mathematics

MDPI

Article

Progressively Multi-Scale Feature Fusion for Image Inpainting

Wu Wen [1], Tianhao Li [1], Amr Tolba [2,*], Ziyi Liu [1] and Kai Shao [3]

[1] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; wenwu@cqupt.edu.cn (W.W.); s210101078@cqupt.edu.cn (T.L.); s220132098@cqupt.edu.cn (Z.L.)
[2] Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia
[3] School of Software, Dalian University of Technology, Dalian 116024, China; shaokai@mail.dlut.edu.cn
* Correspondence: atolba@ksu.edu.sa

Abstract: The rapid advancement of Wise Information Technology of med (WITMED) has made the integration of traditional Chinese medicine tongue diagnosis and computer technology an increasingly significant area of research. The doctor obtains patient's tongue images to make a further diagnosis. However, the tongue image may be broken during the process of collecting the tongue image. Due to the extremely complex texture of the tongue and significant individual differences, existing methods fail to fully obtain sufficient feature information, which result in inaccurate inpainted tongue images. To address this problem, we propose a recurrent tongue image inpainting algorithm based on multi-scale feature fusion called Multi-Scale Fusion Module and Recurrent Attention Mechanism Network (MSFM-RAM-Net). We first propose Multi-Scale Fusion Module (MSFM), which preserves the feature information of tongue images at different scales and enhances the consistency between structures. To simultaneously accelerate the inpainting process and enhance the quality of the inpainted results, Recurrent Attention Mechanism (RAM) is proposed. RAM focuses the network's attention on important areas and uses known information to gradually inpaint image, which can avoid redundant feature information and the problem of texture confusion caused by large missing areas. Finally, we establish a tongue image dataset and use this dataset to qualitatively and quantitatively evaluate the MSFM-RAM-Net. The results shows that the MSFM-RAM-Net has a better effect on tongue image inpainting, with PSNR and SSIM increasing by 2.1% and 3.3%, respectively.

Keywords: tongue image inpainting; MSFM; RAM; tongue image dataset

MSC: 68U10

Citation: Wen, W.; Li, T.; Tolba, A.; Liu, Z.; Shao, K. Progressively Multi-Scale Feature Fusion for Image Inpainting. *Mathematics* **2023**, *11*, 4908. https://doi.org/10.3390/math11244908

Academic Editor: Konstantin Kozlov

Received: 6 November 2023
Revised: 5 December 2023
Accepted: 7 December 2023
Published: 8 December 2023

1. Introduction

With the improvement of science and technology, computer technology is gradually integrated into people's lives, and has a significant impact on medical treatment [1,2], human pose estimation [3,4] and transportation [5,6]. Tongue diagnosis holds a significant position in Traditional Chinese Medicine (TCM) as a pivotal diagnostic technique with a long-standing history of thousands of years. In traditional tongue diagnosis, doctors of TCM diagnose a person's health condition by observing the color, shape and texture characteristics of the tongue. In the past few years, the concept of WITMED has been introduced and increasingly implemented in various medical domains. Nowadays, a various of methods have been applied to tongue image segmentation [7], feature extraction [8] and computer analysis and diagnosis [9]. Since computerized tongue diagnosis is based on analysis of the collected tongue image pictures, this method will have the image blocked or damaged in the process of tongue image acquisition, so the tongue image inpainting has become an indispensable part of computerized tongue diagnosis.

Similar to many computer vision challenges, the advancement of image inpainting precedes the widespread adoption of deep learning techniques. Traditional image inpainting

methods primarily rely on the semantic information available in non-missing regions of the image. They utilize this information to identify pixels and image blocks that exhibit similar features, which can then be used to fill in the missing areas of the image. These methods can be divided into diffusion-based methods [10–12] and patch-based methods [13–15]. However, the aforementioned methods are applicable only to situations where the features are simple and the missing areas are small. When the missing areas become larger or the image textures become complex, these methods cannot generate inpainted content with advanced semantic information. At the same time, the huge model structure often causes a lot of resource consumption and high time delay, which makes the algorithm unable to face daily life [16].

The method of combining deep learning with image tasks has achieved unexpected results. Deep learning-based image inpainting methods can greatly improve the inpainting effect by learning the information of damaged images and the structural features of images. And models based on deep learning can solve tasks that need to be processed on the basis of fast computation [17,18]. To this end, scholars have proposed deep learning-based methods to improve the effectiveness of image inpainting [19–22]. However, tongue images are different from ordinary images and often have extremely complex texture structures, with small differences between the structures of the tongue. At the same time the color and subtle texture of the tongue surface are highly individualized, and each person's tongue is unique. The existing studies are suitable for images with less complex textures, clear structures and little individual differences. Therefore, the above characteristics of the tongue make it impossible for existing studies to accurately inpaint the tongue image. The process of image inpainting is to extract missing image information from known information. Therefore, obtaining both low-level and high-level features of an image simultaneously is beneficial for generating more appropriate semantic information. [23,24]. In addition, since the inpainted image needs to be as consistent as possible with human visual perception, the traces of inpainting that are imperceptible to humans are inpainted. Furthermore, it is necessary to ensure that the speed of image inpainting algorithm is as fast as possible, so the above image inpainting algorithm cannot achieve satisfactory results. Meanwhile, due to limited medical resources and high real-time performance, it is essential to implement a high-performance image inpainting model that can be achieved with few learning samples. Nowadays, many methods have implemented feature learning for few samples to achieve real-time transmission [25,26].

In order to address the issues mentioned above, we have introduced a new deep image inpainting architecture called the Multi-Scale Fusion Module and Recurrent Attention Mechanism Network (MSFM-RAM-Net). In this paper, the Multi-Scale Fusion Module (MSFM) is first proposed to realize the information interaction between image features at different depths. This approach not only captures complex semantic information that is challenging to extract, but also mitigates the issue of shallow feature loss resulting from deepening the network. Moreover, group convolution enables the information between groups to communicate with each other and improve the correlation between image information in the channel direction. It also greatly reduces the number of parameters and computational complexity of the model. To improve the model's ability to inpaint details, we also propose Recurrent Attention Mechanism (RAM). This method is divided into two parts: recurrent mechanism and attention mechanism. The recurrent mechanism takes the output feature map of the network as the input of the entire network to continue learning, achieving progressive inpainting of damaged images and improving the filling ability of detailed information. Compared with multi-stage tasks, the parameters of progressive inpainting are greatly reduced, which can improve the rate of convergence of the network. The attention mechanism reweights feature channels, suppresses irrelevant features in the image and focuses the network on important features.

In summary, the contributions of this paper include the following three aspects.

1. we propose a Multi-Scale Fusion Module (MSFM). This model effectively captures both low-dimensional and high-dimensional image features and greatly preserves a

significant amount of detailed information within the image. Moreover, through the interaction of semantic information in the channel direction, the output feature map has stronger expressive ability;

2. we propose the Recurrent Attention Mechanism (RAM) to make images inpainted greatly. This method gradually achieves image inpainting from the outside to the inside through the known information of the image. And strengthen the attention of important feature information in the channel direction, so that more reasonable texture structures can be inpainted during the inpainting process.

The rest of this paper is organized as follows: Section 2 reviews the related work. Section 3 introduces the overall framework design of the method. Section 4 provides experimental evaluation. Section 5 discusses the work of this paper.

2. Related Work

In this section, we review previous research work.

2.1. Traditional Image Inpainting Algorithms

The traditional image inpainting methods consist of two parts: diffusion-based methods and patch-based methods.

The diffusion-based methods capitalizes on the high similarity between neighboring pixels in an image to match the pixels surrounding the damaged area with the region to be inpainted. By utilizing this matching process, the algorithm generates the inpainted results. Bertalmio et al. [10] proposed an algorithm using partial differential equations in 2000. This algorithm uses individual pixels and propagates known information along the isophotes to inpaint the damaged area. However, due to the lack of consideration for details and consistency in image structure, the inpainting work is not ideal. In 2017, Li et al. [22] presented a method that begins by detecting the diffusion of the inpainting region and subsequently establishes a feature set by considering the local variance of changes within and between channels. These features are utilized to identify and delineate the inpainted area. The diffusion-based inpainting algorithm only considers the relationship between missing areas and adjacent pixels, but cannot achieve ideal inpainted results when it comes to large and rich texture features.

The patch-based methods are to calculate the similarity between patches, and then use patches with high similarity to reconstruct missing areas of the image. In 2009, Barnes et al. [13] proposed a random algorithm for quickly matching approximately adjacent points between image patches. Find matching points through sampling, then use the consistency in natural images to quickly propagate matching points in the surrounding area to find the most suitable patch. In 2018, Liu et al. [27] employed statistical regularization and similarity measures to extract the dominant linear structure of the target region. Subsequently, they employed a Markov random field model to inpaint the missing area. This method ensured the neighborhood consistency and structural consistency of the inpainted area.

Traditional image inpainting methods only achieve good inpainted results in scenes where the missing area is small, the texture structure of image is simple and the similarity between the rest of information and the missing area is high. However, for tasks with large damaged areas and complex texture structures in images, the results obtained are poor. Therefore, people propose to combine image inpainting with deep learning.

2.2. Inpainting Based on Deep Learning

The advent of deep learning has led to significant advancements, and methods rooted in deep learning have progressively emerged as the dominant approach [28–31]. Since deep learning-based image inpainting methods can perform feature-level processing on images, providing assurance for generating semantically continuous and structurally reasonable results for image inpainting tasks, an image inpainting method based on GAN [32] has been proposed.

In 2016, Pathak et al. [33] first applied GAN to the task of inpainting and proposed a inpainting algorithm based on Context Encoder (CE). However, this algorithm cannot obtain a significant amount of semantic information from the vicinity of the missing area, resulting in poor image consistency after inpainting. Iizuka et al. [34] used local and global context discriminators to maintain global semantic consistency. But the resulting images were still blurry and weak in detail and texture features. Liu et al. [21] proposed a model with partial convolution, which focus more on obtaining feature information of the complete region of image than vanilla convolution. However, the inpainted image may experience boundary artifacts and local color differences. Yu et al. [35] proposed gated convolution on the basis of partial convolution. This method provides a dynamic feature selection mechanism that can be learned, and constantly optimizes mask parameters to improve mask flexibility. Nevertheless, in cases where the missing area is substantial, excessive smoothness and blurring may occur. Liu et al. [36] proposed a two-stage network with a coherent semantic attention layer. This layer enhances the semantic coherence between the inpainted area and the known area but may not sufficiently address the inpainting of fine texture details. Zhang et al. [37] proposed a generative adversarial image inpainting algorithm with a parallel variable autoencoder with short+long term attention layer called Pluralistic Image Completion (PICNet), which uses it to achieve multi-style image inpainting. However, the correlation between pixels in the inpainted image is poor, and the texture detail features are weak. In 2020, Lahiri et al. [38] proposed a method based on generative adversarial networks, which considers image inpainting problems as searching for the best potential prior, and then uses pre-training generation models to map to natural images. The method uses iterative inference types to accelerate inference speed and improve visual quality.

2.3. Feature Extraction and Feature Fusion

Since the image inpainting often requires deep semantic information, and the combination of low-level and high-level semantic information is also the key to this task. So feature extraction and feature fusion are very important for inpainting.

Shin et al. [39] and Iizuka et al. [34] used dilated convolutions with different sizes of receptive fields in network structures. Dilated convolution can obtain more feature information with the same parameter quantity as ordinary convolution. Liu et al. [40] generated a probability map from a mask graph by constructing a network, and relied on the probability map to determine whether the generated pixels were certain or uncertain. Zeng et al. [41] proposed a pyramid network based on generative adversarial network. This method gradually learns area correlation from high-level semantic feature maps and transfers the learned attention to the previous low-level feature maps. Nevertheless, the aforementioned methods may fall short in fully extracting the deep-level features of the image, which can potentially result in poor structural consistency of the inpainted results.

So as to fully fuse low dimensional features and high dimensional features of images, researchers have proposed many methods. Quan et al. [42] proposed a method of using large receptive field for rough inpainting and small receptive field for fine inpainting, thus completing the inpainting of main structure and texture details. Shen et al. [43] used the joint inpainting method of multiple U-Net networks [2] and added dense connections between different U-Net layers. This method makes the multi-scale spatial location information better preserved and finally shows better results. Zeng et al. [44] proposed a model for high-resolution inpainting to capture information rich context information. This model uses context transformation from different receptive field to finally achieve feature fusion through cascade and vanilla convolution aggregation. However, the above methods fail to fully fuse features from different depths and scales. The computational costs of the model are also high.

2.4. Progressive Inpainting Algorithm

In recent years, more and more multi-stage progressive algorithms have been applied to inpainting tasks. Liao et al. [45], Xiong et al. [46] and Nazeri et al. [47] firstly inpainted the damaged image boundary map, and then generated the final inpainting image using the boundary map. The above methods reduce the difficulty of the inpainting process and improve the inpainting effect. However, the prediction result of image boundary map seriously affects the effect of inpainting. When dealing with images that contain extensive areas of damage, the boundary maps of these regions can be challenging for the network to accurately predict. As a consequence, generating ideal inpainting results becomes more difficult. Yu et al. [35] proposed a coarse-to-fine two-stage inpainting method, which introduces a contextual attention mechanism in order to capture long-distance image information, but may result in distortion. Shin et al. [39] proposed the Diet-PEPSI, which utilizes adaptive dilated convolution to capture global contextual information. However, this method employs a two-stage network approach, which entails a significant number of parameters and entails a substantial computational cost. Li et al. [48] proposed a plug-and-play module called Recurrent Feature Reasoning (RFR), which can reduce the range of areas to be filled layer by layer and achieve the reuse of model parameters. And the addition of knowledge consistent attention (KCA) makes the details of inpainting more refined. Li et al. [49] improved on the basis of partial convolution by gradually interleaving the edge information and filling information of damaged images, and shared parameters to improve the inpainting effect. In the case of extensive damage to an image, the network faces difficulty in capturing the correlation between the inpainted region and distant information. As a consequence, the detailed features of the image are lost in the inpainted results.

3. Method

3.1. Model Structure

In order to inpaint a tongue image with clear texture and continuous semantics, we propose a recurrent tongue image inpainting algorithm based on multi-scale feature fusion. The MSFM-RAM-Net structure is shown in Figure 1. d represents the expansion rate of the convolution and s represents the stride of the convolution kernel.

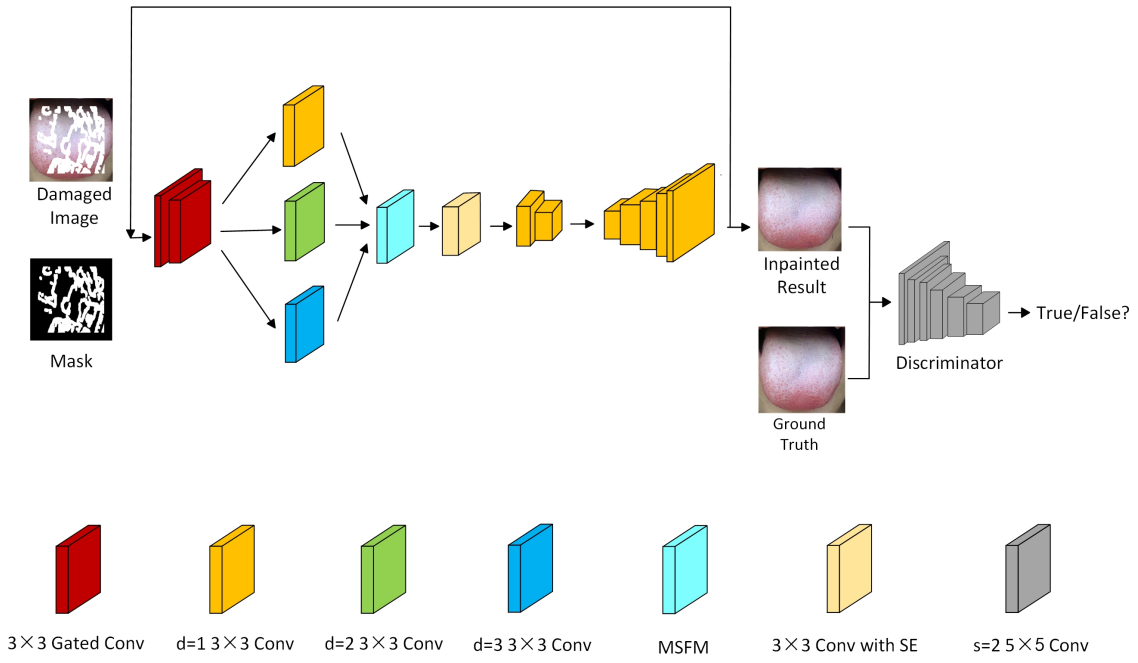

Figure 1. The structure of MSFM-RAM-Net.

In this paper, two layers of gated convolution are used to fully extract the potential information of tongue image. Then three convolution with different dilated rates are used to obtain tongue image features at different scales. We use convolutions with dilated rates of 1, 2, and 3, respectively. A larger dilated rate can effectively increase the receptive field, allowing for better processing of global and local features in the image. However, blindly increasing the dilated rate often leads to redundancy of feature information, which is not conducive to the learning of image features. Afterwards the fusion of these features is achieved through the MSFM module to obtain semantic information combining shallow features and deep features. Subsequently, RAM is used to continue learning the semantic information of the tongue image. There is a Squeeze-and-Excitation module (SE) [50] between the first and second convolutional layers. The SE attention mechanism learns the importance of each channel. It enhances useful feature information, and suppresses the dispensable feature information. Finally, the output feature map is utilized as input for the entire network, allowing for the relearning of image information. The recurrent mechanism can achieve the progressive image inpainting and constantly rely on the known information around the image to fill the missing area.

3.2. MSFM

In order to fuse multi-scale input feature information while reducing the number of parameters,we propose MSFM. This module concatenates the input feature maps and uses channel shuffle to distribute the semantic information of each scale throughout the entire network structure. Following that, group convolution is employed to decrease the parameter count, resulting in a lightweight model that maintains effectiveness during the information fusion stage. Finally fusing the semantic information of each grouping channel through 1×1 convolution operation and 3×3 convolution operation to achieve information exchange between channels. Therefore, MSFM achieves multi-scale fusion and reserves detailed information as much as possible, while reducing the computational complexity of the model and obtaining richer semantic information. The structure of MSFM is shown in Figure 2. W, H and C represent the width, height and number of channels of three input feature maps and one output feature map, respectively.

The network input is composed of three different scales of information, which can be defined as l_i(i = 1, 2, 3) . Firstly, concatenate the three different scales feature maps. Define the concatenation operation C(·) to form a simple fusion information r. The expression is computed as

$$r = C(l_1, l_2, l_3). \tag{1}$$

In order to better achieve the interaction between information on various channels and improve the model's ability to express tongue image information, shuffling the channel of the concatenated feature maps. Channel shuffle can interweave different types of image feature information to achieve information fusion in different scales. However, a single channel shuffle only disrupts the order of features and performs simple feature fusion on the current state of feature information. Therefore, we propose inter-group feature information learning on the feature maps after channel shuffle after channel grouping. For the entire channel convolution, there are drawbacks such as high computational costs and insufficient information fusion. The proposed group convolution presents a reduction in computational complexity for the model, consequently enhancing the execution efficiency of the algorithm. It also provides a prerequisite for subsequent mutual learning of group convolutions. Then, the shuffled feature maps are divided into three groups, and the semantic information of different groups is fused again to obtain the image information. The feature maps of the three groups obtained can be defined as x_i(i = 1, 2, 3). By point-wise addition and the convolution operation of 3×3, the feature maps of different groups are further fused to obtain richer semantic information. The numerical value of each pixel on the feature map is represented as the content information of the current image at that pixel. Adding the information between different groups with point-wise addition, semantic

information on different feature channels on two groups can be obtained simultaneously. Subsequently, 3×3 convolutions are performed on the fused features of the group to extract the corresponding feature information. We process group convolution in three steps, with the first step obtaining the vanilla convolution and the last step fusing all the feature information of the three group convolutions to obtain the most abundant feature maps. The convolution operation is defined as D_j(j = 2, 3) , where y_j(j = 1, 2, 3) represents the output result of each group fusion operation. Therefore, the expression is computed as

$$y_j = \begin{cases} x_j, & j = 1 \\ D_j(x_{j-1} + x_j), & j = 2. \\ D_j(y_{j-1} + x_j), & j = 3 \end{cases} \tag{2}$$

Subsequently, the obtained three feature maps are concatenated together again to form a complete feature map. It can be found that after two concatenated and fusion operations of feature maps, the feature maps of each channel have a closer connection and the output features have semantic information from different receptive fields and depths. This is beneficial for the encoder to fuse information of different scales and depths, combining large receptive fields with small receptive fields. It also ensures strong learning ability for image structures with relatively large missing areas and improves global image consistency. Then, the number of feature channels after concatenation is restored to the number of channels before multi-scale operation through 1×1 convolution operation, which becomes 1/3 of the number of channels after concatenation. Dimensionality reduction of the channel direction achieves feature refusion, enriching the feature map. Finally, the semantic information is further extracted using 3×3 convolution operations to obtain the final multi-scale fused feature map.

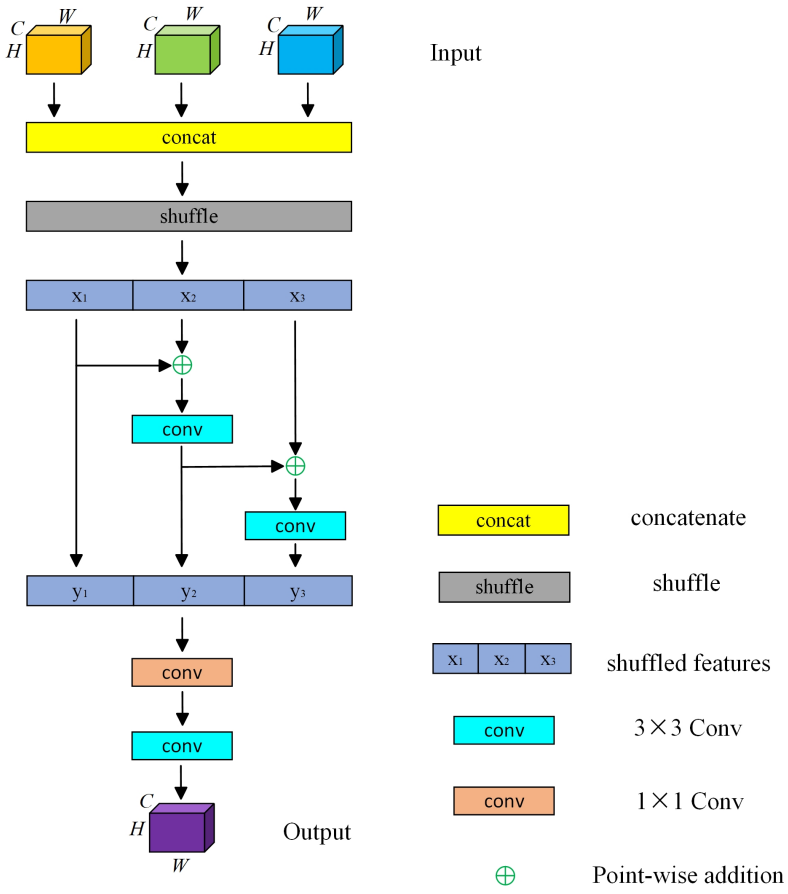

Figure 2. MSFM Structure.

3.3. RAM

Traditional inpainting methods usually complete the image inpainting work at once and have the same attention to information of different importance, which leads to poor effect of image inpainting. To address the above issues, we propose RAM to augment the network's capacity for representing image features. General image inpainting methods usually use the entire image to be inpainted at once, while RAM divides the inpainting process into multiple stages and iterates for inpainting. Each stage makes a partial inpainting of the image and passes the inpainted results to the next stage. RAM can restore the inpainting details gradually, and reduce the impact of error accumulation, making the inpainting effect more accurate. At the same time, RAM adaptively weights different regions according to the content and structure of the image, so that the model pays more attention to important regions and features. This improves the accuracy of the inpainted results and reduces unwanted artifacts and distortions. RAM consists of two parts: recurrent mechanism and SE attention mechanism. The recurrent mechanism obtains semantically continuous inpainting images by relearning feature images, while the SE attention mechanism focuses the network's focus on important features and strengthens the correlation between pixels.

The recurrent mechanism gradually achieves the inpainting of missing images by reusing the output image features as input features of the network. This method performs image inpainting in the feature map space, solving the problem of information distortion caused by repeated mapping between the feature map space and RGB space during the image inpainting. Moreover, the recurrent mechanism reuses parameters, reducing the computational complexity of the model and making the model lightweight. The function can be expressed as

$$O = F(F(\cdots)). \tag{3}$$

O represents the final output result of the network model. F(·) represents the result obtained by the image after a complete network operation.

We only design a simple branch to achieve ideal results. The structure of this model is similar to the residual connection of resnet [51], except that the residual connection avoids network degradation during forward propagation and extracts deeper feature information. The recurrent mechanism can be seen as a reverse connected identity mapping, where the final output feature information is used as input to the network again. While progressive inpainting, it also avoids the problem of feature degradation that may occur in the subsequent inpainting process of the model. This method directly connects the output information with the input information, without adding other additional parameter calculations. Almost all computational costs are generated by the parameters learned by the convolution kernel in the model when extracting image feature information and the operation between the convolution. The recurrent mechanism not only improves the the inpainted results during the progressive image inpainting process, but also improves the ability of the model at the root by combining with the overall model. Instead of the output image, we take the output feature map as the input of the recurrent mechanism. Frequent mapping of images in different information spaces may lead to incorrect image features, ultimately resulting in the inability of the inpainted image to produce the desired results. We control the recurrent mechanism in the feature space of the image, and a single information space will reduce errors during inpainting.

After multi-scale fusion of the feature information of the image through the MSFM module, although the feature information of the image is more abundant at this time, it can simultaneously consider both shallow and deep information, the features expressed on the obtained feature map are confused. At this point, regardless of any information on the image, the attention of the network is consistent. Therefore, we propose the SE attention mechanism during the downsampling process. It can obtain the feature weights of each channel through learning, thereby focusing the network's focus on important parts of the tongue image and suppressing feature information that has little effect on tongue image inpainting. Meanwhile, the attention mechanism simplifies the model structure

and accelerates network operations. The structure of SE attention mechanism is shown in Figure 3. X represents the input feature map. $C^{'}$, $W^{'}$ and $H^{'}$ represent the number of channels, width and height of the feature map, respectively. F_{tr} is a transformation operation, usually a convolution operation. After convolution operation, feature maps U with channel numbers, width and height of C, W and H were obtained. $F_{sq}(\cdot)$ represents the global average pooling operation, which compresses the feature map to generate a vector of $1 \times 1 \times C$. $F_{ex}(\cdot$,W) means to use FC full connection layer, ReLU and Sigmoid activation function to obtain channel weight. $F_{scale}(\cdot,\cdot)$ represents multiplying the generated weights and feature map U element by element to obtain the final feature map $\widetilde{X}$.

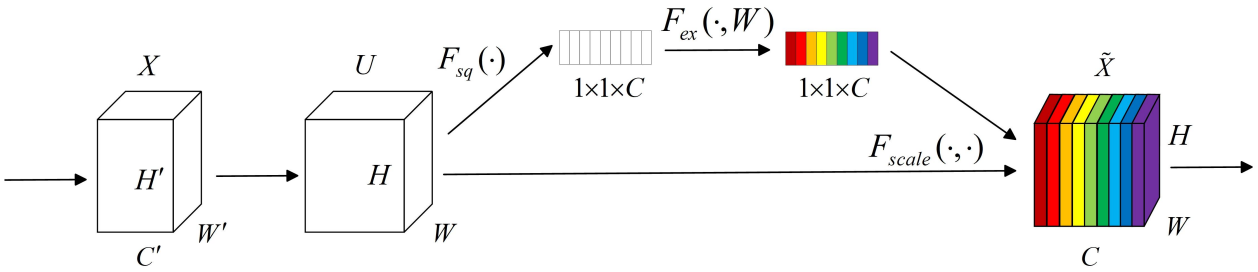

Figure 3. Structure of SE attention mechanism.

At the same time, RAM model structure is also mutually reinforcing within it. The recurrent mechanism takes feature maps that have not been fully learned by the network as input for the entire network to learn image features again. Throughout the entire process, complex and critical features will be continuously learned, improving the efficiency of the SE attention mechanism in the execution process. SE attention mechanism focuses the model's focus on important features of the image. This helps to expedite the convergence of the network and reduce the frequency at which the network learns image features within a recurrent mechanism. In summary, RAM can improve the image inpainting speed of the network while obtaining satisfactory inpainting results.

3.4. Loss Function

In order to ensure the stability of training and inpaint tongue images with clear texture, we use a combination of multiple loss function to determine the effectiveness of image inpainting and optimize the result of the tongue image inpainting.

The real tongue image is represented by x and M represents mask. So tongue image inpainted by the generator can be represented as

$$z = x \odot (1 - M) + G(x, M) \odot M. \tag{4}$$

The reconstruction loss L_{rec} using pixel level L_1 loss represents the difference between real tongue images and inpainted tongue images, with the goal of inpainting the missing area of the image as accurately as possible. By minimizing reconstruction loss, the model can learn how to complete missing pixels or areas and make the inpainting as similar as possible to the original image. The function can be expressed as

$$L_{rec} = \|M \odot (x - G((1 - M) \odot x))\|_1. \tag{5}$$

In order to make the inpainted tongue image approach the real tongue image in perception, we propose perception loss and style loss based on VGG-16 network [52]. perceptual loss is optimized by comparing the differences between images to inpaint the

image and can preserve textures and important features as much as possible for the model. The function can be expressed as

$$L_{per} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|\phi_i(x) - \phi_i(z)\|_1}{H_i \times W_i \times C_i}. \tag{6}$$

$\phi_i(\cdot)$ represents the output feature map of the i-th layer in the VGG network. $H_i \times W_i \times C_i$ represents the size of feature map at the i-th layer.

Style loss defines the distance between image feature Gram matrices, so that the inpainted result more stylistically matches the original image, expressed as

$$L_{style} = \frac{1}{N}\sum_{i=1}^{N}\|GM(x) - GM(z)\|_1. \tag{7}$$

GM$(\cdot)$ represents the calculation of Gram matrix. The function can be expressed as

$$GM(I) = \phi_i^T(I)\phi_i(I). \tag{8}$$

I represents the feature matrix of the image.

Adversarial loss can enable the generator and discriminator to compete with each other, enhancing the authenticity and details of the inpainted image. We use WGAN-GP loss [53] as adversarial loss, expressed as

$$L_{adv} = -D(x). \tag{9}$$

In summary, the specific function for L we propose can be expressed as

$$L = \lambda_{rec}L_{rec} + \lambda_{per}L_{per} + \lambda_{style}L_{style} + \lambda_{adv}L_{adv}. \tag{10}$$

λ_{rec}, λ_{per}, λ_{style} and λ_{adv} represent the weights of reconstruction loss, perception loss, style loss and adversarial loss, respectively. The combined loss has achieved comprehensive optimization of the model from four aspects: pixel, visual, holistic and structural, improving the quality of image inpainting.

4. Experimental Results and Analysis

The experiments are all using the Pytorch deep learning development framework, trained and tested on NVIDIA GeForce RTX 3080 GPU. For the generator and discriminator, we set the initial learning rates to 1×10^{-4} and 4×10^{-4}, respectively. The recurrent number is set to 5 and the batch size is set to 8. We propose to use the Adam algorithm to optimize the model, setting two parameters of the optimizer $\beta_1 = 0.5$ and $\beta_2 = 0.9$. Each weight size of the loss function is set to $\lambda_{rec} = 5$, $\lambda_{per} = 0.05$, $\lambda_{style} = 120$ and $\lambda_{adv} = 0.1$. So as to better evaluate the model, we propose a tongue image dataset. Using the irregular mask dataset with six different mask rates such as $(0.01, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.4]$, $(0.4, 0.5]$ and $(0.5, 0.6]$ proposed by Pconv [21] and central square mask, MSFM-RAM-Net was qualitatively and quantitatively evaluated under the self-built tongue image dataset, and compare with RFR [48], PEPSI++ [39], and PD-GAN [40]. Finally, evaluating the influence of each module through ablation experiments.

4.1. Tongue Image Dataset

Due to the inconvenience of collecting tongue images and the issue of personal privacy, there is currently no standardized dataset for medical tongue images both at home and abroad. Therefore, researchers do not have a universal set of comparative data when comparing models. We establish a complete tongue image dataset based on the needs of tongue image work. This dataset contains a total of 1622 tongue images. Since these tongue images are collected on the Internet and taken under arbitrary conditions using electronic products such as mobile phones and cameras, the size and brightness of the tongue images in this dataset are inconsistent. The tongue images in this dataset are diverse, including

normal and patient tongue images. The shape, size and texture of tongue images in the dataset are different. And for abnormal tongue images, there are also characteristics such as different colors, missing parts of the tongue body and the appearance of other pathological structures. The sample tongue image of this dataset is shown in Figure 4. By using this data for training, it ensures that the network model learns the tongue images' features. It can improve the model's generalization ability for different scenes and image information.

The establishment of a tongue image dataset is also beneficial for future research.

By collecting a sufficient number of tongue images and labeling information including disease type and disease course development, a disease model related to the characteristics of the tongue image can be established. Such datasets can be used for disease diagnosis, prediction and surveillance.

The creation of this new tongue image datasets is essential to drive the application of deep learning and computer vision technologies in this field. This can facilitate the study of automatic analysis of tongue images, feature extraction, disease detection and classification, etc.

Figure 4. Partial tongue images in the tongue image dataset.

4.2. Qualitative Evaluation

To visually demonstrate the inpainting capability of the MSFM-RAM-Net, we conduct experiments on our self-built tongue image dataset. For each experiment, the same original image and mask are used as inputs to inpaint damaged tongue images using the MSFM-RAM-Net, RFR [48], PEPSI++ [39], and PD-GAN [40]. Figures 5 and 6 compare our method with state-of-the-art approaches using self-built dataset under irregular masks and central square mask, respectively.

From Figure 5, it can be seen that when the irregular mask rate is low, all of the above methods achieve good results in inpainted tongue images. However, when the mask rate is large, different algorithms have different results for inpainting. RFR has an overly smooth inpaiting effect to a certain extent, resulting in the loss of some detail features. Therefore, the inpainted image is quite different from the real image. PEPSI++ can generate a more reasonable tongue image structure, but features of redundancy appeared. For example, there is a texture structure in the middle of the tongue image that does not exist in the real tongue image. PD-GAN can inpaint images with large missing areas, but the tongue image still shows incomplete texture details, and the cracks in the tongue groove are not completely inpainted. The MSFM-RAM-Net performs better than other methods in tongue

image inpainting under irregular masks, and can generate more reasonable tongue image structures to ensure the global consistency of tongue images. Visually, the texture details are clearer, greatly improving the inpainted effect of tongue images.

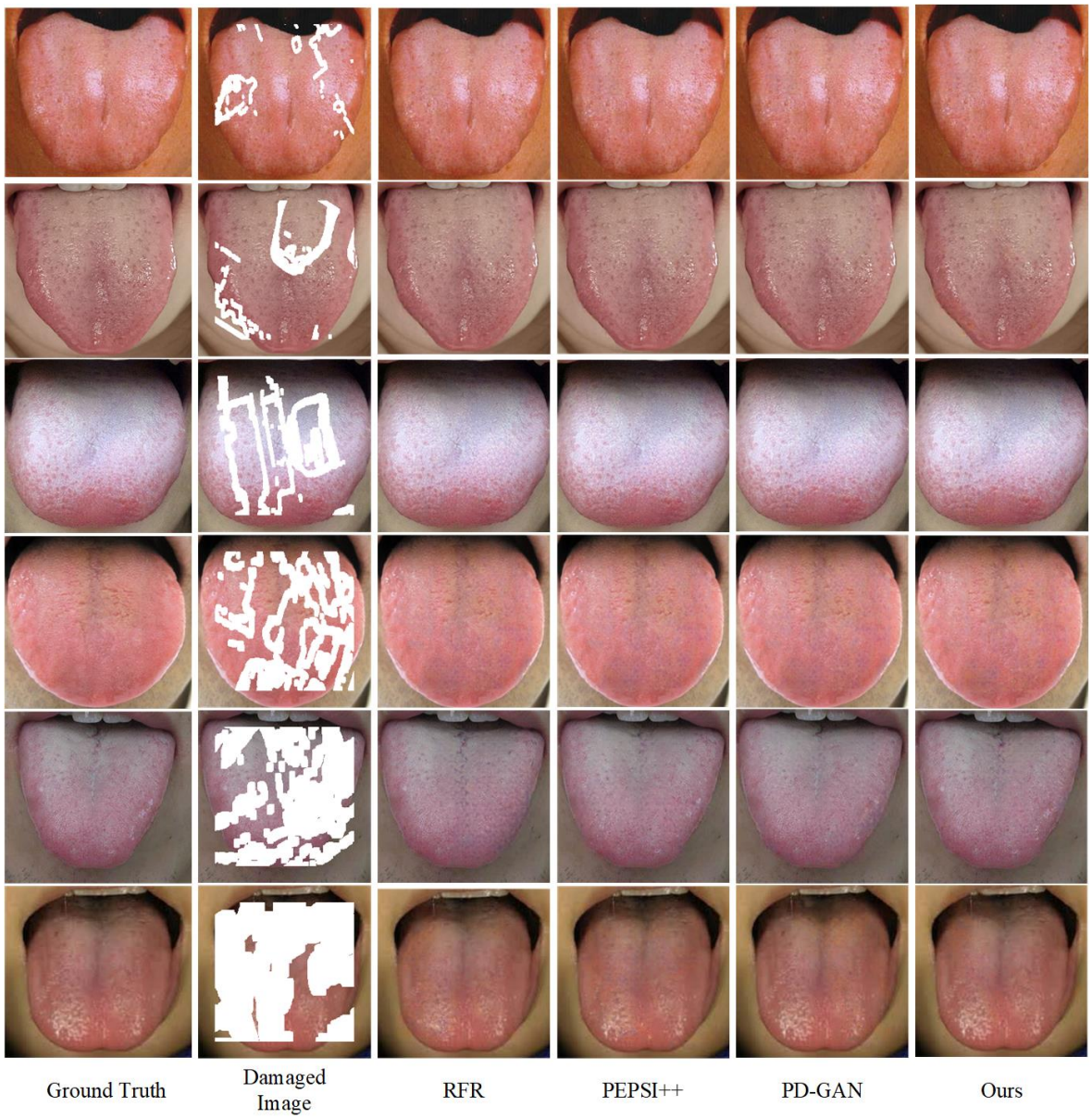

Figure 5. Comparison of inpainting effects under irregular masks.

As depicted in Figure 6, it is evident that for tongue image under the central square mask, RFR results in blurry inpainting boundaries and accompanies by unclear texture structures. PEPSI++ has done a good job in inpainting edge information, but it still lacks detailed information such as texture structure. On the other hand, while PD-GAN is capable of inpainting large damaged areas, it still struggles with unclear textures and inconsistent overall semantics. The comparison demonstrates that the MSFM-RAM-Net exhibits a significantly superior inpainting effect under the central square mask compared to other methods. MSFM-RAM-Net avoids possible boundary blurring and can inpaint images that are more reasonable and have clearer texture details.

Figures 5 and 6 verify that the MSFM-RAM-Net has better inpainting performance under irregular masks and central square mask. This is because on the one hand, the MSFM module fuses the low-level and high-level semantic information of tongue images

and concentrates the texture features of images under different receptive field sizes, so as to improve the detail preservation and visual authenticity of the inpainted image. As the network deepens, there is no loss of low-level features, and texture details can be well preserved. At the same time, the network learns high-level features without losing detailed information. On the other hand, the recurrent attention mechanism focuses the network on important features and strengthens the learning ability of key areas, improving the semantic coherence and authenticity of the inpainted results. By leveraging the edge information of the image, the proposed method progressively enhances the inpainting of tongue images. This approach ensures a more seamless connection between the inpainted area and surrounding pixels, resulting in improved semantic coherence.

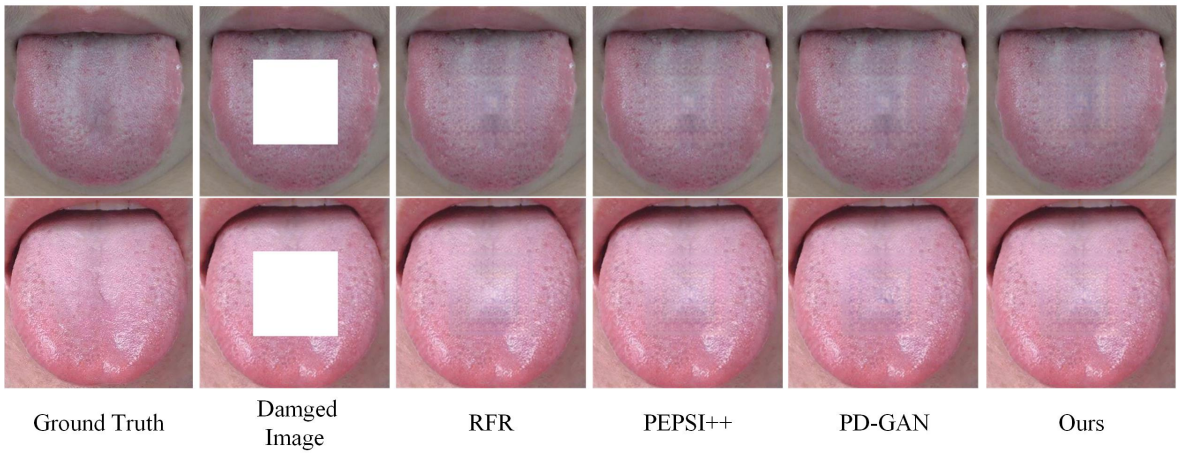

Figure 6. Comparison of inpainting effects under central mask.

4.3. Quantitative Evaluation

To objectively and fairly assess the image inpainting effect, we use Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) and L_1 loss as evaluation indicators. PSNR is utilized for quantifying image distortion, SSIM measures the perceptual similarity between the original image and the inpainted image and indicates the degree of two images, and L_1 represents the difference between pixels. This evaluation standard can be used to evaluate the inpainting results from two dimensions: image features and pixels. The larger the PSNR and SSIM are, the better the inpainting effect is; the smaller the L_1 loss is, the smaller the difference between pixels is.

We uses four methods, including MSFM-RAM-Net, RFR, PEPSI++ and PD-GAN, to inpaint irregular masks and center square masks. The representation of the final results can be observed in both Tables 1 and 2.

From Table 1, it is evident that when the mask rate is small, the structural consistency is not damaged generally, so all methods can achieve good inpainted results. However, when the mask rate continues to increase, the advantages of MSFM-RAM-Net over the other three methods also continue to increase. RFR has insufficient retention of texture features and high computational complexity. For rigorous tongue image medical inpainting tasks, the inpainting of details is insufficient. PEPSI++ can quickly inpaint tongue images due to its lightweight model structure. However, due to this lightweight design, the algorithm lacks sufficient learning ability for complex feature structures. However, the texture structure of tongue images is often complex and diverse, resulting in poor performance in tongue image inpainting. In addition, PSPSI++ requires more training and validation data to ensure its generalization ability under lightweight design. Although PD-GAN can achieve good inpainted results, this method has high computational costs and complexity. Moreover, probabilistic diversity maps may not accurately protect the detailed information of the image, resulting in a lack of realism in the inpainted image and the possibility of some unnatural flaws. The performance of this model in PSNR, SSIM and L_1 loss is significantly

better than the other three models. This is because MSFM-RAM-Net fuses features of different scales, reducing the loss of detailed information. At the same time, the network strengthens the learning of important features and reduces the degree of image distortion. More importantly, the MSFM-RAM-Net has low computational complexity and does not require too much data to achieve good inpainted results.

Table 1. Quantitative comparison of irregular masks with different mask rates.

	Mask Rate	RFR	PEPSI++	PD-GAN	Ours
PSNR	$(0.01, 0.1]$	38.806	38.813	38.972	38.987
	$(0.1, 0.2]$	33.218	33.265	33.920	34.034
	$(0.2, 0.3]$	29.874	29.885	30.014	30.163
	$(0.3, 0.4]$	26.978	26.984	27.165	27.364
	$(0.4, 0.5]$	24.053	24.144	25.370	25.637
	$(0.5, 0.6]$	21.637	21.751	22.688	22.875
SSIM	$(0.01, 0.1]$	0.946	0.948	0.961	0.968
	$(0.1, 0.2]$	0.922	0.926	0.948	0.956
	$(0.2, 0.3]$	0.909	0.906	0.923	0.937
	$(0.3, 0.4]$	0.886	0.882	0.902	0.914
	$(0.4, 0.5]$	0.830	0.827	0.852	0.866
	$(0.5, 0.6]$	0.748	0.743	0.795	0.818
L_1	$(0.01, 0.1]$	0.0050	0.0050	0.0048	0.0046
	$(0.1, 0.2]$	0.0107	0.0106	0.0088	0.0083
	$(0.2, 0.3]$	0.0154	0.0151	0.0132	0.0120
	$(0.3, 0.4]$	0.0272	0.0259	0.0202	0.0185
	$(0.4, 0.5]$	0.0382	0.0378	0.0311	0.0296
	$(0.5, 0.6]$	0.0585	0.0599	0.0470	0.0414

Table 2. Comparison of quantitative evaluation of center square mask.

Model	PSNR	SSIM	L_1
RFR	25.146	0.855	0.0320
PEPSI++	25.314	0.852	0.0316
PD-GAN	26.468	0.875	0.0269
Ours	26.746	0.893	0.0244

From Tables 2, The results clearly demonstrate that the inpainting performance of MSFM-RAM-Net with the central square mask is superior to that of RFR, PEPSI++ and PD-GAN. The central square mask covers the complex part of the middle texture of the tongue image. Large area centralized information loss makes it difficult to obtain effective image known information. Therefore, RFR cannot achieve satisfactory inpainted results with less image information. Although PEPSI++ accelerates the speed of image inpainting, it does not have enough ability to inpaint the tongue images. Due to the lack of large image information in the central area and the lightweight design, the processing ability for complex textures is poor, so the image inpainting effect of this method is still not ideal. PD-GAN uses probabilistic diversity maps to assist with inpainting, generating the final inpainted image. But it fails to effectively grasp the texture detail features and there are still flaws in the detail information. Therefore, for the central large area mask, the ideal result was not achieved in the end. MSFM-RAM-Net is more sensitive to the correlation between pixels and the structure between images. The tongue image features under the central square mask can be gradually inpainted through RAM. The inpainted results have low distortion due to the close relationship between pixels. At the same time, MSFM enables the correlation of image information at various scales of tongue images, improving the structural consistency of inpainted images. Therefore, the pixel information difference of the tongue image inpainted by MSFM-RAM-Net is smaller, and the texture is more delicate.

In summary, MSFM-RAM-Net has better inpainted results compared to RFR, PEPSI++ and PD-GAN. It uses MSFM to fuse features of different receptive field sizes. Meanwhile, it can disrupt the sequence of feature maps in the channel direction, making connections between features with different correlations and enhance the interaction between features. RAM inputs the output feature maps as input into the model, and gradually obtains the image texture results that need to be inpainted using the known surrounding pixel information of the image. At the same time, the MSFM-RAM-Net uses dilated convolution and shared parameters to reduce the number of model parameters and improve the inpainting speed of the network. Tongue image data contains personal privacy and is only collected under specific circumstances, so tongue image data is rare. MSFM-RAM-Net fully captures the low-level and high-level features of tongue images, so it does not require training on a large dataset to achieve good inpainted results. Therefore, we confirm that the MSFM-RAM-Net can achieve better inpainted results in tongue image inpainting.

4.4. Ablation Experiments

This section verifies the significance of the proposed module in tongue image inpainting through ablation experiments. These experiments not only prove the effect of MSFM, but also compares the influence of SE attention mechanism and different recurrent numbers on image inpainted results in RAM module.

In order to prove the effect of MSFM module on image inpainting, we compare the method of directly adding elements by element (DilatedConv) using different expansion rates (expansion rates of 1, 2 and 3, respectively). Figure 7 represents the results of experiments, while Tables 3 and 4 represent quantitative evaluation of the two methods. As shown in Figure 7, DilatedConv cannot ensure the structural consistency of tongue images and loses a lot of texture information. The MSFM module remains more low-level semantic information and makes the texture of the inpainting tongue image clearer and more detailed. From Tables 3 and 4, it is evident that the MSFM module outperforms the approach of directly fusing different scales individually. This is because MSFM enriches the semantic information of output feature maps by densely connecting feature maps of different scales. Meanwhile channel shuffle and information re-fusion between different channel groups enhance information exchange.

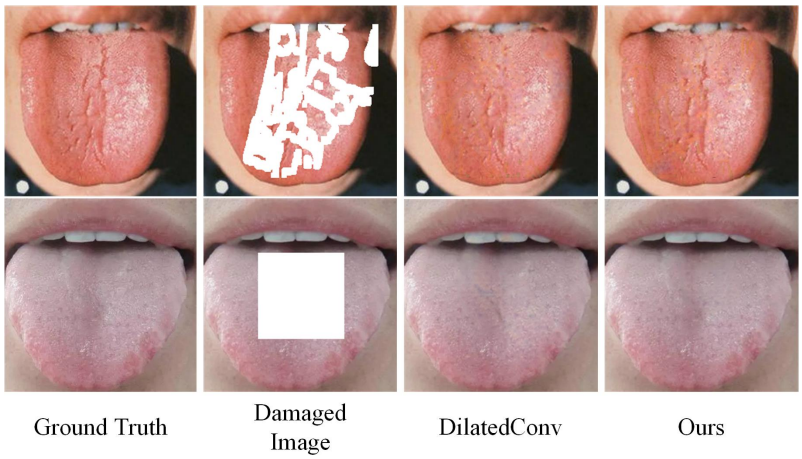

Figure 7. Comparison of inpainting effects of different fusion modules under irregular masks and central square masks.

Table 3. Comparison of inpainting effects of different fusion modules under $(0.4, 0.5]$ mask.

Model	PSNR	SSIM	L_1
DilatedConv	25.108	0.849	0.0335
Ours	25.637	0.866	0.0296

Table 4. Comparison of inpainting effects of different fusion modules under central square mask.

Model	PSNR	SSIM	L_1
DilatedConv	26.231	0.869	0.0280
Ours	26.746	0.893	0.0244

The experimental results of the SE attention mechanism verified at a mask rate of $(0.4, 0.5]$ are shown in Table 5. W/o SE indicates that there is no SE in the model. MSFM-RAM-Net with the SE addition module significantly improves the effectiveness of tongue image inpainting compared to the method without the SE attention module. This is because that the SE attention mechanism readjusts the weight of feature map channels and suppresses areas with little effect on tongue image inpainting. It focuses the attention of the network on important tongue image information, enhancing the model's learning ability, and also improves efficiency of the network.

Table 5. The influence of SE attention mechanism on experimental results.

Model	PSNR	SSIM	L_1
Ours w/o SE	25.493	0.860	0.0303
Ours	25.637	0.866	0.0296

The recurrent mechanism is the process of re-extracting image features. We verify the effectiveness of tongue image inpainting with different recurrent numbers, as shown in Table 6. The model achieved the optimal result when the recurrent number was 5. With the continuous increase in the recurrent number, the performance of the model cannot be improved and reaches saturation. This is because that sufficient recurrent numbers have obtained the most effective shallow and deep features, and increasing the numbers again will only gain redundant features.

Table 6. Comparison of experimental effects with different recurrent numbers.

Nums	PSNR	SSIM
4	25.376	0.858
5	25.637	0.866
6	25.631	0.863
7	25.635	0.862

5. Discussion

In this paper, we propose MSFM-RAM-Net, a new image inpainting algorithm that combines MSFM and RAM for progressive image inpainting. The experimental results demonstrate the effectiveness of MSFM-RAM-Net from several aspects.

First, MSFM allows the model to efficiently capture low-level features and high-level features during the inpainting process. By fusing features at multiple scales, MSFM-RAM-Net can better preserve fine textured details while maintaining the overall consistency of the inpainted image.

In addition, RAM is able to progressively complete the missing areas, which helps to reduce structural distortion and artifacts in the final inpainted image. By starting with small, simple areas and gradually scaling up to larger and more complex areas, MSFM-RAM-Net achieves better structural consistency and overall visual quality compared to other methods.

The MSFM-RAM-Net not only shows excellent performance in the field of tongue image inpainting, but also has potential application value and expansion space.

In the clinical environment, the MSFM-RAM-Net algorithm can be widely used in the field of medical image processing. For example, in the processing of CT or MRI images, the

algorithm can be used to inpaint distorted or missing areas of the image, improving image quality and allowing doctors to better diagnose and treat diseases.

In other image processing fields, the MSFM-RAM-Net algorithm also has a wide range of application potential. For example, in natural scene image processing, this algorithm can be used to inpaint missing image areas and corrupted image details, thereby improving image quality.

In summary, the MSFM-RAM-Net algorithm has a wide range of application prospects, and has important application value and expansion space in medical image processing and other fields. In the future, the potential application of this algorithm in other image processing fields can be further explored, and the development and application scope of image processing technology can be promoted.

However, we also need to recognize that there are some limitations to the algorithm. First of all, the main limitation is to deal with large and complex inpainting scenarios, such as large damage to the tongue. These textures and structures to be inpainted are very complex. In this case, it may be difficult for MSFM-RAM-Net to fully capture and inpaint fine details. Secondly, the algorithm may have limitations on the inpainting of tongue images with extreme forms or abnormal situations, resulting in a decrease in the inpainting effect. In addition, the algorithm has limited ability to deal with noise, and may not be able to completely eliminate the influence of noise on the inpainted result.

To overcome these limitations, the following areas of improvement can be considered. First, future research may focus on developing more sophisticated techniques to solve these challenging inpainting tasks. Second, further research on how to deal with tongue images with extreme morphology or anomalies can improve the inpainted results by introducing richer datasets or using more complex models. In addition, for the noise problem, more advanced noise models can be used to improve the robustness of the algorithm.

In conclusion, the MSFM-RAM-Net algorithm proposed by us proves its superiority in multi-scale feature fusion and progressive inpainting framework. It shows good results in tongue image inpainting tasks. Despite of this, further improvements are needed to address challenging inpainting scenarios and improve the algorithm's performance in handling complex textures and structures.

6. Conclusions

In this study, we propose a Multi-Scale Fusion Module and Recurrent Attention Mechanism Network (MSFM-RAM-Net). MSFM fuses information streams of different scales together, making the feature map contain richer semantic features. At the same time, this model increases the interaction between features and greatly preserves the low-level features that are left during continuous learning. Afterwards, RAM is proposed. This method focuses the features of the network on the key parts of the image and uses the progressive method to achieve the re-learning of the output information to improve the model inpainting effect. Finally, we conduct qualitative and quantitative evaluations of MSFM-RAM-Net on the self-developed dataset. Experiments have shown that MSFM-RAM-Net outperforms other existing models. Especially when the mask rate is high, the PSNR is 22.875, SSIM is 0.818, and L_1 loss is 0.0414, which is a significant improvement compared to existing models and can generate more delicate textures and more reasonable structures.

Author Contributions: Conceptualization, W.W. and T.L.; Methodology, W.W., T.L. and A.T.; Software, W.W., T.L. and Z.L.; Validation, T.L., A.T. and Z.L.; Formal analysis, W.W. and Z.L.; Investigation, W.W., A.T. and Z.L.; Resources, W.W. and T.L.; Data curation, W.W., T.L. and K.S.; Writing—original draft preparation, W.W. and A.T.; Writing—review and editing, T.L. and Z.L.; Visualization, A.T. and K.S.; Supervision, W.W., T.L. and K.S.; Project administration, W.W., T.L. and A.T.; Funding acquisition, A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Researchers Supporting Project Number (RSPD2023R681), King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Ning, Z.; Dong, P.; Wang, X.; Hu, X.; Guo, L.; Hu, B.; Guo, Y.; Qiu, T.; Kwok, R.Y.K. Mobile Edge Computing Enabled 5G Health Monitoring for Internet of Medical Things: A Decentralized Game Theoretic Approach. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 463–478. [CrossRef]
2. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Cham, Switzerland, 5–9 October 2015; pp. 234–241.
3. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5686–5696.
4. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1302–1310.
5. Ning, Z.; Huang, J.; Wang, X.; Rodrigues, J.J.P.C.; Guo, L. Mobile Edge Computing-Enabled Internet of Vehicles: Toward Energy-Efficient Scheduling. *IEEE Netw.* **2019**, *33*, 198–205. [CrossRef]
6. Ning, Z.; Zhang, K.; Wang, X.; Obaidat, M.S.; Guo, L.; Hu, X.; Hu, B.; Guo, Y.; Sadoun, B.; Kwok, R.Y.K. Joint Computing and Caching in 5G-Envisioned Internet of Vehicles: A Deep Reinforcement Learning-Based Traffic Control System. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 5201–5212. [CrossRef]
7. Li, Z.; Yu, Z.; Liu, W.; Zhang, Z. Tongue Image Segmentation via Color Decomposition and Thresholding. In Proceedings of the 2017 4th International Conference on Information Science and Control Engineering (ICISCE), Changsha, China, 21–23 July 2017; pp. 752–755.
8. Huang, C.W.; Chen, Y.J.; Yen, T.T.; Lin, K.Y.; Chen, D.Y. Region-based hierarchical tongue feature extraction. In Proceedings of the 2014 International Conference on Machine Learning and Cybernetics, Lanzhou, China, 13–16 July 2014; pp. 867–870.
9. Fu, S.; Zheng, H.; Yang, Z.; Yan, B.; Su, H.; Liu, Y. Computerized tongue coating nature diagnosis using convolutional neural network. In Proceedings of the 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, 16–17 November 2017; pp. 730–734.
10. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image Inpainting. In Proceedings of the Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New York, NY, USA, 1 July 2000; pp. 417–424.
11. Efros, A.A.; Freeman, W.T. Image Quilting for Texture Synthesis and Transfer. In Proceedings of the Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, New York, NY, USA, 1 August 2001; pp. 341–346.
12. Ballester, C.; Bertalmio, M.; Caselles, V.; Sapiro, G.; Verdera, J. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Process.* **2001**, *10*, 1200–1211. [CrossRef] [PubMed]
13. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D.B. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. *ACM Trans. Graph.* **2009**, *28*, 1–11. [CrossRef]
14. Simakov, D.; Caspi, Y.; Shechtman, E.; Irani, M. Summarizing visual data using bidirectional similarity. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
15. Darabi, S.; Shechtman, E.; Barnes, C.; Goldman, D.B.; Sen, P. Image Melding: Combining Inconsistent Images Using Patch-Based Synthesis. *ACM Trans. Graph.* **2012**, *31*, 1–10. [CrossRef]
16. Ning, Z.; Hu, H.; Wang, X.; Guo, L.; Guo, S.; Wang, G.; Gao, X. Mobile Edge Computing and Machine Learning in The Internet of Unmanned Aerial Vehicles: A Survey. *ACM Comput. Surv.* **2023**. [CrossRef]
17. Ning, Z.; Dong, P.; Kong, X.; Xia, F. A Cooperative Partial Computation Offloading Scheme for Mobile Edge Computing Enabled Internet of Things. *IEEE Internet Things J.* **2019**, *6*, 4804–4814. [CrossRef]
18. Wang, X.; Ning, Z.; Guo, L.; Guo, S.; Gao, X.; Wang, G. Mean-Field Learning for Edge Computing in Mobile Blockchain Networks. *IEEE Trans. Mob. Comput.* **2022**, pp. 1–17. [CrossRef]
19. Xie, J.; Xu, L.; Chen, E. Image Denoising and Inpainting with Deep Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Stateline, NV, USA, 3–8 December 2012; pp. 1–9.
20. Eigen, D.; Krishnan, D.; Fergus, R. Restoring an Image Taken through a Window Covered with Dirt or Rain. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 633–640.
21. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image Inpainting for Irregular Holes Using Partial Convolutions. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 1–23.
22. Li, H.; Luo, W.; Huang, J. Localization of Diffusion-Based Inpainting in Digital Images. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 3050–3064. [CrossRef]

23. Liu, Y.; Yan, H.; Liu, Q.; Zhang, W.; Huang, J. ECO++: Adaptive deep feature fusion target tracking method in complex scene. *Digital Communications and Networks* **2022**, pp. 1–16. [CrossRef]
24. Ma, R.; Zhang, Z.; Ma, Y.; Hu, X.; Ngai, E.C.; Leung, V.C. An improved pulse coupled neural networks model for semantic IoT. *Digit. Commun. Netw.* 2023, *in press*. [CrossRef]
25. Ning, Z.; Chen, H.; Ngai, E.C.H.; Wang, X.; Guo, L.; Liu, J. Lightweight Imitation Learning for Real-Time Cooperative Service Migration. *IEEE Trans. Mob. Comput.* **2023**, pp. 1–18. [CrossRef]
26. Ning, Z.; Zhang, K.; Wang, X.; Guo, L.; Hu, X.; Huang, J.; Hu, B.; Kwok, R.Y.K. Intelligent Edge Computing in Internet of Vehicles: A Joint Computation Offloading and Caching Solution. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 2212–2225. [CrossRef]
27. Liu, J.; Yang, S.; Fang, Y.; Guo, Z. Structure-Guided Image Inpainting Using Homography Transformation. *IEEE Trans. Multimed.* **2018**, *20*, 3252–3265. [CrossRef]
28. Ning, Z.; Sun, S.; Wang, X.; Guo, L.; Guo, S.; Hu, X.; Hu, B.; Kwok, R.Y.K. Blockchain-Enabled Intelligent Transportation Systems: A Distributed Crowdsensing Framework. *IEEE Trans. Mob. Comput.* **2022**, *21*, 4201–4217. [CrossRef]
29. Wang, X.; Ning, Z.; Guo, S.; Wen, M.; Guo, L.; Poor, H.V. Dynamic UAV Deployment for Differentiated Services: A Multi-Agent Imitation Learning Based Approach. *IEEE Trans. Mob. Comput.* **2023**, *22*, 2131–2146. [CrossRef]
30. Shixin, P.; Kai, C.; Tian, T.; Jingying, C. An autoencoder-based feature level fusion for speech emotion recognition. *Digital Commun. Netw.* **2022**, pp. 1–14. [CrossRef]
31. Ning, Z.; Yang, Y.; Wang, X.; Song, Q.; Guo, L.; Jamalipour, A. Multi-Agent Deep Reinforcement Learning Based UAV Trajectory Optimization for Differentiated Services. *IEEE Trans. Mob. Comput.* **2023**, pp. 1–17. [CrossRef]
32. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
33. Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
34. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and Locally Consistent Image Completion. *ACM Trans. Graph.* **2017**, *36*. [CrossRef]
35. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T. Free-Form Image Inpainting With Gated Convolution. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4470–4479.
36. Liu, H.; Jiang, B.; Xiao, Y.; Yang, C. Coherent Semantic Attention for Image Inpainting. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4169–4178.
37. Zheng, C.; Cham, T.J.; Cai, J. Pluralistic Image Completion. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1438–1447.
38. Lahiri, A.; Jain, A.K.; Agrawal, S.; Mitra, P.; Biswas, P.K. Prior Guided GAN Based Semantic Inpainting. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13693–13702.
39. Shin, Y.G.; Sagong, M.C.; Yeo, Y.J.; Kim, S.W.; Ko, S.J. PEPSI++: Fast and Lightweight Network for Image Inpainting. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 252–265. [CrossRef] [PubMed]
40. Liu, H.; Wan, Z.; Huang, W.; Song, Y.; Han, X.; Liao, J. PD-GAN: Probabilistic Diverse GAN for Image Inpainting. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9367–9376.
41. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1486–1494.
42. Quan, W.; Zhang, R.; Zhang, Y.; Li, Z.; Wang, J.; Yan, D.M. Image Inpainting With Local and Global Refinement. *IEEE Trans. Image Process.* **2022**, *31*, 2405–2420. [CrossRef] [PubMed]
43. Shen, L.; Hong, R.; Zhang, H.; Zhang, H.; Wang, M. Single-Shot Semantic Image Inpainting with Densely Connected Generative Networks. In Proceedings of the Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1861–1869.
44. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Aggregated Contextual Transformations for High-Resolution Image Inpainting. *IEEE Trans. Vis. Comput. Graph.* **2023**, *29*, 3266–3280. [CrossRef] [PubMed]
45. Liao, L.; Hu, R.; Xiao, J.; Wang, Z. Edge-Aware Context Encoder for Image Inpainting. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 3156–3160.
46. Xiong, W.; Yu, J.; Lin, Z.; Yang, J.; Lu, X.; Barnes, C.; Luo, J. Foreground-Aware Image Inpainting. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5833–5841.
47. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3265–3274.

48. Li, J.; Wang, N.; Zhang, L.; Du, B.; Tao, D. Recurrent Feature Reasoning for Image Inpainting. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 7757–7765.
49. Li, J.; He, F.; Zhang, L.; Du, B.; Tao, D. Progressive Reconstruction of Visual Structure for Image Inpainting. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5961–5970.
50. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
52. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
53. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved Training of Wasserstein GANs. In Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5769–5779.