNet	VGG16	68.47	76.56	190.7
DeepLabV3	MobileNetV2	71.2	81.39	5.02
PSPNet	ResNet50	80.11	89.62	54.58
DeepLabV3 +	Xception	80.73	89.12	60.43
DANet	ResNet50	78.5	87.64	48.26
SETR	ViT-L	78.84	88.15	260.1
DAFPN	Swin-T	**80.8**	**89.7**	158
MFAM (Ours)	CA_MobileViT	80.25	88.76	9.5

MioU: mean intersection over union; MPA: mean pixel accuracy (MPA). The bold number indicates the highest value among the evaluation indicators

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ij}}{\sum_{j=0}^{k} P_{ij}} \tag{9}$$

$$Dice_i = \frac{2 \times P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji}} \tag{10}$$

where $k+1$ means that there are $k+1$ categories, including the background involved in the segmentation task, and P_{ij} indicates that the pixels of category i are predicted to be the number of categories j.

Comparative experiments and analysis with the existing methods

To better verify the universality and efficiency of the proposed mural segmentation model in segmentation tasks, classic segmentation models and transformer models that have emerged in recent years, such as the FCN [6], SegNet [7], DeepLabV3 [29], PSPNet [8], DeepLabV3 + [10], DANet [12], MC-DM [11], SETR [13] and DAFPN [30], are selected for experimental comparisons conducted on the PASCAL VOC and mural datasets. Additionally, quantitative and qualitative analyses are used to compare the experimental results.

First, the PASCAL VOC dataset is used for training and comparing the proposed approach with SegNet, DeepLabV3, PSPNet, DeepLabV3 +, DANet, SETR, DAFPN and other segmentation networks. The results of the comparative analysis are shown in Table 2. The experiments show that, compared with the DeepLabV3 + model, the proposed model reduces the number of required parameters with almost no reduction in its average intersection–union ratio or average pixel accuracy; it uses only 1/6 of the parameters required by the DeepLabV3 + model. Compared with the DANet model, the proposed model improves the MIoU and MPA by

1.75% and 1.12%, respectively. Compared with those of the DeepLabV3 model, which uses MobileNetV2 [31] as its feature extraction network, under the condition of a small increase in the number of parameters, the MIoU and MPA of the proposed model are greatly improved (by 9.05% and 7.37%, respectively). Furthermore, the number of parameters involved in the model proposed in this study is no more than 1/16 that of the DAFPN model, but it achieves comparable outcomes in terms of the MIoU and MPA. Therefore, the model proposed in this paper achieves good performance in terms of balancing the segmentation effect, the segmentation accuracy and the number of network parameters and has good universality in segmentation tasks.

Second, to verify the effectiveness of the proposed method for use in ancient mural segmentation tasks, the classic semantic segmentation models (the FCN, SegNet, PSPNet, and DeepLabV3+) and the advanced semantic segmentation models (DANet, MC-DM, SETR and DAFPN) developed in recent years are selected for comparison. With the same environmental configuration, all the models are trained and perform prediction on the ancient mural dataset. Upon comparing the segmentation performances of different models, the experimental results obtained in terms of the MIoU, MPA and number of frames per second are shown in Table 3. The proposed method is superior to the other comparison methods in terms of the MIoU and MPA. The MIoU and MPA reach 88.19% and 95.66%, respectively, and the FPS is also higher than those of the other approaches.

Table 3 Performance comparison among different segmentation methods

Model	Backbone	MIoU/%	MPA/%	FPS
FCN	VGG16	77.36		
SegNet	VGG16	82.23	92.84	17.26
PSPNet	ResNet50	85.07	93.8	22.17
DeepLabV3+	Xception	85.75	94.37	41.54
MC-DM	MobileNetV2	84.42	93.71	56.52
DANet	ResNet50	85.2	94.53	**67.31**
SETR	ViT-L	85.14	94.2	15.65
DAFPN	Swin-T	86.82	95.25	26.8
MFAM (Ours)	CA_MobileViT	**88.19**	**95.66**	45.43

The bold number indicates the highest value among the evaluation indicators

Then, the segmentation effect of the model proposed in this study is separately verified on the five groups of images contained in the mural dataset (namely, "animal", "building", "cloud", "disciple" and "buddha"), with the Dice coefficient as the assessment index. According to the results (Table 4), our method outperforms the other models, particularly in terms of its segmentation results produced for the animal and building groups, which have more complex boundary contours than do the other groups of images. This satisfactory effect attained for these groups can be attributed to the use of a multiscale feature fusion module (MFAM), which is able to make full use of the spatial boundary information contained in low-resolution features, thereby optimizing the segmentation details of mural target boundaries.

Additionally, one image is randomly selected from each of the five groups for comparison (deer, pagoda, cloud, disciple and Buddha statue images, respectively). Figure 7 shows the visual comparison between the proposed method and the above image segmentation methods on the five groups of images. The images from left to right are the input image, the label image, the segmentation mask of the compared model and the segmentation mask of the proposed model, and the red rectangular boxes represent the regions where noticeable segmentation effect differences can be observed. For the deer mural, the method proposed in this study and the DAFPN are more accurate than the remaining methods in terms of segmenting limbs. SegNet, MC-DM, DANet, and SETR all exhibit some degree of detail loss. Although SETR and DAFPN yield satisfactory segmentation results for the head area, both methods yield incorrect segmentation results. For the Buddhist pagoda mural, incorrect segmentation results are produced by in the DAFPN and SETR methods, which mistakenly identify the background noise area as a part of the building. The method proposed in this study and DANet achieve the most accurate segmentation effects on the top areas of the pagoda.

Table 4 Comparison among the Dice coefficients of different models

Model	Backbone	Animal	Building	Cloud	Disciple	Buddha
FCN	VGG16	0.812	0.784	0.841	0.825	0.763
SegNet	VGG16	0.854	0.817	0.876	0.86	0.792
DeepLabV3+	Xception	0.873	0.834	0.915	0.871	0.817
MC-DM	MobileNetV2	0.867	0.832	0.872	0.881	0.81
SETR	ViT-L	0.88	0.853	0.876	0.864	0.825
DAFPN	Swin-T	0.907	0.851	0.874	0.906	0.822
MFAM (Ours)	CA_MobileViT	0.935	0.898	0.875	0.915	0.852

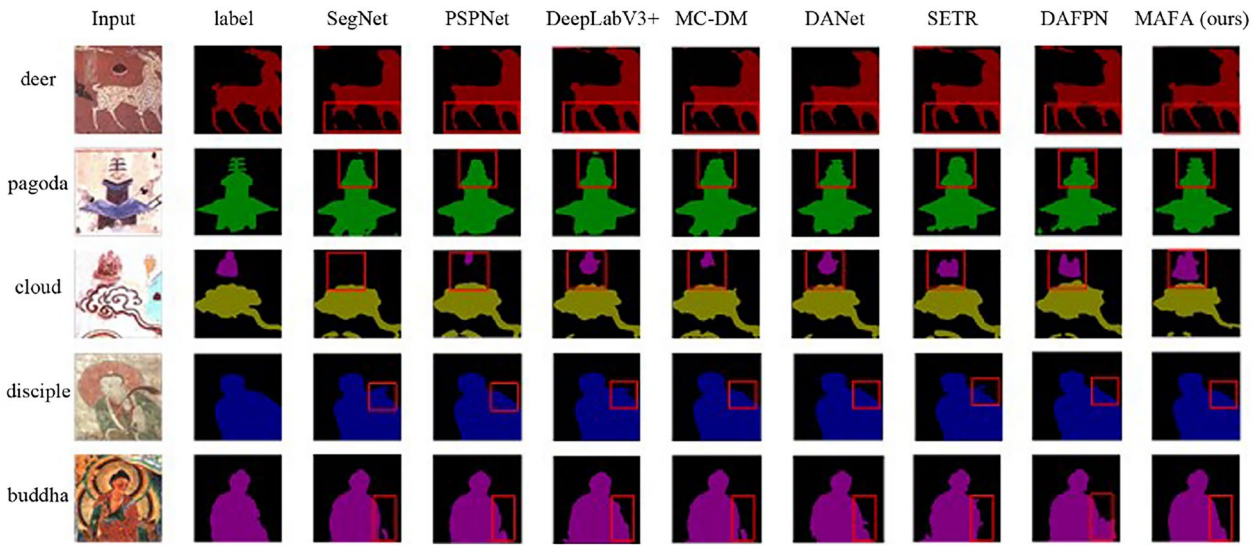

Fig. 7 Visual comparison among the segmentation effects of various models on the mural dataset

For the cloud image, all methods show some degree of segmentation error. For the segmentation of the Buddha statue with significant scale changes, our method yields results that are closest to the tag value, while SegNet even cannot recognize its category. For the disciple mural, the method proposed in this study and DAFPN achieve more accurate edge contour segmentation because of their full utilization of detailed features. For the Buddha mural image, SegNet, PSPNet, and MC-DM exhibit segmentation errors, and SETR segments the edges too roughly. In contrast, DeepLabV3 +, DAFPN, and our method achieve segmentation results that are closest to the authentic values. Based on the overall visual effect analysis and comparison, our method outperforms the other methods in mural image segmentation tasks.

Ablation experiment

To verify the effectiveness of the feature extraction backbone network in this paper, the CA_MobileViT network is classified and compared with the commonly used feature extraction network on the ImageNet-1 k dataset. The model parameters can reflect the model structure and affect the memory or display memory needed for the model to operate. The classification accuracy achieved by the model on the ImageNet-1 k dataset can reflect the ability of the model to extract features to a certain extent. The comparison results are shown in Table 5. Table 5 shows that, compared with those of the Xception, InceptionV3, ResNet-50 and other network models, the number of parameters in the CA_MobileViT network model is only ¼ the total, but its classification accuracy is greater than those of Inception and ResNet-50, and the results are similar to those of Xception. Compared with that of the lightweight MobileNetV2 neural network, the classification accuracy improvement provided by our model is also obvious.

Table 5 Network model comparison

Model	Parameters/M	Top-1/%
Xception [32]	22.9	79
VGG-16 [33]	138	74.4
ResNet-50 [34]	25	75.3
MobileNetV2	3.5	72
CA_MobileViT	5.8	78.5

To further verify the ability of the CA_MobileViT network to extract features from mural images, the above-mentioned feature extraction network and the mural feature map extracted by the CA_MobileViT network are used for a visual comparison. The comparison results are shown in Fig. 8. Figure 8 shows that the mural feature maps extracted by MobileNetV2 and VGG16 are blurred, and the textural details are not sufficiently obvious. Although the ResNet50 network can extract the features of deer and pagodas more completely, its edge feature extraction ability is insufficient. The CA_MobileViT network is more sensitive to boundary coordinate positioning information due to the addition of coordinate attention, and it can fully extract the edge features of the target mural. This experiment verifies the superiority of the CA_MobileViT network as a feature extraction backbone for extracting the features of ancient murals.

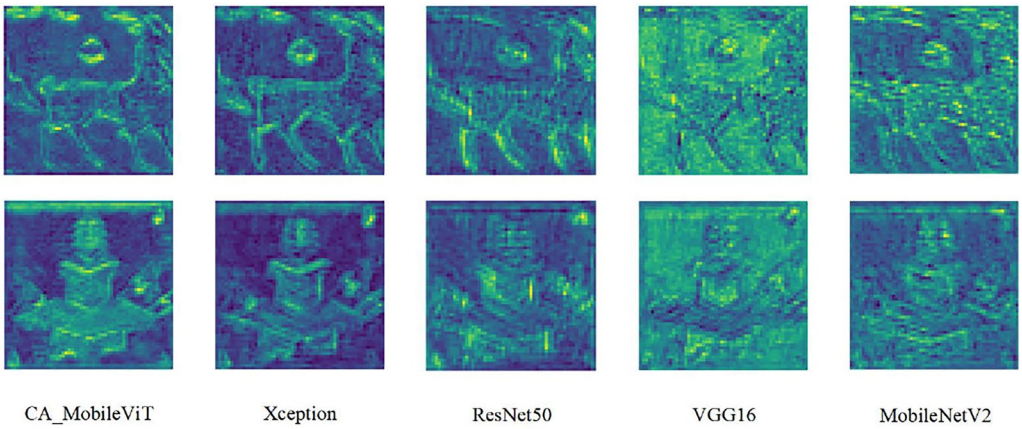

Fig. 8 Comparison among the mural feature map extraction effects of different models

Table 6 Results of ablation experiments conducted on each module

A_R_ASPP	DAFM	MIoU (%)	Δa (%)
		84.78	
√		86.62	1.84
	√	86.7	1.92
√	√	88.19	3.41

To prove the effectiveness of the attention-optimized residual ASPP module (A_R_ASPP) and the dual attention-enhanced feature fusion module (DAFM) proposed in this paper, ablation experiments are carried out, and the results are shown in Table 6. We use the DeepLabV3 + network as the benchmark method, which employs the standard empty space, pyramid pooling structure and feature stacking fusion module. The change involves replacing the basic feature extraction network with the CA_MobileViT network.

When the attention-optimized residual atrous spatial pyramid pooling module (A_R_ASPP) is added to the benchmark method, the MIoU performance index is improved by 1.84%; when the dual attention-enhanced feature fusion module (DAFM) is added to the benchmark method, and performance improve by 1.92%. When the A_R_ASPP module and the DAFM are simultaneously added to the benchmark method, the performance improves by 3.41%, and the average intersection ratio of the network reaches 88.19%. Therefore, the attention-optimized residual atrous spatial pyramid pooling module and the dual attention-enhanced feature fusion module proposed in this paper are effective.

A comparison among the segmentation effects produced for objects with different scales is shown in Fig. 9 (the red rectangular boxes are the regions where noticeable differences in the obtained segmentation effects can be observed). The model proposed in this paper uses the CA_MobileViT network to extract mural features via its feature extraction module and obtains four different levels of feature information. By introducing coordinate attention and self-attention mechanisms, the model can obtain location information, local features and global feature information to improve the accuracy of its judgments concerning mural features. Second, the extracted mural features must be processed by the A_R_ASPP module. This module not only expands the receptive

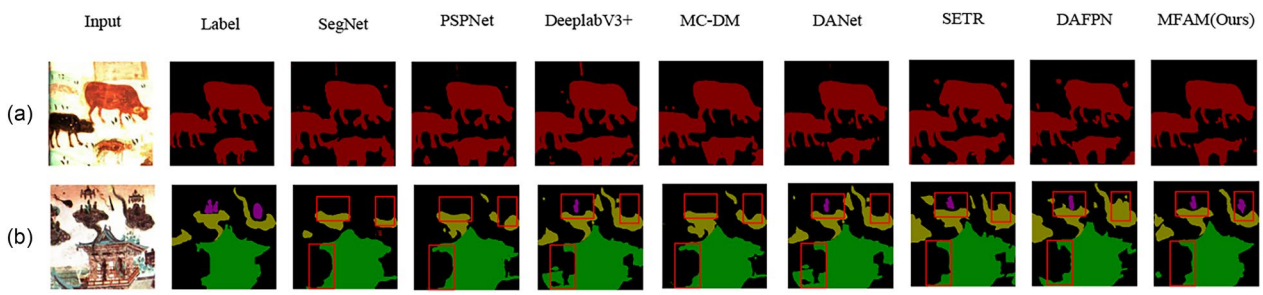

Fig. 9 Comparison among the image segmentation effects produced for objects with different scales

field and enhances the semantic feature information but also effectively alleviates the problem of small target feature losses in murals. Finally, the four different levels of features are fused through the DAFM, which improves the lack of mural contour details during the segmentation process. As shown in Fig. 9a, our method results in fewer segmentation errors and clearer and more accurate contours on mural images containing cows with different scales than do the other methods. For complex mural images, the method proposed in this study can more accurately identify small-scale mural targets for segmentation, and its segmentation results obtained for building edges are more accurate (Fig. 9b).

In summary, the work conducted of this paper is described as follows. First, a CA_MobileViT feature extraction network is inserted into the coordinate information mechanism, and the feature information is globally modeled. Second, an improved A_R_ASPP feature enhancement module effectively compensates for the lack of small target features in murals. A DAFM is used to integrate different levels of features, increase the amount of available feature information and optimize the segmentation boundaries. Compared with other models, the proposed method achieves significant MIoU and MPA improvements, reduces the number of model parameters, and improves the efficiency of mural segmentation. Therefore, this approach has obvious performance advantages in mural image segmentation tasks.

Conclusion

The MFAM model proposed in this paper is an optimization model that is specifically applied to ancient mural image segmentation. The model is improved and optimized to address the fuzzy details, small target losses and low efficiency encountered in ancient mural image segmentation. By introducing a lightweight MobileViT network and adding a coordinate attention mechanism, the model in this paper performs long-distance mural feature information modeling, which improves the feature extraction ability and overall efficiency of the model. Additionally, this paper proposes an attention-optimized residual atrous spatial pyramid pooling module, which can enhance the semantic information of features and solve the problem that small-scale mural targets are easily lost. This paper also proposes a feature fusion module with dual attention enhancement, which is used to fuse multiscale features to compensate for the information gaps between the semantics, locations and boundaries of multiscale features to improve the overall segmentation details. In an experiment, by comparing different models on different datasets, comparing the feature maps extracted by the proposed approach with those of other feature extraction networks and performing qualitative and quantitative analyses of different models and ablation experiments on different modules, the universality and effectiveness of the MFAM mural segmentation model are fully verified. Additionally, on the mural dataset, the MFAM model achieves the highest MIoU and MPA values, reaching 88.19% and 95.66%, respectively. Compared with the classic segmentation models, the MFAM model achieves significant segmentation detail, segmentation accuracy, segmentation efficiency and training time improvements, thus providing a new method for the segmentation of ancient mural images.

However, this method possesses several shortcomings. During the segmentation process, edge blur and oversegmentation problems occur. In Fig. 9, the details of the cattle limbs are blurred, and a small amount of grass is mistakenly identified as a part of the cattle body. These problems are mainly due to the high noise contained in the mural images and the relatively small size of the mural dataset. Consequently, it is impossible to accurately learn mural features. To overcome these problems, future research can consider strengthening the detection and learning of edge information to improve the clarity of the edges in the segmentation results. To improve the accuracy of the network model after training it, the mural dataset can be enriched, and more diverse and higher-quality mural data can be collected for training.

Abbreviations

FCM	Fuzzy C-means clustering
FCN	Fully convolutional neural network
CNN	Convolutional neural network
ASPP	Atrous spatial pyramid pooling
MFAM	Multiscale feature fusion and dual attention-augmented segmentation model
A_R_ASPP	Attention-optimized residual atrous spatial pyramid pooling module
DLA	Deep layer aggregation
OCP	Objective semantic pooling mechanism
RCCA	Repeated cross-cross-attention module
DAFM	Dual attention-enhanced multiscale feature fusion module
Res_ASPP module	Residual atrous spatial pyramid pooling module
GAP	Global average pooling
GMP	Global maximum pooling
SGD	Stochastic gradient descent
CE	Cross-entropy
MIoU	Mean intersection over union
MPA	Mean pixel accuracy
FPS	Frames per second
IoU	Intersection over union

Acknowledgements
Not applicable.

Author contributions
All the authors have contributed to the current work. CJF devised the study plan and led the writing of the article. CZ, CZQ, and WF conducted the experiment and prepared Figs. 1–9. WXH and YZL conducted the analysis, and CJF supervised the whole process and provided constructive advice. All the authors read and approved the final manuscript.

Funding
This study was funded by the National Natural Science Foundation of China (Grant no. 62372397), the Humanities and Social Sciences Research Project of the Ministry of Education (Planning Fund Project) (Grant no. 21YJAZH002), and the Shanxi Province Natural Fund Project (Grant no. 202203021221222).

Availability of data and materials
The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 14 November 2023 Accepted: 7 February 2024
Published online: 16 February 2024

References

1. Chen GQ, Li YF, Liu R, Chai B, Cui Q, Chai Z, et al. An investigation of the blisters on the murals of the Mogao Grottoes. Dunhuang Res. 2016;03:110–6.
2. Fu XY, Li Y, Sun ZJ, Du J, Wang FP, Xu YQ. Digital color restoration of soot covered murals in the Mogao Grottoes at Dunhuang. Dunhuang Res. 2021;01:137–47.
3. Cao JF, Zhang Q, Cui HY, Zhang ZB. Application of improved GrabCut algorithm in ancient mural segmentation. J Hunan Univ Sci Technol. 2020;35:83–9.
4. Wang XP, Wang QS, Jiao JJ, Liang JC. Fuzzy C-means clustering with fast and adaptive non-local spatial constraint and membership linking for noise image segmentation. J Electron Inf Technol. 2021;43:171–8.
5. Venkatachalam K, Reddy VP, Amudhan M, Raguraman A, Mohan E. An implementation of K-means clustering for efficient image segmentation. 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India; 2021. p. 224–9.
6. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. p. 3431–40.
7. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39:2481–95.
8. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2881–90.
9. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. 2017; arXiv Preprint: 1706.05587.
10. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV); 2018. p. 801–18.
11. Cao JF, Tian XD, Jia Y, Yan M. Application of improved DeepLabV3+ model in mural segmentation. J Comput Appl. 2021;41:1471–6.
12. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 3146–54.
13. Zheng SX, Lu JC, Zhao HS, Zhu XT, Luo ZK, Wang YB, et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. https://doi.org/10.1109/cvpr46437.2021.00681.
14. Metha S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. 2021; arXiv Preprint: 2110.02178.
15. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021. p. 13713–22.
16. He J, Wang X, Song Y, Xiang Q. A multi-scale radar HRRP target recognition method based on pyramid depthwise separable convolution network. 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi'an, China; 2022. p. 579–85.
17. Ai QL, Zhang JR, Wu FQ. AF-ICNet semantic segmentation method for unstructured scenes based on small target category attention mechanism and feature fusion. Acta Photonica Sinica. 2023;52:189–202.
18. Zheng YF, Wang XB, Zhang XW, Cao T, Sun M. The self-distillation HRNet object segmentation based on the pyramid knowledge. Acta Electron Sin. 2023;51:746–56.
19. Zhu YF, Wang HT, Li K, Wu HJ. Crack U-Net: towards high quality pavement crack detection. Comput Sci. 2022;49:204–11.
20. Quyen VT, Kim MY. MPNet: Multiscale predictions based on feature pyramid network for semantic segmentation. 2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN), Paris, France; 2023. p. 114–9.
21. Yang L, Han Y, Chen X, Song S, Dai J, Huang G. Resolution adaptive networks for efficient inference. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 2369–78.
22. Zhang ZH, Dong FM, Hu F, Wu YR, Sun SF. Residual based gated Recurrent unit. Acta Automatica Sinica. 2022;48:3067–74.
23. Yu F, Wang D, Shelhamer E, Darrell T. Deep layer aggregation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018. p. 2403–12.
24. Liu Y. A lane line detection method based on squeeze and excitation network. 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), Guangzhou, China; 2022. p. 117–21.
25. Yuan Y, Huang L, Guo J, Zhang C, Chen X, Wang J. OCNet: object context for semantic segmentation. Int J Comput Vision. 2021;129:2375–98.
26. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W. Ccnet: Criss-cross attention for semantic segmentation. Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 603–12.
27. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at Scale. 2020; arXiv Preprint: 2010.11929.
28. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 11534–42.
29. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. 2017; arXiv Preprint: 1704.04861.
30. Lu L, Xiao Y, Chang XJ, Wang XH, Ren PZ, Ren Z. Deformable attention-oriented feature pyramid network for semantic segmentation. Knowl-Based Syst. 2022;254: 109623.
31. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018. p. 4510–20.
32. Chollet CF. Xception: Deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 1251–8.
33. Zheng LY, Dai YX. m-VGG16: a dermoscopy image segmentation method based on VGG16. Third International Conference on Computer Vision and Data Mining (ICCVDM 2022); 2023, p. 271–6.
34. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–8.

Publisher's Note